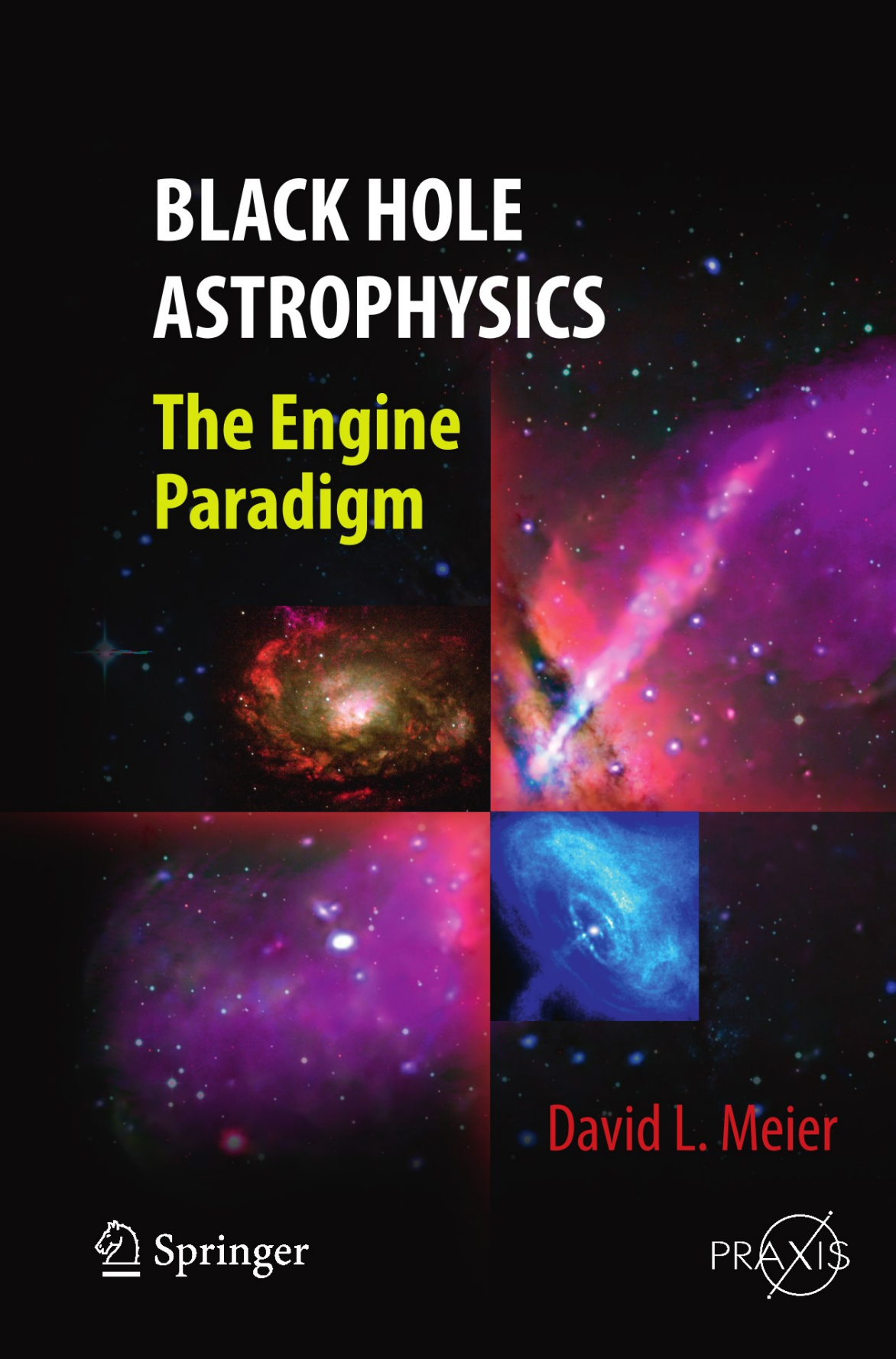


BLACK HOLE ASTROPHYSICS

The Engine Paradigm



David L. Meier

 Springer

PRAXIS

Black Hole Astrophysics

The Engine Paradigm

David L. Meier

Black Hole Astrophysics

The Engine Paradigm

 **Springer**

Published in association with
Praxis Publishing
Chichester, UK

PRAXIS 

Dr. David L. Meier
California Institute of Technology/Jet Propulsion Laboratory
Pasadena
California
U.S.A.

SPRINGER-PRAXIS BOOKS IN ASTRONOMY AND PLANETARY SCIENCES
SUBJECT ADVISORY EDITOR: John Mason, M.B.E., B.Sc., M.Sc., Ph.D.

ISBN 978-3-642-01935-7 ISBN 978-3-642-01936-4 (eBook)
DOI 10.1007/978-3-642-01936-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012943939

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Cover design: Jim Wilkie
Project copy editor: Mike Shardlow
Author-generated LaTeX, processed by EDV-Beratung, Germany

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my grandfather, Otto E. Hennig,
who showed a young ten-year-old the night sky
and all its wonders, right from his back garden.
May I be half the grandfather to my grandchildren
that he was to me.

Contents

Preface xix

List of Figures xxiii

List of Tables xxxi

Part I Observations of Black Hole Engines

1 Recognizing Black Holes 3

 1.1 Energy Output 5

 1.2 Compactness 6

 1.3 Engine Mass..... 7

 1.3.1 Binary X-Ray Star Masses 7

 1.3.2 Masses of Black Holes in the Centers of Stellar Systems ... 7

 1.3.3 Masses from Microlensing of Isolated Black Holes in the Galaxy 10

 1.4 Relativistic Motion 11

2 Macroquasars: Supermassive Black Holes in the Centers of Galaxies 13

 2.1 The Early Days of Active Galactic Nuclei Research and Seyfert Galaxies 14

 2.2 Radio Galaxies and Classical Quasars: Supermassive Black Holes with Jets 17

 2.2.1 Extended Radio Galaxies 17

 2.2.2 Extended Radio QSRs: Unification by Brightness Contrast . 23

 2.2.3 Compact Radio QSRs: Unification with Extended QSRs by Viewing Angle 25

 2.2.4 Cosmic Evolution of Radio Galaxies and QSRs 35

 2.3 Seyfert Galaxies and Quasi-Stellar Objects: Supermassive Black Holes with Weak or No Jets 39

 2.3.1 Unification of Classical Seyfert Galaxies by Viewing Angle 39

 2.3.2 The X-Ray Spectrum of Seyfert Nuclei 39

2.3.3	Narrow-Line Seyfert 1 Galaxies	42
2.3.4	Quasi-Stellar Objects	42
2.3.5	Cosmic Evolution of Seyferts and QSOs	52
2.4	Low-Luminosity Active Galactic Nuclei (LLAGN)	52
2.4.1	Dwarf Seyferts, LINERs, Transition-type, and H II Nuclei ..	52
2.4.2	Radio Cores	53
2.4.3	Weak-Lined Radio Galaxies	53
2.4.4	Sgr A*: The Quiescent Black Hole Engine at the Center of Our Galaxy	54
2.5	“Inactive” Galactic Nuclei	55
2.5.1	The $M_{\bullet}$ - M_{bulge} Relation	56
2.5.2	The $M_{\bullet}$ - σ_V Relation	57
2.5.3	The $M_{\bullet}$ - r_{core} Relation	57
2.6	Macroquasars – A Summary and Synthesis	59
3	Microquasars: Black Holes (and Neutron Stars) of Stellar Mass in Our Galaxy	63
3.1	Isolated Neutron Stars: Pulsars and Magnetars	65
3.1.1	“Normal” Pulsars	65
3.1.2	Magnetars	70
3.1.3	RRATs	71
3.1.4	Geminga and Other Nearby Neutron Stars	72
3.2	Neutron Stars with Companions: Neutron Stars in Binaries	73
3.2.1	Neutron Star X-Ray Binaries	73
3.2.2	The Strange Case of SS433	81
3.2.3	Recycled Pulsars	83
3.2.4	Binary Pulsars	84
3.3	Neutron Stars in Formation: Supernovae	84
3.4	Isolated Stellar-mass Black Holes: Twinkling of a Little Star	87
3.5	Stellar-mass Black Holes with Companions: Black Hole Binary Systems	88
3.5.1	Black Hole X-Ray Binaries	88
3.5.2	Classical Microquasars	92
3.5.3	The Microquasar Explosion and the Relative and Absolute Numbers of Neutron Stars and Black Holes	99
3.6	Stellar-mass Black Holes in Formation: Gamma-Ray Bursts	101
3.6.1	The Long and Short of Gamma-Ray Bursts	101
3.6.2	Long-duration Gamma-Ray Bursts	102
3.6.3	Short-duration Gamma-Ray Bursts: The Neutron Star Merger Model	109
3.6.4	Other Types of Gamma-Ray Bursts	111
3.6.5	The End of the World as We Know It?	112

4	Miniquasars: Elusive Black Holes of Intermediate Mass	115
4.1	Definition of an Intermediate Mass Black Hole (IMBH)	116
4.2	First Type of IMBH Candidate: Ultra-Luminous X-Ray Sources (ULXs) in Star-Forming Galaxies	116
4.2.1	Observational Properties of ULXs in Star-Forming Galaxies	117
4.2.2	Theoretical Reasons for Expecting IMBHs in Regions of Rapid Star Formation	118
4.2.3	Temperature Problems: Black Hole Fever	119
4.2.4	Better ULX Spectral Fits	120
4.2.5	ULXs in Elliptical Galaxies	120
4.3	Second Type of IMBH Candidate: Kinematic and Dynamical Evidence for IMBHs in Globular Cluster Cores in Our Galaxy and the Andromeda Galaxy	121
4.3.1	Motivation	121
4.3.2	Observational Evidence	121
4.3.3	Weighty Problems	123
4.4	Third Type of IMBH Candidate: A Possible IMBH Near the Center of the Milky Way	124
4.4.1	The Galactic Center in the Near-Infrared	125
4.4.2	The Need for a Second Black Hole at the Center of Our Galaxy	125
4.4.3	The Discovery of a Possible Second Black Hole in the Center of Our Galaxy	126
4.5	Fourth Type of IMBH Candidate: Low-Luminosity Hard X-Ray/Radio Sources in the Galactic Bulge and Other Population II Systems	126
4.5.1	Hard X-Ray Sources in the Galactic Bulge	126
4.5.2	Suggested Radio Searches for Low-Luminosity Miniquasars	128
4.6	Summary and Synthesis	128

Part II Physics of Black Hole Engines

5	Physics in Everyday Life: A Review of Newtonian Mechanics	133
5.1	Mechanics of Single Particles: Planetary Systems and Binary Stars	133
5.1.1	Newton's Law of Conservation of Momentum	133
5.1.2	Newton's Law of Gravity	135
5.1.3	General Binary Star Motion in the Center-of-Mass Frame of Reference	136
5.1.4	Motion of a Test Particle in a Gravitational Field	137
5.1.5	Motion of Two Comparable Masses in Each Other's Gravitational Field: Orbits of Binary Planets and Stars	140
5.2	Mechanics of a System of Particles: Fluid Dynamics and the Internal Structure of Stars	143
5.2.1	The Gravitational Force Produced by a Fluid – Inside and Out	144

5.2.2	Newtonian Conservation Laws for a Fluid in the Lagrangian Frame	147
5.2.3	Newtonian Conservation Laws for a Fluid in the Inertial Eulerian Frame	155
5.2.4	Simple Polytropic Stellar Structure	156
5.3	Summary	163
6	Geometry and Physics without Gravity: Special Relativity	165
6.1	Why Geometry?	165
6.2	Two-Dimensional Pythagorean Geometry	166
6.2.1	The Line Element	166
6.2.2	Line Element Has the Same Length in Any Coordinates	167
6.2.3	Matrix Form for the Geometry Equations	169
6.2.4	A Word of Warning: Global vs. Local Transformations	174
6.3	Three-Dimensional Euclidean Geometry	175
6.4	Four-Dimensional Minkowski Geometry	177
6.4.1	The Four-Dimensional Minkowski Line Element	178
6.4.2	Time Dilation	179
6.4.3	The FitzGerald Contraction	179
6.4.4	Minkowski Metric is the Same for All Travelers	180
6.4.5	Working with Minkowski Geometry: Spacetime Diagrams and Light Cones	181
6.4.6	The Meaning of the Proper Distance Δs : Proper Time	182
6.5	Mechanics in Four Dimensions	183
6.5.1	Mechanics of Single Particles in Four Dimensions	183
6.5.2	Fluid Mechanics in Four Dimensions	188
6.5.3	The Doppler Effect, the Doppler Factor, and “Beaming”	188
6.6	Electricity and Magnetism in Four Dimensions	191
6.6.1	Review of Electricity and Magnetism in the 3+1 Form	191
6.6.2	Electricity and Magnetism in Four-Dimensional, Covariant Form	200
6.6.3	Summary of Section 6.6	207
7	Physics in Curved Spacetime: General Relativity and Black Holes	209
7.1	The Clash between Special Relativity and Newtonian Gravity, and its Resolution	210
7.2	Curved Space and the Riemann Curvature Tensor	212
7.2.1	Simple Examples of Curved, and Not Curved, Space	212
7.2.2	Testing for Curvature Using the Riemann Tensor	213
7.3	A Quick Road to Einstein’s Theory of Gravity	217
7.3.1	Determining the Form of the Gravitational Field Equations $\mathcal{G}$	217
7.3.2	Determining the Constant of Proportionality K	219
7.3.3	The Meaning of Gauge in Einstein’s Theory of Gravity	220
7.3.4	Curvature without Local Matter	221
7.3.5	E & M in a General Four-Dimensional Curved Space	221

7.4	Four-Dimensional Schwarzschild Geometry: Non-Rotating Black Holes	222
7.4.1	The Schwarzschild Metric and Schwarzschild Radius	222
7.4.2	Physics in Curved Schwarzschild Geometry	224
7.4.3	Motion in the Schwarzschild Metric	229
7.4.4	Only Schwarzschild–Hilbert Coordinates are Singular at the Horizon, Not the Schwarzschild Metric	232
7.5	Four-Dimensional Kerr Geometry: Rotating Black Holes	234
7.5.1	The Kerr Metric in Boyer–Lindquist Coordinates	234
7.5.2	The Kerr Metric in Horizon–Penetrating Coordinates	242
7.6	Four-Dimensional Kerr–Newman Geometry: Rotating, Charged Black Holes	244
7.7	A Return to Physics in Three-Space Plus Time: Saying Goodbye to the Four-Dimensional Formalism	245
7.7.1	The General Non-Stationary “3+1” Metric	247
7.7.2	Electromagnetism in a Stationary “3+1” Metric	249
7.7.3	Return to Orthonormal Vectors in Stationary “3+1” Metrics: Saying Goodbye to Contravariant Vectors and 1-Forms	250
8	Four-Dimensional Evolving Geometry: Gravitational Waves and Gravitational Collapse	253
8.1	Plane Gravitational Waves	254
8.1.1	Basic Physics of Gravitational Waves	254
8.1.2	Plane Gravitational Waves in a Vacuum	256
8.1.3	Plane Gravitational Waves Generated by a Time-Varying Stress-Energy Distribution	260
8.1.4	Example of a Gravitational Wave Source: A Close Binary System of Point Masses	263
8.2	Nonlinear Gravitational Wave Sources and Numerical Relativity: The Merger of Black Hole Binary Systems	269
8.2.1	The Different Phases of a Binary Merger and Methods Used to Study Them	269
8.2.2	Simulating Colliding Black Holes on Supercomputers	270
8.2.3	Determining the Physics of Black Hole Systems by Comparing Simulations and Observations	274
8.2.4	Astrophysics of Merging Black Holes from Gravitational Wave Simulations: Black Hole Kick Velocities	275
8.3	Basic Gravitational Collapse: Formation of a Black Hole Horizon in Collapsing Matter	277
8.3.1	The Equations for the Structure and Evolution of Spherical, Adiabatic, Relativistic Stars and Gas Clouds	278
8.3.2	Collapse of Pressure-Free Dust to a Black Hole	281
8.3.3	More Realistic Collapse: Adiabatic Stars and Clouds with Pressure	286

9	Nuts and Bolts of the Black Hole Engine: General Relativistic Mechanics	291
9.1	Overview of General Relativistic Mechanics	292
9.1.1	Quantum Mechanics	292
9.1.2	Particle Mechanics	293
9.1.3	Statistical Mechanics	294
9.1.4	Kinetic Theory	295
9.1.5	General Relativistic (Electro-)Magnetohydrodynamics	297
9.2	The Conservation Laws of Relativistic Magnetohydrodynamics	298
9.2.1	Laws of Conservation of Rest Mass and Energy-Momentum	298
9.2.2	Laws of Conservation of Charge and Current	306
9.3	The Equations of State	312
9.3.1	Pressure and Internal Energy of Thermal Gases	313
9.3.2	Pressure and Internal Energy of Nonthermal Gases	320
9.3.3	Thermal Conductivity	322
9.3.4	Particle Viscosity	323
9.3.5	Turbulent Viscosity	324
9.3.6	Radiative Opacity	325
9.3.7	Radiative Heat Transport vs. Thermal Conduction	328
9.4	Optically Thin Radiative Emission	328
9.4.1	Bremsstrahlung (Free-Free Emission)	329
9.4.2	Synchrotron	330
9.4.3	Comptonization	331
9.5	Useful Sets of Magnetohydrodynamic Equations for Solving Black Hole Astrophysical Problems	335
9.5.1	Adiabatic Ideal Magnetohydrodynamics	335
9.5.2	Standard Force-Free Electrodynamics	340
9.5.3	An Alternate Form for FFDE: GRMHD with No Material Forces	346
9.5.4	General Relativistic Hydrodynamics (GRHD)	348
9.5.5	Special Relativistic Magnetohydrodynamics (SRMHD)	348
9.5.6	Stationary, Axisymmetric SRMHD in Newtonian Gravity	349
9.5.7	Non-Relativistic Dynamics	352
9.6	Waves and Instabilities in the Fluid Plasma	355
9.6.1	General Features of Wave Analysis	356
9.6.2	Non-Damped Waves	358
9.6.3	Instabilities in Non-Relativistic, Shearing HD and MHD Flow	370

Part III Astrophysics of Black Hole Engines

10	Assembling the Engine Block: Formation of Black Holes in the Universe	381
10.1	The Formation of Neutron Stars and the Synthesis of Heavy Elements	382

10.1.1	Big Bang Nucleosynthesis	382
10.1.2	Stellar Evolution Nucleosynthesis for Different Star Masses	383
10.2	Formation of Stellar-mass Black Holes in <i>Single</i> Massive Stars	394
10.2.1	Mass Loss and the Production of Wolf–Rayet Stars	395
10.2.2	Current Best Theoretical Estimates for Black Hole Formation	397
10.3	Formation of Neutron Stars and Stellar-mass Black Holes in <i>Binary</i> Star Systems	398
10.3.1	Formation of High-Mass X-Ray Binaries and Binary Relativistic Stars	398
10.3.2	Formation of Close Binary Systems that May Form Neutron Stars and Black Holes	402
10.3.3	Formation of Low-Mass X-Ray Binaries and Microquasars	403
10.3.4	Formation of Neutron Stars and Black Holes via Accretion and Mergers in Close Binary Systems	403
10.3.5	Cataclysmic Variable Binaries: Possibly Type Ia <i>Carbon-Detonation</i> Supernovae, Definitely not Type Ib/c Core-Collapse Supernovae	405
10.3.6	Double-Degenerate Binaries: Some Type Ia Supernovae, an Occasional Neutron Star, and Many Low-Mass Helium Stars	407
10.4	Formation of Intermediate Mass Black Holes	408
10.4.1	Formation of IMBHs in the First Stars Created after the Big Bang	410
10.4.2	Formation of IMBHs in Dense Star Clusters in the Present Universe	415
10.5	Formation of Supermassive Black Holes at the Centers of Galaxies	423
10.5.1	Broad Observational Constraints on SMBH Formation	424
10.5.2	The Current, Sketchy View of Galaxy and SMBH Formation	426
10.5.3	Unsolved Problems and the Possible Role of IMBHs Formed in Compact Star Clusters	431
10.5.4	The Author’s Personal Perspective on Formation of SMBHs	434
11	Fueling and Carburetion: Breakup and Accretion of Stars and Gas Clouds	437
11.1	Accretion of Matter	437
11.1.1	Accretion Basics	438
11.1.2	Generic Bondi–Hoyle–Lyttleton Accretion: Black Holes in the ISM	439
11.1.3	Accretion of Stars onto Massive Black Holes in AGN and Globular Clusters	442
11.1.4	Accretion of Interstellar Gas onto Massive Black Holes in AGN: Tidal Disruption of Newborn Star Clusters	451
11.1.5	Accretion in Stellar-mass Binary Systems	470
11.1.6	Accretion in Collapsing Supernova Cores	475
11.2	Accretion of Angular Momentum: Spinup of the Engine Flywheel	476

11.2.1	Angular Momentum Accretion in Binary Systems	476
11.2.2	Angular Momentum Accretion onto Massive Black Holes in AGN and Globular Clusters	477
11.2.3	Angular Momentum Accretion in Collapsed Supernova Cores	480
12	The Combustion Chamber: Energy Generation by Gravitational Accretion	483
12.1	Spherical Accretion and Winds	485
12.1.1	Physical Structure	485
12.1.2	Thermal Structure and Accretion Luminosity: The Trapping Radius	490
12.1.3	Bondi Accretion vs. Begelman Accretion	491
12.2	Classical Turbulent Accretion Disk Theory	492
12.2.1	Physical Structure	493
12.2.2	Thermal Structure and Continuum Emission Spectrum	498
12.3	Magnetically-Dominated Accretion Disk Theory	530
12.3.1	The Case for Strong Magnetic Fields in the Center of Low-Accretion Rate Flow	530
12.3.2	Magnetically-Advective (Transitional) Accretion Flow	532
12.3.3	Magnetically-Dominated Accretion Flows (MDAFs)	535
12.4	Magnetohydrodynamic Numerical Simulations of Accretion Flows .	547
12.4.1	The Current State of MRI Simulations	547
12.4.2	Local MRI Simulations	549
12.4.3	Semi-Local MRI Simulations	552
12.4.4	Global MRI Simulations	555
12.5	Chapter Summary	560
12.5.1	What We Know	560
12.5.2	What We Do Not Know	562
13	The Thermal Exhaust System: Radiation- and Thermally-Driven Winds and Jets	563
13.1	Radiation-Driven Winds	564
13.1.1	Line-Driven, Sub-Eddington Winds	564
13.1.2	Continuum-Driven, Super-Eddington Winds	567
13.2	Thermally-Driven, ADIOS Winds: A Glimpse into the Disk–Wind Interaction of All Advection-Dominated Accretion Flows	582
13.2.1	ADIOS Basics and Motivation	583
13.2.2	ADIOS Equations	583
13.2.3	ADIOS Models	585
13.3	Thermally-Driven Jets	586
13.3.1	Physical Structure	587
13.3.2	Plasma Conditions, Synchrotron Spectrum, and Comparison with Data	589

14 The Non-Thermal Exhaust System I. Rotating Magnetospheres that Drive the Turbo Exhaust	591
14.1 Isolated Pulsars: Rotating Neutron Star Magnetospheres with No Accretion	592
14.1.1 Pulsar Basics: Dipole Radiation, Spindown Power, and Light Cylinder Radius	592
14.1.2 The Basic Force-Free Pulsar Magnetosphere	594
14.2 Pulsars in Close Binary Systems: Rotating Neutron Star Magnetospheres with Accretion	604
14.2.1 Plasma Flow onto Pulsars in the Accretion Regime	606
14.2.2 Plasma Flow onto Pulsars in the “Propeller” or Wind Turbine Regime	608
14.3 “Isolated” Black Holes with Magnetospheres	609
14.3.1 No Strong Magnetic Fields around “Isolated” Black Holes	610
14.3.2 Singular Surfaces of the Grad–Schlüter–Shafranov Equation	611
14.3.3 The Basic Force-Free Black Hole Magnetosphere	614
14.3.4 Numerical Models of Force-Free Black Hole Magnetospheres	617
14.4 Black Hole Magnetospheres in the Presence of Accretion Flow	622
14.4.1 Origin and Generic Structure of the Black Hole Magnetic Field	622
14.4.2 Structure of a Magnetosphere with an Accretion Disk that Does <i>Not</i> Reach the Black Hole	625
14.4.3 Structure of a Rotating Black Hole Magnetosphere with an Accretion Disk that Reaches into the Ergosphere	635
14.4.4 Relativistic Plasma Outflow from a Pure Split Monopole Magnetosphere	641
14.5 Jet Launching from Accretion Disks Only	642
14.5.1 Magnetocentrifugal (Alfvén) Launching	642
14.5.2 Magnetic Pressure (Fast-Mode) and Gas Pressure (Slow-Mode) Launching	645
14.5.3 The Magnetic Tower Mechanism: Fast-Mode Launching with Closed Field Lines	646
15 The Non-Thermal Exhaust System II. Magnetic Winds and Jets	655
15.1 Magnetized Wind and Jet Theory: The Formation of Jets	655
15.1.1 Non-Relativistic, Cold MHD Wind and Jet Theory	656
15.1.2 Relativistic, Cold MHD Wind and Jet Theory	672
15.1.3 Non-Relativistic, Warm MHD Wind and Jet Theory	678
15.1.4 Relativistic, Warm MHD Wind and Jet Theory	692
15.2 Beyond the Magnetosonic Horizon: Propagation of Ballistic Jets	700
15.2.1 Hydrodynamic Jets: A Model for FR II Sources	701
15.2.2 Magnetohydrodynamic Jets: A Model for Perhaps Many FR I Sources	711

15.2.3 Possible Scenarios for Behavior of Jets beyond the Magnetosonic Horizon	724
16 Putting it All Together: Black Hole Engines of All Sizes	727
16.1 Neutron Stars: $1 M_{\odot}$ Near-Black Holes	727
16.1.1 Neutron Star Formation	727
16.1.2 Isolated Neutron Stars	731
16.1.3 Accreting Neutron Stars in Binary Systems	735
16.2 Stellar-mass ($10-100 M_{\odot}$) Black Holes: Rosetta Stones for Deciphering the Ultimate Engines	743
16.2.1 Stellar-mass Black Hole Formation	743
16.2.2 Black Hole Astrophysics in a Nutshell	746
16.2.3 Isolated Stellar-mass Black Holes Accreting from the Interstellar Medium	760
16.2.4 Stellar-mass Black Holes Accreting in Binary Systems	762
16.3 Intermediate Mass ($10^2-10^4 M_{\odot}$) Black Holes: What They Might Look Like and How to Find Them	784
16.3.1 Formation of IMBHs	784
16.3.2 Accretion-Powered IMBHs in the Local Universe	786
16.4 Massive ($10^5-10^7 M_{\odot}$) Black Holes in Spiral Bulges, Pseudobulges, and Dwarf Ellipticals	789
16.4.1 A Proposed Model for the Central Engines of Seyfert 1s and Quasars, from First Principles	791
16.4.2 Important Clues to How Massive Black Holes Form and Grow	801
16.4.3 Formation, Growth, and Fueling of MBHs in Mergers	802
16.4.4 Formation, Growth, and Fueling of MBHs <i>without</i> Mergers .	807
16.5 Supermassive ($10^8-10^{10} M_{\odot}$) Black Holes: The Ultimate Engines .	810
16.5.1 Formation, Growth, and Fueling of SMBHs	811
16.5.2 Growth of the Elliptical Stellar Bulge	813
16.5.3 Jet Production and Propagation in the Mighty Quasars	815
16.6 Summary: Black Holes in the Universe	823
16.6.1 Neutron Stars	826
16.6.2 Stellar-mass Black Holes	826
16.6.3 Intermediate Mass Black Holes	828
16.6.4 Massive Black Holes	828
16.6.5 Supermassive Black Holes	829

Appendices

A Mathematical Notation Used in this Book	833
A.1 Vector and Tensor Notation for Two- and Three-Dimensional Spaces .	833
A.1.1 Two- and Three-Dimensional Vector and 1-Form Notation . .	833
A.1.2 Tensor Notation	834
A.2 Vector and Tensor Notation for Four-Dimensional Spacetime	834

A.2.1	Vector and 1-Form Notation in Four-Dimensional Spacetime	834
A.2.2	Tensor Notation in Four-Dimensional Spacetime	835
A.3	Miscellaneous Notation	835
B	Derivatives of Vectors and Tensors: Differential Geometry	837
B.1	Covariant Gradients in Curved Spacetime	838
B.2	Divergences in Curved Spacetime	839
B.3	The Metric Has No Gradient or Divergence	840
C	Derivation of the Adiabatic Relativistic Stellar Structure Equations	841
C.1	The Spherical Metric in Mass Coordinates	841
C.2	The Field Equations and Conservation Laws	842
C.3	The Adiabatic, Relativistic Stellar Evolution Equations	844
C.3.1	The Mass Shell Geometric Factor	844
C.3.2	The Density Equation	844
C.3.3	Conservation of Energy Equation	845
C.3.4	Equation of Motion in Mass Coordinates	845
C.3.5	Equation of Motion in Schwarzschild–Hilbert-like Coordinates	846
D	Derivation of the General Relativistic MHD Equations from Kinetic Theory	847
D.1	The Multi-Fluid Equations of General Relativistic Magnetohydrodynamics	847
D.1.1	The Zeroth Moment: Conservation of Particle Number	847
D.1.2	The First Moment: Conservation of Particle Four-Momentum	848
D.2	The One-Fluid Equations of General Relativistic Magnetohydrodynamics	850
D.2.1	Conservation of Rest Mass and Four-Momentum	850
D.2.2	Conservation of Charge and Four-Current	851
E	Derivation of the General Relativistic Grad–Schlüter–Shafranov Equation	855
E.1	The Magnetic Induction Equation	855
E.2	The Electric and Magnetic Field Equations	856
E.3	The Charge and Current Densities	857
E.4	The GSS Equation	858
F	Derivation of the Equations for Stationary, Axisymmetric Ideal SRMHD in Newtonian Gravity	859
F.1	The Axisymmetric, Stationary Equation(s) Parallel to the Magnetic Field	859
F.1.1	Faraday’s and Ohm’s Laws and Conservation of Mass: The Frozen-in Magnetic Field	860
F.1.2	Conservation of Specific Angular Momentum	862
F.1.3	Conservation of Specific Entropy	863

F.1.4	Conservation of Specific Energy	863
F.2	The Axisymmetric, Stationary Equation(s) Normal to the Magnetic Field	864
G	Physical and Astrophysical Constants Used in this Book	865
	References	867
	Glossary	883
	Index of Names	891
	Subject Index	897

Preface

The first time I heard the word “engine” applied to a black hole system was in a talk at the Santa Cruz Workshop on Quasars and Active Galaxies in the summer of 1978. The late Donald Osterbrock, then at UCSC, was explaining how the ionizing ultraviolet continuum from a “central engine” was responsible for producing the strong broad and narrow emission lines observed in the optical and ultraviolet spectra of active galactic nuclei. The only earlier such reference that I have been able to find is in an article by Franco Pacini in another conference proceedings that same year. In those days the term was considered little more than a metaphor. It described the mysterious black box at the center of a radio galaxy or quasar that produced all the energy we observed, but whose inner workings were not understood. One of the purposes of this book is to show how prescient this term was and how close to a standard internal combustion engine black hole systems really are.

The primary purpose of this book, however, is to discuss astrophysical black hole systems of all known masses as a single class of objects, governed by the same basic physics and exhibiting the same basic behavior. This text is divided into three parts: (I) observations, (II) physics, and (III) astrophysics of black hole engines. Here “physics” refers to how a completely isolated black hole is structured, while “astrophysics” refers to how black hole engines interact with their environment.

Despite its attempt to be comprehensive (which it does not quite achieve), the book endeavors to be as simple and straightforward as possible in all the discussions. In general, an expert in a particular subfield of black hole research may be disappointed that his or her subfield is treated with insufficient depth. This is intentional in a book of this breadth. The hope, instead, is that most aspects of black hole research are represented and discussed with sufficient depth that the reader will be able to consult more specialized texts and papers for more detailed information. Here the focus is on first demonstrating and explaining the basic concepts; computational tools also are provided, but there are no worked problems in this edition.

There are many omissions, however – all involving the generation of photon spectra in a variety of astrophysical situations and comparison of that theory with observed data. In essence, this is a book on black hole engine “interiors”, as opposed to the usual kind on black hole engine “atmospheres” (to draw an analogy with

stellar astrophysics). A few of the books that *do* cover this area are: in the radio regime, Pacholczyk's classic book [1] and the late Dave De Young's more recent one [2]; in the infrared (IR), optical, and ultraviolet (UV), Donald Osterbrock and Gary Ferland's book [3] on AGN spectra; in the higher-energy parts of the spectrum (particularly X-rays), Rybicki and Lightman's book on radiative processes [4] is always a very good source. Finally, Julian Krolik's book on AGN [5] not only covers many of these areas, it also is a good all-around reference on all aspects of AGN, as is the Lewin and van der Klis compendium [6] for X-ray binaries.

Part I of this book begins in a deceptively simple manner with a short chapter on how to recognize black holes by various means. It is then followed by two chapters on the most studied classes of black hole systems – the supermassive systems in the centers of galaxies, with black hole masses of millions to billions times the sun's mass, and X-ray binary stars with engine masses in the range $1\text{--}100\text{ }M_{\odot}$. Very little is known about objects that have masses intermediate between these two extremes, so discussion in Chapter 4 centers on attempts to find them, conducted in the past and in the future.

Part II can be considered a tutorial on the fundamental parts of a black hole engine and how they fit together. Chapter 5 on non-relativistic physics is essential for understanding the basics of double and multiple star systems, as well as fluids in both dynamical (flowing) situations and static structures like stars. Chapters 6 and 7 introduce relativity in, hopefully, one of the simplest ways published so far. Every effort has been made to avoid tensor manipulation and as much of the mathematics of curved space as possible. Chapter 7 includes a discussion of rotating *vs.* non-rotating black holes, with all the formulae and parameters that we will need in later chapters. Chapter 8 discusses situations (gravitational waves, gravitational collapse) where spacetime is not static, as in a black hole, but rather changes with time. I am particularly pleased that I was able to cast the derivation of *gravitational* waves in a manner that parallels the derivation of electromagnetic waves in Chapter 6. Chapter 9 on relativistic magnetohydrodynamics (MHD) is, by far, the most difficult of all the chapters. Yet, this material is as necessary for understanding cosmic jets as nuclear physics is for understanding the evolution of stars. I also am pleased that I was able to construct equations (9.100) to (9.117) which are a combination of Professor Koide's very useful "conservative 3+1" form with Professor Komissarov's useful notation that makes the equations look as much like regular non-relativistic MHD and regular electromagnetism as possible. The chapters in Part II are supported by even more mathematical derivations in the appendices, in which only the most dedicated theorist might be interested.

Part III exploits fully the engine paradigm, with chapters on how engine blocks are built (Chapter 10), the fuel and its carburetion (Chapter 11), the accretion disk combustion chamber with three of its exhaust systems (angular momentum transport outward, radiation, and heat and matter loss down the black hole; Chapter 12), the wind thermal exhaust system (Chapter 13), and, finally, the nonthermal jet exhaust system (Chapters 14 and 15). The astrophysics in these chapters is presented either in simple algebraic equations that can be used by readers for computations or in images that illustrate the complicated physical processes being described.

Chapter 16 is a synthesis of the entire field, as the author sees it, and is similar in style to an Annual Reviews article. Some caution should be exercised when reading this last chapter. As with all astrophysicists working on black hole systems, our goal here is to understand objects of all masses as a single class, in both how they are similar and how they differ. Our goal here also is to see how far we can get in this endeavor using just the physics and astrophysics that we know right now and that we presented in the earlier chapters in the book. The answer is that we can get surprisingly far – more so than I had anticipated when this book was begun. This exercise was aided, if made more difficult, by the large number of constraints provided by trying to understand stellar, massive, and supermassive black holes with the same theoretical concepts. The first five sections of Chapter 16 present the model in some detail. The final section of the book summarizes it, warns of its pitfalls, and lists the issues that remain great mysteries.

There are many people whom I would like to thank for their assistance during this project: first and foremost my wife Ann, who has stood by me during these ten years; her support and patience has been amazing. I am grateful, as well, to Dr. Paul Maher, whose angioplastic skill is largely responsible for this project not ending prematurely. I also would like to thank the large number of people in the JPL/Caltech community and elsewhere who have volunteered to read and comment on various chapters: Marshall Cohen, Curt Cutler, David Garofalo, Shinji Koide, Dayton Jones, Oliver King, Walter Max-Moerbeck, David Murphy, Masanori Nakamura, Patrick Ogle, Vasiliki Pavlidou, Tim Pearson, Joey Richards, Jackie Villadsen, and especially Tony Readhead, who organized a reading course at Caltech based on this book. I also am grateful to my editor Clive Horwood at Springer-Praxis for his continued patience and understanding during these many years of missed deadlines, to Ramon Khanna at Springer who suggested the title, to Kip Thorne who graciously approved of that title, and to Rashid Sunyaev for his encouragement. I am grateful as well to the Jet Propulsion Laboratory for its Senior Research Scientist program, which offers periodic sabbaticals; this, more than anything else, was the chief reason I was able to finish this large project. Finally, I am grateful to The University of Texas Department of Astronomy and my three graduate advisors there (David Schramm, Beatrice Tinsley, and Craig Wheeler) as well as to the California Institute of Technology and the Institute of Astronomy in Cambridge, UK, and my three postdoctoral advisors at those places (Kip Thorne, Roger Blandford, and Martin Rees). The breadth of their knowledge in black hole physics, stellar evolution, galaxy evolution, and cosmology built the foundations of my interest in black hole astrophysics that became this book.

This book was typeset by the author in $\text{\LaTeX}$, using the $\text{\TeX}$ Shop interface. Figures without explicit credits in their captions were drawn by the author.

David L. Meier
Pasadena, CA,
April 2012

List of Figures

1.1	Orbits of stars around our Galaxy's central black hole	8
1.2	The discovery of the central black hole in Messier 87	9
1.3	Discovery of a possible isolated black hole using microlensing	11
2.1	Series of exposures of Seyfert 1 galaxy NGC 4151	14
2.2	Elliptical galaxy M87 and its jet	15
2.3	Early aperture synthesis "map" of radio galaxy Cygnus A	18
2.4	Three images of the radio galaxy NGC 6251	18
2.5	Modern radio image of radio galaxy Cygnus A	19
2.6	Typical radio synchrotron spectrum	20
2.7	Typical examples of Fanaroff and Riley Class I and Class II sources	21
2.8	Superluminal expansion of the jet in the compact radio quasar 3C 273	26
2.9	The Conway–Murphy helical jet model for misalignments of jets in quasars	31
2.10	Broad-band spectra of several different kinds of blazars	33
2.11	Fossati's unification diagram for blazar spectra	34
2.12	Antonucci and Miller model of the Seyfert sequence	38
2.13	Two Hubble views of dusty tori	40
2.14	Example of an objective prism / grism field	44
2.15	Two determinations of the QSO radio-to-optical distribution, with different results	48
2.16	Typical spectrum of a BAL QSO with "attached" absorption troughs and the standard BAL QSO model	50
2.17	A peculiar BAL QSO and an alternative bipolar wind model	51
2.18	The black hole mass vs. galaxy bulge luminosity and velocity dispersion relations	56
2.19	The black hole mass vs. galaxy core radius relation	58
2.20	Two schemes for unifying macroquasars	60
3.1	The period–period-derivative diagram for Galactic disk pulsars	67

3.2	Distribution of pulsars in Galactic coordinates	68
3.3	The only two pulsars with measured proper motions and observed jets	69
3.4	X-ray color–color diagrams of a typical Z and atoll source	74
3.5	Cartoons depicting various types of low-mass X-ray binaries	75
3.6	Cartoons depicting different methods of mass transfer in HMXB systems	76
3.7	Production of an X-ray nova by a LMXB	77
3.8	X-ray spectra of rapidly-accreting neutron stars and black holes	78
3.9	Full X-ray spectrum of atoll source 4U 1728-34 in the low/hard state	79
3.10	Schematic diagrams of neutron star and black hole accretion states	80
3.11	The story of SS433	82
3.12	Spectra of the accreting black hole Cygnus X-1	89
3.13	Time evolution of the first superluminal microquasar discovered	93
3.14	The Fender–Belloni–Gallo X-ray color–magnitude	96
3.15	Schematic depiction of the short excursion cycle in the “turtle” eye	98
3.16	γ -ray light curves of a LGRB and SGRB	102
3.17	Viewing a spherical, relativistically-expanding fireball as it decelerates	103
3.18	Light curves of different types of relativistic fireballs	104
3.19	Schematic diagram of the “collapsar” model for a LGRB	106
4.1	The only globular-like objects known to contain central black holes	122
4.2	The 19 original Galactic center infrared sources (IRSs)	124
5.1	Diagram of an unequal mass binary star system in a circular orbit	141
5.2	Cartoon of a fluid mass and a small mass element	147
5.3	Mass shells and coordinates for a spherical star	153
5.4	Density structure comparison of four polytropes	159
6.1	A simple example of a line element	167
6.2	Infinitesimal line element remains unchanged during coordinate transformations	168
6.3	Limitations of local differential coordinate transformations	174
6.4	The Pythagorean theorem and line element in three dimensions	176
6.5	Light cone in Minkowski metrics	181
6.6	Plot of a relativistic jet’s Doppler factor against viewing angle	190
6.7	The mixed polarization case for a plane electromagnetic wave	196
7.1	Specific energy and angular momentum vs. radial coordinate r for a non-rotating black hole	231
7.2	Various important equatorial radii vs. black hole spin	243
8.1	Demonstration of the two gravitational wave polarizations, $+$ and $\times$	258
8.2	The “chirp” gravitational wave signal from an inspiraling binary	265

8.3	Capabilities of several current and planned gravitational wave detectors	267
8.4	Schematic diagram of the three phases of black hole mergers	269
8.5	First numerical simulations of gravitational waves generated by two merging black holes	273
8.6	Orbital evolution of three different binary systems with spinning black holes	274
8.7	Comparison of gravitational waveforms from the simulations in Fig. 8.6	275
8.8	Collapse of a uniform density, pressureless star	284
8.9	Stellar collapse with a realistic equation of state	286
8.10	State of the interior of collapsed star in Fig. 8.9	288
9.1	Path through different fields of physics to derive MHD equations . .	293
9.2	Comparison of different thermal particle distributions	315
9.3	The n - T plane for most equations of state discussed herein	321
9.4	Plot of a sample plasma cooling function	332
9.5	Comptonized spectra for different Compton depths	334
9.6	Demonstration of the difference between phase and group velocity .	357
9.7	Phase polar diagrams for HD and MHD waves	360
9.8	Group polar diagrams for HD and MHD waves	360
9.9	Moving polar diagram, showing how characteristics are generated . .	362
9.10	Characteristic construction from a group polar diagram	367
9.11	Linearly accelerating flow along a magnetic field and the characteristics that are generated	369
9.12	Cartoon demonstration of the MRI	375
10.1	Final stellar and remnant mass vs. initial mass for stars with solar-level metallicity	396
10.2	Evolution of high-mass binary stars	399
10.3	Formation of close binary star systems that can lead to neutron stars or black holes	401
10.4	General relativistic evolution of the merger of two neutron stars . . .	405
10.5	Final stellar and remnant mass vs. initial mass for stars with zero metallicity	409
10.6	Possible IMFs for different star populations	412
10.7	Profile of a King isothermal sphere star cluster model	416
10.8	Evolution of a King model core radius during core collapse	418
11.1	Schematic pictures of stationary and moving accreting stars	441
11.2	Accretion of gas into the central 1 kpc in a spiral disk galaxy	453
11.3	Dynamical friction infall of H I clumps and the creation of a star-formation ring	462
11.4	Inspiral of star clusters via dynamical friction	465
11.5	Two examples of Roche equipotentials in the binary orbital plane . .	471

12.1	Solutions of the Parker wind/accretion equation	487
12.2	Solutions to the relativistic spherical wind/accretion equation	489
12.3	Four different views of the same accretion disk around a $10 M_{\odot}$ black hole	505
12.4	The Σ - $\dot{m}$ plane at a disk radius of $R = 10 r_g$	506
12.5	Example of both secular and thermal instabilities in the “inner” region	509
12.6	Two different views of the same advective accretion flow around a $10 M_{\odot}$ black hole	520
12.7	Comparable to Fig. 12.4, but computed using more accurate numerical models	526
12.8	Schematic representation of different observed accretion states	528
12.9	Structure of a magnetically-dominated accretion flow (MDAF) interior to an ADAF	536
12.10	ADAF vs. MDAF	546
12.11	Two-dimensional, axisymmetric simulations of MHD turbulence generated by the MRI	550
12.12	Three-dimensional simulations of MHD turbulence generated by the MRI	552
12.13	Turbulent power spectra of magnetic and kinetic energy	553
12.14	Vertical structure of a semi-local radiation-pressure-dominated accretion disk simulation	554
12.15	Turbulence development in a three-dimensional MRI simulation of accretion flow	556
12.16	Results of simulation in Fig. 12.15, averaged over the tenth orbit	557
12.17	Effects of adding radiative losses to a global MRI simulation	559
13.1	Example of a super-Eddington wind	569
13.2	Physical structure of different super-Eddington wind cases (A, B)	576
13.2	Physical structure of different super-Eddington wind cases (C, D)	577
13.3	Two possible wind scenarios for the super-Eddington accretion case	581
13.4	The three-dimensional ADIOS wind parameter space	585
13.5	Schematic diagram of the Markoff thermal jet model	587
14.1	Goldreich–Julian cartoon drawing of a pulsar magnetosphere	596
14.2	Comparison between specifying and solving for the current distribution in a pulsar magnetosphere	597
14.3	Development of a steady state, aligned pulsar magnetosphere during a time-dependent FFDE simulation	600
14.4	Structure and angular dependence of oblique rotating pulsars	601
14.5	Primary and secondary gaps in the pulsar magnetosphere	603
14.6	Effect of an intruding accretion disk on an aligned rotating pulsar	604
14.7	An accreting pulsar in the accretion regime	605
14.8	Pulsar accretion flow and impact near the magnetic poles	606

14.9	Accretion structure below and above the threshold for unstable inflow	607
14.10	An accreting pulsar in the propeller regime	608
14.11	MHD accretion onto rotating black holes, as viewed from the north polar axis	613
14.12	The “split monopole” black hole magnetosphere model	615
14.13	Breakdown of the current sheet when there is no equatorial plasma disk	617
14.14	Stabilization of the current sheet when an equatorial plasma disk exists	618
14.15	Wald vacuum solution for a black hole in a uniform magnetic field	619
14.16	Charge- and current-filled solution for a black hole in a uniform magnetic field	620
14.17	The two light cylinders for a rotating Kerr black hole	621
14.18	Hirose’s extension of Blandford’s magnetic field line classification	624
14.19	Vacuum solution for black hole magnetosphere with a non-rotating equatorial plasma disk	626
14.20	Similar to Fig. 14.19, but with an externally applied magnetic field	627
14.21	Charge- and current-filled solution for black hole magnetosphere with a Keplerian disk	628
14.22	General relativistic MHD simulation of black hole with a magnetized Keplerian disk	629
14.23	Charge- and current-filled solutions for Kerr black hole magnetospheres vs. spin parameter	631
14.24	Open field rotating black hole magnetospheres with Keplerian disk	632
14.25	Electromagnetic jet-production power of a black-hole-launched jet vs. black hole spin	634
14.26	Current and magnetic structure of an ergospheric wind around a rotating black hole	637
14.27	GRMHD simulation of a rotating black hole with a Keplerian disk	637
14.28	GRMHD simulation of the development of an ergospheric wind and jet	638
14.29	Two views of the final state in Fig. 14.28D	639
14.30	Disappearance of negative energy torus and ergospheric wind in the steady state	640
14.31	MHD simulation of an extreme Kerr black hole with initially radial field	641
14.32	Contours of effective Roche potential for a point mass in the poloidal plane	644
14.33	Simulation of the sweeping pinch jet-launching mechanism	644
14.34	Magnetic field line evolution in a jet launched from a thin disk	645
14.35	Cartoon of the development of a magnetic tower jet	647
14.36	First simulations of a magnetic tower and the “magnetic switch” effect	649
14.37	Evolution of a field loop in the magnetic tower process	651

14.38	Demonstration of the robustness of the magnetic tower	651
14.39	Formation of a magnetic tower by a rotating black hole and disk . . .	652
14.40	Field lines from four different regions in an accreting black hole simulation	654
15.1	Basic parts of the Blandford–Payne jet acceleration and collimation process	657
15.2	A semi-analytic solution of the BP wind equation	663
15.3	Magnetic field/stream lines for a steady, non-self-similar solution to the BP MHD wind	664
15.4	Magnetic field/stream line plots for three Contopoulos and Lovelace models	665
15.5	Electromagnetic jet-production power of a disk-launched jet vs. black hole spin	671
15.6	Solutions to the LCB equations as calculated by Contopoulos	676
15.7	Causal structure of an accelerating, collimating MHD flow	681
15.8	MHD group polar diagram distorted by acceleration normal to the field/stream line	683
15.9	Causality diagram for a radially self-similar jet model	688
15.10	Components of conserved total energy E per unit mass along a field line	689
15.11	Magnetic field/stream diagram from non-self-similar simulation of MHD outflow	691
15.12	Magnetic field/stream lines on two very different scales for warm MHD winds/jets from black holes	695
15.13	Positions of MSP, AP, and MFP for models of warm, self-similar MHD winds/jets from black holes	697
15.14	Axisymmetric GRMHD simulation of MRI turbulent disk that launches a large-scale jet	697
15.15	Important surfaces for the simulation in Fig. 15.14	699
15.16	Blandford–Rees cartoon of hot spot structure at the end of a supersonic jet	702
15.17	A simple one-dimensional supersonic, relativistic HD “jet” simulation	703
15.18	One of the first numerical simulations of a two-dimensional, axisymmetric hydrodynamic jet	705
15.19	Four different simulations of axisymmetric, relativistic HD jets . . .	706
15.20	Cross-sections of three-dimensional HD jet that becomes unstable along its length	707
15.21	Full three-dimensional realization of the growth of jet turbulence . .	707
15.22	Three-dimensional simulation of a <i>moderately</i> relativistic HD jet that precesses at its base	709
15.23	Three-dimensional simulation of a <i>highly</i> relativistic HD jet that precesses at its base	709
15.24	Example of forming FR I jet flow by disrupting FR II flow	711

15.25	A simple one-dimensional super-slow magneto-acoustic MHD “jet” simulation	712
15.26	Simulations of two-dimensional, axisymmetric MHD jets with strong toroidal magnetic field	714
15.27	Shock conditions for final state in the highly magnetized simulation in Fig. 15.26	715
15.28	Three-dimensional development of the helical kink instability in an MHD jet	718
15.29	Velocity structure for four different three-dimensional, super-slow magneto-acoustic jet simulations	719
15.30	Comparison of MHD jet simulations with different Poynting flux and external pressure gradients	720
15.31	Comparison of relativistic MHD jet propagation without and with sheath of MHD wind surrounding the jet	722
15.32	Three-dimensional GRMHD simulation of MRI turbulent disk that launches a large-scale jet	723
16.1	Formation, evolution, and death of stars, with an eye toward neutron star, black hole, and jet formation	729
16.2	Radio vs. bolometric luminosity for both binary black holes and neutron stars	740
16.3	Photon and power spectra for black hole binary sources in the hard state	763
16.4	Our model for the hard state: geometrically thin, evaporating disk with dominant ADAF corona	765
16.5	Radio/X-ray luminosity (fundamental plane) plot for hard state black hole binaries	767
16.6	The MDAF model for hard state outlier sources	768
16.7	Photon and power spectra for black hole binary sources in the soft state	772
16.8	Typical model for the soft state: geometrically thin disk with a compact corona	774
16.9	Our model for a proposed new, unstable state: SS “inner” region disk with outflow	776
16.10	Photon and power spectra for black hole binary sources in the very high/SPL state	778
16.11	Our model for the SPL state: a moderately super-Eddington “slim” disk	780
16.12	Our model for another new proposed state: the super-soft super-Eddington wind disk state	781
16.13	A model for the super-soft X-ray source P098 in M101	783
16.14	The radio/X-ray luminosity plot for Seyfert galaxies	800
16.15	Spin-dependence of black-hole-launched jet radio power	816
16.16	Evidence that M87 may have a recollimation shock nearly $10^6 r_g$ from the black hole	820

16.17 Schematic diagram of the BL Lacertae jet, over several decades in distance from the black hole 822

16.18 A theorist’s “principal component” diagram for all astrophysical black hole systems 825

List of Tables

3.1	Polarization and elongation of supernova explosions	86
3.2	Relative numbers of known binary X-ray sources	91
5.1	Properties of selected polytropes	159
7.1	Important equatorial radii for Schwarzschild and Kerr black holes . . .	242
8.1	Gravitational wave properties of different types of binary stars	268
10.1	Estimated birth and death rates for different types of close binaries . .	402
10.2	Estimated fraction of different types of doubly-degenerate binaries . .	407
11.1	Tidal disruption of stars and typical AGN accretion rates	450
11.2	Long-term black hole growth by different types of stars and clouds . .	451
15.1	Types of possible jet outflows from accretion inflow models in Chapter 12	668
15.2	Lorentz factors for different jet-launching magnetizations	676
16.1	Secular time scales for high accretion rate states <i>vs.</i> $M_{\bullet}$	749
16.2	Thermal time scale ranges for high accretion rate states <i>vs.</i> $M_{\bullet}$	749
16.3	Typical longitudinal magnetosound crossing times for a relativistic jet <i>vs.</i> $M_{\bullet}$	749
16.4	Models of P098 at two different color temperatures	782
G.1	Physical constants used in this book	865
G.2	Astrophysical constants used in this book	865

Part I

Observations of Black Hole Engines

Chapter 1

Recognizing Black Holes

We all live in a gravitational “hole”, produced by the earth’s gravitational field. By human (but not cosmic) standards, it takes quite a bit of energy to climb out of our hole before we can travel freely in space to other planets or stars. Light, however, has no trouble escaping the earth, so from a distance our planet appears as a pale blue object in reflected sunlight, not black at all.

Except for its strong gravity, a true black hole is very similar to any other object like the earth or the sun. It is simply a planet or star (or group of these objects) that has collapsed in on itself to essentially a point mass. A black hole’s gravity is so strong that nothing, not even light, can escape from it, and no material forces that we know are strong enough to keep the matter that makes up the hole from collapsing in on itself. When viewed from a distance, a black hole will have no visible solid surface, not only because reflected light cannot escape, but also because its surface has shrunk to zero radius. There is, however, a “radius of no return” called the “horizon”, which could be seen as a black spot blotting out the light from distant stars behind the black hole. Compared to the size of the sun or earth, the horizon is quite small. If the sun were compressed to a black hole, its horizon diameter would be only 6 kilometers (about four miles). And if the earth were compressed to a black hole, it would be about the size of a golf ball. Nevertheless, although the compressed sun and earth will no longer function as a star and planet, they still will execute their gravitational duties, keeping the rest of the planets, and the moon, in their respective orbits. This is because, at the same large distance from a gravitating object, it does not matter whether the object is a star, planet, or black hole. Only its mass matters. It is only very near the gravitating object that size matters, and, in this case, small is much more powerful.

Black holes have been around for almost the age of the universe, of course, but our ability to recognize them as such has been around for less than forty years. One of the problems in this regard, and fortunately for the human race, is that black holes are quite distant. While the horizon of the supermassive (3 billion solar mass) black hole (SMBH) in M87 is larger than the orbit of Pluto, at a distance of 15

megaparsecs (Mpc) or 49 million light years (Mly),¹ its angular size is only 8 microarcseconds. This is far smaller than the imaging power of any telescope available now or for the next two decades. The situation for stellar-mass (up to 30 solar mass) black holes in our own Galaxy is even worse. While thousands of times closer (only a few kiloparsecs [kpc] or 10 thousand light years away) they are a hundred million times smaller in radius (less than 100 kilometers), making their angular sizes an extremely small 0.4 nanoarcseconds. This is smaller than the size of a cold virus in London (10 nm) as seen from New York City. The black hole with the largest angular diameter on the sky is our own supermassive one at the center of our Galaxy (4.1 million solar masses at about 8 kiloparsecs distance or about 20 microarcseconds). Because black holes are too small to appear from earth as black spots on the sky, their presence can be inferred *only* by observing their effects on the environment around them – on the stars and gas that reside near the black holes, or on more distant regions of the galaxies that harbor them, and beyond – or by observing the energy output generated by black hole output itself.

The fact that black holes can generate energy – and actually act like an internal combustion engine – initially surprises most people. Nothing can escape a black hole, so how can it produce anything? The answer is that the black hole’s gravity provides the *environment* (just outside the horizon) where the engine can operate at maximum efficiency. At that location, power generated by infalling material still can affect the outside universe and do its job. Of course, once the infalling matter crosses the horizon into the black hole, it no longer can be useful as fuel. Note that the word “engine”, as we use it here, has the same meaning as on earth: a system that consumes fuel, does work of some kind, probably produces heat, expels exhaust, *and* stays intact. That is, while an explosion (like the big bang) accomplishes most of these tasks, it does so only once and does not remain intact; so, by this definition, while the big bang is certainly more powerful than any black hole system, it is not an engine of any kind, let alone the ultimate engine. Many black hole engines are much more powerful than an entire galaxy, so the only other possible known rival (energetically) is a cluster of galaxies. However, even these are not that much more luminous than the brightest quasars, and clusters are not known to do work on other parts of the cosmos. Black hole engines, therefore, are the most energetic and powerful engines in the universe.

Black hole systems can masquerade as a variety of objects which, like the numerous species of living things on earth, first need to be recognized, then cataloged, classified, explained, and unified with all other black hole systems. The task often is not easy. Black hole systems sometimes appear as normal or somewhat peculiar galaxies, perhaps with strong jet-like emission in the radio, but perhaps not. Sometimes they appear as star-like objects with no other properties; but these “stars” are actually at enormous distances. Sometimes they are variable; sometimes they are not.

Many black holes have completely escaped notice in optical telescopes, only revealing themselves in the more exotic regions of the electromagnetic spectrum:

¹ One parsec equals 3.26 light years or 3.086×10^{18} cm.

radio, X-ray, and γ -ray. At first, it was particularly difficult to tell whether the compact X-ray source in a binary system was a neutron star or a stellar-mass black hole. Both have a size of about 10 kilometers, both have very strong gravitational fields, and both can produce X-rays when material accretes onto them. In fact, these two objects are often so similar in their origins and properties that in this book we will include neutron stars as “near-black” holes. They will round out the discussion and make the overall black hole field much more complete. Ultimately, however, it will be important to identify true black holes in binaries because it is they, more so than the neutron stars, that are related to their much more massive cousins that lie at the centers of galaxies and quasars.

1.1 Energy Output

How does one distinguish a black hole engine from one powered by some other object or objects? One of the initial clues to the existence of black holes was energy output. The first quasars discovered were estimated to produce about 10^{46} erg s⁻¹ (one thousand trillion trillion trillion watts), which was difficult to explain by any means other than accretion of matter onto a black hole of at least 100 million solar masses. The reasoning actually came from stellar structure theory. In 1917 Sir Arthur Eddington showed [7] that, in order to not blow itself apart with internal radiation pressure, any star (or engine) of a mass M could not have a luminosity any greater than

$$L_{\text{Edd}} \equiv \frac{4\pi G M c}{\kappa_{\text{es}}} = 1.25 \times 10^{38} \text{ erg s}^{-1} m \quad (1.1)$$

where $G = 6.6738 \times 10^{-8}$ erg cm g⁻² is the universal gravitational constant, $c = 2.997925 \times 10^{10}$ cm s⁻¹ is the speed of light², $\kappa_{\text{es}} = 0.401$ cm² g⁻¹ is the opacity of ionized matter (to electron scattering), and $m = M/M_{\odot}$ is the object’s mass in units of solar masses (the mass of the sun or $M_{\odot} = 1.989 \times 10^{33}$ g). If an object were found to emit a total luminosity L , the “Eddington luminosity” could be used to determine a *minimum* mass for that object of

$$M_{\text{min}} = (L/L_{\text{Edd}}) M_{\bullet} \quad (1.2)$$

where symbol $M_{\bullet}$ represents the black hole mass. This is a crude, but often-effective estimate. In the case of the sun, $M_{\text{min}} = 0.00003 M_{\odot}$. The actual value is, of course, $M = 1.0 M_{\odot}$, which is much greater than the minimum value because the sun produces energy at a rate far below its Eddington luminosity limit. In the case of the first quasars discovered, M_{min} is about $10^8 M_{\odot}$, whereas the actual value is believed to be around $10^9 M_{\odot}$. These quasars, therefore, apparently radiate at about 10% of the Eddington limit.

² About 671 million miles per hour.

Many other objects that look similar to quasars have been found (*e.g.*, Seyfert and radio galaxies) but with much lower luminosities. Nevertheless, using a “guilt-by-association” argument, these objects also are suspected to harbor black holes of similar or lower mass, possibly in proportion to their luminosity if all radiate at 10% of the Eddington limit.

1.2 Compactness

Of course, ten trillion stars like the sun can produce $10^{46} \text{ erg s}^{-1}$, so another important criterion for a candidate black hole engine is that it be compact. Even under extreme conditions, it would be very difficult to pack this many stars into a volume smaller than several parsecs (10 ly) across; they would begin to collide with one another and merge into a supermassive star, whose fate would be to collapse to a black hole anyway. So, if a quasar engine appears to be smaller than several parsecs, this also might indicate a black hole engine.

An object’s size can be estimated if its luminosity happens to vary. Nothing can travel faster than the speed of light. Therefore, if the brightness of an object varies by, say, 10%, in a time Δt , then the region from which that 10% of the light comes can be no larger than

$$r_{\text{max}} = c \Delta t \quad (1.3)$$

For example, if $\Delta t \approx 1$ month, then $r_{\text{max}} \approx 1$ light-month (or 10^{17} cm), so at least 10% of the actual object is probably considerably smaller than a light-month in size.

Another way of inferring compactness is to determine that a large portion of the energy output is emitted in X-rays rather than in the optical or radio. If the object emits like a black body, then its luminosity L , effective surface temperature T_e , and size R are related by the black body relation

$$L = 4\pi R^2 \sigma T_e^4 \quad (1.4)$$

where $\sigma = 5.6704 \times 10^{-5} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ K}^{-4}$ is the Stefan–Boltzmann constant. Knowing the power output L and measuring T_e from the spectrum, we can solve for the size

$$r_{\text{bb}} = \left(\frac{L}{4\pi\sigma T_e^4} \right)^{1/2} \quad (1.5)$$

For example, consider a binary system in our Galaxy with one star emitting, say, $10^{37} \text{ erg s}^{-1}$ in the *optical–UV* ($T_e \sim 10,000$ kelvins). From the Eddington limit argument above, its mass should be at least $0.1 M_\odot$ and probably more like $1.0 M_\odot$. And from the black body argument we would imply a radius of 10^{12} cm for that star – about the size of a normal giant star. However, if most of this power instead came out in the *X-ray* at 1 keV (10^7 K), then r_{BB} would be about 10^6 cm – about the size of a neutron star or stellar-mass black hole. More information (specifically

a more accurate estimate of the mass) would be needed to distinguish between the latter two.

1.3 Engine Mass

In addition to luminosity and compactness, a third method of recognizing a black hole system is to actually measure the mass of the central engine. This technique works even if the black hole is not accreting or producing much light.

1.3.1 Binary X-Ray Star Masses

For X-ray binary (XRB) systems, from the above black body argument we know that the compact X-ray source in the binary must be either a neutron star or a black hole. Given what we know about neutron star matter, any neutron star more massive than $2.5\text{--}3.0\text{ }M_{\odot}$ cannot exist: it will collapse to a black hole in about 100 microseconds. So, *any X-ray star in a binary system with a mass greater than $3.0\text{ }M_{\odot}$ most likely is a black hole.*

With orbits of less than ten solar radii in size, however, it is currently difficult or impossible to resolve orbits of binary X-ray systems with telescopes and deduce all of the orbital parameters. Generally, one can determine at most the orbital period P_{orb} and projected orbital velocity v_{orb} of the binary but not the orbital shape or inclination. This allows us to compute the “mass function” f of the X-ray binary system from these

$$f \equiv \frac{P_{orb} v_{orb}^3}{(2\pi G)} = \frac{M_x^3}{(M_x + M_c)^2} \sin^3 i < M_x \quad (1.6)$$

The middle part of the equation comes from Newton’s laws of gravity and motion, and the right part comes from knowing that $\sin i$ and $M_x/(M_x + M_c)$ are always less than 1.0. Equation (1.6) tells us that, even if we cannot determine the companion mass M_c and the orbital inclination i and solve for M_x uniquely, if the mass function f of a binary X-ray system is greater than $3.0\text{ }M_{\odot}$, we know that the X-ray source is likely to be a black hole.

1.3.2 Masses of Black Holes in the Centers of Stellar Systems

Measuring the mass of compact objects in the centers of globular star clusters and galaxy bulges does not have to be as precise as that in X-ray binaries to conclude that a black hole is there. All turn out to be between ten thousand and ten billion

solar masses. Nevertheless, the field of supermassive black hole mass determination is now a rather precise endeavor, often with errors of only 10–20%. In the case of our own Galaxy, with a $4.1 \times 10^6 M_\odot$ black hole at the center, the orbits of several giant stars around the center can be followed (Fig. 1.1). Each star has a different orbital plane, allowing a good statistical sampling of the central hole’s gravitational field, and, therefore, a good determination of its mass. One star (S0-16) has an extremely eccentric elliptical orbit, passes within only 60 AU³ of the central black hole, and has an orbital velocity of $0.03 c$ or 9000 kilometers per second (20 million miles per hour) on its closest approach.

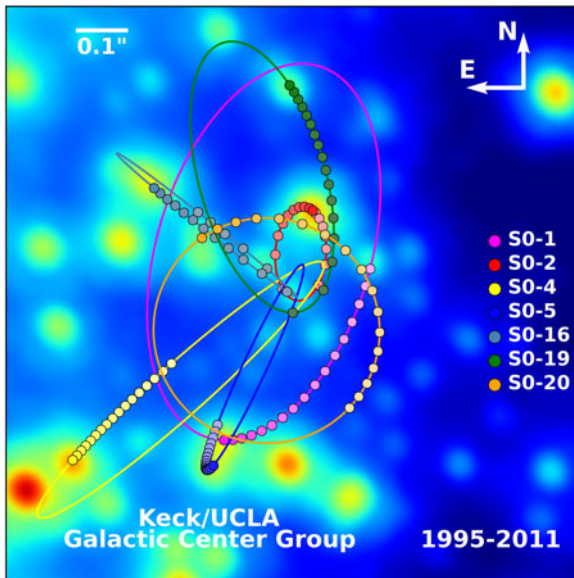


Fig. 1.1: Actual measured orbits of different stars around our Galaxy’s central black hole, determined by monitoring the stars’ positions over 13 years. The background image shows those stars in their 2011 positions in the innermost one square arcsecond around the Galactic center ($\sim 1/30 \text{ pc} \times 1/30 \text{ pc}$). The seven stars listed at right were used to determine the black hole mass, which is approximately $4.1(\pm 0.6) \times 10^6 M_\odot$ [8]. Figure kindly provided by Professor A. Ghez; reproduced with permission.

In other galaxies, the orbits of individual stars cannot be followed, but the aggregate light from millions of stars surrounding the galaxy’s central black hole *can* be measured. In particular, the width of absorption lines in those stars’ light, measured at a given radius from the galaxy’s center, indicates how fast those stars are moving there. Normally, if a black hole were not there, this “velocity dispersion” σ_V would increase as one measured it closer to the center, until one reached the core radius

³ Astronomical unit, the distance from the earth to the sun, or $1.496 \times 10^{13} \text{ cm}$. 60 AU is only half again the distance of Pluto from the sun.

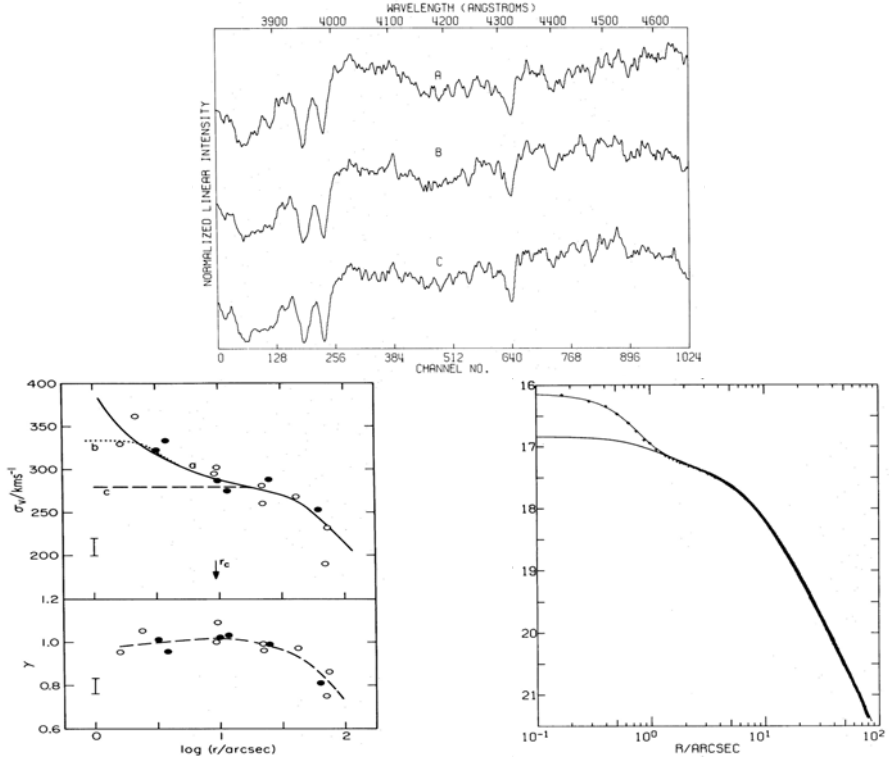


Fig. 1.2: Data from the first determination of the mass of the black hole at the center of the spheroidal galaxy M87 (see Fig. 2.2) [9], [10]. *Top:* Optical spectra of two regions in the core of M87: 70 (B) and 4 (C) arcseconds from the M87 galactic center; (A) is the spectrum of a K0 giant star in our own Galaxy that was used for comparison. Note the broader absorption lines in (C) than in (B). *Lower left:* Plot of velocity dispersion (essentially absorption line width) as a function of distance from the M87 center. Note how σ_v increases dramatically within 10 arcseconds (800 pc) of the center. *Lower right:* Stellar luminosity density (in V magnitudes per square arcsecond) vs. distance from the center. Note the “light cusp” within about 1 arcsecond. The authors determined a black hole mass of $3\text{--}5 \times 10^9 M_\odot$ for M87’s black hole. In practice, very detailed stellar dynamical models have to be fit to data like this before a reliable black hole mass can be determined. Modern estimates, which measure gas motions with the Hubble telescope, obtain $M_\bullet = 3.2 \pm 0.9 \times 10^9 M_\odot$ [11]. Reproduced by permission of the American Astronomical Society (AAS).

r_c (a few hundred parsecs or so). Then the velocity dispersion would be constant inside $r < r_c$. However, with a black hole at the center, σ_v will continue to increase well within the core radius (Fig. 1.2) at a rate given by

$$\sigma_v = K (GM_\bullet/r)^{1/2} \quad (1.7)$$

K is a dimensionless constant of order unity that describes the distribution of stars in the galaxy center. We can solve for the black hole mass $M_\bullet$ and get

$$M_{\bullet} = \langle r \sigma_V^2 \rangle / (K^2 G) \quad (1.8)$$

where the angle brackets indicate an average over different values of the quantity $r \sigma_V^2$ that were measured at different radii r . In practice, a lot of work goes into determining the factor K , and the problem can be complicated by the gravitational influence of other stars in the nucleus of the galaxy. Detailed models of the nucleus are built inside a computer, with thousands of stars in different, and sometimes very unusual, orbits, and models with different black hole masses are tried. The best-fit model gives the black hole mass. However, with current telescopes, the best data can detect black holes in galaxies only greater than about $10^4 M_{\odot}$.

1.3.3 Masses from Microlensing of Isolated Black Holes in the Galaxy

As we shall see in Chapter 7, Einstein's theory of relativity predicts that gravity not only bends the paths of stars as they travel in orbits about black holes, it also bends the path of massless light particles themselves. This prediction was first demonstrated by Sir Arthur Eddington on 29 May 1919 during a solar eclipse [12]: positions of stars on the opposite side of the sun appeared to shift as the sun moved in front of them.

For a distant compact object like an isolated black hole in our Galaxy, and an even more distant point source of light positioned right behind the compact object (like a star in our Galactic bulge or in the Magellanic clouds), this effect can act like a “gravitational lens”, focusing the star's light in our direction and magnifying its intensity. Because isolated black holes will be in orbits about the Galaxy, just like other stars, such a magnification of a star will be only a temporary feature in the sky, brightening as the black hole moves between us and the distant star and then dimming again as it passes (see Fig. 1.3). These events are called “microlenses” or “microlensing events” (to distinguish them from the very massive gravitational lenses that are created by galaxies or clusters of galaxies as they distort the images of extremely distant galaxies and quasars that lie behind them). At first a microlensing event might be mistaken for a common variable star. However, it differs markedly in that the light curve (the variation in light intensity with time) is very symmetric: it decreases in intensity in the same way as it originally increased and does not repeat.

Even after detecting a microlensing event, however, it is difficult to measure the mass of the lens unless the event is so long in duration that the earth has moved appreciably in its orbit during the event. This creates a “parallax effect” that allows observers to also measure the speed of the lens v_{lens} as it passes in front of the distant star. Then the mass of the lens can be computed from the following equation

$$M_{lens} = \frac{v_{lens}^2 \tau_E^2 c^2}{4 G} \left(\frac{D_{star} - D_{lens}}{D_{lens} D_{star}} \right) \quad (1.9)$$

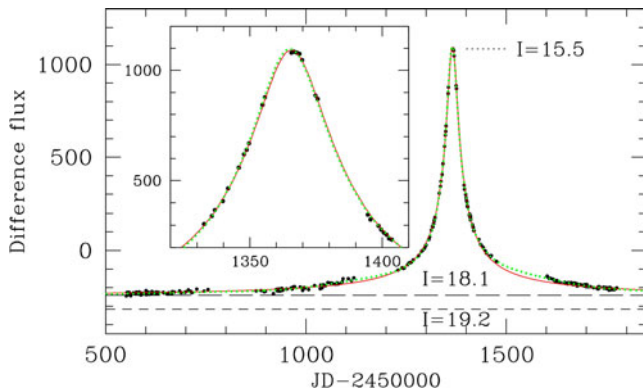


Fig. 1.3: Data from the OGLE collaboration [13] of a microlensing event called OGLE-1999-BUL-32, which was viewed toward our Galaxy’s bulge. (The event also was observed simultaneously by the MACHO group and, hence, also named MACHO-99-BLG-22 [14].) This event had a very long crossing time of $\tau_E = 640$ days, indicating a dark mass greater than a few solar masses and, therefore, possibly an isolated stellar-mass black hole. Reproduced by permission of J. Wiley and Sons, Ltd.

where τ_E is the “Einstein crossing time” for the microlensing event and D_{lens} and D_{star} are the distance to the lens and distant star, respectively. This estimate is accurate to about a factor of 2, so any microlens that is much fainter than a normal star and has a mass of at least $5\text{--}6 M_\odot$ (twice $2.5\text{--}3.0 M_\odot$) is likely to be a black hole. In practice, teams of astronomers must monitor tens of millions of stars with automated equipment in order to obtain a few good black hole microlensing candidates.

1.4 Relativistic Motion

A fourth method of inferring that a black hole possibly lies at the center of an observed object involves measuring some *highly* relativistic *bulk* linear motion in that system, with a velocity very close to the speed of light, which motion can be produced naturally only in a strong gravitational field.⁴ The fastest outflows measured so far (close to a Lorentz factor of $\gamma \approx 700$ or $V = 0.999999 c$) are produced by the γ -ray burst sources; these are believed to be stellar black holes (in distant galaxies) in the process of forming that produce jets which happen to point right at us.

Relativistic motion by itself, however, especially in galactic sources, is not conclusive evidence of a true black hole; it usually must be used in conjunction with one or more of the other above properties. Supernovae (exploding stars), for example, generate blast waves that can reach $0.1c$, but we believe that most supernovae involve neutron stars, not black holes. And pulsar winds can reach very close to

⁴ Of course, particles can be accelerated to highly relativistic *random* velocities by shocks (Fermi acceleration) and magnetic processes.

the speed of light, but we know that pulsars are rotating magnetized neutron stars. Some binary black hole X-ray sources generate outflowing jets of plasma that reach speeds of $0.9c$ or greater. But similar jets of plasma are also seen from some binary neutron star X-ray sources.

In the case of extragalactic sources in the centers of galaxies, relativistic motion is a more direct indicator. Some radio sources are seen to produce jets with components traveling up to $\gamma \approx 20$ or $V = 0.999c$. Since the mass of the central object is thousands to billions of solar masses, it cannot be a neutron star and must be a supermassive black hole.

Chapter 2

Macroquasars: Supermassive Black Holes in the Centers of Galaxies

Once recognized as a black hole system, the process of cataloging, classifying, explaining, and unifying like systems can begin. The next few chapters describe several different varieties of black hole systems. Also discussed are several different criteria for explaining why these systems appear the way they do to observers and for unifying them with other types of black hole system. Ultimately, we shall see that all black hole systems (including the small ones in our own Galaxy) are fundamentally the same. They differ only in the mass of their central black hole, how fast it is rotating, how fast it consumes fuel, the angle at which we view it (relative to the hole's spin axis), and the nature of the environment in which the black hole resides.

There are traditionally three types of black hole systems: active galactic nuclei (AGN, which includes quasars, radio and Seyfert galaxies, and a host of other objects), binary X-ray sources (now called “microquasars”), and the recently recognized “intermediate mass” black holes (IMBH, with masses of one hundred to ten thousand solar masses). In order to unify the terminology, I sometimes will refer to AGN as “macroquasars” (large ones), IMBH as “miniquasars” (small ones), and binary X-ray sources as “microquasars” (tiny ones). This is a further evolution of the word quasar (originally short for “quasi-stellar radio source”), which is already now used to describe even radio-quiet AGN, and a likely direction for the terminology to take as we learn more about the similarities among the structure and behavior of these systems. For the original objects identified as quasi-stellar radio sources, we will use the terms QSRs or RLQs (for radio-loud quasars).¹

Our discussion will begin with the discovery of Seyfert galaxies, which had a profound influence on the understanding of radio galaxies and quasars when the latter were discovered a few decades later. These, in turn, revolutionized the way we think about all active galaxies, leading to unification of the different Seyfert types and of other large groups of AGN.

¹ Throughout much of the book, we shall use the traditional term “supermassive” to describe black holes in the range 10^5 – $10^{10} M_{\odot}$. By Chapter 16, however, it will have become clear that, for several important reasons, this very large class really should be divided into two: the massive (10^5 – $10^7 M_{\odot}$) and the truly supermassive (10^8 – $10^{10} M_{\odot}$). Readers may want to keep this in mind as they proceed through this volume, and note clues to this new classification along the way.

2.1 The Early Days of Active Galactic Nuclei Research and Seyfert Galaxies

At the beginning of the 20th century, astronomers were trying to understand the celestial “nebulae” – diffuse bright blotches in the sky that clearly were not stars. Earlier, in the 1700s Charles Messier, a French naval astronomer, had cataloged about 100 such objects; and in the late 1800s a much larger New General Catalog of Nebulae and Clusters of Stars, with 7840 objects, had been compiled by John L. E. Dreyer of the Armagh Observatory in Ireland. It was not known then that many of these objects were “island universes”, like the Milky Way but many millions of light years distant, let alone that each harbored a central supermassive black hole.

Most of these nebulae had spectra with absorption lines only (now known to be produced by the aggregate light from billions of stars), but in 1908 Edward A. Fath, of Lick Observatory in California, found emission lines in the spectrum of the nucleus (center) of NGC 1068 [15]. Then, in 1917, Vesto M. Slipher, of Lowell Observatory in Arizona, showed that these lines were very broad, indicating fast gas motions in the nucleus [16]. The next year, W. Wallace Campbell and Joseph H. Moore, also of Lick, discovered similar emission in another now-famous galaxy NGC 4151 [17]. The lines in NGC 1068 indicated gas motions up to 3600 km s^{-1} and in NGC 4151, up to 7500 km s^{-1} – 2.5% the speed of light! Fig. 2.1 compares short and long photographic exposures of NGC 4151 to reveal the bright nucleus.

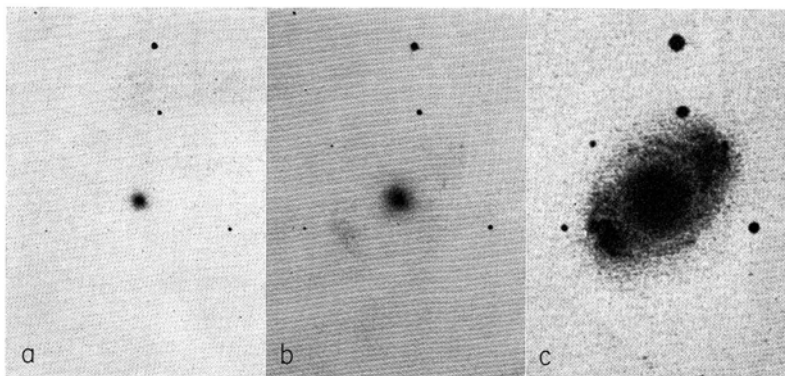


Fig. 2.1: Series of exposures of Seyfert 1 galaxy NGC 4151. Short exposure (left) shows the central unresolved Seyfert nucleus; intermediate exposure shows the ionized material surrounding the nucleus; deep exposure (right) shows the host galaxy [18]. Reproduced by permission of the AAS.

In a seemingly unrelated discovery, also in 1918 and also at Lick, Heber D. Curtis *underexposed* another nebula, Messier 87, uncovering a long slender shaft or jet of light emanating from the nucleus (Fig. 2.2) [19]. The M87 optical jet remained unique until the mid 20th century, when the science of radio astronomy and the study of extragalactic jets were born.

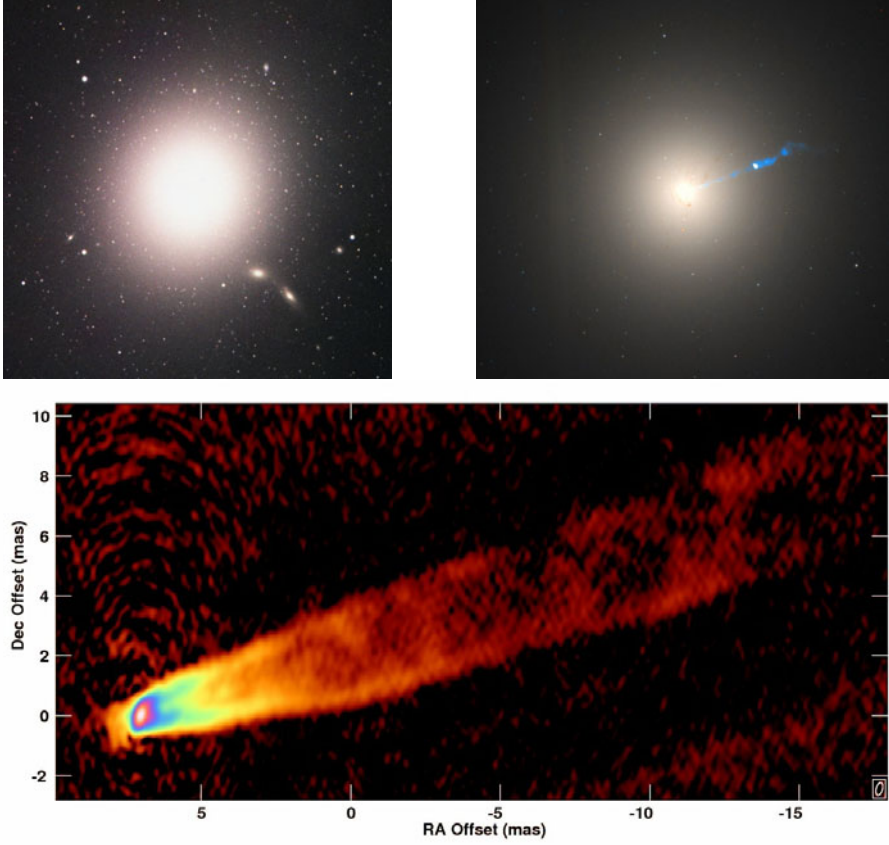


Fig. 2.2: Elliptical galaxy M87 and its jet. *Left:* Ground-based image made by David Malin of the Anglo-Australian Observatory; ©Australian Astronomical Observatory. *Right:* Hubble Space Telescope exposure of the inner region, showing jet and bright core. Reproduced by permission of STScI. *Bottom:* VLBA radio image made by C. Walker at NRAO and colleagues; latter image lies in the core with $\sim 1/10,000$ th the scale of the image at right. Reproduced from Fig. 2 of reference [20], with permission of IOP Publishing Ltd.

In 1925, Edwin P. Hubble, of Mt. Wilson Observatory in California, discovered variable stars in NGC 6822, proving that many of these objects were indeed galaxies millions of light years distant [21], and in 1926 he published his famous “Hubble sequence” classification of galaxies into ellipticals, spirals, and irregulars [22]. Then, in 1929, he announced his famous “Hubble relation”, showing that a galaxy’s distance could be approximated simply by measuring the velocity by which it is moving away from the Milky Way [23]:

$$d = v/H_0 \quad (2.1)$$

where H_0 is Hubble's constant, now known to be about $75 \text{ km s}^{-1} \text{ Mpc}^{-1}$. This paved the way for a much better understanding of the nebulae, including those with very broad emission lines.

By 1943, another astronomer at Mt. Wilson, Carl K. Seyfert, had gathered enough information on the luminosity and spectra of the nuclei of twelve galaxies with broad emission lines (ten spirals and two peculiar ones) to conclude that such galaxies form a distinct class of objects [24]. We now call galaxies of this type "Seyfert galaxies", and they appear to number in the many millions. A particularly rich source of Seyfert candidates came in the surveys of Benik E. Markarian, of the Byurakan Observatory in Armenia, performed from the 1960s to the 1980s [25, 26]. He and his colleagues identified a large number of galaxies that had strong ultraviolet continuum emission, and many turned out to be emission line galaxies like Seyfert's original candidates. Their main differences from other emission line nebulae is that they are galaxies (not gas clouds), the line emission is produced by a very small, almost star-like, nucleus at the center of the galaxy, and the lines are broadened by gas motions at speeds of thousands of kilometers per second.

In 1971 Edward Ye. Khachikian, also of the Byurakan, and Daniel W. Weedman, of Dyer Observatory in Tennessee, showed that there were two distinctly different classes of Seyfert galaxies [27]. Type 1 Seyferts (like NGC 4151) have lines from "permitted" atomic transitions of hydrogen, carbon, magnesium, *etc.* that are distinctly broader than the "forbidden" lines like the 500.7 nm (green) emission line of doubly-ionized oxygen. Type 2 Seyfert galaxies (like NGC 1068) have permitted lines with the same, rather narrow width of the forbidden lines ($500\text{--}1000 \text{ km s}^{-1}$). Here, the word "permitted" really means that the atoms emitting these lines can do so very quickly, before another process (such as a collision with another atom) changes its excited atomic state. Forbidden transitions have a lower probability of happening than permitted transitions; so atoms that emit forbidden lines generally do so only if they are in a low-density environment with a long time between collisions with other atoms.

After more detailed analysis of many Seyfert spectra, the following general model of most Seyfert galaxies emerged. There must be a very small ($<10^{15} \text{ cm}$) central radiation source (called the "central engine") that emits $10^{43\text{--}45} \text{ erg s}^{-1}$ and ionizes the emission line gas, and that gas must be moving very rapidly, either in an explosive outflow or in orbital motion very close to a supermassive object. In both Seyfert 1 and 2 galaxies, there is a fairly tenuous "narrow-line region" (NLR) of gas that is fairly far from the central engine ($100\text{--}1000 \text{ pc}$) and flowing at speeds of many hundreds of kilometers per second (well above the galaxy's escape speed). In Seyfert 1 galaxies, however, there also is another high-density, "broad-line region" (BLR) that is much closer to the central engine ($<0.01 \text{ pc}$) that is in an even more violent state, with velocities up to $0.1 c$. In intermediate and Seyfert 2 systems, this broad line region must be less prominent or missing altogether. But in the late 1970s it was not clear whether the Seyfert $2 \rightarrow 1$ sequence was due to an actual increasing presence of a BLR or whether the BLR simply was obscured somehow in Seyfert 2s and not in Seyfert 1s.

2.2 Radio Galaxies and Classical Quasars: Supermassive Black Holes with Jets

Ideas for solving problems often come from unexpected sources. This was the case for the Seyfert galaxies, where the solution lay in the new field of radio astronomy and the myriad of objects that were discovered there.

2.2.1 *Extended Radio Galaxies*

2.2.1.1 Discovery of the First Radio Galaxies

As is often the case, wartime advances in technology produce additional advances in the peacetime that follows. This was particularly true for the development of radar and radio technology, which led to the rapid development of the science of radio astronomy in the 1950s. Strong radio sources were discovered throughout the sky, and many were shown to be associated with galaxies – often rather nearby, but sometimes quite distant. Initially, the radio sources were given the name of the constellation in which they were found, followed by a letter, with A being the brightest in that constellation, B the second brightest, and so on. Virgo A was identified with the M87 galaxy, only 17 Mpc away. Centaurus A was identified with the southern hemisphere galaxy NGC 5128 – only 4 Mpc away and now known to be the closest radio galaxy². Some radio sources, like Cygnus A and Hercules A, were identified with rather distant galaxies (~ 200 Mpc and ~ 600 Mpc distant, respectively). The luminosity of these objects in radio power alone ranged from 10^{40} erg s⁻¹ for Cen A to 10^{44} erg s⁻¹ for Cyg A.

Because these active galaxies sometimes looked like two galaxies engaging in a violent collision or merger (*e.g.*, Cygnus A), at first the radio emission was thought to be generated directly by interstellar gas in these galaxies, excited by the collision. However, in the mid 1970s modern radio interferometers like the 5-kilometer telescope in Cambridge, UK, shown in Fig. 2.3, were able to show that the radio sources had a dumbbell-like shape with the galaxy halfway between the ends. In 1974, Sir Martin Ryle of the University of Cambridge, UK, was awarded one-half of the Nobel Prize in physics for his pioneering work in developing imaging techniques using multiple-telescope interferometers.

While very suggestive that the jet activity originated in the nucleus of the host galaxy (as proposed by Roger Blandford and Martin Rees [28] at that time), there was no strong evidence from the 5-km telescope observations that this actually was the case. This all changed in 1978 when Anthony Readhead used the new technique of Very Long Baseline Interferometry (VLBI) to make an image of the nuclear ra-

² Or any other type of galaxy with an active nucleus.

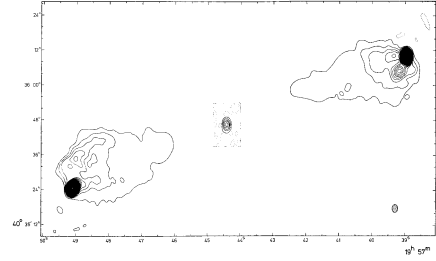


Fig. 2.3: Early aperture synthesis “map” of radio galaxy Cygnus A (right, reprinted from Fig. 2 of [32], by permission of J. Wiley and Sons, Ltd.) using the Ryle 5-kilometer radio telescope near Cambridge, UK (left; ©Mullard Radio Astronomy Observatory).

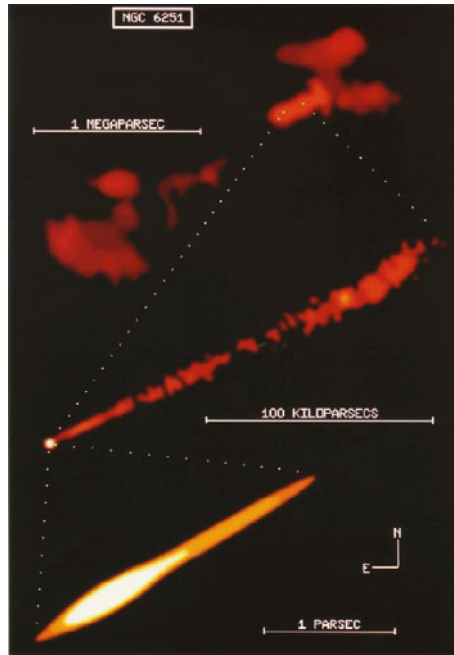


Fig. 2.4: Three images of the radio galaxy NGC 6251, showing its structure over a factor of a million in size scale. *Top:* 1400 MHz image of the outer radio lobes, using the Cambridge one-mile telescope [33]. *Middle:* 2700 MHz image of the jet powering the northwest lobe, using the Cambridge 5-km telescope [33]. *Bottom:* 10 GHz VLBI image of the parsec-scale compact source in the nucleus of the galaxy using a three-telescope array in the USA (Haystack, MA; Green Bank, WV; and Owens Valley, California). Figure is a color version of Fig. 2 in [34], kindly provided by Professor A. Readhead; reproduced with permission.

radio source in the newly-discovered radio galaxy NGC 6251.³ Fig. 2.4 shows three images of this radio galaxy at very different resolutions. The structure in all is well aligned, indicating that something deep in the nucleus is powering the radio lobes that lie three million light years away.

In the 1980s the Very Large Array (VLA) in Socorro, NM, USA, shown in Fig. 2.5, was able to confirm with spectacular radio images that the radio “lobes” at the ends of the dumbbells sources were connected to their nuclear jets by slender radio jets emanating from their galactic centers. In fact, a detailed comparison of the M87 jet, observed with the Hubble Space Telescope (HST) in the optical (Fig. 2.2, right) and with the VLA in the radio shows that Curtis’ original jet in M87 is, in fact, the inner part of a much larger radio jet which is bright enough, and close enough to us, to also be detectable in the optical.

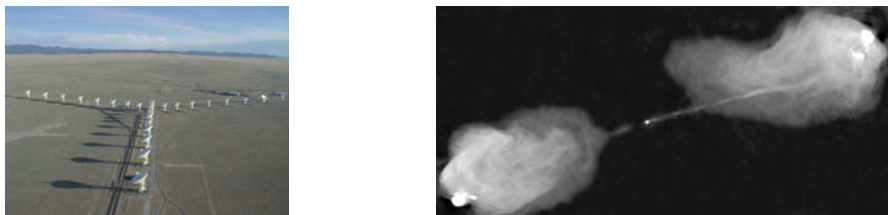


Fig. 2.5: Modern radio image of radio galaxy Cygnus A (right [35]) using the Very Large Array synthesis telescope in New Mexico, USA (left) in several configurations. Images provided by NRAO, an NSF facility, managed by AUI.

The radio and light emitted by these jets is not produced by heat (*i.e.*, not by thermal radiation), unlike that from the atmospheres of the sun and other stars. It is nonthermal synchrotron radiation, caused by highly relativistic electrons spiraling about magnetic field lines in the radio jet. The speeds of these electrons are so high ($v > 0.9999995 c$) that they are best described by the dimensionless Lorentz factor

$$\gamma_e \equiv \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \quad (2.2)$$

which becomes infinite as the electron speed approaches exactly c . Relativistic electrons in radio galaxies have Lorentz factors of $\gamma_e \sim 10^3$ – 10^5 , and the kinetic energy (and mass) of each electron is given by $\gamma_e m_e c^2$, which is much greater than the mass of the electron itself! Nonthermal radio spectra, when displayed in power f_ν per unit frequency ν , typically rise rapidly at low frequencies, reach a peak brightness, then become slowly less bright toward higher frequencies until a “break” frequency where they become fainter much more rapidly with increasing frequency (Fig. 2.6).

³ VLBI had just been developed in the late 1960s and early 1970s [29, 30, 31]; while unable to see much extended structure, VLBI could produce images of objects with most of their emission concentrated in a few *thousandths* of an arcsecond (“milliarcseconds”), allowing astronomers to compare structure in a single source on scales varying by a factor of a million or more in size.

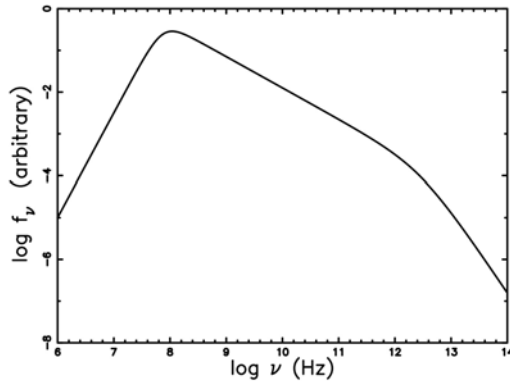


Fig. 2.6: Typical radio synchrotron spectrum, with $\nu_{\max} \approx 10^8$ Hz and $\nu_b \approx 10^{12}$ Hz.

Their spectra have the approximate shape given by

$$\begin{aligned}
 f_\nu &\propto \nu^{2.5} & \nu < \nu_{\max} \\
 &\propto \nu^{-0.75} & \nu_{\max} < \nu < \nu_b \\
 &\propto \nu^{-2} & \nu_b < \nu
 \end{aligned} \tag{2.3}$$

For the M87 jet the break frequency is higher than the optical band ($\nu > 10^{15}$ Hz) so its jet is optically bright. This tells us a lot about jet physics. Such a high ν_b indicates that the electrons have been accelerated to very high speeds in the last 20 years or so. Even traveling at nearly the speed of light, the jet must have taken at least 5000 years to reach its current position ~ 1500 pc from the M87 nucleus. This travel time is much longer than the time it takes the electrons to radiate all of their relativistic energy, so they must have been accelerated *in situ* – at the site where we are now observing them. That is, *the energy we see in radio and optical emission must come from the kinetic energy of the jet itself*, not from some acceleration process that occurred when the jet was first ejected from near the black hole. The process that accelerates the particles is believed to be shock waves that are internal to the jet. Internal shocks appear to play a major role in many different relativistic jet flows.

2.2.1.2 The 3C Radio Survey and Fanaroff and Riley Radio Source Classes

In the early 1960s the Cavendish radio astronomy group in Cambridge, UK, surveyed the northern sky ($> -5^\circ$ declination) and published the Revised Third Cambridge (3CR) catalog of all 328 radio sources with radio fluxes greater than 9 jans-

skys ($9 \times 10^{-26} \text{ W Hz}^{-1} \text{ m}^{-2}$) at a radio frequency of 178 MHz [36].⁴ Some of these sources, like 3C 405 (Cygnus A), were known radio galaxies, and others were new ones that had not been known previously.

The compilation of a complete sample of radio sources with known distances and luminosities allowed Bernard L. Fanaroff and Julia M. Riley at the Cavendish Laboratory to recognize in 1974 that most radio sources fit neatly into two classes [38]. The less powerful ones had complex structure, and the sky position of the brightest radio emission occurred near the optical galaxy, with the remainder of the radio source billowing outward in a broad expanding flow. (See Fig. 2.7, left.) On the other hand, the more radio luminous radio sources all had very straight, double structure (like Cygnus A), and the brightest regions of the radio emission were at the far ends of the lobes. (See Fig. 2.7, right.) These two classes, called “FR Class I” and “FR Class II” had a sharp boundary at the luminosity of about $10^{41} \text{ erg s}^{-1}$.

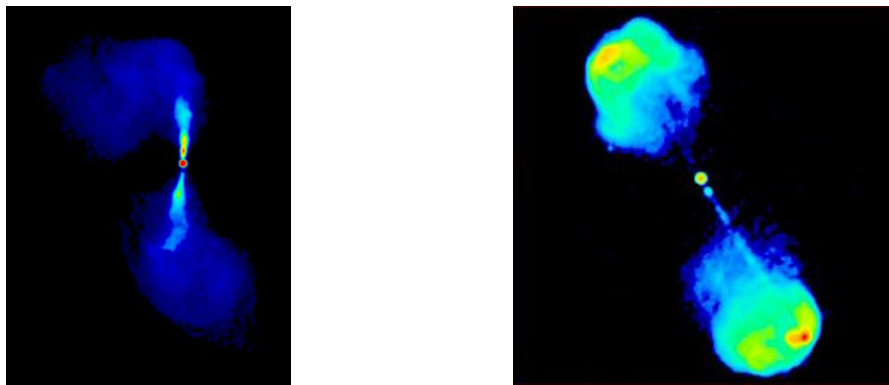


Fig. 2.7: Typical examples of Fanaroff and Riley Class I (left) and Class II (right) sources. The FR I source [39] is radio galaxy Messier 84 (a.k.a. 3C 272.1 or NGC 4374), and is a companion of M87 in the Virgo cluster (about 17 Mpc distant). The FR II source is quasar 3C 47 [40] which, at a redshift of 0.425, lies at the fairly great distance of 1400 Mpc (4.5 billion light years). Images reproduced from the Jodrell Bank Observatory atlas of double radio sources [41], with permission.

The modern interpretation of the difference between FR I and FR II objects is that the jet in the latter is much more powerful, and faster, than the jet in the former. This not only results in a stronger source, it also allows the jet to “plow through” the interstellar medium (ISM) of the host galaxy without being disrupted. The bright head of the jet in the FR II sources is the “working surface” shock front, where the fast-moving jet material impinges on the medium outside the galaxy. This head slowly

⁴ The original 3C survey was done at 159 MHz, went down to -22° declination, and had 471 sources in order of increasing right ascension. The 3CR (now often referred to as just 3C) lost several sources when the declination limit was raised to -5° and added a few, like 3C 390.3, in between the original 3C sources. In 1983 the catalog was revised a second time (3CRR) [37] in order to include sources like NGC 6251, which had been missed because it was too large on the sky ($>1^\circ$) to be classified as a single radio galaxy.

moves outward as the low-density intergalactic medium (IGM) is pushed aside by the even lower density jet. The FR I sources, on the other hand, have less powerful and slower jets. Below a critical luminosity these jets fail to flow unabated through their own galaxy's ISM, decelerating to subsonic speeds and becoming billowing subsonic clouds.

In the early 1990s Frazer Owen of the National Radio Astronomy Observatory (NRAO) and his colleagues showed that the FR I–FR II “break” actually occurs at different radio luminosities depending on how massive the radio galaxy is [42, 43]. The more massive a galaxy is, the harder a jet has to work to get through it. Geoffrey Bicknell of the Australian National University has shown that this trend is probably caused by the heavier galaxies containing more gas in between their stars (a denser ISM) than less massive galaxies [45]. This additional gas makes it more difficult for the jet to plow through the galaxy, so that a source with an FR II structure cannot occur unless the jet power is correspondingly higher.⁵

2.2.1.3 Radio Galaxies with Bright Optical Nuclei

In the late 1960s it was discovered that some of the 3C radio galaxies, like 3C 371 and 3C 390.3, had a very bright and compact optical nucleus and so were named “N”-galaxies. In the mid 1970s a further examination of the spectrum of these bright nuclei by Osterbrock [48, 49] revealed two distinct classes, just like the Seyfert galaxies. One group had very broad permitted emission lines (with widths implying velocities of $>10,000 \text{ km s}^{-1}$) and narrow ($<1000 \text{ km s}^{-1}$) forbidden lines, while the other group's permitted and forbidden lines were both relatively narrow. The two classes were given the names broad- and narrow-line radio galaxies (BLRGs and NLRGs), respectively, very analogous to Seyfert 1 and 2 objects. Just as in Seyferts, at the time it was not clear whether the difference between the two was due to the actual lack of a BLR in the NLRGs or that there still was a BLR in those objects, but it was obscured by dust absorption in the nucleus because it was simply so much smaller than the NLR. What was clear, however, was that *all* N-galaxies appeared to have FR II morphology, and they tended to occur in ellipticals only. They are among the most powerful of all radio galaxies and are clearly distinct from Seyfert galaxies. The latter are never strong, double-lobed radio galaxies (although

⁵ The author of this book also had a physical model (called the “magnetic switch”) for the FR I–FR II break, which also explained the trend with galaxy mass quite well [46]. The idea, however, involved differences in how the jets were accelerated in the central engines of FR I and FR II radio galaxies, rather than differences in the way the jets propagate through the galaxy. This idea no longer appears to be the principal explanation for the FR I–FR II break, due to the discovery of “hybrid morphology” radio sources or HYMORs [47]. These are radio sources that have one jet with FR I structure and one with FR II structure. This unbalanced structure, on scales of hundreds of thousands of light years, apparently persists for millions of years. It is extremely unlikely that a central radio galaxy engine could have an imbalance that persists for anywhere near that long. It is much more likely that the galaxy's interstellar medium itself has an imbalance, probably more dense on one side than the other, making it more difficult for the jet on the FR I side to plow through the ISM. See also Section 15.2.1.

a handful have been found with weak radio lobes), rarely appear in radio surveys, and tend to occur primarily in spiral galaxies. In the late 1990s, however, after it was recognized that Seyferts and N-galaxies were simply different types of the same phenomenon, the term “N-galaxy” fell out of favor; all such galaxies now are placed in the more general category of AGN.

2.2.2 *Extended Radio QSRs: Unification by Brightness Contrast*

2.2.2.1 **Discovery of the First Classical Quasars**

Many sources in the 3CR catalog did not have a galaxy in the vicinity of the radio source position. In those days, radio telescopes had poor resolution (only about 0.1 degree), so it was difficult to find an optical counterpart for these sources using the Cambridge data alone. Some, like 3C 48, were tentatively identified as “radio stars”, because a star-like object with a peculiar emission line spectrum was located near the center of the radio emission.

However, in 1962 Cyril Hazard, of the Jodrell Bank observatory in Manchester, UK, and the University of Sydney, and his colleagues were able to use the technique of “lunar occultation” (watching a radio source as it disappears behind the moon) to determine a precise position for one of the unidentified sources, 3C 273 [50]. The radio source was found to be an elongated (jet-like) source and to coincide with a star-like object with a most peculiar optical spectrum: its emission lines corresponded to those of no known element. The mystery was solved by Maarten Schmidt of the California Institute of Technology [51], who noticed that the spectrum could be understood if it were the normal spectrum of hydrogen, shifted toward the red by an enormous 15.8%. The implication was that the star-like object associated with 3C 273 was receding from us at $\sim 16\%$ the speed of light.

By this time, of course, the Hubble expansion of the universe had been discovered. However, a star-like object receding at a redshift of 0.158 implied that 3C 273 was at a distance of ~ 600 Mpc – similar to the furthest known galaxies at that time. Were these redshifts Doppler shifts like those of galaxies, and were the quasars really at the implied cosmological distances? Other “radio stars” were re-examined in light of this discovery, producing even higher redshifts: 3C 48 at $z = 0.367$ [52], 3C 9 at an astonishing $z = 2.012$ [53], *etc.* This new class of objects was given the name “quasi-stellar radio sources” or “quasars” and sometimes referred to as QSRs.⁶

The high apparent optical and radio brightness of the quasars and their enormous distances implied very high absolute optical and radio luminosities (*e.g.*, for 3C 273, $\sim 10^{46}$ and 10^{44} erg s⁻¹, respectively) and their variability implied sizes of

⁶ The name quasar is still used today, often even for radio-quiet objects of a similar nature, and also for radio-loud stellar-mass black holes in binary star systems. In this book we further expand the term to include all black hole engines. The acronym QSR will be reserved for objects like 3C 273 and 3C 48.

the central engine that were much less than a light year. Some investigators argued that these luminosities were unrealistically high and that interpreting the redshifts as being due to the cosmological expansion of the universe must be wrong. Others, *e.g.* John A. Wheeler of Princeton at the First Texas Symposium on Relativistic Astrophysics in 1963, argued that such a power output could be generated by matter falling into a “gravitational singularity” (black hole; see Hong-Yee Chiu’s review of the symposium [54], p. 13). A 10^8 solar mass black hole accreting stars and gas at the modest rate of two solar-type stars per year easily could generate typical quasar luminosities.

2.2.2.2 QSR Radio and Optical Properties

The types of radio structures displayed by different 3C QSRs varied greatly. Some were extended like radio galaxies. Others were very compact, appearing as radio point sources. However, the optical counterpart to all 3C QSRs appeared the same: it was star-like, with an emission line spectrum much like that of a BLRG, but often with a very high redshift. The permitted lines were very broad, and forbidden lines were much narrower, although still nearly 1000 km s^{-1} .

In addition to the emission lines, the spectrum had bright continuum emission that extended throughout the ultraviolet-optical-infrared (UVOIR) region. Like the radio continuum, the UVOIR continuum appeared to be nonthermal. It had a “steep” spectrum, similar to that from a synchrotron gas made of relativistic electrons ($f_\nu \propto \nu^{-1}$). Detailed analysis of the emission lines showed that it was the continuum that was responsible for exciting and ionizing the atoms that produced the broad and narrow lines. Because this continuum emission varied, unlike the lines, it was concluded that whatever produced the ionizing continuum must be very small – near or at the central engine producing the energy for the optical and radio emission. The optical QSR, therefore, had a hierarchical structure, with the ionizing source at the center (much less than a parsec in size), the BLR surrounding that (somewhat less than a parsec in size), and the NLR outside that at perhaps 100 pc. In no case, however, did a quasar have a detectable galaxy surrounding the optical source. In fact, that became the definition of a QSR: if a galaxy could be detected, the radio source was called an N-galaxy; if not, it was called a QSR.

2.2.2.3 Unification of BLRGs and Extended QSRs

Except for the lack of any detectable galaxy, some QSRs like 3C 47 (Fig. 2.7) appeared strikingly similar to radio galaxies, particular the BLRGs. Like the BLRGs, extended QSRs all were very powerful FR II sources. (FR I QSRs were not discovered until the turn of the millennium.) They had a straight dumbbell shape, with peaks of emission at the ends of the dumbbell lobes. They both also had broad line and narrow line regions, and the line widths, strengths, and ionization levels all were similar. The implication was that BLRGs and QSRs were the same type of engine,

but in the quasar case the central engine was so bright that the galaxy around it simply could not be seen. But, without the detection of a surrounding galaxy, the possibility remained open that QSRs were isolated engines, with no galaxy at all, while BLRGs were similar engines that occurred in galaxies.

An answer to this question began to emerge slowly during the 1970s and 1980s, with careful ground-based optical observations (particularly those by John Hutchings of the Dominion Astrophysical Observatory in British Columbia, Canada) of QSRs under good seeing conditions [55, 56]. Several known low-redshift quasars were observed to get detailed information on the “fuzz” that surrounded these objects, particularly density profiles and colors. Its properties appeared to be consistent with those of normal elliptical and spiral galaxies, although the redshift of this fuzz still could not be measured because it was so faint.

Recognizing that BLRGs and QSRs were the same type of object was the first step in the “unification” of suspected black hole systems. In this case the unifying parameter was the *ratio of quasar to galaxy optical brightness*. However, the fact that different systems had different brightness ratios could be because either the central quasar was a brighter object than most BLRG nuclei, or there was some obscuration of the central optical source in BLRGs that does not occur in QSRs (or both). A resolution to this question would help unite all extended radio sources, including objects like Cyg A and Cen A.

2.2.3 Compact Radio QSRs: Unification with Extended QSRs by Viewing Angle

A few of the 3C QSRs were quite different from the extended doubles. 3C 273 is a prime example; it was a compact source, clearly elongated but only by a few arc minutes. And several QSRs, like 3C 345, were so compact they were completely unresolved, even at the 5-km telescope in Cambridge. The spectrum of these sources also was a bit different. Unlike the steep-spectrum extended sources, the radio spectrum of the compact sources was rather flat, often $f_\nu \sim \text{constant}$ ($f_\nu \propto \nu^0$), leading to the name flat spectrum radio QSRs or FSRQs. Their high polarization indicated synchrotron emission, but it was difficult to imagine a single population of relativistic electrons that produced a flat spectrum.

Bright, unresolved sources, however, were ideal candidates for observing with VLBI. The resulting images revealed a radio structure quite different from the extended radio quasars (see Fig. 2.8, center): nearly all FSRQs contained a core (less than one milliarcsecond in size and unresolved even with VLBI) and a single radio jet perhaps a few tens of milliarcseconds long. The core presumably marked a site close to or at the position of the central engine, and the jet must be a small version, or the inner portion, of the giant jets seen in extended QSRs. Indeed, in 3C 273 it pointed in the same direction as the larger radio and optical jets (Fig. 2.8, left). In some cases the VLBI jet could be resolved into several components, each having a different classical synchrotron spectrum. The “flat” spectrum of the source was

apparently due to the superposition of several synchrotron spectra that made up an expanding, conical jet of relativistic plasma. But there was no hint of a “counterjet” in the other direction. Either the counterjet did not exist, or it was so faint that it could not be detected.

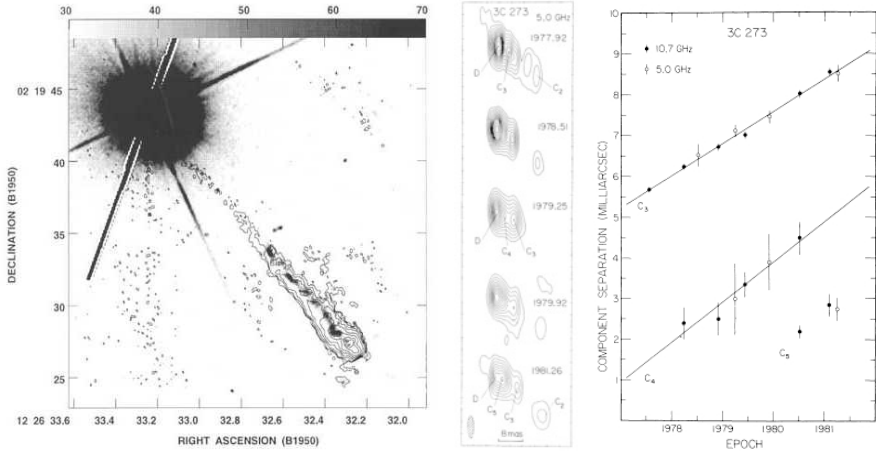


Fig. 2.8: Superluminal expansion of the jet in the compact radio quasar 3C 273. *Left:* Composite Hubble optical (gray scale) and MERLIN radio (contours) image of the quasar jet [57]. *Center:* Successive VLBI maps of the central milliarcsecond jet over the period 1977–1981; note motion of jet components C_3 and C_4 with respect to the central core D. *Right:* separation of jet components from the compact core vs. time [58]. At a redshift of 0.158 (distance of 670 Mpc), these apparent component velocities are about $9c$. Reproduced by permission of the AAS.

2.2.3.1 Relativistic Beaming and Superluminal Motion

In a landmark paper, Roger Blandford and Arieh Königl (then at Caltech) proposed in 1979 that radio sources with extended jets and bright compact radio sources were one and the same type of object, differing only by their “viewing angle” to the jet axis [59]. The former were viewed at a large angle to the line of sight, while the latter were viewed almost end-on to the jet. The extreme brightness of the compact jet, and faintness of the counterjet in FSRQs⁷, could be explained if the jet plasma were flowing at very near the speed of light *and* if the jet were pointed nearly at the earth. Such a “bulk” relativistic jet would emit most of its radio emission not in every direction like a star, but rather in a small cone in the direction of the jet flow. And, because of the relativistic motion, that radiation would be strongly Doppler boosted to very high apparent brightness. This process is called “relativistic beaming”. Like

⁷ Unification by viewing angle also explained many other properties of compact radio sources, including their radio flux, angular structure, variability, and polarization.

extended QSRs, compact QSRs still could have two opposing jets, but we should only see the one pointing toward us, not the one pointing away from us.

Before going on, it is important to distinguish between the random relativistic motion of the electrons, for which we have used the Lorentz factor γ_e , and the bulk relativistic motion of the jet, for which we will use

$$\gamma_{\text{jet}} \equiv \left(1 - \frac{v_{\text{jet}}^2}{c^2}\right)^{-1/2} \quad (2.4)$$

γ_e refers to the small-scale, disordered motion of the electrons and is similar to thermal motion of particles in a gas (but at very relativistic, nonthermal speeds). γ_{jet} refers to bulk motion of the jet and is similar to a blowing wind or other gas flow. The former can reach values of 10^3 – 10^5 , while the latter gets no higher than 100 or so, and often is much less.

Concluding that the jets were pointed nearly along our line of sight initially raised the “Copernican” problem – that somehow we lie at the center of the universe with all FSRQs pointing at us. Relativistic beaming solved this problem, however. With so many QSRs distributed randomly in the universe, a few jets should be pointing at us⁸. If those jets were traveling near the speed of light, they would appear as some of the brightest radio sources in the sky, and we would tend to pick them out first in early radio surveys like the 3CR sample.

If relativistic beaming were the explanation, then the jets also should show another property, predicted by Martin Rees of the Institute of Astronomy in Cambridge, UK, in 1966: they should exhibit “superluminal motion” [60]. That is, if we measure the apparent transverse distance traveled by the jet, and divide by the time it takes to go that distance, we will get a speed that is *faster* than the speed of light! Superluminal motion is an optical illusion, of course; nothing can travel faster than light. It is caused by a computational mistake on our part. We measure the correct distance that the jet travels, but we divide by the wrong time (the time that has lapsed on earth) to calculate the speed. Rees’s prediction was confirmed in the 1970s by radio VLBI observations done by Marshall Cohen and his colleagues at Caltech [61, 62] with radio visibility data, and then in 1981 by the same team (led by Tim Pearson [63]) with a time series radio VLBI images, from which actual speeds could be measured. Fig. 2.8 shows the brightest components in the 3C 273 jet (components #3 and #4) and how they moved away from the core during the period 1978 to 1981. The distance traveled by these is ~ 8 parsecs (26 ly), for a derived speed of $\sim 8.7 c$. This is, indeed, superluminal motion.

The reason the jet appears to be traveling faster than the speed of light, then, is that it is actually traveling very close to c and pointed very close to our line of sight. Our calculation above failed to take into account that the radio waves detected in 1978 traveled toward us only a little faster than the jet itself. By the time those waves reach us, it is almost time for the second set of radio waves, which we detected in

⁸ An approximately equal number of other jets also will be pointing at civilizations in other galaxies.

1981, to reach us also. The distance the jet traveled between those two points is

$$\Delta x = v_{\text{jet}} \sin \theta \Delta t$$

where θ is the angle, in radians, of the jet's axis to our line of sight. But the time we measure between those two events is not Δt . Because the second set of radio waves arrives on the heels of the first, the apparent time between the two images is compressed by the difference in speeds (toward us) of the waves (c) and the jet ($v_{\text{jet}} \cos \theta$)

$$\Delta t_{\text{app}} = \left(1 - \frac{v_{\text{jet}}}{c} \cos \theta \right) \Delta t$$

which is smaller than Δt when θ is small. (If the speed of light had been infinite, or v_{jet} had been small, then Δt_{app} and Δt would have been virtually equal.) The apparent speed we calculate is the ratio of Δx and Δt_{app}

$$v_{\text{app}} = \frac{v_{\text{jet}} \sin \theta}{1 - \frac{v_{\text{jet}}}{c} \cos \theta} \quad (2.5)$$

This expression has a maximum for a given v_{jet} when the observer's viewing angle is given by $\cos \theta = v_{\text{jet}}/c$. At that angle, the value of the apparent velocity is $v_{\text{app}}^{\text{max}} = \gamma_{\text{jet}} v_{\text{jet}}$, which can be much greater than the speed of light when $v_{\text{jet}} \rightarrow c$. *This is the standard explanation of superluminal motion*: QSR jets are traveling close to the speed of light, so the viewing angle must be small ($\theta \approx \cos^{-1}(v_{\text{jet}}/c) \ll 1$).

Jet motion toward or away from an observer is often described by the “Doppler factor”

$$\delta_{\text{jet}} \equiv \gamma_{\text{jet}}^{-1} \left(1 - \frac{v_{\text{jet}}}{c} \cos \theta \right)^{-1} \quad (2.6)$$

It is this factor that determines the relativistic beaming of both the jet ($\theta \approx 0$) and counterjet ($\theta \approx \pi$). For a QSR, with a power law synchrotron spectrum of

$$f_{\nu} \propto \nu^{\alpha}$$

the brightness of its jets will be amplified (or diminished) by a factor

$$f_{\nu, \text{app}} = f_{\nu} \delta_{\text{jet}}^{3-\alpha}$$

In the ratio of jet to counterjet brightness the γ_{jet} factors cancel, giving

$$R \approx \left(\frac{1 + \frac{v_{\text{jet}}}{c} \cos \theta}{1 - \frac{v_{\text{jet}}}{c} \cos \theta} \right)^{3-\alpha}$$

For jets also traveling near the speed of light, and close to our line-of-sight, $R \approx (4\gamma_{\text{jet}}^2)^{3-\alpha}$. So if a QSR jet has a Lorentz factor of $\gamma_{\text{jet}} = 10$ ($v_{\text{jet}} = 0.995c$) and a spectrum with $f_{\nu} \propto \nu^0$ ($\alpha = 0$), then the jet-to-counterjet ratio R is 64 million! No

wonder no counterjet was detected in these compact radio sources. Even such a jet not pointing toward us (say, at $\theta \sim 45^\circ$) will be 200 times fainter than one pointing nearly at us.

2.2.3.2 End-on Lobes: Unification of Compact and Extended QSRs

In addition to superluminal motion, if a compact QSR is, in fact, the inner portion of an extended QSR jet seen nearly end-on, we also should see emission at lower radio frequencies associated with the extended source. In some cases we still will see two radio lobes, but with the distance between them foreshortened by the small angle of the source to our line of sight. Some objects, however, should be so end-on that their extended radio lobes should appear as an amorphous blob surrounding the jet that is pointing at us. David Murphy of the Jet Propulsion Laboratory, Ian Browne of Jodrell Bank, and Rick Perley of NRAO studied several FSRQs with the VLA telescope and indeed found amorphous emission surrounding the sources [64]. Using the low-frequency emission as a measure of a source's "unbeamed" extended luminosity, and the high-frequency emission as a measure of the beamed jet portion, they were able to show that the compact QSRs are, in fact, similar to ones like 3C 47, with most of their radio luminosity in the extended emission. This is a good example of *unification by angle to the line of sight*, another parameter that has turned out to be even more important than brightness contrast. Extended and compact QSRs appear different only because of the direction from which we view them.

Murphy, Browne, and Perley's results not only supported the beaming hypothesis, it also showed that FSRQs, while probably FR II sources seen end-on, were not always surrounded by the most powerful of extended sources. In fact, they tended to be ones only somewhat above the FR I–FR II break and more like moderately powerful FR II radio galaxies than QSRs. This was another indication that radio galaxies and QSRs are probably the same type of object, and that viewing angle might have something to do with whether or not we see a compact optical source.

2.2.3.3 Helical Jets

As VLBI and larger-scale (MERLIN and VLA) images of sources began to accumulate, Readhead and his colleagues noticed an interesting effect in 1978 [65]. While many sources had their VLBI and large-scale jets well aligned, a nearly equal number had the compact and extended jets *highly* misaligned. In sources like 1823+568 (Fig. 2.9, top), the main axis of the helix was traced by the extended jet, so one could measure the angle between the two (extended and compact jets) to determine the "misalignment angle". Furthermore, the effect was not random; most sources were either well-aligned (0° difference) or highly misaligned (90° difference) [66, 67]. (See the histogram in Fig. 2.9, lower left.)

In 1993 John Conway of Caltech and David Murphy proposed that this misalignment was caused by helicity in the jet motion. The amount of helicity does not have to be high in order to explain the strong misalignment of the jets, *if* the jets are viewed nearly end-on and undergoing relativistic motion (Fig. 2.9, lower right). A simple explanation of this helicity effect is as follows. We only see those parts of the relativistic jet that are pointing within a few degrees of our line of sight. When parts of the jet point more than that away from us, they are difficult or impossible to detect – almost as faint as a counterjet. In the case of a helical (corkscrew) jet, viewed nearly along the axis of the helix, we see only a short portion of the helix, and that visible portion will appear perpendicular to the helix main axis.

While the beaming hypothesis and jet helicity explained misalignments, what produces the jet helicity was not yet understood. In addition to QSRs, objects like M87, when imaged with very high-resolution VLBI from the ground (Fig. 2.2) or space, show that the inner portions of their jets have helical structure (“wiggles”).

2.2.3.4 Blazars: Looking “Down the Pipe”

Long before astronomers knew that other galaxies existed, they had developed catalogs of stars, particularly ones that were variable. Many of these variable stars (eclipsing, pulsating, exploding) provided important clues to the formation and evolution of all stars. Some, however, turned out not to be stars at all. One in particular, BL Lacertae⁹, turned out to be coincident with a flat spectrum, compact radio source at a redshift of $z = 0.069$. It lay at the center of an active galaxy, whose jet happened to be pointing right at us and emitting optical synchrotron light. Variations in the jet flow produced variations in the optical light, making the object appear as a variable star.

Since its identification by Bryan Andrew, John MacLeod, and John Schmitt (of David Dunlap Observatory in Ontario, Canada) as a peculiar radio source in 1968 [72, 73], many other “BL Lac” objects have been discovered. Their defining characteristics are a stellar appearance, like QSRs, but with a very red polarized optical continuum spectrum and weak or non-existent emission lines. Occasionally the continuum weakens, perhaps because the jet shuts down or temporarily points away from our direction, allowing emission lines from a quasar-like object to be detected [74]. Otherwise, redshifts to these objects can be determined only by detecting absorption lines from gas in the galaxy hosting the BL Lac object.

Because their optical properties make them hard to distinguish from the myriads of red stars in the sky, BL Lac objects are usually discovered at other wave-

⁹ According to a convention worked out by Friedrich W. A. Argelander in 1862, variable stars in a given constellation, like Lacerta, are given names in order of their discovery starting with the letter R [70]. The second discovered is S Lacertae, and so on. The tenth begins the two-letter designation RR, RS, *etc.*, which continues through SS (the 19th) and on to ZZ (the 54th). Strangely, the sequence starts over at AA (the 55th variable star observed in that constellation), continues to through BB (the 79th) and on to QZ (the 334th), always excluding the letter J. Beyond QZ, one uses V335, V336, *etc.* BL Lac was the 89th variable “star” discovered in the constellation Lacertae (by Cuno Hoffmeister in 1929 [71]).

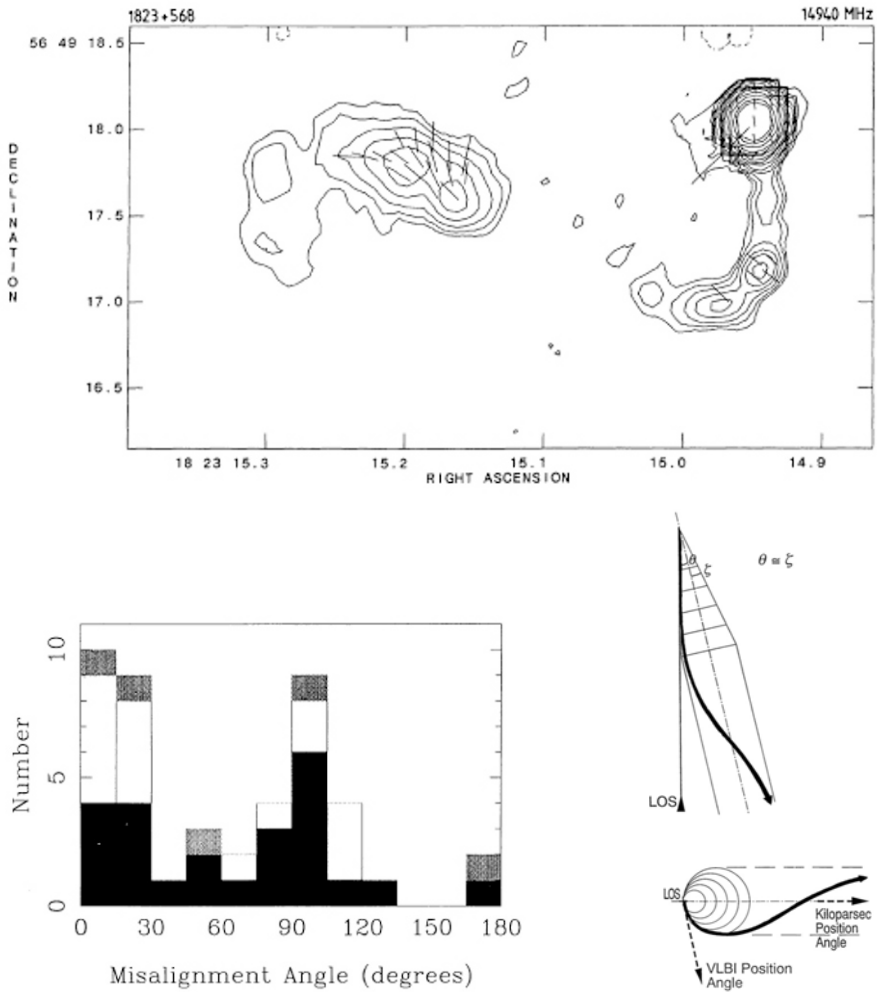


Fig. 2.9: The Conway–Murphy helical jet model for misalignments of jets in quasars. Unlike NGC 6251 and 3C 273 (Figs. 2.4 and 2.8), some quasar jets on the parsec scale appear to be dramatically misaligned with the jet structure observed on the kiloparsec scale. *Top:* VLA map of 1823+568 [68], showing a 90° misalignment between the compact structure (pointing South) and the extended structure (pointing East). *Lower left:* Histogram of 45 quasars showing predominantly sources that are aligned, or 90° misaligned (from [69] using data in [67]). *Lower right:* Helical jet model showing how primarily orthogonal misalignments occur in helical jets that point almost directly toward the earth: unless the line of sight (LOS) is directly along the jet, the large-scale kiloparsec structure will appear to one side; but the helical jet components beamed most strongly toward the observer will lie along an angle approximately normal to that extended structure (from [69]). Reproduced by permission of the AAS.

bands – radio or X-ray – and then identified optically. It was soon recognized that radio-selected BL Lacs and X-ray-selected BL Lacs were different types of objects. Complicating the picture, some also were detected not only in the radio to X-ray wavebands, but also at million-electronvolt (MeV) γ -ray energies, at GeV (billion eV) energies, or at TeV (trillion eV) energies. Clearly very high-energy photons could be produced by a doubly-relativistic jet, made of a synchrotron gas with particles moving randomly in a magnetic field at Lorentz factors of 10^3 - 10^5 , and with bulk relativistic motions toward us at Lorentz factors of 10 or more.

The resolution of all these observations represents one of the triumphs of black hole research of the 20th century. While many investigators contributed observations and ideas, it was the work of a team of Italian investigators (Giovanni Fossati, Laura Maraschi, Annalisa Celotti, Andrea Comastri, and Gabriele Ghisellini, from Trieste, Milan, and Bologna) that was primarily responsible for sorting out all the data and developing the current model [75, 76]. This field also has given rise to a general class of active galactic nuclei with one of the more clever names in astronomy: blazars, combining the name of the original BL Lac object and the name “quasar”, plus conveying the physical nature of these blazing objects, as powerful beams of relativistic particles which, by chance, happen to be pointed right at us. Because of their special orientation, the sources we observe as blazars comprise less than 1/10,000th of all active galaxies with jets (and those are only about 10% of all active galaxies and quasars). Yet, blazars reveal a wealth of knowledge about jets that is applicable to all radio sources and, perhaps, ultimately even to radio-quiet objects and Galactic objects as well.

Fig. 2.10 shows the spectrum of several different types of blazars, including a “normal” FSRQ. Note the horizontal axes: emission is detected from radio frequencies through the infrared, optical, ultraviolet, X-ray, and γ -ray, up to several TeV.¹⁰ Though peaking in different wavebands, the spectra are actually very similar, composed of a low-frequency peak, somewhere in the far-infrared to X-ray range, and a high-frequency peak, somewhere in the MeV to TeV range. The first peak is due to direct synchrotron emission by the relativistic electrons. The second is due to the Compton effect: the fast electrons collide with photons emitted by the broad line region or by dust in the nucleus of the QSR’s host galaxy, bumping the photon up to γ -ray energies.

There also is a trend for the highest power objects (FSRQs and radio-selected BL Lac objects) to have peaks at lower frequencies (synchrotron peak in the far-infrared and synchro-Compton peak in the 10 MeV to 1 GeV range). These are called Low-frequency BL Lac objects (LBLs for short) or “red” blazars. The less luminous objects (mainly X-ray selected BL Lacs) have their synchrotron peaks in the ultraviolet or soft X-ray energies and synchro-Compton peaks at 100 GeV to TeV energies. The latter objects (high-frequency BL Lacs) are referred to as HBLs or blue blazars.

¹⁰ Interestingly, the wavelength of a single TeV photon is about 10^{-18} meters (one attometer) and its energy is about 1.6 ergs (about the kinetic energy of a large crawling insect). This energy may be small by human standards, but by subatomic standards it is extremely large.

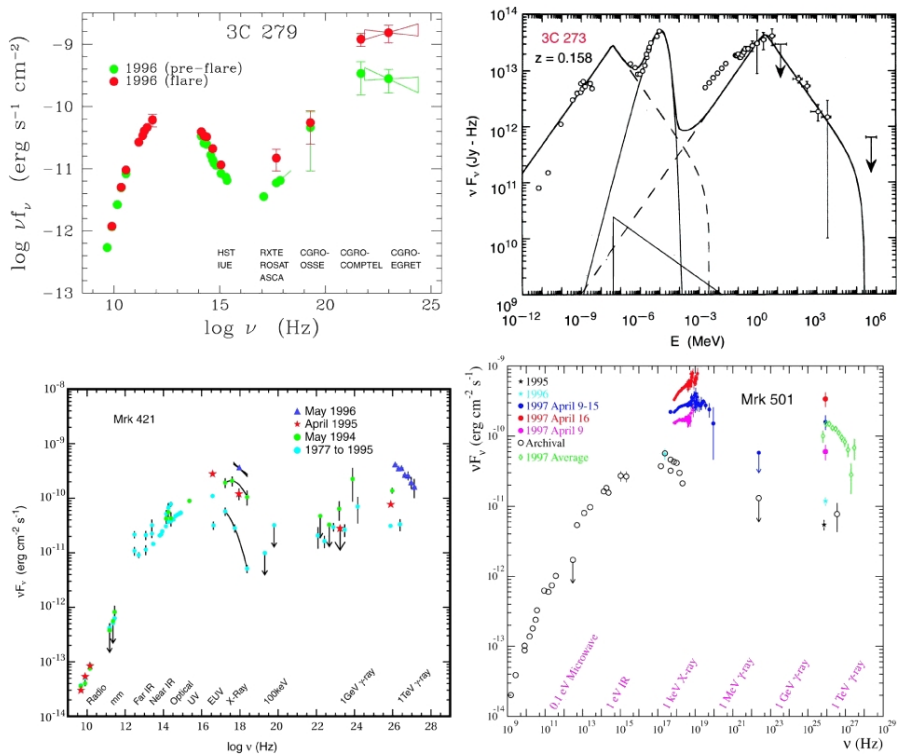


Fig. 2.10: “Broad-band” spectra of several different kinds of blazars, from radio to TeV γ -rays: 3C 279 (LBL), 3C 273 (FSRQ); Mrk 421 (HBL); Mrk 501 (HBL). Credit: A.E. Wehrle/M.A. Catanese/J.H. Buckley/Whipple Collaboration. Reprinted from Fig. 3.1 of the GLAST Science Document [77, 78].

As we can see from Fig. 2.11, the color designation describes not only the positions of the emission peaks but also their optical colors near 10^{15} Hz (steeply decreasing with frequency for red blazars, flat or steeply rising for blue blazars). This color–luminosity trend appears to be due to a more efficient “cooling” of the jet by the synchro-Compton effect as the jet increases in power. Low-power jets are not cooled well, leaving their electrons “hotter” (more relativistic) and emitting at a higher frequency. Because all sources with radio jets appear to be unified, these properties probably describe jets in extended radio galaxies and QSRs, even though we cannot observe the relativistic motion and γ -ray emission in them.

2.2.3.5 IDV Blazars

Over many decades of time nearly all blazars are seen to vary in their synchrotron emission. However, a few are seen to vary on time scales less than one day. These

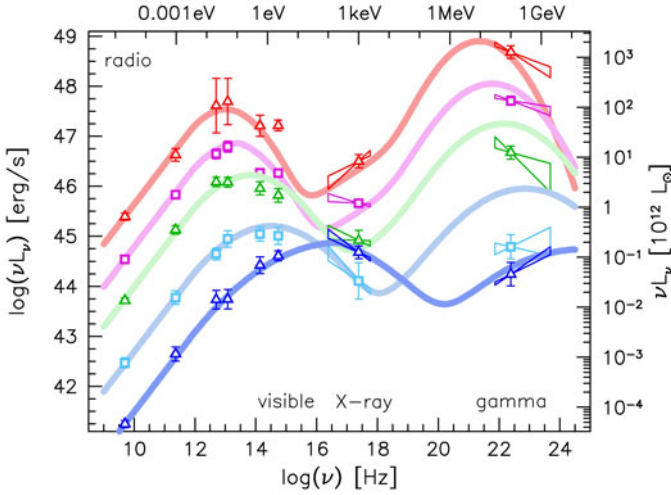


Fig. 2.11: Fossati’s unification diagram for blazar spectra. FSRQs (red blazars) are the most luminous and lowest in energy peaks, followed by the intermediate LBLs, and finally the least luminous, but highest energy peaked, spectra of the HBLs (blue blazars). Kindly prepared especially for this book by G. Fossati; reproduced with permission.

are the “intra-day variables” or IDVs. If this variability were intrinsic to the jet itself, it would imply enormous bulk Lorentz factors for these jets. The argument is as follows. Changes on less than an hour imply that the jet is less than a light-hour (10^{14} cm) in size. From the measured flux of the IDV radio source, we would deduce a brightness temperature of 10^{18-21} K. This is enormous, since a synchrotron emitting, relativistic gas is expected to produce intrinsic emission at about $T_{b,intr} = 10^{11}$ K and certainly no higher than 10^{12} K [79]. One way in which high apparent brightness temperatures could be produced is if the jet were moving toward us at relativistic speeds. In that case, the apparent brightness temperature deduced from variability would be [59]

$$T_{b,app} = \delta_{jet}^{3-\alpha} T_{b,intr}$$

While we do believe bulk relativistic motion to be occurring, the deduced Lorentz factors from this equation are uncomfortably high – several hundred or more. This is far faster than any other observed jet source.

The best explanation for the rapid variability, however, is not that the radio jet itself is varying on intra-day time scales. Instead, the source is simply so small that it is scintillating as its light passes through our Galaxy’s interstellar medium – in much the same way as stars twinkle as we view them through our atmosphere. The jet still must be compact, a size on the order of 5–8 microarcseconds, but not the few nanoarcseconds implied by the variability, bringing the implied apparent brightness temperature down to about 10^{13-14} K. A non-varying jet flowing toward us at a Lorentz factor of 5–50 could produce brightness temperatures of this magnitude.

While still high, these values for γ_{jet} are only a factor of a few larger than the largest observed superluminal motions and also less than the speeds inferred for those other fast black hole jets, the γ -ray bursts. IDVs, therefore, may represent the extremes of viewing angle close to the line of sight or, alternatively, the extremes of jet speed in extragalactic radio sources.

2.2.4 Cosmic Evolution of Radio Galaxies and QSRs

2.2.4.1 Evolution of Extended Radio QSRs

As more and more redshift distances were obtained for 3C QSRs, an important trend became apparent. They were not distributed uniformly in space or time. There apparently were more at higher redshifts ($z > 1$), when the universe was young billions of years ago, than at the present epoch ($z < 0.2$). One measure of this apparent evolution in space density is the “ V/V_{max} ” test of Peter Kafka (of the Max Planck Institute for Physics and Astrophysics) and Maarten Schmidt [80, 81]. In this test one chooses a quasar from the complete sample of 3C quasars and computes the volume of the universe V from the earth (and the Milky Way) out to that quasar. One also computes for that quasar the *maximum* volume V_{max} that it could have in the 3C sample, given the limits on the radio and optical fluxes placed on the sample.¹¹ The ratio of these two, V/V_{max} , gives a measure of how close that quasar is to the Milky Way compared to all other quasars of the same luminosity.

For example, if a quasar were at a redshift of $z = 0.3$, then its volume V would be

$$V = \frac{4}{3}\pi R^3 = \frac{4}{3}\pi \left(\frac{cz}{H_0} \right)^3 \quad (2.7)$$

where R is the distance to the quasar and we have used $H_0 = 75 \text{ km s}^{-1} \text{ Mpc}^{-1}$ as the Hubble constant of cosmic expansion. And, if that quasar had a measured flux of 4 janskys, while the entire sample had been surveyed to a limiting flux of 1 Jy, then we could give that particular quasar a just barely visible flux of 1 Jy = 4 Jy $(R/R_m)^2$, if we moved it out to a maximum distance $R_m = 2R$. The maximum volume then would be

$$V_{\text{max}} = \frac{4}{3}\pi (2R)^3 \quad (2.8)$$

For this quasar, then, $V/V_{\text{max}} = 0.125$. Equations (2.7) and (2.8) actually are valid only in a flat universe or for $z \ll 1$ (i.e., $R \ll c/H_0$). For high redshift galaxies and QSRs, the more sophisticated equations of cosmology need to be used for V and V_{max} .

¹¹ Depending on the quasar’s actual optical and radio luminosities, as one mentally moves that quasar outward, further than its actual distance, it gets fainter in both apparent radio flux and apparent optical magnitude. At some point, either the apparent radio or optical brightness reaches the minimum value for the quasar sample, and at that point the maximum volume is reached.

For a sample of objects that are uniformly distributed in space, the average of $\langle V/V_{max} \rangle$ over all objects in the sample should be 0.5, even when we take into account the complexities of cosmology and the expanding universe. For the entire sample of 3C QSRs, however, Maarten Schmidt found an average of [81]

$$\langle V/V_{max} \rangle = 0.70 \pm 0.04 \quad (2.9)$$

Equation (2.9) implies that, even after dividing the density of QSRs at each redshift by $(1+z)^3$ (to take into account the fact that the universe was much smaller in the past than it is now), the “co-moving” density of QSRs still increases with redshift by a large factor $\rho(z) \propto (1+z)^6$ or, in terms of the “lookback” time $\tau_0 - t(z)$,

$$\rho(z) = \rho_0 e^{[\tau_0 - t(z)]/\tau} \quad (2.10)$$

where ρ_0 is the density of QSRs right now (at $z = 0$, within a billion light years of the Milky Way), $\tau_0 \approx 13\text{--}14$ Gyr is the present age of the universe (1 Gyr = 1 billion years), and $t(z)$ is the age of the universe at redshift z . The time constant $\tau \approx 2.7$ Gyr is a measure of how fast the density of QSRs evolves with time.

2.2.4.2 Evolution of Extended Radio Galaxies

Schmidt also studied the evolution of the extended radio galaxy population with time and found a relation like equation (2.10), but where τ is different for galaxies of different radio luminosity [82]. For low-powered FR I sources, there appeared to be almost no evolution ($\tau > 34$ Gyr!), but for the most luminous FR II sources there was rapid evolution ($\tau \approx 1.7$ Gyr). High-power radio galaxies not only look like radio QSRs, they also evolve like radio QSRs. *In the distant past, high-power radio sources were as numerous (or more numerous) than low-power ones today.*

2.2.4.3 Evolution of Compact QSRs

Because compact QSRs have a much flatter spectrum ($\alpha > -0.5$) than that of extended QSRs ($\alpha < -0.7$), a much better place in the radio to look for FSRQs is in the high-frequency bands – 2.7 GHz, 5.0 GHz, and even 15 and 22 GHz. Surveys in these bands have been done at NRAO, Parkes Radio Observatory in Australia, Max Planck Institute for Radio Astronomy in Germany, and the Mullard Radio Astronomy Observatory in Cambridge, UK.¹² Surprisingly, it was found that flat-spectrum compact QSRs do not seem to evolve as fast as steep spectrum, extended sources –

¹² A new “C-band All Sky Survey” is now underway using radio telescopes at Owens Valley Radio Observatory in California and Hartebeesthoek Radio Astronomy Observatory (HartRAO) in South Africa; it is a joint collaboration among Caltech, Universities of Oxford and Manchester, HartRAO, and the King Abdulaziz City for Science and Technology (KACST) in Saudi Arabia.

with an average $\langle V/V_{max} \rangle$ of only 0.52 (essentially no evolution at all) [83]. This result can be understood easily as an effect of relativistic beaming [84, 85]. Beaming tends to pick out those jets pointing close to our line of sight, but not necessarily the most powerful radio sources. A weaker radio galaxy or QSR, if beamed toward us, would appear brighter at high frequency than a very powerful steep spectrum FR II source. Because τ is longer for the less powerful radio sources, statistically we expect those selected at high frequency to appear to evolve more slowly.

Part of this slower evolution for flat spectrum sources also might be due to a beaming angle effect. If the jet speed γ_{jet} is faster for the more powerful radio sources, then their beaming angle ($\theta_b = 1/\gamma_{jet}$) will be narrower than that for weaker sources. This would make it even more likely that we would pick out less powerful sources in a high-frequency survey. And, if sources were more powerful in the past than now, it also would make it less likely that we would pick out compact sources at high redshifts, making it appear that compact sources were less numerous early in the universe than extended sources.

2.2.4.4 Evolution of Blazars: Unification with Extended Radio Galaxies

Because models of relativistic synchrotron-emitting jets pointed toward us work so well in explaining blazars, they clearly must be part of a much larger population of jet sources, which are *not* pointed at us. But what is the parent population: FR I or FR II sources only, all radio galaxies and QSRs (FR Is + FR IIs), or all radio-loud and -quiet macroquasars perhaps? The answer appears to be the second choice, all radio sources. In fact, most rather bright blazars seem to be rather modest FR I radio galaxies, just viewed down the jet “pipe”. The low-frequency extended radio emission of BL Lac, for example, would put it in the weak FR I class. Beaming of the jet makes it appear much brighter than it really is. Furthermore, BL Lac objects as a class do not appear to evolve much in space density over cosmological time scales. Again, this is similar to the FR I sources.

Of course, if most blazars are actually weak radio galaxies and if a few of them are modestly powerful FSRQs, then a very few of them probably are the very strong FR II QSRs seen end-on. If we can identify only those, then we should see a strong evolutionary trend with most of them occurring at high redshift and dying out in recent times, just as we see happening to extended FR II radio galaxies and QSRs.

In addition to cosmological evolution, one also can investigate evolution of the sources on shorter time scales – specifically the turning on or off of their jets on time scales of only 10 million years. In that regard, it is tempting to identify the sequence FSRQ→LBL→HBL, or its inverse, as an evolutionary one. However, it also is possible that this sequence is one of viewing angle (HBL being a jet viewed very close to the line of sight), or simply a sequence of different jet powers. Disentangling viewing angle, evolutionary state, and intrinsic effects, and being able to infer true jet luminosities, are important goals for blazar research in the next few years.

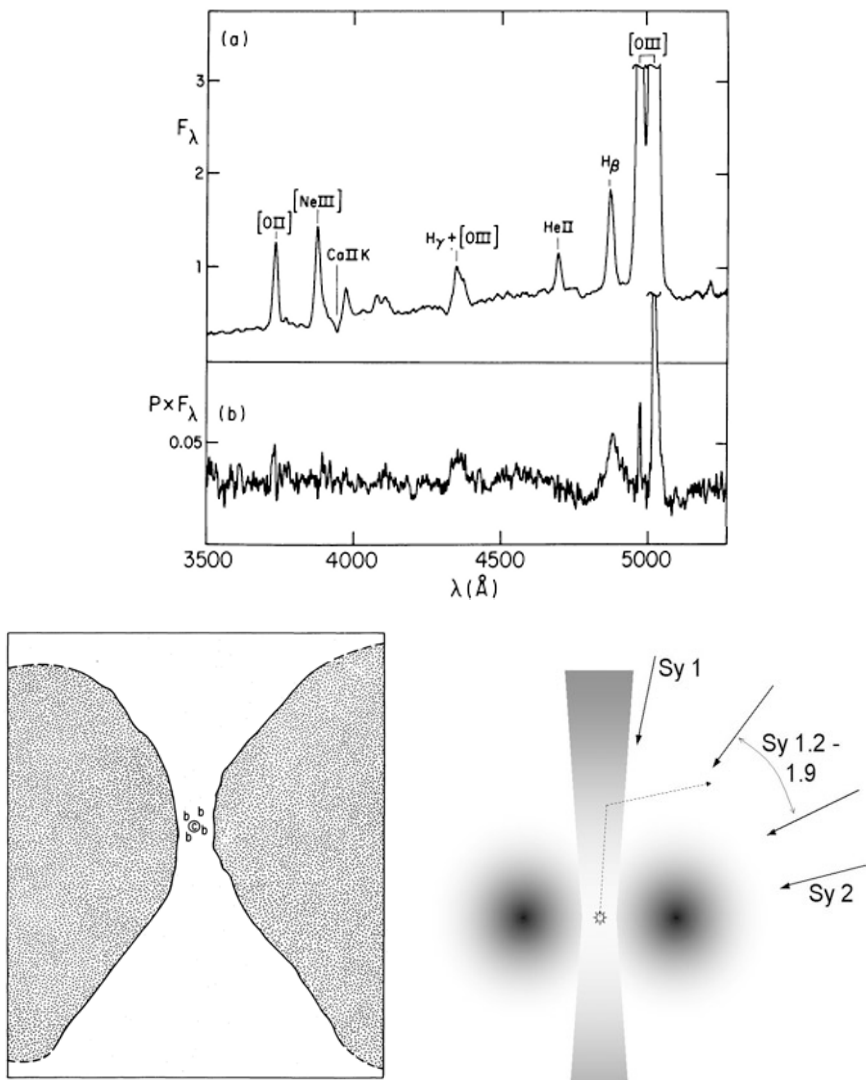


Fig. 2.12: Antonucci and Miller model of the Seyfert sequence [86]. *Top:* Spectrum of Seyfert 2 galaxy NGC 1068 in unpolarized light (upper) and polarized light (lower). The polarized spectrum resembles a Seyfert 1 spectrum (note broad $H\beta$ line at $4861 \text{ \AA}$). *Lower left:* Original Antonucci and Miller concept for an obscuring dusty torus; C represents the central engine and the “b”s represent broad line-emitting clouds. *Lower right:* Modern view of the Seyfert sequence, with scattering medium along the torus axis; dashed line represents scattered light from the central Seyfert 1 nucleus. Seyfert 1 galaxies have an unobstructed view of the central source and narrow-line-emitting gas above and below the torus, while Seyfert 2 galaxies have only the narrow-line emitting gas directly visible with the central source seen in weak scattered light only. Years ago, various intermediate Seyfert galaxies were classified as Seyfert 1.2–1.9 objects, and thought to be observed at intermediate angles. However, that method of classification is no longer widely used; one now speaks only of “Type 1” or “Type 2” AGN. Reproduced by permission of the AAS.

2.3 Seyfert Galaxies and Quasi-Stellar Objects: Supermassive Black Holes with Weak or No Jets

2.3.1 Unification of Classical Seyfert Galaxies by Viewing Angle

In 1985, shortly after the proposed unification of compact and extended radio quasars by viewing angle by Blandford and Königl, Robert Antonucci (then at NRAO) and Joseph Miller of Lick Observatory proposed a similar, and equally important, unification of all Seyfert galaxy types [86]. Many of the properties of the Seyfert sequence could be explained if they consisted of a standard Seyfert 1 nucleus whose very small broad line region was surrounded by a “dusty torus” – a doughnut-shaped cloud of interstellar gas and dust – that was perhaps 100 pc or so in radius (see Fig. 2.12), but the much larger narrow-line region lay outside that torus, probably in an “ionization cone” along the axis of the torus. All Seyfert nuclei, then, had basically the same structure, but were viewed from different polar angles. Seyfert 1 nuclei, for example, were viewed from above the torus, where there was little absorption of the central nucleus by the dusty torus. Seyfert 2 galaxies were viewed from a high inclination angle, so that the torus completely obscured the broad line region and only allowed viewing of the large narrow-line region.

Three important discoveries confirmed the Antonucci and Miller picture. First, Antonucci and Miller showed that, for the Seyfert 2 galaxy NGC 1068, a very compact broad line region is still visible *in scattered light*, even though it is obscured directly by the dusty torus. Light scattered by free electrons tends to be polarized. So, if only the polarized light is studied in a Seyfert 2 nucleus, one often can see a Seyfert 1 spectrum (see Fig. 2.12). The second important result was the direct optical imaging of the narrow-line, ionization cone. This was done first from the ground (sometimes aided by interferometric techniques) [87, 88] and then with direct imaging by the Hubble Space Telescope [89]. The top panel of Fig. 2.13 [90] shows another Seyfert 2 galaxy, NGC 5728, in the [O III] 500.7 nm line. This confirmed not only that the NLR was extended, but also that it was created by ionizing radiation *and* was shadowed by a dusty torus everywhere except along the torus’ axis. The third result that confirmed this model was direct imaging (in reflection and also by Hubble) of the dusty torus itself in objects such as NGC 4261 (see bottom panel of Fig. 2.13) [91]. This last confirmation occurred in 1993 – eight years after the existence of a dusty torus was first suggested by Antonucci and Miller.

2.3.2 The X-Ray Spectrum of Seyfert Nuclei

2.3.2.1 Comptonized Power Law, Reflection Spectrum, and Iron $K\alpha$ Emission

The observed X-ray spectrum of Seyfert galaxies is rather “hard”, consisting of a power law ($f_{\nu X} \propto \nu^{-1}$) with a high-energy cutoff near a few hundred keV. Such a

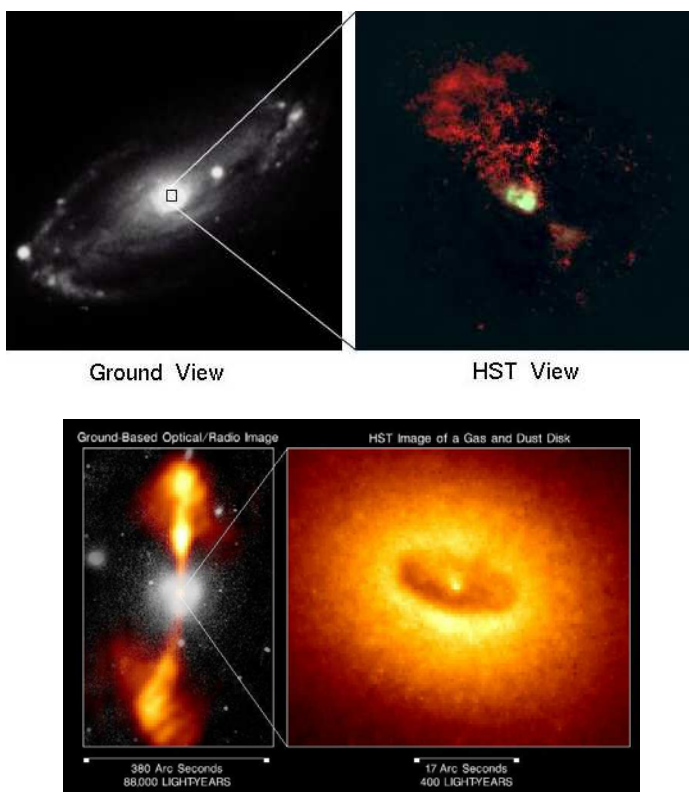


Fig. 2.13: Two Hubble views of dusty tori. *Top two panels:* Seyfert 2 galaxy NGC 5728 [90]; the right-hand HST view shows the bipolar ionization cone (red narrow-line region) and the reflection of the central Seyfert nucleus (bright greenish area); also note the obscuration by the dusty torus at the center of the cone and perpendicular to it. *Bottom:* Radio galaxy NGC 4261 [91]; here the dusty torus is viewed from above and, without a strong Seyfert nucleus, can be seen in reflected light from the surrounding stars and gas emission. Reproduced by permission of STScI.

spectrum is thought to be produced when photons emitted by an optically thick, cool (10^{6-7} K) disk are Compton up-scattered by a hot, optically thin (10^9 K) plasma. In Seyfert 1 galaxies there also appears to be a ‘reflection’ component, peaking at about 20–30 keV ($50\text{--}72 \times 10^{17}$ Hz), which is produced when the very hard X-rays from the power-law source strike cold gas and then are down-scattered. In Seyfert galaxies of types 1.5–2, there is considerable absorption at lower X-ray energies (a few keV) due to a large number of cold atoms along the line of sight between the observer and the X-ray emitter ($N_H > 10^{22} \text{ cm}^{-2}$). This result is consistent with the above dusty torus model. Seyfert 1 nuclei, on the other hand, rarely show any X-ray absorption by cold material.

In addition to the continuum radiation, there is strong line emission from iron at about 6–7 keV. Iron has a $K\alpha$ fluorescence line in this energy range, occurring when

hard X-rays knock out an inner K shell electron, and a remaining outer shell L shell electron then falls back down to the K shell, emitting a photon as the newly-ionized atom readjusts. For neutral iron, this photon is at 6.4 keV. As the iron becomes more ionized, the $K\alpha$ fluorescence line's energy approaches 6.9 keV. Iron also has recombination lines in this energy range, with a particularly strong one at 6.7 keV that can complicate the issue. However, for the lower-luminosity sources ($L < 10^{44}$ erg s $^{-1}$), the iron lines are consistent with 6.4 keV (neutral iron fluorescence). In the higher-luminosity sources, approaching QSO luminosities, the iron line occurs at energies closer to 7 keV and is significantly weaker, which is consistent with a hotter (and smaller) fluorescing iron component.

Models of these Seyfert spectra [92, 93] generally assume that the original hard power law emission is produced by a geometrically thick, hot disk region or corona near the center of an accretion disk around the black hole. A cooler disk, possibly on the equator, then reflects and reprocesses this emission, producing the 20–30 keV bump and the fluorescent iron emission. The presence of the reflection bump in Seyfert 1 spectra, in which the disk should be viewed almost face on, and the lack of such a bump in Seyfert 2 spectra, in which the accretion disk should be more edge-on (or even obscured), are also consistent with the unified Seyfert model.

In 1995 Yasuo Tanaka and his colleagues at the Institute of Space and Astronautical Science in Tokyo, Japan, used the Japanese X-ray satellite ASCA (Advanced Satellite for Cosmology and Astrophysics) to observe the Seyfert 1 galaxy MCG-6-30-15 [94]. They found that, while the emission line peaked at the expected 6.4 keV and had little emission above that, the line had an unusually broad red wing that extended down to below 4 keV. This broad emission is most likely caused by the relativistic gravitational redshift in the deep potential well of the central black hole and by the Doppler redshift of the material rotating in the relativistic accretion disk around the hole. This confirmed that the iron line in Seyfert 1 galaxies is indeed formed very near the black hole in the accretion disk. In addition, detailed analysis of this and other Seyferts with broad red iron line wings indicates that rotational velocities of the disk are so high that the disk must approach extremely close to the black hole – well within the last stable orbit radius for a non-rotating black hole (6 GM/c ^{2}). As we shall see in Chapters 7 and 12, if this analysis is correct, these observations indicate that the central black hole is spinning rather rapidly.

2.3.2.2 The Warm Absorber

Because the iron line emission is a product of reflection and/or reprocessing, one also might expect it to be weak or absent for Seyfert 2 galaxies. Instead, just the opposite is true: the iron line is much stronger in Seyfert 2s than in Seyfert 1s. There is, therefore, an indication that some, and maybe all, of the observed iron $K\alpha$ emission in Seyfert 2 galaxies might be produced in cold material that lies on the inner far side of the dusty torus rather than very near the black hole. When the torus is viewed in Seyfert 2s from the side, this material is seen in reflection, while the $K\alpha$ emission produced near the black hole may be attenuated or absorbed entirely.

When viewed in Seyfert 1s from above the dusty torus, the reflection will be much less, as the $K\alpha$ emission produced near the black hole will not be absorbed.

Such ‘warm’ ($T < 10^5$ K) absorbing gas was seen in Seyfert 1s in the mid 1980s with the Einstein X-ray observatory and was given the name “warm absorber” by Jules Halpern (then at Caltech and the Smithsonian Astrophysical Observatory in Cambridge) [95]. With the Chandra observatory, this gas is now becoming well studied. While there is clearly iron absorption above 6.4 keV, most of the absorption lines occur well below 3 keV from a variety of atomic species (sulfur, aluminum, magnesium, iron, neon, *etc.*). They indicate that the gas is, indeed, flowing out toward us with a velocity of a few hundred kilometers per second and with a similar (probably turbulent) velocity dispersion. And, as expected, in Seyfert 2s this warm absorber gas is seen in emission when the ionization cone is viewed from the side.

2.3.3 *Narrow-Line Seyfert 1 Galaxies*

A recent addition to the Seyfert family of galaxies (discovered in 1985 by Osterbrock and Richard Pogge at Lick) are those whose nuclear spectra have the same *kind* of lines as those in Seyfert 1 nuclei, but their permitted lines are rather narrow (well under 2000 km s^{-1} , compared to several thousand km s^{-1} for normal Seyfert 1s) [96]. These galaxies are more than just unusually narrow-lined Type 1s. They are a separate class, distinguished by their unusually soft X-ray emission, peaking around 0.2 keV (5×10^{16} Hz), making them more like extreme ultraviolet sources than X-ray sources. One attractive interpretation of this new class is that these narrow-line Seyfert 1 galaxies (NLSy1s) are in a high-luminosity “soft” X-ray state, similar to the soft state of binary X-ray sources (see Chapter 3). Classical X-ray hard Seyfert galaxies, then, would be black holes in a lower-luminosity hard accretion state. If this suggestion is correct, then we might expect Seyfert nuclei to undergo transitions between the intermediate and high/soft states, just as we see in the lower-mass black holes in binary star systems. That is, we might expect to see some classical, hard X-ray Seyfert 1 galaxies turn into NLSy1s and the latter sometimes change into classical Seyfert 1s. However, whereas we often see such state changes occur over a half-hour for a $10 M_{\odot}$ black hole, we might expect this to take a million times as long, 10–100 years, for a ten million solar mass black hole at the center of a Seyfert galaxy.

2.3.4 *Quasi-Stellar Objects*

2.3.4.1 BSOs and Type 1 QSOs: More Unification by Brightness Contrast

Shortly after the discovery of radio QSRs, a large effort was made to try to find similar objects using optical techniques only. Images of regions of the sky were

taken in different color filters, and stellar objects with colors bluer than normal stars were selected from those images. (A blue color indicated that the “star” likely had strong ultraviolet emission, just like most radio QSRs.) Many of these “blue stellar objects” or BSOs turned out to be indeed identical to QSRs optically, but most had no detectable extended radio emission.

Invariably, however, depending on redshift, the BSOs had strong, broad optical and ultraviolet emission lines at the frequencies of (redshifted) hydrogen Balmer β (486.1 nm or 6.2×10^{14} Hz), Lyman α (121.6 nm or 2.5×10^{15} Hz), Carbon IV (154.9 nm or 1.9×10^{15} Hz), Silicon IV (139.7 nm or 2.1×10^{15} Hz), and Magnesium II (270.0 nm or 1.1×10^{15} Hz). This suggested even more powerful QSO search techniques, developed in the mid 1970s by Patrick Osmer and Malcolm Smith of the Cerro Tololo Observatory in Chile, that selected objects for their broad emission lines rather than their blue continuum emission [97]. (Emission line searches are more powerful than continuum searches because, when they exist in a spectrum, emission lines are usually much brighter than the continuum.) To perform emission line searches, a prism (and sometimes a finely-grooved metal grating as well) was placed in front of an imaging telescope to disperse the light of an entire portion of the sky, $5^\circ \times 5^\circ$ in size. These “objective prism” and “grism” surveys allowed the light from each star in the field to be spread into a short rainbow. Objects with broad emission lines easily stood out, and their spectra could even be scanned right off the photographic image to obtain approximate redshifts. (See Fig. 2.14.)

By the early 21st century, these and other optical techniques had discovered over 10^4 QSOs, ranging up to redshifts of 6.3. Believe it or not, these most distant quasars are receding at a cosmological expansion speed that is almost twice the speed of light!¹³ Their calculated distances are more than 4 Gpc and lookback times more than 13 Gyr. With an age of the universe of about 14 Gyr, these quasars are being viewed when the universe was only 5% as old as it is now.

2.3.4.2 Type 2 QSOs, LIRGs, ULIRGs, HyLIRGs, and SMGs

QSOs (which include all such objects selected using optical techniques) turned out to be a remarkably homogeneous set of objects. Until the 21st century (see [99] and references therein), there were no Type 2 QSOs known – only Type 1 objects, virtually identical to Seyfert 1s, but without the obvious galaxy host. The rarity of Type 2 QSOs was somewhat of a mystery until the launch of the Spitzer Space Telescope (JPL’s Michael W. Werner, project scientist (P.S.)) in 2003. However, there was no lack of explanations for it. One explanation offered was that any Seyfert nucleus bright enough to be a QSO also was bright enough to evaporate the dust and ionize

¹³ A complete discussion of why real (not apparent) superluminal speeds can be achieved in cosmology is not in the scope of this book. However, it is important to know that, in the case of the expansion of the universe, depending on how we define distance, this is indeed possible. Superluminal cosmic expansion does not violate relativity; indeed, the general relativistic theory of the universe demands it. This is a rather unknown fact among astronomers, since they tend to work with the cosmological redshift (which will remain finite even when $v > c$).

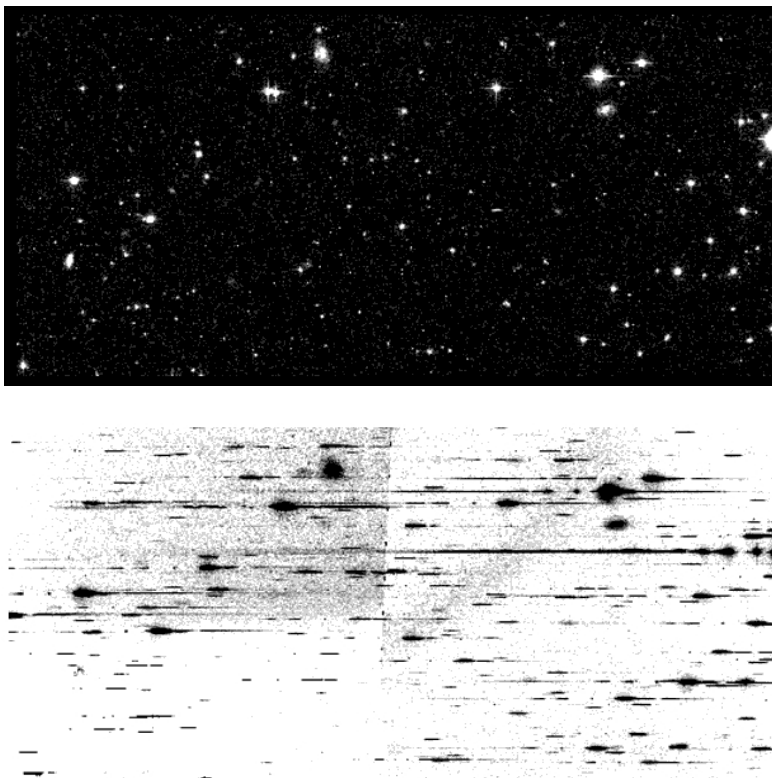


Fig. 2.14: Example of an objective prism / grism field. *Top:* A direct image of a sky field. *Bottom:* Negative of a dispersed image of the same field, using the HST grism. Note how some objects' spectra are dominated by bright features (emission lines). Reproduced from Figs. 6.10 and 6.11 of [98], by permission of STScI.

the gas in any torus around it. Another suggestion was that Type 2 QSOs *do* exist, and they are some of the famous objects discovered by the InfraRed Astronomy Satellite (IRAS) in the 1980s called Luminous InfraRed Galaxies or LIRGs.¹⁴

LIRGs are galaxies that are enormously bright in the infrared, with luminosities above $10^{11} L_{\odot}$ ($> 4 \times 10^{44} \text{ erg s}^{-1}$) or several times that of an average galaxy. ($L_{\odot} = 3.839 \times 10^{33} \text{ erg s}^{-1}$ is the average solar photon luminosity.) In fact, among high-luminosity objects, LIRGs are the most abundant species, exceeding the numbers of Seyferts and QSOs that have similar power output [100]. If a large amount

¹⁴ Occasionally, some authors may use Roman numerals (I and II) to designate optical AGN classes instead of Arabic numerals (as in Seyfert 1 and 2). This new terminology can, of course, cause considerable confusion with the Fanaroff and Riley classification scheme for radio-loud AGN, so one should be careful to avoid it and refer to "Type 1" or "Type 2" AGN, whether they are Seyferts or QSOs. Some of this usage still occurs in the literature, however, so when perusing papers on AGN, the reader should understand *very* clearly to which types of object the Is and IIs refer.

of infrared radiation is detected from a celestial source, it usually indicates the presence of a dust-enshrouded object or objects. When dust intercepts the ultraviolet and optical radiation from stars or quasars, the radiation is absorbed, heating the dust to a temperature of 20–100 K. The dust then radiates as a black body, at this temperature, causing the original optical–UV emission to be re-radiated in the far-infrared. The term Ultra-Luminous InfraRed Galaxy (ULIRG) refers to even more luminous infrared galaxies with output above $10^{12} L_{\odot}$ ($> 4 \times 10^{45} \text{ erg s}^{-1}$), and HyLIRG (hyper-luminous) refers to those rare objects above $10^{13} L_{\odot}$ ($> 4 \times 10^{46} \text{ erg s}^{-1}$). This is similar to the optical luminosity of many QSRs and QSOs, so some LIRGs could be enshrouded accreting black hole systems. On the other hand, this amount of emission also could be produced by a large stellar system that is undergoing a massive amount of star formation (“one million Orions”) in the LIRG galactic nucleus [101, 102]. Because a LIRG’s radiation is so thoroughly reprocessed by the dust, it is difficult to tell whether a “starburst” or QSO is buried there, or, in fact, whether *both* are going on at the same time. The current view is that the lower-luminosity LIRGs are simply large starbursts, creating 10–100 stars per year in the nucleus, while the higher-luminosity objects must be at least partially powered by a central black hole. However, given that almost every large galaxy appears to have a black hole in the center, because of all the gas and dust around the LIRG nucleus, it is likely that every LIRG’s black hole is active, although maybe not always the dominant contributor to the luminosity.

Objects similar to the ULIRGs were discovered by the Submillimeter Common-User Bolometer Array (SCUBA) instrument used on the James Clerk Maxwell Telescope (JCMT) operated in Hawaii by the UK, Canada, and the Netherlands [103]. These galaxies are at redshifts $z > 1$, and also have dust emission that peaks at about $100 \mu\text{m}$ in their rest frame. However, because of the redshift, this infrared emission appears in the submillimeter region (wavelengths of several hundred micrometers) when it is observed; so they are called SubMillimeter Galaxies or SMGs. The SMGs are ULIRGs at high redshift, but their space density at those redshifts is far higher than the space density of ULIRGs today by a factor of several hundred, similar to that of the Lyman break galaxies that are seen at optical wavelengths [104].¹⁵

Mid-infrared observations by JPL’s Spitzer Space Telescope in the first decade of the third millennium confirmed that many LIRG-type objects are, indeed, obscured Type 2 AGN. More importantly, they also put us on a path toward understanding how galaxy construction and the growth of supermassive black holes are linked, and they gave us insight into the major events in the life of a quasar. One of the big mysteries in X-ray astronomy has been the nature of the extragalactic X-ray background. Its spectral shape might be explained by many QSOs in the early universe, but the numbers of detected quasars (by the year 2000) fell far short in this endeavor – by a factor of four or five! It was predicted that 75–80% of QSOs must be hidden from view of optical telescopes and must lie at $z > 1.5$ (see, *e.g.*, [107]).

¹⁵ SMGs, therefore, seem to be the submillimeter primeval galaxies predicted by Michelle Kaufman in 1976 [105], just as the Lyman break and Lyman alpha galaxies are the optical–UV primeval galaxies predicted by this author in the same year [106].

When Spitzer came on line, it was quickly realized by Daniel Stern of JPL that known optical AGN and QSOs could be recognized by their mid-IR colors using the $3.6\mu\text{m}$, $4.5\mu\text{m}$, $5.8\mu\text{m}$ and $8.0\mu\text{m}$ bands [108]. (This selects *against* objects dominated by star formation.) However, it also quickly became apparent that there were many more detected Spitzer sources with these same properties that were *not* visible in the optical [108]. These “obscured AGN” were, indeed, a factor of 3–4 more numerous than the unobscured kind that had been discovered in the 1960s [109, 110]. Even more recently, the all-sky survey of JPL’s Wide field InfraRed Survey Explorer mission [111] (UCLA’s Edward L. Wright, principal investigator (P.I.); JPL’s Peter M. E. Eisenhardt, P.S.) has identified about a thousand HyLIRG candidates out to redshifts beyond $z \sim 2$ [112, 113]. And, based on rough estimates, there may be at least a few thousand times more than that (*i.e.*, several millions) of ULIRGs throughout that same volume.

While this is all very encouraging for solving the X-ray background, the “smoking guns” (good hard X-ray spectra of many AGN) still have not been collected. This will be the job of the JPL/Caltech NuSTAR satellite (Caltech’s Fiona A. Harrison, P.I.; Daniel Stern, P.S.), to be launched in 2012.

It is fairly clear, now, that galaxies and black holes grow together, largely during phases of very rapid star formation [114]. And it also appears that most rapidly accreting black holes (*i.e.*, AGN/QSOs) do so, at least initially, deep within their star-forming parent galaxy, heavily enshrouded in dust created by the continually occurring supernova explosions. The quasars that were originally discovered in the 1960s, therefore, seem to be ones that already have lived out 75–80% of their lives and are on their way to becoming optically faint, low-luminosity AGN (Section 2.4).

2.3.4.3 Type 3 QSOs?

Physically, there is a clear difference, of course, between a QSO whose central emission is blocked simply by a dusty torus along our line of sight (a true Type 2 object) and one that is completely enshrouded in dust (as may, indeed, be the case with HyLIRGs, ULIRGs, and perhaps even many LIRGs). For this reason, it has been suggested [115, 116] that QSOs whose high-ionization optical emission is hidden from view, no matter from which direction they are observed, be placed in a new class called Type 3 AGN.

Type 3 objects are defined observationally [115] as those that are selected as candidate luminous AGN using *infrared* colors, but having observed *optical* spectra that are more typical of low-ionization LINERs or H II nuclei (see Section 2.4.1 below). At the present time, however, this new classification scheme is not in wide use, and efforts are still ongoing to establish whether a genuine Type 3 class exists or not. This author believes that this is a very useful evolution in the terminology and supports continued work to refine it. Its use would clear up confusion in discussions of hidden AGN by distinguishing those objects that may have an observable NLR (Type 2) and all the effects associated with it (*e.g.*, beaming of radiation *not* in the

direction of the observer) from those that may not and, instead, may be radiating in the infrared, more or less, isotropically (Type 3).

2.3.4.4 Radio-Loud and Radio-Quiet QSOs: The Radio Loudness Ratio

As large samples of QSOs were discovered, the objects were subsequently observed in the radio using the VLA and other radio telescopes. Some original radio QSRs were re-discovered. However, it soon became clear that the QSRs comprised only a small fraction (10%) of all optically-selected QSOs. Most supermassive accreting black hole systems, it seems, do not produce a powerful jet. This fact can be made quantitative if we calculate for each QSO a ratio R of radio to optical emission

$$R \equiv S_r/S_o \quad (2.11)$$

where S_r and S_o are the radio and optical flux densities of the quasar (typically at 1.4 GHz and at the optical B or i bands, respectively), both measured in janskys.¹⁶ “Radio-loud” quasars (RLQs or just QSRs) are defined as $R > 100$ while “radio-quiet” quasars (RQQs, which sometimes is used as a name for QSOs) are those with $R < 100$. This definition works well when we consider only the low-frequency extended radio emission below 1 GHz. However, at frequencies above 5 GHz, the effects of beaming can be significant, so the R value will appear different for an object viewed from one direction than from another. Fig. 2.15 shows the R distribution function for QSOs. Early in the study of QSOs, it was claimed that the R distribution function may be “bimodal”, with a peak in the radio-quiet region and a peak in the radio-loud region. This issue still has not been resolved fully. The figure shows one determination of the R distribution with no bimodality, although it does show a long tail at high R . On the other hand, another determination shows that R may, indeed, be bimodal, although not nearly at the level that was originally thought. The final resolution of this is important: if bimodality were to be confirmed, it would mean that powerful radio QSRs are different types of objects from the radio-weak QSOs, while no bimodality would mean that all QSOs are essentially the same, but have a wide range of radio to optical luminosity, with no preferred ratio of jet power to accretion power.

2.3.4.5 Host Galaxies of QSOs

Because of the similarity of RQQs and RLQs to Seyfert and N-galaxies, it had been suspected since their discovery that the former were simply brighter versions of the

¹⁶ Note that R is *not* the ratio of luminosity in the radio to luminosity in the optical. S_r and S_o measure luminosity *per unit frequency*, so there is an additional factor of $\nu_o/\nu_r \approx 10^5$ in equation (2.11). For example, for $R = 100$, the luminosity ratio is only $L_r/L_o = 10^{-3}$. Even the most radio-loud quasars, with R close to 10^4 , still only have about 10% of their luminosity in the radio, and that is not counting the energy radiated in X-rays.

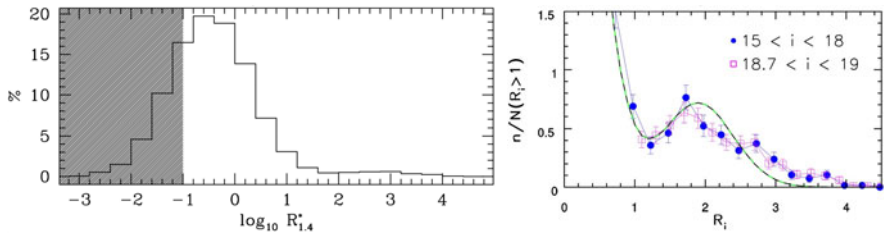


Fig. 2.15: Two determinations of the QSO radio-to-optical distribution, with different results. *Left:* Analysis by the SISSA group at Trieste, Italy, derived from 195 optically-selected QSOs from the 2nd Hubble Deep Field, the Large Bright Quasar Survey, and the Palomar-Green survey. (Reproduced from Fig. 4a of [117], by permission of J. Wiley and Sons, Ltd.) Note the very broad tail at very high R ratios (10^{2-5}); the results are consistent with the R -distribution being monotonic (no separation between the two curves). *Right:* Analysis by the Sloan Digital Sky Survey team of about 10,000 SDSS QSOs. In this case some bimodality is seen. Reproduced from Fig. 2b of reference [118], by permission of the ASP.

latter. However, confirming that quasars are indeed active galactic nuclei proved very difficult using ground-based observations, as the galaxies in true quasars were overwhelmed by the brightness of the central source. As discussed earlier, many such ground-based studies were done in the 1970s and 1980s. Some quasar hosts were detected, but many were not. Because of this, some people remained unconvinced that QSRs and QSOs were simply active galaxies with a particularly strong optical nucleus.

Because of this on-going controversy, one of the first tasks accorded to the Hubble Space Telescope after its repair in late 1993 was to confirm the presence of faint host galaxies surrounding most quasars. Initially, it seemed that the ground-based observations had been wrong; host galaxies could not be found. But this turned out to be a data-reduction problem. The question was finally settled in 1994–96 when John Hutchings’ team and another team led by John Bahcall, of the Institute for Advanced Study, alternately submitted papers that presented detailed images of galaxies around QSOs, including some at high redshift [119, 120, 121, 122, 123]. The clarity and resolution of the HST observations convinced most skeptics that quasars are indeed AGN. In some cases even the redshift of the host galaxy could be measured, using the Keck Observatory in Hawaii, and the QSR and galaxy redshifts were indeed the same, proving that quasar redshifts are cosmological. The luminosities of quasars are indeed as large as 10^{46} – 10^{47} erg s $^{-1}$, and Wheeler’s suggestion that a black hole was powering the engine was upheld as a very possible explanation for these luminosities.

The general picture that has emerged from these studies is that host galaxies of RLQs tend to be very bright (similar in magnitude to giant ellipticals), while the hosts of RQQs tend to be 0.5–1 magnitude fainter (more typical of giant spirals).

2.3.4.6 BAL QSOs: Yet More Unification by Viewing Angle?

Another intriguing sub-class of QSOs is the Broad Absorption Line or BAL QSOs. These comprise about 10% of optically selected QSOs (*i.e.*, roughly the same percentage as the QSRs), and show broad *absorption* lines in their spectra, indicating outflows with velocities of 5000–50,000 km s⁻¹. Some BAL QSOs have “P Cygni” profiles (named after the famous star with a strong outflow, P Cygni); these absorption troughs are *attached* to the broad emission lines of the QSO. In many cases, however, the broad absorption line is *detached* from the emission line – several thousand km s⁻¹ wide, with 10,000 km s⁻¹ or more separating the two. (See Fig. 2.16 and Fig. 2.17.) Generally these absorption lines are from high-ionization atomic species (C IV, Si IV, N V, O VI) only; these are called HiBALs. But in about 10% of the BAL cases (1% of QSOs), low-ionization lines of Mg II or Al III also appear, and these are referred to as LoBALs. Even more rare are the “iron LoBALs”, which show Fe II absorption in addition to the other low-excitation lines.

X-ray observations of BAL QSOs help to elucidate the difference between HiBAL and LoBAL QSOs. HiBALs (*i.e.*, *most* BAL QSOs) are consistent with the outflowing gas and the X-ray absorber being one and the same medium. In LoBAL QSOs, however, there appears to be additional absorption of the X-rays, which is due to additional gas and dust beyond that creating the absorption troughs. LoBAL QSOs, therefore, may be those viewed through a thin, dusty (but not obscuring) torus and, therefore may be another candidate for the missing Type 2 QSOs. If this is the case, the dusty torus in QSOs still exists, but it has a much smaller relative size than it does in Seyfert galaxies.

The nature of the BAL QSOs, and their relation to the entire QSO population, is perhaps more uncertain and controversial than that of the QSRs. The standard model for BAL QSOs (proposed in 1995 by Norman Murray, James Chiang, and their colleagues at the Canadian Institute for Theoretical Astrophysics (CITA) in Toronto) is that these objects are normal QSOs that are viewed nearly edge-on to the accretion disk that is feeding the black hole (Fig. 2.16) [125]. An outflowing wind ablates matter off the disk forming the absorption line clouds. This model has a number of attractive features. Analysis of the absorption lines indicates that the absorbing gas surrounds only about one-fifth of the central QSO – a covering factor of 0.20. This means that there must be five times as many QSOs with such winds (or ~50% of all QSOs) in which we will not see absorption lines. So half, and maybe all, QSOs may have strong winds, but we only observe as BALs those with their disks edge-on. Detailed numerical simulation of disk winds driven by radiation pressure from their surface by Daniel Proga (now at University of Nevada, Las Vegas) and his colleagues shows that this model can, indeed, produce the right sort of wind, including the generation of “hitchhiking gas” needed to shield the absorbing clouds from the central X-ray emitting region [126, 127].

However, the radiatively-driven disk wind model does have some problems explaining many BAL QSO properties. The first BALs to be discovered were all radio-quiet. That is, they did not seem to be drawn from the same population as optically-selected QSOs. In recent years, moderately radio-loud BAL QSOs, and even a cou-

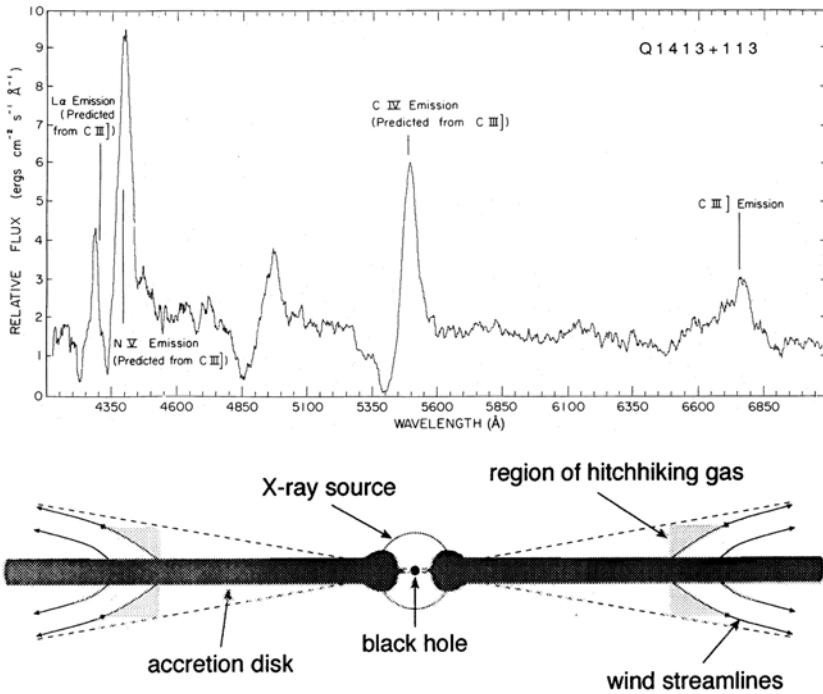


Fig. 2.16: Typical spectrum of a BAL QSO with “attached” absorption troughs and the standard BAL QSO model. *Top:* Note outflow in Q1413+113 at up to 12,000 km s⁻¹ (reprinted from [124]). The line analysis of Turnshek *et al.* indicates full absorption along the line of sight, but that the “covering factor” or “covering fraction” is small (<0.35). One suggestion is that many or all QSOs may be BAL objects, but we only view a few of them with the outflowing absorbing clouds between us and the central continuum source. *Bottom:* A radiation pressure-driven disk wind model for QSOs. This model assumes that BAL objects and QSOs are the same, but BALs are those viewed nearly along the disk surface. (Reproduced from Fig. 1 of [125], by permission of the AAS.)

ple of FR II BAL QSOs, *have* been discovered, but still in a much lower percentage than all QSOs. Clearly, the BAL QSOs and FR II QSRs tend to avoid each other in ways that may be difficult to explain by an equatorial wind. Moreover, the existence of some BAL QSOs with detached troughs (Fig. 2.17) requires the low-velocity material to *not* be along the line of sight, a difficult thing to accomplish. An alternative, but less popular, model involves a *bipolar* magnetohydrodynamically-driven wind model (Fig. 2.17). A bipolar wind model has most of the attractive features of the disk wind model, is more consistent with modern accretion models and simulations, *and* explains the radio-loud/BAL avoidance: QSRs with very powerful jets are likely to have much narrower bipolar winds and, therefore, a much lower incidence of BAL detections.

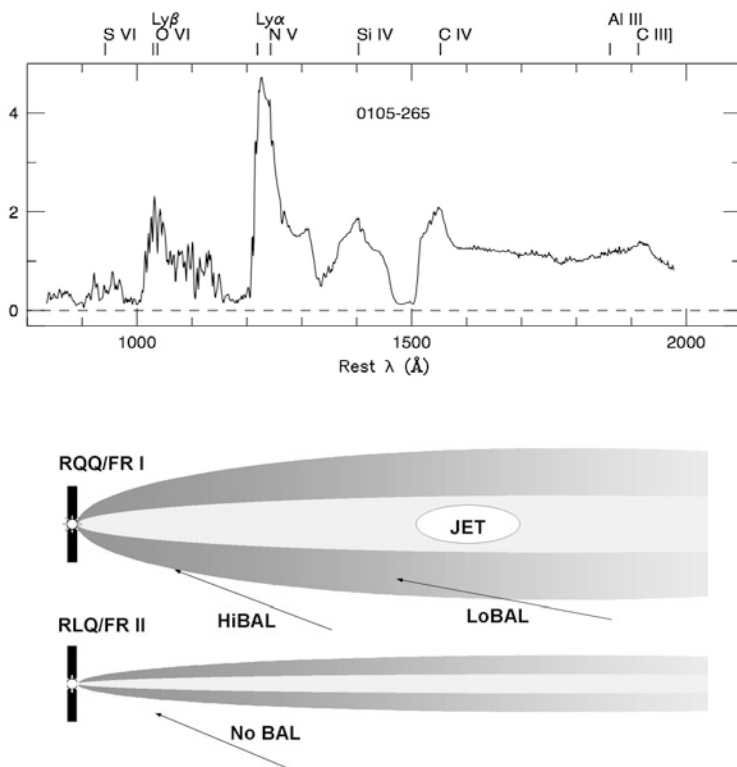


Fig. 2.17: A peculiar BAL QSO and an alternative bipolar wind model (suggested by D. Turnshek, private communication). *Top:* A detached trough BAL QSO [128]. *Bottom:* A bipolar MHD wind model with RQQ/FR I and RLQ/FR II structures. Such objects would appear as LoBALs when viewed with the lower-ionization wind material (Al III, Mg II, Fe II) obscuring the central source, as HiBALs when viewed through the higher-ionization wind material (C IV, S IV, N V), and as a normal QSR when viewed at a large angle from the jet axis. The probability of viewing a strong FR II jet through the optically obscuring material would be very small, if any such thermal material even existed in a powerful FR II jet. Reproduced by permission of the AAS.

2.3.4.7 The X-Ray Spectrum of Radio-Quiet QSOs

While the observed X-ray properties of QSRs are dominated by properties of the beamed jet, that of radio-quiet QSOs is dominated by the disk emission, just as it is in the related Seyfert 1 objects. The basic X-ray spectrum is similar – $f_{\nu X} \propto \nu^{-1}$ power law above about 1 keV, with a high-energy cutoff – but the reflection components (20–30 keV bump and iron line) appear to be greatly diminished or absent. And when an iron line is detected, its energy is closer to 6.7 keV than to 6.4, similar to high-luminosity Seyfert 1s [129]. The reduced reflection bump could be an indication that there is less cool gas in these higher-luminosity sources than in

lower-luminosity Seyferts. And, as in Seyfert 1s, a higher energy emission line is also consistent with a hotter, and perhaps smaller, fluorescing iron component.

2.3.5 Cosmic Evolution of Seyferts and QSOs

Maarten Schmidt's 1972 studies on quasar evolution also investigated the optically-selected QSO population. Its evolution appeared similar to that of the radio-selected population in equation (2.10) (essentially pure density evolution with a time constant of about 2.7 Gyr). Currently, little is known about the cosmological evolution of Seyferts and other lower-luminosity AGN. By analogy with radio galaxies, however, one might expect the lower-luminosity radio-quiet population to evolve much more slowly than its high-luminosity counterparts (luminosity-dependent density evolution). This question may be resolved by deriving the time-dependence of the X-ray luminosity function from the X-ray background. But, until it is, the cosmological evolution of the Seyfert population will remain an important missing constraint on models for grand unification of all macroquasars.

2.4 Low-Luminosity Active Galactic Nuclei (LLAGN)

In the 1980s and 90s, as instruments became more sensitive and as the QSO picture unfolded, astronomers began to look for lower-luminosity counterparts to the powerful quasars. To their surprise, the fainter they looked, the more active galactic nuclei they found. Whereas QSOs span the range 10^{45-47} erg s⁻¹ and Seyfert galaxies roughly 10^{42-44} erg s⁻¹, low-luminosity AGN (or LLAGN) are generally below these luminosities. They occur in up to half of all known galaxies and are thousands of times more common than QSOs [131]. Some very familiar objects, often observed by amateur astronomers, are in the LLAGN category, including M81, M84, M104, and the nucleus of our own Galaxy. Another famous one is NGC 4258, in which we can see an edge-on disk of gas that allows us to very accurately measure the black hole mass (36 million solar masses).

2.4.1 Dwarf Seyferts, LINERs, Transition-type, and H II Nuclei

As the luminosity of an AGN drops, so apparently does the ionization state of the gas near the black hole. Lines of singly-ionized oxygen become stronger than doubly-ionized, for example, and even un-ionized species of atoms become important. The brighter of these objects (*i.e.*, with luminosities of 10^{41-42} erg s⁻¹) are classified as “dwarf” Seyfert galaxies. But the majority of these objects (*i.e.*, those $<10^{41}$ erg s⁻¹) are classified as Low-Ionization Nuclear Emission-line Regions or

LINERs – another clever AGN name, first proposed by Timothy Heckman of Johns Hopkins University [130]. As the central engine becomes even weaker, it becomes more difficult to distinguish it from the surrounding dense nucleus of stars. In fact, if there is some star formation in the nucleus, the ultraviolet light from new hot O and B stars can ionize the surrounding gas to a low state, just as in H II regions like the Orion nebula. (This problem of distinguishing between AGN and starburst is reminiscent of the LIRG $\rightarrow$ HyLIRG sequence, but now at the low-luminosity end of the active galaxy spectrum instead of the high end.) Objects in which it is clear that both star formation and a weak AGN are operating are called “transition-type” nuclei [131]. About 11% of nearby galaxies are considered dwarf Seyferts, about 20% are LINERs, and about 13% are LINER/H II transition nuclei. The other 55% are either pure H II nuclei or have no discernible sign of nuclear activity at all.

LINERs share many features with their more luminous cousins. Like Seyferts, they appear to divide into Type 1 and Type 2 objects. The former have broad emission line winds indicating gas velocities of 5000 km s^{-1} or higher – velocities that generally can occur only a small fraction of a parsec from a central black hole. In some Type 2 objects a broad wing is detected, but only in polarized (reflected) light, indicating that a mini-Seyfert-like nucleus is there, but it is hidden by a dusty torus.

2.4.2 *Radio Cores*

Most LLAGN, including our own Galactic nucleus, have some detectable radio emission. Such objects are often referred to as “radio cores”. The origin of this radio emission is the subject of a hot debate: is it a weak radio jet, like powerful radio galaxies, or does the emission come from the accretion flow itself, either originating in the corona or in a bloated optically transparent torus? Resolving this controversy will be difficult. It will take VLBI observations of extraordinary sensitivity and resolution. The latter is becoming rapidly available as instruments probe shorter millimeter and submillimeter wavelengths. However, great sensitivity must await large intercontinental arrays with many radio telescopes.

2.4.3 *Weak-Lined Radio Galaxies*

In addition to the radio cores, some LLAGN are, in fact, radio-loud and have powerful kiloparsec-scale jets. M87 is a notable example and was discussed previously. The optical luminosity of these weak-lined radio galaxies (or WLRGs) lies below those of N-galaxies in the same way as LINERs lie below Seyferts. The optical properties of WLRGs are quite similar to those of LINERs, so the former may be the radio-loud counterpart to the latter. As do all radio-loud galaxies, WLRGs occur exclusively in giant elliptical or S0 galaxies – ones with a large spheroidal bulge of stars.

2.4.4 Sgr A*: The Quiescent Black Hole Engine at the Center of Our Galaxy

Even by LLAGN standards, our Galaxy's central black hole engine is extraordinarily weak and anemic. Shining at a rate of only $10^{-9} L_{\text{Edd}}$ ($\sim 10^{35.7} \text{ erg s}^{-1}$), it is less luminous than many X-ray binary black holes and even many regular stars.¹⁷ The central black hole is often referred to as “Sgr A*”, because it produces the star-like radio emission in the brightest radio source in the constellation Sagittarius. Several scientific meetings and books have been devoted to the Galactic center. Here we only will touch on those properties that help us relate Sgr A* to the much more powerful black hole engines in other galaxies.¹⁸

If one were to travel to the Galactic center, about 100 pc from the black hole one would encounter a system of dense molecular clouds. Embedded within this cloud system is the “nuclear star cluster”, several tens of parsecs in radius. This star system is composed of cool, middle-aged stars (late A, early F stars with ages > 1 Gyr and masses $0.5\text{--}2 M_{\odot}$); bright horizontal branch and red supergiant stars would be very prominent. The stellar density of this system increases inward toward the black hole similar to an isothermal sphere distribution ($\rho \propto r^{-2}$). At a distance of only a few (1.5–4) parsecs from the black hole lie cool, dense molecular cloud streamers (the “circum-nuclear disk” or CND) orbiting the central regions.

Inside the central parsec (which turns out to be approximately the black hole sphere of influence r_h) the gas is ionized to $T \sim 10^6 \text{ K}$, creating a low-density “central cavity”. The cavity also has a set of cooler ($\sim 10^4 \text{ K}$) gas filaments (the “minispiral”) that orbit the central object. And, deep inside that, at a distance of only $\sim 0.03\text{--}0.1$ pc from the hole, lies at least one disk of O/B and Wolf–Rayet stars orbiting that central object in the opposite sense to the Galactic rotation. The “Bondi” accretion rate (Section 11.1.2) onto the black hole should be as high as $\sim 0.1 \dot{M}_{\text{Edd}}$, but it is four orders of magnitude lower. It is possible that these mass-losing stars are responsible for driving out much of the plasma that otherwise could fuel the central engine. These stars also are only 2–8 Myr old. It is not clear how they formed, and that has created a “paradox of youth”: it is very difficult to form stars so close to the tidal forces of the black hole, but the stars also are too young to have simply drifted into the center from further out in the nuclear star cluster.

Furthermore, in the central $1/30$ pc (7000 AU) of this disk lies a more spherized and randomized distribution of many tens of young B stars. It is these stars that are used to determine the central black hole's mass (see Fig. 1.1).

¹⁷ One might think that this inactivity is very fortunate for life on earth, but this makes no difference. Even if our central engine were to throttle up to full power ($L_{\text{Edd}} = 10^{44.7} \text{ erg s}^{-1}$), at a distance of 7 kpc, its brightness here on earth would be no more than the sun is when viewed from the inner Oort cloud (*i.e.*, 100 times the distance of Pluto); and much of our central black hole's energy would have been absorbed by the intervening dust and gas in the Galactic plane.

¹⁸ Much of the information discussed in this subsection is gleaned from the review of Reinhard Genzel, Frank Eisenhauer, and Stefan Gillessen of the Max Planck Institute for Extraterrestrial Physics in Garching, Germany [132].

Finally, we shall discuss briefly some models for the accreting plasma near the black hole itself. It is fairly well accepted that the accretion onto the central hole is a radiatively inefficient accretion flow (RIAF) of some sort. The accretion rate is probably around $\dot{m} = \dot{M}/\dot{M}_{\text{Edd}} \sim 10^{-5}$, which would produce $L/L_{\text{Edd}} \sim [(10^{-5})^2/0.1] \sim 10^{-9}$ (equation (12.58)). As we approach the hole, then, the electron temperature of the RIAF gas rises to $\sim 10^9$ K about $1000 r_g$ out (~ 4000 AU) and then saturates. The ion temperature, however, continues to rise to $\sim 10^{12}$ K just outside the hole and right before the plasma enters the horizon. RIAFs are well known for their ability to drive jet outflows, particularly when the accretion is driven by magnetic turbulence (see Sections 12.4 and 15.1). Therefore, a hybrid RIAF-jet model is probably the most likely description of the immediate environs of Sgr A*. However, unlike most black hole X-ray binaries and AGN, the jet in Sgr A* appears to not produce a shock or other feature that accelerates charge particles in the jet to highly relativistic energies [133].

Shortly before this book went to press, Gillessen and his colleagues announced the discovery of dramatic changes in a small gas cloud (only a few earth masses) near Sgr A* that they had been following since 2008 [134]. It now is being tidally shredded by the central hole (very much like Jupiter shredded Comet Shoemaker-Levy 9 in 1994), and in 2013 this cloud will reach its closest approach to the monster of only about 3000 black hole radii. Now, it may take many years, if not centuries, for most of this gaseous material to make its way down to the hole’s horizon, whereupon our Galaxy’s central engine will reach its peak brightness in this event. However, a few earth masses is 100,000 times as much material as our fuel-starved central black hole currently consumes in a year. So, even if only a minuscule fraction of this cloud begins to reach the hole in 2013, after many thousands or more years of being dormant, we could witness in our lifetime a sudden re-awakening of our central black hole, from inactive to (fairly) active galactic nucleus. Closely following our central black hole in the next few decades could reveal a lot about how the central engines of AGN operate.

2.5 “Inactive” Galactic Nuclei

As mentioned in the previous chapter, the presence of a black hole and its mass can be inferred simply by measuring the velocities of the central stars in a galactic nucleus. From equation (1.7), a black hole of $10^9 M_{\odot}$ in a galaxy 10–100 Mpc away would produce stellar velocities of order several hundred kilometers per second within a fraction of an arcsecond of the observed galactic center. With a resolution of better than 0.1 arcsecond, HST has allowed the measurement of many black hole masses in nearby galaxies, often in cases where no nuclear activity is seen [135]. These studies produced a surprising and important result: *black holes in weakly active or inactive galaxies are every bit as massive as those inferred to be powering the mighty quasars* [138]. Their inactivity must be due not to the lack of a large central engine, but simply to the current lack of fuel being fed to that engine.

2.5.1 The $M_{\bullet}$ – M_{bulge} Relation

It appears that the black hole mass in any galaxy depends *primarily* on the properties of the spheroidal bulge of stars in that galaxy. (See Fig. 2.18.) That is, giant elliptical galaxies of similar mass have similar mass black holes. And giant spiral galaxies with small bulges have similar mass black holes to those in bare dwarf ellipticals with similar small-mass bulges but no spiral disk. The mass of the gas and stellar *disk* in spirals apparently does not count much in black hole formation and growth. Bulges of higher mass have more massive black holes with an approximate linear relation of

$$M_{\bullet} \approx 0.0013 M_{\text{bulge}} \quad (2.12)$$

This is sometimes called the *Magorrian relation* after the first discoverer of a $M_{\bullet}$ – M_{bulge} correlation [135]. While there is much scatter in it, the relation is now considered sufficiently universal that some observers sometimes estimate the mass of the central black hole simply by measuring the luminosity of the bulge, computing the mass of the bulge from known mass–luminosity relations for elliptical galaxies

$$M_{\text{bulge}} = 1.9 \times 10^{10} M_{\odot} (L_{\text{bulge}}/10^{10} L_{\odot})^{1.25} \quad (2.13)$$

where L_{bulge} is the luminosity of the bulge in red light [139, 140], and then taking the black hole to be 0.0013 of that bulge mass.

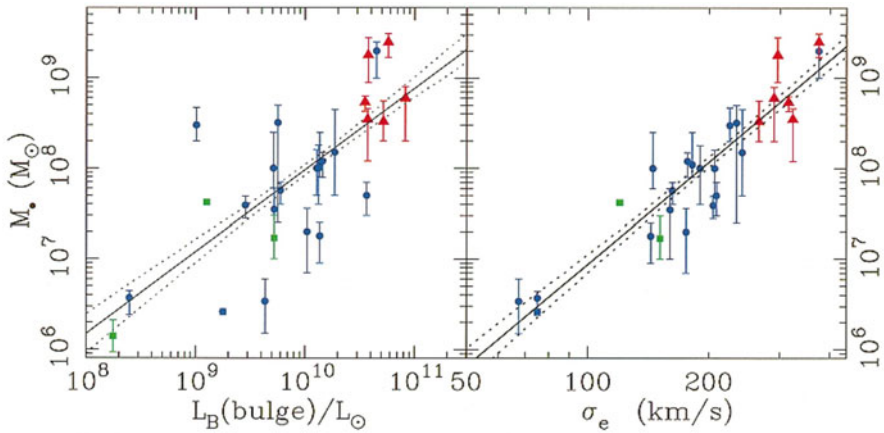


Fig. 2.18: The black hole mass vs. bulge luminosity (left) and velocity dispersion (right) relations. The considerable scatter in the $M_{\bullet}$ – L_{bulge} relation indicates that more than just galaxy bulge mass may influence black hole formation. The $M_{\bullet}$ – σ_{bulge} relation (often called the “ M – σ ” relation) is considered much tighter, as it has scatter consistent with the vertical error bars [137]. Reproduced by permission of the AAS.

2.5.2 The $M_{\bullet}$ – σ_V Relation

The black hole mass correlates even better with the velocity dispersion of the stars in the outer part of the bulge (right panel of Fig. 2.18) [136, 137]

$$M_{\bullet} = 1.3 \times 10^8 M_{\odot} (\sigma_{\text{bulge}}/200 \text{ km s}^{-1})^{4.0} \quad (2.14)$$

However, σ_{bulge} is harder to measure than L_{bulge} . So, in order of difficulty, the easiest way to estimate a black hole mass in the center of a galaxy is to (1) measure the bulge luminosity and use equations (2.12) and (2.13), (2) measure the velocity dispersion of the stars *in the bulge* and use equation (2.14), (3) measure the velocity dispersion of the stars *very near the black hole* and use equation (1.8), (4) measure the velocity structure of the broad line clouds surrounding the black hole using reverberation mapping (also using equation (1.8)).

In recent years Luis Ho of the Carnegie Observatories has measured masses of black holes in the centers of dwarf spheroidal galaxies using method #3. Such galaxies have masses of only $\sim 10^{7-9} M_{\odot}$ and velocity dispersions down to 20–60 km s^{−1}. The resulting mass determinations extend the $M_{\bullet}$ – M_{bulge} and $M_{\bullet}$ – σ_{bulge} relations down to $M_{\bullet} < 8 \times 10^4 M_{\odot}$. This is significant for several reasons. First, it indicates that the process of black hole formation in galactic nuclei must be very similar for spheroidal galaxies over a mass range that stretches six orders of magnitude. Second, for several years no black hole below a million solar masses had been discovered; Ho’s investigations broke the $10^6 M_{\odot}$ barrier, penetrating the region formerly called the “intermediate mass” black hole regime. Now, because these much smaller mass black holes are so similar to their gigantic cousins, we have re-defined in this book the IMBH range to be roughly 10^2 – $10^4 M_{\odot}$. (See Chapter 4.)

2.5.3 The $M_{\bullet}$ – r_{core} Relation

While some galaxies and/or bulges have stellar densities that continue to increase as a power law inward (hence the name “power-law” profile galaxies), many have an identifiable core where the stellar density flattens out within a few hundred parsecs or less of the galactic center.¹⁹ (This core in the galaxy’s stellar distribution is not to be confused with the much smaller radio jet core that was described earlier in this chapter.) In 1999 Roeland van der Marel, of the Space Telescope Science Institute studied how the properties of these core galaxies and their black holes correlate and found a relation between the mass of their galaxy cores and the masses of their black holes ($M_{\bullet} \propto M_{\text{core}}^{1/2}$) [141]. In Fig. 2.19 we have re-plotted his data, comparing the core *radius* (in parsecs) with the black hole mass. To this data we have added

¹⁹ See, for example, Fig. 1.2, lower right. While M87 has a discernible cusp within 1 arcsecond of the center, even that cusp does not display a true power-law profile. There is a distinct flattening of the overall stellar density that occurs within 6–7 arcseconds, giving a core radius of $r_c \approx 450$ –500 pc.

some dwarf spheroidal core galaxies (see Fig. 4.1). Note that the slope black hole mass–core radius relation over many orders of magnitude in black hole mass is significantly different from that at the high-mass end (where there is only one order of magnitude of data). Formally, the overall fit gives a relation

$$\log r_c = 0.52 \log M_\bullet - 2.06$$

with r_c in parsecs and $M_\bullet$ in solar masses. This is close to the relation

$$r_c = 422 \text{ pc} \left(\frac{M_\bullet}{10^9 M_\odot} \right)^{1/2} \quad (2.15)$$

As we shall see in Chapter 11, this implies a core mass–black hole mass relation of

$$M_{\text{core}} \approx 7 M_\bullet \approx 0.009 M_{\text{bulge}}$$

that is, $M_{\text{core}} \propto M_\bullet$. Of course, this modified relation needs to be confirmed with data on more lower-mass core galaxies and bulges. But if it holds, then black hole, core, and galaxy masses all are, in fact, approximately proportional to each other.

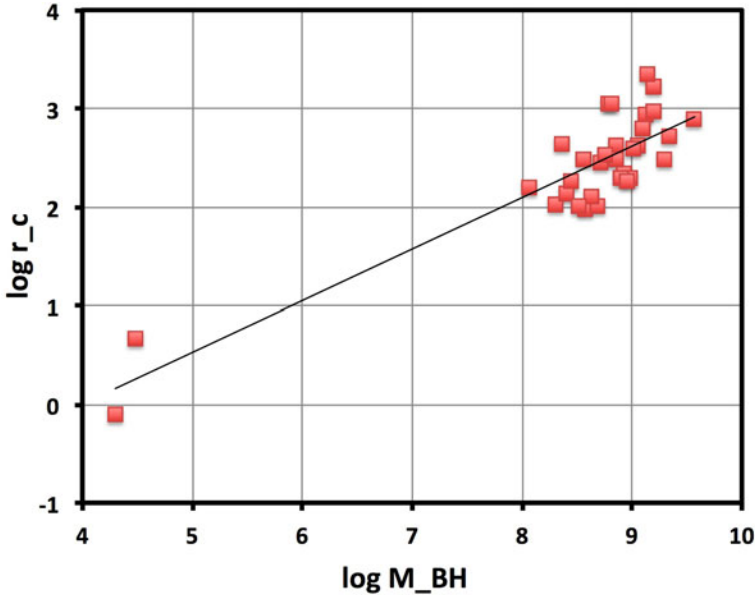


Fig. 2.19: Data from the literature for core galaxies on the variation of core radius with black hole mass. Points at upper right are from reference [141]; low-mass points at lower left are for the dwarf galaxies in Fig. 4.1.

2.6 Macroquasars – A Summary and Synthesis

Before proceeding on to lower-mass black hole systems, it is important to make sense of the large number of different macroquasars that exist. The discovery that black hole masses are the same for all galaxies with similar bulges is perhaps the most important result. The fact that one galaxy of a given type can have a bright QSO in its center, while another will appear normal, must be largely statistical or perhaps environmental. Some recent event that is probably independent of the central nuclei of both galaxies has triggered the activity in one and not the other. Many millions of years from now, their roles may be reversed. The currently quiet galaxy may become active while the active one may become dormant.

A second important conclusion comes from the comparison of host galaxies of the brightest QSOs and the optically similar, but fainter, Type 1 Seyfert nuclei. The brightest QSOs occur in galaxies with the largest spheroidal bulges, while Seyferts occur in galaxies (spirals) with smaller bulges. Their large difference in luminosity, then, is not due solely to their fueling rate but rather to their difference in black hole mass: more massive black holes simply are larger engines and can produce more power. Therefore, objects that appear similar to QSOs (high luminosity relative to the stellar bulge, broad emission lines, *etc.*) must be fueled at a rate near the maximum for a given black hole mass. Indeed, their luminosities are within a factor of 10 of the Eddington maximum limit (equation (1.1)) for their black holes. Above this luminosity radiation pressure on the free electrons in the accreting gas will blow out all matter that tries to fall into the hole.

The third important conclusion comes from a study of the host galaxies of radio-loud and radio-quiet QSOs, led by Ross McLure of Oxford and James Dunlop of Edinburgh University. They found that the brightest QSRs and QSOs both occur exclusively in giant elliptical galaxies [142]. A bright QSR and a bright QSO must, therefore, have similar black hole masses, similar host galaxies, and similar fueling rates. Yet, they differ in their ability to produce a powerful, extended radio jet by a factor of 10,000! Such a difference cannot be due to beaming, because even an unbeamed jet would produce a strong extended radio source. Clearly, there must be another property of the central black hole system that determines whether or not a jet can be produced and yet does not affect the black hole's ability to produce optical emission. Several authors, including the present one, have suggested that this property is the black hole's spin, and that is the one of the main theses of this book.

While it is clear that a third parameter is needed to control the jet power, it still is the case that fueling rate and black hole mass play an important part. The most rapidly-fueled objects (QSRs) have the most powerful FR II jets. And Dunlop and McLure's team, as well as another led by Mark Lacy, now at the Spitzer Science Center in California, have concluded that the maximum jet power appears to increase with the black hole mass at a higher rate than expected

$$L_{\text{jet}} \propto M_{\bullet}^2 \quad (2.16)$$

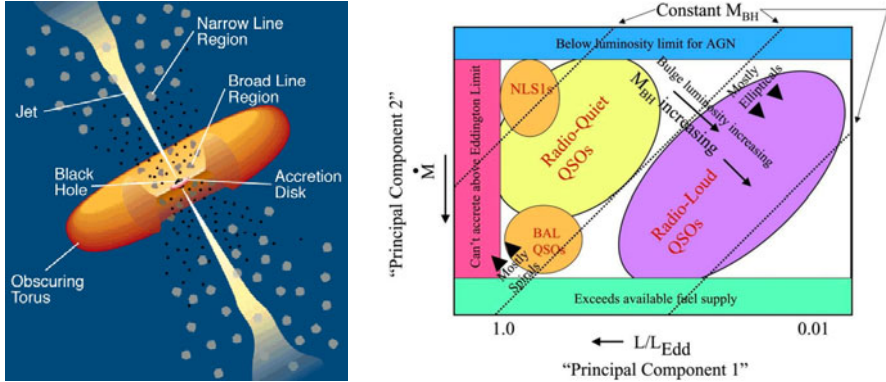


Fig. 2.20: Two schemes for unifying macroquasars. *Left:* Phenomenological model for an active galactic nucleus [143, 144]. Most AGN probably have the components pictured, although in varying degrees (*e.g.*, Seyfert and QSO jets may be weak or absent). Credit: C.M. Urry and P. Padovani. *Right:* Diagram for macroquasars in the accretion rate–Eddington ratio plane [145]. From left to right, diagonal lines are at approximately 10^6 , 10^8 , and $10^{10} M_{\odot}$. Lines of constant luminosity are horizontal, as $L \propto \dot{M} c^2$ with roughly constant efficiency. Reproduced by permission of the AAS.

rather than the expected $L_{\text{jet}} \propto M_{\bullet}$ [146, 147]. Apparently, more massive black holes are more efficient at producing strong radio jets. This is undoubtedly related to why radio galaxies and QSRs occur exclusively in elliptical galaxies with very massive black holes.

Dunlop and McLure’s work nevertheless supports unification of QSRs and FR II radio *galaxies* by orientation. Their host galaxies and their radio jets appear to be very similar. And in some cases, FR II radio galaxies are found to have a hidden broad line quasar in reflected polarized light. FR II radio galaxies, therefore, are probably QSRs whose accreting black hole is hidden by a dusty torus.

Many people have attempted to combine all active galaxies into a single unified scheme. Two different approaches are shown in Fig. 2.20. The first (left panel in the figure), by C. Megan Urry of Yale University and Paolo Padovani of the University of Rome has unified radio galaxies, blazars, and QSRs into a single phenomenological picture [144]. A supermassive black hole lies at the center of the source, with an accretion disk surrounding it that produces the big blue bump and perhaps the broad line clouds as well. An obscuring dusty torus surrounds this central source, with the narrow line region and jet lying roughly along the torus’ rotation axis. Seyferts and QSOs can be unified similarly, but without a powerful jet in the picture. This is essentially the Antonucci and Miller picture extended to radio-loud sources as well.

The second approach (right panel in Fig. 2.20), by Todd Boroson of NOAO [145], is based mainly on observations and their straightforward analysis and classifies objects on the basis of their black hole mass, accretion rate (inferred from the object luminosity), and “Eddington ratio” (from luminosity and mass)

$$\ell_{\text{rad}} \equiv \frac{L_{\text{rad}}}{L_{\text{Edd}}} \quad (2.17)$$

The figure shows that, while RLQs tend to have larger black hole mass than RQQs, they both tend to have similar absolute accretion rates ($\dot{M}$), suggesting that radio-loud objects accrete at normalized rates ($\dot{m}$) somewhat below the Eddington limit. BAL QSOs have high Eddington ratios and masses nearly as large as some RLQs. NLSy1s also lie near the Eddington limit, while lower-luminosity Seyfert nuclei would lie just to the right and above the NLSy1s (at the low-mass end of the RQQs).

These, and other unification schemes, are very useful for beginning detailed theoretical investigations of the physical processes that may be at work in macroquasar systems.

Chapter 3

Microquasars: Black Holes (and Neutron Stars) of Stellar Mass in Our Galaxy

So far we have concentrated on supermassive black holes ($10^5\text{--}10^6 M_\odot$) at the centers of galaxies. Except during merger events, we expect, at most, one of these per galaxy and with a mass of $\sim 10^{-3}$ of that galaxy's spheroidal bulge mass. The existence of black holes much larger than this is very unlikely, both theoretically and observationally: there is no reasonable evolutionary scenario by which such an object could be made, and we have no observational evidence for such hypermassive objects. However, there are many reasons for suspecting that black holes much less massive than ten thousand solar masses might exist, and that there may be many such objects in each galaxy, with especially large numbers in the stellar-mass range ($1\text{--}100 M_\odot$). The work of the late Subrahmanyan Chandrasekhar, of the University of Chicago, on the structure of white dwarf stars showed that the collapse of such dense remnant stars is inevitable, if they have masses greater than $\sim 1 M_\odot$ [148] (now known to be closer to $1.4 M_\odot$; see Section 5.2.4). We also do not find stars much more massive than $100 M_\odot$. So it is natural to postulate the existence of collapsed stars and black holes somewhere in this mass range.

Chandrasekhar received one-half the Nobel Prize in 1983 (along with William Fowler) for his theoretical work on stars in the mid 1900s. Curiously, most of the astrophysics-related Nobel Prizes have been for theoretical or observational work related to these compact objects.¹

One year before the discovery of radio-loud quasars, another astronomical revolution was taking place, at the *upper* end of the electromagnetic spectrum. Riccardo Giacconi, Herbert Gursky, and Frank Paolini, of American Science and Engineering in Cambridge, Massachusetts, and Bruno Rossi of the Massachusetts Institute of Technology, were using Aerobee sounding rockets to survey the universe in a wavelength region never before explored – X-rays [149]. Fortunately for life on our planet, but unfortunately for astronomers, the atmosphere is opaque to X-rays. Unlike radio and optical astronomy, X-ray astronomy can be done only from space,

¹ The noted exceptions are the 1967 Prize for stellar nuclear reactions and two Prizes for studies of the Big Bang microwave radiation (in 1978 for its discovery and in 2006 for determining its blackbody form and anisotropy). The 1983 Prize was for stellar evolution and nucleosynthesis work, which *is* related to neutron star and black hole formation.

either with sub-orbital sounding rockets or, as is now the case, with X-ray satellite telescopes. The year 1962 is significant for the discovery of two important compact cosmic X-ray sources²: Scorpius X-1, now believed to be a neutron star of about $1.5 M_{\odot}$, and Cygnus X-1, believed to be a black hole of about $8.7 \pm 0.8 M_{\odot}$ [150]. Neutron stars are “near-black” holes, composed primarily of neutrons throughout much of the star’s interior, with a gravitational field nearly as strong as that of a true black hole. A neutron star of $1.5 M_{\odot}$ has a radius of 10 km, while a black hole of the same mass would be about 4.5 km in radius. Both can produce X-rays when matter accretes onto them, but they do not do so in quite the same way. An important difference is that a neutron star, as compact and strong gravitationally as it is, still has a solid surface at that 10 km radius. A $1.5 M_{\odot}$ black hole, on the other hand, has no such surface; instead, it is surrounded by a horizon that lies 4.5 km from the center of the otherwise point-like object. All the heat radiation produced by accretion must eventually escape a neutron star’s solid surface, while the black hole can swallow much of the X-rays (as well as matter) and add them to its mass. Although it was suspected in 1962 that the X-rays in these sources were produced by such accretion onto very compact objects, there was as yet no direct evidence of companions that could donate the accreting material. The X-ray positions were too poorly known for the optical counterparts of the X-ray sources to be identified and for any possible companions to be examined. Finally in that same year, the sounding rockets also discovered a “background” X-ray glow that pervaded the entire sky. It is now believed that most, if not all, of this background is due to distant quasars producing X-rays in a manner much like Cyg X-1. In 2002 Giacconi received one-half of the physics Nobel Prize for his discovery of compact and background X-ray sources.

Two years after these initial results it was discovered that the Crab Nebula, in the constellation Taurus, was an extended and steady X-ray source [151]. The Crab had had a long history before 1964, however. Discovered with optical telescopes in the 1700s, and designated as the first object in Messier’s catalog, the Crab Nebula had been identified as the remnant of a stellar explosion witnessed by the Chinese in 1054 A.D. It also was identified with the extended radio source Taurus A by John Bolton, of the Council for Scientific and Industrial Research in Australia, in 1948 [152]. The source of energy for its optical, radio, and X-ray emissions, however, was a mystery. Then in 1968 Antony Hewish, S. Jocelyn Bell, and their colleagues at Mullard discovered pulsations coming from a compact star near the center of the nebula, once every 1/30th of a second [153]. (For this discovery, Hewish received one-half of the physics Nobel Prize (sharing with Ryle) in 1974.) This pulsar was soon identified as an isolated spinning neutron star, with the radio pulses coming from its magnetic field of 10^{13} G (gauss; *i.e.*, about 10^{13} times stronger than the earth’s magnetic field). The gradual spinning-down of this pulsar, and the pumping of that energy into the nebula through the rotating magnetic field, precisely

² As in radio astronomy, the first few sources discovered in a constellation were named after that constellation, but later the community converted to the International Astronomical Union designation using sky position (right ascension and declination). This convention is still in use today, but often with more digits added to the position, as surveys become fainter and discover more objects in a given area of sky.

accounted for the nebular emissions. The lack of strong compact X-ray emission from the pulsar is evidence that there is no companion star to donate accretion fuel. So, while the Crab Pulsar and Sco X-1 neutron star are very similar, one is faint in X-rays while the other is very bright solely because of their very different environments.

Another space astronomy discovery at the close of the 1960s helped to complete the story of stellar-mass black holes. In order to monitor the 1963 nuclear test ban treaty, the USA launched two Vela satellites to search for γ -rays from terrestrial explosions. The satellites did detect powerful explosions, not from earth, but from distant reaches of the cosmos. These “ γ -ray bursts”, or GRBs, were a mystery until the late 1990s when they were determined to be the most powerful objects in the universe, even outshining the brightest quasars by tens of thousands of times for a few brief seconds. It is now generally agreed that *a GRB represents the celestial herald of the birth of a stellar-mass black hole*.

At the end of the 1960s all the pieces of the stellar-mass black hole story were in place, save one: astronomers had discovered neutron stars in binary systems, isolated neutron stars (pulsars), neutron stars in formation (supernovae (SN)), black holes in binaries, and black holes in formation (GRBs). Only one piece had eluded observers, and it did so until the turn of the new millennium – the identification of isolated black holes of stellar mass. With no companion to donate accreted material, no magnetic field of their own, and no significant heat to radiate, isolated black holes in our Galaxy only have their gravity to give themselves away.

3.1 Isolated Neutron Stars: Pulsars and Magnetars

3.1.1 “Normal” Pulsars

The compact radio source at the center of the Crab Nebula was known as a 1 Jy low-frequency radio source long before it was known to be a pulsar. Pulsars, therefore, are not discovered simply by detecting the source in the sky. Rather, they are found by detecting their pulsations, which can range in period from several seconds to as short as only a millisecond. Modern pulsar searches are done by using computers to analyze data obtained by scanning large areas of the sky. As the telescope scans, the computers concentrate on the section of sky just observed and test a single possible pulse period, adding together many possible pulses to get good signal-to-noise. Then a new period is tested, and so on. With Fourier analysis techniques, many possible pulse periods at different radio frequencies can be tested at once. Possible pulsar candidates are recorded, and then the computer turns its attention to the next region of the sky. Afterward, follow-up observations are performed on each candidate to confirm the pulsar discovery and to accurately measure its position by timing the arrival of the pulses at different times in the year.

3.1.1.1 Distances

Approximate distances to pulsars are measured by measuring the “dispersion” – the variation of arrival time – of the pulses with radio frequency. The distance is given by the equation [154]

$$D = 240 \text{ pc} \frac{\Delta t_{12}}{\langle n_e \rangle} \frac{1}{\nu_1^{-2} - \nu_2^{-2}} \quad (3.1)$$

where the frequencies ν_1 and ν_2 are measured in GHz, the time delay Δt_{12} in seconds, and the *average* electron density along the line of sight to the pulsar $\langle n_e \rangle$ in cm^{-3} . For example, for a time delay of 10 ms in pulse arrival time between two different frequencies $\nu_1 = 425 \text{ MHz}$ and $\nu_2 = 435 \text{ MHz}$, with an assumed average galactic electron density of $\langle n_e \rangle = 0.02 \text{ cm}^{-3}$, the pulsar distance is estimated to be about 480 pc. Clearly, $\langle n_e \rangle$ carries the largest uncertainty here. If great care is taken to model the Galaxy electron density distribution, including spiral structure and important nebulae that lie in the way, distances accurate to 30% are possible [155, 156].

3.1.1.2 Periods, Spindown Rates, Ages, and Magnetic Fields

Pulsar periods, and the rate at which they spin down, can be measured to extremely high accuracy. Because each pulsar keeps such good time, observers can count exactly how many pulses have occurred between its discovery and its next observation! Over a decade, this amounts to many billions of pulses and, with sub-millisecond timing, can reach averaged pulse periods accurate to better than a femtosecond (10^{-15} s). In the not too distant future, this highly accurate timing will allow pulsars to be used as aids to gravitational wave detection. Spinning neutron stars will be used to find colliding supermassive black holes.

The long time base also allows accurate determination of a pulsar’s spindown rate $\dot{P}$, which is caused by dipole radiation by the spinning pulsar magnetic field. $\dot{P}$ can be used to estimate a pulsar’s age (or spindown time) in seconds

$$\tau = \frac{P}{2\dot{P}} \quad (3.2)$$

and the neutron star’s surface magnetic field

$$B = 3.2 \times 10^{19} \text{ G} \left(P \dot{P} \right)^{\frac{1}{2}} \quad (3.3)$$

(see Section 14.1.1) where, in both cases, P is in units of seconds and $\dot{P}$ in seconds per second. For example, the Crab Pulsar has $P = 0.033 \text{ s}$ and $\dot{P} = 4.23 \times 10^{-13}$, which gives a characteristic age of $\tau = 1200 \text{ yr}$ (close to its real age of about 950 yr) and a magnetic field of $3.8 \times 10^{12} \text{ G}$. Fig. 3.1 shows the $P - \dot{P}$ plane for all

radio pulsars in the Galactic plane known in 2007. The Crab is a young member of the class of “normal” pulsars, which occupy the large collection of objects in the middle-right portion of the diagram. These typically have periods of order 0.1–3.0 s and spindown rates of $\dot{P} = 10^{-17}$ – 10^{-13} , therefore yielding ages of about 10^5 – 10^8 yr and magnetic fields of 10^{11} – 10^{13} G. The other types of pulsars will be discussed in the next subsections below.

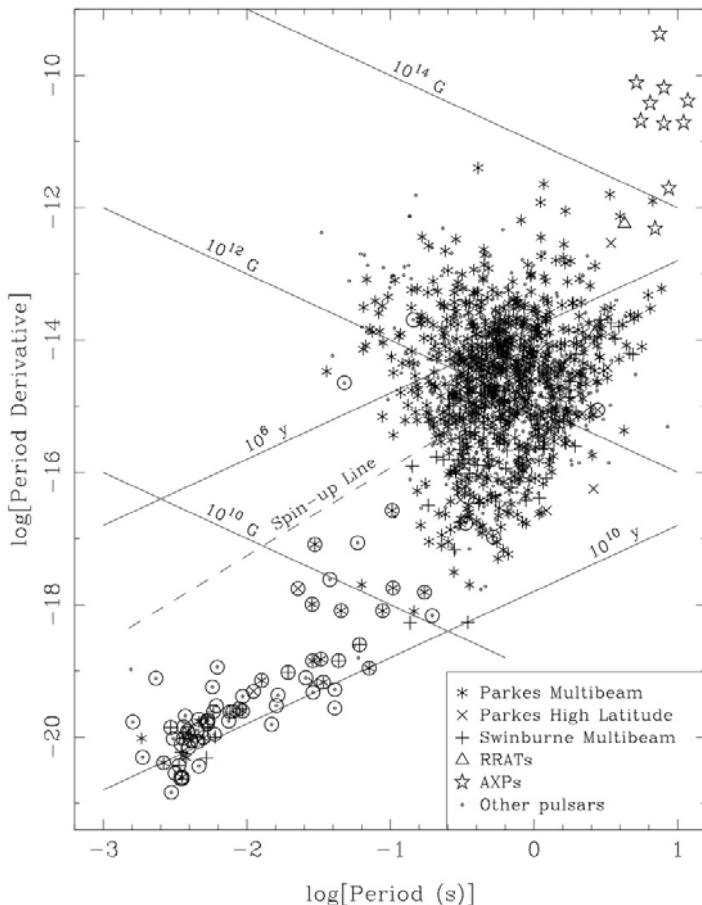


Fig. 3.1: The period–period-derivative diagram for all known Galactic disk pulsars, as of about 2007 [157, 158]. Most symbols indicate at what observatory the radio pulsars were discovered (asterisks [*], crosses [×], and pluses [+] at Australian observatories; dots (·) at other observatories). Two symbols, however, indicate the physical type of pulsar: stars (★) indicate anomalous X-ray pulsars (AXPs) and soft γ -ray repeaters (SGRs), neither of which were discovered in the radio; the delta (Δ) indicates a rotating radio transient (RRAT). Any symbol surrounded by a circle indicates a pulsar in a binary star system. Corresponding lines of age τ (eq. (3.2)) and magnetic field strength B (eq. (3.3)) are shown. Reproduced from Fig. 1 of reference [158], with permission of CSIRO.

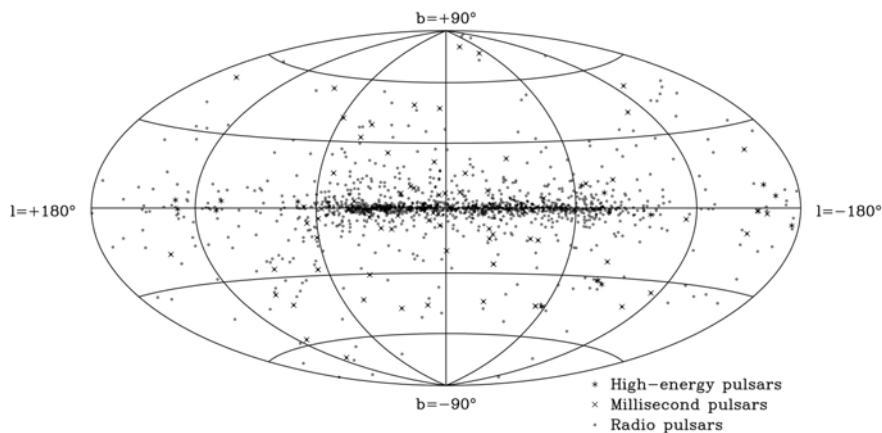


Fig. 3.2: Distribution of pulsars in Galactic coordinates, with the Galactic center at the origin [166]. Most pulsars lie within ± 0.5 kpc of the Galactic plane. (Many of the ones that lie at high latitude are, in fact, near the solar system and, therefore, also near the Galactic disk.) This limited distribution is believed to be due to a short pulsar lifetime (few million years). After they die, pulsars continue to shoot away from the Galactic plane at $\sim 150 \text{ km s}^{-1}$, filling the entire Galaxy with large numbers of “dead” pulsars through its $\sim (20 \text{ kpc})^3$ volume. Reproduced by permission of the AAS.

3.1.1.3 Pulsar Proper Motions and Cosmic Rockets

Shortly after the discovery of the neutron by Sir James Chadwick (in 1932) [159],³ Walter Baade and Fritz Zwicky proposed (in 1934) that neutron stars could be formed in supernova explosions of massive stars [160]. And in 1942 Baade and Rudolph Minkowski (nephew of Hermann Minkowski of flat-space metric fame (Section 6.4)) suggested that the faint star at the center of the Crab Nebula might indeed be the remnant of that explosion [161, 162]. The discovery that this star was a pulsar in 1968 confirmed the suggestion, but the subsequent discovery of more pulsars raised a puzzle. If pulsars were fairly young remnants of supernova explosions, then their galactic distribution should be similar to that of their parents, massive stars that begin life at about $10\text{--}20 M_{\odot}$. It is well known that massive stars are confined to the Galactic plane: because they do not live long, all currently observed massive stars must have been born recently in Galactic molecular clouds. However, James Gunn and Jeremiah Ostriker of Princeton University found that the pulsar distribution extends much higher, and below the Galactic plane (see Fig. 3.2) [163]. (Modern estimates of the pulsar scale height range from 120 pc [163], to ~ 400 pc [164], to over 600 pc for the oldest (millisecond) pulsars [165], with the older ones at higher latitudes.) So, if pulsars are a few million years old on the average, and

³ For which he won the 1935 Nobel Prize for Physics.

extend ~ 500 pc from the Galactic plane where they were born, they must have traveled at speeds of about 150 km s^{-1} to get to that height. This is nearly 10^{48} erg of kinetic energy for each pulsar – 0.1–1% of the supernova explosion energy itself that created each runaway object. Accurate proper motion measurements of a few key pulsars confirm that all such motions are away from the Galactic plane. Therefore, many pulsars must be born in supernovae and given a “kick” that propels them above and below the Galactic plane. However, after traveling only ~ 500 pc from their points of origin (and long before reaching their maximum distance from the plane of perhaps 20 kpc), the pulsar dies and becomes a non-radio-emitting neutron star, orbiting the Galaxy repeatedly but not pulsating. The 500 pc distribution, then, represents not the full extent of the neutron stars’ orbits, but rather simply how far they can get from their birth places before they quit being pulsars and turn into dead neutron stars.

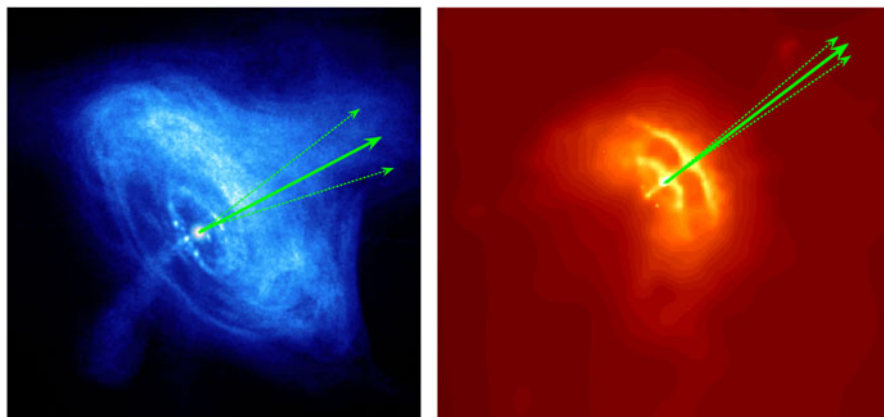


Fig. 3.3: The only two pulsars with measured proper motions and observed jets each have both these features well aligned. *Left:* The Crab Pulsar (center bright component) observed with the Chandra X-ray telescope [169]. The jet is most visible pointing toward the southeast (lower left). (Also visible in X-rays is the complex “pulsar wind nebula” or PWN.) The thick green arrow shows the pulsar proper motion at $180 \pm 30 \text{ km s}^{-1}$, nearly parallel (on the sky) with the inner jet. The dotted arrows show the $\pm 10^\circ$ error in the proper motion direction. Credit: NASA/CXC/ASU/J. Hester et al. [170]. *Right:* The Vela Pulsar, also observed with Chandra. Again the jet is most visible to the southeast, and again the pulsar proper motion (at $130 \pm 8 \text{ km s}^{-1}$) is parallel to the jet, within the rather small errors. Credit: NASA/PSU/G.Pavlov et al. [171].

With the launch of the Chandra X-ray observatory by NASA in 1999, another important property of pulsars was discovered that linked them with black holes. They produce jets. Both the Crab Pulsar and the one in the Vela supernova remnant (which is about 10^4 yr old) have oppositely-directed, twin jets with speeds of at least half the speed of light. Furthermore, those pulsars appear to be moving with speeds of about $100\text{--}150 \text{ km s}^{-1}$ *along the same direction as their jets are pointing* [167, 168]. (See Fig. 3.3.) A possible explanation of this result is that one of the twin jets may be, or may have been, stronger than the other, providing a net thrust to

each pulsar. As these speeds are typical of all pulsars, perhaps most pulsars act like cosmic rocket engines. However, while the present Crab Pulsar's jets have produced more than enough *energy* in the last thousand years (10^{49} erg) to give the neutron star its present kinetic energy, they have not produced enough *momentum*. With the emissions being mostly radiation and light particles, they have produced only 1% of the $>10^{40}$ g cm s $^{-1}$ required to explain the observed motion. Therefore, if this mechanism is responsible, a much more powerful rocket, occurring when the pulsar was first born, must have accelerated the pulsar to its current velocity. Because that rocket must have pointed in the same direction as the pulsar jets now point, it is tempting to suggest that even more powerful jets occurred during the supernova explosion. In the discussion of supernovae below we see that this is indeed appears to be the case!

3.1.2 Magnetars

One millisecond is about as short of a period as has been measured for a pulsar. Much shorter than that and the neutron star begins to approach the “breakup” or Keplerian period of 0.46 ms. If we take 1 ms as the shortest typical pulsar period possible, then we can ask, what would happen if a pulsar were born with such a period and a much greater magnetic field than normal pulsars – say 10^{15} G, which is strong enough to affect the structure of the neutron star core? From equations (3.2) and (3.3) the time for the period of such a star to double is $3\tau = 25$ minutes! In only a week, such a pulsar would be spinning at the rate of the Crab Pulsar. In a thousand years it would have a spin period of 8 seconds and not likely be producing any pulsar dipole radio radiation. Two related classes of X-ray stars have been discovered which are likely to be these high magnetic field “magnetars”.

The first type of star discovered is called a “soft γ -ray repeater” or SGR. These actually were observed in the prior decade during searches for GRBs [172, 173]. However, they were not recognized as a separate class of objects until 1986 when it was discovered that, unlike GRBs which never repeat, these sources have repeated soft γ -ray (or very hard X-ray) flashes [174, 175]. In 1992–95 Christopher Thompson, now at the Canadian Institute for Theoretical Astrophysics in Toronto, and Robert Duncan, now at The University of Texas at Austin [176, 177, 178] suggested that SGRs were magnetars and presented a detailed model of how neutron stars could generate such strong magnetic fields. SGRs also pulsate in their X-rays, but generally only as their soft γ -ray flux is decreasing after the burst. The pulse periods are 5–8 seconds, and the spindown rates are of order $\dot{P} \sim 10^{-10}$. These and other properties have led to the conclusion that SGRs are magnetars with fields of $2\text{--}4 \times 10^{14}$ G and ages of only a few thousand years [179]. At the time of writing there are ~ 6 SGRs known.

The second type of magnetar is the “anomalous X-ray pulsar” or AXP. These objects also pulsate in X-ray, with periods of around six seconds, they appear to *not* occur in binary systems, and yet they emit much more energy than they should

given their slow rotation (hence the term “anomalous”). This extra energy is thought to come from the breakdown of the neutron star’s enormous magnetic field. Six such stars are known, and they are all associated with supernova remnants. In 2002 it was discovered that at least one of the AXPs is also an SGR, confirming the magnetar hypothesis for AXPs as well. What causes a magnetar to sometimes be a soft γ -ray repeater is not yet clear. It could be related to age or magnetic field strength; there are arguments for and against both of these ideas. At the time of writing there are ~ 9 AXPs known.

Both of these types of pulsars have been combined into the AXP, or magnetar, class, and ten of them are shown in the upper right portion of Fig. 3.1. Most of these objects are only a few thousand years old, and have magnetic fields of 10^{14-15} G.

A third type of pulsar that has been identified as a possible magnetar is the X-ray Dim Isolated Neutron star, or XDIN. As of this writing, this group consists of only seven (the “magnificent seven”, as they are called) radio-quiet and very close X-ray pulsars (only ~ 100 – 400 pc away) [180]. They are significantly weaker in X-ray luminosity than the AXPs. Because they are all so close, even though only a few have been found, the implications are that this is a very large population of strong-field neutron stars – as many as there are normal radio pulsars, maybe more. Fields and ages measured for two of these are of order 2 – 3×10^{13} G and 1 – 2 Myr. While they have weaker fields (and are older) than AXPs, these two are still a factor of ten more magnetized than normal pulsars. Their periods are also much longer (3 – 11 s), which is consistent with being the low-field end of the magnetar range (or with being rather old).

3.1.3 RRATs

An initially rather elusive class of intermittent pulsars was given the appropriate name RRATs (Rotating Radio Transients) [181]. This new class of pulsars is characterized by short radio bursts of 2 – 30 ms duration, separated by time intervals ranging from 4 minutes to 3 hours. Generally, each RRAT is visible in the radio for only one second per day! In spite of RRATs’ fleeting nature, these observers were able to painstakingly fold together these fragments of data and measure a pulsar period for ten of the eleven RRATs they discovered and $\dot{P}$ for three of those. Two of the sources have normal magnetic fields (3 – 6×10^{12} G) and ages (2 – 3 Myr) and one RRAT has a ten-times stronger magnetic field and is only a hundred thousand years old. Most interestingly, however, was that half of the sources had periods similar to the AXPs (over 4 seconds). At present this class is so new that there are *too many* possible explanations for the phenomenon (dying pulsars, pulsars with an asteroid belt blocking the pulses most of the time, or even an entirely new class of pulsars). Again, there are some indications that there may be as many, or even more, RRATs as there are normal pulsars. And, because RRATs are significantly farther away than XDINs (2 – 7 kpc), but their numbers may be similar, some investigators have suggested that RRATs and XDINs may be the same population of objects. That is, we

cannot detect X-rays in most RRATs because they are so far away; and we have not yet been able to detect the short radio bursts in XDINs yet because we simply have not tried hard enough yet, or because the pulsar fan beam does not intersect the earth. Sorting this all out will be important ultimately for understanding how neutron stars are made and how they evolve.

3.1.4 Geminga and Other Nearby Neutron Stars

If pulsars live for a few million years, and if their rate of production in the Galaxy is about equal to the supernova rate (about one per 100 years), then there must be a few tens of thousand active radio pulsars in our Galaxy, not to mention a few hundred million dead ones. Let us suppose that the active pulsars are evenly distributed throughout the Galactic disk, which is roughly 1000 pc thick and 15 kpc in radius. Then, from any random point in that disk, including from our own star, there must be at least one active radio pulsar only 200–300 pc away.

In 1974 the SAS-2 satellite discovered the second brightest γ -ray source in the sky in the constellation of Gemini [182, 183]. It was given the name Geminga, which means “it is not there” in Milanese. The source remained a mystery until 1992 when Jules Halpern of Columbia University and Stephen Holt of NASA’s Goddard Space Flight Center discovered 0.24 s X-ray pulsations with the ROSAT X-ray satellite [184]. Its magnetic field and spindown age are 1.6×10^{12} G and 3.4×10^5 yr, respectively, and its distance appears to be only about 157 pc. It appears to be a normal radio pulsar whose fan beam of *radio* emission does not intersect the earth. Its very close distance indicates that Geminga may be that one of the few active radio pulsars that are the closest to the sun. The fact that the XDINs are at a similar distance, but are older and slower, opens the intriguing possibility that they represent a later (perhaps dying) stage in the life of Geminga and other normal pulsars as they slow down.

How close might the nearest *dead* pulsar be? Because these are in permanent orbits about the Galaxy that take them far from the plane, the cloud of dead pulsars must be about as extensive as the spherical Galactic halo. So their numbers near the sun will be much less than if they were confined to a 1 kpc thick disk. Nevertheless, there still should be many more dead pulsars near the sun than active ones. The nearest one is probably no more than 40 pc (130 light years) away. And there must be as many as two or three hundred isolated dead pulsars within the same distance as Geminga and the XDINs!

3.2 Neutron Stars with Companions: Neutron Stars in Binaries

Neutron stars in binary systems are important to the study of true black holes because they behave similarly to the latter, both in their accretion and jet-producing properties. The similarities of the two, and their differences, will shed light on the behavior of true black holes. In addition, accreting neutron stars can grow considerably in mass, sometimes very quickly, as would happen when merging with a companion neutron star. If it becomes more massive than the critical $2.5\text{--}3.0\text{ }M_{\odot}$, it will collapse violently to a black hole in about 100 ms. Near-black holes can become true black holes.

3.2.1 Neutron Star X-Ray Binaries

The Russian astrophysicist Iosif Shklovskii, of the Sternberg Astronomical Institute in Moscow, suggested in 1967 that Sco X-1 could be a neutron star powered by accretion of matter from a binary companion [185]. However, the short duration of sounding rocket observations prevented identification of any periodic variations in the X-rays that would reveal their binary nature. In December 1970, however, the Uhuru satellite (a.k.a. Small Astronomy Satellite-1 or SAS-1) was launched. This allowed X-ray astronomers to monitor sources for much longer than the few minutes possible with sub-orbital rockets. By 1972 it was known that sources such as Centaurus X-3 and Hercules X-1 were periodically eclipsed *and* that they also pulsed every few seconds [186, 187]. The pulsations were not unlike those of the Crab Pulsar, but they were in X-rays, not radio. Moreover, the pulsations were speeding up over time, indicating that something was spinning up the neutron star, not spinning it down. These observations confirmed that (1) accretion from a companion indeed was occurring and probably was responsible for producing the X-rays, (2) the X-ray star was probably a magnetized neutron star, and (3) the accreting matter was probably orbiting that neutron star in a disk, spinning up the X-ray pulsations as the disk's orbital motion pushed onto the pulsar's magnetic field.

3.2.1.1 Z and Atoll Sources

Some neutron star X-ray sources have spectra similar to Cen X-3 or Her X-1 but do not show strong X-ray pulsations. These have been classified according to how their X-ray emission behaves when one compares high-energy and low-energy X-ray colors at different times in their evolution. A “Z” source, for example, produces a zigzag pattern in the color–color diagram as it changes over a period of hours to days (Fig. 3.4a) [188]. These changes are believed to be produced by variations in the accretion rate, and therefore the luminosity, of the source: the X-ray flux is at its

weakest in the hard, upper horizontal branch (HB), increasing as the source travels down the normal branch (NB), and at its strongest along the flaring branch (FB). Z sources have magnetic fields of $\sim 10^{10}$ G and (on the FB) are accreting near the Eddington limit of $\sim 10^{18}$ g s $^{-1}$ for their $\sim 1.5 M_{\odot}$ neutron star masses [189, 190]. An “atoll” source produces island-shaped and/or boomerang-shaped patterns in the color–color diagram (Fig. 3.4b) and appear to be accreting much more slowly (only a few percent Eddington). Again, the upper “island” state (IS) in the color–color diagram has the weakest luminosity and the lower “banana” state has the strongest luminosity. The banana state is normally divided into two parts: the lower banana (LB; with moderate luminosity) and the upper banana (UB; higher luminosity). While the atoll source color–color diagram might *look* like the NB/FB part of a Z source diagram, it must be remembered that the absolute luminosity of an atoll (and therefore its accretion rate) is much lower than that of a Z source on the NB and FB. Michiel van der Klis (of the University of Amsterdam) has suggested that atoll sources have even weaker magnetic fields ($\sim 10^8$ G) [189, 190]. In 2002 Rob Fender, then at the University of Amsterdam, showed that the Z and atoll sources both produce jets [191], apparently in a manner similar to the isolated pulsars and not unlike the powerful radio galaxies and quasars. Jets appear to be a common phenomenon in any source that rotates rapidly and possesses a magnetic field.

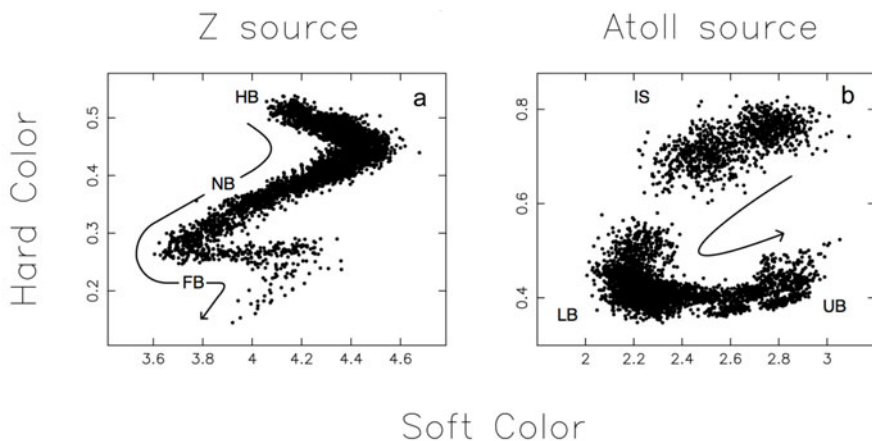


Fig. 3.4: X-ray color–color diagrams of a typical Z source (left) and a typical atoll source (right). Z sources are the most luminous, with the accretion rate increasing toward the Eddington limit as the source evolves along the arrow through the horizontal, normal, and flaring branches. The atoll sources are lower in luminosity (several percent Eddington or less), with the upper island state (IS) being lower than the banana states by a factor of 5 or more [193]. Reprinted from Fig. 1 of reference [192], Copyright (2001) with permission from Elsevier.

3.2.1.2 High-Mass and Low-Mass X-Ray Binaries

There are more than 300 binary X-ray sources known [194, 195]. These divide roughly equally into four classes, based on whether the X-rays are persistent or transient and whether the binary companion is a low-mass or high-mass star. Low-mass X-ray binaries (LMXBs, as they are known) typically have a K or M dwarf companion, or sometimes another type of low-mass star that is $1 M_{\odot}$ or less in mass. High-mass X-ray binaries (or HMXBs) typically have an O or B star companion, $10\text{--}20 M_{\odot}$ in mass. Several cartoons depicting different types of LMXBs and HMXBs are shown in Fig. 3.5 and Fig. 3.6, respectively.

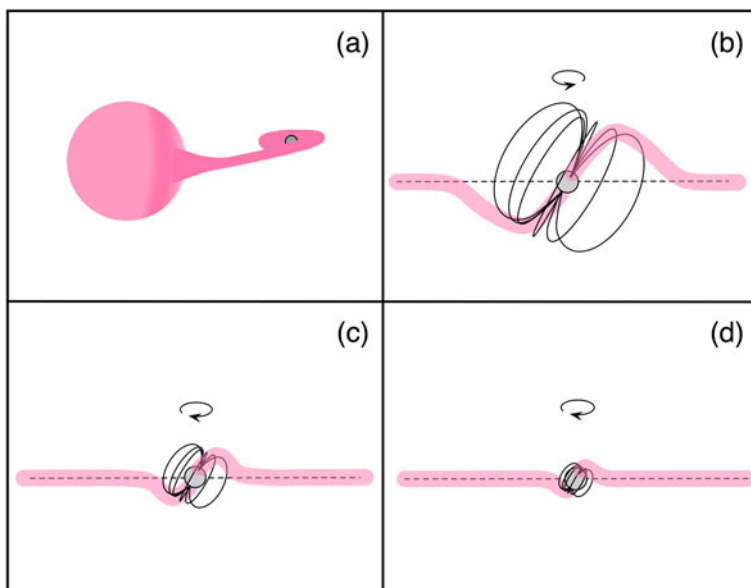


Fig. 3.5: Cartoons depicting various types of low-mass X-ray binaries: (a) general “Roche lobe” overflow, the only method by which a LMXB neutron star (small gray sphere) accretes mass from the low-mass companion red star; note that the accreting material must leave the accretion disk midplane (dotted line) in order to reach the neutron star surface at the poles; (b) X-ray pulsar with a strong ($\sim 10^{12}$ G) magnetic field. Remaining panels use van der Klis’s magnetic classification scheme: (c) Z source, moderate ($\sim 10^{10}$ G) field; (d) atoll source, very weak ($\sim 10^8$ G) field.

In both cases the X-rays are produced when matter is accreted from the companion onto the X-ray star. However, in the LMXB case it is due exclusively to “Roche lobe overflow” (Fig. 3.5a). This occurs when the companion star expands and spills over into the gravitational potential of the X-ray star. Her X-1, all the Z sources (including Sco X-1), and atoll sources are LMXBs. A typical X-ray pulsar, with its strong $\sim 10^{12}$ G magnetic field, is depicted in Fig. 3.5b. In order for the accreting material to reach the neutron star surface, it must reach the polar regions of the magnetized pulsar; the impacting plasma near the poles creates X-ray luminous,

hot spots that rotate in and out of the observer's view, creating the pulsations. A Z source, with its weaker $\sim 10^{10}$ G field is depicted in Fig. 3.5c, and an atoll source with an extremely weak $\sim 10^8$ G magnetic field in Fig. 3.5d. In the latter case, the field is so weak that it probably does not affect the accretion flow much at all.

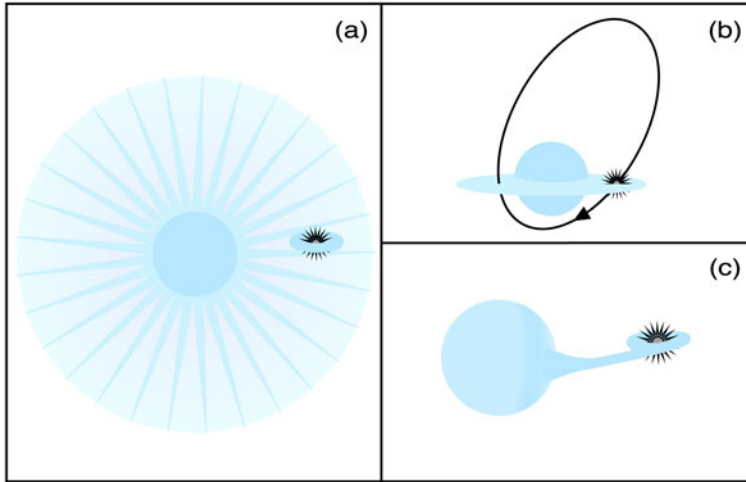


Fig. 3.6: Cartoons depicting different methods of mass transfer from a high-mass companion blue star to the neutron star in HMXB systems: (a) a radiation-driven wind (outflowing streamers) from the blue star is intercepted by the neutron star, creating an accretion disk around the latter; (b) intermittent accretion when the orbiting neutron star crashes through a Be star disk; (c) Roche lobe overflow, although the enormous mass flow from the massive star in this latter case probably will swamp the compact neutron star and quench any X-ray emission.

In the HMXB case, on the other hand, there are two other methods of mass transfer: accretion of part of a powerful stellar wind emitted by the O/B star (Fig. 3.6a) or accretion from an equatorial disk of gas surrounding a companion Be (emission B-type) star when the orbiting neutron star crashes temporarily into that disk (Fig. 3.6b). Roche lobe overflow of the high-mass star also is possible (Fig. 3.6c), but this generally tends to produce enormous accretion rates of $10^{19-22} \text{ g s}^{-1}$ (see Chapter 11) – far greater than the Eddington limit. Cen X-3 is an example of a HMXB and, not surprisingly, is a wind-accreting HMXB.

3.2.1.3 Persistent and Transient X-Ray Sources

Both LMXBs and HMXBs can be “persistent” or “transient” sources. Transient sources are not merely variable. Rather, they appear bright in X-rays for a short time (a few days or weeks) and then disappear. Some transients return years later; some still have not come back after erupting decades ago.

In the high-mass case the X-rays can be transient when the X-ray star is in an elliptical orbit (Fig. 3.6b). One such example is the object LSI +61°303. In this system a Be star periodically expels an equatorial disk of gas, and the neutron star is in a highly elliptical orbit about the Be star [196, 197]. Every orbit the neutron star crashes into the Be disk and produces radio and X-ray emission. A re-ejection of more disk material can cause this emission to have a very complicated time-dependent behavior.

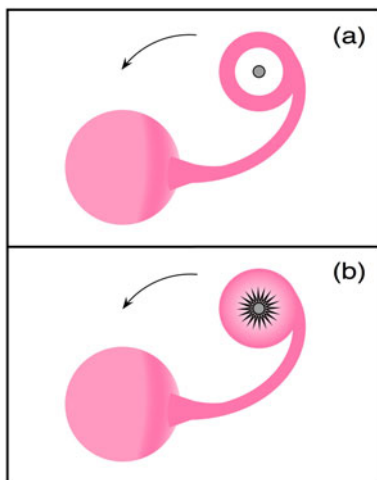


Fig. 3.7: Production of an X-ray nova by a LMXB. The binary system is viewed from the system's north pole, rather than from near its equator (as in Fig. 3.5). (a) For very low accretion rates ($<10^{16} \text{ g s}^{-1}$) the accreting material collects in a ring about the neutron star and does not reach the surface immediately. (b) At some point the collected material becomes turbulent and accretes onto the neutron star, producing a brief, but bright X-ray nova.

In the low-mass case transience is produced in an entirely different way, as has been suggested by Jean-Marie Hameury of the Paris Observatory, Andrew King (Leicester University), and Jean Pierre Lasota (Institut d'Astrophysique de Paris) [198]. For accretion rates below about 10^{16} g s^{-1} , the accretion disk around the X-ray source does not accrete steadily. Instead, matter from the Roche lobe overflow collects in a ring far from the X-ray star (Fig. 3.7a). At some point that ring becomes dense enough to be turbulent and then accretes rapidly onto the neutron star (Fig. 3.7b), but only briefly as the ring is quickly emptied. This burst is so bright, and its onset so rapid, that these sources are often called “X-ray novae”. The ensuing explosion is also relatively cool – for X-ray sources that is (only about 20 million kelvins or 2 keV) – so these are also sometimes called “soft X-ray transients” or SXTs. These outbursts produce typically 10^{43-44} ergs – equivalent to the accretion of only a very small fraction of the earth's mass in gas onto the X-ray star. Famous LMXB neutron star transients include Circinus X-1, Aquila X-1, and the “Rapid Burster” 1730-335 [199]. The Rapid Burster is also famous for its short (seconds

to minutes) bursts of X-rays that are produced as accreting material rapidly drips through the central pulsar's magnetic field.

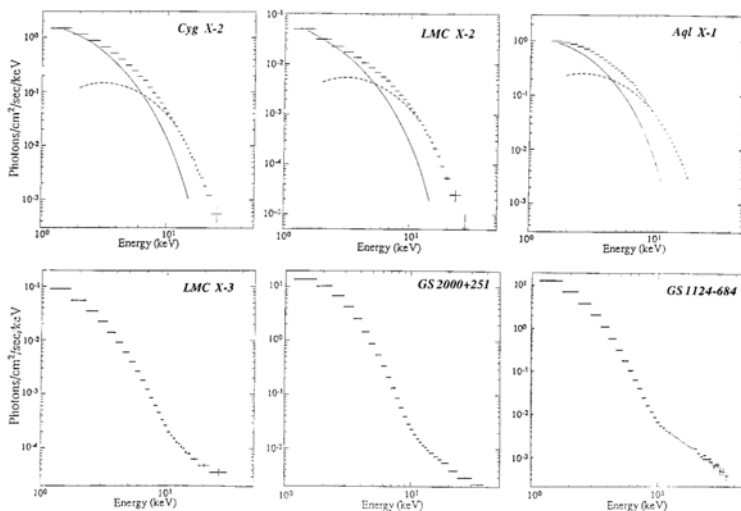


Fig. 3.8: X-ray spectra of three rapidly-accreting neutron stars (top) and three rapidly-accreting black holes (bottom). Note the two-component fits to the neutron star spectral data (horizontal dashes): a soft, accretion disk component (solid line) and a higher-temperature black body component, from the boundary layer where the accretion impacts the neutron star surface (dotted line). Note that the black hole spectra have an even softer disk component (chiefly because black holes have higher mass; see Chapter 12) and no hint of a harder boundary-layer component. They also have a hard power-law X-ray tail that stretches well beyond 10 keV. Reprinted from Fig. 1 of reference [199], ©1996, by permission from Annual Reviews.

3.2.1.4 Signature of an Accreting Neutron Star: How to Tell One from a Black Hole

The X-ray spectrum of a neutron star accreting near the Eddington limit makes it easy to tell that the X-ray star is not a true black hole. It is made up of two thermal components (Fig. 3.8, top). The first is a soft, multi-temperature spectrum that one expects from accretion disks. We will discuss accretion disks in detail in Chapter 12. The hottest temperature in the disk is about 1.5 keV or 17 MK (million kelvins). In addition to the disk emission, however, there is an even hotter component, about 2.4 keV or 28 MK, that is produced when the accreting plasma strikes the neutron star surface, either at the magnetic poles, as in an X-ray pulsar, or in a boundary layer near the equator (if the magnetic field is very weak). (See Fig. 3.10a.) This hotter thermal component is considered to be the signature of an accreting (hard-surfaced) neutron star. Black holes, which do not have hard surfaces, do not have this hotter

thermal X-ray spectral component, as is shown in the bottom of Fig. 3.8. The black hole spectra also have a hard, power-law component above 10 keV; in the past this was thought to be an exclusive signature of a black hole. However, that is no longer the case; rapidly-accreting neutron star systems (Z sources) also show similar hard power-law tails extending beyond 10 keV [200]. In both cases the power-law tail appears to correlate with the radio emission from a relativistic jet produced by the system. This suggests that the hard emission and jet have the same origin, perhaps a hot Comptonizing corona in or above the accretion flow.

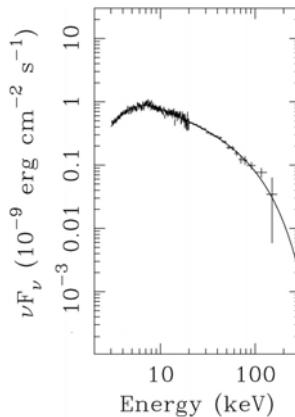


Fig. 3.9: Full X-ray spectrum (3–200 keV) of the atoll source 4U 1728-34 when it was in the slowly-accreting low/hard state ($\sim 2\%$ Eddington). (Note that the vertical axis is different from that in Fig. 3.8: each plotted value is multiplied by ν^2 [or E^2], so that a flat spectrum now has *equal* energy per log ν .) A significant portion of the emission in this state is at a temperature of 35 keV or 400 MK and comes from a very hot, bloated toroidal region near the center of the accretion flow. Credit: Fig. 3.10b of reference [201], reproduced with permission ©ESO.

3.2.1.5 Accretion States

The accretion disk X-ray spectrum described above is usually referred to as the “high, soft accretion state” or just the “high/soft state”,⁴ and it occurs whenever the accretion luminosity in the X-ray binary exceeds about 10^{37} erg s^{−1}. As the accretion rate drops, however, sending the luminosity below this value, the spectrum changes markedly – to a simple power law ($f_\nu \propto \nu^\alpha$), with $\alpha = -0.5$ to -1.0 and a cut-off at very high temperatures (approaching 100 keV; see Fig. 3.9). This hard component is believed to be produced by Compton scattering of cool photons (from the optically thick disk) by hot electrons in an optically thin corona or inner bloated

⁴ These are sometimes called “black hole accretion states”, which is misleading, as such state changes also occur in neutron star systems. Furthermore, the different states are due to changes in the *accretion flow* only, not in any properties of the compact object itself.

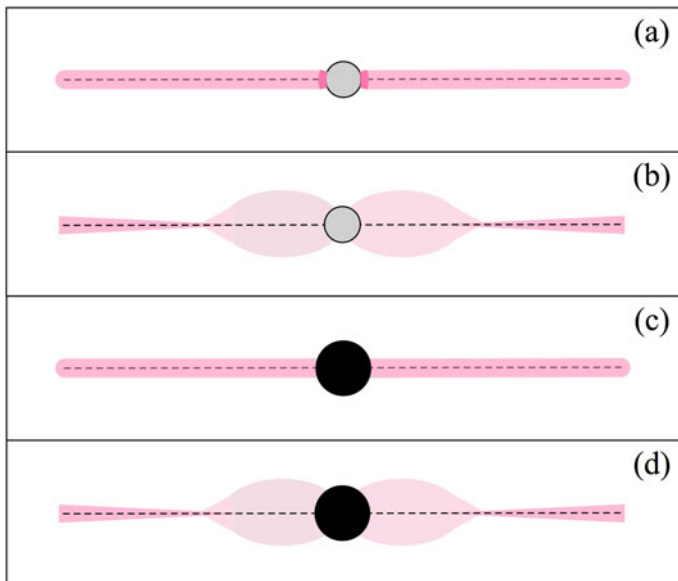


Fig. 3.10: (a) Schematic diagrams of the neutron star “high/soft” accretion state seen in the top of Fig. 3.8. Note the warmer, optically thick boundary layer in the high/soft state, in addition to the cooler accretion disk. (b) The “low/hard” accretion state seen in Fig. 3.9, with the very hot bloated torus in the low/hard state that produces the 400 million K emission. (c, d) similar states for a $5 M_{\odot}$ black hole (with an horizon 15 km in radius). Note the striking similarity to neutron star accretion flow, save for the absence of boundary layer emission.

torus (see Fig. 3.10b). This accretion state is called the “low/hard state” and is a signature of an accretion rate that is well below 10% of the Eddington limit. In the low/hard state there tends to be an increase in “flickering” – high-frequency noise in the X-ray spectrum, with time scales ranging from seconds down to milliseconds.

3.2.1.6 Quasi-Periodic Oscillations

In addition to the random noise in the disk, there appear to be oscillatory, almost periodic, variations in the X-ray flux as well. These “quasi-periodic oscillations”, or QPOs, tend to be of two kinds: a slow variation near a frequency of one oscillation every few seconds (tenths of a Hz) and a very rapid variation near 1000 Hz.

QPOs give clues to the structure of the accretion disk, but we do not yet know quite how to read those clues. At first one might connect the kHz QPO with the rotation of the neutron star, but these pulsations are much less stable than those in the classic millisecond pulsars (see below). Furthermore, a similar high-frequency QPO is seen in black holes as well. A more likely hypothesis is that the kHz HF QPO is associated with the Keplerian orbital frequency

$$\nu_K = \frac{1}{2\pi} \left(\frac{GM}{R^3} \right)^{\frac{1}{2}} \quad (3.4)$$

at a distance of less than twice the neutron star’s radius of about 10 km. One feature that *is* unique to neutron star systems is that there are actually *two* closely-spaced kHz QPOs in these sources – a doublet. It has been proposed that the *difference* between the two oscillations is indeed the pulsar frequency. In fact, this difference is of order 200–300 Hz (corresponding to a period of 3–5 ms), which is reasonable for a pulsar being spun up by accretion. However, while this “beat frequency” difference remains stable at times in some sources, it often tends to decrease when the QPOs get faster for some reason. There is currently no model that can explain all of these and other properties of kHz QPOs in neutron star X-ray binaries [202, 203].

The low-frequency (few Hz) QPOs occur in Z and atoll sources when their disks are in the normal “horizontal branch” accretion state (essentially the low/hard state). These oscillations also are a mystery. If related to Keplerian rotation, then they must occur at a much larger distance from the neutron star – 100–200 km instead of only 20 km for the kHz QPOs. Curiously, though, LF QPOs occur simultaneous with the faster variety, *and* the two types are correlated: sources with more rapid kHz QPOs also have more rapid LF QPOs. The eventual understanding of QPOs should shed considerable light on the process of disk accretion in these systems.

3.2.1.7 Weak Radio Jets

Another important piece of the puzzle is that radio emission from an X-ray binary and the “C-type” LF QPOs often occur together [204, 205]. It is tempting, therefore, to suggest that at least some LF QPOs occur in the accretion disks and are directly related to the acceleration of a radio jet. This trend appears to occur in black hole binaries as well and, therefore, may be an important clue to how accretion disks make jets in all black hole and neutron star systems.

3.2.2 The Strange Case of SS433

In 1979 Bruce Margon, then at the University of California Los Angeles, and his colleagues announced the discovery of a remarkable object, SS433 [206, 207]. Its optical spectrum was truly amazing, with emission lines of hydrogen at enormous positive and negative velocities (of order one-quarter the speed of light; Fig. 3.11a). And these emission lines moved back and forth in the spectrum by many thousands of kilometers per second, with a sinusoidal period of about 164 days. For a short time, much of the astronomical world, observers and theoreticians alike, dropped everything to investigate this very strange beast. Radio observations showed that around the source lay the large supernova remnant W50, with peculiar “ears” (Fig. 3.11b). Inside that remnant was a “corkscrew” jet-like flow (Fig. 3.11c) that changed on the

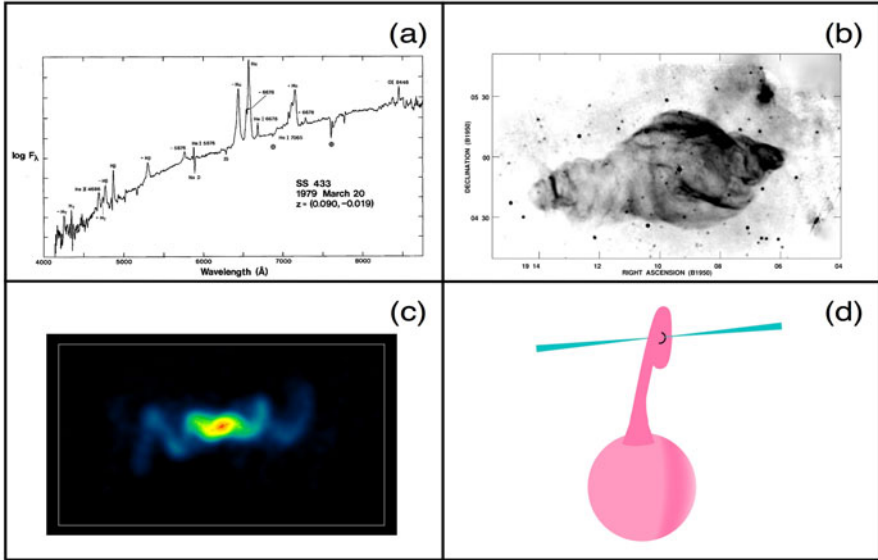


Fig. 3.11: The story of SS433. (a) An early 1979 spectrum [207] showing *simultaneous* emission lines at $+27,000 \text{ km s}^{-1}$ away from the earth (e.g., $-\text{H}\alpha$), -5700 km s^{-1} toward the earth ($+\text{H}\alpha$), and at rest ($\text{H}\alpha$). (b) A VLA image of the supernova remnant W50; SS433 is the strong radio source at the center [208]. Note the “ears” to the east and west (indicating jet-like outflow from the center of the SNR) and also the twisted, helical structure of the nebula. (Top two panels reproduced by permission of the AAS.) (c) High-resolution radio image (NRAO/AUI/NSF) of the SS433 point source seen in panel (b) [209]. The twisting structure in the inner core can be fit to a kinematic model where a jet is continuously ejected both east and west, at $0.26 c$, and that jet precesses around a central axis with a 164-day period. (d) The LMXB model of SS433. Unlike Fig. 3.5a, however, the accretion rate in SS433 appears to be well above the Eddington limit; the compact object is hidden from view.

same time scale as the emission lines did. Surprisingly, this energetic source was weak in X-rays.

When the dust had settled, and dozens of theories had been discarded, the model in Fig. 3.11d was deemed the most likely explanation for the source: a neutron star or black hole is accreting from a companion at a rate much, much greater than the Eddington limit, creating a temporary star-like envelope around the compact object and hiding from our view the X-rays that are produced by that accretion. (This is much the same as the situation inside our own sun, whose interior also produces X-rays, but then whose outer layers absorb this emission and re-radiate it in the visible.) Jets are generated and collimated, probably by a spinning magnetic field and in a manner similar to how X-ray binaries generate jets in much less dense environments. The spinning compact object that produces the jets and the accretion disk must have different rotation axes. This causes the former to precess like a top with a 164-day period, causing the jets’ emission lines to move about the spectrum and the radio jet to take on a corkscrew appearance.

Objects like SS433 are extremely rare; indeed, it is the only one of its kind known in the Galaxy. Because it likely is in rapid Roche lobe mass transfer at perhaps 10^{-6} – $10^{-5} M_{\odot} \text{ yr}^{-1}$ ($10^{20-21} \text{ g s}^{-1}$), its lifetime in this state will be short – only of order 10^{5-6} yr . This is several hundred times shorter than that of a normal X-ray binary, so we should expect to find one SS433 (or fewer) compared to the several hundred X-ray binaries now known. This appears to be the case.

Because of the very rapid mass accretion onto the compact object in SS433, any X-ray emission from the compact object itself is hidden from view. We, therefore, do not know for sure if this source harbors an accreting neutron star or a black hole, although the latter is currently favored.

3.2.3 Recycled Pulsars

In 1982 Donald Backer and his colleagues at the University of California Berkeley announced the discovery of a pulsar with a period of 1.5 ms – PSR 1937+21 [210]. This period is close to the rotation rate at which the spinning neutron star becomes unstable, and the rotation speed at its equator is $0.1 c$. Measurement of its spindown rate, however, indicated an old object ($\tau \sim 2 \times 10^8 \text{ yr}$), implying a magnetic field of only $4 \times 10^8 \text{ G}$ and a $\dot{P}$ of 10^{-19} . This is quite different from the normal pulsars with young ages and high magnetic fields. The most reasonable explanation was that somehow the pulsar was spun *up* by accretion in an X-ray binary system long ago. However, PSR 1937+21 has no evidence of a binary companion larger than a very small planet.

The “millisecond pulsar” remained somewhat of a mystery until the late 1980s and early 1990s when a new class of fast binary pulsars was discovered [165]. The new pulsars are *not* strong X-ray sources; their companions are other neutron stars or white dwarfs, so accretion onto the pulsar from its companion is minimal at the present time. Instead, they are very fast radio pulsars with magnetic fields of only $\sim 10^8 \text{ G}$. These binary systems seem to be ones that not only survived the supernova explosion that formed the pulsar, but also the second event that formed the companion neutron star or white dwarf.

Several clues indicate, however, that accretion must have played an important role in the past. First of all, these are close binary systems in which mass transfer could have taken place. Secondly, the pulsars with the *fastest* spin periods also have the *weakest* magnetic fields. Accretion, therefore, not only must have spun up the pulsar, it also must have helped to destroy the magnetic field in the process. Thirdly, there are some binary radio pulsars with companion stars so low in mass that they could not have formed by any astrophysical means, other than being slowly consumed by the accreting pulsar about which they orbit. Finally, a few more isolated millisecond pulsars have been discovered, lending credence to the idea that these had been spun up by a mass-donating companion in the past, but *ultimately they consumed that companion entirely*.

3.2.4 *Binary Pulsars*

While all pulsars discussed in this section are in binary star systems, the term “binary pulsar” is usually reserved for those unique and rare systems that appear to have two neutron stars. So far, only one of these stars is detected as a pulsar; the other is not, either because the second one is dead or because its fan beam does not sweep the earth. These binary pulsar systems have proven to be rich physics laboratories for testing Einstein’s theory of general relativity, and Russell Hulse and Joseph Taylor of Princeton University were awarded the Nobel Prize in Physics in 1993 for their work in this area.

Hulse and Taylor discovered the first binary pulsar (PSR 1913+16) in 1974 [211]. The two neutron stars in this system are orbiting each other with a period of only 7.75 hours, a semi-major axis of 1.95×10^{11} cm, and a high eccentricity of 0.62. Their closest approach, therefore, is only 1.1 solar radii, making the gravitational interaction of the two compact stars quite strong. Over the next decade Hulse and Taylor painstakingly and repeatedly measured the orbital period and found that it was decreasing at a rate of about 76 microseconds per year. This was precisely the rate that was predicted by general relativity, confirming the existence of a new kind of cosmic radiation that is emitted by strongly-interacting bodies – gravitational radiation. In some cosmic events, such as the formation of neutron stars and black holes, gravitational radiation is expected to be the dominant energy loss. As of this writing, no actual gravitational waves have been detected unambiguously, but several gravitational wave telescopes are being built or planned. The emerging field of gravitational wave astronomy will revolutionize the study of black holes in the next century.

Several additional binary pulsars have since been discovered, including another two or so that are very similar to PSR 1913+16 (PSR 1534+12 and PSR 2127+11C). Each of these has slowly shrinking orbits, whose rate of shrinking will increase with time. After about 300 million years (approximately one orbit about the Galaxy) each of these binaries will coalesce and form a black hole of a few solar masses. During that brief moment, the event is expected to appear much like another object discussed below, the γ -ray burst.

3.3 Neutron Stars in Formation: Supernovae

The identification of the Crab Pulsar and supernova remnant with an actual stellar explosion witnessed by the Chinese almost one thousand years ago confirmed Baade and Zwicky’s suggestion that supernovae can form neutron stars. Observationally, supernovae explosions are divided into a number of classes, depending on what elements are detected in their spectra [212]:

1. Type I – no hydrogen detected in the explosion
 - a. Type Ia: iron lines detected at late times (6 months after the explosion)

- b. Type Ib: some helium also detected; oxygen and calcium detected at late times
 - c. Type Ic: no helium detected; oxygen and calcium detected at late times
2. Type II – hydrogen detected in the explosion
- a. Type IIP: “plateau” light curves: the supernova brightness remains at the same value for 1–3 months after the explosion
 - b. Type IIL: “linear” light curves: the supernova brightness decreases linearly with time after the explosion begins
 - c. Type IIn: “nebular” spectra: narrow emission line cores inside broad emission line wings (like a Seyfert spectrum)
 - d. Type I Ib: similar to Type Ib, but weak hydrogen lines detected⁵

These spectra give clues to the type of star that exploded – a supergiant star for Type II supernovae, a helium-rich Wolf–Rayet star for Type Ib, and an oxygen-rich Wolf–Rayet star for Type Ic.

Further clues to the origins of supernovae come from their environments. Only one type of supernova occurs in most types of galaxy, including elliptical galaxies – the Type Ia supernovae. All other types occur *only* in late type galaxies (spirals and irregulars), and then only in or near regions of recent star formation. All supernovae except Type Ia, therefore, must come from massive stars with very short lifetimes that have recently formed only in the last few million years. Type Ia supernovae must come from old white dwarf stars and, unlike other stellar explosions, produce a lot of iron.

While the precise manner in which most supernovae explode is still controversial, so much theoretical work has been done on the late stages of stellar evolution that we now can say much more about the astrophysics of the above supernova classes. Type Ia supernovae appear to occur in white dwarf stars with carbon–oxygen cores that undergo partial, but probably not complete, violent nuclear burning of those elements to iron (Sections 10.1.2 and 10.3.) The ignition is triggered in this several-billion-year-old system by accretion from a low-mass companion that only recently expanded to fill its Roche lobe and began dumping matter onto the white dwarf at a rapid rate. No appreciable neutron star remnant is left, so these supernovae are of no interest to us in this book. However, Type Ia supernovae have become very interesting to the field of cosmology. Their explosions are all so similar that they can be used as “standard candles” in measuring distances to even high redshift galaxies in which they may occur.

All of the other supernovae (Type Ib/c, Types II) must be “iron core collapse” supernovae, similar to the Crab explosion (Sections 10.1–10.3). Where does the iron core come from? Unlike the sun, in which the helium ash from hydrogen fusion burning simply collects at the stellar center, in a 10–20 $M_{\odot}$ O or B star the helium

⁵ I Ia or not Ia: Some supernova astrophysicists lament the fact that there is a Type I Ib class, but not Ia. One alternate scheme would be to lump all those with strong hydrogen lines (IIP, IIL, IIn, etc.) into a Type Ia. While some astronomers have proposed this scheme, as of this writing it is not generally accepted in the field. Instead, capital letter suffixes are used when the classification is based on the light curve, and lower-case letters are used when it is based on spectroscopic characteristics (emission/absorption lines), resulting in at least four separate Type II classes.

eventually ignites, because of the higher temperatures and pressures, burning to carbon. Then that carbon ash subsequently ignites and burns to oxygen; the oxygen, in turn, burns to silicon; and finally the silicon ignites and burns to iron. Iron, however, cannot fuse to make even heavier elements; it is stable to both nuclear fusion and fission. It simply collects in the center of the massive star as a hot, white dwarf-like core until it reaches the Chandrasekhar critical mass of $\sim 1.4 M_{\odot}$ (see 5.2.4). Then the iron core collapses in a fraction of a second to a hot proto-neutron-star (about 50 km in radius), which quickly cools by neutrino emission to a degenerate neutron star about 10 km in radius. This releases about 5×10^{53} erg of binding energy, much of which goes into neutrinos and other forms of emission. Only a small fraction of this energy, about 10^{50-51} erg, is tapped to eject the stellar envelope and cause the supernova explosion that we detect. (See Chapter 10.) The mechanism by which that is done is still unclear. However, it is clear that we see different types of core-collapse supernovae simply because they occur in different types of envelope, some small and some large.

Table 3.1: Polarization and elongation of supernova explosions

Supernova type	Envelope type and size	Observed polarization	Inferred elongation
II	Hydrogen; 10^{13} cm	1%	2 : 1
IIf	Helium; 10^{12} cm	2%	2.5 : 1
Ib/c	Oxygen; 10^{11} cm	3–5%	>3 : 1

Very interesting recent work on the polarization of light from supernova explosions, especially that done by Lifan Wang of The University of Texas at Austin and his colleagues, may provide important clues to what is causing the explosion. Light becomes polarized when the object that is being viewed is not symmetric, and the interpretation of this polarization is that the explosion is elongated like a cigar, not spherical. Furthermore, Wang’s work has shown that the degree of polarization *increases* as the size of the envelope around the exploding core *decreases* (see Table 3.1) [213]. The conclusion is that *core-collapse supernova explosions are highly asymmetric in the interior and powered (or at least strongly affected) by a jet* that is produced in the collapse. If the explosion occurs in a very large envelope (like a Type II), the jet action is absorbed by that envelope, resulting in a fairly spherical explosion. However, with most of the envelope gone, and only a small Wolf–Rayet oxygen mantle remaining, the jet action has a more visible effect, resulting in a highly-elongated explosion.

Very recently, a new class of supernovae, Type Ic-BL (for “broad lines”) has generated a lot of interest [214].⁶ Only a few are currently known, and they are characterized by expansion velocities that are significantly above the typical $0.02 c$ speeds of normal supernovae and by the ejection of large amounts (up to $0.5 M_{\odot}$)

⁶ Various earlier names for these objects were *hypernovae* or Type Id. These names have fallen out of favor for the purely observational name Type Ic-BL.

of iron peak elements, particularly radioactive nickel (^{56}Ni), which is formed in the explosion and then decays to iron (^{56}Fe) in a few months. Some investigators have argued that the Type Ic-BL supernovae also have very large explosion energies, approaching 10^{53} ergs, but the observations can be explained equally well with a very elongated expansion and a more typical few times 10^{51} ergs instead [213]. The event that caused all the excitement is SN 1998bw⁷, which has been shown to be associated with the γ -ray burst GRB 980425. As γ -ray bursts produce highly relativistic jets, and are thought to signal the birth of a black hole, not a neutron star, the association of one with a supernova suggests a connection between black hole and neutron star formation. We will discuss this more fully in Section 3.6.

Jets, therefore, appear to be important even in the formation of neutron stars, and they continue to be important throughout the life of the pulsar no matter in what environment it ultimately finds itself.

3.4 Isolated Stellar-mass Black Holes: Twinkling of a Little Star

The irony of black holes, of course, is that while they themselves are black, they power some of the brightest and most powerful engines in the universe. However, without fuel for that engine, they are truly black and emit essentially no detectable light. Finding an isolated black hole, therefore, is difficult. In 2002 two groups performing gravitational microlensing observations have, perhaps, done just that. They have observed a few black hole candidates passing between earth and a distant star.

One of these groups, the MAssive Compact Halo Objects (MACHO) project, is a collaboration between Mt. Stromlo and Siding Springs Observatories (MSSSO) in Australia and several facilities in California in the USA (the Center for Particle Astrophysics in Santa Barbara, U.C. San Diego, U.C. Berkeley, and Lawrence Livermore National Laboratory). Their main purpose is to look for large chunks of dark matter in the Galactic halo (brown dwarfs, planets, black holes, *etc.*). The other group, the Optical Gravitational Lensing Experiment (OGLE), is a similar collaboration between Warsaw University Observatory in Poland and Princeton University and the Carnegie Institute of Washington in the USA. In 1999 these groups' patient efforts in monitoring tens of millions of stars in the Magellanic clouds and Galactic center bulge were rewarded with the discovery of the object co-named OGLE-

⁷ Supernovae of all types in all galaxies searched are given names in order of their discovery, using the year of discovery followed by one or more Roman letters. The famous supernova 1987A in the nearby Large Magellanic Cloud (LMC), for example, was the first discovered that year. Counting continues with the capital letters B through Z as each new one is discovered. When more than 26 are observed in one year (consistently the case since 1988), the two-letter combinations are used: aa for #27, ab for #28, ba for #53, and so on. This scheme can accommodate up to 702 supernovae per year – only about twice their present rate of discovery. It is likely that sometime in the early 21st century, the scheme will have to be extended again. Indeed, the total number of supernovae produced per year in the observable universe is several tens of millions!

1999-BUL-32 and MACHO-99-BLG-22 [13, 215]. This was a microlensing event that lasted 641 days (see Fig. 1.3), yielding a long enough time base for a good parallax effect to be detected, giving a lens mass of $37 M_{\odot}$, although with considerable uncertainty. The MACHO team found two more objects 96-BLG-5 ($3\text{--}16 M_{\odot}$ in mass) and 98-BLG-6 ($3\text{--}13 M_{\odot}$). Because of their implied large masses and no detectable emission, it is possible that these objects are black holes of several solar masses or more. If so, then a considerable fraction of the Galaxy's mass, perhaps as high as 1%, might be locked up in one hundred million isolated stellar-mass black holes (SBHs). This is nearly as many as there are neutron stars. We will get a better estimate of this ratio below from the numbers of binary X-ray stars with true black holes in them. The rate implied here is about a factor of 4 greater than that derived below, so perhaps only one of these three microlensing events may be real.

3.5 Stellar-mass Black Holes with Companions: Black Hole Binary Systems

3.5.1 Black Hole X-Ray Binaries

Historically, the first X-ray source to be recognized as a binary system was not an accreting neutron star, but instead the binary black hole object Cyg X-1. And determining this binary nature was done not with X-ray observations but rather with optical observations of the star HDE 226868, which coincided with the radio position of Cyg X-1. In 1972, shortly before Uhuru discovered Cen X-3 to be a binary, B. Louise Webster and Paul Murdin, of the Royal Greenwich Observatory in the UK, showed that the star was a spectroscopic binary with a period of 5.6 days [216]. However, until the 1990s, there were very few other stellar-mass binary black hole candidates known. The great majority of known X-ray binaries in the Galaxy were believed to be neutron stars. Only a few had mass measurements that implied compact objects in excess of $3 M_{\odot}$. As discussed in Chapter 1, this is a good indicator that the accreting X-ray star is, in fact, a black hole.

3.5.1.1 Signature of an Accreting Black Hole: How to Tell One from a Neutron Star

Another major black hole indicator is the spectrum of the X-ray source. As discussed earlier, in contrast to neutron star X-ray binaries, when rapidly accreting, the black hole candidates have somewhat softer accretion disk X-ray spectra (with a peak temperature of 1.2 keV or 14 MK) and no hint of emission from thermal material impacting a hard stellar surface. (See Fig. 3.8.) This softer disk component (indicating a more massive accreting X-ray star) and the lack of additional warmer boundary layer emission (indicating no hard neutron star surface) are considered

signatures of an accreting black hole. Schematic structures for black hole high/soft and low/hard states are depicted in Fig. 3.10c and d, respectively.

3.5.1.2 Accretion States

As with the accreting neutron stars, accreting black holes also display a variety of accretion states. Cygnus X-1's principal states are the high/soft and low/hard states (Fig. 3.12). When the accretion rate is well above $\sim 10^{17} \text{ g s}^{-1}$, the accretion flow displays a soft X-ray spectrum (red curve), with most of its emission below 10 keV, but also showing a hard power-law tail at very high MeV energies. When the accretion rate drops well below this value, Cyg X-1 goes into a low/hard state, in which the X-ray spectrum is dominated by hard X-ray emission in the 20–300 keV region, and the power-law tail in the γ -ray region is surprisingly steeper (blue curve). Clearly at least part of the cool (14 MK) accretion disk in the soft state has been heated to a billion kelvins or more by some mechanism in the hard state.

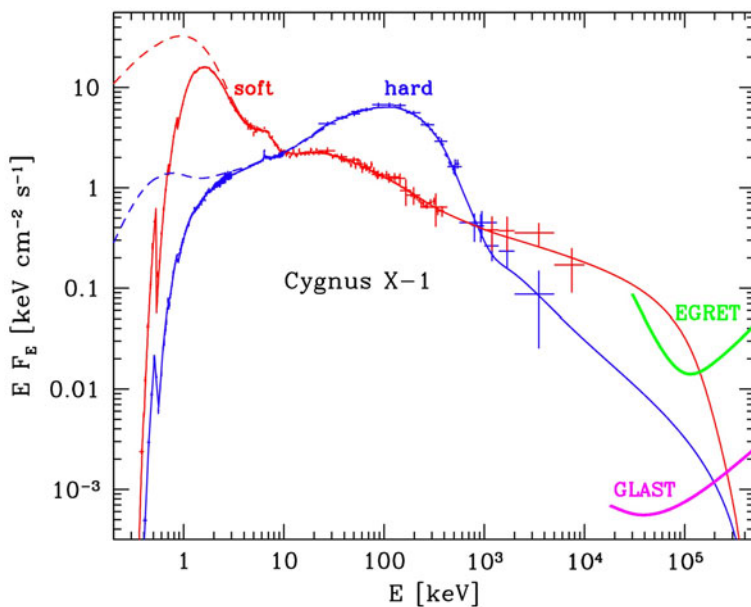


Fig. 3.12: Spectra of the accreting black hole Cygnus X-1 in the high/soft (red) and low/hard (blue) states, again as a νf_ν (or $E f_E$) plot [217]. The solid curves show models of the spectra [218, 219], with the dashed portions showing the predicted intrinsic source emission (without absorption by our interstellar medium). The figure is taken from reference [220] and also shows the approximate γ -ray sensitivities of GLAST and of the EGRET instrument on the Compton Gamma-Ray Observatory. Reproduced by permission of the AAS.

By observing a variety of black hole binaries, other accretion states have been identified, although the exact set of states and their characteristics are still under investigation. The *intermediate state* lies between the low/hard and high/soft states in luminosity and, therefore, in accretion rate. It is distinguished mainly by the types of QPOs produced in this state. Another state, called the *very high, unstable state*, occurs when the accretion rate approaches the Eddington limit. The disk then undergoes dramatic changes in character, often cycling through the low/hard, intermediate, and high/soft states in a matter of minutes, and then repeating. Some observers prefer to identify each individual structure through which the disk passes as a separate state. Because the other states are accretion rate based, however, and because the average accretion rate is near-Eddington in this case ($\sim 10^{19} \text{ g s}^{-1}$ or $10^{-7} M_{\odot} \text{ yr}^{-1}$), we will choose the former identification (*i.e.*, one single, very high state or VHS). The different structures that the accretion disk assumes in this very high/unstable state will be called “sub-states”.

3.5.1.3 Quasi-Periodic Oscillations

QPOs also occur in accreting binary black hole X-ray sources, just as they do in neutron star systems. They also come in low-frequency and high-frequency flavors, *and* the periods of these two different types of oscillations appear to be correlated from source to source. However, there are some key differences from neutron star systems.

Many of these differences can be accounted for by the different masses of neutron stars and black holes ($1.4 M_{\odot}$ vs. $\sim 10 M_{\odot}$), because characteristic time scales (orbital periods, thermal oscillation times, *etc.*) tend to scale proportionately with the compact object mass. Low-frequency QPOs are above 10 Hz for neutron star systems, while they are only about 1 Hz for black holes. And high-frequency QPOs are on the order of 1 kHz vs. 100 Hz, respectively. However, in neutron star systems there appear to be even higher frequency variations than one would expect from this simple scaling, and in black hole systems we do not see QPO doublets at high frequency. The absence of doublets in black hole candidates is another indication that those seen in neutron star systems may be due to the beating of the spinning neutron star magnetic field with a single high-frequency QPO producing in the orbiting disk.

Another key difference in black hole systems is that the LF QPO and HF QPO *do not occur simultaneously*. The LF QPO occurs only in the low/hard state and the HF QPO only in the very high soft state. Because the inner disk radius is large in the hard state, and small in the soft state, both QPOs may be related to the Keplerian frequency in each state. But why the accretion returns to exactly the same very high state, with the same high-frequency QPO, is still not understood.

QPOs involve a substantial portion of the accretion flow: LF QPOs can involve 1–10% of the hard power luminosity in the hard state. They clearly contain a great deal of information on the accretion flow itself, but we are not yet able to read that information.

Table 3.2: Relative numbers of known binary X-ray sources

a. Neutron star X-ray binaries (85%)		
X-ray binary mass class	Persistent	Transient
HMXB	23%	21%
LMXB	~27%	~14%

b. Black hole X-ray binaries (15%)		
X-ray binary mass class	Persistent	Transient
HMXB	1.5%	<1%
LMXB	<1%	12%

3.5.1.4 High-Mass, Low-Mass, Persistent, and Transient Black Hole X-Ray Binaries

Black hole X-ray binaries, like neutron star systems, can be divided into HMXB and LMXB, and those can be divided into persistent and transient sources (see Table 3.2). However, black hole X-ray binaries are distributed quite differently from neutron star systems. *High-mass black hole systems are rare, and persistent ones like Cygnus X-1 are very rare.* This is somewhat surprising, since this source was the first stellar-mass black hole candidate to be discovered. The only two other similar objects known are in the Large Magellanic Cloud, LMC X-1 and LMC X-3. In fact, transient black hole HMXBs are also rare, with only two possible candidates (J0421+560 and J0052.9-7158) in the Small Magellanic Cloud (SMC). It would appear that high-mass systems make black hole binaries in fewer numbers (about 20 times less often) than neutron star binaries. This could be either because black holes are made less often or because the formation of a black hole is so violent that it unbinds the binary.

Similarly, there are very few low-mass binary black hole systems with persistent X-ray emission. One such object, GX 339-4, is sometimes classified as a transient: it is currently in the very high/unstable state, with an average accretion rate very near the Eddington limit, causing rapid changes in its accretion disk structure. Another famous persistent LMXB is the “Galactic Center Source” 1E 1740.7-2942, discovered by the Einstein X-ray satellite. Radio images made by Felix Mirabel (of the French Atomic Energy Commission and the Institute for Astronomy and Space Physics in Argentina) and Luis Rodriguez of the University of Mexico (UNAM) show a double-lobed structure very similar to that of radio quasars [221]. These authors coined the term “microquasar” for the source, a name that now describes an entire class of objects.

As shown in Table 3.2, by far the great majority of black hole X-ray binary systems known are the LMXB transients – the black hole X-ray novae. These transients appear to be very similar to the neutron star SXTs discussed earlier, and also probably operating by Lasota’s mechanism (Fig. 3.7a). While lasting only a short period of time, these are nevertheless very important objects. In fact, it is the steady de-

cline of famous X-ray novae like A0620-00 (Nova Monoceros 1975), GS 2000+251 (Nova Vulpeculae 1988), GS 1124-684 (Nova Muscae 1991) and J0422+32 (Nova Persei 1992) that have allowed a greater understanding of how the accretion state changes with accretion rate for black hole systems.

The reason why most black hole binary systems were not discovered with sounding rockets is now clear. Such flights last only a few minutes and so have little chance of catching a transient X-ray nova explosion. Only the rare persistent sources like Cyg X-1 were observable. However, after the launch of X-ray satellites, beginning with Uhuru in late 1970, it became possible to regularly scan the whole sky for new sources and discover that most of the black holes were hiding as LMXB transients.

3.5.2 *Classical Microquasars*

The year 1992 marks an important turning point in the history of black hole research. In that year, and in the decade that followed, it was discovered that accreting stellar-mass black holes are not merely curious low-mass objects that appear similar to the giant AGN and quasars. Instead, they are identical miniature copies of the macroquasars. In addition to producing thermal heat emission in their accretion disks, black hole X-ray binaries also produce jets. These jets, in turn, propagate at nearly the speed of light and produce miniature radio lobes. And, if pointing reasonably close to our line of sight, the jets can appear to be moving superluminally, just like the macroquasars. Mirabel and Rodriguez' persistent source 1E 1740.7-2942 was the first. While it looks very similar to a typical double-lobed quasar, it is only 7 kpc distant and its lobes extend only 2 pc.

3.5.2.1 GRS 1915+105: The Rosetta Stone

Then, in mid 1992, a team led by Alberto Castro-Tirado, then at the Danish Space Research Institute at Lyngby, used the Russian Space Research Institute's γ -ray satellite GRANAT to discover what would turn out to be the missing link between micro- and macroquasars: GRANAT Source 1915+105 (Nova Aquilae 1992; a.k.a. V* V1487 Aql) [222, 223].

An entire book could be written about GRS 1915+105.⁸ During its many outbursts since 1992 it has passed through essentially all the accretion states exhibited by other binary X-ray sources, and, when in a very high-luminosity state, it can cy-

⁸ Some of my colleagues have (gently) chastised me for emphasizing the importance of GRS 1915+105. To a large extent, they are correct. The source is not typical of most other microquasars in its behavior or X-ray properties. Nevertheless, its historical significance cannot be underestimated, and it is in that spirit that the source is discussed here. Moreover, if there *are* two or more kinds of microquasars, then theoretical efforts to understand, *and unify*, them should be the ultimate goal of this field of research. Perhaps some day Nova Aql 1992 will become as much of a scientific Rosetta Stone as it has been historically.

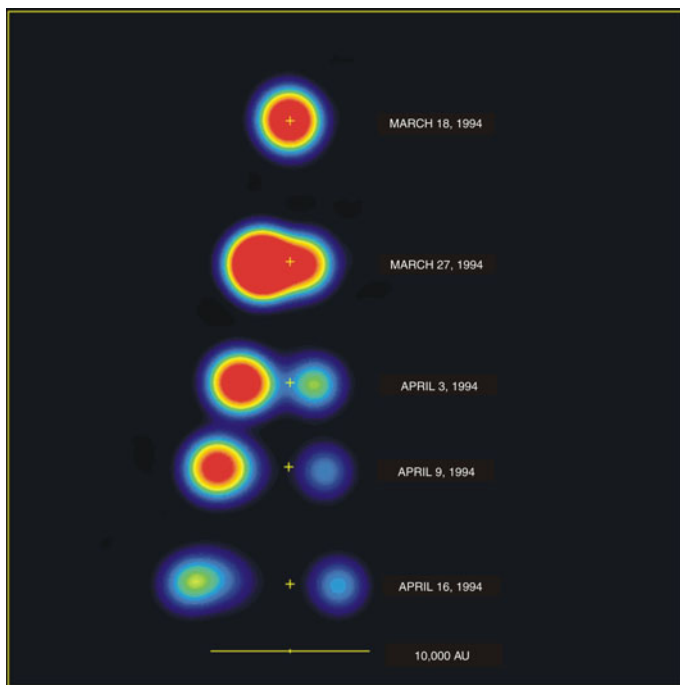


Fig. 3.13: Time evolution of the first superluminal *microquasar* discovered in our Galaxy, as observed with the VLA. On March 18, 1994 it was essentially a point source, but by April 16 the two expanding jet lobes had separated from each other by 10,000 AU (1.9 light-months in about 29 days) – an apparent speed of $v_{app} \sim 2c$. Reprinted by permission from Macmillan Publishers Ltd: [224], copyright (1994).

cle through numerous accretion sub-states in minutes or hours. At a distance of 12.5 kpc (40,000 light years), its luminosity reaches or exceeds $10^{39} \text{ erg s}^{-1}$, the Eddington limit for a $10 M_{\odot}$ black hole. In 1994 Mirabel and Rodriguez, again using the VLA, discovered that not only does the source eject jets like the Galactic Center Source, these jets also produce superluminal motion and occur in powerful outbursts (Fig. 3.13) [224]. The jets are apparently pointed at an angle of about 30° to our line of sight and traveling at a Lorentz factor somewhere between 2.5 and 5, similar to quasar jet speeds. And, in its low/hard state, the source produces a steady, weak jet, similar to that of 1E 1740.7-2942 and to that produced by the supermassive black hole at the center of our Galaxy and to other LLAGN.

In 1996, Jochen Greiner of the Max Planck Institute for Extraterrestrial Physics in Garching, Germany, and Edward Morgan and Ronald Remillard of MIT used the newly-launched Rossi X-ray Timing Explorer to discover a large variety of time-dependent oscillations in GRS 1915+105 [225]. Prior to this date observers could study only long-term behavior of the source, but RXTE allowed the detection of oscillations with microsecond accuracy and good sensitivity. As classified by Tomaso Belloni, of the Osservatorio Astronomico di Brera in Italy, there are three separate

sub-states that this source cycles through when in its very high/unstable accretion state: sub-state A (similar to the high/soft state), B (similar to the very high state), and C (similar to the low/hard state) [226]. Low-frequency QPOs (1–10 Hz) occur in sub-state C only, while two high-frequency QPOs (41.5 Hz and 69 Hz) occur only in sub-state B. Quick jumps between sub-states can occur in any direction but one: a jump from C to B (from a low accretion rate to a very high one) is not allowed; apparently the disk must pass through the high/soft sub-state A on its way to the very high state B, even though it can directly drop from B to C.

A key link between the X-ray and radio ejections was noticed by Mirabel and his colleagues: at certain times the X-ray emission would “dip” or decrease, and shortly thereafter a radio outburst would be detected [227]. Specifically, the inner portion of the relatively cool disk in sub-states A or B would disappear, then the disk would drop into the hot state C and remain there for about 15 minutes. After that, it would suddenly eject a superluminal radio jet component, and then return to a softer disk state about 3 minutes later. After another 5 minutes, the cycle would repeat, for a total cycle time of about half an hour. A similar X-ray dip behavior was later seen in the Seyfert galaxy 3C 120 [228], whose black hole mass is about 30 million solar masses – two million times the value of $14 \pm 4 M_{\odot}$ measured by Greiner for GRS 1915+105 [229]. The cycle time is not proportional to the black hole mass:

$$\tau_{cycle} \propto M^{0.70}$$

so there is a lot to explain here. These observations are some of the most important clues to how jets form in quasars of all types. In particular, the ejection of a jet (strong and superluminal or weak and steady) always occurs when the disk is in a hot state – the same state that is associated with low-frequency, quasi-periodic oscillations in the hard X-ray flux. *Apparently LF QPOs, hot accretion, and jet production are very closely linked.*

3.5.2.2 GRO J1655-40

In 1994 S. Nan Zhang and his colleagues at the Marshall Spaceflight Center in Huntsville, Alabama, discovered another LMXB nova, in the constellation of Scorpius, using the Compton Gamma-Ray Observatory satellite [230]. Shortly thereafter Steven Tingay of MSSSO, and colleagues at ATNF and the Jet Propulsion Laboratory in the USA, imaged the object with a trans-Australian array of radio telescopes and announced that GRO J1655-40 also was a superluminal radio source [231]. B. Alan Harmon, and the Compton Gamma-Ray Observatory team at Marshall Space Flight Center in the USA, showed that, like GRS 1915+105, the radio and hard X-ray emission were correlated [232]. Finally, the late Robert Hjellming and Michael Rupen of NRAO showed that the source has a Lorentz factor similar to that of GRS 1915+105 ($\gamma \approx 2.6$ or about $0.92 c$) [233].

Observations with RXTE revealed that there are two high-frequency QPOs, one at 300 Hz and one at 450 Hz. The faster one oscillates at higher X-ray energies,

indicating that both QPOs must be related to the Keplerian orbital frequency at two different radii. For a measured black hole mass of about $6 M_{\odot}$, these radii would be about 60 km and 45 km, respectively, from the black hole. The latter radius is smaller than the standard disk inner radius for a non-rotating black hole of $6 GM/c^2 = 53$ km (see Section 7.4.1). This is an indication that the black hole may be rotating rapidly, at as much as 50% of its maximum possible rotation speed.

At first it was thought that perhaps the ratio of the high-frequency QPOs in GRO J1655-40 and in GRS 1915+105 ($450/69 = 6.5$) indicated the ratio of black hole masses between the two systems. When the two masses were measured, however, the mass ratio was found to be only 2.3 ± 0.7 , nearly three times smaller than the QPO ratio. The difference is believed to be due to the two black holes having different relative spins. In fact, the predicted radii for the 41.5 and 69 Hz QPOs in GRS 1915+105 are about 300 and 200 km, respectively, and are significantly *outside* the last stable orbit radius (125 km) for a non-rotating $14 M_{\odot}$ black hole. As we shall see in Chapter 7, having an inner disk radius so far outside the last stable orbit radius still may be an indication that the black hole is spinning rapidly, but in the *opposite direction* to its accretion disk!

3.5.2.3 XTE J1550-564

The reader may have noticed that the above QPO pairs are in approximate integer ratios of 2:3 for GRO J1655-40 and 3:5 for GRS 1915+105, suggesting that the two oscillations may be resonating with one another. The discovery of another X-ray nova with radio jets (XTE J1550-564) revealed yet another source with integer QPO ratios: 184 and 276 Hz (2:3). This prompted Włodek Kluźniak of Zielona Góra University in Poland and Marek Abramowicz of Chalmers University in Sweden to suggest that the accretion disks in all three objects were vibrating near their inner edges, with vibrations perpendicular to the disk resonating with vibrations in the radial direction [234, 235]. Such a model might even explain HF QPOs in the neutron star X-ray binaries better than the current pulsar spin beat model. As of this writing, however, no one model has been confirmed for either type of system, let alone for both neutron star and black hole QPOs.

3.5.2.4 The Fender–Belloni–Gallo Diagram

In 2004 Fender, Belloni, and Elena Gallo, also of the University of Amsterdam (UvA), published a paper that may do as much for understanding black hole evolution as Hertzsprung and Russell's work about 100 years before did for the understanding of the evolution of stars [236]. They collected observations of GRS 1915+105, XTE J1550-564, and two other similar sources (GX 339-4 and XTE J1859+226) and, as their colleagues a century ago, they created a color–magnitude diagram, but this time in X-rays. As with the stellar color–magnitude diagram, the FBG diagram can be depicted with real data, or as a cartoon. We reproduce their

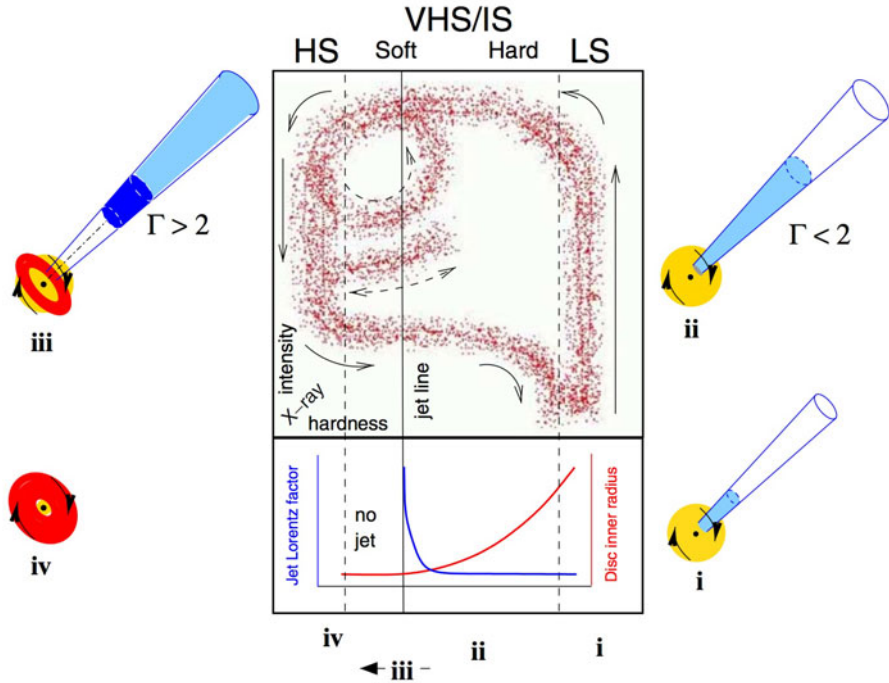


Fig. 3.14: The Fender–Belloni–Gallo X-ray color–magnitude diagram, along with suggested associated disk–jet morphologies [236]. A detailed description is in the text. Reproduced by permission of J. Wiley and Sons, Ltd.

latter rendition in Fig. 3.14, which also shows their association of states, or changes in state, with different types of jet outflow. Note that, while luminosity increases vertically, unlike the HR diagram, the hotter color is to the right, rather than to the left.

The accretion states discussed earlier are all represented in Fig. 3.14 – high/soft at upper left, low/hard at lower right, very-high and intermediate states in the upper middle. The dark red scattered points represent the regions of the diagram that are occupied by most accreting binary black hole systems. Unlike stellar evolution, however, which proceeds on time scales of millions of years or more, binary black hole systems evolve on time scales of weeks and sometimes even minutes. The authors, therefore, were able to trace the evolutionary paths in the FBG diagram and show that black hole evolution generally follows a hysteresis-type pattern (arrows):

- **Region i:** Slowly accreting sources at lower right remain hard when their accretion rate increases. While in this hard state, black hole sources do produce jets, but, FBG argued, these jets are rather slow (few tenths c) and weak.
- **Region ii:** When a rather high X-ray luminosity is reached, the increasingly active source suddenly makes a left turn in the diagram, remaining at about the same apparent intensity in the observed X-ray band, but becoming softer. This

flat evolutionary track is probably due to the limited size of the observing X-ray band; that is, had the much softer X-rays emitted by the source not been absorbed by the interstellar medium (see dashed red curve in Fig. 3.12) the total detected X-ray flux would have continued to rise during this leftward excursion. In this state a jet continues to be produced, and it increases in power, as well as speed. (See schematic plot of jet speed vs. hardness at the bottom of the diagram (blue line).) Region **ii** is terminated by the “jet line”, shown to the left of middle of the diagram.

- Transition **iii**: According to FBG, the crossing of the jet line at high X-ray luminosity produces strong, explosive and fast (relativistic) jet outbursts. This is case for GRS 1915+105 and the other black hole binaries, and may also apply to explosive quasar jet ejections as well.
- Region **iv**: softward of the jet line the source appears to produce no detectable jet – by a factor of 30 in radio luminosity or more. During a full hysteretic cycle, a black hole source then maintains a soft spectrum as it decreases in luminosity, until it reaches a *lower* threshold X-ray intensity. Then it makes another sharp turn, this time back toward the low/hard state, to either remain there or begin another cycle when another accretion episode starts again.

Another curve to note is the red one at bottom. It (schematically) plots the inner radius of the optically thick accretion disk seen in Fig. 3.10d – that is, the *outermost* extent of the optically thin torus also seen in that previous figure. A large “inner radius” indicates the low/hard state with a large torus, while a small one indicates the high/soft state in Fig. 3.10c.

In addition to the main loop, some sources, like GRS 1915+105 take short excursions in and out of the soft state. Each time they re-cross the jet line they eject another strong, explosive jet component. Since its inception, the FBG diagram and related concepts have evolved a little, but the basic ideas remain: black hole sources *do* travel through the X-ray color–magnitude diagram, and their position and transitions therein *do* appear to be correlated with the production of jets by those sources.⁹

Fig. 3.15 shows a schematic evolutionary sequence that fits the basic properties of both X-ray states and jet production. This scheme has been proposed by the author [237], but should be considered in the same manner as cartoons like Fig. 3.10: illustrative and phenomenological, with real physics needing to be filled in and checked to see if these ideas are close to being correct. Only the black magnetic field lines producing a jet are new, however; the interplay between disk and hot torus is well known and has been simulated to some extent before [238]. (See Fig. 12.5.)

The turtle “eye-cycle” begins at lower left in Fig. 3.15 in the soft state. Unlike the larger loop, however, a transition to the hard state (b) occurs, not because of a change in overall accretion rate, but because of local secular and thermal instabilities (which are discussed in detail in Chapter 12). This produces a hard sub-state; even though the accretion rate remains very high, a bloated torus develops anyway. The situation cannot remain, however; as more material accretes in toward the black

⁹ The FBG diagram in Fig. 3.14 is sometimes called the “turtle diagram”, with these short excursions making up the eye and smile of the turtle, and the low/hard state its neck.

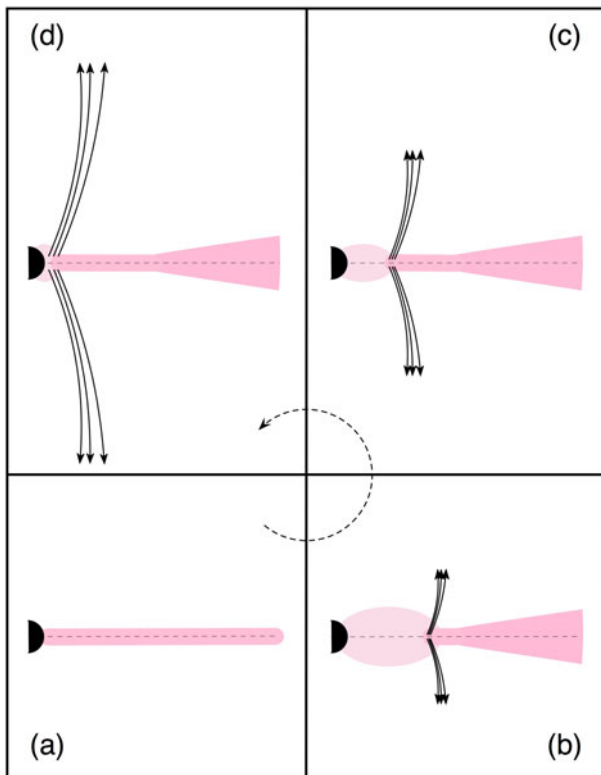


Fig. 3.15: Schematic depiction of the short excursion cycle in Fig. 3.14 that makes up the “turtle” eye [237]. The direction of the circular arrow is the same as in the previous figure. Note the unusual placement of panels (a)–(d). See text for details.

hole, the geometrically thin, cool outer disk eats its way into the hot torus from outside, causing it to shrink in size (c). Eventually, this inward motion of the “inner disk radius” reaches the black hole horizon (d), restoring the situation with which we began in (a).

The jet-producing properties highlighted by the FBG diagram can be understood qualitatively by simply requiring that the jet be produced at roughly at the inner disk radius. That is, there is strong evidence that the jet is launched from or near the hot torus region itself. The jet speed increases as the torus shrinks because it is being launched from deeper within the black hole gravitational potential, and it must attain a faster speed to escape. In (d) the jet speed will be relativistic, producing a shock in the slower jet material that was ejected earlier, just as FBG suggested. Much more on the launching and acceleration of jets is discussed in Chapters 14 and 15.

3.5.3 *The Microquasar Explosion and the Relative and Absolute Numbers of Neutron Stars and Black Holes*

There are now over thirty LMXB black hole systems known or suspected. A great many, and possibly all of them, produce radio emission of some kind and, therefore, probably jets. The name “microquasar” has now been applied to all objects of this type, and the field’s first meeting (involving only a couple dozen people) was held in 1996 at NASA’s GSFC in Maryland. At the present time more than seven such conferences have been held, and they have grown to well over one hundred people and include papers on nearly every type of object discussed in this chapter. In addition, jets have been discovered not only from the LMXB transients, but also even from HMXB persistent sources like Cygnus X-1. The microquasar class has grown to include all jet-producing stellar-mass compact objects, including neutron star X-ray binaries, SS433, and even γ -ray bursts. This book is the first to apply the term consistently to all stellar-mass black hole and neutron star systems.

As discussed above, the absolute numbers of neutron stars (including dead pulsars) in the Galaxy can be estimated easily. The SN rate in galaxies like our own is about one every 50 years. About half of these are Type Ia that obliterate the exploding white dwarf and half are core-collapse supernovae that produce neutron stars. If this rate remains constant over the $\sim 10^{10}$ -year lifetime of the Galaxy, then there should be at least 100–200 million neutron star remnants in our Galaxy at this time.¹⁰

3.5.3.1 Theoretical Estimates of the Galactic Black Hole Population

Now, if we can estimate the relative numbers of black holes and neutron stars, we then can obtain an estimate of the absolute number of stellar-mass black holes in our Galaxy as well. One method of doing this is to use the theory of stellar evolution and death. Current theoretical models of massive stars indicate that ones in the range $10\text{--}20 M_{\odot}$ make neutron stars and those above $20 M_{\odot}$ make black holes (Sections 10.1 and 10.2). Using estimates of the stellar initial mass function (IMF), such as that derived in 1955 by Edwin Salpeter [239], when at Australian National and Cornell Universities, massive stars should be making black holes at a rate about only 2.5 times less often than neutron stars, implying that the current number of black holes in the Galaxy is about 80 million. However, that does not seem to be the case.

¹⁰ And, if the star formation (and supernova) rate were higher in the past (when the Galaxy first formed), that number could be much higher. For example, if the Salpeter IMF was typical over the history of the Galaxy, then there would be about one neutron star produced for every 35 average stars like the Sun. This would imply that as many as 7 billion old neutron stars exist here – 70 times more than if the SN rate had been constant.

3.5.3.2 Observational Estimates of the Black Hole Population

Another method of estimating the numbers of black holes in the Galaxy is to directly compare the number of neutron stars and black holes currently known. At first glance, Table 3.2 would seem to imply that there are about 6 neutron stars for every black hole. But the distributions of types of sources are so different that this result probably is an oversimplification. In fact, the formation and evolution of LMXBs and HMXBs are so different that it probably is not correct to lump them together at all. The latter were made recently, in only the past few million years, whereas the former binaries were created early in the history of the Galaxy, *billions* of years ago. The LMXB statistics also are difficult to interpret. There are about 3 times as many LMXB neutron star systems as black hole systems, but there are about equal numbers of transient X-ray novae. This would imply a rather broad range in the numbers of black holes – from 0.3 to 1.0 times the numbers of neutron stars in the Galaxy, or between 60 and 200 million.

On the other hand, the HMXBs seem to be a rather homogeneous class, with active lifetimes about equal to the few-million-year lifetime of a massive star. It appears, then, that at the present time, there are at least 20 times more recently-made neutron stars in high-mass X-ray binaries than black holes. Unless something catastrophic happens when first forming a black hole to immediately unbind most such binaries, this ratio indicates that massive stars currently make black holes about 20 times less often than they make neutron stars. This would mean that the current number of stellar-mass black holes in the Galaxy is only 10 million and that stars do not begin producing black holes until their initial masses are above $\sim 50 M_{\odot}$.

Why are the theoretical and HMXB estimates of the current black hole formation rates so different? The most likely explanation is that the theoretical estimate does not take into account the high stellar wind mass loss that occurs in present-day (Population I) high-mass stars. (See Chapter 10.) Stars with a lot of heavy elements in their atmosphere absorb the highly luminous radiation emanating from hot massive stars, thereby driving a strong wind and reducing a star's mass by many solar masses over its lifetime. When it comes time for the core to collapse, the star may behave more like a lower-mass $10\text{--}20 M_{\odot}$ one than a $20\text{--}50 M_{\odot}$ one, producing only a neutron star instead of a black hole (Section 10.2 and Fig. 10.1).

So, the current rate of black hole formation is somewhere between one every 250 years and one every 2000 years. If this rate had been constant over the life of the Galaxy, the current number of black holes built up would be roughly 5–40 million. However, if the star and black hole formation rate were much higher when the Galaxy was first formed, these numbers could be significantly higher. For the present, we will take a reasonable number for the stellar-mass black hole count in the Galaxy to be $10^{7.2 \pm 0.3}$, or about 15 million.

3.6 Stellar-mass Black Holes in Formation: Gamma-Ray Bursts

3.6.1 *The Long and Short of Gamma-Ray Bursts*

The year 1997 marks another dramatic turning point in the history of stellar mass black hole studies. Before that time γ -ray bursts were known only as quick bursts of γ -rays in the sky that faded too fast for astronomers operating telescopes at other wavelengths (X-rays, optical, radio) to search that region of the sky and find a counterpart. After 1997, γ -ray bursts (or GRBs) were known to be the most powerful explosions in the universe, annihilating an entire solar mass of matter in a few seconds, far outshining the nearest competitor (the supernovae) by a factor of one hundred million. And that energy is emitted predominantly in lethal high-energy γ -rays, not optical and ultraviolet light like supernovae.

The heroes in this story are the Beppo-Sax X-ray satellite and the team of Italian and Dutch engineers and scientists, led by the late Jan van Paradijs of the University of Amsterdam and University of Alabama, which launched it in 1996. Its Wide Field Cameras had the unique ability to detect and locate a GRB in X-rays within only an arc minute on the sky, instead of the large one-degree errors of γ -ray telescopes, allowing follow-up observations at X-ray and optical wavelengths only a few hours after the burst when it was still bright enough to be observed [240]. For the first time then, a team of Caltech astronomers, led by Mark Metzger, was able to measure the burst optical absorption line spectrum in GRB 970508¹¹ [241]. Optical lines give redshifts (GRB 970508 was at $z \geq 0.835$); redshifts give distances; and distances give burst energies. With few exceptions, the redshifts of most GRBs turned out to be very large, lying in the range $0.10 < z < 8.2$. Their distances, therefore, are as great as the quasars (gigaparsecs), and their luminosities, though short-lived, are up to a million times greater than the brightest quasar. If the GRB explosions that are detected occurred isotropically, then they must produce as much as 3×10^{54} erg in γ -rays alone (more than the entire mass-energy of the sun) and do so in only a few hundred seconds. This result temporarily generated a γ -ray burst “energy crisis”, because in astronomical events, observers tend to see only 1–10% or so of the actual amount of energy generated. As much as 15–150 $M_{\odot}$ of matter, therefore, needed to be annihilated in a few minutes in order to reasonably explain what was being observed. This crisis was resolved with long-term follow-up optical observations of the burst afterglow. Current γ -ray satellites now all have X-ray burst detectors on them, and future ones also will have optical instruments to immediately study the event.

Some things were known about GRBs before optical counterparts were found. The Burst And Transient Source Experiment (BATSE) aboard the now-retired Compton Gamma-Ray Observatory (CGRO) showed that about two or so bursts

¹¹ This burst occurred on 8 May, 1997, hence its name. If there had been more than one GRB on that day, they would have been designated 970508a, 970508b, etc.

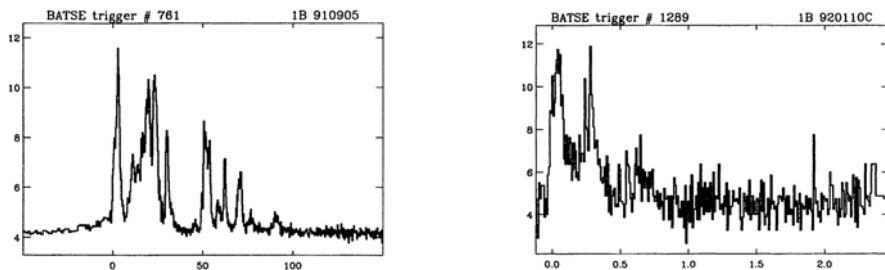


Fig. 3.16: The γ -ray light curves of a LGRB (left) and a SGRB (right) from CGRO BATSE observations [242]. The vertical axes show the γ -ray count rate in thousands per second, and the horizontal axis the time lapse in seconds from when the BATSE observation was first triggered. Note the difference in duration by a factor of 100 between the two, as well as the considerable structure (sputtering) in the GRB activity. Reproduced by permission of the AAS.

are detected per day, and they are distributed isotropically in the sky (a strong indication that GRBs are, in fact, cosmological events) [243]. They also appeared to divide into two distinct classes: *long-duration* GRBs, or LGRBs, (lasting longer than 2 seconds) and *short-duration* GRBs, or SGRBs,¹² (< 2 s; see Fig. 3.16). Both emit a nonthermal γ -ray spectrum in the 100 keV to few MeV region, but the latter tend to be harder (have relatively more high-energy γ -rays). But, for several years, it was only the long-duration variety that were detected at other wavelengths. The SGRBs decayed so fast that, until recently, not even orbiting X-ray telescopes had time to obtain a good position, thus keeping optical observers from finding the object and measuring a redshift distance. Most GRBs show very short time scale variations of a significant fraction of their light, down to milliseconds and possibly shorter. This indicates that the engine that generates this enormous luminosity in γ -rays is less than a light-millisecond (300 km) in size.

3.6.2 Long-duration Gamma-Ray Bursts

Because they were the first to be investigated at other wavelengths, much is known about LGRBs. Many have high redshifts and measured “isotropic” energies up to thousands of times that of supernovae. While the γ -rays last only several minutes at most, their “afterglows” at X-ray, optical, and radio wavelengths can be studied for days and sometimes months, allowing ample time for detailed observations.

¹² An alternate name for the shorter variety is Short Hard Bursts, or SHB, and this term is used fairly often in the literature. However, its opposite (Long Soft Bursts or LSB) has *never* been used in the literature, as far as this author can ascertain. Therefore, in the interest of symmetry, we will use here the increasingly common pair, LGRB and SGRB.

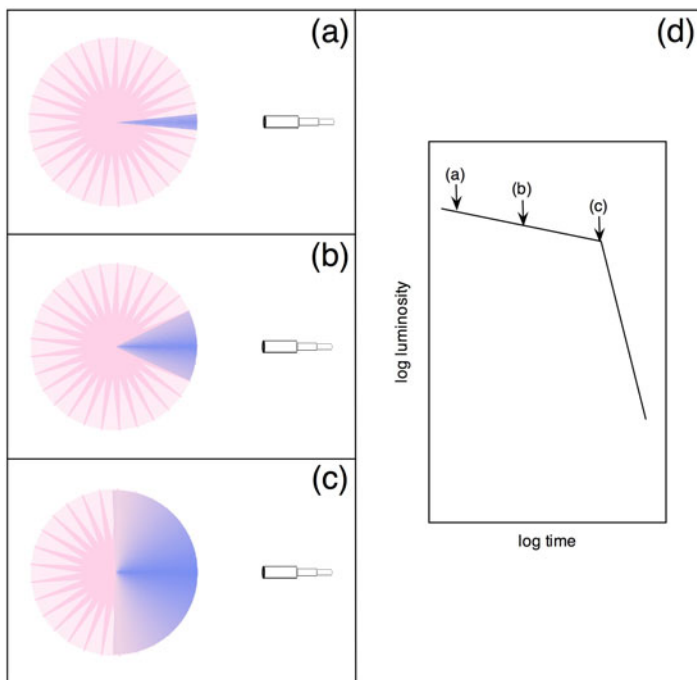


Fig. 3.17: Viewing a spherical, relativistically-expanding fireball as it decelerates through different Lorentz factors: (a) $\gamma \sim 10$; (b) $\gamma \sim 2.5$; (c) $\gamma \sim 1$. The telescope signifies the observer, the blue-shifted portion the amount of the fireball that he actually can observe, and the red-shifted portion the amount that cannot be observed. Note how an increasing amount of the fireball facing the observer can be seen as the expansion slows, overcoming to a great extent the rapid decrease in luminosity of the fireball (d), until the entire structure becomes visible to the observer.

3.6.2.1 The Relativistic Fireball

One of the most puzzling aspects of GRBs was the existence of the high γ -ray luminosities themselves. The small size of the burst, its high γ -ray photon flux, and its apparently optically thin synchrotron spectrum were contradictory. So many photons at such high energies should interact amongst themselves, producing electron-positron pairs, which would thermalize and produce an optically thick relativistic thermal spectrum, not a synchrotron spectrum. The problem was solved by the late Bohdan Paczyński (Princeton University), Peter Mészáros (Pennsylvania State University), and Martin Rees (Cambridge University), who showed that the observations were consistent with an optically thin fireball of electrons, positrons, and magnetic field that was expanding at a Lorentz factor of $\gamma_{\text{fireball}} \sim 300$ ($0.999995 c$)¹³

¹³ It is interesting to note that much of this work was done in the decade *preceding* the discovery that GRBs were at cosmological distances and, therefore, had γ -ray luminosities much greater than supernovae. And an equal body of work was done by other theorists in anticipation that the sources were relatively nearby and much less energetic. So, when observers finally discovered the

[244, 245, 246]. The actual photon energy densities then are lower than those implied by the observations by a factor of $\gamma^6 = 10^{15}$. The very rapid, sub-millisecond, variations in the γ -ray luminosity are believed to be due to irregularities in the outflow, caused perhaps by sputtering of the inner engine or by turbulence in the flow as it interacts with the external medium. These irregularities create shock waves that are *internal* to the flow. And these internal shocks cause the outflow to slow down as they convert kinetic energy into accelerated particles and γ -rays.

This deceleration has an interesting effect (Fig. 3.17). Initially any observer watching it from *any* direction can view only that small portion expanding nearly directly toward him or her (with a half-opening angle of $\theta \sim 1/\gamma$). As the fireball slows, this angle broadens, allowing more of the fireball to be seen and making the rapidly-dimming object appear to dim rather slowly (Fig. 3.17d). Once the entire object is revealed, however, it suddenly shows its true rate of luminosity decrease, rapidly dimming into obscurity. This effect was crucial in deciphering the true nature of γ -ray burst outflows.

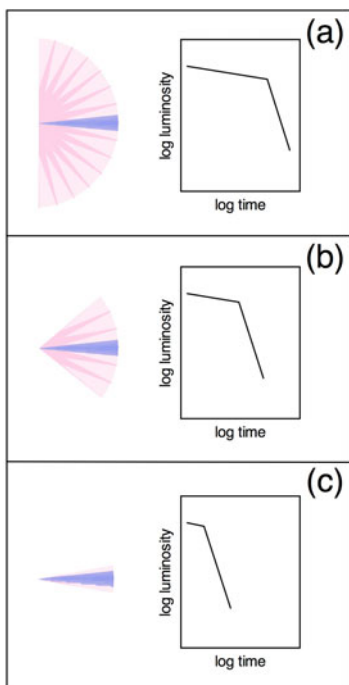


Fig. 3.18: Light curves of different types of relativistic fireballs, with the same color coding as in Fig. 3.17: (a) hemisphere; (b) cone; (c) jet. The hemisphere light curve is no different from the full sphere (Fig. 3.17). However, a conical flow shows rapid decline in brightness much earlier, and a narrow jet shows a beaming break very soon after the initial burst.

true nature of GRBs, theorists were ready for a rather detailed explanation of the phenomenon no matter what it turned out to be!

3.6.2.2 The “Beaming Break”: A Relativistic Jet Reveals Itself

The important property of a relativistically expanding flow coming toward an observer is that one cannot tell what the nature of the flow is initially. A flow of $\gamma \sim 300$ has a visible half-angle of only $\theta \sim 3 \times 10^{-3}$ radians ≈ 10 arc minutes. The fireball itself could be spherical, hemispherical, cone-shaped, or very slender like a jet, and the observer could not determine *which* it was until it slowed down and he or she could see significantly more of it (see Fig. 3.18). This is the same relativistic beaming effect discussed in Section 2.2.3. (See also Fig. 6.7.)

The first burst to reveal its true nature was GRB 990123. For about two days after the initial burst of γ -rays, its optical afterglow flux decayed at a rate of

$$F \propto t^{-1.1}$$

indicating that the small portion of the GRB fireball observed was decelerating and expanding. However, as γ_{fireball} decreased, the amount of the fireball that could be seen ($\pi \theta^2 = \pi \gamma_{\text{fireball}}^{-2}$ steradians) increased, so the decay rate was slowed by observers’ ability to see more of it as time went on. Then, suddenly, after two days, the decay rate abruptly *increased* to approximately

$$F \propto t^{-2}$$

indicating that observers had seen *all* of the fireball that they were going to see. Any further decrease in γ_{fireball} would reveal no more of it, and its flux would decay more rapidly now because only the effects of deceleration and expansion were at play. A team led by Alberto Castro-Tirado (now at the Instituto de Astrofísica de Andalucía, in Granada, Spain) computed the shape of the fireball [248] and confirmed the predictions by James Rhoads (Kitt Peak National Observatory in Arizona) [247] that the flow must have a small opening angle, like a jet, of about 0.1 radians or only 6° . Several other bursts have shown similar “beaming breaks” that also indicate small outflow angles. Because the total solid angle subtended by the outflow is only a fraction $\theta^2/4$ of the entire 4π steradians in a sphere, the inferred total γ -ray energy in the jet is much less than the isotropic values of 10^{53-54} ergs. And more sophisticated modeling of the conical jet flow as being fastest along the cone’s axis, and decreasing away from that axis, has revealed an important result. *Almost all LGRBs produce about the same amount of energy in γ -rays, $\sim 10^{51}$ ergs, similar to that of stronger supernovae.* The resolution to the γ -ray burst energy crisis, then, is that GRBs are slender jets, not spherical explosions. *And there must be many more such bursts (greater than $4/\theta^2 \sim 400$ times more) that we do not see because their jets are pointed away from our line of sight.*

3.6.2.3 Stages in the Brief Life of a GRB

While no γ -ray burst is typical, Tsvi Piran of Hebrew University in Israel has identified several stages in the brief life of a LGRB through which many such objects pass [249] (see Fig. 3.19). They are as follows:

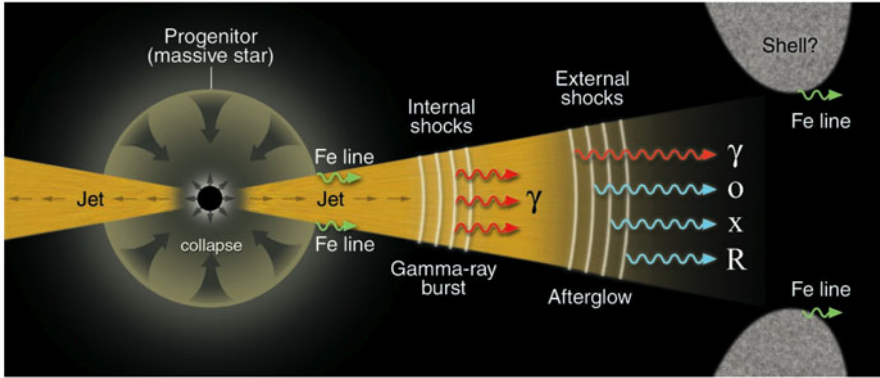


Fig. 3.19: Schematic diagram of the “collapsar” model for a LGRB, also showing the four stages of GRB development: engine (central black hole), energy transfer (jet), burst phase (internal shocks), and afterglow (emitting at many wavelengths). A LGRB is believed to be caused by the formation of a stellar-mass black hole in the center of a massive star, followed by very rapid accretion of central stellar material onto it at hypercritical rates ($>10^{34} \text{ g s}^{-1}$). From Fig. 3 of reference [250]. Reprinted with permission from AAAS.

1. **Inner Engine Activity.** In the first few seconds of the initial explosion, a relativistic jet flow is generated by some means deep in the heart of the GRB central engine. The engine varies (“sputters”) on a time scale δt of milliseconds or less, but stays turned on for the duration T of the GRB (up to several minutes). So, the size of the engine is less than $3 \times 10^7 \text{ cm}$, but the jet outflow itself can be much longer, $L = cT \approx 10^{13} \text{ cm}$ or about 1 AU. This large ratio of engine “on” time to dynamical time is typical of the other jet sources we have discussed in this and previous chapters. It indicates that *the actual engine event is not explosive at all, but rather a relatively benign, quasi-stationary process.*
2. **Energy Transfer Phase.** When the jet flow starts out near the engine, it is not necessarily initially traveling at ultra-relativistic speeds. Instead, it probably is accelerated to the observed speeds ($\gamma_{\text{jet}} \sim 300$) over distances considerably greater than 10^7 cm . During its journey from 10^7 to 10^{13} cm , the jet material is optically thick to γ -rays. These photons interact, produce e^+e^- pairs, and come into thermal equilibrium with those pairs. Because thermal emission from the surface of the jet is not very bright, the GRB will not be visible to observers in this stage. The opening angle of the jet will be $\theta \sim 0.1$ radians, so its width in the outer portion of the jet will be about 10^{12} cm .

3. **Gamma-Ray Burst Phase.** Beyond 10^{13} cm, the jet becomes optically thin to pair production and releases its γ -rays in the direction of the relativistic flow. The irregular nature of the engine creates internal shocks in the jet that accelerate the electrons and positrons. Those particles, in turn, generate relativistic synchrotron emission that ultimately dissipates much of the kinetic energy in the flow. The jet therefore slows down to $\gamma_{\text{jet}} \sim 10$ at a distance of $R \sim 10^{15}$ cm – one (light-) day after the initial engine began firing. Initially, when the jet has reached only 10^{13} cm, we see only a portion of the jet $\pi \gamma_{\text{jet}}^{-2} \sim 3 \times 10^{-5}$ steradians in solid angle. However, as it decelerates this increases to $\pi \theta^2 \sim 3 \times 10^{-2}$ steradians by the time the jet reaches 10^{15} cm. At that point, it also is like a bullet in its rest frame, with a length of $\gamma_{\text{jet}} cT$ and a width $R \theta$ that are about the same size, 10^{14} cm. In our rest frame, however, we see it as a flying pancake only 10^{13} cm thick. Also at this point, $\theta \sim \gamma_{\text{jet}}^{-1}$, and the optical light begins to decrease faster.
4. **Afterglow Phase.** The flow now enters an expansion phase, where it expands sideways at nearly c . The external (bow) shock created by the jet heats the interstellar medium as it sweeps it up, generating X-rays and optical emission that decay rapidly at a rate of $\sim t^{-2}$. In the late stages of this expansion phase, some GRBs show radio emission that initially scintillates, due to the scattering of the light from this point source by interstellar turbulence in our own Galaxy. After about 1 month, the scintillation ceases, indicating expansion to a size of about 10^{17} cm (one light-month). This is independent confirmation that the jet flow in GRBs is relativistic. During the expansion phase the flow decelerates to sub-relativistic speeds, expands to $\sim 10^{16}$ cm laterally, but only $cT \approx 10^{13}$ cm thick in its direction of motion. This non-relativistic flying pancake now acts like a portion of an expanding supernova shell, sweeping up more material and becoming unstable and turbulent. Eventually it is halted altogether.

3.6.2.4 The Supernova Connection

When a GRB host galaxy is detected, the long-duration ones are found to be located near a region of star formation, not in the old galactic halo, for example. This is a strong indication that GRBs come from young, and therefore massive, stars. Furthermore, the energy in a GRB jet appears to be comparable to that in a core-collapse supernova, which is also associated with a massive star. Fair questions to ask, then, are, “Are all LGRBs associated with supernovae? Does every SN produce a GRB?”

The answer to the first question appears to be yes, at least for some events. For example, GRB 980425 and a registered supernova (SN 1998bw) occurred in the same part of the sky and at the same time. They appear to be the same event. This GRB is very close ($z = 0.0086$ or only 34 Mpc away), and this may not be a coincidence; being close would make the light from the supernova easier to detect on its own. It also is significant that SN 1998bw was a Type Ic-BL. Its expansion velocities, and possibly explosion energy, were much greater than typical Type Ib/c supernovae.

Other bursts, like GRB 011121 at a redshift of $z = 0.36$ showed a brightening and then fading of the optical light curve a month or so after the original burst – a “bump” [251]. These appear to be faint, *unregistered* supernova events where the burst and stellar explosion occurred simultaneously, but the light from the burst initially dominated the detected optical emission. As the GRB faded rapidly, even before the SN peaked in brightness, the light curve went from being dominated by the burst to dominated by the supernova light. So the light curve brightened again as the stellar explosion began to peak, then faded away permanently as the SN died out as well. The actual supernova associated with GRB 011121 (*i.e.*, SN 2001ke) was identified several months later, and it indeed occurred in a galaxy at $z = 0.36$.

The answer to the second question, however, appears to be no. Every SN cannot produce a GRB. First of all, supernovae occur at a rate of about one per 50–100 years in galaxies like ours. Even with a generous beaming factor of 1000 ($\theta \sim 0.06$ radians), one would expect to see a GRB every 5×10^4 – 10^5 years per galaxy if every supernova produced a GRB. Yet, the actual observed rate of GRBs appears to be only a few percent of this expected rate – one every 10^6 years or so per galaxy – implying that SN that produce GRBs only occur at a rate of about one per 1000 years in a galaxy like ours. This is close to the rate of black hole formation in our Galaxy that we derived in the previous subsection (one every 250–2000 years). Secondly, the highly relativistic jet flow is more consistent with a black hole central engine than a neutron star. The Crab Pulsar jet, for example, has a speed of only $\sim 0.5c$. Finally, though the evidence is still preliminary, those SN that are associated with GRBs appear to be of the rare Type Ib/c-BL form, not just any type. While some supernovae can produce LGRBs, the evidence suggests that it is only a small fraction of most supernovae that do so – those that leave black holes as remnants.

3.6.2.5 The Collapsar Model: Black Hole Formation in a Massive Star

The model that is consistent with most of the above evidence is the “collapsar” or “failed supernova” model, first suggested by Andrew MacFadyen and Stan Woosley of the University of California Santa Cruz and Lick Observatory [252]. They proposed that a LGRB is nothing less than a jet that is created during the formation of a stellar-mass black hole in the center of a massive star (Fig. 3.19). For some reason the supernova fails to explode, allowing its mantle to rain down onto the newly-formed proto-neutron star (PNS). After a time of several hours to a few days, the PNS accumulates the necessary additional $1.5 M_{\odot}$ or so to drive it to collapse to a black hole. This newly-formed hole continues to accrete, at a prodigious rate, the mantle material that has collected in a disk around it, producing neutrino emission (in much the same manner as a quasar produces light) and jets.

The collapsar model explains the association of GRBs with star-formation regions in their host galaxies and with supernovae, the reason that these bursts have a long duration (essentially the time it takes for the black hole to accrete the disk material), the ultra-relativistic flow, and the low event rate compared to most supernovae. Neutron stars are made by massive stars that begin their life with 10–20

$M_{\odot}$ and end in a classical core-collapse supernova, while black holes are made by massive stars above $20 M_{\odot}$ in a GRB-producing supernova. *It appears that all violent (and even non-violent¹⁴) stellar deaths are accompanied by the formation of a compact object and of a powerful jet.*

3.6.3 Short-duration Gamma-Ray Bursts: The Neutron Star Merger Model

At the beginning of the new millennium, there still were no detected X-ray and optical counterparts to the SGRBs. It was thought that their basic mechanism was much the same: formation of a black hole, followed by rapid accretion of matter at a prodigious rate that produces an ultra-relativistic jet flow and ultimately a brief fireball. Their short duration was thought to be indicative of much less available material for the engine to accrete – one hundred to one thousand times less. This was exactly what was expected for the other proposed models for GRBs, the merger of two binary neutron stars (the ultimate fate of binary pulsars like PSR 1913+16) or the merger of a neutron star–black hole binary. Whereas a collapsar may accrete several solar masses of material from the stellar mantle for up to a few minutes, in a NS-NS or NS-BH merger the coalescence is expected to occur in a fraction of a second, with only the small amount of material in the neutron star outer layers forming the accretion disk. If this was the case, then we would expect SGRBs, if they turned out to be associated with galaxies at all, to be rather old objects that were not associated with massive stars. They should be found in galactic halos, not in star-formation regions, and often in early-type (elliptical) galaxies, not just in late types (spirals and irregulars). And they should be good sources of gravitational radiation that telescopes like the advanced LIGO might be able to detect.

Of course, these were only predictions based on what was believed possible in very compact binary systems. But investigators believed them enough to stake their careers and reputations on building an entirely new satellite observatory. The new set of instruments potentially *could* solve the problem of SGRBs, *if* they behaved like LGRBs and produced a detectable, if brief, afterglow at a variety of wavelengths (especially X-ray and optical) that would allow one to determine the SGRB position and, ultimately, a redshift distance. Everything about this new satellite was fast: its conception, its design and launch, its operation, and its name: Swift. The new spacecraft contained not only a γ -ray burst detector like BATSE, it also had an X-ray and an optical telescope on board. And the latter two could be turned toward a newly-found GRB in less than a minute after it was detected by the γ -ray instrument. The entire process of finding a burst by BATSE, notifying the ROSAT X-ray astronomers so they could then command ROSAT to observe it and obtain a position, and then finally alerting optical astronomers around the world to train their telescopes on it so

¹⁴ Stars much less than $9 M_{\odot}$, like the sun, will not explode. But they do form white dwarf remnants and expanding planetary nebulae. In recent years an increasing number of these nebulae have been found to be bipolar or jet-like – perhaps benign versions of their larger cousins. See, e.g., Fig. 16.1.

they could observe the afterglow was all replaced – with a single orbiting laboratory of electronic machinery [253].

And Swift did not disappoint. The story was very similar to that for the LGRBs; it just occurred eight years later. This time the heroes were the Swift engineering and science teams, the latter led by Neil Gehrels of Goddard Space Flight Center. Launched in late 2004, after a brief checkout phase, Swift quickly began to study GRBs earlier in their development than ever before. Then, during the period 5–7 May 2005 the SGRBs 050509B and 050709 were detected in γ -rays, and then in X-rays minutes later. After only a few hours they were too faint to be seen. GRB 050509B never produced a detectable optical afterglow. However, this still was enough to obtain a position and a redshift: as expected, this SGRB was located in the outer regions of an elliptical galaxy at a redshift of 0.225 [254]. This is exactly what was expected from the NS–NS merger model for SGRBs.

GRB 050709, however, *did* produce a detectable optical afterglow, and it occurred in a $z = 0.160$ star-forming galaxy [255], which is still consistent with the merger model, as binary neutron stars can exist in all types of galaxies. Furthermore, the detection of an optical afterglow is consistent with this star-forming galaxy having more interstellar gas for the burst jet to plow into. The result of several additional discoveries of low-redshift SGRBs, however, did detect a surprising pattern: their burst energies were considerably less than those of LGRBs – only 10^{48-49} erg, a factor of ~ 300 less energetic [256]. Some of these SGRBs clearly were highly beamed, relativistic jets. However, others appeared to not be so; they were simply wide sprays of γ -ray-emitting material. This, again, was consistent with a merger and the formation of a “bare” black hole: there is no large supernova envelope around the engine to collimate the bursts (or to stop cold ones that were not at all collimated to begin with).

In addition to the low-redshift population of SGRBs, there now seems to be a high-redshift ($z = 0.4-1.1$) population as well, with γ -ray brightnesses similar to those of LGRBs [257]. At least 1/3 of all SGRBs detected are at high redshift and perhaps more, which again is consistent with the NS–NS merger model. While it is possible that these distant SGRBs are intrinsically more energetic than their low-redshift counterparts, another possible scenario is that the high-redshift ones produce the typical SGRB 10^{48-49} erg explosions, but we see only those that are highly beamed toward us. If this is the case, then the numbers of distant SGRBs must be considerably more than we currently are seeing. This has interesting implications for gravitational wave detections. Edo Berger of the Carnegie Observatories and his colleagues estimate that, with the current upgrade of the “advanced” LIGO instrument, 2–6 NS–NS mergers per year should be detectable in gravitational waves.

3.6.4 Other Types of Gamma-Ray Bursts

3.6.4.1 Ghosts

Of those LGRBs that can be caught quickly enough to get a position and to study at other wavelengths, about 90% have been detected in X-rays. However, only half of these have detected optical emission. The 40% of the GRBs that are X-ray-only are called dark GRBs or “ghosts” [258]. Ghosts are to be expected, of course. There are many ways in which optical emission can be suppressed. If LGRBs occur in star-formation regions, then many exploding massive stars may be enshrouded in dense, dusty molecular clouds. The clouds are transparent to γ -rays and X-rays, although the latter should show high-column-density absorption; but they are opaque to optical light. It is possible that many GRBs occur in ULIRGs – entire galaxies that are greatly enshrouded in dust. A third possibility is that some ghosts may be low Lorentz factor jets that enter the rapid-decay phase very early. A final possibility is that some ghosts are just extremely distant, perhaps at redshifts of 10 or greater. If so, then their redshifted optical afterglows would be remain largely undetected until JWST comes on line. With or without JWST, however, detailed studies of ghosts’ X-ray emission and absorption could reveal important information about the early universe that we would never obtain from optical observations.

3.6.4.2 X-Ray Flashes and Low-Energy GRBs

Another peculiar class of bursts, the X-ray flashes¹⁵, are not primarily bursts of γ -rays at all, although they do appear in GRB catalogs [259]. Instead, they are intense flashes of X-rays, with weaker accompanying γ -ray emission. Curiously, many X-ray flashes are also ghosts, making up about half of the non-optically-emitting GRBs. Explanations offered for the former are similar to those for ghosts, plus an additional one. It has been suggested that the X-ray flashes are heavily loaded with protons and, therefore, only mildly relativistic fireballs. This would cause the emitted synchrotron emission to be about 100 times lower in energy – in the X-ray region instead of the γ -ray region. It is generally believed that there actually is a continuum of sources between the strongest γ -ray bursts and the X-ray flashes weakest in γ -rays [249].

A third peculiar class of GRB, which also may have low Lorentz factors ($\gamma_{\text{jet}} < 10$), are the long-lag/wide-pulse GRBs [260]. Because these all have long spectral lags (the high-energy γ -rays arrive up to several seconds earlier than low-energy γ -rays), they are often called “spectral lag” GRBs. The presence of a long spectral lag indicates that the beaming factor is low (*i.e.*, the jet is broad and slow). GRB 980425 is a member of this class, and its relation to the Type Ic-BL SN 1998bw

¹⁵ A better name would have been “X-ray bursts”, but that name has been used already for a different type of phenomenon (the much lower energy release of X-rays when matter drips through a neutron star’s magnetic field during accretion, like the Rapid Burster 1730-335).

has prompted suggestions that perhaps all spectral lag GRBs come from Type Ib/c supernovae. Indeed, all long spectral lag GRBs seem to be relatively nearby (within 100 Mpc), and, if their beaming factors are of order unity ($\theta \sim 1$ radian or larger), then the rate of these GRBs and the rate of SN Ib/c are comparable. Long-lag bursts tend to be underluminous compared to other GRBs, but they still are considered to be “high-luminosity” objects.

A new class of truly low-luminosity GRBs [261] has been suggested, since the number of GRBs below luminosities $< 10^{47} \text{ erg s}^{-1}$ increases to perhaps a hundred times or more events per cubic gigaparsec, compared to normal GRBs. At this stage, little is known about them other than there are a lot, and they have very low luminosities (for GRBs). It is much too early to tell whether or not they are produced by a different progenitor or mechanism from, say, the massive star/collapsar one.

The relation among the X-ray flashes, low-luminosity GRBs, and SN Ib/c is not understood yet. However, it is important to remember that these latter supernovae all have little or no hydrogen and helium outer envelope and that they are the most elongated of the supernova explosions (see Table 3.1). The jet could easily break out of the star in some cases and be observed from a large range of viewing angles as a proton-loaded X-ray flash or low- γ_{jet} , long spectral lag GRB. These peculiar types of GRBs, therefore, may be the missing links between the elongated supernovae, which seem to produce neutron stars, and the ultra-relativistic GRBs, which appear to produce black holes. Which remnant the low- γ_{jet} GRBs leave is still unclear.

3.6.5 The End of the World as We Know It?

It is no secret that there are a number of terrestrial and cosmic catastrophes possible that could cause global devastation on the earth of the kind never before seen by man. Global warming (*i.e.*, climate change), nuclear winter, and cometary/asteroidal impacts are all familiar possibilities. And many people are aware that in approximately 5 billion years the sun will begin its death throes and essentially destroy the earth. The end will occur with the sun first entering the red giant phase (scorching the planet and boiling away the oceans), then bathing the solar system in hot (10,000 K) planetary nebula plasma, and finally plunging the earth into an eternal cold as the sun’s tiny white dwarf remnant slowly cools to a black dwarf. Nevertheless, in cosmic terms, all of these events are not that fearful. The first four can be weathered fairly well by the biosphere; the earth has survived most of them before, after all. And plans can be made for the solar death, well in advance, even if it takes thousands of years to colonize planets orbiting a nearby younger star.

Unpredictable events like supernovae and γ -ray bursts, however, are more of a concern. They are enormously energetic, and planning for or thwarting them is extremely difficult. Even early warning systems in the outer reaches of the solar system will be useless, as the speed of the advancing catastrophe is equal to the speed of the signals carrying the warnings (the speed of light). It is important, therefore, to seriously assess the danger and risk of these events and, if significant, to begin to

discuss the issue. Such work has already begun, most notably by Stephen Thorsett in 1995 (then at Princeton), who assessed the effects of a nearby GRB on the earth's biosphere [262]. Many others have looked into this issue, with a variety of conclusions. Here we present only a rough assessment of the problem.

3.6.5.1 Supernova Catastrophes

Supernovae produce about 10^{51} erg of optical–UV emission over a few months (10^7 s) of time, for an average luminosity of about 10^{44} erg s^{-1} – about the brightness of a strong Seyfert galaxy nucleus. While their spectra are a bit more energetic than that of the sun, to a first approximation a supernova will not become a catastrophic threat until its flux at earth is, say, one hundred times that of the sun's – about 10^8 erg $s^{-1} cm^{-2}$ or 100 kilowatts per square meter. To produce a similar $100 L_{\odot}$ energy flux, a supernova would have to be about 3×10^{17} cm or 0.1 parsec away from earth, much closer than the nearest star. This is an extremely unlikely possibility. What is the closest we can expect a supernova to occur in, say, the next 5 billion years? At a rate of about one SN every fifty years in our Galaxy, and with a standard Galactic disk model 1 kpc thick and 15 kpc in radius, we can expect only one of the next one hundred million supernovae, occurring between now and when the sun dies, to explode as close to us as 20 pc. This would make it 400 times *fainter* than our sun – a spectacular event, but not a global killer. So, even though supernovae are fairly common events, in the last billion years our biosphere has survived perhaps *twenty million* of them in our Galaxy. Even the 'minor' mass extinction events probably were caused by other cosmic and terrestrial occurrences, not supernovae.

3.6.5.2 Gamma-Ray Burst Catastrophes

A γ -ray burst, on the other hand, is much more worrisome. It also produces 10^{51} erg, but does so in an intensely narrow beam in less than 100 seconds. This gives an equivalent isotropic luminosity of 10^{52} erg s^{-1} , or 10^8 times brighter than a typical supernova. Ignoring the fact that this energy also is emitted in lethal γ -rays for the moment, we find that a GRB will equal 100 times the sun's energy flux at a distance 10^4 times farther than is the case for a supernova, or about 1 kpc! Our defined lethal distance of 1 kpc is larger than the thickness of the Galaxy itself. So, any GRB in the solar neighborhood (a sphere 500 pc in radius around the sun) probably would kill a substantial fraction of all life on earth, if its jet pointed toward us.

How often might we expect a GRB to point its destructive jet in our direction? Simply from the GRB statistics measured throughout the universe, without needing to know the jet beaming angle, it can be estimated that, in a galaxy like ours, one high-luminosity GRB may erupt and point its jet at us every 10 million years. Most of these events in our Galaxy will be more distant than 1 kpc, but about every 2 billion years ($[15 \text{ kpc} / 1 \text{ kpc}]^2 \times 10^7 \text{ yr}$), one GRB will point at the earth from this lethal distance. This is consistent with life having been allowed to flourish largely

unabated over the last billion years, but it is short enough to indicate that GRBs may be one of the greatest known threats to earth's biosphere. Each powerful burst could cause a complete regeneration of multi-cellular life forms, perhaps more than once, before the sun dies.

Can the current biosphere save itself by colonizing other nearby stars? If the beaming angle is of order 0.1 radians, then the path of destruction is a circle of about 100 pc in radius. And the direction from which the burst would be coming probably would not be known. (It might even come from a massive star hidden deep within a dusty molecular cloud.) In order to escape the event, therefore, colonies would have to be dispersed in a *spherical* volume more than a three hundred light years in radius. While many of them also would be affected by the event, some of those colonies would survive to propagate our species. To be safe, such colonization would have to be accomplished in a fraction of the expected time between lethal events, say in only 100 million years. While this sounds like a large amount of time, the speed of colonization alone (three hundred light years in one hundred million years) is several kilometers per second – nearly as great as the fastest speed that humans have ever traveled during the space shuttle and lunar landing missions.

3.6.5.3 A Cosmic Stay of Execution?

Recent discoveries of the *type* of galaxy in which GRBs occur (particularly the most energetic LGRBs) may make the situation less worrisome. The basic assumption here is that all galaxies produce γ -ray bursts at the same rate. If, however, the Milky Way produced them at a significantly *lower* rate than the average rate per galaxy in the universe, then the time between lethal events may be so long as to exceed the time between now and when the sun becomes a red giant. Work done recently by Krzysztof Stanek of Ohio State University implies that this indeed might be the case [263]. The most powerful GRBs appear to occur preferentially in rather small star-forming galaxies that are poor in heavy elements, *i.e.*, *irregular* galaxies. The rate of production of GRBs by spiral galaxies like our own appears to be much smaller. This result fits with some theories of how GRBs form: that is, it is not simply the collapse of *any* $>20 M_{\odot}$ star that produces a strong GRB, but only those that occur in much more metal-poor galaxies or those that are much more massive than $20 M_{\odot}$ (see Sections 3.5.3 and 10.4). This would put the time between GRB events destroying the solar system to one every 5–10 billion years or more, making at least this method of destruction an unlikely scenario for earthlings.

Whatever the case, it clearly is to our advantage to learn as much about these black hole formation events as possible, and, as with cometary/asteroidal impacts, if there might be any physical evidence of prior GRB damage on earth or elsewhere in the solar system.

Chapter 4

Miniquasars: Elusive Black Holes of Intermediate Mass

Confident that black holes of stellar ($5\text{--}15\ M_{\odot}$) and super ($10^5\text{--}10^{10}\ M_{\odot}$) mass do indeed exist, astrophysicists began naturally to wonder if black holes outside of these mass ranges also existed. Ones smaller than stellar mass (perhaps very small) may have formed in the early universe – primordial black holes. However, up until now there have been few serious attempts to search for them. The same is true of holes much larger than $10^{10}\ M_{\odot}$; if such objects did exist, they would have to be quite isolated from normal galactic matter to have escaped our detection.

However, the intermediate mass range offers some real chance of study and, indeed, some connection with the stellar and supermassive ranges. Are there perhaps black holes of $100\ M_{\odot}$ that form in exotic stars or in compact star systems? Do globular clusters, which look like miniature spherical galaxies of $10^5\text{--}10^7\ M_{\odot}$, have black holes in their centers that comprise 0.1–0.2% of their mass ($10^2\text{--}10^4\ M_{\odot}$)? This is the realm of the intermediate mass black holes, or IMBHs. Finding and studying such black holes could greatly aid our understanding of extreme processes in stellar, cluster, and galaxy evolution and perhaps also help understand how supermassive black holes form.

The search for IMBHs, however, has been fraught with problems. Some observations of objects that first were identified as IMBHs have since been explained, or at least strongly challenged, as something less exotic. These difficulties are similar to those that plagued the early studies of quasars and microquasars, but as better instruments were developed, the supermassive picture became clear. At present, the best evidence that at least some IMBHs exist is right in the center of our Galaxy – in the shadow of the large 3 million solar mass SMBH that lives there. In the next decade or two, the same will happen for IMBHs as well: better instruments will, one way or another, resolve the question of whether or not IMBHs are numerous in the universe and whether or not they have a strong role to play in the scheme of things.

4.1 Definition of an Intermediate Mass Black Hole (IMBH)

The lower end of the IMBH range is normally taken to be somewhere between 15 and $50 M_{\odot}$. A black hole remnant of up to $10\text{--}15 M_{\odot}$ can be left by massive “Population I” stars forming today (*i.e.*, those having a heavy element abundance like that in our sun), whose “main sequence” mass can be as much as $\sim 100 M_{\odot}$. (See Section 10.2.) As the amount of metals¹ decreases in a star, the ability to make a large black hole increases. The most metal-poor stars (the hypothetical Population III) could, in principle, produce black holes as large as $50 M_{\odot}$ if they made significant numbers of stars up to $\sim 140 M_{\odot}$. (See Section 10.4.)

Until recently, the upper end of the IMBH range was taken to be $\sim 10^6 M_{\odot}$, because that represented the smallest black hole masses that had been detected in dwarf elliptical galaxies and bulges of spirals. However, black holes of $< 10^5 M_{\odot}$ have now been found at the centers of dwarf spheroidal galaxies (Section 2.5), so there has been some ambiguity as to whether these are IMBHs or not. This is a bit like the question of whether Pluto is a planet or not, and here we will take a similar scientific, rather than historical, approach. We shall define IMBHs to be black holes that form by processes different from normal massive star evolution *and* objects that are not the single, anchoring black hole at the center of a bona fide galaxy. Admittedly, this definition emphasizes more of what IMBHs are not, rather than what they are. However, that really is the key element of the definition. IMBHs are objects that potentially can illuminate new evolutionary processes in the universe that are either completely unknown now or, at best, only beginning to be understood.

The mass range of IMBHs, therefore, will be taken to be $50 < M_{\text{IMBH}}/M_{\odot} < 10^4$ (or, for short, $10^2\text{--}10^4 M_{\odot}$). Each galaxy may have many IMBHs as isolated objects, in exotic binary systems, as anchors in its star clusters, or even clustered around the central SMBH at the galactic center. The challenge in this area of black hole research is to determine which of the suggested IMBH candidates, if any, are correct and how important black holes of intermediate mass may be in the evolution of stars, galaxies, and supermassive black holes.

4.2 First Type of IMBH Candidate: Ultra-Luminous X-Ray Sources (ULXs) in Star-Forming Galaxies

As discussed in Section 1.1, one method of obtaining a minimum value for an object’s mass is to measure its luminosity at most important wavelengths and then divide by the Eddington luminosity (equation (1.1)). For example, a star of $50 M_{\odot}$ has an Eddington luminosity of $\sim 6 \times 10^{39} \text{ erg s}^{-1}$. So, any object that radiates more than this luminosity must be more massive than $50 M_{\odot}$. Furthermore, if that object

¹ In astrophysics the term “metals” means all substances that are not hydrogen or helium.

is a strong X-ray source, then it is likely to be compact and, therefore either a neutron star or black hole. At $>50 M_{\odot}$, such an object is probably not only a likely black hole candidate, it also possibly one of intermediate mass. In addition, generally the X-rays observed from an X-ray source make up only 10–30% of its total luminosity, so a source becomes interesting as an IMBH candidate when its 2–10 keV luminosity exceeds a few times $10^{39} \text{ erg s}^{-1}$.

Sources that are this bright, and are clearly not supermassive black holes in the nuclei of galaxies are called “ultra-luminous X-ray sources” or ULXs.² Non-nuclear sources of this type that have X-ray luminosities of $>10^{41} \text{ erg s}^{-1}$ often are called hyper-luminous X-ray sources or HLXs, and they may be black holes of 1000–3000 $M_{\odot}$. So far, only a few such HLXs are known: ULX-1 in the starburst galaxy M82 and source #11 in the Cartwheel Galaxy (ESO 350-40), with luminosities of $\sim 10^{41} \text{ erg s}^{-1}$. So, the practical range of ULX luminosities is approximately

$$10^{39} \text{ erg s}^{-1} < L_X^{ULX/HLX} < 10^{41} \text{ erg s}^{-1}$$

which yields a potential range of *minimum* black hole mass of 30–3000 $M_{\odot}$. In fact, while these sources are indeed rare in any given galaxy, it still is possible that they are not accreting at the full Eddington rate, but rather at a rate of 10–30% $\dot{M}_{\text{Edd}}$, which is more common for bright X-ray sources. This gives another factor of three or more, pushing the potential mass range to $10^{2-4} M_{\odot}$, precisely the IMBH range in which we are interested. If these bright X-ray sources are, indeed, black holes, then they would be deserving of the name “miniquasars”, as they would lie intermediate between the original giant quasars and the microquasars in X-ray binary systems.

The ULX field has evolved a lot in the past ten years, and many statements made in the early 2000s are no longer valid in 2011. Recent short reviews by Douglas A. Swartz of NASA/MSFC are available [264, 265] and give a good, brief introduction to the state of the field in that year. Giuseppina Fabbiano reviewed ULXs as well as other types of X-ray sources in galaxies in 2006 [266], and there are earlier 2004 reviews that are still mostly consistent with what is known in 2011 [267, 268]. These all concentrate on properties of ULXs in spiral and other star-forming galaxies.

4.2.1 Observational Properties of ULXs in Star-Forming Galaxies

X-ray sources as bright as this were noted in 1981 by Knox Long of Columbia University (and now at STScI), who used the Einstein X-ray satellite to survey

² The modifier “ultra”, of course, refers to beyond normal X-ray binary luminosities, which are $L_X < 10^{39} \text{ erg s}^{-1}$. In fact, the ULX sources have intermediate, not ultra, X-ray luminosities between normal X-ray binaries and quasars, which can reach $L_X = 10^{44-47} \text{ erg s}^{-1}$. For a while a different, and more descriptive, term was used for these sources: Intermediate X-ray luminosity Objects or IXOs. This term never fully caught on, however, perhaps partly because many astronomers believe that these sources simply represent the bright end of the X-ray binary population and not really IMBHs at all.

galaxies beyond our own [269]. Several non-nuclear sources in spiral arms with $L_X > 10^{39} \text{ erg s}^{-1}$ were found, but it was not clear if these were single sources or groups or even expanding remnants of supernova explosions (SNRs). In the 1990s, the ROSAT satellite found many more ULXs, with at least 1 in every 5 galaxies surveyed and 5–10 in *each* starburst galaxy observed [270]. In 1999, using observations with the ASCA satellite, Andrew Ptak and Richard Griffiths of Carnegie Mellon University reported that X-1 in M82 varied by up to a factor of four, confirming that this bright source was, indeed, a single object [271]. Over half of the discovered ULXs are known to be variable, ruling out the multiple source or SNR hypotheses for these objects.

The association of ULX sources with regions of rapid star formation is an important, if ambiguous, clue to their nature. It means that, whatever process is forming these very luminous black holes (of intermediate mass or otherwise), it is enhanced considerably when many stars, particularly ones of large mass and size, form simultaneously. And the production of the high-brightness X-ray sources goes away as fast as the star formation subsides.

Recent studies comparing Chandra and optical data have found additional clues to the nature of ULXs in star-forming galaxies. In addition to being quite variable, they do not seem to be associated with large star clusters and, in a few cases, a companion O or B star has been found. These properties tend to point toward many ULXs being rather normal HMXBs, which perhaps are undergoing rapid Roche lobe overflow.

4.2.2 Theoretical Reasons for Expecting IMBHs in Regions of Rapid Star Formation

In 2002 two theorists, Simon Portegies Zwart and Steven McMillan (of the University of Amsterdam and Drexel University, respectively), were studying computer simulations of the evolution of the star cluster around the source ULX-1 in M82 [272]. They constructed a star cluster of similar mass inside a computer and allowed it to evolve according to Newton's laws of gravity. They also took into account the masses and sizes of the various stars in the cluster, and the physics that describes stellar collisions and mergers. The result was quite unexpected. Near the cluster center several large, massive, and young OB stars collided and merged to form a very massive star (VMS) of more than $100 M_\odot$. The object continued to grow in mass to about 0.1% of the cluster mass (several thousand $M_\odot$) and then collapsed to a black hole of a similar mass. Since their pioneering work, several other groups have confirmed this as a general process of IMBH formation in young, compact star clusters. While this theoretical work, by itself, does not confirm the IMBH nature of ULXs, it gives a solid physical foundation for how IMBHs might form and why they might exist in regions of very rapid star formation, where there tends to be many large, young OB stars. This physical process is discussed in more detail in Section 10.4.2.

4.2.3 Temperature Problems: Black Hole Fever

Despite theoretical progress on the IMBH formation front, there are some serious problems with the actual identification of all ULX sources with black holes of many hundreds or thousands of solar masses. In Section 12.2.1 we will see that the temperature of the emitted radiation that is expected of a black hole accreting near the Eddington limit is (equation (12.11))

$$T_{rad} \approx 0.23 \text{ keV} \left(\frac{\dot{M}}{\dot{M}_{\text{Edd}}} \right)^{1/4} \left(\frac{M}{1000 M_{\odot}} \right)^{-1/4}$$

However, the color temperature measured for most ULX sources using the ASCA satellite data were 1–2 keV, or 4–9 times *higher* than expected for accreting IMBHs. So, while the intermediate mass black hole hypothesis explained the ULX luminosities, it did not explain their spectra. They simply appeared too hot [273].

A variety of explanations ensued to try to solve this problem, and all pointed *away* from these objects being any more massive than $50 M_{\odot}$. Some of the ones that do not require allowing super-Eddington luminosities are as follows:

- The IMBH could be spinning [273]. As a simple demonstration, if $R = r_g$ in equation (12.11), rather than the $6 r_g$ used, then we would gain a factor of roughly $6^{3/4} \approx 4$ in temperature.
- The object also could be in slightly super-Eddington accretion with $\dot{m} \approx 3$, say. For $\alpha = 0.03$, the slim disk accretion flow would be optically thin to absorption and a factor of ~ 6 hotter because we see the radiation produced deep inside the disk (equations (12.53)).

On the other hand, if one *could* lower the assumed black hole mass in the source (thereby allowing super-Eddington luminosities), one could increase the predicted source temperature by a factor of 1.78 for every factor of 10 lower in mass allowed. Two suggestions that allow super-Eddington luminosities are:

- In 2001, Andrew King of the University of Leicester and his colleagues proposed that the X-rays could, in fact, be “beamed”, either by reflection off the top of a thick accretion disk or by a relativistic jet pointing toward the earth (similar to the current model for FSRQ jets) [274]. In some cases this beaming factor would have to be 100 or more. However, for most sources a factor of 10, plus the assumption of Eddington-limited accretion, would be enough to explain the observed “ultra” X-ray luminosities.
- In 2002, Mitchell Begelman, of the Joint Institute of Laboratory Astrophysics in Boulder, Colorado, showed that magnetized accretion disks could, in fact, produce luminosities somewhat *greater* than the Eddington luminosity – up to a factor of 10 greater for microquasar-type black holes [275].

Taken together, these latter radiative processes could make a factor of 100 difference in what we derive for black hole mass. That is, sources with $L_X \sim 10^{39} \text{ erg s}^{-1}$ ($L_{bol} \sim 3 \times 10^{39-41} \text{ erg s}^{-1}$) could be radiating isotropically at only 1/10th of this

power (because of the beaming), *and* we really should divide this new, smaller L_{bol} by $10^{39} \text{ erg s}^{-1}$ (not 10^{38}) to obtain a lower limit on the black hole mass. If this is indeed correct, the lower limits obtained for all ULXs would be in the range $0.3 - 30 M_{\odot}$, right in the microquasar mass range.

4.2.4 Better ULX Spectral Fits

Since the early days, proponents of the IMBH model for ULXs have countered with better, and lower-energy, spectral data from the large X-ray telescopes of Chandra and XMM. Increasingly, observers are finding that there often is, indeed, a “soft X-ray excess” or peak near 0.1 keV in many ULXs, particularly the brightest ones [276]. This may solve the high-temperature problem for many ULXs and revive the IMBH proposal.

Nevertheless, the possibility remains that many of the high masses may be overestimated by a factor of 10 or more. Independent estimates of the isotropic luminosity (*e.g.*, from reflection by, or ionization of, a surrounding nebula) will be needed to counter the beaming argument. And independent estimates of the black hole masses (*e.g.*, via the motion of a binary or additional stellar neighbors) may be needed to finally settle the controversy.

4.2.5 ULXs in Elliptical Galaxies

Given the association of ULX sources with rapid star formation, it was somewhat of a surprise that some have been found in elliptical galaxies. Although the statistics are small, there appear to be two types of elliptical galaxy ULXs.

The highest-luminosity ULXs in ellipticals appear to occur in globular clusters that orbit those galaxies. They also tend to be quite variable. These may be IMBH candidates, but other explanations also are possible. For example, a team led by Thomas Maccarone of the University of Southampton has discovered the first black hole in a globular cluster: RZ 2109 in the giant elliptical galaxy NGC 4472 [277]. It emits $4.5 \times 10^{39} \text{ erg s}^{-1}$ in 0.2–12 keV X-rays with a soft temperature of $T \sim 0.2 \text{ keV}$. If radiating at the Eddington limit, its mass could be $35 M_{\odot}$, perhaps somewhat lower if the opacity of heavy elements (and not just electron scattering) contributes to the radiation pressure force. If radiating at much less than the Eddington limit, its mass could be far greater.

A team led by Stephen Zepf of Michigan State University observed the oxygen lines produced by the black hole source in RZ 2109 and found them to be very broad, about 2000 km s^{-1} [278]. Such a high velocity in a globular cluster, with a velocity dispersion of perhaps only $10\text{--}20 \text{ km s}^{-1}$, this high-velocity flow might be due to a super-Eddington wind generated by the black hole accreting at much greater than the Eddington limit. (See Sections 13.1.2 and 16.2.3.) The implications

of this observation are significant. The time for a $35 M_{\odot}$ object to sink to the center of a cluster is only a couple of million years (equation (10.11)). So, if there were an intermediate mass black hole at the center, it is possible that it would have ejected this stellar-mass black hole from the cluster and certainly have tidally disrupted the binary system that is feeding the latter. Zepf and his team concluded, therefore, that there is no IMBH in this globular cluster.

The lower-luminosity ULXs in ellipticals do not seem to be associated with any optical object (other than the galaxy itself) and do not vary much. Those properties are more consistent with microquasars like GRS 1915+105 – old, high-luminosity LMXBs that orbit the galaxy alone.

A large census of ULXs in galaxies of all types in 2004 [279] has changed some of the basic ideas about the demographics of ULXs, particularly in elliptical galaxies. Almost half of the ULXs in early-type galaxies (particularly those thought to be in their halos) appear to be background sources, *i.e.*, quasars, *etc.*, not associated with the galaxy at all. So, previous statements that most ULXs in ellipticals lay in their halos were incorrect; instead, most actually lie in the body of the galaxy itself.

4.3 Second Type of IMBH Candidate: Kinematic and Dynamical Evidence for IMBHs in Globular Cluster Cores in Our Galaxy and the Andromeda Galaxy

4.3.1 Motivation

Two important properties of black holes discussed earlier have prompted serious and detailed investigations into looking for kinematic and dynamical (velocity dispersion and light cusp) evidence for IMBHs in the centers of globular clusters. One motivation is the discovery of bright ULX sources in some globular clusters. If this is indicative of active IMBHs in those clusters, then many other globulars may harbor a *dormant* IMBH in their cores also. The second motivation is the discovery that dwarf galaxies harbor a supermassive black hole in their cores with a mass smaller than $\sim 8 \times 10^4 M_{\odot}$, *and* that those galaxies also lie on the $M-L_{\text{bulge}}$ and $M-\sigma$ curves in Fig. 2.18. (See discussion at the end of Section 2.5.) This implies that the spheroidal globular clusters (only 10–100 times smaller than dwarf spheroidal galaxies) also may harbor a black hole in their cores. These arguments suggest that those black holes should be about 0.1% of their clusters' mass, or $\sim 10^{3-4} M_{\odot}$.

4.3.2 Observational Evidence

The first globular to show evidence of an IMBH was the very massive ($1.5 \times 10^7 M_{\odot}$) cluster G1 that orbits the Andromeda Galaxy M31. In 2002 Karl Geb-

hardt, of the University of Texas at Austin (UT), and his colleagues announced the detection of a $\sim 2 \times 10^4 M_\odot$ dark mass in the center of this cluster [280]. The methods used were similar to those the team had used to find inactive black holes in the centers of normal galaxies. Then, a few months later, Joris Gerssen of STScI and his colleagues published the possible detection of a $\sim 1700 M_\odot$ black hole in our own Galaxy's average mass ($4.5 \times 10^5 M_\odot$) globular cluster M15 [281].³ These black hole-to-bulge mass ratios, along with the globular clusters' velocity dispersions ($\sigma_V = 25 \text{ km s}^{-1}$ and 12 km s^{-1} , respectively) placed these new discoveries squarely on the curves in Fig. 2.18.



Fig. 4.1: The only globular-like objects known to contain central black holes: a large cluster orbiting the Milky Way (ω Centauri, a.k.a. NGC 5139; left) and one orbiting the Andromeda galaxy (G1, a.k.a. Mayall II; right). They contain black holes of $3 \times 10^4 M_\odot$ and $2 \times 10^4 M_\odot$, respectively. Unfortunately, neither is a true globular cluster; each appears to be a dwarf spheroidal galaxy that was captured in the past by its parent spiral galaxy. Images by Loke Tan (ST-11000XM with Takahashi FSQ-106) and Michael Rich (HST [283]), respectively. Reproduced with permission.

A few years later, James Ulvestad of NRAO, and his colleagues, discovered a radio source at the center of G1, with an $L_{\text{radio}}/L_{X\text{-ray}}$ ratio similar to that of other accreting black holes [284] (0.1% of the Bondi accretion rate; see Chapter 11). This made the identification of the central object as an IMBH much more likely.

In 2008 Eva Noyola (also of UT) and her colleagues announced the detection of a $\sim 4 \times 10^4 M_\odot$ black hole at the center of the Milky Way's largest globular cluster ω Centauri [285], using methods like those shown in Fig. 1.2. This cluster is very similar to Andromeda's G1 cluster and also quite different from the other globular clusters in our Galaxy or Andromeda. The detection appears pretty firm, and also fits nicely on the supermassive $M_\bullet$ – σ_{bulge} relation.

³ The value of $1700 M_\odot$ was a corrected value from an earlier, somewhat higher estimate [282].

4.3.3 Weighty Problems

As with the ULX sources, however, the IMBH explanation for these observations is not unique. For example, using detailed N-body simulations of M15, a team led by Holger Baumgardt, of the University of Tokyo, showed that the dark mass in the globular cluster core could be explained by the process of “mass segregation” – heavy ($>1 M_{\odot}$) white dwarfs and neutron stars sinking to the center of the cluster [286]. Mass segregation of these dark stars puts a lot of mass and little light near the center, mimicking the effects of an IMBH. These investigators also pointed out that, if there *were* an IMBH at the center of a globular cluster, it would not necessarily have a collapsed core like these candidates. Indeed, their simulations showed that clusters with a black hole in the center could have a normal distribution of stars, with a fairly large cluster core radius. The black hole certainly would have an effect on the density of stars near it, but only at a radius well inside 1% of the globular cluster core radius and only affecting a few tens of stars or so. If correct, this would call into question all dynamical evidence for IMBHs presented so far. This does not mean that an IMBH is not at the center of many or most globular clusters, only that proving so will be much more challenging than originally anticipated.

In the case of G1 and ω Centauri, the problem is not with the black hole detections. (They appear solid.) Instead, it is with the identification of the two as globular clusters themselves. These two star systems have more in common with the cores of dwarf elliptical galaxies than with standard globular clusters. For this reason, it has been suggested that the two are, in fact, partially stripped dwarf ellipticals that have been captured by their respective spiral hosts. Each is, by far, the most massive “globular cluster” in its respective galaxy, and each has several populations of stars, unlike a normal globular, which has only a homogeneous Population II [287, 288]. If, indeed, ω Centauri and G1 are captured dwarf cores, then detection of $\sim 10^4 M_{\odot}$ black holes in their centers sheds no new light on the question of IMBHs in true globular cluster systems. We only learn that supermassive black holes can form in the centers of dwarf galaxies, which agrees with Ho’s work (Section 2.5).

Finally, in 2008, Frank Bash (of The University of Texas) and his colleagues surveyed three globular clusters (including M15) in the radio and did not detect a central black hole radio source in *any* – down to a Bondi accretion rate of 0.01% [289]. This is 100 times lower (in Bondi units) than Ulvestad’s detection of G1. Therefore, there still is no evidence of any kind that true globular clusters harbor IMBHs in their centers that are anywhere near a mass of 0.1–0.2% of the cluster mass.

4.4 Third Type of IMBH Candidate: A Possible IMBH Near the Center of the Milky Way

At present, the most promising evidence that IMBHs exist comes from observations of the center of our own Galaxy. This possible discovery is *not* to be confused with the evidence for a *super*massive black hole (SMBH) there of $4.1 \times 10^6 M_{\odot}$, which was discussed in Section 1.3.2. This is a second black hole of $\sim 1000 M_{\odot}$ which, if confirmed, could have far-reaching consequences for black hole research. The story below is told in three parts: first, some historical background; then, the theoretical work that predicted that such a black hole should exist; and, finally, the tentative discovery of the black hole itself.

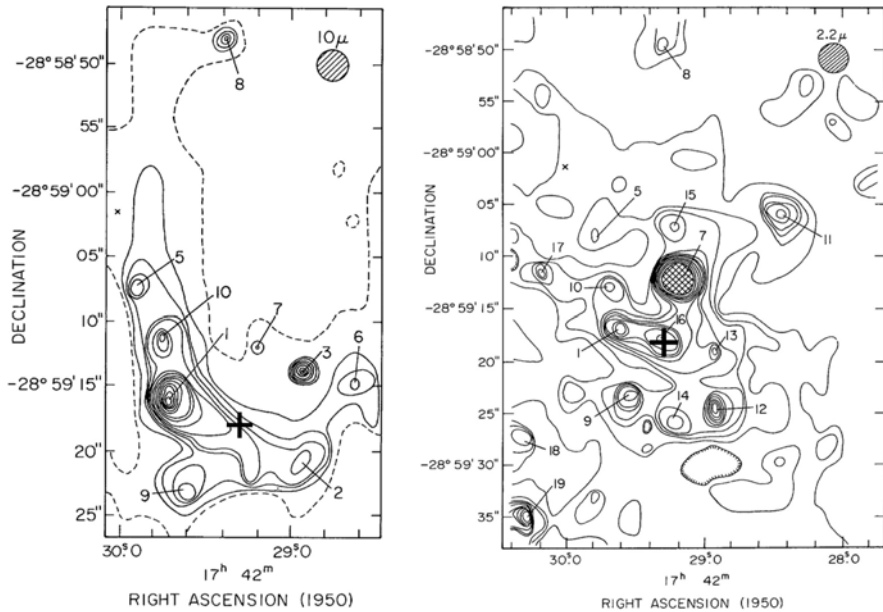


Fig. 4.2: The 19 original Galactic center infrared sources (IRSs) identified by Becklin and Neugebauer [290]. The first ten discovered at $10 \mu\text{m}$ wavelength are at left, and the second nine found at $2.2 \mu\text{m}$ are at right. The supermassive black hole in our Galactic center (crosshairs) is located at approximately $\alpha = 17^{\text{h}} 42^{\text{m}} 29^{\text{s}}.3$, $\delta = -28^{\circ} 59' 18''$ (1950 coordinates), a few arcseconds southwest of IRS 16. Fig. 1.1 showed a $1'' \times 1''$ region centered about that supermassive black hole, which is much smaller than the resolution on either of the above figures. Reproduced by permission of the AAS.

4.4.1 *The Galactic Center in the Near-Infrared*

In 1975 Eric Becklin and Gerald Neugebauer, both then at Caltech, mapped the center of the Galaxy at wavelengths of $10\ \mu\text{m}$ and $2.2\ \mu\text{m}$ (Fig. 4.2) and with a resolution of $2''.5$ [290]. They identified 19 “infrared sources” or IRSs. Even at that time, they realized that much higher resolution was needed to understand the true nature of these sources. With the development of high-resolution infrared imaging techniques a quarter of a century later, with resolution as good as $0''.5$, these sources were resolved into many individual stars by groups at the Max Planck Institute for Extraterrestrial Physics (led by Reinhard Genzel) and at UCLA (led by Andrea Ghez) [291, 292]. The clusters that are of interest are IRS 13 ($\sim 0.2\ \text{pc}$ from the central SMBH), IRS 16 ($\sim 0.1\ \text{pc}$ from the center), and a group of bright stars that is roughly centered on the SMBH itself ($< 0.2\ \text{pc}$ in size). All these stars appear to be very young – only a few million years old – which presented a real puzzle: stars should not be able to form inside $10\ \text{pc}$ of the strong, tidal gravity of the supermassive black hole, and it would take billions of years for stars formed more than $10\ \text{pc}$ away to migrate inward. How did those very young stars wind up so close to the giant black hole?

4.4.2 *The Need for a Second Black Hole at the Center of Our Galaxy*

In 2003 two theorists, Brad Hansen of UCLA and Miloš Milosavljević of Caltech, took a hard look at the data on, and properties of, the two innermost clusters: IRS 16 and the cluster of stars orbiting the central SMBH [293]. They concluded that these young stars indeed could have migrated into the center in a few million years, but only if they remained bound in a compact, very heavy cluster of stars that rapidly sank toward the center of the Galaxy. However, if that cluster had been composed *only* of stars, it could not have survived as it sank inside about $1\ \text{pc}$ from the giant black hole; it would have been torn apart by the SMBH’s tidal forces well beyond where its remnant massive stars now lie. Instead, the only way that it could have survived its plunge inside $1\ \text{pc}$ is for there to have been another very massive black hole *in the center of that cluster* that held it together as it entered the strong tidal region. The mass of the black hole that would have been needed is $\sim 10^{3-4}\ M_\odot$.

The proposed scenario, then, is as follows. A particularly massive ($\sim 10^6\ M_\odot$) star cluster formed much farther than $10\ \text{pc}$ from the central SMBH. In addition to stars, that cluster also formed its *own* IMBH of $\sim 1000\ M_\odot$ in its center, as predicted by Portegies Zwart and McMillan (see Section 4.2.2). Then, over a few million years, the heavy cluster sank toward the central supermassive black hole via dynamical friction as it encountered many, many stars in the galactic center on its journey inward. Somewhere inside $0.1\text{--}0.2\ \text{pc}$, the tidal forces of the giant SMBH became too much for even the second smaller black hole. The latter no longer could

hold onto all of its stars, and the cluster was torn apart. This created the ring of young stars orbiting the central SMBH, all in a time shorter than the few million year lifetimes of those young stars.

4.4.3 The Discovery of a Possible Second Black Hole in the Center of Our Galaxy

At the same time as theorists were predicting an IMBH orbiting our Galaxy's central monster SMBH, a team headed by Jean Pierre Maillard of the Institute of Astrophysics in Paris was studying in detail a third cluster of stars, IRS 13E, with very high angular resolution ($0''.04$) imaging in the near-infrared [294]. The team resolved this cluster into 20 distinct stars, which, as in the previous two clusters, were hot, massive, young stars orbiting the central SMBH (with an orbital speed of $\sim 280 \text{ km s}^{-1}$).

Unlike IRS 16, however, IRS 13E is very compact and is not being tidally disrupted into a ring of stars by the central SMBH. The stars remain clumped together, held together by a strong gravitational field. Maillard and his colleagues followed the internal motions of several stars in the cluster and calculated that a dark mass greater than $1000 M_{\odot}$ – the mass of an IMBH – must lie at the center.

Because this detection uses accurate measurements of individual star motions, it is more reliable than either the ULX or the globular cluster observations that also suggested the existence of IMBHs. Indeed, this result has received only minor challenges since it was published in 2004. Currently, the main concern is that the conclusion relies on observations of only a few stars (4 or 5); more observations of more stars are needed to strengthen this possible detection of an IMBH. Based on this preliminary result, Portegies Zwart and his group have predicted that there may be as many as 50 *additional* one-thousand solar mass IMBHs within 10 pc of our central SMBH. They further point out that, over the course of the history of our Galaxy, IMBHs alone, collecting and merging at the Galactic center, could account for the giant monster that is there now.

4.5 Fourth Type of IMBH Candidate: Low-Luminosity Hard X-Ray/Radio Sources in the Galactic Bulge and Other Population II Systems

4.5.1 Hard X-Ray Sources in the Galactic Bulge

Intermediate mass black holes do not need to be very bright, soft sources (like the ULXs), nor do they need to be emitting no detectable radiation (like the globular

cluster cores or the IMBH candidate in IRS 13E). Black holes of thousands of solar masses, either isolated or in wide binary systems, could be roaming the interstellar medium, sweeping up small amounts of interstellar gas and shining faintly. In Chapter 11 we will learn that the Bondi accretion rate for this process of sweeping up gas is (equation (11.9))

$$\dot{M} \approx 10^{17} \text{ g s}^{-1} \left(\frac{M}{10^3 M_\odot} \right)^2 \left(\frac{\sigma_V}{70 \text{ km s}^{-1}} \right)^{-3} \left(\frac{n}{100 \text{ cm}^{-3}} \right) \quad (4.1)$$

where we have used $1000 M_\odot$ as a typical IMBH mass, 70 km s^{-1} as the typical Galactic bulge velocity dispersion, and 100 cm^{-3} as the typical number density of hydrogen atoms. Even at 100% conversion efficiency, this would produce only $10^{38} \text{ erg s}^{-1}$ (or $\dot{M}/\dot{M}_{\text{Edd}} \approx 10^{-3}$), and in the ADAF model for accretion at low rates (see Section 12.2.2) the actual efficiency of accretion would be more like $\epsilon_{\text{acc}} \approx 10\% (\dot{M}/0.1\dot{M}_{\text{Edd}})$ or $0.1\% (\dot{M}/10^{-3}\dot{M}_{\text{Edd}})$. The luminosity of these freely-roaming IMBHs in this case, then, would be low

$$\begin{aligned} L_X &= \epsilon_{\text{acc}} \dot{M} c^2 \\ &\approx 6 \times 10^{34} \text{ erg s}^{-1} \left(\frac{M}{10^3 M_\odot} \right)^3 \left(\frac{\sigma_V}{70 \text{ km s}^{-1}} \right)^{-6} \left(\frac{n}{100 \text{ cm}^{-3}} \right)^2 \end{aligned} \quad (4.2)$$

This estimate, of course, is strongly dependent on the parameters in parentheses. For example, if the space velocity of the black hole were increased to 150 km s^{-1} , then the luminosity estimate should be *decreased* by two orders of magnitude. On the other hand, if the IMBH were not isolated, but rather at the center of a tidally stripped star cluster of order $10^4 M_\odot$, then the luminosity estimate would *increase* by three orders of magnitude. (The entire cluster could partake in sweeping up the interstellar matter.)

What should these weak X-ray sources look like? By analogy with both weak binary X-ray sources and LLAGN, they should be *hard* X-ray sources, with a large fraction of their emission in the 10–100 keV region of the spectrum. In fact, according to the ADAF model, they could remain hard X-ray sources until they reached accretion rates as high as $\dot{M}/\dot{M}_{\text{Edd}} \approx 0.03\text{--}0.1$. So, they would appear as hard X-ray sources with $L_X \approx 10^{34\text{--}37} \text{ erg s}^{-1}$. If their IMBH status could be confirmed, we would classify them as low-luminosity miniquasars.

Hard X-ray and γ -ray surveys of the Galactic center region, using the INTEGRAL (INTErnational Gamma-Ray Astrophysics Laboratory) and Chandra satellites [295, 296], show ~ 60 hard sources down to $\sim 10^{35} \text{ erg s}^{-1}$ within about 1 kpc of the Galactic center, and 2000 sources down to luminosities of $10^{31} \text{ erg s}^{-1}$ within ~ 20 kpc of the central black hole. If Portegies Zwart's team's estimates are correct, then several hundred of these hard, weak X-ray sources (10–20% of them) could be IMBHs. This fraction would not contradict current ideas that most of these faint sources are cataclysmic variable binaries with a magnetized white dwarf and a main sequence star.

The NASA hard X-ray mission NuSTAR (to be launched in 2012) will be able to detect sources within 10 kpc of the earth in the 10–30 keV and 30–60 keV bands, with luminosities as weak as 10^{32} and 10^{33} erg s $^{-1}$, respectively. As a deep survey within a few hundred parsecs of the Galactic center is one of the key programs in the NuSTAR project, the mission should be able to characterize and classify these hard Galactic center sources and attempt to identify any as IMBHs.

4.5.2 Suggested Radio Searches for Low-Luminosity Mini quasars

Tom Maccarone and his team also have suggested that an even better method of finding low-luminosity mini quasars would be to search for them in the radio region of the spectrum [297, 298]. They discussed this idea in terms of globular cluster X-ray sources, but we can equally apply it to the Galactic center sources. The equation for the predicted radio flux at a frequency of 5 GHz is

$$F_{5\text{ GHz}} = 9.4 \text{ mJy} \left(\frac{L_X}{6.5 \times 10^{34} \text{ erg s}^{-1}} \right)^{0.6} \left(\frac{M_\bullet}{10^3 M_\odot} \right)^{0.78} \left(\frac{d}{8 \text{ kpc}} \right)^{-2} \quad (4.3)$$

Such a flux would be quite easily detectable at the Very Large Array in New Mexico, especially if we already have a hard X-ray candidate to observe.

4.6 Summary and Synthesis

Several candidate sources and objects have been suggested as harboring intermediate mass black holes:

- ultra-luminous and hyper-luminous *soft* X-ray sources (with more than a quarter million solar luminosities in X-rays alone), both in regions of very rapid star formation and in old globular clusters;
- possible dark concentrations of matter of perhaps $10^{3-4} M_\odot$ at the centers of globular clusters;
- a dark concentration of dark matter at the center of the compact, young star cluster IRS 13E that lies within 0.2 pc of our own Galaxy’s *supermassive* hole;
- faint, *hard* X-ray sources within 20 pc of our Galactic center, and presumably near the centers of other galaxies as well.

All of these have very plausible alternative explanations. The strongest identification appears to be the IRS 13E one, although additional observational work on the source has produced only “tantalizing but presently unconvincing” evidence of the presence of an IMBH there [132]. Data will be needed to confirm the detection. The latter observation has direct dynamical evidence, in addition to the indirect arguments that such a large secondary black hole is needed to drag in the very young stars

observed near the Galactic center. The next strongest identifications are the HLX sources (the few with luminosities exceeding 10^{41} erg s $^{-1}$) found in regions of rapid star formation. Their properties strain the beaming and super-Eddington emission alternative explanations and, taken again with the theoretical expectation of IMBHs in young, compact star clusters, the identification of the very brightest ULXs with intermediate mass black holes seems very plausible.

If the IRS 13E result can be confirmed, it will be of tremendous significance to astrophysics. It will mean that

- IMBHs *do* exist;
- they indeed can be formed at the center of massive, compact star clusters, as predicted theoretically;
- such black holes can, and do, get well within 1 pc of a supermassive black hole;
- and these IMBHs and their stars probably can merge with the SMBH (toward which they are inspiraling) at a rate that could account for much of the central SMBH's mass itself.

Furthermore, the existence of an IMBH in a very young star cluster will mean that, while many of the ULX sources found in star-forming galaxies may be regular high-mass X-ray black hole binaries, some indeed probably *are* IMBHs in newly-formed compact star clusters – most likely the brightest and softest HLX ones, with luminosities approaching 25 million solar luminosities in soft X-rays alone.

The field of IMBH study has the potential to link stellar and supermassive black holes, not only in mass and luminosity but also in their processes of formation and evolution. It is a brand new field, only a few years old, and should see significant growth in the next decade.

Part II

Physics of Black Hole Engines

Chapter 5

Physics in Everyday Life: A Review of Newtonian Mechanics

Having surveyed the current state of black hole engine observations, our task in the second part of the book is to gather the tools that we will need to analyze these black hole engines. This will, of course, require that we gain a knowledge of relativistic physics and astrophysics, including Einstein's theory of gravity. Nevertheless, from time to time we will need Newton's classical theory of gravity, either as a stepping stone to Einstein's or, on its own, to calculate some astrophysical phenomena that do not require the accuracy of Einstein's full set of equations. This brief chapter collects together those aspects of Newtonian mechanics that are necessary to discuss Einstein's theory in the next few chapters.¹ Furthermore, most of these aspects will be repeated in Chapters 6, 7, and 8 in their full relativistic forms. Finally, the remaining physics that will be needed to understand black hole formation, fueling, accretion, and outflows will be presented in Chapter 9, after we have spent a fair amount of time discussing relativity.

5.1 Mechanics of Single Particles: Planetary Systems and Binary Stars

5.1.1 *Newton's Law of Conservation of Momentum*

All Newtonian mechanics begins with the three laws of motion stated by Sir Isaac Newton nearly 350 years ago. Roughly translated and paraphrased from his Latin writings, these are [299]:

1. An object will remain in its present state of motion (*i.e.*, velocity), in the same direction, unless an external and unbalanced force causes that state to be changed.
2. The rate of change of motion of an object is proportional to the net force acting on that object and is in the same direction as that force is applied.

¹ For additional information on these topics, the reader may want to consult [300, 301, 302].

3. To every action there is always an opposite and equal reaction: that is, all forces occur in pairs, and these are equal in magnitude but opposite in direction.

Not only did these laws lay down the fundamentals of how bodies move, they also established that they are *vector* laws, not scalar ones. The *direction* of the applied force, and of the *rate of change* of velocity, are as important as the magnitudes of these quantities.

All three of these laws can be stated simply in a single equation: the conservation of momentum. If we define the momentum vector of a body to be the product of that body's (scalar) mass m and its (vector) velocity $\mathbf{V}$

$$\mathbf{P} \equiv m \mathbf{V} \quad (5.1)$$

then the rate of change of $\mathbf{P}$ is proportional to the (vector) force applied to that body

$$\boxed{\frac{d\mathbf{P}}{dt} = \mathbf{F}} \quad (5.2)$$

Note that the proportionality constant in equation (5.2) is unity, and this is no accident of nature. Physicists have decided to *measure* forces in units of mass times distance, divided by time squared (e.g., dyne = g cm s⁻²), so we do not have to include another constant factor in equation (5.2).

We can recover Newton's first two laws by assuming that the mass of the body does not change with time ($dm/dt = 0$), so

$$m \frac{d\mathbf{V}}{dt} = \mathbf{F}$$

which is the famous $\mathbf{F} = m\mathbf{a}$ of Newton's second law, with $\mathbf{a} = d\mathbf{V}/dt$ as the (vector) acceleration. A special case occurs when there is no force applied ($\mathbf{F} = 0$); this gives us $\mathbf{V} = \text{constant}$, which is Newton's first law.

Newton's third law is more fundamental, and more sweeping. It states that any time *two* bodies interact, they *each* exert a force on the other, and of equal magnitude but opposite direction. So we always need to consider the *system* encompassing *both* bodies, not just the force of one acting on the other. One example of this that is often quoted is rocket propulsion. We actually have two "bodies": the rocket itself (object #1) and the fuel that is ejected from the engine (object #2). The equations (5.2) for each object are

$$\frac{d\mathbf{P}_1}{dt} = \mathbf{F}_1 \quad \frac{d\mathbf{P}_2}{dt} = \mathbf{F}_2$$

Now, these bodies "interact" when burnt fuel is ejected at high velocity $\mathbf{V}_2$ from the rocket motor. If there are no other external forces, then $\mathbf{F}_1 = -\mathbf{F}_2$, so

$$\frac{d\mathbf{P}_1}{dt} = -\frac{d\mathbf{P}_2}{dt} \quad (5.3)$$

Total momentum ($\mathbf{P} \equiv \mathbf{P}_1 + \mathbf{P}_2$) of the rocket, plus fuel, is conserved (because there are no other external forces). However, the rocket m_1 will have a reaction (increase in its momentum $d\mathbf{P}_1/dt$) to the action of accelerating the rocket fuel out the back of the rocket motor nozzle ($d\mathbf{P}_2/dt$).

5.1.2 Newton's Law of Gravity

Newton's theory of gravitation is an extension of his laws to explain Kepler's kinematic laws of planetary (and lunar) motion. Newton stated that two bodies m_1 and m_2 still can interact even though they lie at a considerable distance r from each other. That *attractive* force of gravity on one object, say object #2, is proportional to the *product* of their two masses, *inversely* proportional to the square of the distance between them, *and* in a direction from that mass toward the other body (m_1)

$$\mathbf{F}_{\text{GRAV},2} = -\frac{G m_1 m_2}{r^2} \mathbf{e}_{\hat{r}2} \quad (5.4)$$

where

$$\mathbf{e}_{\hat{r}2} = \frac{\mathbf{r}_2 - \mathbf{r}_1}{r}$$

is the unit vector from m_1 to m_2 , and $r \equiv |\mathbf{r}_2 - \mathbf{r}_1|$ is the distance between them. Newton's law of gravity also satisfies his third law of motion, since the opposite force (on m_1) is

$$\mathbf{F}_{\text{GRAV},1} = -\frac{G m_1 m_2}{r^2} \mathbf{e}_{\hat{r}1} \quad (5.5)$$

since $\mathbf{e}_{\hat{r}1} = -\mathbf{e}_{\hat{r}2}$. Note that, if we want to measure mass in grams (or kilograms), distance in centimeters (or meters), and time in seconds, we *must* have a proportionality constant that is not unity²

$$G = 6.6738 \times 10^{-8} \frac{\text{cm}^3}{\text{g s}^2}$$

² In studies of general relativity it is customary to measure both mass and time in units of length: one centimeter is the time it takes light to travel $1 \text{ cm} = 3.336 \times 10^{-11} \text{ s}$ or 33.4 ps – and one centimeter is also the mass of a non-rotating black hole with a (half-) radius of $GM/c^2 = 1 \text{ cm} = 1.347 \times 10^{28} \text{ g}$ or about twice the mass of the earth. This allows us to set $G = 1$ and the speed of light also will be unity (one centimeter per centimeter). In this book we have chosen *not* to use these units and, instead, be more careful in carrying forward factors of the fundamental constants G and c . The reader then can choose what system of units he/she would like to use.

5.1.3 General Binary Star Motion in the Center-of-Mass Frame of Reference

How we perceive the motion of two gravitating bodies depends on what frame of reference we choose. For example, if we choose our origin at the center of one of the bodies, say m_1 , and if the mass of m_2 does not change with time, then equation (5.4) becomes the equation of motion for m_2 orbiting about m_1

$$\frac{d\mathbf{V}_2}{dt} = -\frac{G m_1}{r^2} \mathbf{e}_{\hat{r}_2} \quad (5.6)$$

On the other hand, if we choose to observe from the point of view of mass m_2 , we obtain a similar but not identical equation

$$\frac{d\mathbf{V}_1}{dt} = -\frac{G m_2}{r^2} \mathbf{e}_{\hat{r}_1}$$

Neither one of these frames of reference, however, is an *inertial* one, because in each the observer experiences an acceleration caused by the other body. Non-inertial frames of reference are perfectly valid to use, but one must take into account the fact that acceleration of the observer also is taking place, and that complicates the calculations.

To obtain an inertial frame of reference, we could choose a random point in space that is far enough from the two masses that the observer would not be affected by their gravitational effects. However, there is an even better inertial system. Because the forces on m_1 and m_2 are equal and opposite, there is no force on the combined system

$$\frac{d(\mathbf{P}_1 + \mathbf{P}_2)}{dt} = \frac{d(m_1 \mathbf{V}_1 + m_2 \mathbf{V}_2)}{dt} = 0$$

(from equation (5.3)). So a frame of reference with the observer (and the origin of the coordinate system) at the center of mass of the system does not experience a net acceleration. It travels at a constant velocity

$$\mathbf{V} = \frac{m_1 \mathbf{V}_1 + m_2 \mathbf{V}_2}{M} \quad (5.7)$$

where the total binary mass is

$$M = m_1 + m_2 \quad (5.8)$$

and the rate of change of the momentum of the center of mass is zero

$$0 = \frac{d\mathbf{P}}{dt} = \frac{d(M \mathbf{V})}{dt} = M \frac{d\mathbf{V}}{dt} \quad (5.9)$$

The position of the center of mass of the binary is given by

$$\mathbf{R} = \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2}{M}$$

and this allows us to define new position and velocity vectors

$$\mathbf{r}'_1 = \mathbf{r}_1 - \mathbf{R} = \frac{m_2}{M} (\mathbf{r}_1 - \mathbf{r}_2) \quad (5.10)$$

$$\mathbf{r}'_2 = \mathbf{r}_2 - \mathbf{R} = \frac{m_1}{M} (\mathbf{r}_2 - \mathbf{r}_1) \quad (5.11)$$

$$\mathbf{V}'_1 = \mathbf{V}_1 - \mathbf{V} = \frac{d\mathbf{r}'_1}{dt} \quad (5.12)$$

$$\mathbf{V}'_2 = \mathbf{V}_2 - \mathbf{V} = \frac{d\mathbf{r}'_2}{dt} \quad (5.13)$$

From equations (5.10) and (5.11) we see that at every moment in time, the ratio of the distance of the two objects from the center of mass is simply

$$\boxed{\frac{r'_1}{r'_2} = \frac{m_2}{m_1}} \quad (5.14)$$

This turns out to be a very important and fundamental property of binary star motion.

From equations (5.7) to (5.13) we can form *new* equations of motion for a binary system that *are* in an inertial reference frame

$$\frac{d\mathbf{V}'_1}{dt} = -\frac{G M}{r^2} \frac{\mathbf{r}'_1}{r} \quad (5.15)$$

$$\frac{d\mathbf{V}'_2}{dt} = -\frac{G M}{r^2} \frac{\mathbf{r}'_2}{r} \quad (5.16)$$

These show that each mass orbits around the center of mass of the system with a force still proportional to the inverse square of the distance between them. But they lie at fractional distances $|\mathbf{r}'_1|/r = m_2/M$ and $|\mathbf{r}'_2|/r = m_1/M$ from that center of mass.

5.1.4 Motion of a Test Particle in a Gravitational Field

A simple application of the above binary star equations is the orbit of a test particle m_2 , where $m_2 \ll m_1$. This is simpler than the orbit of two comparable masses and often occurs in astrophysics (planetary orbits, stars about the Galaxy, and even galaxies in clusters of galaxies). When m_1 dominates, we have $M \approx m_1$, $|\mathbf{r}'_1| \approx 0$, $|\mathbf{r}'_2| \approx r$, and $\mathbf{V}'_1 \approx 0$. And we need to be concerned only with the motion of the test mass m_2 in the gravitational field of m_1 (equation (5.16)).

5.1.4.1 Free Fall of a Test Particle

Free fall (motion only along $\mathbf{r}'_2$) toward the larger mass m_1 is the easiest problem to solve first. With

$$\mathbf{V}'_2 = \frac{dr'_2}{dt} \mathbf{e}_{\hat{r}'_2} = \frac{dr'_2}{dt} \frac{\mathbf{r}'_2}{r_2}$$

the motion of the test particle becomes a simple scalar equation for the distance $r'_2 \approx r$

$$\frac{dV'_2}{dt} = -\frac{GM}{r^2} \quad (5.17)$$

We will see this kind of equation several times, so it is useful to briefly go through the standard way of solving it. There are two steps. First, the time derivative is replaced by a spatial derivative and then integrated over r :

$$\frac{d}{dt} = \frac{dr'_2}{dt} \frac{d}{dr'_2} = V'_2 \frac{d}{dr}$$

Equation (5.17) then becomes $d[(V'_2)^2/2]/dr = -GM/r^2$, which has the solution

$$V'_2 = -\left(\frac{2GM}{r_0}\right)^{1/2} \left(\frac{r_0}{r} - 1\right)^{1/2} \quad (5.18)$$

where $r_0 \equiv r(t = 0)$, and we have assumed that the initial velocity is zero ($V'_2(t = 0) = 0$). This first step, then, gives us V'_2 as a function of r . The second step, integration of equation (5.18), will give us r as a function of time t – actually each as a function of a third parameter called the “collapse parameter” v (*upsilon*).³ Setting V'_2 in equation (5.18) equal to dr/dt produces a simple integral over dt and a transcendental integral over dr . The latter is most conveniently performed by making a variable substitution $r = r_0 \cos^2(v/2)$ or

$$r(v) = \frac{r_0}{2} (1 + \cos v) \quad (5.19)$$

which actually tells how r varies with the collapse parameter. The integrals over t and v of equation (5.18), then, give us an expression for how *time* varies as a function of v

$$t(v) = \frac{\tau_{\text{ff}}}{\pi} (v + \sin v) \quad (5.20)$$

where

³ The parametric approach to solving gravitational collapse is very useful. We shall see it again in Chapter 8 when we discuss relativistic collapse to a black hole.

$$\tau_{\text{ff}} \equiv \left(\frac{\pi^2 r_0^3}{8 G M} \right)^{1/2} \quad (5.21)$$

is the “free-fall” time – the time it takes to fall from $r = r_0$ ($v = 0$) to $r = 0$ ($v = \pi$) in equation (5.19). The solution for the velocity also can be parameterized

$$V_2'(v) = - \left(\frac{2 G M}{r_0} \right)^{1/2} \tan \left(\frac{v}{2} \right) \quad (5.22)$$

So, as $t \rightarrow \tau_{\text{ff}}$, $V_2' \rightarrow \infty$. This is a consequence of assuming point masses, however. Speeds never truly reach infinite velocity: in free fall toward a planet or a star, the impact occurs at a finite velocity at the object’s surface, and in free fall toward a black hole the velocity approaches the speed of light (equation (7.36) and following).

For the even simpler case of free fall from rest at infinity ($r_0 \gg r$), equation (5.18) becomes the standard expression for the free fall (and escape) velocity in Newtonian gravity

$$V_{\text{ff,NEWT}} = -V_{\text{esc,NEWT}} = - \left(\frac{2 G M}{r} \right)^{1/2} \quad (5.23)$$

5.1.4.2 Circular Orbit of a Test Particle

Circular orbital motion is another easy problem, and will be the last test particle problem that we will consider in Newtonian mechanics. We will use the same equation of motion (5.16) for the test mass m_2 , but this time we will assume a circular orbit, such that $r_2' = r$ is a constant in time. The *vector* $\mathbf{r}_2'$ is then given by

$$\mathbf{r}_2' = r \mathbf{e}_{\hat{r}'2}$$

where, again, $\mathbf{e}_{\hat{r}'2}$ is the unit vector from mass m_1 ($= M$) to m_2 . Since r is constant for this circular orbit, however, the only part of $\mathbf{r}_2'$ that can vary is $\mathbf{e}_{\hat{r}'2}$, so equation (5.16) becomes

$$r \frac{d^2 \mathbf{e}_{\hat{r}'2}}{dt^2} = - \frac{G M}{r^2} \mathbf{e}_{\hat{r}'2} \quad (5.24)$$

The easiest way to compute the left-hand side of equation (5.24) is to write $\mathbf{e}_{\hat{r}'2}$ in terms of Cartesian unit vectors that do not vary with time, but whose coefficients may:

$$\mathbf{e}_{\hat{r}'2} = \cos \phi \mathbf{e}_{\hat{x}'2} + \sin \phi \mathbf{e}_{\hat{y}'2} \quad (5.25)$$

This assumes that the circular orbital motion is in the x - y plane, that it is about the z axis, and that the “orbital phase angle” ϕ can vary with time but only linearly

$$\frac{d\phi}{dt} = \Omega$$

where Ω is a constant ($d\Omega/dt = 0$). From equation (5.25) the second time derivative of $\mathbf{e}_{\hat{r}'2}$ on the left-hand side of equation (5.24) becomes $-\Omega^2 \mathbf{e}_{\hat{r}'2}$, reducing the conservation of momentum to simply the circular Keplerian orbit equation

$$\Omega = \Omega_K \equiv \left(\frac{GM}{r^3} \right)^{1/2} \quad (5.26)$$

where Ω_K is called the Keplerian orbital angular velocity. We also can calculate the magnitude of the constant *linear* orbital velocity

$$|V'_2| = r \left| \frac{d\mathbf{e}_{\hat{r}'2}}{dt} \right| = r \Omega$$

so

$$V_{\text{orb,NEWT}} = \left(\frac{GM}{r} \right)^{1/2} \quad (5.27)$$

We, therefore, have derived the well-known result that at any radius r , the free fall and orbital speed in Newtonian gravity are related by a factor of $\sqrt{2}$. (Compare equations (5.23) and (5.27).)

5.1.5 Motion of Two Comparable Masses in Each Other's Gravitational Field: Orbits of Binary Planets and Stars

We now are in a position to look at *binary* systems, where both objects have significant mass. This is a much easier problem, now that we have investigated the motion of test masses in a gravitational field. We again must use the center-of-mass equations (5.15) and (5.16) to describe the motion; and equation (5.14) will tell us the ratio between the distances of the two objects from that center of mass.

5.1.5.1 Circular Orbits in a Binary System

As above, we will consider orbits only where the separation between the two stars or planets stays fixed. Because r will be constant, and because the right-hand side of equation (5.14) is constant also, both r'_1 and r'_2 also will be constant as well. So,

each object orbits the center of mass in its own circular orbit, with

$$r'_1 = a_1 \quad r'_2 = a_2 \quad r = a$$

where the primary a_1 and secondary a_2 orbital radii are related by

$$a_2 = \frac{m_1}{m_2} a_1 \quad (5.28)$$

and the total orbital radius is given by

$$a = a_1 + a_2 \quad (5.29)$$

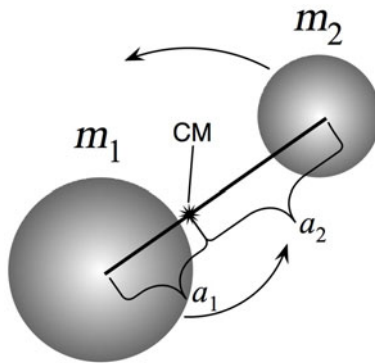


Fig. 5.1: Diagram of an unequal mass binary star system in a circular orbit, viewed from the system's north pole. As in the text, $m_1 > m_2$ is assumed. Each star orbits around the system center of mass, which is closer to the more massive star ($a_1 < a_2$).

This time both objects' vectors from the center of mass are given by the (rotating) unit vectors (see Fig. 5.1)

$$\mathbf{r}'_1 = a_1 \mathbf{e}_{\hat{r}'1} \quad \mathbf{r}'_2 = a_2 \mathbf{e}_{\hat{r}'2}$$

And, assuming a coordinate system with the z axis normal to the orbital plane, both above unit vectors can be written in Cartesian coordinates with orbital phase angles ϕ_1 and ϕ_2 . Because the two stars remain diametrically opposed,

$$\phi_2 = \phi_1 + \pi$$

both unit vectors must rotate at the same angular frequency (*i.e.*, $d\phi_1/dt = d\phi_2/dt$). And equations (5.15) and (5.16) both show that this is indeed the case

$$a_1 \Omega_1^2 = \frac{G M}{a^3} a_1$$

$$a_2 \Omega_2^2 = \frac{G M}{a^3} a_2$$

So, the angular frequency for a binary orbit is simply the Keplerian frequency, with the *sums* of the properties of the individual masses (*i.e.*, M and a) used to compute it

$$\Omega_{\text{bin}} = \left(\frac{G M}{a^3} \right)^{1/2} \quad (5.30)$$

This simple equation allows us to derive the “mass function” equation for X-ray binaries (1.6) that was quoted in the beginning of the book. Suppose that we know the circular velocity of one of the stars, say $V_1 = a_1 \Omega$. In terms of the binary period P , we have

$$V_1 = \frac{2\pi a_1}{P} = \frac{2\pi a m_2}{P M}$$

from equations (5.28) and (5.29). Cubing this equation and combining it with the square of equation (5.30) gives us

$$\frac{P V_1^3}{2\pi G} = \frac{m_2^3}{M^2} \quad (5.31)$$

This is the theoretical period-velocity-mass relation for binary stars, and a similar one exists for V_2 . In order to make it more useful for observational study, we identify the companion and X-ray star as $m_1 = M_c$ and $m_2 = M_x$, respectively, measure the radial velocity of m_1 with optical line observations, and realize that the measured velocity $V_{\text{rad}} = V_1 \sin i$, where i is the inclination angle of the binary relative to the plane of the sky. Then the above equation becomes

$$\frac{P V_{\text{rad}}^3}{2\pi G} = \frac{M_x^3}{(M_x + M_c)^2} \sin^3 i$$

which is equation (1.6). The left-hand side includes only information that we know about the binary, while the right-hand side contains the unknowns. If we estimate M_c based on its spectral type and independently determine i (say, 90° for an eclipsing binary), then we can uniquely determine M_x .

5.1.5.2 General Orbits in a Binary System

Because a binary system can be described so well without referring to the individual masses m_1 and m_2 specifically, it often is useful to cast the equations entirely without them. Instead, one discusses a binary system in terms of its total mass M

and its *reduced mass*

$$\mu \equiv \frac{m_1 m_2}{M} = \frac{a_1 a_2}{a^2} M \quad (5.32)$$

which can be thought of as the “secondary” mass counterpart to total mass M . The magnitude of the gravitational force holding the binary together, then, can be written as

$$F_{\text{GRAV}} = -\frac{G M \mu}{r^2} \quad (5.33)$$

Furthermore, because equation (5.14) holds for any (elliptical, parabolic, *etc.*) orbit, then we always will have

$$\frac{r'_1}{a_1} = \frac{r'_2}{a_2} = \frac{r}{a}$$

where a_1 and a_2 now are semi-major axes of the orbit. Therefore, equations (5.15) and (5.16) can be subtracted (vectorially) to form a single conservation of momentum for the reduced mass μ about total mass M

$$\mu \frac{d^2 \mathbf{r}}{dt^2} = -\frac{G M \mu}{r^2} \mathbf{e}_{\hat{r}} \quad (5.34)$$

In other words, the reduced mass can be considered to be a fictitious test mass, lying at a distance r from the total mass M , which in turn, lies at the center of mass. *The binary system in the μ – M notation acts like a simple Keplerian orbit, with the equations now written in the inertial center-of-mass frame of reference.* Further analysis (*e.g.*, elliptical, parabolic, and hyperbolic orbits) of binary systems now can proceed using the same techniques as one uses for test masses in a gravitational field.

5.2 Mechanics of a System of Particles: Fluid Dynamics and the Internal Structure of Stars

Newton’s basic physical laws that describe the motions of planets and stars also can be used to described those objects’ behavior – from the weather in their atmospheres and oceans to the nuclear burning and element creation in their centers. The only real change in the treatment of these problems is that we write the equations in terms of the mass *density* of the material that composes them

$$\rho = \frac{\delta m}{\delta \mathcal{V}} \quad (5.35)$$

rather than in treating the masses of each individual particle that makes up that matter. Here $\delta\mathcal{V}$ is a small volume of the material where conditions (*e.g.*, velocity, temperature) are essentially the same, and δm is the total mass inside that small volume.

Even though we envision $\delta\mathcal{V}$ to be small (*e.g.*, a cubic centimeter of air or water on the earth, a cubic meter of the sun's interior, or a cubic earth radius of the interstellar medium), it nevertheless will contain an extraordinarily large number of particles – often Avogadro's number ($N_A = 6.02214 \times 10^{23}$) or more. Even with today's computers, it would be impossible to follow every particle's motions. So we instead follow the motion of thousands to millions of small fluid elements, whose properties can change with position in space as well as with time.

Our basic physical variables, therefore, are the density, velocity, and *internal* energy of each δm , and these are assumed to vary smoothly in space and time

$$\begin{aligned}\rho &= \rho(\mathbf{r}, t) \\ \mathbf{V} &= \mathbf{V}(\mathbf{r}, t) \\ \varepsilon &= \varepsilon(\mathbf{r}, t)\end{aligned}$$

The internal energy is a measure of the motions (temperature) of the local fluid particles and will be important for describing the thermodynamics of that material. The solution to a particular physical or astrophysical fluid problem, then, is considered to be the determination of how the above quantities vary in time and space. This is done by solving the conservation laws for a fluid either analytically, semi-analytically, or numerically using large simulation computer programs. Below we derive those equations by first considering Newton's laws for a typical fluid element and then either dividing by δm and/or $\delta\mathcal{V}$ or by allowing these small quantities to go to zero, giving us partial differential equations that govern the motion of fluids.

5.2.1 The Gravitational Force Produced by a Fluid – Inside and Out

As before, the gravitational force produced by a single small mass $\delta m'$ on another fluid element δm is governed by an inverse square law

$$(\delta m) \delta \mathbf{a}' = - \frac{G (\delta m) (\delta m')}{|\mathbf{r} - \mathbf{r}'|^2} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (5.36)$$

Now, because the force changes only with separation, it is very useful to write that force on δm as being derived from the gravitational potential energy (or, simply, “potential”) $\delta\psi'$ produced by fluid element $\delta m'$

$$(\delta m) \delta \mathbf{a}' = -\delta m \nabla(\delta\psi') \quad (5.37)$$

where $\delta\mathbf{a}'$ is the acceleration of δm by $\delta m'$,

$$\delta\psi' = -\frac{G(\delta m')}{|\mathbf{r} - \mathbf{r}'|} \quad (5.38)$$

$\mathbf{r}$ and $\mathbf{r}'$ are the vector locations of masses δm and $\delta m'$, respectively, and the operator ∇ is the vector gradient

$$\nabla = \mathbf{e}_{\hat{x}} \frac{\partial}{\partial x} + \mathbf{e}_{\hat{y}} \frac{\partial}{\partial y} + \mathbf{e}_{\hat{z}} \frac{\partial}{\partial z} \quad (5.39)$$

As is customary, the potential energy is defined to be zero as $|\mathbf{r} - \mathbf{r}'| \rightarrow \infty$, and to be negative as one approaches closer to the mass $\delta m'$. Equation (5.37) is identical to equation (5.36) in both magnitude and direction. We therefore can study the gravitational field by concentrating on its potential (a scalar field) rather than on the force (a more complex vector field).

The *total* gravitational force on the small δm mass, then, can be determined by summing all the $\delta\mathbf{a}'$ components to obtain a total acceleration $\mathbf{a}'$ and by summing the $\delta\psi'$ components to obtain a total potential ψ'

$$\delta\mathbf{F}_{\text{GRAV}} = (\delta m)\mathbf{a}' = -(\delta m)\nabla\psi' = -\rho\nabla\psi'(\delta\mathcal{V}) \quad (5.40)$$

As $\delta m \rightarrow 0$, the integral for the total gravitational potential is, then,

$$\psi' = \int d\psi' = -G \int \frac{dm'}{|\mathbf{r} - \mathbf{r}'|}$$

or

$$\boxed{\psi' = -G \int_{\mathcal{V}'} \frac{\rho'}{|\mathbf{r} - \mathbf{r}'|} d\mathcal{V}'} \quad (5.41)$$

where $\mathcal{V}'$ is the entire volume containing fluid that is interacting gravitationally with δm . Of course, we have not specified *where* δm is located, so it could be inside or outside $\mathcal{V}'$. Therefore, equation (5.41) describes the gravitational potential *everywhere*, inside or outside the matter that has the density distribution $\rho(\mathbf{x}', t)$. We have placed no restrictions on the density, other than it occupy a finite (not infinite) volume $\delta\mathcal{V}'$. So this equation can be used for complex mass distributions like flattened spiral galaxies, asymmetrical exploding stars, *etc.*

In 1813 French mathematician and physicist Siméon-Denis Poisson showed that the integral for ψ in equations like (5.41) actually is a general solution to the second-order differential equation

$$\boxed{\nabla^2\psi = 4\pi G\rho} \quad (5.42)$$

which is now called “Poisson’s equation” in his honor. And the form in equation (5.41) is called a “Green’s function” general solution to the Poisson equation, after

the English physicist and mathematician George Green who studied the solution of inhomogeneous partial differential potential equations some two decades after Poisson's work. In Cartesian and spherical-polar coordinates, the scalar "Laplacian" operator is

$$\begin{aligned}\nabla^2 &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \\ \nabla^2 &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial}{\partial \theta} \left(\sin^2 \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2}\end{aligned}$$

Poisson's equation and the use of Green's functions are enormously important, not only because they describe gravity in the interiors of planets, stars, and galaxies⁴, but also because they will lead us to the Einstein field equations for gravity in extreme conditions, and to gravitational waves.

Note that equation (5.42) says that, "the *local* second-order derivative of the gravitational potential is proportion to the *local* mass density". Gravity, however, is action at a distance; so we will not know the gravitational force locally, due to all matter in a given region of space, until we have solved Poisson's equations for $\psi(\mathbf{r})$ over that space and beyond, with boundary conditions, *and* until we have computed the gravitational acceleration $\mathbf{a} = -\nabla\psi$ from that solution.

Because equation (5.42) is valid everywhere, it also is valid outside the matter, where $\rho = 0$

$$\nabla^2\psi = 0 \tag{5.43}$$

This is Laplace's equation for the potential of an inverse square law force when there are no sources locally. However, ψ is not necessarily zero simply because there is no matter to generate more gravitational field. For example, outside a spherical planet or star ψ varies only with distance r from the center of the object, or

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) = 0$$

which has the solution $\psi \propto -1/r$. The boundary condition at the surface $r = R$ of the object ($\psi = -GM/R$) is what gives us the familiar solution for ψ for all space *outside* the object

$$\psi = -\frac{GM}{r}$$

⁴ Not to mention electrostatic fields of charged matter.

5.2.2 Newtonian Conservation Laws for a Fluid in the Lagrangian Frame

We now will derive the basic equations of fluid dynamics, which will turn out to be important in the discussion of Einstein's theory of gravity. In addition to the conservation of momentum, however, we also must add the equations for conservation of mass and of internal energy. The former will give us an equation for how the density ρ evolves, and the latter will allow us to determine the fluid pressure, which is a new force that affects the flow of fluids and gases, in addition to gravity.

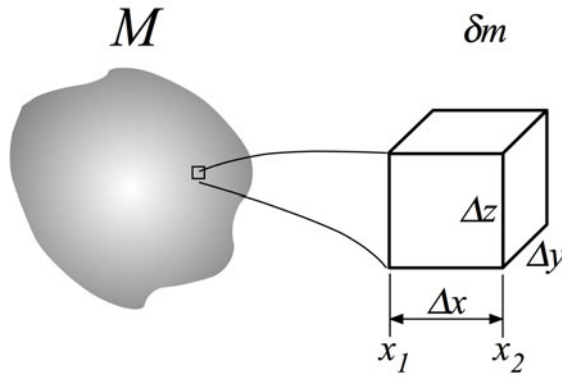


Fig. 5.2: In the case of a fluid, instead of two masses we have multiple small mass elements δm that together make up the total mass M . Depending on the circumstances, we can require the volume $\Delta x \Delta y \Delta z$ to always enclose the same material (Lagrangian fluid dynamics), which may require the volume element to vary in size as the local density varies. *Or* the volume element can remain fixed in size and fixed in space, but contain fluid of differing density (and therefore mass) at different times (Eulerian fluid dynamics).

5.2.2.1 Conservation of Mass for a Fluid

In deriving the fluid equations it is useful to consider the small volume δV to be a simple three-dimensional box in Cartesian coordinates with lower corner at (x_1, y_1, z_1) and upper corner at (x_2, y_2, z_2) (see Fig. 5.2)

$$\delta V = \Delta x \Delta y \Delta z$$

where $\Delta x = x_2 - x_1$, *etc.* describe the lateral dimensions of the box. The time derivative of this small volume will be

$$\frac{d(\delta V)}{dt} = \frac{d(\Delta x)}{dt} \Delta y \Delta z + \Delta x \frac{d(\Delta y)}{dt} \Delta z + \Delta x \Delta y \frac{d(\Delta z)}{dt} \quad (5.44)$$

Now, in order to compute mass conservation, we must not let any matter enter or escape the box, so we must have

$$\frac{d(\delta m)}{dt} = 0$$

If we allow the sides of the box to “go with the flow”, so the amount of mass inside the box is kept fixed, how does this change the box volume? Defining the speeds of the box sides to be

$$V_{x1} \equiv \frac{dx_1}{dt} \quad V_{y1} \equiv \frac{dy_1}{dt} \quad \dots$$

equation (5.44) can be re-written as

$$\frac{d(\delta \mathcal{V})}{dt} = \delta \mathcal{V} \left[\frac{1}{\Delta x} \Delta(V_x) + \frac{1}{\Delta y} \Delta(V_y) + \frac{1}{\Delta z} \Delta(V_z) \right]$$

So, as we consider smaller and smaller boxes of fluid ($\Delta x \rightarrow 0, \dots$), we can write the above as a differential equation

$$\frac{d(\delta \mathcal{V})}{dt} = \delta \mathcal{V} (\nabla \cdot \mathbf{V}) \quad (5.45)$$

That is, in order to keep the mass inside $\delta \mathcal{V}$ the same, its fractional volume must be increased or decreased with time at a rate equal to the divergence of the velocity of that fluid.⁵

Inserting $\delta \mathcal{V} = \delta m / \rho$ from equation (5.35) into equation (5.45), and dividing by (the now constant δm), we obtain an equation for the evolution of the density of a fluid or gas

$$\boxed{\frac{d\rho}{dt} = -\rho (\nabla \cdot \mathbf{V})} \quad (5.46)$$

This is the fundamental conservation of mass or “continuity” equation for fluid dynamics. It must be remembered, of course, that it holds true only if matter in each element of fluid is neither created nor destroyed.

Equation (5.46) was derived using a simple Cartesian coordinate system. However, it is equally valid in spherical-polar and any other three-dimensional coordinate system. As is the case with such derivations, if the result can be cast into a coordinate-free form, it will be valid in any system. Also note that the frame of reference that we have chosen is one that moves with the fluid element in question. This is called the “Lagrangian fluid frame”, after the French mathematician, physicist, and astronomer Joseph-Louis Lagrange, who studied the mechanics of fluids and solids, among many other things, in the mid-late 1700s. This frame is very use-

⁵ In some situations, the fluid will be very stiff and unable to be compressed or expanded easily. These are referred to as incompressible fluids, and $\nabla \cdot \mathbf{V} = 0$ can be assumed.

ful for considering local physics of the fluid, and any type of one-dimensional flow (where fluid elements do not exchange places, mix, swirl, or become turbulent.) However, this frame is less useful for computing the global evolution of a complex three-dimensional fluid body.

5.2.2.2 Conservation of Momentum for a Fluid

Using similar thinking, we now can derive easily the conservation of momentum for fluids. We start with Newton's law applied to each fluid element

$$\frac{d(\delta m \mathbf{V})}{dt} = \delta \mathbf{F} = \delta \mathbf{F}_{\text{PRESS}} + \delta \mathbf{F}_{\text{GRAV}} \quad (5.47)$$

There are now *two* external forces that act on each fluid element: gravity and pressure. Pressure is the force, per unit area, that is exerted on each element by its *neighboring* six fluid elements, and is due *solely to the internal motions* of the particles in those fluid elements. It is not due to the dynamical motion of nearby elements toward or away from $\delta \mathcal{V}$. (That action will be taken into account automatically, as we shall see.)

Since we now are dealing with a total of six different element neighbors (whose conditions all can be different), the pressure at x_2 can be different from that at x_1 , and so on. Therefore, the *net* force on δm in the x direction is

$$\begin{aligned} \delta F_{\text{PRESS},x} &= -[p(x_2) \Delta y \Delta z - p(x_1) \Delta y \Delta z] \\ &= -\delta \mathcal{V} \frac{\Delta p}{\Delta x} \end{aligned}$$

with similar expressions for the y and z net forces. As $\delta \mathcal{V} \rightarrow 0$, we can write the vector force as $\delta \mathbf{F}_{\text{PRESS}} = -(\delta \mathcal{V}) \nabla p$. So, with equation (5.40), the conservation of momentum (equation (5.47)) then becomes

$$(\delta \mathcal{V}) \frac{d(\rho \mathbf{V})}{dt} + \rho \mathbf{V} \frac{d(\delta \mathcal{V})}{dt} = -(\delta \mathcal{V}) \nabla p - (\delta \mathcal{V}) \rho \nabla \psi$$

And we can get rid of $\delta \mathcal{V}$ everywhere by folding in equation (5.45) to get the final version for the conservation of momentum

$$\boxed{\frac{d(\rho \mathbf{V})}{dt} + \rho \mathbf{V} (\nabla \cdot \mathbf{V}) = -\nabla p - \rho \nabla \psi} \quad (5.48)$$

or, if we combine this with equation (5.46), we obtain the simple $m\mathbf{a} = \mathbf{F}$ form

$$\rho \frac{d\mathbf{V}}{dt} = -\nabla p - \rho \nabla \psi$$

5.2.2.3 Conservation of Energy for a Fluid

Equation (5.48) depends on ρ , V , p , and ψ . It is *the* equation for V , and we have equations (5.46) and (5.42) for ρ and ψ , respectively. We therefore still need an equation for the thermodynamic quantity p .

The First Law of Thermodynamics and the Full Energy Equation

In reality, there are actually *three* equations that determine p : two equations of state that relate pressure and internal energy to variables like density and temperature T

$$p = p(\rho, T) \quad \varepsilon = \varepsilon(\rho, T)$$

and an evolution equation for ε . The latter is determined from the first law of thermodynamics, applied to an element of fluid

$$d(\delta Q) = p d(\delta V) + d(\delta E) - \mu_{\text{chem}} d(\delta \mathcal{N})$$

where $d(\delta Q)$ is the heating (+) or cooling (−) that occurs in the fluid element; $p d(\delta V)$ is the work done by (+) or on (−) the fluid element; $d(\delta E)$ is the change in internal energy of the fluid element; μ_{chem} is the chemical potential of each particle that makes up the fluid; and $d(\delta \mathcal{N})$ is the change in the number of particles enclosed in the volume δV . For thermal gases, where a single temperature T can be defined, the change in heat content can be expressed as a change in entropy S

$$d(\delta Q) = T d(\delta S)$$

Note that the above version of the first law includes particle creation and annihilation processes that occur in chemical and nuclear reactions. However, in the majority of cases in this book, we shall assume that the number of particles is conserved – *i.e.*, that $d(\delta \mathcal{N}) = 0$, reducing the first law to the more familiar form

$$d(\delta Q) = p d(\delta V) + d(\delta E) \tag{5.49}$$

Since δQ and δE are in units of energy, not energy per unit volume, we can define

$$\delta Q \equiv \delta m q \quad \delta E \equiv \delta V \varepsilon$$

where q is the heating/cooling per unit *mass* of fluid, and ε is the internal energy per unit *volume*, whose evolution equation we are currently deriving. Taking the differentials in equation (5.49) to be time differentials, the first law becomes

$$\rho(\delta V) \frac{dq}{dt} = (p + \varepsilon) \frac{d(\delta V)}{dt} + \delta V \frac{d\varepsilon}{dt}$$

and, with equation (5.45), we can get rid of the $\delta\mathcal{V}$ factor to obtain the conservation of energy law

$$\boxed{\frac{d\varepsilon}{dt} = -(p + \varepsilon) \nabla \cdot \mathbf{V} + \rho \frac{dq}{dt}} \quad (5.50)$$

The term $\rho dq/dt$ is often referred to as the “volume heating” (or “volume emissivity”, if negative); however, it is *not* the time derivative of the heat per unit volume ($d(\rho q)/dt$). It is, instead, the time derivative of the heat *per unit mass*, multiplied by the mass density per unit volume (*i.e.*, the time derivative of ρq *at constant volume*). This distinction must be remembered by those who calculate heating and cooling rates dq/dt for radiative and other processes.

Coefficients of Specific Heat and the Adiabatic Index

To characterize the thermal properties of the fluid or gas, chemists and physicists often use two different specific heat coefficients C_V and C_p , which give the rate of change of the heat $d(\delta Q)$ as the temperature of the gas is changed. The first specific heat is valid when the volume $\delta\mathcal{V}$ is held constant; the second describes the change in heat when the pressure p is held fixed. Let us consider the constant volume case (C_V) first. With $d(\delta\mathcal{V}) = 0$, equation (5.49) becomes

$$d(\delta E_V) = d(\delta Q_V) \equiv \frac{\delta\mathcal{N}}{N_A} C_V dT$$

where $N_A = 6.02214 \times 10^{23} \text{ mol}^{-1}$ is Avogadro’s number. This defines C_V .

Now let us consider the change in heat content when the temperature changes and the *pressure* is held fixed. For this case we will need the perfect gas law $p\delta\mathcal{V} = [(\delta\mathcal{N})/N_A] \mathcal{R} T$, which gives us the differential relation (for constant pressure)

$$p d(\delta\mathcal{V}) = \frac{\delta\mathcal{N}}{N_A} \mathcal{R} dT$$

where $\mathcal{R} = N_A k = 8.31446 \times 10^7 \text{ erg mol}^{-1} \text{ K}^{-1}$ is the universal gas constant and $k = 1.38065 \times 10^{-16} \text{ erg K}^{-1}$ is the Boltzmann constant. The change in internal energy from equation (5.49) is

$$\begin{aligned} d(\delta E_p) &= d(\delta Q_p) - p d(\delta\mathcal{V}) \\ &= \frac{\delta\mathcal{N}}{N_A} (C_p - \mathcal{R}) dT \end{aligned}$$

where we have used the definition $d(\delta Q_p) \equiv [(\delta\mathcal{N})/N_A] C_p dT$.

Now comes an important step in our reasoning. The internal energy of a volume of perfect gas δE_V depends *only on its temperature* $\delta E_V = (3/2) [(\delta\mathcal{N})/N_A] \mathcal{R} T$. So the *change* in internal energy $d(\delta E_V)$ depends *only on the change in the temper-*

ature

$$d(\delta E_V) = \frac{3}{2} \frac{\delta \mathcal{N}}{N_A} \mathcal{R} dT$$

It does not matter, therefore, whether we change the temperature while holding the volume constant or while holding the pressure constant: the change in *internal energy will be the same*. That is, $d(\delta E_V) = d(\delta E_p)$, or

$$\boxed{C_p - C_V = \mathcal{R}} \quad (5.51)$$

So, the two specific heats differ only by a constant: the gas constant $\mathcal{R}$.

The *ratio* of the specific heats

$$\boxed{\Gamma \equiv \frac{C_p}{C_V}} \quad (5.52)$$

is called the “adiabatic index”, and is a particularly important quantity in thermodynamics, because it describes how the pressure and density of a particular type of gas are related when $\delta Q_p = 0$. The two specific heats can be written in terms of Γ and the gas constant as

$$C_p = \frac{\Gamma}{\Gamma - 1} \mathcal{R} \quad C_V = \frac{1}{\Gamma - 1} \mathcal{R}$$

The Adiabatic Energy Equation and the Polytropic Equation of State

While all gases are heated and cooled by radiation, conduction, viscosity, *etc.*, many astrophysical situations are dominated by the work term $(p + \varepsilon) \nabla \cdot \mathbf{V}$, which takes into account the compression and expansion of the gas. In these situations we can ignore dq/dt and work with the *adiabatic* first law of thermodynamics. Combining equation (5.50) and (5.46), this can be written for a fluid as

$$d\varepsilon = \frac{(p + \varepsilon)}{\rho} d\rho \quad (5.53)$$

Now, it often occurs in thermodynamics that the pressure and internal energy are related by a single constant factor n

$$\varepsilon = n p \quad (5.54)$$

(where n does *not* have to be an integer). Examples are a non-relativistic ($n = 3/2$) ideal gas or a relativistic ($n = 3$) ideal gas. Combining equations (5.53) and (5.54), we find that, in an adiabatic environment where there is no heating or cooling, the pressure must follow the relation

$$p = K_n \rho^{\frac{n+1}{n}} = K_\Gamma \rho^\Gamma \quad (5.55)$$

where K_n is a constant that does not vary with density ρ . This is called the “polytropic” equation of state, and, at once, gives us an expression for p to use in the conservation of momentum equation (5.48). The adiabatic index Γ and polytropic index n describe the same adiabatic process and are related in the following manner:

$$\Gamma = \frac{n+1}{n} \quad (5.56)$$

so that $p = K_\Gamma \rho^\Gamma$ is equivalent to equation (5.55).

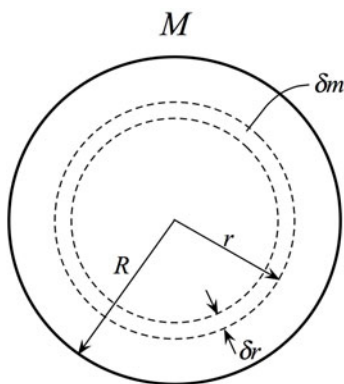


Fig. 5.3: When spherical symmetry is an appropriate assumption for a stellar model, many of the volume elements in Fig. 5.2 can be collected into a single spherical shell, also called δm . The entire star then is composed of concentric spherical shells of radial thickness δr , extending from $r = 0$ to $r = R$. Since shells do not pass through each other, the “mass coordinate” m (eq. (5.57)) can be used instead of r . Labeling shells with m in this manner, and then *computing* r , is a form of Lagrangian fluid dynamics.

5.2.2.4 The Equations of Stellar Structure and Dynamics in Mass Coordinates

Equations (5.46), (5.48), and (5.50) are the conservation laws for mass, momentum, and energy that govern fluid flow in a (Lagrangian) frame of reference that moves with the fluid. A particularly important use of these equations is the study of the interiors of *spherical* stars, both those that are static (with $\mathbf{V} = 0$) and those that are dynamic and potentially explosive. The flow inside such stars is one-dimensional, in the radial direction, so each fluid element is a complete spherical shell of mass δm and thickness δr , rather than simply a small box (see Fig. 5.3). A particularly clever and useful change of coordinates is to use the *mass interior to radius* r

$$m = \int_0^r 4\pi r^2 \rho dr \quad (5.57)$$

as the primary *independent* coordinate, rather than r itself. For this choice of coordinate the conservation mass ($\delta m = \text{constant}$ for each mass shell) is included *implicitly* in the equations themselves. Note that the range of the independent variable is $0 < m < M$ (the total stellar mass) rather than $0 < r < R$ (the stellar radius).

All that remains of the conservation of mass equation, then, is an equation that allows us to calculate the density ρ from the knowledge of how the inner and outer radii of the thin shell differ across it. The continuity equation in mass coordinates is

$$\frac{1}{\rho r^2} \frac{d(\rho r^2)}{dt} = - \frac{dV_{\hat{r}}/dm}{dr/dm}$$

which, because $V_{\hat{r}} = dr/dt$, can be integrated over dt to obtain the simple relation

$$\boxed{\rho = \frac{1}{4\pi r^2 (dr/dm)}} \quad (5.58)$$

Note that we write dr/dm , not dm/dr , as m is now the independent coordinate.

The radius r at the edge of each shell now is a *dependent* variable, and its evolution is found from two equations: the definition of radial velocity

$$\boxed{\frac{dr}{dt} = V_{\hat{r}}} \quad (5.59)$$

and the conservation of momentum (5.48), combined with (5.46), now written in mass coordinates

$$\boxed{\frac{dV_{\hat{r}}}{dt} = -4\pi r^2 \frac{dp}{dm} - \frac{Gm}{r^2}} \quad (5.60)$$

Again, m and dm are simply constants for each mass shell, but r and p are variables that evolve with time.

Finally, we also need to convert equation (5.50) to the convenient mass coordinate system

$$\boxed{\frac{d\epsilon}{dt} = -4\pi p \frac{d(r^2 V_{\hat{r}})}{dm} + \dot{q}} \quad (5.61)$$

where $\epsilon = \varepsilon/\rho$ is the internal energy per unit mass and $\dot{q} \equiv dq/dt$ is the heating and cooling per unit mass. Equations (5.58) to (5.61) are the basic formulae used in almost all computer codes that calculate one-dimensional structure and evolution of stars. We shall make use of these later in several discussions on the formation of stellar-mass black holes from otherwise normal stars.

5.2.3 Newtonian Conservation Laws for a Fluid in the Inertial Eulerian Frame

While the Lagrangian frame follows the fluid flow (as is the case in *smooth particle hydrodynamics*, or SPH, simulations), it is sometimes useful to choose a frame of reference that watches the fluid ebb and flow past fixed points in inertial space. This is called the “Eulerian fluid frame”, after the Swiss mathematician and physicist Leonhard Euler who made significant contributions to mathematics, mechanics, and optics in the mid 1700s.

5.2.3.1 The Vector Partial Differential Form for the Euler Equations

The fluid equations in the inertial Eulerian frame are the same as the Lagrangian ones, but the total time derivative is replaced by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \quad (5.62)$$

This substitution takes into account that mass, momentum, and internal energy can vary at a given point in space not only because of physical processes, but also simply because new fluid elements flow into view and replace old ones. The fluid equations then take on a full partial differential nature in both three-dimensional space and time. This allows them to be solved with standard analytical and numerical methods. The full Eulerian equations for fluid flow, then, are modified versions of equation (5.46)

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0} \quad (5.63)$$

equation (5.48)

$$\boxed{\frac{\partial(\rho \mathbf{V})}{\partial t} + \nabla \cdot (\rho \mathbf{V} \mathbf{V}) = -\nabla p - \rho \nabla \psi} \quad (5.64)$$

and equation (5.50)

$$\boxed{\frac{\partial \varepsilon}{\partial t} + \nabla \cdot (\varepsilon \mathbf{V}) = -p \nabla \cdot \mathbf{V} + \rho \dot{q}} \quad (5.65)$$

Equation (5.63) is said to be in a “fully conservative” form, as it has no sources or sinks of matter on the right-hand side. In equation (5.64) $\mathbf{V} \mathbf{V}$ is the dyadic or “outer” product that creates a tensor from the multiplication of two vectors (as opposed to the inner or “dot” product $\mathbf{V} \cdot \mathbf{V}$, which produces a scalar). Together with Poisson’s equation (5.42), these Euler equations can treat a very large number

of multidimensional astrophysical problems, including stellar and galactic structure and evolution in three dimensions, merging binary stars, jet flows, and many others.

5.2.3.2 The Tensor Form for the Eulerian Conservation of Momentum

Using a more conservative form for equation (5.64), the dynamical forces can be written as being derived from the divergence of the three-dimensional *stress tensor* $\mathbf{T}$

$$\frac{\partial(\rho \mathbf{V})}{\partial t} = -\nabla \cdot \mathbf{T} - \rho \nabla \psi \quad (5.66)$$

where

$$\begin{aligned} \mathbf{T} &= \rho \mathbf{V} \mathbf{V} + p \mathbf{I} \\ &= \begin{pmatrix} \rho V^x V^x + p & \rho V^x V^y & \rho V^x V^z \\ \rho V^y V^x & \rho V^y V^y + p & \rho V^y V^z \\ \rho V^z V^x & \rho V^z V^y & \rho V^z V^z + p \end{pmatrix} \end{aligned}$$

Here $\mathbf{I}$ is the three-dimensional identity tensor

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

At this point we now see that “dynamical” pressure (ρV^2), which is caused by motion of one fluid element toward another, is automatically taken into account in the fluid equations. It has been in the equations all along, but is most easily seen in the above matrix, where those terms occupy the diagonal elements of $\mathbf{T}$, along with the pressure p . The off-diagonal terms are of a similar nature, and represent momentum in one direction brought to a local point in space by flow in another direction.

The tensor form in equation (5.66) is most useful when the stress tensor $\mathbf{T}$ has additional off-diagonal terms, such as those that occur in viscous flows. The tensor form also will be very important in the study of relativistic fluid flow in the next few chapters.

5.2.4 Simple Polytopic Stellar Structure

A particularly good demonstration of the usefulness of the Newtonian fluid equations is the study of stellar structure using the polytropic equation of state (5.55). These were the first stellar models studied, in the early 1900s, before the advent of digital computers in the mid 20th century. While they do not have the ability to

simulate the generation of heat, synthesis of the elements, and evolutionary properties of real stars, polytropes are nevertheless very useful for understanding the basic structure of stars and other astrophysical objects.

5.2.4.1 The Lane–Emden Equation

We begin by specifying that our star must have spherical symmetry ($\partial/\partial\theta = 0$, $\partial/\partial\phi = 0$) and not vary with time ($\partial/\partial t = 0$, including no internal motions [$V_{\hat{r}} = 0$]). Because of these assumptions, we once again can use equation (5.57) for the conservation of mass. Momentum conservation reduces to the equation of hydrostatic equilibrium

$$\frac{dp}{dr} = -\rho \frac{Gm}{r^2} \quad (5.67)$$

and the conservation of energy is simply the polytropic equation of state (5.55). All three conservation laws can be combined to form a single second-order differential equation

$$\frac{1}{r^2} \frac{d}{dr} \left[r^2 \frac{d(\rho^{1/n})}{dr} \right] = - \left[\frac{4\pi G}{K_n (n+1)} \right] \rho \quad (5.68)$$

If we now make variable substitutions for ρ and r

$$\begin{aligned} \rho &\equiv \rho_c \theta^n \\ r &\equiv a_n \xi \end{aligned}$$

where

$$a_n \equiv \left[\frac{K_n (n+1) \rho_c^{\frac{(1-n)}{n}}}{4\pi G} \right]^{1/2} \quad (5.69)$$

and ρ_c is the central mass density (which can be a free parameter), we obtain the Lane–Emden equation [303, 304]

$$\boxed{\frac{1}{\xi^2} \frac{d}{d\xi} \left(\xi^2 \frac{d\theta}{d\xi} \right) = -\theta^n} \quad (5.70)$$

for the structure of a star with a polytropic equation of state of index n . The equation is named after American and Swiss astrophysicists Jonathan Homer Lane and Robert Emden, who lived 60 years apart but worked on similar mathematical studies of stellar structure in the mid–late 1800s and early 1900s. The structures described by this equation are called “polytropes” and have two boundary conditions, both at the center of the polytropic star ($\xi = 0$): one on the value of θ and one on its

derivative there

$$\begin{aligned}\theta(0) &= 1 \\ \left. \frac{d\theta}{d\xi} \right|_{\xi=0} &= 0\end{aligned}$$

The first arises because $\rho = \rho_c$ at the center, and the second because spherical symmetry demands that all quantities have the same value at r and $-r$.

5.2.4.2 Radius, Mass, and Central Concentration of a Polytrope

There are three important properties that we need to know about a polytrope: its radius, mass, and degree of central concentration. Together these tell us a great deal about a star's internal structure. We can determine the polytrope's radius by finding the point $\xi = \Xi$ where the density θ^n reaches its (first) zero

$$\theta(\Xi) = 0$$

This is a condition for finding Ξ , not a boundary condition. With the linear scaling of equation (5.69), the true radius then is

$$\boxed{R_n = \Xi a_n} \quad (5.71)$$

The stellar mass can be found by multiplying equation (5.68) by r^2 and integrating over dr from $r = 0$ to $r = R_n$

$$\boxed{M_n = 4\pi \left[\frac{K_n (n+1)}{4\pi G} \right]^{3/2} \rho_c^{\frac{(3-n)}{2n}} \left(-\xi^2 \frac{d\theta}{d\xi} \right) \Big|_{\xi=\Xi}} \quad (5.72)$$

Finally, the degree of central concentration (the ratio of the central to average densities $\rho_c/\langle\rho\rangle$) can be found from the averaged scaled polytrope density $\langle\theta^n\rangle$, which we get by integrating equation (5.70) times ξ^2

$$\frac{\langle\rho\rangle}{\rho_c} = \langle\theta^n\rangle = \frac{\int_0^\Xi \xi^2 \theta^n d\xi}{\int_0^\Xi \xi^2 d\xi}$$

which gives

$$\boxed{\frac{\rho_c}{\langle\rho\rangle} = \frac{\Xi^3}{3 \left(-\xi^2 \frac{d\theta}{d\xi} \right) \Big|_{\xi=\Xi}}} \quad (5.73)$$

These solutions, therefore, have three free parameters: two for the equation of state (its coefficient K_n and the index n) and the mass M of the star (or, alternatively, ρ_c). From these one can completely determine the structure of the polytrope.

Table 5.1 gives the adiabatic index, scaled radius, scaled mass, and degree of central concentration for four polytropes that are important in this book [301]. And the density profile ρ/ρ_c of each is plotted in Fig. 5.4, on both linear and log–log scales, as a function of r/R . We will discuss these below in more detail.

Table 5.1: Properties of selected polytropes

n	$\Gamma = 1 + 1/n$	Ξ (outer radius)	$\left(-\xi^2 \frac{d\theta}{d\xi}\right)_{\xi=\Xi}$	$\rho_c/\langle\rho\rangle$
1	2	π	π	$\pi^2/3$
3/2	5/3	3.65375	2.71406	5.99071
3	4/3	6.89685	2.01824	54.1825
∞	1	∞	∞	∞

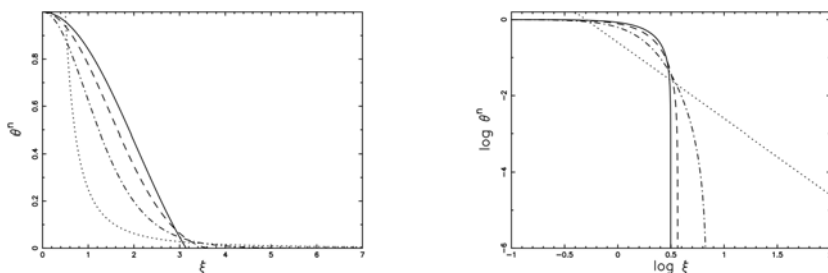


Fig. 5.4: Density structure comparison of the four polytropes in Table 5.1. *Left:* Linear profiles. *Right:* Log–log profiles. The different curves are: $n = 1$ (solid); $n = 3/2$ (dashed); $n = 3$ (dot-dashed); $n = \infty$ (dotted). Note that the stellar envelope becomes more extended as n increases, but the stellar core also becomes more compact.

5.2.4.3 The $n = 1$ Polytrope: A Simple Analytic Stellar Model

While the $n = 1$ polytrope is not important astrophysically (it corresponds to a gas with an unusually large adiabatic index of $\Gamma = 2$), it is interesting because it has a particularly simple analytic solution that satisfies equation (5.70) and the boundary conditions that we require

$$\theta(\xi) = \frac{\sin \xi}{\xi}$$

The first zero of this sinc function is at $\Xi = \pi$, which gives us a stellar radius *independent* of the central density

$$R_1 = \left(\frac{\pi K_1}{2G} \right)^{1/2}$$

The scaled mass $(-\xi^2 d\theta/d\xi)|_{\xi=\Xi} = \pi$ also, which gives a total star mass that *does* depend on central density (linearly)

$$M_1 = 4\pi^2 \left(\frac{\pi K_1}{2G} \right)^{3/2} \rho_c$$

In this case, the star is not very centrally condensed ($\rho_c/\langle\rho\rangle \approx 3.3$).

5.2.4.4 The $n = 3/2$ Polytrope: The Isentropic, Perfect Gas Star

A star composed of an ideal gas, all with the same entropy, is an $n = 3/2$ polytrope ($\Gamma = 5/3$ gas). The solution to the Lane–Emden equation is not analytic, unlike that for the $n = 1$ case. However, it can be integrated numerically (in fact, even by hand) and is plotted also in Fig. 5.4. It is a little more distended ($\Xi = 3.65$) than the $n = 1$ polytrope, but also more centrally condensed ($\rho_c/\langle\rho\rangle = 5.99$). And its radius and mass both depend on the central density

$$R_{3/2} = 3.65 \left(\frac{5K_{3/2}}{8\pi G} \right)^{1/2} \rho_c^{-1/6}$$

$$M_{3/2} = 10.86\pi \left(\frac{5K_{3/2}}{8\pi G} \right)^{3/2} \rho_c^{1/2}$$

A strange feature of this solution is that it predicts a mass–radius relation that is *very* different from what is observed for stars like the sun: $R \propto M^{-1/3}$, instead of the observed $R \propto M^{+0.65}$. The reason is that, unless they are convective, stars are not isentropic (they are heated in their centers by nuclear burning), and stars for different masses have different thermal energy content (*i.e.*, $K_{3/2}$ varies with mass M).

5.2.4.5 The $n = 3$ Polytrope: White Dwarf Stars, the Chandrasekhar Mass Limit, and the Formation of Neutron Stars and Black Holes

An $n = 3$ polytrope ($\Gamma = 4/3$ gas) has a very interesting mass–radius relation. Again, the solution for $\theta(\xi)$ is not analytic, but can be numerically integrated, and used to obtain Ξ and $(-\xi^2 d\theta/d\xi)|_{\xi=\Xi}$. This produces

$$R_3 = 6.90 \left(\frac{K_3}{\pi G} \right)^{1/2} \rho_c^{-1/3}$$

$$M_3 = 8.07\pi \left(\frac{K_3}{\pi G} \right)^{3/2}$$

from equations (5.71) and (5.72). This solution has a very high central mass concentration ($\rho_c/\langle\rho\rangle \approx 54$). Furthermore, *only one value of the stellar mass is allowed for any central density*, and *that mass depends only on the coefficient of the equation of state K_3* . If a star should ever be isentropic throughout, and have a $\Gamma = 4/3$ equation of state, it could have only one particular stellar mass. If mass were added, or removed, the star no longer could be an $n = 3$ polytrope; its equation of state must change in some way – a phase change.

The center of a white dwarf star is composed of helium or heavier elements, whose electrons are so tightly packed that they comprise a degenerate, relativistic electron gas that is highly conducting and, therefore, essentially isentropic. The equation of state for such a gas is a function of density only (equation (9.67)) and it has a $4/3$ adiabatic index

$$p = 1.24 \times 10^{15} \text{ dyn cm}^2 \left(\frac{\rho}{\mu_e} \right)^{4/3} \quad (5.74)$$

where $\mu_e = N_A/n_e$ is the “average molecular weight per electron” of the white dwarf material (in units of grams per mole), and n_e is the electron density. Generally, $\mu_e \approx 2$ for helium and heavier elements. With this physical polytropic equation of state, the mass of a white dwarf composed entirely of material whose electrons are degenerate and relativistic must be

$$M_3 = M_{\text{CH}} \equiv \frac{5.81 M_\odot}{\mu_e^2} \approx 1.45 M_\odot \quad (5.75)$$

This quantity is known as the Chandrasekhar mass or Chandrasekhar limit, after the American astrophysicist Subrahmanyan Chandrasekhar, who was the first to show that this was a crucial mass value in the evolution of stars.

What happens when the mass of a white dwarf is not equal to $1.45 M_\odot$? When $M_{\text{WD}} < M_{\text{CH}}$, the electrons in the interior are under less pressure and, outside some radius $r_{\text{NR}} < R_3$, they behave as a *non-relativistic*, but still degenerate, electron gas with an equation of state (9.66)

$$p = 1.00 \times 10^{13} \text{ dyn cm}^3 \left(\frac{\rho}{\mu_e} \right)^{5/3} \quad (5.76)$$

(i.e., an $n = 3/2$ polytrope) for $r_{\text{NR}} < r < R$. When $M_{\text{WD}} \rightarrow M_{\text{CH}}$, the entire star is held up by degenerate, relativistic electron pressure and, therefore, is an $n = 3$ polytrope, with the equation of state given by (5.74). However, when $M_{\text{WD}} > M_{\text{CH}}$, something strange must happen. The electrons may become more relativistic,

but the equation of state still must be given by (5.74). In short, this situation cannot remain stable. When more mass is added to a white dwarf of Chandrasekhar mass, the equation of state *must* change from a relativistic, degenerate electron gas to something else.

This phase change, in fact, turns out to be quite catastrophic, and of great importance to the formation of neutron stars and stellar-mass black holes. When $M_{\text{WD}} > M_{\text{CH}}$, electrons are forced to combine with protons to form neutrons, the central electron pressure holding the star up drops, and the object then collapses to a completely new type of star – either a neutron star or a black hole. How this occurs is discussed in more detail in Chapters 8 and 10.

5.2.4.6 The $n = \infty$ Polytrope: The Singular Isothermal Sphere and the Structure of Globular Star Clusters

Another polytrope that is very useful for astrophysics is one in which the pressure is simply proportional to density

$$p = K_{\infty} \rho$$

This occurs in ideal gases that are held at a uniform temperature (isothermal). Clusters of thousands to millions of stars are observed to behave as an isothermal gas (with each star considered a single thermal particle). So the $n = \infty$ polytrope also is very useful in the study of the dynamics of stellar systems (“stellar dynamics”).

Because of mathematical problems with setting $n = \infty$ in equation (5.70), we must re-cast the Lane–Emden equation for this case without the scaling length a_n or central density ρ_c

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d(\ln \rho)}{dr} \right) = -\frac{4\pi G}{K_{\infty}} \rho \quad (5.77)$$

This new, $n = \infty$ polytrope equation has the simple solution

$$\rho = \left(\frac{K_{\infty}}{2\pi G} \right) \frac{1}{r^2} \quad (5.78)$$

This is called the “singular isothermal solution”, because everything (ρ_c , M_{∞} , R_{∞}) appears to be infinite, even though the density itself is finite at any one radius r .

The situation is not as bad as it seems, however. Even though ρ_c is infinite, the mass interior to radius r vanishes as $r \rightarrow 0$

$$m(r) = \left(\frac{2K_{\infty}}{G} \right) r$$

And, if we cut off the structure at some outer radius R_t , we still obtain a finite mass $M_t = 2K_{\infty} R_t/G$. In the case of globular clusters, R_t is called the “tidal radius”,

because stars beyond that have been stripped away from the cluster by tidal forces of the parent galaxy. Furthermore, as $r \rightarrow 0$, the assumption of having a “fluid” composed of many stars breaks down, leaving only one or two stars at the center. So the central density of an isothermal star cluster remains finite, even though the simple mathematical solution becomes infinite there.

The development of a singular isothermal structure in a star cluster is called “core collapse”, and may have some applications to the formation of intermediate mass black holes in star clusters (Section 10.4.2).

5.3 Summary

Despite their simplicity, Newton’s laws of motion and gravity are nevertheless very powerful and are the basis of most of classical physics that is used to describe many terrestrial and astrophysical phenomena. However, they only can describe objects that may eventually become black holes; they cannot describe the black holes themselves. For that we will need Einstein’s general laws of motion and gravity.

Chapter 6

Geometry and Physics without Gravity: Special Relativity

The next two chapters deal with Einstein's theory of relativity, which will be needed to discuss gravitational waves and collapse in Chapter 8 and relativistic mechanics in Chapter 9. This is, by no means, a book on relativity, so we will try to keep the treatment and mathematics of relativity as simple as possible and yet still construct all the concepts and tools that are needed to tinker with black hole engines.¹ Knowledge of relativity requires a knowledge of geometry – and considerably more than high school or college geometry. Nevertheless, that is where we will start, and then develop only enough to understand metrics and four-dimensional spacetime (which is crucial to even begin understanding black holes). We then shall introduce the basic concepts of mechanics like time dilation, Lorentz transformations, four-vectors, and relativistic beaming. And, finally, we will begin our extended discussion on fluid dynamics and electromagnetism, but deal here with flat spacetime only. Once we have covered Einstein's theory of gravity in Chapter 7, will we return to this subject and go into more depth in Chapter 9.

6.1 Why Geometry?

Ever since Albert Einstein published his theories of special and general relativity, in 1905 and 1915, respectively [307, 308, 309], it has been recognized that geometry is important not only to mathematics, but also to physics as well. There are two major reasons for this. First, it has become very apparent that the fundamental number of dimensions that describe our (macroscopic) world is four, not three. We perceive three spatial dimensions, of course, as we travel to and fro on the earth's surface,

¹ The material herein is based loosely on elementary plane geometry, special relativistic aspects of Misner, Thorne, and Wheeler's classic book on gravitation [305], and standard electromagnetics, as discussed in texts like Jackson's [306]. However, the use of matrix, rather than summation, notation is primarily the author's own approach. It suffices for demonstration purposes in most places, except in Chapter 7 when we need the Riemann 4-tensor (which cannot be printed as a four-dimensional matrix on these pages).

climb stairs, build buildings, and perform many other activities that require motion in three perpendicular directions. And we perceive some motion through time as we live out our lives. The special theory of relativity, however, showed that these are united into a single four-dimensional geometry, and one that is somewhat different from that of Pythagoras or Euclid. Motion at high speeds simply *cannot* be understood without considering both space and time together.

The second reason that geometry is important to physics is the discovery that, contrary to what Isaac Newton had suggested, gravity does not result simply from two masses attracting each other. Rather, gravity is the result of matter bending (or “warping”) the geometry of space and time, so that it is no longer the “flat” geometry of special relativity. In fact, as we shall see, the pull of gravity that we feel is the true “time warp”. That is, if we travel toward a planet or star in a spaceship with the power turned off, unless we turn the engines back on or alter our course in some other way, the object’s gravity will determine our future motion (toward the object) and perhaps our ultimate destiny (crash landing). In the case of a black hole, the warping of time is as extreme (there is a point of no return) as the warping of space (the surface area of a black hole is proportional to its radius, for example, not the square of its radius as we are used to in normal flat three dimensions). So, while physics is not *all* geometry, nevertheless an understanding of geometry is the key to understanding physics.

6.2 Two-Dimensional Pythagorean Geometry

6.2.1 The Line Element

We begin our discussion of black hole geometry with a review of high school and college geometry, using the mathematical language that we will need for black holes. The most fundamental equation of plane geometry is the Pythagorean theorem for a right triangle

$$h^2 = x^2 + y^2 \quad (6.1)$$

where x and y are the two sides of the triangle and h is the hypotenuse. Because the x and y coordinates in a graph are at right angles to each other, this theorem is important in determining the distance Δh between two points on a graph

$$(\Delta h)^2 = (\Delta x)^2 + (\Delta y)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 \quad (6.2)$$

where Δx refers to the difference between two x values, x_1 and x_2 . (See [Fig. 6.1](#).)

When mathematicians describe geometry it is customary (and usually necessary) to deal with *very small* coordinate differences dx rather than finite differences Δx . The Pythagorean theorem for a two-dimensional plane then becomes

$$\boxed{dh^2 = dx^2 + dy^2} \quad (6.3)$$

where dx^2 is $(dx)^2$, not $d(x^2)$. Mathematically the differentials dx , dy , and dh are formally as close to zero distance as possible without actually being zero. In practice, however, equation (6.3) works well with the differentials being fairly large. For example, on the surface of the earth, equation (6.3) is accurate to a part in a thousand in laying out streets for a city 6 km (four miles) in size. So, “infinitesimal” distances like dx should be considered small in relation to a characteristic scale length (like the radius of the earth here).

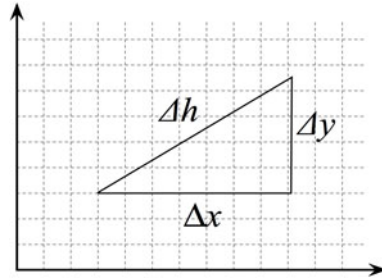


Fig. 6.1: The distance between two points on a two-dimensional graph Δh is a simple example of a line element. The three sides of the triangle are related by the Pythagorean theorem (eq. (6.2)), which also is the equation for the finite line element Δh .

Equation (6.3) is called the “line element” that describes plane geometry. The line element is a *global statement* about a *local property* of the geometry. That is, equation (6.3) is valid anywhere in an infinite plane (a global statement), but the line element describes only how the length of a small rod dh can be broken down into separate pieces in the different coordinates x and y (a local property). In general, the line element does *not* describe how to compute the distance between two *widely* separated points (although in the special case of planar geometry and Cartesian coordinates, the equations for large and small distances have the same form).

6.2.2 Line Element Has the Same Length in Any Coordinates

An obvious, but important, feature of equation (6.3) is that the distance dh remains the same value even when we change the coordinate system used to identify points in the infinite plane. For example, a translation of the origin to a new location (x_0, y_0) produces the coordinate transformation $x' = x - x_0$ and $y' = y - y_0$, but dh is still given by (see Fig. 6.2b)

$$(dh')^2 = (dx')^2 + (dy')^2 = (dx)^2 + (dy)^2 = (dh)^2 \quad (6.4)$$

since dx_0 and dy_0 are zero. Similarly, a *rotation* of the axes by a *constant* angle α

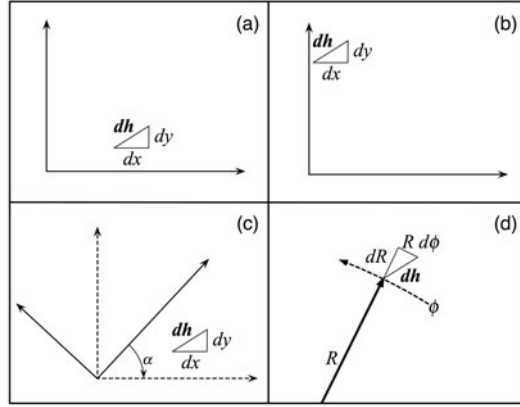


Fig. 6.2: Graphical demonstration that the infinitesimal line element dh remains unchanged during any coordinate transformation. Original line element and coordinate system are in (a). The line element stays the same for a translation (b) or rotation (c) of the coordinate frame. Even re-labeling the coordinates to polar (d), for example, does not change the line element – provided all differentials are small compared the size of the system (such as the distance R from the origin). The arrows show the R and ϕ vector directions in the new coordinates, and eq. (6.7) give the transformation between (x, y) and (R, ϕ) .

$$\begin{aligned} x'' &= x \cos \alpha + y \sin \alpha \\ y'' &= -x \sin \alpha + y \cos \alpha \end{aligned} \quad (6.5)$$

also always yields the same distance dh (see Fig. 6.2c)

$$\begin{aligned} (dh'')^2 &= (dx'')^2 + (dy'')^2 \\ &= (dx)^2 \cos^2 \alpha + 2 dx dy \sin \alpha \cos \alpha + (dy)^2 \sin^2 \alpha + \\ &\quad (dx)^2 \sin^2 \alpha - 2 dx dy \sin \alpha \cos \alpha + (dy)^2 \cos^2 \alpha \\ &= dx^2 + dy^2 = dh^2 \end{aligned} \quad (6.6)$$

A third kind of coordinate change, from Cartesian (x, y) to polar (R, ϕ) , with $x = R \cos \phi$ and $y = R \sin \phi$ does not move the origin or rotate the axes; it simply re-draws the coordinate lines in the plane (“changes the coordinate labels”). The differentials are

$$\begin{aligned} dx &= \cos \phi dR - R \sin \phi d\phi \\ dy &= \sin \phi dR + R \cos \phi d\phi \end{aligned} \quad (6.7)$$

If we combine equations (6.3) and (6.7) the line element in polar coordinates becomes

$$\begin{aligned} dh^2 &= \cos^2 \phi dR^2 - 2R \sin \phi \cos \phi dR d\phi + R^2 \sin^2 \phi d\phi^2 + \\ &\quad \sin^2 \phi dR^2 + 2R \sin \phi \cos \phi dR d\phi + R^2 \cos^2 \phi d\phi^2 \\ &= dR^2 + R^2 d\phi^2 \end{aligned} \quad (6.8)$$

dh is still the length between two closely spaced points, and equation (6.8) tells us how to use the Pythagorean theorem to calculate it from polar coordinates. But whether we express the coordinates as (R, ϕ) or as (x, y) , equations (6.3) and (6.8) will obtain the same answer for dh .

We therefore have deduced a key point: the *geometry* of a surface is described by an equation giving the distance between any two closely-spaced points (the line element dh), and that geometry remains unchanged during any coordinate transformation. *Geometry, therefore, transcends the coordinates that we use to describe it.* This statement remains true for other geometries as well. In spherical geometry, for example, the shape and curvature of the surface of a sphere (like that of the earth) remain the same no matter what coordinates we use to find our way on it.

6.2.3 Matrix Form for the Geometry Equations

6.2.3.1 The Metric

The geometry equations can be written in a matrix form. Of course, if the full matrices are written out, this can look rather messy. But if we use a single variable to represent an entire matrix, the results are very simple, elegant, and easier to work with than the differential notation above. We will use as an example the previous case of conversion from Cartesian to polar coordinates. This will help illustrate the sophisticated concepts of metrics, vectors, 1-forms, and tensors. Equation (6.3) can be written in matrix form as²

$$\boxed{dh^2 = \mathbf{dx}^T \cdot \mathbf{g} \cdot \mathbf{dx}} \quad (6.9)$$

where

$$\mathbf{dx} = \begin{pmatrix} dx \\ dy \end{pmatrix} \quad \mathbf{dx}^T = (dx \ dy)$$

are the differential displacement matrix and its transpose, and

$$\mathbf{g} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (6.10)$$

is the *metric of a flat two-dimensional plane in Cartesian coordinates*. Do not be confused by $\mathbf{g}$ being equal to the well-known constant identity matrix

$$\mathbf{I} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

² While it is not absolutely necessary to use the dot ($\cdot$) symbol for the scalar product, it will make these and many equations below much easier to understand as both matrix and vector equations.

$\mathbf{g}$ is a general expression for the metric of a geometry and just happens to be equal to $\mathbf{I}$ in this case.

The metric is an $N \times N$ matrix, where N is the number of dimensions in the geometry (here two), and its entries give the coefficients of the products of pairs of differentials in the line element equation. (In equation (6.3) the coefficients of dx^2 and dy^2 are 1, while the coefficients of $dx dy$ and $dy dx$ are 0.) The terms “metric” and “line element” often are used interchangeably, since the coordinates are secondary. But, formally, the term “metric” really refers to the geometric coefficients of the differentials in the line element equation, and “line element” refers to the differential dh on the left-hand side of that equation.

Equation (6.8) can be written similarly as

$$dh^2 = dx'''^T \cdot \mathbf{g}''' \cdot dx''' \quad (6.11)$$

where

$$dx''' = \begin{pmatrix} dR \\ d\phi \end{pmatrix} \quad \mathbf{g}''' = \begin{pmatrix} 1 & 0 \\ 0 & R^2 \end{pmatrix} \quad (6.12)$$

Note that, like $\mathbf{g}$, $\mathbf{g}'''$ also is the metric of a flat plane but *expressed in polar coordinates*.

6.2.3.2 Coordinate Transformations: Vectors and 1-Forms

Coordinate transformations also can be written in matrix notation. For the conversion from Cartesian to polar coordinates, we have

$$dx''' = \mathbf{L} \cdot dx \quad (6.13)$$

where the coordinate transformation matrix $\mathbf{L}$ is

$$\mathbf{L} = \frac{1}{R} \begin{pmatrix} R \cos \phi & R \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix}$$

Note that the *inverse* coordinate transformation also can be expressed in matrix form

$$dx = \mathbf{L}^{-1} \cdot dx''' \quad (6.14)$$

where

$$\mathbf{L}^{-1} = \begin{pmatrix} \cos \phi & -R \sin \phi \\ \sin \phi & R \cos \phi \end{pmatrix}$$

and that the transformation and its inverse are normalized to a unit matrix

$$\mathbf{L}^{-1} \cdot \mathbf{L} = \mathbf{I} \quad (6.15)$$

Note that $\mathbf{L}$ is a *differential* transformation that operates on local vectors and 1-forms, changing their components in one coordinate system to those in another. *It does not operate on the coordinates themselves.* See Section 6.2.4.

Any quantity whose matrix of components transforms like dx in equation (6.13) is called a “contravariant vector” or just “vector”.³ In the transformation equation for the components of any vector $\mathbf{V}$, the matrix of the vector’s components $\mathbf{V}'''$ in the new coordinate system is on the left-hand side, and the matrix of components in the old system (and the transformation matrix between the two) are on the right

$$\mathbf{V}''' = \mathbf{L} \cdot \mathbf{V} \quad (6.16)$$

The matrix of a vector is written as a column

$$\mathbf{V} = \begin{pmatrix} V^x \\ V^y \end{pmatrix}$$

with *superscripts* for the coordinate indices, not subscripts.

In advanced geometry there also is a conjugate quantity to the contravariant vector, called the “covariant vector” or “1-form”, whose components transform in the opposite manner to those of a contravariant vector

$$\mathbf{v}''' = \mathbf{v} \cdot \mathbf{L}^{-1} \quad (6.17)$$

We usually will write vectors with capital bold letters and 1-forms with lower-case bold letters. The *inverse* of $\mathbf{L}$ is now on the right-hand side, but this is *not* just an inverse transformation like (6.14). The matrix for $\mathbf{v}'''$ in the new coordinate system still is on the left-hand side in equation (6.17), and that for $\mathbf{v}$ in the old system is on the right. $\mathbf{v}$ and $\mathbf{V}$ are, therefore, different geometric quantities (covariant and contravariant vectors) that transform in different ways. The 1-form is written as

$$\mathbf{dx}^T = (v_x \ v_y)$$

i.e., a *row* matrix with *subscripts* for the indices in this case.

Because it is confusing and wordy to use the adjectives contravariant and covariant, and because the term “vector” is already familiar to us, from now on we will use the terms “1-form” for covariant vector and “vector” for contravariant vector.

6.2.3.3 Dot Products and Magnitudes of Vectors

Equation (6.9) has another important implication. The magnitude of any vector, including dx , should not change if we simply change coordinates. Now, the magnitude of dx is, in fact, the line element dh , but from equations (6.9) and (6.11) we see that

³ In fact, dx represents the components of a special vector called the “basis 1-forms”, which allow 1-forms to be expressed as true geometric objects, with magnitude and direction. Similarly, there is a matrix of “basis vectors” $\mathbf{e}$, which allow us to express vectors as geometric objects.

it cannot be simply the dot product of dx with itself; *i.e.*, $dh^2 \neq dx^T \cdot dx$. The magnitude dh is invariant only if we compute it by inserting $\mathbf{g}$ *between* the dx^T and dx , as in equation (6.9).

Similarly, in order to compute the magnitude of all vectors $\mathbf{V}$, we also must use the metric (unless we are using Cartesian coordinates with $\mathbf{g} = \mathbf{I}$)

$$\boxed{V^2 = \mathbf{V}^T \cdot \mathbf{g} \cdot \mathbf{V}} \quad (6.18)$$

It is easy to show that the quantity

$$\mathbf{v} \equiv \mathbf{V}^T \cdot \mathbf{g} \quad (6.19)$$

transforms according to equation (6.17)

$$\mathbf{v}' = \mathbf{v} \cdot \mathbf{L}^{-1} \quad (6.20)$$

and is a 1-form. So, *the quantity that is invariant under coordinate transformations is the dot product of a vector with its corresponding 1-form, not with itself*

$$\boxed{V^2 = \mathbf{v} \cdot \mathbf{V}} \quad (6.21)$$

Also, because equation (6.19) can be written as

$$\mathbf{V} = \mathbf{g}^{-1} \cdot \mathbf{v}^T \quad (6.22)$$

the dot product in equation (6.18) also can be expressed with 1-forms

$$\boxed{V^2 = \mathbf{v} \cdot \mathbf{g}^{-1} \cdot \mathbf{v}^T} \quad (6.23)$$

6.2.3.4 Coordinates are NOT Vectors

When first learning vector analysis, one is often taught that the line connecting two points, particularly between the origin and a point in space, is the position vector

$$\mathbf{r} = \begin{pmatrix} x \\ y \end{pmatrix}$$

But x and y are *not* vectors in general. Consider the following *incorrect* transform:

$$\begin{pmatrix} x \\ y \end{pmatrix} \neq \mathbf{L} \cdot \begin{pmatrix} R \\ \phi \end{pmatrix}$$

While the units on each side of this equation are correct, it most decidedly is not a correct equation; it does not convert polar coordinates to Cartesian ones. Differential transformations like $\mathbf{L}$ convert components of local quantities like vectors and 1-forms, not coordinates or long distances between points. (See also Section 6.2.4.)

6.2.3.5 Tensors

We end our discussion of two-dimensional geometry by looking at how the metric itself transforms. We can combine equations (6.9), (6.11), and (6.13) to solve for $\mathbf{g}'''$ in terms of $\mathbf{g}$ to get

$$\mathbf{g}''' = (\mathbf{L}^{-1})^T \cdot \mathbf{g} \cdot \mathbf{L}^{-1} \quad (6.24)$$

$\mathbf{g}$ therefore, transforms in a manner similar to a 1-form (with $\mathbf{L}^{-1}$), but with another factor of the inverse transpose in front. An object like $\mathbf{g}$ is called, therefore, a “covariant tensor of rank 2”, and we express its components with lowered indices

$$\mathbf{g} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \quad (6.25)$$

$\mathbf{g}$ has another special property that not all tensors have. It is always symmetric

$$\mathbf{g}^T = \mathbf{g} \quad (6.26)$$

Note that $\mathbf{g}$ is specifically not called a 2-form. This is because *true* 2-forms must possess another property – antisymmetry – whereas $\mathbf{g}$ is symmetric.

The metric also can be converted to a contravariant form by taking its inverse. It is easy to show that $\mathbf{g}^{-1}$ transforms in a manner similar to a (contravariant) vector, but with an added factor of $\mathbf{L}^T$

$$(\mathbf{g}''')^{-1} = \mathbf{L} \cdot \mathbf{g}^{-1} \cdot \mathbf{L}^T \quad (6.27)$$

and therefore is called a “contravariant tensor of rank 2”, or just “tensor of rank 2”, or simply a “2-tensor”. $\mathbf{g}^{-1}$ also is symmetric, and is used to convert 1-forms into vectors (equation (6.22)) and general covariant tensors $\mathbf{t}$ into contravariant ones $\mathbf{T}$

$$\mathbf{T} = \mathbf{g}^{-1} \cdot \mathbf{t} \cdot \mathbf{g}^{-1} \quad (6.28)$$

$\mathbf{g}^{-1}$ is also the corresponding 2-tensor for $\mathbf{g}$ because

$$\mathbf{G} = \mathbf{g}^{-1} \cdot \mathbf{g} \cdot \mathbf{g}^{-1} = \mathbf{g}^{-1} \quad (6.29)$$

Because we usually use the same letter of the alphabet for expressing components of corresponding vectors and 1-forms (and tensors and covariant tensors), but write the indices in either the upper or lower positions, respectively, we also use the same letter “g” for components of $\mathbf{g}^{-1}$

$$\mathbf{g}^{-1} = \begin{pmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{pmatrix} \quad (6.30)$$

6.2.4 A Word of Warning: Global vs. Local Transformations

Because it is quite common to use the differential (equation (6.7)) or matrix (equation (6.13)) forms for coordinate transformations, as we usually do in this book, one has to be careful to recognize when the coordinate transformation being made is a truly global one or when it is simply a local one. What is the difference?

A global coordinate transformation is one that can express the new coordinates as complete and independent functions of the old coordinates. That is,

$$x' = f(x, y) \quad y' = g(x, y) \quad (6.31)$$

with f and g being independent functions of x and y (as in equation (6.5)). Furthermore, while a global transform can be written in a differential or matrix manner

$$dx' = a(x, y) dx + b(x, y) dy \quad dy' = c(x, y) dx + d(x, y) dy$$

these expressions *must be total differentials*, expressible in the form

$$dx' = d[f(x, y)] \quad dy' = d[g(x, y)]$$

for the transformation to be truly global. Global transformations allow the points in an entire space (of any number of dimensions) to be re-labeled from, say $(x, y, z, \dots)$ to $(x', y', z', \dots)$: they work everywhere in the space.

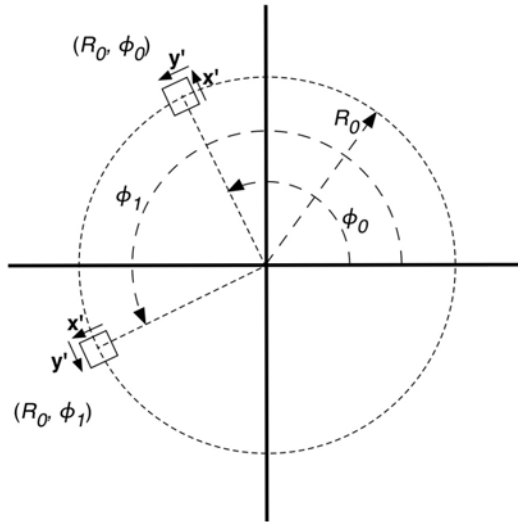


Fig. 6.3: Demonstration of the limitations of local differential coordinate transformations, using the one in eq. (6.32). The orientation of coordinate patch (x', y') changes with the position where it is sited, showing that the new locally Cartesian coordinates cannot be used to span the entire plane for all points (R, ϕ) .

A local transformation, on the other hand, is one that may be expressed in differential or matrix form, but cannot be integrated into the global form (6.31). A simple example would be the following transformation from polar coordinates (R, ϕ) to a local orthonormal system (x', y') at a single point

$$dx' = dR \quad dy' = R d\phi \quad (6.32)$$

This transformation looks valid, but it cannot be integrated into the form (6.31), so it must be only a local transformation. What does this look like graphically? Equations (6.32) erect a small $x'-y'$ grid at a single point (R, ϕ) with the x' axis parallel to a line from the origin to this point and the y' axis perpendicular to that line. (See Fig. 6.3.) The “coordinates” x' and y' are useful for describing the positions of points very near (R, ϕ) in a Cartesian manner, but they *cannot* be used to label points throughout the entire space for all R and all ϕ . For example, if we erect a small $x'-y'$ orthonormal grid at (R_0, ϕ_0) , the new Cartesian axes will have a specific orientation in the plane. On the other hand, if we choose a new (R_0, ϕ_1) , then we will get a different set of $x'-y'$ axes, with a different orientation along the same circle of radius R_0 . The two different sets of (x', y') coordinates are not the same and cannot be used to span the entire plane. So, each different (x', y') system is valid only locally, and the transformation to these axes must be only a local one.

A proper global differential transformation to replace equations (6.32) would be that in equations (6.7) (with primes on the Cartesian coordinates), which can be integrated to $x' = R \cos \phi$ and $y' = R \sin \phi$.

6.3 Three-Dimensional Euclidean Geometry

We can easily see how two-dimensional Pythagorean geometry can be extended to three dimensions by drawing a third axis (z) perpendicular to the x and y axes and then a plane that includes the z axis and the hypotenuse in equation (6.2) (see Fig. 6.4). In this plane the Pythagorean theorem also holds true, so a distance $\Delta\ell$ in that plane is given by

$$\Delta\ell^2 = \Delta h^2 + \Delta z^2 \quad (6.33)$$

Combining this with equation (6.2) we obtain

$$\Delta\ell^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 \quad (6.34)$$

or, in differential form,

$$\boxed{d\ell^2 = dx^2 + dy^2 + dz^2} \quad (6.35)$$

Since this is true for any dx , dy , and dz , equation (6.35) gives the line element for three-dimensional flat space in Cartesian coordinates. The metric associated with this line element is an extension of equation (6.10)

$$\mathbf{g} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (6.36)$$

From the time of Euclid until the latter part of the 19th century, it was thought that this metric represented the geometry of our entire universe.

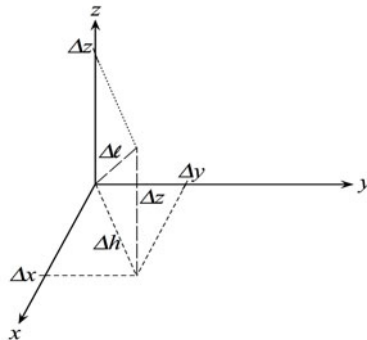


Fig. 6.4: Graphical demonstration of the Pythagorean theorem (eq. (6.34)) and the line element $\Delta\ell$ in three dimensions.

Flat Euclidean space can be described by other coordinate systems: cylindrical

$$d\ell^2 = dR^2 + R^2 d\phi^2 + dz^2 \quad (6.37)$$

which is just an extension of the two-dimensional polar coordinate system of equation (6.8) into the z direction, and spherical-polar coordinates

$$d\ell^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (6.38)$$

which consists of one coordinate with dimensions of length (r) and two angular coordinates (θ, ϕ). The construction of coordinate transformations, vectors, 1-forms, and tensors in these three-dimensional coordinate systems is similar to that in two dimensions. And, as before, the scalar product of two vectors or 1-forms must involve the metric

$$V^2 = \mathbf{V}^T \cdot \mathbf{g} \cdot \mathbf{V} = \mathbf{v} \cdot \mathbf{g}^{-1} \cdot \mathbf{v}^T \quad (6.39)$$

6.4 Four-Dimensional Minkowski Geometry

Although the ability to accelerate matter to speeds close to that of speed of light was not achieved until the 20th century, the accurate *measurement* of the speed of light was routinely done near the end of the 19th. These experiments revealed an amazing result: the measured speed of light was the exactly the same value to a high degree of accuracy in the direction of the earth's motion about the sun (which is about 30 km s^{-1}) as it was perpendicular to that motion, no matter what time of year or what direction the earth was traveling. This was in stark contrast to previous physical experience, where speeds of motion and waves added or subtracted in vector fashion. Until the late 1800s, it had been thought that light and radio waves propagated in an invisible sea of “aether” that permeated the universe, just like sound waves on the earth propagate through a sea of air. If one moved fast enough through that sea, the speed of the waves should be different in the direction of motion than perpendicular to it. In extreme cases, the speed could be negative as we catch up to or overtake the waves that are generated by the experiment. The constancy of the speed of light in any direction indicated that the earth did not drift with respect to the aether, just like air in the closed cockpit of a supersonic jet fighter does not drift with respect to the airplane. Either the universal aether somehow followed the earth around in its orbit (an anthropocentric view), or there simply was no aether at all.

It eventually became apparent that this behavior of light was due not to any unique properties of light itself, but rather to the extreme speed at which the light “particles” travel. *Any* such beam of *any* kind of particles generated at the speed of light by a moving observer would be received by a stationary observer at that same speed – regardless of how fast the two observers were moving relative to one another (see Fig. 6.5). It mattered only that the *particles* were moving at a speed of about $c \equiv 299,792.5 \text{ km s}^{-1}$. The constancy of the speed of particles moving this fast in any frame of reference prompted physicists to re-think their ideas about the quantities that are used to calculate that speed – distance and time. Perhaps these were not the same in different frames of reference, unlike what had been previously thought. That is, perhaps a second and a centimeter for one person are not the same as a second and a centimeter for another person who is traveling at a different speed.

There was a big problem, however. Who has the correct time? Let us use travelers in spaceships as concrete examples. Suppose that there are two spaceships, each traveling past the earth at different speeds. If, as viewed from the earth, clocks tick at one rate in one spaceship and at a different rate on the other, how do the people in the two spaceships view each other? That is, which space traveler decides that he has the slower-running clock, and which decides that he has the faster clock? The answer is neither, if one considers the “special principle of relativity”, which Albert Einstein used to develop the special theory of relativity in 1905. This principle states that the laws of physics are the same for any traveler moving at a constant speed, *i.e.*, in an inertial frame of reference. Therefore, it does not matter how fast each spaceship is going with respect to the earth, only how fast each spaceship is going relative to the other. Each space traveler will get the same answer (that the other traveler's clock is running slower than his own).

6.4.1 The Four-Dimensional Minkowski Line Element

The relative nature of space and time was very unsettling when it was first discovered. The entire notion of Pythagorean/Euclidean geometry was destroyed, because distances measured between points in space came out *differently* depending on how fast one was traveling. The lengths of vectors were no longer constant from one reference frame to the next. This prompted physicists to seek a new order to things – a new geometry – that was consistent with relative time and distance and yet also consistent with the old Euclidean geometry. The answer was proposed by Hermann Minkowski around 1907. Just as we extended the two-dimensional Pythagorean line element to include the third direction of height (z) in equation (6.33), so Minkowski proposed that the three-dimensional Euclidean line element should be extended to include a fourth dimension ($w = ct$) in the direction of time. However, unlike adding another Euclidean dimension to the line element, which would have a coefficient of $+1$, this new time dimension should have a coefficient of -1 . So the line element that Minkowski proposed to describe the universe better than Euclidean geometry is

$$\Delta s^2 = -\Delta w^2 + \Delta \ell^2 \quad (6.40)$$

or, combining with equation (6.34),

$$\Delta s^2 = -c^2 \Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2 \quad (6.41)$$

The universe then, is made up of four dimensions, not three, and points (now called “events”) in this universe have a fourth coordinate, the time t . The vector magnitude that remains the same for any traveler is not the spatial distance $\Delta \ell$, but instead the “proper distance” Δs , which includes a term that takes into account the time difference Δt between two events.

Equation (6.41) has some big surprises. Sitting in our chairs, reading this book, and apparently not moving ($\Delta x = \Delta y = \Delta z = 0$), the proper distance between a point in space (say, the seat of our chair) at one time and that same point at a later time *is not zero*. We still have a motion through a fourth dimension, and the speed of motion is the speed of light

$$\Delta s^2 = -(c \Delta t)^2 \quad (6.42)$$

With each one-second tick of the clock, everyone in the universe travels about 300,000 kilometers in a direction that we really cannot perceive. This is why the speed of light is so special. It is woven into the fabric of the cosmos. And, in each person’s own rest frame, this motion at the speed of light is the same, 24 hours per day, 7 days per week, 52 weeks per year, waking or sleeping, and no matter how fast he is moving with respect to other persons.

Minkowski’s metric explained the constancy of the speed of light in the following way. If light particles travel past us along some straight path ℓ in space, their speed $v = \Delta \ell / \Delta t$ will be measured by us to be the speed of light c . According to equation (6.40), their proper distance is then $\Delta s^2 = (-c^2 + v^2) \Delta t^2 = 0$. This “null” proper

distance traveled that we measure for those light particles should be the same for any other observer traveling at any other speed and measuring Δs for those same light particles. That is, another observer should also obtain $0 = \Delta s^2 = (-c^2 + (v')^2)(\Delta t'^2)$ and also conclude that the particle's speed is $v' = c$. The speed of light particles, therefore, is constant no matter who is measuring the speed of those particles. This is also true for other particles traveling at c . It is not light itself that is special, it is the speed with which it travels.

6.4.2 Time Dilation

Traveling through space in a spaceship does not change our personal perception of our own motion through time, but it *does* change *others'* perception of our motion through time, if they are in another spaceship traveling at a different speed. For us, the proper distance between the seat of our chair at time t and at time $t + \Delta t$, as measured by us, is given by equation (6.42). That is, from our point of view, the chair has not moved in our space, so its only motion is the motion through time. However, for people in another spaceship, who see us moving in their reference frame in the direction of some straight line ℓ' , the proper distance between the seat of our chair at time t' and at time $t' + \Delta t'$ in their reference frame is given by equation (6.40) with $\Delta \ell' = v' \Delta t'$

$$\Delta s^2 = -c^2 (\Delta t')^2 + v'^2 (\Delta t')^2 = -c^2 \left(1 - \frac{v'^2}{c^2}\right) (\Delta t')^2 \quad (6.43)$$

But, since the proper distance Δs should be the same in both frames, we can equate (6.42) and (6.43) to obtain

$$\Delta t' = \gamma' \Delta t \quad (6.44)$$

where the “Lorentz factor” γ' (always greater than or equal to 1) is given by

$$\gamma' \equiv \left(1 - \frac{v'^2}{c^2}\right)^{-\frac{1}{2}}$$

That is, as the speed between our two spaceships begins to approach c , the people in the other spaceship perceive *our* clocks to take longer to tick by a factor γ' . This effect is called “time dilation” (lengthening of time) and is very important for the study of jets in radio sources and γ -ray burst sources.

6.4.3 The FitzGerald Contraction

The principle of relativity also states that people in both spaceships must measure their relative speed to be the same. That is, the magnitude of the speed of the other

spaceship v , as measured by us as it passes by, must equal the speed that they measure for us v' . Seeing that our clocks tick more slowly, but knowing that our speed is still v' , the people in the other spaceship conclude that in our reference frame, our spaceship must have traveled only a distance

$$\Delta \ell = v' \Delta t = \frac{v' \Delta t'}{\gamma'} = \frac{\Delta \ell'}{\gamma'} \quad (6.45)$$

in the time interval Δt that we measure. So, to the other space travelers, it looks as if our distances are all shorter by a factor of $1/\gamma'$. This is called the FitzGerald contraction, after the Irish physicist George FitzGerald, who proposed this process in 1889. Of course, from our point of view, our distances have not changed.

6.4.4 Minkowski Metric is the Same for All Travelers

But, does this work in reverse? By the principle of relativity, *we also* should see the clocks in the *other* spaceship tick more slowly – the opposite to what they perceive – and causing us to think that their distances have shortened in their direction of motion. Our calculation for the equivalent factor γ proceeds in the same way as for γ' above, but *we* now follow a chair in *their* spaceship. Again, to the occupants of the other spacecraft, their chair does not move in space, so the proper distance moved according to them is only

$$(\Delta S')^2 = -(c \Delta T')^2$$

just like equation (6.42), but now with capital letters because we are computing a different problem. But for us, the chair moves a proper distance

$$(\Delta S')^2 = -c^2 \left(1 - \frac{v^2}{c^2}\right) \Delta T'^2$$

so we now find that the time that we measure for their ticking clocks ($\Delta T'$) is a factor

$$\gamma \equiv \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}} \quad (6.46)$$

greater than what they would measure ($\Delta T'$). We also find a similar result for the FitzGerald contraction: a distance $\Delta L'$ in the other ship (along the direction of motion) is measured by us to be $\Delta L = \Delta L'/\gamma$. These time dilation and length contraction results are the same equivalent, but opposite, *relative* ones that the other ship perceives, if the speeds v and v' (i.e., γ and γ') that we measure are the same.

These strange effects actually have real-life consequences when we study relativistic jets in quasars and, particularly, in γ -ray bursts. An observer traveling with a $\gamma = 300$ jet (which has only a finite length and lifetime) would see it as a long

spindle of plasma. However, observers on earth see it as a flying pancake (or, more precisely, a *series* of flying pancakes, as it is broken up into multiple pieces) [249].

We see, then, that the Minkowski metric has several important properties:

1. It works for observers moving at any speed up to c .
2. Equation (6.41) is the same for travelers moving at different speeds.
3. It is an extension of the Euclidean metric, although definitely not simply four-dimensional Euclidean.
4. The true fabric of the universe includes both space and time.

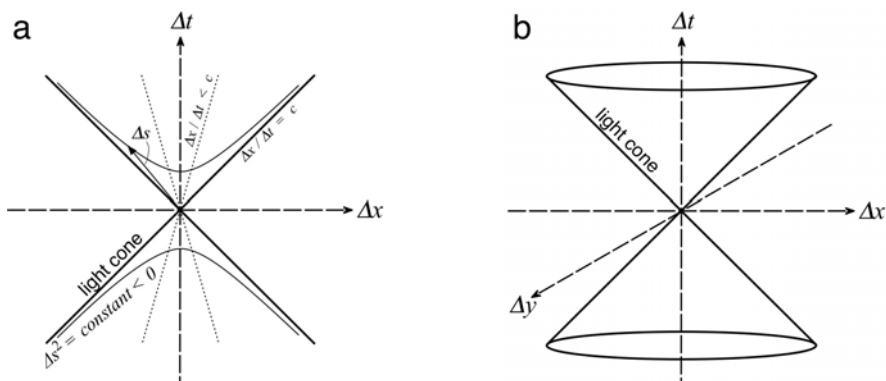


Fig. 6.5: Light cone in a 1+1 Minkowski metric (one space plus one time dimension) at left and in a 2+1 metric at right. The light cone also is called the “null surface”, because the line element $\Delta s = 0$ there. This cone is a limiting case for two different curves: constant velocity (dotted lines) and uniform proper distance $\Delta s^2 = \text{constant}$ (thin hyperbolic lines).

6.4.5 Working with Minkowski Geometry: Spacetime Diagrams and Light Cones

Much of the nature of the Minkowski metric can be understood by returning to two dimensions, but with only one of those dimensions being a spatial one (say, Δx); the other one will be time (Δt). Fig. 6.5a shows a simple spacetime diagram, similar to the two-dimensional planar diagrams in Pythagorean geometry in Figs. 6.1 and 6.2, but with time along the vertical axis instead of Δy . The origin at (0,0) represents “now”; positive Δt points toward the *future*; and negative Δt points into the *past*. Straight lines that intersect the origin represent constant velocity motion that either impacts us from the past (like an object thrown *at* us) or begins now and recedes away from us into the future (like an object thrown *away by* us). The speed of light is drawn at a 45° angle in both the $+\Delta x$ and $-\Delta x$ directions. If we were to add

another spatial dimension coming out of the paper in Fig. 6.5b (Δy , for example), the speed of light would be drawn as a cone in our three-dimensional (Δt , Δx , Δy) diagram rather than an open triangle. Therefore, the 45° angle lines representing the speed of light in Fig. 6.5a and b are often called the “light cone”, even though the first figure is a two-dimensional one. The metric in this 2-D spacetime diagram is given by equation (6.41) with $\Delta y = \Delta z = 0$

$$\Delta s^2 = -c^2 \Delta t^2 + \Delta x^2 \quad (6.47)$$

Equation (6.2) tells us that lines of constant distance Δh from some origin are circles in two dimensions (or spheres in three). However, from equation (6.47) we see that in Minkowski geometry the lines of constant distance Δs are hyperbolas. In fact, the light cone is a special limiting case of an hyperbola (an open triangle) that touches the origin. Everything along the light cone has $\Delta x = \pm c \Delta t$ or $\Delta s = 0$. This zero Δs all along this line is responsible for another name often used for the light cone: the “null surface”, where here null means zero proper distance Δs .

6.4.6 The Meaning of the Proper Distance Δs : Proper Time

What is the meaning of the proper distance Δs , and how can it be that light particles that are seemingly very far from us (in Δx) have zero proper distance Δs ? The first thing to remember is that points in spacetime are not points in space; they are *events*, each separated from us in time as well as in space. As we see below, the proper distance in spacetime, then, is actually the *time* between two events, as measured by some observers, like ourselves.

The best way to see events on your own light cone is to go outside at night, look up at the sky, and take in all the things happening in one brief moment. The light arriving right now from each meteor blazing in the earth’s atmosphere, each comet crashing into a planet in the solar system, each distant pulsating star, and each exploding quasar *that you see* is affecting you right at that moment. Of course, the *spatial* distance of each of these events is different. Some are thousands to billions of light years away; others are very nearby. But for all of these events, the *spacetime* distance Δs from you is zero because they all are happening to you right now.

Any events inside your light cone (closer to the Δt axis in Fig. 6.5) have $\Delta s^2 < 0$. They can affect what happens now if they are events in the past ($\Delta t < 0$) or they can be affected by what you do now if they are events in the future ($\Delta t > 0$). A line connecting your position now with those events has a slope $\Delta x / \Delta t$ that is slower than the speed of light. Because those lines are nearer to the Δt axis than the light cone, they are called “timelike” lines or “timelike geodesics”.

Events outside the light cone (nearer the Δx axis) cannot affect you now, but they could do so in the future as your light cone marches forward in time. These events have $\Delta s^2 > 0$, and straight lines that connect them to the origin are called “spacelike geodesics”. An example of an event that is “spacelike related” would be

a supernova explosion occurring at the present time in a galaxy that is ten million light years away. We do not see it now, but in ten million years we will. (Or, rather, it will be seen by those who are on the earth at that time.)

Therefore, Δs^2 is related to the time difference between when an event occurs in our reference frame and when it affects (or is affected by) us. We will call that time difference the “proper time” $\Delta\tau$. In fact, because Δs^2 is negative, this time difference is given by

$$\Delta\tau^2 = -\Delta s^2/c^2 \quad (6.48)$$

so, within a factor of c , the proper time $\Delta\tau$ and the proper distance Δs are the same. Outside our light cone, events have imaginary $\Delta\tau$, and inside our light cone these events are real. Events that occur along the hyperbolas in Fig. 6.5a all occur at the same proper time $\Delta\tau$ before or after our current time.

6.5 Mechanics in Four Dimensions

6.5.1 Mechanics of Single Particles in Four Dimensions

6.5.1.1 Four-Velocity

Because spacetime is four-dimensional, vectors like velocity therefore must have four components. But, because it would be a little awkward (although quite valid) for the time component of velocity to have different units (seconds per second) from the spatial components (centimeters per second), we shall define a new time coordinate w with units of distance

$$dw \equiv c \, dt \quad (6.49)$$

so that the Minkowski metric becomes (in differential form)

$$ds^2 = -dw^2 + dx^2 + dy^2 + dz^2 \quad (6.50)$$

or, in matrix form,

$$\mathbf{g} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (6.51)$$

and the four-velocity vector $\mathbf{U}$ must have four components

$$\mathbf{U} \equiv \begin{pmatrix} U^w \\ U^x \\ U^y \\ U^z \end{pmatrix} \quad (6.52)$$

Note that we now use the italic $\mathbf{g}$ for the four-metric that includes time, but we shall continue to use the roman $\mathbf{g}$ for two- and three-metrics that have only spatial components.

Earlier in this chapter, when we discussed three-dimensional Euclidean geometry, we recalled that the components of vectors can change with coordinate transformations, but the *magnitudes* of vectors remain unchanged. The same is true in four-dimensional Minkowski geometry

$$\mathbf{U}^2 = \mathbf{U}^T \cdot \mathbf{g} \cdot \mathbf{U} = -c^2 \quad (6.53)$$

That is, the square of the magnitude of the four-velocity for anyone at rest ($V^x = V^y = V^z = 0$) in the Minkowski metric (including us) is $-c^2$. Since the square magnitude of our velocity is the same for any coordinate frame, it *always* must be $-c^2$.

For someone in a spaceship moving past us with a three-velocity

$$\mathbf{V} = \begin{pmatrix} V^x \\ V^y \\ V^z \end{pmatrix}$$

in order to satisfy equation (6.53) the four-velocity must be

$$\mathbf{U} = \begin{pmatrix} \gamma c \\ \gamma V^x \\ \gamma V^y \\ \gamma V^z \end{pmatrix} \quad (6.54)$$

where γ is the Lorentz factor from equation (6.46). Equation (6.54) for $\mathbf{U}$ tells us several things. First, if we want to associate a speed V^w with our motion through time, that speed is *positive*

$$V^w = c$$

not $-c$ or ic . *We really are always moving in the w direction at the speed of light.* Second, there must be a corresponding velocity 1-form $\mathbf{u}$ to the velocity vector $\mathbf{U}$

$$\mathbf{u} = \mathbf{U}^T \cdot \mathbf{g} = (-\gamma c \quad \gamma V^x \quad \gamma V^y \quad \gamma V^z) \quad (6.55)$$

that contains the minus sign in the time component. It is no longer the convention to set $V^w = ic$ from equation (6.53), nor does one use an imaginary coordinate $dw' = ic dt$ to make the Minkowski metric look Euclidean. These steps would mask the important null properties of the light cone and make the time coordinate w , and our constant motion through it, less real.

6.5.1.2 Lorentz Transformations

It is now apparent that, because U^2 remains the same value ($-c^2$) for everyone, no matter in what speed or direction a traveler doing the measurement is moving, a change in speed is nothing more than a coordinate transformation in four dimensions. The time coordinate changes, as does the spatial coordinate, in the direction of motion (x). But the metric (equation (6.50) or (6.51)) remains the same Minkowski one. So, in addition to the types of coordinate transformations that we already have seen in three-dimensional geometry (rotations, translations, and label or variable changes), in Minkowski space we now can add “boosts” (or instantaneous changes in velocity) as another. These are often called Lorentz transformations, after the famous physicist Hendrik Lorentz who, in 1895, proposed that time passed at different rates in different moving frames and, in 1904, developed a mathematical transformation to describe these changes. For a single boost in x , the four-dimensional transformation matrix is

$$\mathcal{L} = \begin{pmatrix} \gamma & -\frac{v}{c}\gamma & 0 & 0 \\ -\frac{v}{c}\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (6.56)$$

and its inverse

$$\mathcal{L}^{-1} = \begin{pmatrix} \gamma & \frac{v}{c}\gamma & 0 & 0 \\ \frac{v}{c}\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (6.57)$$

It is easy to show that, because of the definition of the Lorentz factor (equation (6.46)), for a Lorentz boost, the Minkowski metric (equation (6.51)) remains unchanged

$$g' = (\mathcal{L}^{-1})^T \cdot g \cdot \mathcal{L}^{-1} = g \quad (6.58)$$

Of course, we can boost in x , y and z , as well as rotate and change variables, all in one single operation if we use a complicated $\mathcal{L}$. Such complicated $\mathcal{L}$'s are called “generalized Lorentz transformations” and are simply general coordinate transformations in four-dimensional space.

6.5.1.3 Energy and Momentum: Conservation of Four-Momentum

Four-Momentum

Of course, if the magnitude of the four-velocity is conserved, then the magnitude of the four-momentum of a moving body with “rest mass” m_0

$$\mathbf{P} \equiv m_0 \mathbf{U} = \begin{pmatrix} \gamma m_0 c \\ \gamma m_0 V^x \\ \gamma m_0 V^y \\ \gamma m_0 V^z \end{pmatrix} \quad (6.59)$$

also must be conserved, because m_0 is just a constant. (Note that the rest mass is the mass a body has when that body is not moving.) If we set the velocity to zero in equation (6.59) and compute its magnitude using an equation like (6.53), we get

$$P^2 = -m_0^2 c^2 \quad (6.60)$$

The x , y and z components of the four-momentum look just like those of the old three-momentum we know from Newtonian mechanics (equation (5.1)), but with the mass of the body increased by the Lorentz factor

$$m = \gamma m_0 \quad (6.61)$$

This greater mass not only resists acceleration, it also exerts a greater gravitational pull, as we shall see in the next chapter. However, does the w component of the four-momentum have a counterpart in classical physics, or is it a new quantity? We can find out the answer by expanding γ for small velocities ($v \ll c$)

$$P^w = \gamma m_0 c \approx \frac{1}{c} \left(m_0 c^2 + \frac{1}{2} m_0 v^2 \right)$$

That is, the w component of the four-momentum appears to be $1/c$ times the sum of the kinetic energy associated with motion through space at velocity v and an energy

$$\mathcal{E}_0 = m_0 c^2 \quad (6.62)$$

associated with our motion through time at velocity c . So, for particles with mass m_0 , P^w appears to be the total energy of the body $\mathcal{E}$, including rest energy, divided by the speed of light. That is, *the time component of the four-momentum is simply the quantity we know from freshman physics as energy.*

We actually can generalize this concept for particles of any mass (including zero-mass photons) if we define

$$\mathbf{P} \equiv m_0 \mathbf{U} = \begin{pmatrix} \mathcal{E}/c \\ P^x \\ P^y \\ P^z \end{pmatrix} \quad (6.63)$$

which has a magnitude of $P^2 = -\mathcal{E}^2/c^2 + P^2$, where $P^2 \equiv (P^x)^2 + (P^y)^2 + (P^z)^2$ is the magnitude of the classical three-momentum. When we combine this with equation (6.60), we obtain the famous conservation law for the absolute value of the energy-momentum

$$\boxed{\mathcal{E}^2 = P^2 c^2 + m_0^2 c^4} \quad (6.64)$$

which is true not only for bodies with mass but also for those with $m_0 = 0$. Even massless photons still have momentum $\mathcal{E} = Pc$ because they have energy of motion.

Newton's Second Law of Motion and the Covariant Gradient

Finally, Newton's second law of motion (equation (5.2)) also has a four-dimensional counterpart

$$\frac{d\mathbf{P}}{d\tau} = \mathbf{F}$$

where $\mathbf{F}$ is the force four-vector and $d\tau = ds/c$ is the “proper time” interval. Written in a general covariant form, this equation of motion is

$$\mathbf{U}^T \cdot \nabla \mathbf{P} = \mathbf{F} \quad (6.65)$$

where the bold symbol

$$\nabla \equiv \left(\frac{\partial}{\partial w} \quad \frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \quad \frac{\partial}{\partial z} \right)$$

is the four-dimensional gradient vector in Cartesian coordinates. Note that, if we were working in a curvilinear coordinate system, rather than a Minkowskian one, the quantity $\nabla \mathbf{P}$ in equation (6.65) would have pseudo-force terms (centrifugal and Coriolis forces) in addition to the simple derivatives $\partial P^\alpha / \partial x^\beta$. These terms arise because the gradient of a vector ($\mathbf{P}$) includes derivatives of not only the components P^α but also of its unit vectors $\mathbf{e}_\alpha$ in the four curvilinear dimensions. In a general coordinate system, therefore, we will write the *covariant gradient* as

$$\nabla \equiv \left(\nabla_0 \quad \nabla_1 \quad \nabla_2 \quad \nabla_3 \right) \quad (6.66)$$

Appendix B presents the details of how to compute the covariant gradients (and divergences) of scalars, vectors, and tensors, and Chapter 7 shows how it is used to generate the gravitational “force” in a curved spacetime.

Four-Force

The three spatial components of $\mathbf{F}$ are, of course, the standard three-dimensional forces that we are familiar with. However, what is the “energy force” F^w ? Just as the three-dimensional force is composed of processes that change momentum, F^w is made up of those processes that change energy: conduction, convection, radiation, and advection. (Advection is the carrying-away of energy by moving particles, such as a person blowing on a bowl of soup or a star ejecting a powerful wind.)

6.5.2 Fluid Mechanics in Four Dimensions

A much more powerful form for the conservation laws, however, is the four-dimensional version of equation (5.66, without gravity)

$$\nabla \cdot \mathcal{T}^T = \nabla \cdot \mathcal{T} = 0 \quad (6.67)$$

(which can be written routinely without the transpose notation because $\mathcal{T}$ is symmetric). For example, for a perfect fluid with rest mass density ρ , isotropic pressure p , internal energy ε , and velocity $\mathbf{U}$, the four-dimensional stress-energy tensor is given by

$$\mathcal{T}_{\text{FL}} = \left[\rho + \frac{(\varepsilon + p)}{c^2} \right] \mathbf{U} \mathbf{U} + p \mathbf{g} = \begin{pmatrix} [\rho + \frac{(\varepsilon+p)}{c^2}] U^w U^w - p \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^x U^w \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^y U^w \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^z U^w \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^w U^x + p \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^x U^x + p \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^y U^x \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^z U^x \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^w U^y \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^x U^y \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^y U^y + p \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^z U^y \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^w U^z \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^x U^z \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^y U^z \\ [\rho + \frac{(\varepsilon+p)}{c^2}] U^z U^z + p \end{pmatrix} \quad (6.68)$$

(The time derivative, of course, becomes a w derivative in equation (6.67).) Just like $\mathbf{T}$, $\mathcal{T}$ also is symmetric, even in the wx , wy , and wz components. In the four-dimensional world, momentum density (the left column) and energy flux (the top row of $\mathcal{T}$) are equivalent. When expressed in a coordinate system that travels with the fluid [$\mathbf{U} = (c, 0, 0, 0)$], the perfect fluid stress-energy tensor reduces to the familiar form

$$\mathcal{T}_{\text{FL}} = \begin{pmatrix} (\rho c^2 + \varepsilon) & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix} \quad (6.69)$$

The stress on a fluid at rest is simply the pressure, and the fluid's energy is simply its rest mass-energy plus its internal energy.

6.5.3 The Doppler Effect, the Doppler Factor, and “Beaming”

So far we have been concerned with physical quantities that we measure (like space and time or energy and momentum) that depend *only* on the relative speed v between

two travelers. However, some things that we measure (like the pitch of a horn or the color and intensity of a laser pointer aimed at us from the other spaceship) depend not on the *total* speed, but only on the *component* of that speed that is approaching or receding away from us. That is, many things depend on $v \cos \theta$, where θ is the angle between the other travelers' velocity vector and our line of sight to them.

A good example is the Doppler effect. Everyone is familiar with the sound of an automobile horn as the vehicle passes by us on the street. As it approaches us the pitch sounds a little bit higher than it would if we were in the car, and after it passes the pitch abruptly drops. When the auto approaches, the sound waves emitted by the horn strike our ears at a slightly higher rate than if the car were not moving; and when the auto recedes, the waves come at a slightly lower rate because the car is pulling away. If f_0 is the actual pitch of the car's horn, then the pitch we hear is given by the equation

$$f = \delta f_0 \quad (6.70)$$

where δ is called the "Doppler factor", after the physicist Christian Doppler who discovered this effect in 1845. For sound waves, the Doppler factor is given by $\delta = 1 + (v/c_s) \cos \theta$, where c_s is the speed of sound in air and the angle θ ranges from 0, when exactly approaching us, to 180° (or π radians) when exactly receding from us. When the car's speed is exactly perpendicular to our line of sight ($\theta = 90^\circ$), $\cos \theta = 0$, and, for a brief moment, we hear the car's horn at the original pitch f_0 .

As long as the speed v remains much less than the speed of light c , the situation is similar for light waves, like those produced by a laser pointer, but with c_s replaced by c in the above expressions. However, if the vehicle is a spaceship traveling near the speed of light, rather than in a slow-moving automobile, then we know that the above simple expression for the Doppler effect cannot be correct. Of course, $\delta = 1 + (v/c) \cos \theta$ is zero when the spaceship is receding from us at $v = -c$, as we expect. However, laboratory experiments tell us that δ should become very large as v becomes close to $+c$, but here (when the spaceship is approaching us) δ only becomes 2. Furthermore, because of the time dilation effect, even when $\theta = 90^\circ$ we expect the frequency with which light waves hit us from the other spaceship to be reduced by a factor of $1/\gamma$, but here we just get $\delta = 1$. Clearly, the simple expression does not take relativistic motion into account.

The true Doppler factor that works for both slow and fast speeds is [59]

$$\delta = \frac{1}{\gamma [1 - (v/c) \cos \theta]} \quad (6.71)$$

When approaching ($\theta = 0$) or receding ($\theta = 180^\circ$), $\delta = [(1 \pm v/c) / (1 \mp v/c)]^{1/2}$ (which is approximately $(2\gamma)^{\pm 1}$ as $v \rightarrow c$); and when passing by ($\theta = 90^\circ$), $\delta = 1/\gamma$. That is, when neither approaching nor receding, the frequency of a wave still is reduced by $f = f_0/\gamma$. (This is the time dilation effect, equation (6.44).) For speeds near the speed of light, the difference between approaching and receding Doppler effects can be enormous. The ratio of frequencies f_+ and f_- in the two cases is

$$f_+/f_- \approx 4\gamma^2$$

Even for a moderate Lorentz factor of $\gamma = 10$ ($v = 0.995 c$) this ratio is 400.

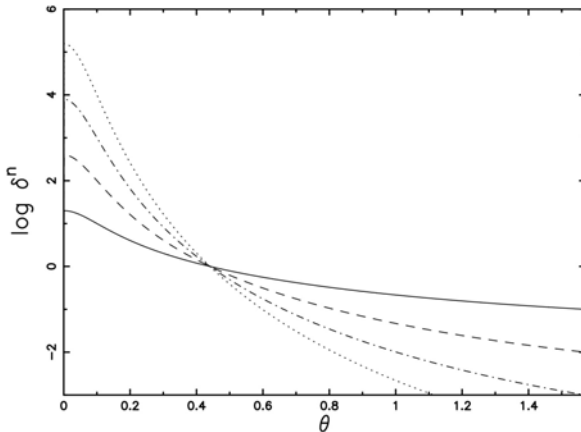


Fig. 6.6: Plot of the log of a $\gamma = 10$ relativistic jet's Doppler factor δ raised to a power n plotted against viewing angle θ . (Zero viewing angle has the jet pointed directly at the observer.) The curves are $n = 1$ (solid); $n = 2$ (dashed); and $n = 3$ (dash-dot); and $n = 4$ (dotted). Note that the latter two drop by an order of magnitude or more after only $\theta \approx 1/\gamma = 0.1$ radian.

The tremendous enhancement in the Doppler effect for relativistic motion is called “Doppler boosting” or “Doppler beaming”. It affects many physical quantities (besides just frequency of the emitted light or radio waves) that are measured when astronomers observe relativistic jet flow. Some, such as brightness, depend on the Doppler factor to a high power like 3 or 4. Fig. 6.6 shows a plot of different powers of the Doppler factor as a function of the jet viewing angle θ for $\gamma = 10$. The Doppler factor does not drop formally to a value of 1 until it reaches an angle given by $\cos \theta = [(\gamma - 1)/(\gamma + 1)]^{1/2}$ (or $\theta = 0.440$ radians for $\gamma = 10$). However, we see from the figure that the effective angle of the Doppler enhancement for an exponent of $n = 3$ or 4 is only about $\theta \approx \theta_b = 0.1$ radians, where

$$\theta_b \equiv 1/\gamma \tag{6.72}$$

is the effective “beaming angle”. For travelers, or extragalactic radio jets, moving very near the speed of light with respect to us, it is very difficult to see them if they are moving at an angle $\theta > \theta_b$ to our line of sight.

6.6 Electricity and Magnetism in Four Dimensions

6.6.1 Review of Electricity and Magnetism in the 3+1 Form

The quest by physicists in the 19th century to understand the electromagnetic field helped to create the theory of relativity in the early 20th century. And relativistic E & M, in turn, set the stage for developing gauge theories of all other forces in nature – the classical theory of gravity (general relativity) and relativistic quantum theories of the weak and strong nuclear forces. The basic properties of all physical fields (sources, field equations, field constraints, gauge conditions, and the possibility of plane wave solutions) have analogies in the electromagnetic field and are fairly easy to understand in that context. So, discussing the electromagnetic field not only will help us to understand accretion disks and jets in Chapters 12, 14 and 15, it also will help us to understand waves generated by the gravitational field (Chapter 8).

We therefore first will review classical E & M in this chapter, in a manner that can be understood by upper division undergraduate physics majors, but with special emphasis on those properties that have a direct analogy in general relativity and the study of black holes. In this first subsection we will use the “3+1” approach, with derivative operators ∇ and $\partial/\partial t$, rather than the four-gradient ∇ .

6.6.1.1 Maxwell’s Equations for the Field Components

In Gaussian/cgs units Maxwell’s field equations are [306]

$$\nabla \cdot \mathbf{B} = 0 \qquad \frac{\partial \mathbf{B}}{c \partial t} + \nabla \times \mathbf{E} = 0 \qquad (6.73)$$

$$\nabla \cdot \mathbf{D} = 4\pi \rho_q \qquad -\frac{\partial \mathbf{D}}{c \partial t} + \nabla \times \mathbf{H} = \frac{4\pi}{c} \mathbf{J} \qquad (6.74)$$

where $\mathbf{B}$ and $\mathbf{E}$ are the magnetic induction and electric field, ρ_q is the local charge density, $\mathbf{H}$ and $\mathbf{D}$ are the magnetic field and electric displacement, and $\mathbf{J}$ is the current density three-vector. We will assume here, and throughout this book, that the electric permittivity and magnetic permeability are unity – essentially those of free space. So, $\mathbf{D} = \mathbf{E}$ and $\mathbf{H} = \mathbf{B}$ in this discussion.

6.6.1.2 Evolution Equations vs. Constraint Equations and the Generation of Conservation Laws

A glance at equations (6.73) and (6.74) reveals a small puzzle. They comprise 8 total equations, but have only 6 unknowns (the components of $\mathbf{E}$ and $\mathbf{B}$). However, a deeper examination shows that the 8 equations are not independent. Indeed, if we take the divergence of the homogeneous vector equation on the right in (6.73), we

obtain the following simplified version

$$\frac{\partial(\nabla \cdot \mathbf{B})}{\partial t} = 0$$

because $\nabla \cdot (\nabla \times \mathbf{V}) = 0$ for any vector $\mathbf{V}$. That is, if the solenoidal constraint equation ($\nabla \cdot \mathbf{B} = 0$) is true at some time in the evolution of the field, then Faraday's evolution equation ($\partial \mathbf{B} / c \partial t + \nabla \times \mathbf{E} = 0$) will ensure that the solenoidal constraint will be met *for all time* and for any electric field $\mathbf{E}$. This is called “constrained transport” of the magnetic field. *The solenoidal constraint is not independent of the evolution equation, so there really are only the three independent Faraday's law equations to determine the three components of $\mathbf{B}$.*

In the case of the inhomogeneous equations (6.74), the same redundancy is true, but a new twist is added. The divergence of the right-hand equation (Ampère's law) gives

$$-\frac{\partial(\nabla \cdot \mathbf{D})}{\partial t} = 4\pi(\nabla \cdot \mathbf{J})$$

and a substitution from Gauss's law (the left equation of (6.74)) produces

$$-\frac{\partial \rho_q}{\partial t} = \nabla \cdot \mathbf{J} \quad (6.75)$$

This is simply the law of conservation of charge. *So, the electric constraint equation is not independent of the corresponding evolution equation law either, if charge is conserved.* We again have only Ampère's three independent evolution equations for the three components of $\mathbf{E}$. With 6 equations and 6 unknown fields, all seems well now.

There is a very important lesson to be learned here. By eliminating the two constraint equations, we have gained two conservation laws: the conservation of electric charge (6.75) and the conservation of magnetic monopoles (which seem to be zero in number in our universe). That is, *redundant field constraint equations are really conservation laws for the field sources and not field equations at all.* This is a universal and fundamental property of *all* gauge fields (electromagnetic, gravitational, quantum-chromodynamic, *etc.*).

6.6.1.3 Energy and Momentum Conservation for Single Charged Particles

Electric and magnetic fields can modify the energy and momentum of charged particles. In three dimensions these processes add electromagnetic terms to the energy and momentum evolution equations that are described by the familiar expressions for work done by a single particle of charge q on (or by) the electric field and for the Lorentz force on that charged particle

$$\left[\frac{d\mathcal{E}}{dt} \right]_{\text{EM}} = q \mathbf{V} \cdot \mathbf{E} \quad (6.76)$$

$$\mathbf{F}_{\text{EM}} = q \left(\mathbf{E} + \frac{\mathbf{V}}{c} \times \mathbf{B} \right) \quad (6.77)$$

In Newtonian mechanics, equation (6.77) provides an additional force term on the right-hand side of equation (5.2), allowing both gravitational and electromagnetic fields to affect the motion of the particle.

6.6.1.4 Energy and Momentum Conservation for a System of Charged Particles

When we are dealing with a system of particles (a fluid or plasma), these equations are expressed as energy change or force *per unit volume*, and involve the current and charge density. They are the Joule heating and Lorentz force terms

$$\rho \dot{q} = \mathbf{J} \cdot \mathbf{E} \quad (6.78)$$

$$\mathfrak{F}_{\text{EM}} = \rho_q \mathbf{E} + \frac{1}{c} \mathbf{J} \times \mathbf{B} \quad (6.79)$$

where $\dot{q}$ is the local heating (or cooling) rate per unit mass (as defined after equation (5.50)), and $\mathfrak{F}$ denotes vector force $\mathbf{F}$ per unit volume. These equations express the conservation laws of energy and momentum as the electromagnetic field interacts with the surrounding matter. In Newtonian mechanics, these provide additional heating and force terms on the right-hand side of equations (5.50) and (5.48), respectively.

6.6.1.5 Maxwell's Equations in Potential Form and the Concept of Gauge Freedom

Equations (6.73) are often expressed in an alternate form by defining a scalar electric potential (Φ) and a vector magnetic potential ($\mathbf{A}$) that automatically satisfy equations (6.73)

$$\mathbf{B} = \nabla \times \mathbf{A} \quad \mathbf{E} = -\nabla \Phi - \frac{\partial \mathbf{A}}{c \partial t} \quad (6.80)$$

Substituting these expressions into equations (6.74) we obtain Maxwell's equations in "potential" form

$$\nabla^2 \Phi + \frac{\partial(\nabla \cdot \mathbf{A})}{c \partial t} = -4\pi \rho_q \quad (6.81)$$

$$\nabla^2 \mathbf{A} - \frac{\partial^2 \mathbf{A}}{c^2 \partial t^2} - \nabla \left(\nabla \cdot \mathbf{A} + \frac{\partial \Phi}{c \partial t} \right) = -\frac{4\pi}{c} \mathbf{J} \quad (6.82)$$

So we now have four unknowns (Φ , $\mathbf{A}$) and four equations. Is everything okay? *No*. As before, equation (6.81) is a constraint equation and is therefore not independent of equation (6.82).

The introduction of the potentials has led to more unknowns (4) than independent equations (3). In this case we are completely free to choose another *additional* (scalar) relation between Φ and $\mathbf{A}$ to replace equation (6.81) in order to get a complete set of 4 equations. This freedom to choose any relation that we want is called “gauge freedom”, and the specific choice that we decide to make is called the “gauge”. No matter how we construct Φ and $\mathbf{A}$, however, the fields themselves will behave the same and, therefore, will be “invariant” under a gauge transformation.

6.6.1.6 The Lorenz Gauge

The choice of gauge that will be the most useful here, and in our study of gravitational waves, is the one suggested by the Danish physicist Ludwig Lorenz in the mid 1800s.⁴ If we choose

$$\nabla \cdot \mathbf{A} + \frac{\partial \Phi}{c \partial t} = 0 \quad (6.83)$$

(or equals any constant other than zero), then equations (6.81) and (6.82) then *do* become independent (wave) equations

$$-\frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} + \nabla^2 \Phi = -4\pi \rho_q \quad (6.84)$$

$$-\frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} + \nabla^2 \mathbf{A} = -\frac{4\pi}{c} \mathbf{J} \quad (6.85)$$

with the conservation of charge now embodied in equation (6.83) itself.

Equation (6.83) is often referred to as the “Lorenz gauge” or the “Lorenz gauge condition”. In fact, there are an infinite number of Lorenz gauges, since we can choose an arbitrary function Z and construct a new set of potentials

$$\Phi' = \Phi + \frac{\partial Z}{c \partial t} \quad (6.86)$$

$$\mathbf{A}' = \mathbf{A} + \nabla Z \quad (6.87)$$

and still satisfy Maxwell’s equations *and* the Lorenz condition with Φ replaced by Φ' and $\mathbf{A}$ replaced by $\mathbf{A}'$ in equation (6.83). This remains true for any Z so long as that function satisfies its *own* homogeneous wave equation

$$-\frac{\partial^2 Z}{c^2 \partial t^2} + \nabla^2 Z = 0 \quad (6.88)$$

⁴ Lorenz suggested his gauge when the Dutch physicist Hendrik Lorentz was only a young teen and decades before the latter developed the concept of Lorentz invariance. However, because of the similarity in their names, Lorentz is often erroneously credited with this gauge instead of Lorenz.

6.6.1.7 Linearly Polarized Plane Electromagnetic Waves in a Vacuum

Studying the behavior of E & M waves propagating through a region with no sources ($\mathbf{J} = 0$, $\rho_q = 0$) will help greatly in the understanding of gravitational waves later on. As is customary, we will study waves using the complex form for a single oscillating electromagnetic field potential with angular frequency ω and propagation vector $\mathbf{k} = k \hat{\mathbf{e}}_z$ in the z direction

$$\Phi = \Re \{ \varphi \exp [i(kz - \omega t)] \} \quad (6.89)$$

$$\mathbf{A} = \Re \{ \mathcal{A} \exp [i(kz - \omega t)] \} \quad (6.90)$$

where $\Re$ is the operator that extracts the real part of the complex exponential expression. Note that φ and $\mathcal{A}$ are *constant* amplitudes; *all the spatial and time-dependence of Φ and $\mathbf{A}$ are contained in the complex sinusoidal exponential factors*. Furthermore, we assume also that φ and $\mathcal{A}$ are *real* values, which restricts our discussion here to linear polarization only. Finally, with no x and y dependence, equations (6.89) and (6.90) describe an infinite plane wave propagating in the z direction.

With no sources, substituting equations (6.89) and (6.90) into equations (6.84) and (6.85) gives, simply, the dispersion relation

$$\omega^2 = k^2 c^2$$

and the Lorenz condition becomes a condition on the component of the electromagnetic potential along the propagation vector

$$\mathcal{A}^z = \pm \varphi$$

Furthermore, equation (6.88) indicates that the Lorenz gauge function also is sinusoidal

$$Z = \Re \{ \mathcal{Z} \exp [i(kz - \omega t)] \}$$

and the gauge transformations become conditions on the amplitude components

$$\begin{aligned} \varphi' &= \varphi \pm i k \mathcal{Z} \\ \mathcal{A}'^x &= \mathcal{A}^x \\ \mathcal{A}'^y &= \mathcal{A}^y \\ \mathcal{A}'^z &= \mathcal{A}^z \pm i k \mathcal{Z} \end{aligned}$$

(which preserves the Lorenz condition, since $\mathcal{A}'^z = \pm \varphi'$). We see, therefore, that we can arbitrarily choose $\mathcal{A}^z$ (and therefore φ) but not $\mathcal{A}^x$ and $\mathcal{A}^y$.

What choice might we make for the scalar potential and the z component of the vector potential? Let us look at the electric and magnetic fields themselves. Equations (6.89) and (6.90) mean, of course, that the field functions are also sinusoidal

$$\mathbf{E} = \Re \{ \mathcal{E} \exp [i(kz - \omega t)] \} \quad (6.91)$$

$$\mathbf{B} = \Re \{ \mathcal{B} \exp [i(kz - \omega t)] \} \quad (6.92)$$

which lead to expressions for their amplitudes from the definitions of the potentials (6.80)

$$\begin{aligned} \mathcal{E}^x &= i \frac{\omega}{c} \mathcal{A}^x & \mathcal{B}^x &= -i k \mathcal{A}^y \\ \mathcal{E}^y &= i \frac{\omega}{c} \mathcal{A}^y & \mathcal{B}^y &= i k \mathcal{A}^x \\ \mathcal{E}^z &= 0 & \mathcal{B}^z &= 0 \end{aligned}$$

The fields, therefore, depend only on the values of the constant potential amplitudes $\mathcal{A}^x$ and $\mathcal{A}^y$, so we are then free to choose

$$\mathcal{A}^z = \pm \varphi = 0 \quad (6.93)$$

as our specific Lorenz gauge. This is called the *transverse* Lorenz gauge, because the fields *and* the potentials in the longitudinal (z) direction are all zero.

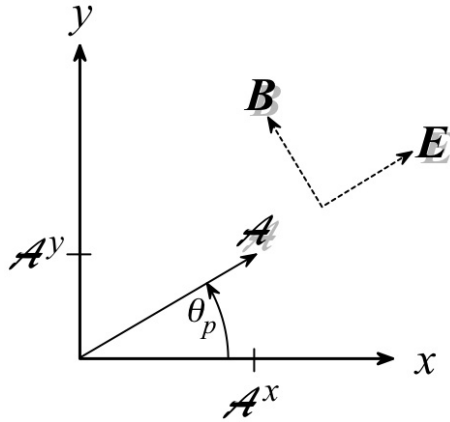


Fig. 6.7: Illustration of the *mixed* polarization case ($\mathcal{A}^x \neq 0, \mathcal{A}^y \neq 0$) for a plane electromagnetic wave traveling out of the page toward the reader. The electric and magnetic vectors are plotted at zero phase (e.g., $z = 0, t = 0$). (They, of course, oscillate with time and distance z according to eqs. (6.91) and (6.92).) The polarization angle θ_p is defined as the angle between $\mathbf{E}$ and the x axis at zero phase.

The two amplitudes $\mathcal{A}^x$ and $\mathcal{A}^y$ represent the two (pure) *polarization modes* possible for a propagating plane electromagnetic wave. If $\mathcal{A}^x > 0$ and $\mathcal{A}^y = 0$, then $\mathcal{E}$ will lie along $+x$ and $\mathcal{B}$ will lie along $+y$ (see Fig. 6.7). On the other hand, if $\mathcal{A}^x = 0$ and $\mathcal{A}^y < 0$, then $\mathcal{E}$ will lie along $-y$ and $\mathcal{B}$ along the $+x$ axis. Waves of mixed polarization will be rotated by the polarization angle

$$\theta_p = \arctan \left(\frac{\mathcal{A}^y}{\mathcal{A}^x} \right)$$

and have an amplitude at that angle of

$$\mathcal{A} = \left([\mathcal{A}^x]^2 + [\mathcal{A}^y]^2 \right)^{1/2}$$

However, $\mathcal{E}$ and $\mathcal{B}$ always will be perpendicular to each other.

Finally, the Poynting energy flux per unit area of the wave always will be in the direction of the propagation of the wave

$$\begin{aligned} \mathbf{S} &= \frac{c}{4\pi} \langle \mathbf{E} \times \mathbf{B} \rangle_t \\ &= \frac{c}{8\pi} \mathcal{E} \times \mathcal{B}^* \\ &= \frac{\omega^2}{8\pi c} \mathcal{A}^2 \hat{\mathbf{e}}_z \end{aligned} \quad (6.94)$$

where $\langle \rangle_t$ is an average of the squared sinusoidals over time (which produces the factor of $1/2$ above), and $\mathcal{B}^*$ is the complex conjugate of the $\mathcal{B}$ in equation (6.92).

6.6.1.8 Electromagnetic Waves Generated by Time-Varying Charge and Current Distributions

General Green's Function Dipole Solution

We are now, finally, able to describe the character (particularly the amplitude and power loss) of waves generated by a time-varying charge distribution. For this we will use the *inhomogeneous* Maxwell's equations (*with* charge and current) in equations (6.84) and (6.85), which have the general Green's function solutions using the retarded (delayed) charge and current distributions [306]

$$\begin{aligned} \Phi(\mathbf{x}, t) &= \int d^3\mathbf{x}' \frac{\rho_q(\mathbf{x}', t - r'/c)}{r'} \\ \mathbf{A}(\mathbf{x}, t) &= \frac{1}{c} \int d^3\mathbf{x}' \frac{\mathbf{J}(\mathbf{x}', t - r'/c)}{r'} \end{aligned}$$

where $r' \equiv |\mathbf{x} - \mathbf{x}'|$ is the distance between the observer at $\mathbf{x}$ and a part of the source at $\mathbf{x}'$.

Let us ignore $\Phi(\mathbf{x}, t)$ for the moment (we will not need it) and concentrate on $\mathbf{A}(\mathbf{x}, t)$. The factor $1/r'$ can be expanded in Legendre polynomials about $1/r$

$$\frac{1}{r'} = \sum_{n=0}^{\infty} \frac{1}{r^{n+1}} r'^n P_n \left(\frac{\mathbf{x} \cdot \mathbf{x}'}{r r'} \right)$$

or

$$\frac{1}{r'} = \frac{1}{r} + \frac{\mathbf{x} \cdot \mathbf{x}'}{r^3} + \frac{3}{2} \frac{\mathbf{x} \cdot (\mathbf{x}' \mathbf{x}' - r'^2 \mathbf{I}/3) \cdot \mathbf{x}}{r^5} + \dots$$

where $r \equiv |\mathbf{x}'|$, $\mathbf{x}' \mathbf{x}'$ is a dyadic product, and $\mathbf{I}$ is the 3×3 identity tensor. The first term on the right is the *monopole* term, the second the *dipole*, and the third is the (“reduced”) *quadrupole*. Keeping only the term of lowest order in $1/r$ (the *far field* or radiation zone approximation), we find that

$$\mathbf{A}(\mathbf{x}, t) = \frac{1}{cr} \int d^3 \mathbf{x}' \mathbf{J}(\mathbf{x}', t - r'/c) \quad (6.95)$$

Now comes a trick of mathematics and physics that converts this integral for $\mathbf{A}(\mathbf{x}, t)$ into the simple and familiar dipole formula. If we first apply Gauss’s theorem

$$\int \nabla \cdot [\mathbf{J}(\mathbf{x}', t - r'/c) \mathbf{x}'] d^3 \mathbf{x}' = \oint \mathbf{J}(\mathbf{x}', t - r'/c) \mathbf{x}' \cdot d\mathbf{S}' = 0$$

(which vanishes because of the factor of $\mathbf{x}'$ and because the current source is confined to a small region and vanishes at infinity faster than $1/|\mathbf{x}'|$), and if we then apply the chain rule for the gradient operator to the left side of the above equation

$$\begin{aligned} \int \nabla \cdot [\mathbf{J}(\mathbf{x}', t - r'/c) \mathbf{x}'] d^3 \mathbf{x}' &= \int \mathbf{x}' [\nabla \cdot \mathbf{J}(\mathbf{x}', t - r'/c)] d^3 \mathbf{x}' \\ &\quad + \int \mathbf{J}(\mathbf{x}', t - r'/c) d^3 \mathbf{x}' \end{aligned}$$

we then obtain

$$\mathbf{A}(\mathbf{x}, t) = -\frac{1}{cr} \int \mathbf{x}' [\nabla \cdot \mathbf{J}(\mathbf{x}', t - r'/c)] d^3 \mathbf{x}' \quad (6.96)$$

from equation (6.95). Finally, if we apply the conservation of charge

$$\nabla \cdot \mathbf{J}(\mathbf{x}', t - r'/c) = -\frac{\partial \rho_q(\mathbf{x}', t - r'/c)}{\partial t} \quad (6.97)$$

we then can write the equation (6.96) for $\mathbf{A}$ as

$$\mathbf{A}(\mathbf{x}, t) = \frac{1}{cr} \frac{d(\mathfrak{D})}{dt} \quad (6.98)$$

where $\mathfrak{D}$ is the time-dependent, *retarded* dipole moment of the *charge* distribution

$$\mathfrak{D} = \int \mathbf{x}' \rho_q(\mathbf{x}', t - r'/c) d^3 \mathbf{x}' \quad (6.99)$$

(The partial time derivative becomes a total time derivative because the dipole moment $\mathfrak{D}$ is independent of x .)

Poynting Flux and Power Generated by a Simple Oscillating Dipole

For example, for a simple sinusoidally-oscillating dipole we have

$$\mathfrak{D} = \Re \{ \mathfrak{D} \exp [-i\omega(t - r/c)] \}$$

which gives a sinusoidally-oscillating potential of

$$\mathcal{A} = \Re \{ \mathcal{A} \exp [-i\omega(t - r/c)] \}$$

From equation (6.98), the wave amplitude vector, *in the frame of the source*, is

$$\mathcal{A} = \frac{\omega}{c r} \mathfrak{D}$$

Now, if we choose to describe the wave in the *observer's* frame ($\mathcal{A}'$, $\mathfrak{D}'$, etc.) with the transverse Lorenz gauge (equation (6.93)), we will need neither Φ nor the component of $\mathcal{A}$ in the direction of propagation of the wave. We will need only the projections of the potential amplitude $\mathcal{A}'$, and therefore the dipole amplitude $\mathfrak{D}'$, in the observer's x and y directions. In the observer's frame, then, the projected wave amplitude vector will be

$$\mathcal{A}' = \frac{\omega}{c r} \mathfrak{D}' \quad (6.100)$$

with

$$\mathfrak{D}' = (\mathfrak{D}'^x \ \mathfrak{D}'^y \ 0)$$

and

$$\begin{aligned} \mathfrak{D}'^x &= \mathfrak{D} \sin \theta \cos \alpha \\ \mathfrak{D}'^y &= \mathfrak{D} \sin \theta \sin \alpha \end{aligned} \quad (6.101)$$

where $\mathfrak{D} = |\mathfrak{D}|$ is the absolute value of the total dipole vector amplitude (in the frame of the source), θ (given by $\cos \theta = \mathbf{x} \cdot \mathfrak{D} / r \mathfrak{D}$) is the angle between the dipole vector amplitude and the radial vector from the dipole to the observer, and α is the arbitrary angle of rotation of the x - y axes about the propagation direction relative to the projection of the dipole vector in the x - y plane. (When $\alpha = 0$, the projected dipole lies along the observer's x axis.)

From equations (6.94), (6.100), and (6.101), the Poynting power flow density in the wave passing by the observer is

$$S^{EM} = \frac{\omega^2}{8\pi c} \mathcal{A}'^2 \quad (6.102)$$

$$= \frac{\omega^4}{8\pi c^3 r^2} \left([\mathcal{D}'^x]^2 + [\mathcal{D}'^y]^2 \right) \quad (6.103)$$

$$= \frac{\omega^4}{8\pi c^3 r^2} \mathcal{D}^2 \sin^2 \theta \quad (6.104)$$

Because of the $\sin^2 \theta$ dependence, the radiation pattern is not isotropic in the frame of the oscillating dipole. Integrating the Poynting flux over a sphere around the dipole at a radius r gives the total power emitted

$$\mathcal{P}^{EM} = \frac{\omega^4}{3c^3} \mathcal{D}^2 \quad (6.105)$$

which is only 2/3 of what would have been obtained if the dipole had radiated at peak intensity into a full 4π steradians.

General Dipole Formulae

Because equation (6.104) represents only one Fourier component ω of a possible spectrum of oscillating dipoles in a source, we can write general expressions for the Poynting flux and the power output as

$$S^{EM} = \frac{1}{4\pi c^3 r^2} \left\langle \left| \frac{d^2 \mathfrak{D}}{dt^2} \right|^2 \right\rangle_t \quad (6.106)$$

$$\mathcal{P}^{EM} = \frac{2}{3c^3} \left\langle \left| \frac{d^2 \mathfrak{D}}{dt^2} \right|^2 \right\rangle_{t\Omega} \quad (6.107)$$

where the averages $\langle \rangle$ are over solid angle and/or several oscillation cycles. Note that the factors of $1/4\pi$ and $2/3$ in the above equations are, indeed, correct: the extra factor of $1/2$ in equations (6.104) and (6.105) arises *after* we take the time average of the sinusoidally-oscillating part of $|\mathfrak{D}|$.

6.6.2 Electricity and Magnetism in Four-Dimensional, Covariant Form

The preceding section on E & M is, indeed, fully relativistic. However, it is not expressed in the language of four-dimensional spacetime. We therefore will recap the above discussion using covariant notation. This will be extremely important for understanding the equations of general relativity and gravitational waves later on.

6.6.2.1 Maxwell's Field Equations in Covariant, Tensor Form

The four-dimensional formulation of E & M begins by recognizing that the three-dimensional fields are actually parts of a four-dimensional antisymmetric electro-magnetic tensor called the Faraday tensor. In the rest frame (*i.e.*, for zero velocity), this tensor and its corresponding covariant form are

$$\mathcal{F} = \begin{pmatrix} 0 & E^x & E^y & E^z \\ -E^x & 0 & B^z & -B^y \\ -E^y & -B^z & 0 & B^x \\ -E^z & B^y & -B^x & 0 \end{pmatrix} \quad (6.108)$$

$$f \equiv g \cdot \mathcal{F} \cdot g = \begin{pmatrix} 0 & -E^x & -E^y & -E^z \\ E^x & 0 & B^z & -B^y \\ E^y & -B^z & 0 & B^x \\ E^z & B^y & -B^x & 0 \end{pmatrix} \quad (6.109)$$

Note that, while g and $\mathcal{T}$ are symmetric, $\mathcal{F}$ is antisymmetric and, therefore, has zeros on the diagonal. In this regard, f is a true 2-form. But, just like g and $\mathcal{T}$, $\mathcal{F}$ can be transformed to a new frame with a Lorentz transformation, giving us the electric and magnetic fields that would be measured by someone in that new frame (*cf.* equation (6.27) for contravariant tensors)

$$\mathcal{F}' = \mathcal{L} \cdot \mathcal{F} \cdot \mathcal{L}^T \quad (6.110)$$

For example, equation (6.110) correctly describes how a moving magnetic field can generate an electric field – an electromotive force or emf.

It is also useful here to define a related tensor, the Maxwell tensor, which is a re-worked version of the Faraday one

$$\mathcal{M} = \begin{pmatrix} 0 & -B^x & -B^y & -B^z \\ B^x & 0 & E^z & -E^y \\ B^y & -E^z & 0 & E^x \\ B^z & E^y & -E^x & 0 \end{pmatrix}$$

$$m = \begin{pmatrix} 0 & B^x & B^y & B^z \\ -B^x & 0 & E^z & -E^y \\ -B^y & -E^z & 0 & E^x \\ -B^z & E^y & -E^x & 0 \end{pmatrix}$$

Finally, we also can define a four-dimensional current density (and its corresponding 1-form), with the w component being the charge density ρ_q itself, *flowing in the time direction at the speed of light*

$$\mathbf{J} = \begin{pmatrix} \rho_q c \\ J^x \\ J^y \\ J^z \end{pmatrix} \quad \mathbf{j} = \begin{pmatrix} -\rho_q c & J^x & J^y & J^z \end{pmatrix} \quad (6.111)$$

Maxwell's eight equations (6.73) and (6.74), then can be conveniently written as two four-dimensional vector equations

$$\boxed{\nabla \cdot \mathcal{F}^T = \frac{4\pi}{c} \mathbf{J}^T} \quad (6.112)$$

$$\boxed{\nabla \cdot \mathcal{M}^T = 0} \quad (6.113)$$

Now, *any* antisymmetric tensor $\mathcal{F}$ automatically satisfies the four-dimensional identity

$$\nabla \cdot \left(\nabla \cdot \mathcal{F}^T \right)^T = 0 \quad (6.114)$$

This identity is closely related to the vector identities $\nabla \cdot \nabla \times \mathbf{A} = 0$ and $\nabla \times \nabla \Phi = 0$. Equations (6.112) and (6.114) immediately lead to a conservation law for charge

$$\nabla \cdot \mathbf{J} = 0 \quad (6.115)$$

which is the covariant form of equation (6.75). Because equation (6.114) is a *geometric* identity (called a *Bianchi* identity) that is valid for all antisymmetric tensors, it is much more fundamental than just a random property of the magnetic field. It is the ultimate reason for electromagnetism being a gauge theory. The fact that the E & M field can be expressed as an antisymmetric tensor (equation (6.108)) whose divergence is given by the source current (equation (6.112)), leads to an automatic satisfaction of the constraints with time (constrained transport), to gauge invariance, and to conservation of the sources (charge).

6.6.2.2 Conservation of Energy-Momentum for Single Charged Particles

The equations for the energy and momentum evolution of charged particles ((6.76) and (6.77)) are, in fact, covariant. So, the only change to the Newtonian equation of particle motion is in the definition of the momentum. The four-dimensional covariant form of the relativistic equations of motion, therefore, is equation (6.65), with the electromagnetic force term on the right-hand side

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}_{\text{EM}} = \frac{q}{c} \mathbf{U}^T \cdot \mathcal{F} \quad (6.116)$$

6.6.2.3 Conservation of Energy-Momentum for a Fluid in an Electromagnetic Field

The covariant form of equations (6.78) and (6.79), then, is

$$\boxed{\mathcal{F}_{\text{EM}} = \frac{1}{c} \mathbf{J}^T \cdot \mathcal{F}} \quad (6.117)$$

where $\mathcal{F}_{\text{EM}}$ is the four-dimensional electromagnetic force per unit volume. These forces, and how they interact with matter, can be included in a general conservation law like equation (6.67)

$$\nabla \cdot \mathcal{T}_{\text{FL}} = \mathcal{F}_{\text{EM}} \quad (6.118)$$

However, we can write equation (6.118) in a much simpler, and more powerful, form by defining an additional electromagnetic stress-energy tensor component given by

$$\boxed{\mathcal{T}_{\text{EM}} = \frac{1}{4\pi} \left[\mathcal{F} \cdot \mathbf{f}^T - \frac{1}{4} (\mathcal{F} : \mathbf{f}) \mathcal{I} \right] \cdot \mathbf{g}^{-1}} \quad (6.119)$$

where the double-dot product

$$\mathcal{F} : \mathbf{f} = \mathcal{F}^{ww} f_{ww} + \mathcal{F}^{wx} f_{wx} + \mathcal{F}^{wy} f_{wy} + \cdots + \mathcal{F}^{zz} f_{zz}$$

is the sum of the product of all corresponding tensor elements, and $\mathcal{I}$ is the *four*-dimensional identity tensor. With the electromagnetic stress-energy tensor written in this way, it can be shown that $\mathcal{F}_{\text{EM}}$ is actually equivalent to the (negative) divergence of $\mathcal{T}_{\text{EM}}$

$$\mathcal{F}_{\text{EM}} = -\nabla \cdot \mathcal{T}_{\text{EM}} \quad (6.120)$$

So, equation (6.118) becomes, simply,

$$\boxed{\nabla \cdot (\mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{EM}}) = 0} \quad (6.121)$$

That is, the conservation laws of energy and momentum, including the electromagnetic field, can be written in a simple four-dimensional divergence-free form. This has great importance in Einstein's theory of relativistic gravity.

6.6.2.4 Maxwell's Equations in Four-Potential Form: A Model Prescription for All Gauge Fields

Just as the electric and magnetic fields are derived from scalar and vector potentials (equations (6.80)), so the Faraday 2-form can be written as being derived from a four-dimensional vector potential

$$f_{\alpha\beta} = -\frac{\partial a_\alpha}{\partial x^\beta} + \frac{\partial a_\beta}{\partial x^\alpha} \quad (6.122)$$

where the four-potential is composed of the electric (scalar) and magnetic (vector) potentials

$$\mathbf{A} = \begin{pmatrix} \Phi \\ A^x \\ A^y \\ A^z \end{pmatrix} \quad \mathbf{a} = \mathbf{g} \cdot \mathbf{A} = (-\Phi \ A^x \ A^y \ A^z) \quad (6.123)$$

Note from equation (6.122) that the 2-form $\mathbf{f}$ satisfies the relation

$$\frac{\partial f_{\alpha\beta}}{\partial x^\gamma} + \frac{\partial f_{\gamma\alpha}}{\partial x^\beta} + \frac{\partial f_{\beta\gamma}}{\partial x^\alpha} = 0 \quad (6.124)$$

or, in a simpler index notation

$$\partial_\gamma f_{\alpha\beta} + \partial_\beta f_{\gamma\alpha} + \partial_\alpha f_{\beta\gamma} = 0$$

This is actually identical to equation (6.113). So equation (6.122) creates a set of potentials that automatically satisfy the homogeneous Maxwell equations.⁵

Because the Maxwell tensor $\mathcal{M}$ that is derived from this $\mathbf{f}$ automatically satisfies the *homogeneous* equations (6.113), we will no longer need them. This leaves us with only the four inhomogeneous ones

$$\nabla^2 \mathbf{A} - \nabla (\nabla \cdot \mathbf{A}) = -\frac{4\pi}{c} \mathbf{J} \quad (6.125)$$

(i.e., the four-dimensional version of equations (6.84) and (6.85)). Note, now, that the divergence of the left-hand side of equation (6.125) vanishes

$$\nabla \cdot [\nabla^2 \mathbf{A} - \nabla (\nabla \cdot \mathbf{A})] = 0$$

for *any* vector $\mathbf{A}$. This allows us to recover the conservation of charge (equation (6.115)) from this set of four field equations written in potential form. However, as before, we really have only three independent field equations, but four components of $\mathbf{A}$ that we need to determine, giving rise to gauge freedom once again.

The four-potential form for the electromagnetic field equations illustrates the properties and analysis of *all* relativistic gauge theories better than any other form. In general, one proceeds as follows:

1. First, define the “generalized *electromagnetic* (not electric) vector” $\mathbf{E}$

⁵ In geometric form equations (6.122) and (6.124) are written as

$$\begin{aligned} \mathbf{f} &= \mathbf{d}\mathbf{a} \\ \mathbf{d}\mathbf{f} &= \mathbf{d}(\mathbf{d}\mathbf{a}) = 0 \end{aligned}$$

This new differential notation opens up the whole new field of differential forms, which is beyond the scope of this book. So, instead of introducing an entirely new subject, we have chosen here to use the cumbersome index notation, with partial derivatives shown.

$$\mathbf{E} \equiv \nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \quad (6.126)$$

Note that the roman bold $\mathbf{E}$ is a new four-vector derived from the four-potential vector $\mathbf{A}$; it is not the three-vector $\mathbf{E}$ (written in italic bold), which often is referred to as the electric field.

2. Second, using equation (6.125), we then can rewrite Maxwell's *inhomogeneous* equations as, simply,

$$\mathbf{E} = \frac{4\pi}{c} \mathbf{J} \quad (6.127)$$

3. Third, note that the number of equations n (here = 4) and number of unknowns m (= 4, for the four components of $\mathbf{A}$) appear to be equal.
4. Recognize, however, that the electromagnetic vector field $\mathbf{E}$ automatically satisfies a certain number ℓ (= 1) geometric or “Bianchi” identities

$$\nabla \cdot \mathbf{E} = 0 \quad (6.128)$$

which reduce the number of independent equations to only $n' = n - \ell$ (= 3).

5. However, these redundant equations are not useless. Instead, they split off into a set of conservation laws for the field sources. That is, equations (6.127) and (6.128) imply

$$\nabla \cdot \mathbf{J} = 0 \quad (6.129)$$

which is the law of conservation of charge.

6. With only $n' < m$ field equations, we are free to choose ℓ additional gauge conditions to replace the field equations lost as conservation laws. In electromagnetics this often is the Lorenz condition (6.83)

$$\nabla \cdot \mathbf{A} = 0 \quad (6.130)$$

but it can be any other equation involving the four components of $\mathbf{A}$ that is not redundant with the field equations themselves.

7. Finally, when the field equations are linearized, they should admit propagating wave solutions. In a quantized version of the theory, these waves will be the quanta (photons, gravitons, *etc.*) whose exchange will carry the force itself. In the electromagnetic case, the field equations are already linear in the potentials, so with the Lorenz condition, the wave equations are

$$\nabla^2 \mathbf{A} = -\frac{4\pi}{c} \mathbf{J} \quad (6.131)$$

6.6.2.5 The Four-Dimensional Description of Plane Electromagnetic Waves in a Vacuum

We again will treat (more briefly this time) plane waves, without sources, propagating in the z direction, but now using four-dimensional language rather than 3+1 language. We now write the propagation vector as

$$\mathbf{k} = \begin{pmatrix} k^w \\ 0 \\ 0 \\ k^z \end{pmatrix}$$

where $k^w \equiv \omega/c$, and the four-potential as a sinusoidally-varying complex function

$$\mathbf{A} = \Re \{ \mathbb{A} \exp [i(\mathbf{k} \cdot \mathbf{x})] \}$$

where, again, $\mathbb{A}$ is a uniform and constant four-vector wave amplitude.

With no sources ($\mathbf{J} = 0$), the wave equation (6.131) implies that

$$\mathbf{k} \cdot \mathbf{k} = -k^{w2} + k^{z2} = 0$$

(recall that the dot product of two four-vectors must now involve the metric $\mathbf{g} = \boldsymbol{\eta}$). This equation is an interesting result: it says that the four-vector $\mathbf{k}$ is a *null* vector, with zero length in four dimensions. And it says that k^z must have the value

$$k^z = \pm k^w = \pm \frac{\omega}{c}$$

This is a result of the propagation speed of the wave being the speed of light.

The Lorenz condition (6.130), therefore, is simply a requirement that the four-potential be transverse to the propagation four-vector

$$\mathbf{k} \cdot \mathbb{A} = 0$$

or

$$k^w \mathbb{A}^w = k^z \mathbb{A}^z$$

Since we are free to choose $\mathbb{A}^w$ and, since the Faraday tensor does not depend on $\mathbb{A}^w$ nor on $\mathbb{A}^z$, we choose $\mathbb{A}^w = \mathbb{A}^z = 0$. The four-potential for a plane wave propagating in z , then, is simply

$$\mathbb{A} = \begin{pmatrix} 0 \\ \mathbb{A}^x \\ \mathbb{A}^y \\ 0 \end{pmatrix}$$

which reflects the two orthogonal polarizations $\mathbb{A}^x$ and $\mathbb{A}^y$.

6.6.3 *Summary of Section 6.6*

The purpose of Section 6.6 was to familiarize the reader with several important concepts that will be used in Chapter 7:

- Section 6.6.1 introduced gauge field theories and how electromagnetism, for example, can be written in the common language of three space dimensions plus time (3+1). That subsection also showed that electromagnetism can be written in terms of the fields themselves ($\mathbf{E}$, $\mathbf{B}$) or in terms of their potentials (ϕ , $\mathbf{A}$). We also discussed linearly polarized plane waves in this language, as well as the generation of electric dipole radiation.
- Section 6.6.2 showed how the same gauge theory can be written in a four-dimensional, covariant form with tensors and four-vectors, and how it still embodies the concept of gauge freedom (most easily seen when written in four-potential form).

Chapter 7

Physics in Curved Spacetime: General Relativity and Black Holes

The special theory of relativity can describe an enormous amount of physics. It can handle forces due to fluid or gas pressure, flow speeds up to the speed of light, electric and magnetic forces, and even viscous forces. One aspect that is completely missing, however, is gravity. And the gravitational force cannot be added easily: it cannot be written in a stress-energy tensor form for use in equation (6.121). This is true even in Newtonian mechanics, where it is included as an add-on “body force”. In order to include gravity in the theory of relativity, Einstein reasoned that gravity must be a pseudo-force, arising not from another stress-energy component, but from the gradient operator itself (∇) in equation (6.121). In other words, because gravity occurs when matter is present, somehow matter must cause four-dimensional space to be curved, rather than flat. This curvature then gives rise to additional terms in the equations of motion that we interpret as the force of gravity. The addition of curvature to the spacetime metric, plus the realization that energy is mass, and therefore can partake in the generation of the gravitational field, led to a complete and consistent theory of gravity that we now know as Einstein’s general theory of relativity [308, 309]. Constructing a theory of gravity that was consistent with four-dimensional spacetime, however, was a monumental task that took Einstein nearly ten years to fully work out. And it has taken the century following that for the rest of us to determine its implications.¹

¹ The material in this chapter is the author’s distillation of essentials of general relativity [305, 310] that are needed to discuss the astrophysics of black holes. The author has added a new, more intuitive, expression for $r_{\text{isco,BL}}$ (in Boyer–Lindquist coordinates), an emphasis on the angular momentum parameter j being negative as well as positive (see Fig. 7.2), and a strong advocacy of using the symbol j for the dimensionless J , instead of the a or a^* or $\bar{a}$ or α or ... used in the literature.

7.1 The Clash between Special Relativity and Newtonian Gravity, and its Resolution

After Einstein published his relativistic version of Newton's laws of motion at the beginning of the 20th century [307], it was realized that Newton's theory of gravity also was sorely lacking. Newtonian gravity includes *no propagation speed*. For example, equation (5.42) implies that, as the earth–moon system rotates each month, the changes in the gravitational field caused by this rotation are felt instantaneously, and simultaneously, at Mars, Jupiter, Pluto, and even at the stars, light years away. Clearly this cannot be correct, because gravitational (or any other) effects should not be able to propagate faster than the speed of light.

Several attempts were made in the 1910s to construct a relativistic *scalar* theory of gravity – most notably by Gunnar Nordström of Finland, and Einstein himself. Like Newton's theory, these contained only one potential and one field equation, but they did include time-dependence and finite propagation of field effects at the speed of light. Such scalar theories were abandoned, however, because they are not covariant and, more importantly, because they do not predict correctly the precession of the perihelion of Mercury's orbit about the sun, nor the gravitational bending of light by the sun itself. The same thing can be said of vector theories of gravity.

Mathematically, and physically, the lack of covariance of Newton's and other scalar and vector theories was a real problem. The source term for the gravitational field – mass density – is only one component of the stress-energy tensor, and mass density alone has different values in different Lorentz frames. Scalar theories of gravity imply that the law of gravity should be different for different observers. This is not a good property for a relativistic theory.

Electromagnetism solves the covariance problem by using a four-vector as the source – the current $\mathbf{J}$. The components of the current may be different in different frames, but the magnitude of the vector itself remains the same. And the field equations written in terms of geometric objects (equation (6.126)) are the same in any frame. In order for this to be the case for a relativistic theory of gravity, the source term would have to be the entire stress-energy tensor $\mathcal{T}$, and the gravitational field would have to be a tensor field. That is, the field equations should be something like

$$\mathcal{G} \propto G\mathcal{T}$$

Since $\mathcal{T}$ is symmetric 2-tensor, the gravitational tensor $\mathcal{G}$ must be a symmetric 2-tensor also, so there must be 10 field equations and 10 gravitational potentials. Furthermore, because $\mathcal{T}$ is divergence-free in any frame, then $\mathcal{G}$ must be globally divergence-free also. That is, the gravitational field tensor must possess a Bianchi identity (differential symmetry) like the electromagnetic field vector $\mathbf{E}$ does in equation (6.126), making gravity a gauge theory just like electromagnetism. The difficult part comes in determining answers to the following questions: what are the 10 gravitational potentials, what are their significance, and what is the exact form of the field equations?

Recognizing that, in order to be covariant, gravity must be a tensor theory was one of Einstein's breakthroughs in the development of general relativity. The other was the realization of how gravity arises as a force in the relativistic conservation laws. Unlike electromagnetism, gravity is not expressed by adding another component to the stress-energy tensor. Instead, it is a pseudo-force that arises from the gradient operator itself in equation (6.121), in a manner similar to the centrifugal and Coriolis pseudo-forces. However, in order to be a covariant law, that pseudo-force (unlike centrifugal and Coriolis forces) must be a general feature of spacetime for all observers. The easiest way to produce a global pseudo-force is to make spacetime *curved* instead of flat. Curvature then generates a pseudo-force throughout the domain, for all observers. Einstein therefore concluded that gravity arises because the presence of matter curves the otherwise flat Minkowski space. This produces additional terms in the equations of motion that are perceived as a gravitational force. Free fall, then, is the true inertial frame, and this statement becomes the generalization of Newton's first law: a body that experiences no external forces will follow a geodesic in the curved spacetime set up by the presence of matter.

These ideas, along with the character of Newton's theory itself, give clues to what the 10 potentials should be and to how the field equations should be constructed. The potentials must be the independent components of the symmetric metric tensor $\mathbf{g}$. And, because curvature in the spacetime metric can be expressed as second derivatives of the metric coefficients, the tensor $\mathcal{G}$ must be an expression of the metric curvature. Newton's theory essentially embodies this idea already: Poisson's equation implies that the curvature (second derivative with respect to space ∇^2) of the gravitational potential (ψ) is caused by the presence of matter ρ (equation (5.42)).

To summarize mathematically, then, the true metric describing local spacetime is no longer the simple Minkowskian one (equation (6.51)). Rather, it is a general four-dimensional symmetric tensor of the form

$$\mathbf{g} = \begin{pmatrix} g_{00} & g_{01} & g_{02} & g_{03} \\ g_{01} & g_{11} & g_{12} & g_{13} \\ g_{02} & g_{12} & g_{22} & g_{23} \\ g_{03} & g_{13} & g_{23} & g_{33} \end{pmatrix} \quad (7.1)$$

containing all we need to know about gravity. Its 10 independent components are found by solving a tensor field equation of the form

$$\mathcal{G} = K G \mathcal{T} \quad (7.2)$$

where G is the gravitational constant from Newton's theory, and K is a constant that we will need to determine once we understand $\mathcal{G}$. The equations of motion still will be given by equation (6.65)

$$\mathbf{U}^T \cdot \nabla \mathbf{P} = \mathbf{F} \quad (7.3)$$

for a single particle and by equation (6.67)

$$\nabla \cdot \mathcal{T} = 0 \quad (7.4)$$

for a continuous medium. However, the covariant gradient now will generate not only pseudo-forces associated with any curvilinearity of the metric but also those associated with the curvature of spacetime itself, *i.e.*, gravity. How that happens is discussed below.

7.2 Curved Space and the Riemann Curvature Tensor

7.2.1 Simple Examples of Curved, and Not Curved, Space

Before we discuss curved spacetime metrics in four dimensions, we should take a quick look at simpler curved metrics that we can visualize. Up until now all the metrics we have considered have been flat – the two-dimensional Pythagorean one and all its coordinate variations, the three-dimensional Euclidean ones, and the four-dimensional Minkowski metric. None of these involved curvature, let alone mass and gravity.

7.2.1.1 The Curved Surface of a Sphere

What does it mean to have a curved space, and how can we tell whether or not a metric has curvature? An easy way to answer these questions is to consider the surface of a sphere and, in particular, the surface of the earth. Three-dimensional Euclidean space itself is flat, but the *two-dimensional surface* of a sphere embedded in that space is curved. We can derive the two-dimensional metric for the surface of a sphere by setting $dr = 0$ in the three-dimensional spherical-polar metric, equation (6.38)

$$dh^2 = d\ell^2 = r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (7.5)$$

Because r is a constant, this metric is a function of only two coordinates θ and ϕ . On this spherical surface, the classical Pythagorean theorem (equation (6.3)) does *not* hold true, and that is due solely to the factor $\sin^2 \theta$. There is no way to convert equation (7.5) to equation (6.3) by transforming to a new x and y . There always will be a nonlinear function of x and/or y multiplying at least one of the squared differentials.²

² One might think that such a Cartesian grid could be created by transforming

$$dx' = r d\theta \quad dy' = r \sin \theta d\phi$$

However, it is easy to see that this is a local, not global, transformation. It cannot be integrated into the total differential form (6.31), so it is valid at only one point (r, θ) . This is well known in the field of surveying. A local system of perpendicular streets and avenues in a single city cannot be used to describe the addresses of all houses on the earth's surface.

Another method of detecting curvature in a two-dimensional surface is to cut it open and lay it flat in an attempt to try to project it onto a flat two-dimensional surface. If that projection produces no distortions in the size of portions of that surface (like continents on the earth, for example), then the original surface was flat.

7.2.1.2 The Flat Surface of a Cylinder

As another example of this idea of curvature, let us look at the surface of a cylinder. We can derive the metric for it by setting $dR = 0$ in equation (6.37)

$$dh^2 = d\ell^2 = R^2 d\phi^2 + dz^2 \quad (7.6)$$

This metric also appears curved because of the R^2 . However, it is not; R is not a function of ϕ or z . We can cut the cylinder down its length, open it up and lay it flat on a table. Also, because R is a constant, equation (7.6) can be re-written as

$$dh^2 = [d(R\phi)]^2 + dz^2 \quad (7.7)$$

which is flat in the circumferential coordinate $C = R\phi$. Because the metric passes our tests, it is, therefore, flat.

7.2.2 Testing for Curvature Using the Riemann Tensor

These simple examples are illustrative, but they do not give us good tools to determine whether or not even three-dimensional metrics are curved, let alone four-dimensional ones. Of course, it is not possible to physically cut out a piece of complex spacetime and try to lay it flat in a higher-dimensional space. And trying to show that a particular metric cannot be converted to a flat one with any global coordinate transformation is not always practical either. To handle complex metrics we need a good mathematical tool. The best way to test for all possible types of curvature is to compute the *Riemann curvature tensor* from the metric.

7.2.2.1 The Riemann Curvature Tensor and its Symmetries

By the mid 1800s the German mathematician Georg Friedrich Bernhard Riemann of the University of Göttingen had developed a general theory of metric spaces, which we now know as Riemannian geometry. In order to test for curvature, a special tensor known as the Riemann curvature tensor $\mathfrak{R}$ was developed: if all components of this tensor were zero, the space was flat (no curvature along any direction in that space). Of course, the tensor has many other uses as well.

Unfortunately, it is not a trivial exercise to compute the Riemann tensor in most cases. While slope requires only the first derivative of the metric components, de-

termining curvature requires computing second derivatives. In a space of N dimensions, there are $N \times N$ possible second derivatives of each of the $N \times N$ components of the metric

$$\frac{\partial^2}{\partial w^2}, \frac{\partial^2}{\partial w \partial x}, \frac{\partial^2}{\partial x \partial w}, \frac{\partial^2}{\partial w^2}, \dots, \frac{\partial^2}{\partial z \partial y}, \frac{\partial^2}{\partial z^2}$$

So a tensor with these derivatives of the metric tensor (which has N^2 components itself) potentially has N^4 components in all – 16 in two dimensions, 81 in three, and 256 in four. Fortunately, because of numerous symmetries in $\mathfrak{R}$ and in the metric itself, the amount of computation necessary turns out to be far less.

The Riemann Tensor in Index Notation

In Appendix B we discuss the Christoffel symbols, which are used to construct covariant gradients and divergences in curvilinear and curved spaces. These actually express all possible *first* derivatives of the metric

$$\Gamma_{\alpha\beta\gamma} = \frac{1}{2} \left(\frac{\partial g_{\alpha\beta}}{\partial x^\gamma} + \frac{\partial g_{\alpha\gamma}}{\partial x^\beta} - \frac{\partial g_{\beta\gamma}}{\partial x^\alpha} \right) \quad (7.8)$$

So it is natural that the Riemann curvature tensor can be expressed as a covariant-like derivative of the Christoffel symbols

$$\mathfrak{R}_{\alpha\beta\gamma\delta} = \frac{\partial \Gamma_{\alpha\beta\delta}}{\partial x^\gamma} - \frac{\partial \Gamma_{\alpha\beta\gamma}}{\partial x^\delta} + \sum_{\mu, \nu} g^{\mu\nu} [\Gamma_{\mu\alpha\delta} \Gamma_{\nu\beta\gamma} - \Gamma_{\mu\alpha\gamma} \Gamma_{\nu\beta\delta}] \quad (7.9)$$

Note that the Christoffel symbols Γ are not true tensors. They can have their indices raised and lowered (using g), but they do not transform between coordinate systems like tensors do. $\mathfrak{R}$, however, is a true tensor.

At this point it has become obvious that our simple matrix notation has failed, chiefly because these objects have more than two indices. In order to continue we would need Γ to be a *three*-dimensional matrix and $\mathfrak{R}$ to be *four*-dimensional (not because of the dimensions of spacetime but because of their number of indices). We could, for example, try to write

$$\mathfrak{R} = \partial \Gamma - (\partial \Gamma)^T + g^{-1} : \Gamma \Gamma - [g^{-1} : \Gamma \Gamma]^T$$

However, this has little computational usefulness, because we do not know which indices are being summed and which are being transposed.³ So, we have chosen instead to use the index notation above to discuss curvature.

³ There *do* exist mathematical formalisms (e.g., differential forms) that express these quantities in a geometrical form, without indices, and yet also have computational power. However, introducing and mastering them in this book would require more effort than briefly resorting to the messy index notation, as we have done here.

Algebraic Symmetries of the Riemann Tensor

The Riemann tensor possesses several algebraic symmetries. An obvious one is antisymmetry when the last two indices are interchanged:

$$\mathfrak{R}_{\alpha\beta\delta\gamma} = -\mathfrak{R}_{\alpha\beta\gamma\delta}$$

from equation (7.9). And, if we plug equation (7.8) into (7.9), we also find symmetry on interchanging adjacent *pairs* of indices as well as antisymmetry on the *first* two indices:

$$\mathfrak{R}_{\gamma\delta\alpha\beta} = \mathfrak{R}_{\alpha\beta\gamma\delta} \qquad \mathfrak{R}_{\beta\alpha\gamma\delta} = -\mathfrak{R}_{\alpha\beta\gamma\delta}$$

With equation (7.9), this shows that any repeated index in the first or second pairs results in the Riemann tensor component being identically zero

$$\mathfrak{R}_{\alpha\alpha\gamma\delta} = \mathfrak{R}_{\alpha\beta\gamma\gamma} = 0$$

Finally, there is another algebraic symmetry, sometimes called the first Bianchi identity, that relates components of permuted indices

$$\mathfrak{R}_{\alpha\beta\gamma\delta} + \mathfrak{R}_{\alpha\delta\beta\gamma} + \mathfrak{R}_{\alpha\gamma\delta\beta} = 0 \tag{7.10}$$

For dimensions $N < 4$, equation (7.10) is redundant with the other algebraic symmetries above, and the number of potentially non-zero components is only

$$\mathfrak{N} = \frac{1}{8} (N^4 - 2N^3 + 3N^2 - 2N) \qquad N < 4$$

However, for $N \geq 4$, equation (7.10) has non-redundant conditions that involve $\alpha \neq \beta \neq \gamma \neq \delta$, such as the one in four dimensions

$$\mathfrak{R}_{0123} + \mathfrak{R}_{0312} + \mathfrak{R}_{0231} = 0$$

These further reduce the number of independent components to

$$\mathfrak{N} = \frac{1}{8} (N^4 - 2N^3 + 3N^2 - 2N) - \frac{N!}{4!(N-4)!} \qquad N \geq 4$$

In $N = 2, 3$, and 4 dimensions, then, the Riemann tensor has, respectively, only 1, 6, and 20 unique, potentially non-zero components.

Differential Symmetry of the Riemann Tensor: The Bianchi Identities

The Riemann tensor also possesses differential symmetries, sometimes called the second Bianchi identity or simply the Bianchi identity. These are

$$\nabla_\gamma \mathfrak{R}_{\mu\nu\alpha\beta} + \nabla_\beta \mathfrak{R}_{\mu\nu\gamma\alpha} + \nabla_\alpha \mathfrak{R}_{\mu\nu\beta\gamma} = 0 \quad (7.11)$$

This identity is similar to equation (6.124) for the electromagnetic field, but with co-variant gradients used instead of simple derivatives. This identity does not decrease the number of components of the tensor that we need to calculate, but it does have profound importance for Einstein's theory of gravity and for fundamental laws of physics.

7.2.2.2 Example Riemann Tensors

Most everything we need to demonstrate about the Riemann tensor can be done in two dimensions. Below are a few examples of curvature in simple metrics.

The Pythagorean Metric in Polar Coordinates

For equation (6.8) only one component of $\mathfrak{R}$ needs to be computed

$$\mathfrak{R}_{1212} = \frac{\partial \Gamma_{122}}{\partial x^1} + g^{22} \Gamma_{212} \Gamma_{221}$$

Now, $\Gamma_{212} = \Gamma_{221} = -\Gamma_{122} = R$; and $g^{22} = 1/R^2$. So we have $R_{1212} = -1 + 1 = 0$. That is, even in polar coordinates, the Pythagorean metric is still flat, as we expected.

The Surface of a Cylinder

In the case of equation (7.6), because R is a constant (not a function of the coordinates ϕ and z), all $\Gamma_{\alpha\beta\gamma}$ are zero, as are all $\mathfrak{R}_{\alpha\beta\gamma\delta}$. So the surface of a cylinder is flat, as we deduced before.

The Surface of a Sphere

For equation (7.5) we again need to compute only one component of $\mathfrak{R}$. However, in this case, it is *not* zero

$$\mathfrak{R}_{1212} = r^2 \sin^2 \theta \quad (7.12)$$

since $\Gamma_{212} = \Gamma_{221} = -\Gamma_{122} = r^2 \sin \theta \cos \theta$, and $g^{22} = 1/(r^2 \sin^2 \theta)$. Initially, this looks strange: does the curvature change with position (θ) on the sphere? No, the issue is simply with how we express the Riemann tensor. If we write it as a *mixed* tensor, then the only non-zero component $\mathfrak{R}_{12}{}^{12} = 1/r^2$ is constant over the

sphere.⁴ So the surface of a sphere is *not* flat: the Riemann tensor has at least one non-zero component.

Other Examples

Similar, though somewhat more involved, calculations can be done by applying equations (7.8) and (7.9) to the three-dimensional curvilinear metrics in equations (6.37) and (6.38). After some effort one finds that, as in the polar Pythagorean metric above, all components of $\mathfrak{R}$ are zero. So three-dimensional Euclidean space, even in curvilinear coordinates, is flat.

7.3 A Quick Road to Einstein's Theory of Gravity

A full treatment of the general theory of relativity is well beyond the scope of this book. However, we can give an outline of the theory that will be sufficient for our discussion of black holes, and do so without having to deal with too much of the detailed mathematics of tensor calculus, differential geometry, and differential forms. At this point, in fact, we are in a position to draw the same conclusions that Einstein did about the relativistic gravitational field and arrive at the same field equations.

The goal is to develop a tensor gauge theory of gravity with a structure similar to the relativistic vector theory of electromagnetism outlined in Section 6.6.2.4. The 10 metric coefficients are our 10 potentials, and the 10 sources are the independent components of the symmetric stress-energy tensor. The field equations must have the form

$$\mathcal{G} = K G \mathcal{T} \quad (7.13)$$

where K is a constant we need to determine, and G is the gravitational constant from Newton's theory.

7.3.1 Determining the Form of the Gravitational Field Equations $\mathcal{G}$

$\mathcal{G}$ must be a 2-tensor that somehow embodies the curvature of the metric g , and it must possess a differential symmetry condition

$$\nabla \cdot \mathcal{G} = 0 \quad (7.14)$$

⁴ Furthermore, in the next section we will discuss the contracted version of $\mathfrak{R}$, the Ricci tensor $\mathcal{R}$ and its trace (the Ricci scalar $\mathcal{R}$). In mixed form, these have the values $\mathcal{R}_1^1 = \mathcal{R}_2^2 = \mathcal{R}/2 = 1/r^2$ for the spherical surface of radius r .

in order to preserve the conservation laws

$$\nabla \cdot \mathcal{T} = 0 \quad (7.15)$$

which now must also be true in curved spacetime. The conservation laws (equations of motion), therefore, are a consequence of gravity being a gauge field and that field $\mathcal{G}$ having differential Bianchi identities. *The laws of conservation of energy and momentum are not accidents of nature; they derive directly from the gauge nature of spacetime.*

The gravitational curvature 2-tensor must somehow be related to the Riemann curvature 4-tensor $\mathfrak{R}$. There are many ways to extract a 2-tensor from a 4-tensor, but only a few will inherit the properties of the Bianchi identities (equations (7.11)). One method is to *contract* $\mathfrak{R}$ with the inverse metric

$$\mathcal{R} \equiv g^{-1} : \mathfrak{R} \quad (7.16)$$

If the double-dot product is on the first and third indices of $\mathfrak{R}$, i.e., if

$$\mathcal{R}_{\alpha\beta} = \sum_{\mu,\nu} g^{\mu\nu} \mathfrak{R}_{\mu\alpha\nu\beta} \quad (7.17)$$

then the resulting tensor $\mathcal{R}$ inherits a differential symmetry property from the Riemann tensor's Bianchi identity (equation (7.11))

$$\nabla_\alpha \left(\sum_\lambda g^{\alpha\lambda} \mathcal{R}_{\lambda\beta} \right) - \frac{1}{2} \nabla_\beta \mathcal{R} = 0$$

or

$$\boxed{\nabla \cdot (g^{-1} \cdot \mathcal{R}) - \frac{1}{2} \nabla \mathcal{R} = 0} \quad (7.18)$$

where

$$\begin{aligned} \mathcal{R} &\equiv \sum_{\beta,\lambda} g^{\beta\lambda} \mathcal{R}_{\lambda\beta} \\ &= g^{-1} : \mathcal{R} \end{aligned} \quad (7.19)$$

$\mathcal{R}$ is called the Ricci curvature tensor, after the Italian mathematician Gregorio Ricci-Curbastro, who developed tensor calculus in the late 1800s. And $\mathcal{R}$ is the trace of the Ricci tensor, usually called the “Ricci scalar”.⁵

Now, $\mathcal{R}$ is not divergence-free, as we see from equation (7.18); so *the Ricci tensor cannot be the gravitational tensor $\mathcal{G}$ that we are looking for*. However, we *can* use this contracted Bianchi identity of the Ricci tensor, and the property of the

⁵ For the simple two-dimensional example of the surface of a sphere (equation (7.12)) we find the Ricci tensor components and scalar to be $\mathcal{R}_{11} = 1$, $\mathcal{R}_{22} = \sin^2 \theta$, and $\mathcal{R} = 2/r^2$, respectively.

metric that it is divergence-free (Appendix B), to construct another 2-tensor that *is* divergence-free

$$\mathcal{G} \equiv \mathcal{R} - \frac{1}{2} \mathcal{R} g \quad (7.20)$$

This is sometimes referred to as the “Einstein tensor” in his honor, even though the symbol $\mathcal{G}$ is used to express it. It actually is the *only* 2-tensor that can be derived from the Riemann tensor and that has all the necessary properties to be the $\mathcal{G}$ in equation (7.13) [305]. In index notation, the Einstein tensor is usually written in contravariant tensor form

$$\mathcal{G}^{\alpha\beta} = \mathcal{R}^{\alpha\beta} - \frac{1}{2} g^{\alpha\beta} \mathcal{R}$$

where the contravariant form of the Ricci tensor is

$$\mathcal{R}^{\alpha\beta} = \sum_{\mu,\nu} g^{\alpha\mu} g^{\beta\nu} \mathcal{R}_{\mu\nu}$$

7.3.2 Determining the Constant of Proportionality K

The constant of proportionality is determined by requiring that Einstein's theory of gravity (equation (7.13)) reduce to Newton's theory (equation (5.42)) when the gravitational field is weak and the speed of fluid in the gravitational field is slow ($V \ll c$). In this case we can write the metric as

$$g = \begin{pmatrix} -1 + h_{ww} & 0 & 0 & 0 \\ 0 & 1 + h_{xx} & 0 & 0 \\ 0 & 0 & 1 + h_{yy} & 0 \\ 0 & 0 & 0 & 1 + h_{zz} \end{pmatrix}$$

where the h 's are weak perturbations on the Minkowski metric. We really will need only h_{ww} and the three Christoffel symbols

$$g^{ii} \Gamma_{iww} \approx -\frac{1}{2} \frac{\partial h_{ww}}{\partial x^i}$$

From equation (B.5) the conservation of momentum in, *e.g.*, the x direction is

$$\begin{aligned} 0 &= (\nabla \cdot \mathcal{T})^x \\ &= \sum_{\beta} \nabla_{\beta} \mathcal{T}^{x\beta} \\ &= \sum_{\beta} \frac{\partial \mathcal{T}^{x\beta}}{\partial x^{\beta}} + \sum_{\mu\beta} g^{xx} \Gamma_{x\mu\beta} \mathcal{T}^{\mu\beta} + \sum_{\mu} \mathcal{T}^{x\mu} \sum_{\beta\lambda} g^{\beta\lambda} \Gamma_{\lambda\mu\beta} \end{aligned}$$

If $V \ll c$, then the only large component of $\mathcal{T}^{\alpha\beta}$ is

$$\mathcal{T}^{ww} = \rho c^2$$

so we are left with only one pseudo-force term in the full equation

$$0 = \frac{\partial(\rho \mathbf{V} c)}{\partial w} + \nabla \cdot (\rho \mathbf{V} \mathbf{V}) + \nabla p - \frac{1}{2} \nabla(h_{ww}) \rho c^2$$

The last term in this equation of motion must be the negative of Newton's gravitational force $\rho \nabla \psi$, so we now can determine h_{ww} and $\mathcal{G}^{ww}$

$$h_{ww} = -2\psi/c^2 \quad \mathcal{G}^{ww} = 2\nabla^2 \psi/c^2$$

Therefore, if Newton's and Einstein's laws of gravitation are, respectively,

$$\nabla^2 \psi = 4\pi G \rho \quad \text{and} \quad 2\nabla^2 \psi/c^2 = K G \rho c^2$$

then, in order for the two theories to agree in the weak gravity limit, the constant of proportionality in Einstein's theory of gravity must be

$$K = \frac{8\pi}{c^4}$$

The final Einstein field equations, then, are

$$\boxed{\mathcal{G} = 8\pi \frac{G}{c^4} \mathcal{T}} \quad (7.21)$$

7.3.3 The Meaning of Gauge in Einstein's Theory of Gravity

The fact that $\mathcal{G}$ possesses the differential symmetry (7.14) means that the number of useful field equations in (7.21) is reduced from 10 to 6. The theory, therefore has gauge freedom that allows us to construct four independent conditions on the metric coefficients to replace the four field equations lost as conservation laws. What is the interpretation of gauge in the general theory of relativity?

The answer is that the gauge is simply the coordinate system itself. Just as we have the freedom to choose how we lay out a coordinate grid in three-space by specifying three variables ($[x, y, z]$, $[r, \theta, \phi]$, *etc.*), so we have the freedom to specify the four coordinates used to span spacetime. *The gauge is the coordinate system, and the gauge transformation is the generalized Lorentz transform.*

7.3.4 Curvature without Local Matter

Because the Einstein tensor is formed by summing Riemann tensor components, $\mathcal{G}$ does not always have the same properties as the latter. To be sure, if the spacetime is flat, and all the Riemann components are zero, then all the components of $\mathcal{G}$ will be zero as well. However, the reverse is not true. *If all the components of $\mathcal{G}$ are zero, then it is not necessarily true that the spacetime is flat.* Non-zero components of $\mathfrak{R}$ can sum to zero when we calculate $\mathcal{G}$.⁶

This sounds like it might be a rare occurrence, but it is very common. It happens everywhere in the vacuum of space where $\mathcal{T} = 0$, *i.e.*, where the simpler equation

$$\mathcal{G} = 0 \quad (7.22)$$

gives the gravitational field. Obviously, space is not flat outside a planet or a black hole, so the components of $\mathfrak{R}$ cannot all be zero there. But where is no matter, $\mathcal{G}$ *must* be zero. In fact, when we discuss black holes in this book, we often will refer to solutions of equation (7.22) and not to ones of the more general equation (7.21).

7.3.5 E & M in a General Four-Dimensional Curved Space

The beauty and simplicity of using geometry to describe curved space is very obvious for electromagnetic theory. The same geometric equations that we derived in flat space (equations (6.112) and (6.113)) still describe the electromagnetic field

$$\nabla \cdot \mathcal{F}^T = \frac{4\pi}{c} \mathbf{J} \quad (7.23)$$

$$\nabla \cdot \mathcal{M}^T = 0 \quad (7.24)$$

and the equations of motion can also be expressed in the same way

$$\nabla \cdot \mathcal{T}_{\text{FL}} = \frac{1}{c} \mathbf{J}^T \cdot \mathcal{F}$$

or, by defining the electromagnetic stress-energy tensor (equation (6.121)),

$$\nabla \cdot (\mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{EM}}) = 0$$

The only difference now is that the curvature of space and the presence of gravity are hidden in the geometry of the gradient operator ∇ , which is determined by how the metric is expressed in a coordinate system. The equations of E & M are valid in any frame – flat or curved, accelerating or not – and, therefore, are generally covariant.

⁶ Indeed, in the earlier two-dimensional example of the surface of a sphere (equation (7.12)), we find that the components of the Einstein tensor are $\mathcal{G}_1^1 = \mathcal{G}_2^2 = 0$, even though the scalar curvature $\mathcal{R} = 2/r^2$ is clearly non-zero.

7.4 Four-Dimensional Schwarzschild Geometry: Non-Rotating Black Holes

We are now in a position to discuss black holes and how curved spacetime generates gravity. Up until now we have been dealing with only one metric, the Minkowskian one for flat spacetime. It is important to realize, however, that for each matter distribution $\mathcal{T}$, and even each type of black hole vacuum metric, the solution of either equation (7.21) or (7.22) for g will be different. Each metric will create a different geometry and, therefore, some different physics.

7.4.1 The Schwarzschild Metric and Schwarzschild Radius

In 1916, just a few months before he died, Karl Schwarzschild of the Potsdam Astrophysical Observatory showed that a static, spherical metric is a solution to equation (7.22) for the spacetime around a point mass M that is not rotating [311]⁷

$$g_{\text{SH}}^{\text{SCH}} = \begin{pmatrix} -c^2 \left(1 - \frac{2GM}{c^2 r}\right) & 0 & 0 & 0 \\ 0 & \left(1 - \frac{2GM}{c^2 r}\right)^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \theta \end{pmatrix} \quad (7.25)$$

where the spatial coordinates are now spherical-polar, and the full set (t, r, θ, ϕ) , with the metric in this form, is often called the “Schwarzschild” coordinate system. However, it actually was David Hilbert who re-derived this metric in the above form [312, 313]. For this reason, and for reasons discussed in Section 7.4.4, we will refer to these *coordinates* as Schwarzschild–Hilbert (SH) ones but still will refer to the geometry of a spherical and stationary black hole as the Schwarzschild metric (SCH).

Note that, even in the Schwarzschild metric, a sphere of constant radius ($dr = dt = 0$) has the same two-dimensional metric as that of the surface of a sphere in Euclidean space (equation (7.5)). Therefore, the circumference and area of a sphere of any value r that encloses the black hole are related in the same way as in Euclidean geometry: $A = C^2/\pi$. But Schwarzschild space is not Euclidean, so r is not really a true, or proper, distance from the center of the black hole; it is, instead, an areal or circumferential coordinate:

$$r \equiv \left(\frac{A}{4\pi}\right)^{1/2} = \frac{C}{2\pi}$$

⁷ Schwarzschild also was the father of astrophysicist Martin Schwarzschild and died just a few weeks before his son’s fourth birthday at the age of 42.

Nevertheless, the proper distance between two closely-spaced spheres at r and $r + \Delta r$ can be computed by setting $dt = d\theta = d\phi = 0$ in the Schwarzschild line element, yielding

$$\Delta s \approx \frac{1}{(1 - \frac{r_S}{r})^{1/2}} \Delta r \quad (7.26)$$

where r_S is the Schwarzschild radius

$$r_S \equiv \frac{2 G M}{c^2} \quad (7.27)$$

If we integrate this as a differential, from the horizon outward, we find that the proper radial distance from the *horizon* at r_S to a point at r is given by the complicated expression

$$s = r \left(1 - \frac{r_S}{r}\right)^{1/2} + r_S \ln \left\{ \left(\frac{r}{r_S}\right)^{1/2} \left[1 + \left(1 - \frac{r_S}{r}\right)^{1/2}\right] \right\}$$

When $r \rightarrow r_S$ we have $s = 0$, of course. And as $r \rightarrow \infty$, the first term dominates and we have $s \rightarrow r$. So, when r is large, it is a good approximation to the proper distance, but when r gets close to r_S , the differential proper distance Δs becomes much larger than Δr itself (equation (7.26)).

There are several things to note about the Schwarzschild metric. First, far away from the black hole ($r \rightarrow \infty$) the line element becomes

$$ds^2 = -c^2 dt^2 + dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (7.28)$$

which is just the spherical-polar form of the Minkowski line element (equation (6.50)).⁸ The same is true if the mass goes away ($M \rightarrow 0$). So, when the gravitational influence of M disappears, the Schwarzschild metric reduces to the flat Minkowski metric. That is, *the Schwarzschild metric is consistent with special relativity*.

Second, the metric is diagonal (no $t\phi$ or $r\theta$ terms, for example), and it is curved in time and radius. Furthermore, those curvature terms are functions only of the r coordinate. So, relativistic spherical gravity is still a radial force only.

Finally, and most interestingly, strange things seem to occur when we approach closer to M than the distance $r = r_S$. The quantity $1 - 2 G M / c^2 r$ changes sign and the Schwarzschild metric becomes

⁸ Note that, unlike the Cartesian-like Minkowski metric in Chapter 6 (where we used four linear coordinates x , y , z , and even $w = ct$ for time), we now have only one linear coordinate (r). The remainder are two angular coordinates (θ , ϕ), with units of radians, and a true time coordinate (t).

$$g_{\text{SCH}}^{\text{SCH}} = \begin{pmatrix} c^2 \left(\frac{2GM}{c^2 r} - 1 \right) & 0 & 0 & 0 \\ 0 & - \left(\frac{2GM}{c^2 r} - 1 \right)^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \theta \end{pmatrix}$$

Inside the Schwarzschild radius the distance metric component is negative and the time component is positive. Time and space seem to have changed places. While outside the black hole we are constrained to always move forward in time, inside r_S we are now constrained to move toward the mass M . Even light must fall toward the central object; there is no escape. This is the essence of a black hole horizon. It is not defined merely by the fact that the escape velocity is the speed of light there. It is that interior to the horizon, time and space are so warped that the future of anything there, including light, is predestined to fall into the center.

7.4.2 Physics in Curved Schwarzschild Geometry

7.4.2.1 Coordinate Systems

Calculating how objects move near a black hole (orbits, for example) now becomes a little tricky. Not only does the flow of time change with how fast an object is moving, but because g_{ww} changes with r , time's flow also changes with how close that object is to the black hole. In which coordinate system do we write the equations? We have several choices:

1. *The moving-body system (MOV).* Things are rather simple in this case – in fact, too simple. Just as in any local coordinate system that follows a moving body (including the reader now sitting in his chair), the metric is Minkowski and the four velocity is $\mathbf{U} = (c, 0, 0, 0)$. Even though the body may be moving with respect to the black hole, the spatial velocity of the moving body with respect to this coordinate system is zero, by definition. Using such a system is equivalent to the observer traveling in a closed box surrounding the body, never aware of the black hole metric outside. So nothing is learned about the body's motion around the black hole. This moving frame sometimes is called the local Lorentz observer system and is useful for expressing microphysics, such as gas pressure, temperature, and density, but not motion.
2. *The fixed local Lorentz system (FIX).* This system is a little better. At least the body has a velocity in this system $[\mathbf{U} = (\gamma c, \gamma V^{\hat{x}}, \gamma V^{\hat{y}}, \gamma V^{\hat{z}})]$. The system is equivalent to placing an (open) box at a fixed point somewhere near the black hole and watching the body as it moves through the box. We need to decide how to orient the box, of course. Generally we would choose the new $\hat{x}$ along r , $\hat{y}$ along θ , and $\hat{z}$ along ϕ , so that $\mathbf{U} = (\gamma c, \gamma V^{\hat{r}}, \gamma V^{\hat{\theta}}, \gamma V^{\hat{\phi}})$. The hats (^) are usually added to the subscripts to emphasize that these are not Schwarzschild–Hilbert coordinates, even though they point in the same direction. They are a

locally-Lorentz system, still with a Minkowski metric. The problem with this system, of course, is that it is not unique. There is a different coordinate system, oriented in a different direction, for every point around the black hole. And, of course, there is no single time coordinate that describes all the points around the black hole; time flows differently at different radii r , so it will flow differently in different boxes. This is sometimes called the Fiducial Observer or FIDO system.

3. *The Schwarzschild–Hilbert system (SH).* The Schwarzschild–Hilbert coordinate system (t, r, θ, ϕ) is global, so it does not suffer the above problems. The body has motion through the system, so $[\mathbf{U} = (U^t, U^r, U^\theta, U^\phi)]$. There is a unique time coordinate (here we have used t) that is usable throughout space and time, and we have a single spatial coordinate system (r, θ, ϕ) as well. The problem here is that the four-velocity $\mathbf{U}$ is difficult to interpret. The units on most of the components are strange (U^t is unitless [seconds per second] and U^θ and U^ϕ are in radians per second), and the components of $\mathbf{U}$ contain the Lorentz factor γ as well as the metric coefficients. Then, there is the added confusion of whether to use vectors or 1-forms for the velocity and other vector quantities, because now the metric is not Minkowskian.

The usual solution to this dilemma is to employ a hybrid system that uses the best aspects of all three. Derivatives are expressed in the *global* (here SH) system, so solutions will be given as functions of (t, r, θ, ϕ) . On the other hand, velocity, electric and magnetic field, and the charge/current four-vectors are expressed in the fixed local Lorentz (FIX) system, and we usually will use the standard three-velocity $\mathbf{V} = (V^{\hat{r}}, V^{\hat{\theta}}, V^{\hat{\phi}})$, with components in centimeters per second, instead of $\mathbf{U}$. Finally, thermodynamic scalar quantities such as density and pressure are expressed in the moving-body system (MOV), where the familiar thermodynamic laws and equations of state are valid. This prescription, is sometimes called the “3+1 system”, and is discussed in much more detail at the end of this chapter and in Chapter 9. Its great advantage is that variables are expressed in familiar terms and the equations look very similar to the laws of physics that we already know here on the earth.

But how is the 3+1 system accomplished in practice? The answer is through generalized Lorentz transformations. For example, let us consider the velocity. In the moving-body system $\mathbf{U}_{\text{MOV}} = (c, 0, 0, 0)$ and the metric is given by equation (6.51). A Lorentz transformation to the fixed local system produces $\mathbf{U}_{\text{FIX}} = (\gamma c, \gamma V^{\hat{r}}, \gamma V^{\hat{\theta}}, \gamma V^{\hat{\phi}})$, with the metric remaining Minkowskian. Finally, to transform from the fixed local system to the global Schwarzschild–Hilbert system, we use a transformation that turns equation (6.51) into the Schwarzschild metric

$$\mathcal{L}_{\text{FIX} \rightarrow \text{SH}} = \begin{pmatrix} \frac{1}{c(1 - \frac{2GM}{c^2 r})^{1/2}} & 0 & 0 & 0 \\ 0 & (1 - \frac{2GM}{c^2 r})^{1/2} & 0 & 0 \\ 0 & 0 & 1/r & 0 \\ 0 & 0 & 0 & 1/(r \sin \theta) \end{pmatrix}$$

so that $(\mathcal{L}_{\text{FIX} \rightarrow \text{SH}}^{-1})^T \mathbf{g}_{\text{FIX}}^{\text{SCH}} \mathcal{L}_{\text{FIX} \rightarrow \text{SH}}^{-1} = \mathbf{g}_{\text{SH}}^{\text{SCH}}$. This means that the four-velocity in Schwarzschild–Hilbert coordinates is

$$\mathbf{U}_{\text{SH}} = \mathcal{L}_{\text{FIX} \rightarrow \text{SH}} \mathbf{U}_{\text{FIX}} = \begin{pmatrix} \left(1 - \frac{2GM}{c^2 r}\right)^{-1/2} \gamma \\ \left(1 - \frac{2GM}{c^2 r}\right)^{1/2} \gamma V^{\hat{r}} \\ \gamma V^{\hat{\theta}}/r \\ \gamma V^{\hat{\phi}}/(r \sin \theta) \end{pmatrix} \quad (7.29)$$

with its corresponding 1-form

$$\begin{aligned} \mathbf{u}_{\text{SH}} &= \mathbf{U}_{\text{SH}}^T \mathbf{g}_{\text{SH}}^{\text{SCH}} \\ &= \left(-\left(1 - \frac{2GM}{c^2 r}\right)^{1/2} \gamma c^2 \quad \left(1 - \frac{2GM}{c^2 r}\right)^{-1/2} \gamma V^{\hat{r}} \quad \gamma V^{\hat{\theta}} r \quad \gamma V^{\hat{\phi}} r \sin \theta \right) \end{aligned}$$

We then insert $\mathbf{U}_{\text{SH}}$ into the four-dimensional equations of motion and arrive at a set of laws expressed in both the variables *and* coordinates that we want.

Before continuing, we should take a good look at the velocity components in the Schwarzschild–Hilbert system. As pointed out before, only one (U^r) has units of centimeters per second. The rest have strange units. U^ϕ is the most familiar. It is the angular velocity in the ϕ direction, measured in radians per second, but increased by the Lorentz factor. And U^θ is the Lorentz-enhanced angular velocity in the θ direction.

U^r is curious. It is Lorentz-enhanced, *but gravitationally retarded* by the factor $(1 - 2GM/c^2 r)^{1/2}$. Near the Schwarzschild radius, this factor approaches zero, so the radial velocity measured in the Schwarzschild–Hilbert system slows down to a crawl! How can this be? Should not speed increase toward the speed of light as one approaches the black hole horizon? The answer is that it depends on what frame one is in when one measures speed. In the fixed system, this increase in speed occurs: $V^{\hat{r}}$ approaches c and the proper velocity $U^{\hat{r}}$ approaches infinity. However, the Schwarzschild–Hilbert coordinate system uses the time t as measured by someone far away from the black hole, and the metric in equation (7.25) tells us that time ticks more slowly in this coordinate system near the black hole. So objects falling in will appear from outside a black hole to stop near the horizon and never cross it.

Finally, the fact that U^t goes to infinity at the horizon should not be any cause for alarm. The same thing happens in flat space as $\mathbf{V}$ approaches the speed of light. This is the time dilation effect, and it reflects the fact that time appears to tick more slowly for someone near the horizon when viewed from outside.

7.4.2.2 The Equations of Motion and Gravitational Force

At the beginning of this chapter we pointed out that gravity is generated by the curvature in the metric (*e.g.*, equation (7.25)), through the curvilinear terms generated by the covariant gradient ∇ in Newton's second law of motion. But, how is this specifically accomplished mathematically? The answer is that gravity enters through the Christoffel symbol terms in the covariant derivative. Applying the methods for taking covariant gradients that are outlined in Appendix B to the equation of motion for a single particle (7.3) we can write a general expression for those equations as

$$\left(\frac{d\mathbf{P}}{d\tau}\right)^\alpha = (\mathbf{U} \cdot \nabla \mathbf{P})^\alpha = \sum_\beta U^\beta \left[\frac{\partial P^\alpha}{\partial x^\beta} + \sum_\mu \Gamma^\alpha_{\mu\beta} P^\mu \right] = F^\alpha \quad (7.30)$$

It is the second term on the left that generates gravity. The process is identical to the generation of the centrifugal pseudo-force from derivatives of $g_{\phi\phi}$, but in this case the gravitational force comes mainly from g_{tt} . For metrics like the Schwarzschild one, which is diagonal with g_{tt} a function of r only, Newton's second law becomes for radial motion ($V^{\hat{\theta}} = V^{\hat{\phi}} = 0$)

$$U^t \frac{\partial P_{\text{FIX}}^{\hat{r}}}{\partial t} + U^r \frac{\partial P_{\text{FIX}}^{\hat{r}}}{\partial r} - \frac{P^t U^t}{g_{rr}^{1/2}} \frac{1}{2} \frac{\partial g_{tt}}{\partial r} = F^{\hat{r}} \quad (7.31)$$

where the right-hand side $F^{\hat{r}}$ contains true forces like electromagnetism and the left-hand side is the rate of change of momentum in the curved Schwarzschild spacetime. The first two terms on the left are simply the advective time derivative ($d/d\tau$) in the moving frame of $P_{\text{FIX}}^{\hat{r}} = \gamma m_0 V^{\hat{r}}$ for a body of rest mass m_0 . Using equation (7.29) to evaluate the third term and assuming that there are no other forces in the system ($F = 0$), the equation of motion becomes

$$\frac{d(\gamma m_0 V^{\hat{r}})}{d\tau} = - \frac{\gamma}{\left(1 - \frac{2GM}{c^2 r}\right)^{1/2}} \frac{GM\gamma m_0}{r^2} \quad (7.32)$$

This equation looks very much like Newton's law of motion, equation (5.17), except: (1) the mass is increased by its relativistic motion to γm_0 ; (2) the force has another factor of γ (from time dilation) because the time derivative $d/d\tau$ is in the frame of the moving body; and (3) there is a factor of $1/(1 - 2GM/c^2 r)^{1/2}$ multiplying the gravitational force. The latter factor says that, at the black hole horizon, the moving body feels an infinite acceleration relative to a local fixed frame – an irresistible force that is impossible to counteract. It will enter the black hole permanently.

7.4.2.3 Time Warps

It is often said about general relativity that matter warps space and creates the gravitational field. This is not really correct. The presence of matter warps *spacetime*, with the emphasis on the time part. From equation (7.31) we see that the radial gravitational force in Schwarzschild geometry comes directly from the warping of the time component of the metric g_{tt} . *The true time warp is gravity itself.*

This point is obvious to a space traveler. In the absence of stars or planets, the traveler moves in a straight line. However, as he nears one of these objects, his world line is modified to bend toward the star or planet. If his speed is above the escape velocity of that object, the traveler will simply have his course altered. But if it is less than the escape velocity, his world line will impact the object, altering his future.

We on the earth are trapped in a similar time warp. Except for the few astronauts who have orbited the earth or gone to the moon, the world lines of every human being always point toward the center of the earth. It is only the upward force of the earth's crust, whose matter is slightly compressed as we stand upon it, that interrupts our otherwise certain future to fall toward the center.

7.4.2.4 Conserved Quantities

In Section 6.2.3 we introduced 1-forms as the complements of vectors, and now we appear to have gotten rid of them in the 3+1 description of the conservation laws and expressed everything in terms of orthonormal vector components (with hatted indices). The reader must be wondering, “What good are the 1-forms anyway?” In the case of the momentum 1-forms, the answer is that they comprise some very important and familiar items: the conserved quantities.

Checking equation (7.25) again, we notice that the Schwarzschild metric is not a function of the coordinates t or ϕ . Then, in addition to not having any other forces $F^{\hat{r}}$, as we are assuming for the moment, the ϕ component of Newton's second law also has no gravitational force. In the Schwarzschild–Hilbert system it becomes

$$U^t \frac{\partial P_{\text{SH}}^\phi}{\partial t} + U^r \frac{\partial P_{\text{SH}}^\phi}{\partial r} + U^\theta \frac{\partial P_{\text{SH}}^\phi}{\partial \theta} + U^\phi \frac{\partial P_{\text{SH}}^\phi}{\partial \phi} + P_{\text{SH}}^\phi U^r \frac{2}{r} + P_{\text{SH}}^\phi U^\theta \frac{2 \cos \theta}{\sin \theta} = 0$$

or, because $r^2 \sin^2 \theta$ also is not a function of t or ϕ , simply

$$\frac{d \left(P_{\text{SH}}^\phi r^2 \sin^2 \theta \right)}{d\tau} \equiv \frac{d(p_{\phi, \text{SH}})}{d\tau} = 0$$

just like equation (7.32), but without any gravitational force. The quantity in parentheses is the azimuthal (ϕ) momentum 1-form and, under these conditions it is a constant for any motion around the black hole. We know it as the familiar *angular momentum*

$$\boxed{\mathcal{L} \equiv p_{\phi, \text{SH}} = \gamma m_0 V^{\hat{\phi}} r \sin \theta} \quad (7.33)$$

We can perform the same procedure for the t component of Newton's second law and find another conserved quantity, the *energy*

$$\boxed{\mathcal{E} \equiv -p_{t, \text{SH}} = \left(1 - \frac{2GM}{c^2 r}\right)^{1/2} \gamma m_0 c^2} \quad (7.34)$$

The constant $\mathcal{E}$ is sometimes called the “energy at infinity” because it is equal to $\gamma m_0 c^2$ as $r \rightarrow \infty$, but it is the same at any radius r and so can be used to calculate $\gamma(r)$. A moving body is bound to the black hole if $\mathcal{E} < m_0 c^2$ and unbound otherwise.

7.4.3 Motion in the Schwarzschild Metric

7.4.3.1 Free Fall

As an example of motion near a black hole, imagine the hole immersed in a cloud of dust – tiny particles, each of mass m_0 , which do not interact ($\mathbf{F} = 0$). After a while the ensemble of falling particles will reach a steady state in which the time derivative of their momenta at a given radius $\partial P_{\text{FIX}}^{\hat{r}} / \partial t$ in equation (7.31) is zero. The equation now can be integrated easily over dr to produce

$$\gamma \left(1 - \frac{2GM}{c^2 r}\right)^{1/2} = \mathcal{E} / m_0 c^2 \quad (7.35)$$

which is just the conservation of energy obtained by another means. In the case of free fall from rest at infinity ($\mathcal{E} = m_0 c^2$) we can solve for the velocity in a fixed frame at some distance r

$$V_{\text{ff,SH}}^{\hat{r}} = - \left(\frac{2GM}{r} \right)^{1/2} \quad (7.36)$$

For even a Schwarzschild black hole, then, the free-fall speed is precisely the same as we had for Newtonian gravity (equation (5.23)). And, as the dust crosses the black hole horizon, it reaches an infall speed of exactly the speed of light

$$V_{\text{ff,SH}}^{\hat{r}}(r_S) = - \left(\frac{2GM}{r_S} \right)^{1/2} = -c$$

Does $V^{\hat{r}}$ also reach this speed when free fall does not start from rest? Suppose we shoot a particle at a black hole at nearly the speed of light ($\mathcal{E} \gg m_0 c^2$). Does it still cross the horizon going at c ? The answer is yes, of course. We can solve equation (7.35) in this case as well to get the general result

$$V^{\hat{r}} = -c \left(\frac{m_0 c^2}{\mathcal{E}} \right) \left[\left(\frac{\mathcal{E}}{m_0 c^2} \right)^2 - 1 + \frac{2GM}{c^2 r} \right]^{1/2} \quad (7.37)$$

So, at $r = r_S$, $V^{\hat{r}}$ still exactly equals c . The same thing is true even if we drop an object from a radius very close to the horizon. It still is accelerated to exactly c when it crosses the horizon. In this case the object is bound, but $\mathcal{E}$ is still positive ($0 < \mathcal{E} / m_0 c^2 < 1$). So equation (7.37) still has $V^{\hat{r}} = c$ at $r = r_S$. *Anything crossing the horizon must do so at exactly the speed of light* (in Schwarzschild–Hilbert coordinates).

7.4.3.2 Photon Orbits

Material accreting onto a black hole does not necessarily fall directly in, however. Because it may have some angular momentum, the material may settle into a time-independent orbit about the hole. We again can use the conservation of momentum to investigate this as well.

For a circular orbit we will set $V^{\hat{r}} = V^{\hat{\theta}} = 0$, $\sin \theta = 1$ to work in the equatorial plane, and the time derivative of $V^{\hat{\phi}}$ to zero to study stable orbits. The equation of motion then becomes the usual balance between gravity and centrifugal force

$$V_{\text{orb,SH}}^{\hat{\phi}} = \left[\frac{GM}{r - r_S} \right]^{1/2} \quad (7.38)$$

Equation (7.38) has a surprising result. Unlike the radial infall speed, the orbital velocity is not the speed of light at the horizon. It is infinite! Instead, the orbital speed reaches c at a larger radius, called the “photon orbit radius”

$$r_{\text{ph,SH}} \equiv \frac{3GM}{c^2} \quad (7.39)$$

At this radius massless particles like photons are *completely* deflected by the black hole’s gravity, orbiting around the black hole. Any closer, however, and even light will be pulled in. Any further out than this, and the photons will spiral outward. Therefore, light circles around the hole in an unstable orbit at the radius $r_{\text{ph,SH}}$.

7.4.3.3 Particle Orbits and the Innermost Stable Circular Orbit

For particles with mass, however, we must always have $V^{\hat{\phi}} < c$, so there must be an orbital Lorentz factor corresponding to the azimuthal velocity in equation (7.38)

$$\begin{aligned} \gamma_{\text{orb,SH}} &= \left(1 - \frac{V_{\text{orb,SH}}^2}{c^2} \right)^{-1/2} \\ &= \left[\frac{r - r_S}{r - \frac{3}{2}r_S} \right]^{1/2} \end{aligned}$$

which becomes infinite at $r_{\text{ph,SH}}$, of course. Inserting this Lorentz factor in equations (7.33) and (7.34), we have the orbital angular momentum and energy for a particle that has mass

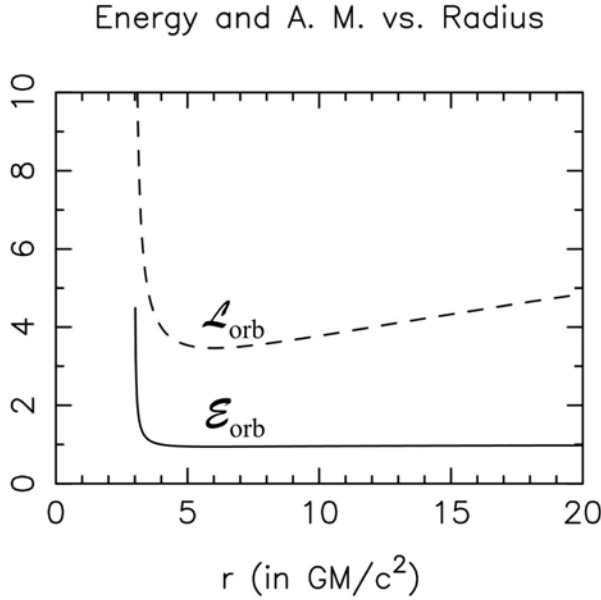


Fig. 7.1: Plot of particle specific energy $\mathcal{E}_{\text{orb}}/m_0 c^2$ (eq. (7.41)) and specific angular momentum $\mathcal{L}_{\text{orb}}/m_0 c$ (eq. (7.40)) for circular orbits as a function of the Schwarzschild–Hilbert radial coordinate r . Both functions have a minimum at $r = 3 r_S = 6 GM/c^2$ (eq. (7.42), the innermost stable circular orbit). At $r = 2 r_S$ (eq. (7.44), the marginally bound orbit) the particle energy climbs back to $\mathcal{E}_{\text{orb}} = m_0 c^2$. And at $r = \frac{3}{2} r_S$ (eq. (7.39), the photon orbit) both functions climb to infinity, since the Lorentz factor diverges there as well.

$$\mathcal{L}_{\text{orb,SH}} = \left(\frac{r_S}{2r - 3r_S} \right)^{1/2} m_0 r c \quad (7.40)$$

$$\mathcal{E}_{\text{orb,SH}} = \frac{(r - r_S)}{[r(r - \frac{3}{2}r_S)]^{1/2}} m_0 c^2 \quad (7.41)$$

Surprisingly, each of these has a minimum at a radius known as the “innermost stable circular orbit” (ISCO; see Fig. 7.1)

$$r_{\text{isco,SH}} \equiv \frac{6GM}{c^2} = 3 r_S \quad (7.42)$$

The minimum angular momentum is $\mathcal{L}_{\text{isco,SH}} = 2\sqrt{3} G M m_0/c$ and the minimum energy is $\mathcal{E}_{\text{isco,SH}} = 2\sqrt{2} m_0 c^2/3 = 0.9428 m_0 c^2$. This includes the rest mass energy, or course, so the binding energy of a particle at the ISCO is

$$\mathcal{E}_{\text{b,SH}} = -(\mathcal{E}_{\text{isco,SH}} - m_0 c^2) = 0.0572 m_0 c^2 \quad (7.43)$$

So, ignoring any further release of binding energy inside the ISCO, matter accreting onto a Schwarzschild black hole can give up about 6% of its rest energy as other forms (heat, radiation, electromagnetic fields or kinetic energy) before plunging into the hole. This is considerably more than produced by thermonuclear burning of an equivalent amount of matter (only $\sim 0.7\%$). Except for particle/anti-particle annihilations, accretion onto a black hole is the most efficient form of energy production known.

7.4.3.4 The Marginally Bound Orbit

There is one additional orbital radius of interest. Because $\mathcal{E}_{\text{orb,SH}}$ has a minimum at $r_{\text{isco,SH}}$ and then rises as r decreases further, there must be radius $r_{\text{mb,SH}} < r_{\text{isco,SH}}$ where the orbital energy (equation (7.41)) rises back to the rest energy of the particle $m_0 c^2$ (see Fig. 7.1). This radius is given by

$$r_{\text{mb,SH}} = 4 \frac{GM}{c^2} \quad (7.44)$$

Between $r_{\text{isco,SH}}$ and $r_{\text{mb,SH}}$ orbits can exist, but they are unstable. Interior to $r_{\text{mb,SH}}$ the binding energy is negative: more than the rest energy of the particle is needed just to keep it in a circular orbit, making these orbits very difficult to achieve. Like all orbits with $r < r_{\text{isco,SH}}$, these orbits also are unstable. Generally, all particles in this region with finite mass will spiral into the black hole, unless a very powerful acceleration can move them into higher, safer orbits.

7.4.4 Only Schwarzschild–Hilbert Coordinates are Singular at the Horizon, Not the Schwarzschild Metric

The apparent singularity in the r coordinate of the Schwarzschild metric is a big problem, especially in computer simulations that simulate gas flows around, and into, black holes. While matter in *physical* space can flow easily into black holes, matter in simulations in Schwarzschild–Hilbert coordinates never reaches the horizon. Often numerical problems prevent the simulation from continuing past, say, a hundred or so Schwarzschild times

$$\tau_S \equiv r_S/c$$

However, while strange things do happen near the horizon in the Schwarzschild metric, a physical singularity is not one of them. The problem turns out to be all in the coordinates (particularly in time, actually) that are chosen to express the geometry, not the geometry itself.

We can see this easily by choosing a different time coordinate t' , such that

$$dt' = dt + \frac{r_S}{c(r - r_S)} dr$$

This coordinate transformation is a global one (because the second term above can be integrated over r). It converts the Schwarzschild metric from equation (7.25) to

$$g_{\text{HP}}^{\text{SCH}} = \begin{pmatrix} -c^2 \left(1 - \frac{r_S}{r}\right) & \frac{r_S}{r} c & 0 & 0 \\ \frac{r_S}{r} c & \left(1 + \frac{r_S}{r}\right) & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \theta \end{pmatrix} \quad (7.45)$$

where now the first row and column are the $g_{t'\mu}$ coefficients of the line element, not $g_{t\mu}$. The coordinate singularity at r_S now has “magically” gone away. However, the physical singularity at $r = 0$ remains, as does the horizon ($g_{t't'} = 0$) at $r = r_S$. This form of the metric also approaches the spherical-polar form of the Minkowski metric as r becomes much larger than r_S . But, for small r we now have a new, off-diagonal coefficient: $g_{t'r} = (r_S/r)c$ that has helped remove the coordinate singularity.

The most surprising thing about this transformation is that r is *still* the same radial coordinate as before; only the definition of time has changed. Yet the proper distance no longer diverges at $r = r_S$

$$s' = r \left(1 + \frac{r_S}{r}\right)^{1/2} + r_S \ln \left\{ \left(\frac{r}{r_S}\right)^{1/2} \left[1 + \left(1 + \frac{r_S}{r}\right)^{1/2}\right] \right\}$$

which has the value $s' = [\sqrt{2} + \ln(1 + \sqrt{2})] r_S = 2.296 r_S$ there. Instead, s' now measures the distance from the *physical* singularity at the black hole center out to the radius r , not from the Schwarzschild radius. Why is the proper distance different even though r is the same coordinate? The reason is that s and s' are measured on a “time slice” where $dt = 0$ or $dt' = 0$. These latter two different conditions, in two different coordinate systems, generate two different equations for the proper distance on a time slice. Numerical simulations that use this new Schwarzschild metric will have no trouble with matter flowing into the black hole.

The metric in equation (7.45) is an example where the coordinates “drift” with time. This occurs whenever any of the $g_{t'i}$ components are non-zero. In this case only $g_{t'r}$ is non-zero, so the drift must be in r only; it turns out that the drift here is inward. Had we chosen $dt' = dt - [r_S/c(r - r_S)] dr$, we would have obtained an *outgoing* system. See Section 7.7 for more details.

This transformation to ingoing coordinates illustrates an important point that sometimes confuses people who are first introduced to the field. The Schwarzschild metric is usually written in what are sometimes called “Schwarzschild” (rather than “Schwarzschild–Hilbert”) coordinates (equation (7.25)). That is, the same name of-

ten is used to describe both the metric (the geometry) and the coordinate system, and the two then are used interchangeably. However, as we discovered earlier, even for flat metrics, geometry and coordinates are very different. Indeed, equation (7.45) still represents the geometry of the Schwarzschild metric, but it is written in inward-drifting coordinates. Here we have used the same superscript (SCH) on g' to indicate that the geometry is still Schwarzschild, but used a prime mark and “HP” to indicate that the coordinates have been transformed to horizon-penetrating ones.

7.5 Four-Dimensional Kerr Geometry: Rotating Black Holes

Like other stars, black holes can rotate. The difference here, is that black hole rotational speeds can be near the speed of light, causing still more changes in the metric and in the way matter moves near a black hole. Space itself around a black hole can rotate, and it is this rotation of space that, very possibly, is the ultimate cause for some of the powerful jets of matter that we see being ejected from quasars and other black hole systems. The equations in this case will be more complicated than in the non-rotating case, but the reward is that we will be able to understand how black holes can produce the various kinds of engines discussed in Chapters 2 to 4.

7.5.1 The Kerr Metric in Boyer–Lindquist Coordinates

The metric for a rotating black hole was worked out by New Zealand mathematician Roy Kerr, then at the University of Texas at Austin, in 1963 [314]. It is sufficiently complicated that here we need to write it in terms of intermediate functions of the basic spherical-polar coordinates (t, r, θ, ϕ)

$$g_{\text{BL}}^{\text{KER}} = \begin{pmatrix} -\left(1 - \frac{2r_g r}{\rho^2}\right) c^2 & 0 & 0 & -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} \\ 0 & \frac{\rho^2}{\Delta} & 0 & 0 \\ 0 & 0 & \rho^2 & 0 \\ -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} & 0 & 0 & \frac{\Sigma^2}{\rho^2} \sin^2 \theta \end{pmatrix} \quad (7.46)$$

where ρ is a radial distance, like r , but surfaces of constant ρ are oblate

$$\rho \equiv (r^2 + j^2 r_g^2 \cos^2 \theta)^{1/2}$$

Σ is an area quantity related to the equatorial area

$$\Sigma \equiv \left[(r^2 + j^2 r_g^2)^2 - j^2 r_g^2 \Delta \sin^2 \theta \right]^{1/2}$$

and $\Sigma \sin \theta / \rho$ is the cylindrical distance from the rotation axis. Δ is a slightly more complicated expression for the Schwarzschild factor in the tt and rr components of equation (7.25)

$$\Delta \equiv r^2 - 2r_g r + r_g^2 j^2$$

ω is the angular velocity of space itself at different points around the black hole

$$\omega \equiv \frac{2 r_g^2 r c}{\Sigma^2} j \quad (7.47)$$

j is the dimensionless angular momentum parameter ($-1 \leq j \leq 1$) that measures how much angular momentum J the black hole has relative to its maximum value

$$j \equiv \frac{J}{(GM^2/c)}$$

r_g is the “gravitational radius”, equal to one-half Schwarzschild radius

$$r_g \equiv \frac{GM}{c^2}$$

and M is again the mass of the black hole, which includes both the matter that originally fell into the black hole (the “irreducible” mass) and the rotational energy that hole has (the “reducible” mass).⁹

The Kerr metric (KER) is the rotating generalization of the Schwarzschild metric, and the form above is written in Boyer–Lindquist coordinates (BL, the generalization of Schwarzschild–Hilbert coordinates) after Robert H. Boyer (University of Liverpool, UK) and Richard W. Lindquist (Wesleyan University, USA) who published a detailed analysis of this metric in 1967 while visiting The University of Texas [315].¹⁰

⁹ We have a few important notes on the j parameter. First, there is little consistency in the literature and books for what to call the dimensionless angular momentum. Some define an $a \equiv J/(Mc)$, which has units of length, and then a dimensionless $a^* \equiv a/r_g$. Others drop the asterisk on the latter parameter and use a for the dimensionless quantity, thereby conflicting with others in the literature. Here we use j throughout the book (and in papers we publish) as the dimensionless expression of J , just as m is for M , $\dot{m}$ is for $\dot{M}$, *etc.* Second, most authors take the range of j to be $0 \leq j \leq 1$ and then treat orbits of individual particles as being prograde or retrograde. However, in this book we find it both mathematically, and astrophysically, useful to envision the black hole as being at the center of a stationary (*e.g.*, Boyer–Lindquist) coordinate system in which many other objects (stars, gas, *etc.*) are moving. In that case the black hole can be rotating in a prograde ($0 \leq j \leq 1$) or retrograde ($-1 \leq j \leq 0$) fashion in that system, and orbits *still* can be prograde or retrograde also within that system. Therefore, here we allow j to have the full range from -1 to $+1$. This also still handles all four combinations of ($j < 0, j > 0$) and ($\Omega_{\text{orb}} < 0, \Omega_{\text{orb}} > 0$).

¹⁰ It is not widely known that Dr. Boyer, age 33, was one of the many victims of the infamous sniper attack on August 1, 1966, in Austin, TX – less than two weeks after this paper was submitted. The

This is a rather complicated metric to work with, so we should discuss the simple things first. In the special case of a non-rotating black hole ($\omega \propto j = 0$), we find that

$$\rho = r \quad \Sigma = r^2 \quad \Delta = (r - 2r_g)r$$

which gives us the Schwarzschild metric (equation (7.25)), as we expect.

7.5.1.1 The Horizon and Ergosphere

The first thing we want to do is locate the horizon, the surface of the black hole. For the Schwarzschild metric it was where $g_{tt} = 0$ and $g_{rr} \rightarrow \infty$, *i.e.*, where $(1 - \frac{2GM}{c^2 r}) = 0$. But for the Kerr metric these two conditions do not occur at the same place. The condition $g_{rr} \rightarrow \infty$ occurs when $\Delta = 0$, or at the radius

$$r_H = \left(1 + \sqrt{1 - j^2}\right) r_g \quad (7.48)$$

which is, in fact, the horizon. So, even for a rotating black hole, the horizon is still spherical. On the other hand, $g_{tt} = 0$ at the radius

$$r_E \equiv \left(1 + \sqrt{1 - j^2 \cos^2 \theta}\right) r_g \quad (7.49)$$

which is larger than the horizon radius for all θ and ϕ , except at the poles ($\theta = 0$ or π) where the two are equal. Note that at the equator, $r_E = 2r_g$ always, for *any* black hole spin.

Something funny has happened to time in the region $r_H < r < r_E$. Both g_{tt} and g_{rr} are positive in that region; there is no negative diagonal component. So what happened to time? Is this a four-dimensional Euclidean region? No, time is alive and well and pointing in the ϕ direction (in the direction of the black hole rotation), and is being carried by the *off*-diagonal element $g_{\phi t}$. *In this region, called the “ergosphere”, everything must rotate in the same direction as the black hole*, just as interior to the horizon everything must fall toward the black hole. Anything not rotating in this region is moving against the rotating metric at a speed faster than c , which is impossible. Unlike the horizon, however, in the ergosphere it still is possible to escape the black hole’s gravity; but for a rapidly-rotating black hole one must travel faster than $0.91c$ in order to do so!

The rotation of space with the black hole is not confined to the ergosphere; it is just inevitable there. However, the rotation rate ω does drop off as $1/r^3$ beyond the ergosphere, so most of the action occurs near or inside this region.

paper was published posthumously. The name of the sniper often is publicized, but it is rare that the victims themselves are honored. We do the latter here.

7.5.1.2 Reducible and Irreducible Mass

If rotational energy were to be extracted from a spinning black hole, it not only would spin down, it also would lose the gravitational mass of that lost spin energy. When this process is complete, and the hole is left in a non-spinning state, the mass remaining is its “irreducible” mass¹¹

$$M_{\text{ir}} = M \left[\left(1 + \sqrt{1 - j^2} \right) / 2 \right]^{1/2} \quad (7.50)$$

which is proportional to the square root of the horizon area $4\pi(r_{\text{H}}^2 + j^2 r_g^2)$. The total available rotational energy that can be extracted, therefore, is $M_{\text{red}} c^2$, where $M_{\text{red}} \equiv M - M_{\text{ir}}$ is the amount of “reducible” mass in the spinning hole. Note that M_{red} is not the same as the mass that can be formed from the (angular) momentum of the spinning black hole: $M_P = Jc/(2GM_{\text{ir}})$. The latter adds in quadrature with M_{ir} to give $M^2 = M_P^2 + M_{\text{ir}}^2$, similar to equation (6.64).

7.5.1.3 Physics in Curved and Rotating Kerr Geometry

In analyzing physics in Kerr geometry, we note that the concept of a FIX reference frame is now somewhat broader. The frame still is fixed with respect to the local spacetime, but we now recognize that this is true even if the spacetime is dragging its inertial frames. That is, as viewed from a large distance from a Kerr black hole, a fiducial observer at (r, θ, ϕ) will appear to rotate with the spacetime at the rate ω , given by equation (7.47). Despite this apparent rotation, however, within the Kerr metric this observer will have *no* measured angular momentum with respect to the black hole. For this reason, such a coordinate system is often called a “locally non-rotating reference frame” (LNRF), and Kerr fiducial observers are often called a ZAMOs (“zero angular momentum observers”) [305]. Remember that each FIX/ZAMO frame still is a locally-Lorentz one, with its own local Minkowski metric. While vector quantities typically will be expressed in this frame, their derivatives will be taken in a global frame like the Boyer–Lindquist one.

The “Observer-at-Infinity/Synchronous” Coordinate System (OIS)

In addition to the three systems introduced for Schwarzschild geometry, we now have a fourth coordinate system in which to express the conservation of energy and momentum. This system rotates with the space around the black hole, is *local* to each radius r , and has a diagonal metric; but that metric is not an orthonormal Lorentz/Minkowskian system. We can transform the Kerr metric to this system with

¹¹ Although not very common, it is possible to think of M_{ir} as the invariant “rest mass” of a given black hole and the gravitational mass $M = M_{\text{ir}}/[(1 + \sqrt{1 - j^2})/2]^{1/2}$ as a quantity that increases or decreases as the spin j of that black hole changes.

a simple local Lorentz transformation that changes the ϕ variable only: $d\phi' = d\phi - \omega dt$. Shinji Koide, of Kumamoto University in Japan, calls this the “observer at infinity/synchronous” or OIS frame. The Lorentz transformation for this is simply

$$\mathcal{L}_{\text{BL} \rightarrow \text{OIS}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\omega & 0 & 0 & 1 \end{pmatrix}$$

so that $(\mathcal{L}_{\text{BL} \rightarrow \text{OIS}}^{-1})^T \mathbf{g}_{\text{BL}}^{\text{KER}} \mathcal{L}_{\text{BL} \rightarrow \text{OIS}}^{-1} = \mathbf{g}_{\text{OIS}}^{\text{KER}}$, yielding

$$\mathbf{g}_{\text{OIS}}^{\text{KER}} = \begin{pmatrix} -\frac{\rho^2}{\Sigma^2} \Delta & 0 & 0 & 0 \\ 0 & \frac{\rho^2}{\Delta} & 0 & 0 \\ 0 & 0 & \rho^2 & 0 \\ 0 & 0 & 0 & \frac{\Sigma^2}{\rho^2} \sin^2 \theta \end{pmatrix}$$

We see now that in this system, g_{tt} and g_{rr} indeed do change sign at the horizon (where $\Delta = 0$). The metric is nice and diagonal in this frame, making it easy to write down the velocity vector and 1-form in terms of quantities in the moving-body (MOV) or fixed systems (FIX)

$$\mathbf{U}_{\text{OIS}} = \begin{pmatrix} \gamma \Sigma / (\rho \Delta^{1/2}) \\ \gamma V^{\hat{r}} \Delta^{1/2} / \rho \\ \gamma V^{\hat{\theta}} / \rho \\ \gamma V^{\hat{\phi}} \rho / (\Sigma \sin \theta) \end{pmatrix}$$

$$\mathbf{u}_{\text{OIS}} = \left(-\gamma \frac{\rho \Delta^{1/2}}{\Sigma} \quad \gamma V^{\hat{r}} \frac{\rho}{\Delta^{1/2}} \quad \gamma V^{\hat{\theta}} \rho \quad \gamma V^{\hat{\phi}} \frac{\Sigma \sin \theta}{\rho} \right)$$

However, this is not a good system in which to work. It is not global, so derivatives are valid only locally. And the ergosphere is hidden in this frame because ω is gone from $\mathbf{g}_{\text{OIS}}$. So we quickly leave the OIS system and transform the vectors and 1-forms back to the Kerr system to find

$$\mathbf{U}_{\text{BL}} = \mathcal{L}_{\text{BL} \rightarrow \text{OIS}}^{-1} \mathbf{U}_{\text{OIS}} = \begin{pmatrix} \gamma \Sigma / (\rho \Delta^{1/2}) \\ \gamma V^{\hat{r}} \Delta^{1/2} / \rho \\ \gamma V^{\hat{\theta}} / \rho \\ \frac{\gamma V^{\hat{\phi}} \rho}{\Sigma \sin \theta} + \frac{\gamma \Sigma \omega}{\rho \Delta^{1/2}} \end{pmatrix} \quad (7.51)$$

$$\mathbf{u}_{\text{BL}} = \left(-\left[\gamma c^2 \frac{\rho \Delta^{1/2}}{\Sigma} + \gamma V^{\hat{\phi}} \frac{\Sigma \sin \theta}{\rho} \omega \right] \quad \gamma \frac{\rho}{\Delta^{1/2}} V^{\hat{r}} \quad \gamma \rho V^{\hat{\theta}} \quad \gamma \frac{\Sigma \sin \theta}{\rho} V^{\hat{\phi}} \right) \quad (7.52)$$

As in the Schwarzschild case, the three-velocity components $\mathbf{V} = (V^{\hat{r}}, V^{\hat{\theta}}, V^{\hat{\phi}})$ are all in units of centimeters per second and evaluated in the local FIX orthonormal

Minkowski frame. But that frame rotates with the black hole, so we need to include a rotation (proportional to ω) in the angular velocity U^ϕ .

Conserved Quantities and “Negative Energy” Trajectories

Also, as in the Schwarzschild case, because the Kerr metric is not a function of ϕ and t , we can read off the conserved quantities from the azimuthal and time momentum 1-forms in equation (7.52). For a particle of mass m_0 moving in the global Kerr frame they are

$$\mathcal{L} \equiv m_0 u_{\phi, \text{BL}} = \gamma m_0 V^{\hat{\phi}} \frac{\Sigma}{\rho} \sin \theta \quad (7.53)$$

$$\mathcal{E} \equiv -m_0 u_{t, \text{BL}} = \frac{\rho \Delta^{1/2}}{\Sigma} \gamma m_0 c^2 + \gamma m_0 V^{\hat{\phi}} \frac{\Sigma \sin \theta}{\rho} \omega \quad (7.54)$$

Equation (7.53) for the angular momentum is essentially the Schwarzschild one (7.33) but with the new cylindrical radius, and it is indeed zero when $V^{\hat{\phi}}$ (in the ZAMO frame) is zero. The first term of equation (7.54) for the energy at infinity is essentially equation (7.34). However, the second term is new in the Kerr metric and is due to the rotation of space itself. If a particle falling toward the black hole has a velocity $V^{\hat{\phi}}$ in the azimuthal direction that is in the *same* direction as the hole’s rotation, then as seen from afar that motion adds a little rotational energy to the black hole system (in addition to the Lorentz enhancement). When that particle merges with the black hole, that rotational energy becomes part of the hole. On the other hand, if $V^{\hat{\phi}}$ is in the opposite direction to ω , then that motion will *reduce* the rotational energy of the system a bit. But, as long as $\mathcal{E} > 0$, accretion of the particle by the black hole will increase the mass of the black hole.

A curious thing happens, however, when the velocity $V^{\hat{\phi}}$ in the ϕ direction of an infalling particle is largely negative, that is, when

$$-\frac{V^{\hat{\phi}}}{c} > \frac{\rho^2 \Delta^{1/2}}{\Sigma^2 \sin \theta} \frac{c}{\omega} = \frac{\rho^2 \Delta^{1/2}}{2j r_g^2 r \sin \theta}$$

The particle’s energy at infinity $\mathcal{E}$ is negative and *accretion of this particle will decrease the black hole’s mass*. An infalling particle can get into one of these “negative energy trajectories” only by being violently scattered or accelerated into that state. *But in that process, the particle will have given up more than all of its rest energy $m_0 c^2$ to the world outside the black hole system.* For a rotating black hole, the efficiency of accretion can be greater than 100% for some of the material accreted. The extra energy, of course, comes from the hole as it loses a bit of its rotational mass. Because $V^{\hat{\phi}}$ cannot be greater than c , these negative energy trajectories can exist only within the radius where $\rho^2 \Delta^{1/2} < 2j r_g^2 r \sin \theta$, or $r < r_E$, given by equation (7.49). So negative energies can exist inside, and only inside, the ergosphere of any

rotating black hole (with $j > 0$). The process of extracting rotational mass-energy from the ergosphere of a black hole was first recognized by Roger Penrose of the University of Oxford. As a result, it often is referred to as the “Penrose Process”.

7.5.1.4 Motion in the Kerr Metric

Free Fall

The trajectory of a freely falling body is nearly as easy to compute in the Kerr case as it was in the Schwarzschild case. Because $\mathcal{L}$ is conserved, if the body has $V^{\hat{\phi}} = 0$ when it starts falling, it will have $V^{\hat{\phi}} = 0$ always. But we now need to be a little careful. $V^{\hat{\phi}}$ is the azimuthal velocity measured relative to the rotating space around the Kerr black hole. So, the falling body does pick up some angular velocity – exactly that of the rotating space. That is, as seen from afar, the angular velocity of the body is the ϕ component of equation (7.51)

$$U_{\text{ff,BL}}^{\phi} = \frac{\gamma \Sigma \omega}{\rho \Delta^{1/2}} \quad (7.55)$$

Having determined the azimuthal velocity from the conservation of angular momentum, we also can calculate the free fall radial velocity from the conservation of energy

$$V_{\text{ff,BL}}^{\hat{r}} = - \left(\frac{2GM}{r} \right)^{1/2} \frac{r (r^2 + j^2 r_g^2)^{1/2}}{\Sigma} \quad (7.56)$$

which is now slightly different from the Newtonian/Schwarzschild velocity. If we plug equation (7.48) for the horizon radius into equation (7.56), we find that indeed the infall velocity equals c everywhere along the horizon. However, as the body falls toward the black hole, its approach to this asymptotic value is a little different than in the non-rotating black hole case (equation (7.36)). The reason is that the horizon for a rotating hole is smaller in radius (in units of r_g) than that of a non-rotating one.

Orbits in the Equatorial Plane

Computing orbital motion in Kerr geometry is much more difficult than free fall, even for the simple case of circular orbits ($V^{\hat{r}} = V^{\hat{\theta}} = 0$) in the equatorial plane ($\theta = \pi/2$). It is done in the same manner as we did for Schwarzschild black holes: balance centrifugal force and gravity in Newton’s second law (written in Kerr geometry) and solve for the orbital velocity. After much careful algebra, best done these days with a computer program like Mathematica or Maple, we obtain¹²

¹² S. Koide, private communication.

$$\frac{V_{\text{orb,BL}}^{\hat{\phi}}}{c} = \frac{\Sigma^2}{\Delta^{1/2} (r^3 - j^2 r_g^3)} \left[\pm \sqrt{\frac{r_g}{r}} - j \frac{r_g^2}{r^2} \right] - \frac{2j r_g^2}{r \Delta^{1/2}} \quad (7.57)$$

The first term in the brackets is the familiar Keplerian rotation, in the rotating space of the black hole. It can be positive (prograde orbits in the global coordinate system) or negative (retrograde). The second term in the brackets comes from an outward centrifugal effect due to the rotation of space. And the last term on the right arises because we measure $V^{\hat{\phi}}$ in the rotating frame, so the velocity of the metric must be subtracted off. Setting $j = 0$ in equation (7.57) recovers the Schwarzschild orbital velocity in equation (7.38).

The photon orbit radius occurs when $V^{\hat{\phi}} = \pm c$ or

$$r_{\text{ph,BL}} = 2r_g \left\{ 1 + \cos \left[\frac{2}{3} \cos^{-1} (\mp j) \right] \right\} \quad (7.58)$$

For $j = 0$ this reduces to the Schwarzschild value of $r_{\text{ph,BL}} = 3r_g$, but for $j \neq 0$ we see that there also are *two* values, corresponding to the prograde photon orbit (the upper [negative] sign) and the retrograde orbit (the lower [positive] sign). For a maximal Kerr hole ($j = 1$) these are $r_{\text{ph,BL}} = r_g$ and $4r_g$, respectively. (For a hole spinning in the opposite direction [$j = -1$], they are just the opposite, $4r_g$ and r_g .) A photon in a prograde orbit is assisted in its orbital motion by the rotation of space, so it easily escapes until very close to the hole. In a retrograde orbit, the photon travels against the black hole's rotation, so it is captured at a much larger radius. Interior to r_{ph} a photon spirals into the hole and outside it spirals outward.

Moving outward from the photon orbit, we find the marginally-bound particle orbit, which also must have a prograde and retrograde value. Setting $\mathcal{E} = m_0 c^2$ in equation (7.54) we find

$$r_{\text{mb,BL}} = \left(1 + \sqrt{1 \mp j} \right)^2 r_g \quad (7.59)$$

where the upper sign is again the prograde orbit case. For $j = 1$, these radii are at $r_{\text{mb,BL}} = r_g$ and $5.828 r_g$, respectively.

The innermost stable circular orbit radius takes more effort to compute: equation (7.57) for $V^{\hat{\phi}}$ must be inserted into the angular momentum equation (7.53) (including into γ), the derivative of the entire expression set to zero, and solved for r_{isco} . The result is fairly complicated, as one might expect

$$r_{\text{isco,BL}} = \left[h \mp \left(\frac{-h^3 + 9h^2 - 3(6 - j^2)h + 7j^2}{h - 3} \right)^{1/2} \right] r_g \quad (7.60)$$

where the half-way point between the prograde (−) and retrograde (+) innermost stable orbits is

$$h \equiv \frac{r_{1/2}}{r_g} = 3 + \left\{ 3 + j^2 + (1 - j^2)^{1/3} \left[(3 - j)(1 + j)^{1/3} + (3 + j)(1 - j)^{1/3} \right] \right\}^{1/2}$$

in units of r_g . For a maximally spinning ($j = 1$) Kerr hole $r_{1/2} = 5r_g$, and the prograde and retrograde values are $r_{\text{isco,BL}} = r_g$ and $9r_g$. For $j = 0$ they all three degenerate to $6r_g$.

No black hole is likely to reach exactly the maximum spin, but some might reach very close to it, say $j = 0.998$, as suggested by Kip Thorne in 1974 [316]. [Table 7.1](#) and [Fig. 7.2](#) compare the various important radii *in the equatorial plane* in units of r_g for holes of different rotation rate.

Table 7.1: Important radii for Schwarzschild and Kerr black holes for $\theta = \pi/2$

Radius	Schwarzschild ($j = 0$)	Kerr ($j = 0.998$)		Kerr ($j = 0.99999$)	
r_H	2	1.0632		1.004472	
r_E	(2)	2		2	
	Prograde / Retrograde	Prograde	Retrograde	Prograde	Retrograde
r_{ph}	3	1.0739	3.9982	1.005168	3.999991
r_{mb}	4	1.0914	5.8250	1.006335	5.828410
r_{isco}	6	1.2370	8.9944	1.035232	8.999972

Technically, a non-rotating black hole has no ergosphere, but the equatorial ergospheric radius r_E is exactly equal to the Schwarzschild radius r_S for all black holes no matter how rapidly they rotate.

7.5.2 The Kerr Metric in Horizon-Penetrating Coordinates

As with the Schwarzschild metric, it is possible to remove the coordinate singularity in r at the horizon (where $\Delta = 0$) and allow the coordinate system to extend all the way to the center of the black hole. The problem coefficient is g_{rr} , which has Δ in its denominator. To perform such a singularity-removing coordinate conversion, however, requires transformation of both time *and* azimuthal angle

$$dt = dt' - \frac{2r_g r}{c \Delta} dr$$

$$d\phi = d\phi' - \frac{j r_g}{\Delta} dr$$

which again is a global coordinate transformation, because each second term above is integrable in r . These coordinate changes compress time near the black hole. The

Important Radii in Rotating Black Hole Structure

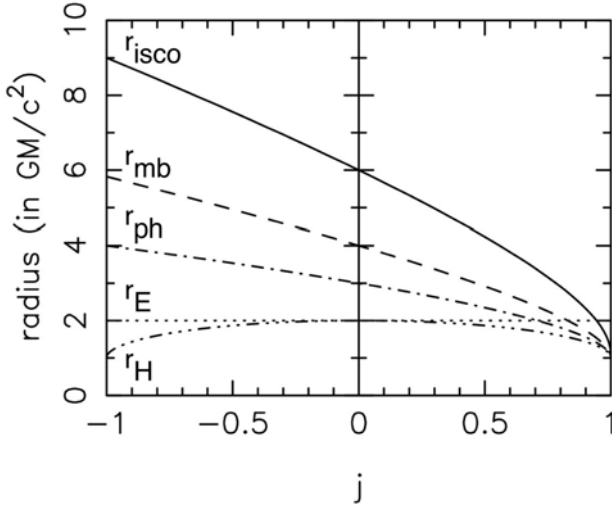


Fig. 7.2: Various important equatorial radii vs. black hole spin: ISCO (solid line; eq. (7.60)), marginally bound (dashed; eq. (7.59)), photon orbit (dash-dot; eq. (7.58)), ergosphere (dotted and always $2r_g$ on the equator; eq. (7.49)), and horizon (dash-double-dot; eq. (7.48)).

resulting metric is the Kerr generalization of equation (7.45)

$$g'_{\text{HP}}^{\text{KER}} = \begin{pmatrix} -\left(1 - \frac{2r_g r}{\rho^2}\right) c^2 & \frac{2r_g r}{\rho^2} c & 0 & -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} \\ \frac{2r_g r}{\rho^2} c & 1 + \frac{2r_g r}{\rho^2} & 0 & -j r_g \sin^2 \theta \left(1 + \frac{2r_g r}{\rho^2}\right) \\ 0 & 0 & \rho^2 & 0 \\ -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} & -j r_g \sin^2 \theta \left(1 + \frac{2r_g r}{\rho^2}\right) & 0 & \frac{\Sigma^2}{\rho^2} \sin^2 \theta \end{pmatrix} \quad (7.61)$$

Note that the only coefficients that have changed are $g_{t'r}$, g_{rr} , and $g_{r\phi'}$. This metric still is not a function of t' or of ϕ' , so energy $-p_{t',\text{HP}}$ and angular momentum $p_{\phi',\text{HP}}$ are still conserved quantities in this space.

7.6 Four-Dimensional Kerr–Newman Geometry: Rotating, Charged Black Holes

Microscopic black holes of, say 10^{13} g (the mass of a large hill of earth three football fields high), will have a Schwarzschild radius of ~ 1 femtometer ($\sim 10^{-13}$ cm) – the size of an elementary particle. Such black hole particles not only could have mass and spin, they also may have other quantum numbers, such as electric charge, color charge, flavor, *etc.* And the addition of each of these properties to a black hole will alter its stationary metric, just as adding spin converted the Schwarzschild metric into the Kerr one.

The same also is true of the microscopic black holes' much more massive cousins – the astrophysical black holes that house the engines discussed in this book. Because these engines appear to be electromagnetically driven, it will be necessary to consider at least the possibility that astrophysical black holes may have an electric charge. The generalization of the Kerr metric for a rotating, charged black hole is the Kerr–Newman (KER–NEW) one, named jointly after E. T. Newman who, with his colleagues at the University of Pittsburgh, derived it in 1965 [317].

To begin our discussion of this even more general metric, let us rewrite (7.46) to emphasize the quantity Δ

$$g_{\text{BL}}^{\text{KER-NEW}} = \begin{pmatrix} -\frac{(\Delta - j^2 r_g^2 \sin^2 \theta) c^2}{\rho^2} & 0 & 0 & -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} \\ 0 & \frac{\rho^2}{\Delta} & 0 & 0 \\ 0 & 0 & \rho^2 & 0 \\ -\frac{\omega \Sigma^2 \sin^2 \theta}{\rho^2} & 0 & 0 & \frac{\Sigma^2}{\rho^2} \sin^2 \theta \end{pmatrix} \quad (7.62)$$

and rewrite the equation for ω as

$$\omega \equiv \frac{r_g (r^2 + j^2 r_g^2 - \Delta) c}{\Sigma^2} j \quad (7.63)$$

So far, this is still the Kerr metric and still written in Boyer–Lindquist coordinates. To change it to the Kerr–Newman metric for a rotating, *charged* black hole, we need add only an additional charge term to Δ

$$\Delta_{\text{BL}}^{\text{KER-NEW}} \equiv r^2 - 2r_g r + (j^2 + q^2) r_g^2 \quad (7.64)$$

where the dimensionless charge is defined as the total charge Q on the black hole divided by the maximum value possible

$$q \equiv \frac{Q}{G^{1/2} M}$$

Note that, while q does lie in the range $-1 \leq q \leq 1$, just like j , both parameters cannot be ± 1 at the same time. The two are limited so that their quadrature sum is no more than

$$j^2 + q^2 \leq 1 \quad (7.65)$$

Furthermore, if the black hole is both spinning and charged, it must have a magnetic moment, and the value of that quantity is

$$\mu = \frac{Q J}{M c} = q j \frac{G^{3/2} M}{c^2} \quad (7.66)$$

Since $q \leq (1 - j^2)^{1/2}$, the maximum magnetic moment occurs when $q = j = 1/\sqrt{2}$, so $\mu_{max} = G^{3/2} M / (2 c^2)$.

Finally, the Kerr–Newman metric also can be written in horizon-penetrating coordinates, just like its Schwarzschild and Kerr special cases. As with the Kerr metric, we must transform both time and azimuthal angle to new versions

$$dt = dt' - \frac{(r^2 + j^2 r_g^2 - \Delta)}{c \Delta} dr$$

$$d\phi = d\phi' - \frac{j r_g}{\Delta} dr$$

This transformation again is global and removes the singularity in Boyer–Lindquist coordinates (equation (7.62)). But the Kerr–Newman metric in these coordinates has slightly more complex terms than in the regular Kerr version (equation (7.61)) because of electric charge effects.

As it turns out, it is very difficult for black holes in astrophysical environments to maintain enough charge to affect the operation of the black hole engine. (See Section 14.3.) So we will not have a lengthy discussion of the modification to the physics of particle orbits and electromagnetism that having a charged black hole might bring.

7.7 A Return to Physics in Three-Space Plus Time: Saying Goodbye to the Four-Dimensional Formalism

The four-dimensional formalism that we have been using in Chapter 6 and this chapter has many advantages. It is the true description of physics in our four-dimensional universe; it illustrates what things are the same no matter in which frame they are

measured (four-momentum, proper distance, *etc.*); and it shows that things thought previously to be invariant (time, length, *etc.*) depend, in fact, on who is measuring them. However, the four-dimensional formalism has a number of disadvantages. It introduces new four-dimensional tensors, like $\mathcal{F}$ and $\mathcal{T}$, whose components are sometimes difficult to visualize. Furthermore, it does not lend itself easily to numerical simulation or even semi-analytic computation: even in four dimensions, there always is a time-like coordinate through which structure in the other three dimensions evolves. Why spend so much time on a four-dimensional formalism if one still ends up having to divide it into three space-like dimensions plus a time-like one?

In recent years many physicists and astrophysicists have returned to thinking about physics in three-dimensional space and evolving in a global time coordinate – but with the new realization that our perception of space and time still depends on where the observer is and how fast he is moving. This “3+1” concept was first introduced for the very complex field of the numerical simulation of the time-dependent Einstein equations (Chapter 8) by Richard Arnowitt (Texas A & M University), Stanley Deser (Brandeis University), and Charles Misner (University of Maryland) in their 1962 book [318]. It is often called the ADM formalism after these three authors. The 3+1 formalism was championed by Kip Thorne, Richard Price, and Douglas Macdonald in their 1986 book on electrodynamics in stationary metrics (ones where no component of g is a function of the time coordinate t) [319]. Their particular version, however, had some features (*e.g.*, Lie derivatives) that made the equations somewhat complex and non-standard. In 2001 Shinji Koide (then at Toyama University, Japan) and his colleagues (Kazunari Shibata [Kyoto University], Takahiro Kudoh [NAOJ], and this author) published a significant simplification of the 3+1 formalism for stationary metrics whose three-dimensional portion g is diagonal [320, 321]. Finally, in 2004 Serguei Komissarov (Leeds University in Britain) published a further simplification of Koide’s equations, valid for more general stationary metrics, that made the fully general relativistic equations of electrodynamics look exactly like the original Maxwell equations [322].

Our treatment below is a hybrid of the formalisms of Koide and Komissarov. We use Koide’s form for the metric (rather than the ADM one used by Komissarov), because all of the physical equations from Chapter 9 onward will use that version. However, initially we will begin with a more general (non-diagonal) 3+1 metric, similar to Komissarov’s, and will simplify Maxwell’s equations to their standard form using the former’s definitions of the displacement D and magnetic induction B in stationary metrics. Furthermore, because the Kerr three-metric in Boyer–Lindquist coordinates is diagonal in this 3+1 form, we ultimately will be able to define expressions for orthonormal vector operators (*e.g.*, divergence and curl) that are identical to those used in undergraduate mathematics and physics courses. This not only allows us to use all of the physical intuition that we have gained from laboratory and real life experience (and familiar concepts like electric and magnetic fields), it also allows us to use the more familiar mathematics of three-dimensional vector calculus.

After some final usage of the four-dimensional formalism in the next chapter on gravitational waves and collapse, we will say goodbye to the use of contravariant

vectors and 1-forms and express all the equations of physics with the more familiar *orthonormal* vectors (indicated by hatted indices). This will allow us to compare directly with the non-relativistic physics of Chapter 5 and better understand the effects that the relativistic properties of black holes have on the engines they operate.

7.7.1 The General Non-Stationary “3+1” Metric

The “return to three-space” begins by writing the 4×4 spacetime metric as a symmetric 3×3 tensor $\mathbf{g}$ plus some time and spacetime components:

$$\mathbf{g}_{3+1} = \begin{pmatrix} -\alpha^2 c^2 (1 - \beta^2) & -\beta_1 \alpha c & -\beta_2 \alpha c & -\beta_3 \alpha c \\ -\beta_1 \alpha c & \mathbf{g}_{11} & \mathbf{g}_{12} & \mathbf{g}_{13} \\ -\beta_2 \alpha c & \mathbf{g}_{12} & \mathbf{g}_{22} & \mathbf{g}_{23} \\ -\beta_3 \alpha c & \mathbf{g}_{13} & \mathbf{g}_{23} & \mathbf{g}_{33} \end{pmatrix} \quad (7.67)$$

This metric is fully general, but has been split into space and time parts. $\mathbf{g}_{ij}$ is the three-dimensional metric at a local point in spacetime that describes the curvature of space only; β is the three-dimensional unit “drift” 1-form¹³ (with corresponding contravariant “drift vector” we will call $\mathcal{B}$) that describes in what direction the metric drags time through space (β is the magnitude of β); and α is the “lapse function” that describes the rate at which time passes locally. The determinant of this 3+1 metric has a nice, simple form

$$\sqrt{-g} = \alpha c g^{1/2} \quad (7.68)$$

where the scalar $g = ||\mathbf{g}||$ is the determinant of the three-metric.

The line element for this form of the metric is

$$ds^2 = -\alpha^2 c^2 dt^2 + \mathbf{dx}^T \cdot \mathbf{g} \cdot \mathbf{dx} \quad (7.69)$$

where the three-dimensional differential *vector* $\mathbf{dx}$ is

$$\mathbf{dx} = \begin{pmatrix} dx^1 - \mathcal{B}^1 \alpha c dt \\ dx^2 - \mathcal{B}^2 \alpha c dt \\ dx^3 - \mathcal{B}^3 \alpha c dt \end{pmatrix}$$

where $\mathcal{B} = \mathbf{g}^{-1} \cdot \beta$. It also is useful to have on hand the *inverse* metric in this general 3+1 notation

¹³ In the ADM formalism the g_{0i} components of the metric are *positive* in sign; the quantity associated with these then is called the “shift” 1-form, and it has an associated shift vector. The difference is one of definition and interpretation: the ADM shift vector describes how to *shift* spacetime coordinates so they do not *drift* with the dragging of inertial frames.

$$g_{3+1}^{-1} = \begin{pmatrix} -\frac{1}{\alpha^2 c^2} & -\frac{\mathcal{B}^1}{\alpha c} & -\frac{\mathcal{B}^2}{\alpha c} & -\frac{\mathcal{B}^3}{\alpha c} \\ -\frac{\mathcal{B}^1}{\alpha c} & g^{11} - \mathcal{B}^1 \mathcal{B}^1 & g^{12} - \mathcal{B}^1 \mathcal{B}^2 & g^{13} - \mathcal{B}^1 \mathcal{B}^3 \\ -\frac{\mathcal{B}^2}{\alpha c} & g^{21} - \mathcal{B}^2 \mathcal{B}^1 & g^{22} - \mathcal{B}^2 \mathcal{B}^2 & g^{23} - \mathcal{B}^2 \mathcal{B}^3 \\ -\frac{\mathcal{B}^3}{\alpha c} & g^{31} - \mathcal{B}^3 \mathcal{B}^1 & g^{23} - \mathcal{B}^3 \mathcal{B}^2 & g^{33} - \mathcal{B}^3 \mathcal{B}^3 \end{pmatrix} \quad (7.70)$$

where the g^{ij} are the elements of the 3×3 inverse spatial metric $\mathbf{g}^{-1}$.

The Kerr metric in Boyer–Lindquist coordinates, for example, can be written in this form with the following definitions¹⁴

$$\mathbf{g} = \begin{pmatrix} \frac{\rho^2}{\Delta} & 0 & 0 \\ 0 & \rho^2 & 0 \\ 0 & 0 & \frac{\Sigma^2}{\rho^2} \sin^2 \theta \end{pmatrix} \quad \mathbf{g}^{-1} = \begin{pmatrix} \frac{\Delta}{\rho^2} & 0 & 0 \\ 0 & \frac{1}{\rho^2} & 0 \\ 0 & 0 & \frac{\rho^2}{\Sigma^2 \sin^2 \theta} \end{pmatrix} \quad (7.71)$$

$$\mathcal{B} = \begin{pmatrix} 0 \\ 0 \\ \omega/(\alpha c) \end{pmatrix} \quad \beta = \mathcal{B}^T \cdot \mathbf{g} = \begin{pmatrix} 0 & 0 & \frac{\omega \Sigma^2 \sin^2 \theta}{\alpha \rho^2 c} \end{pmatrix} \quad (7.72)$$

$$\beta = (\beta \cdot \mathcal{B})^{\frac{1}{2}} = \left(\mathcal{B}^T \cdot \mathbf{g} \cdot \mathcal{B} \right)^{\frac{1}{2}} = \frac{\omega \Sigma \sin \theta}{\alpha \rho c} \\ \alpha = \frac{\rho}{\Sigma} \Delta^{\frac{1}{2}} \quad (7.73)$$

Note that the three-dimensional portion of the Kerr metric is diagonal in this 3+1 system, and recall that ω is the rotation rate of space about the spinning black hole. The Schwarzschild metric is, in turn, a subset of this, with $\alpha = (1 - 2r_g/r)^{1/2}$, $\beta = 0$, and g_{ij} diagonals of $(1 - 2r_g/r)^{-1}$, r^2 , and $r^2 \sin^2 \theta$. Flat Minkowski spacetime has $\alpha = 1$, $\beta = 0$, and $g_{ij} = \delta_{ij}$.

¹⁴ Remember that the Boyer–Lindquist coordinates here are spherical-polar so, even in the three-metric portion, only $dx^1 = dr$ has units of length. $dx^2 = d\theta$ and $dx^3 = d\phi$ both are in radians, so $\mathcal{B}^3 \alpha c dt$ also is in radians, not length.

7.7.2 Electromagnetism in a Stationary “3+1” Metric

We now can attach some familiar meaning to the components of the Faraday tensor $\mathcal{F}$. We will *define* the *three*-dimensional electric field $\mathbf{E}$, magnetic field $\mathbf{H}$, electric displacement $\mathbf{D}$, and magnetic induction $\mathbf{B}$ vectors (with corresponding *three*-dimensional 1-forms $\mathbf{e}$, $\mathbf{h}$, $\mathbf{d}$, and $\mathbf{b}$) such that

$$\mathcal{F} = \begin{pmatrix} 0 & D^1/(\alpha^2 c) & D^2/(\alpha^2 c) & D^3/(\alpha^2 c) \\ -D^1/(\alpha^2 c) & 0 & h_3 \mathbf{g}^{-1/2} & -h_2 \mathbf{g}^{-1/2} \\ -D^2/(\alpha^2 c) & -h_3 \mathbf{g}^{-1/2} & 0 & h_1 \mathbf{g}^{-1/2} \\ -D^3/(\alpha^2 c) & h_2 \mathbf{g}^{-1/2} & -h_1 \mathbf{g}^{-1/2} & 0 \end{pmatrix}$$

$$\mathbf{f} = \begin{pmatrix} 0 & -e_1 c & -e_2 c & -e_3 c \\ e_1 c & 0 & B^3 \mathbf{g}^{1/2} & -B^2 \mathbf{g}^{1/2} \\ e_2 c & -B^3 \mathbf{g}^{1/2} & 0 & B^1 \mathbf{g}^{1/2} \\ e_3 c & B^2 \mathbf{g}^{1/2} & -B^1 \mathbf{g}^{1/2} & 0 \end{pmatrix}$$

$$\mathcal{M} = \begin{pmatrix} 0 & -B^1/(\alpha c) & -B^2/(\alpha c) & -B^3/(\alpha c) \\ B^1/(\alpha c) & 0 & e_3/(\alpha \mathbf{g}^{1/2}) & -e_2/(\alpha \mathbf{g}^{1/2}) \\ B^2/(\alpha c) & -e_3/(\alpha \mathbf{g}^{1/2}) & 0 & e_1/(\alpha \mathbf{g}^{1/2}) \\ B^3/(\alpha c) & e_2/(\alpha \mathbf{g}^{1/2}) & -e_1/(\alpha \mathbf{g}^{1/2}) & 0 \end{pmatrix}$$

$$\mathbf{m} = \begin{pmatrix} 0 & \alpha h_1 c & \alpha h_2 c & \alpha h_3 c \\ -\alpha h_1 c & 0 & D^3 \mathbf{g}^{1/2}/\alpha & -D^2 \mathbf{g}^{1/2}/\alpha \\ -\alpha h_2 c & -D^3 \mathbf{g}^{1/2}/\alpha & 0 & D^1 \mathbf{g}^{1/2}/\alpha \\ -\alpha h_3 c & D^2 \mathbf{g}^{1/2}/\alpha & -D^1 \mathbf{g}^{1/2}/\alpha & 0 \end{pmatrix}$$

and

$$\mathbf{J} = \frac{1}{\alpha} \begin{pmatrix} \rho_q c \\ J^1 \\ J^2 \\ J^3 \end{pmatrix}$$

where $\mathbf{g}$ is the determinant of the three-metric $\mathbf{g}$. These definitions look rather awkward and unwieldy, but when plugged into equations (7.23) and (7.24), they give us the very familiar-looking, original Maxwell’s equations

$$\boxed{\nabla \cdot \mathbf{B} = 0 \qquad \frac{\partial \mathbf{B}}{c \partial t} + \nabla \times \mathbf{E} = 0} \quad (7.74)$$

$$\boxed{\nabla \cdot \mathbf{D} = 4\pi \rho_q \qquad -\frac{\partial \mathbf{D}}{c \partial t} + \nabla \times \mathbf{H} = \frac{4\pi}{c} \mathbf{J}} \quad (7.75)$$

with the relations

$$\boxed{\mathbf{D} = \frac{1}{\alpha} \mathbf{E} + \boldsymbol{\beta} \times \mathbf{B} \quad \mathbf{B} = \frac{1}{\alpha} \mathbf{H} - \boldsymbol{\beta} \times \mathbf{D}} \quad (7.76)$$

Curved spacetime acts a bit like a nonlinear dielectric/diamagnetic medium, with the permittivity and permeability equal to α^{-1} . Note that the time derivatives and gradients are taken with respect to the global coordinates (*e.g.*, Boyer–Lindquist coordinates for the Kerr metric). And, while the fields $\mathbf{E}$ and $\mathbf{H}$ are measured in the fixed frame, through which fiducial observers are moving with velocity $-\boldsymbol{\beta}$, the displacement $\mathbf{D}$ and induction $\mathbf{B}$ are measured in the frame of those moving observers. Note that, even though they use three-dimensional space plus time, *these equations work in any frame of reference in any stationary metric*. They are covariant in stationary metrics.

7.7.3 Return to Orthonormal Vectors in Stationary “3+1” Metrics: Saying Goodbye to Contravariant Vectors and 1-Forms

Splitting up the metric into space and time parts accomplishes only half the task. We still need to define how to perform vector divergences and curls in a curved three-space. And then, there is the difficult task of working with two kinds of vectors: contravariant and covariant (1-form) ones. Can we return to working with single orthonormal vectors, as we did in Chapter 5?

Shinji Koide showed [321] that, if we use metrics like Kerr in Boyer–Lindquist coordinates, with diagonal sub-three-metrics (equations (7.71) to (7.73)), we can return to using all of the standard vector calculus that we learned as undergraduates in college. The Kerr three-metric and its inverse are diagonal can be written as

$$\mathbf{g} = \begin{pmatrix} h_1^2 & 0 & 0 \\ 0 & h_2^2 & 0 \\ 0 & 0 & h_3^2 \end{pmatrix} \quad \mathbf{g}^{-1} = \begin{pmatrix} h_1^{-2} & 0 & 0 \\ 0 & h_2^{-2} & 0 \\ 0 & 0 & h_3^{-2} \end{pmatrix} \quad (7.77)$$

where

$$h_1 = \frac{\rho}{\Delta^{1/2}} \quad h_2 = \rho \quad h_3 = \frac{\Sigma}{\rho} \sin \theta \quad (7.78)$$

We now can define three-vectors in an orthonormal manner and only have to deal with one vectorial quantity instead of two. That is, we now write

$$\mathbf{V} = (V^{\hat{1}} \quad V^{\hat{2}} \quad V^{\hat{3}}) \quad (7.79)$$

where the orthonormal components are related to the contravariant and covariant components as

$$V^{\hat{i}} = h_i V^i = v_i / h_i \quad (7.80)$$

Finally, we now can use equations (7.78) to write vector derivatives with the familiar formulae of regular curvilinear vector calculus

$$\boxed{(\nabla\psi)^{\hat{i}} = \frac{1}{h_i} \frac{\partial\psi}{\partial x^i}} \quad (7.81)$$

$$\boxed{\nabla \cdot \mathbf{V} = \sum_i \frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial x^i} \left(\frac{h_1 h_2 h_3}{h_i} V^{\hat{i}} \right)} \quad (7.82)$$

$$\boxed{(\nabla \times \mathbf{V})^{\hat{i}} = \sum_{j,k} \frac{h_i}{h_1 h_2 h_3} \epsilon_{ijk} \frac{\partial}{\partial x^j} (h_i V^{\hat{k}})} \quad (7.83)$$

where the antisymmetric symbol

$$\begin{aligned} \epsilon_{ijk} &= +1 && \text{(if } [ijk] \text{ are an even permutation of } [123]) \\ &= -1 && \text{(if } [ijk] \text{ are an odd permutation of } [123]) \\ &= 0 && \text{(otherwise [if } i = j, i = k, \text{ etc.])} \end{aligned}$$

simply gives us a convenient way of writing the curl that we all remember. We also will be able to drop the transpose notation for scalar and vector products, since orthonormal vectors are both contravariant and covariant by nature. However, we still shall write component indices as superscripts, not subscripts, and we shall write them as “hatted” indices to indicate that the components are orthonormal.

Of course, we cannot forget that we now are working in a curved *spacetime*, not simply curvilinear coordinates. The lapse function α will be important, as will be the (now orthonormal) drift vector $\beta^{\hat{i}} = h_i \mathcal{B}^i = \beta_i / h_i$ or

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & \frac{\omega \Sigma \sin \theta}{\alpha \rho c} \end{pmatrix} \quad (7.84)$$

These will appear explicitly in the equations, when they are needed, and perform their functions of slowing down time and rotating spacetime near the black hole.

We will begin using and discussing some of this new 3+1 notation in the next chapter on gravitational waves and black hole formation and then fully convert to it by the end of Chapter 9, where all the physics necessary for analyzing black hole engines is presented.

Chapter 8

Four-Dimensional Evolving Geometry: Gravitational Waves and Gravitational Collapse

The Minkowski, Schwarzschild, and Kerr metrics introduced in Chapters 6 and 7 represent stationary geometries – ones that do not change with time. However, when the distribution of matter that generates the gravitational field changes rapidly with time (such as when two bodies orbit each other or when a large amount of matter collapses to form a black hole), the gravitational field is *not* stationary. In this case we must solve the time-dependent Einstein field equations (7.21) themselves in order to determine how the metric evolves. This is one of the most difficult problems ever attempted in classical physics. However, understanding how relativistic gravitating systems change with time has enormous consequences for understanding many objects in the universe, and so is well worth the effort.

A metric describes the spacetime both close to and far from the object. If things change rapidly near the black hole, then those changes eventually will be felt far from the hole, propagating outward from the system through the spacetime at the speed of light. These “gravitational waves”, predicted by Einstein in 1918 [323], will carry information about the system that generates them, just as electromagnetic waves carry information about the atoms and particles that generate them. By measuring the properties of gravitational waves that pass by the earth, we will be able to measure the properties of black hole orbits, extract the spins and masses of the black holes themselves, learn about the details of gravitational collapse, and begin to unravel the mysteries of what role black hole collisions and mergers play in the formation of black holes themselves.

When the metric is no longer constant with time, or with azimuthal angle ϕ , strange physics begins to take place. Familiar conservation laws like the conservation of energy or angular momentum are no longer valid. Of course, stress-energy-momentum still will be conserved (equation (7.15)), but now we must take into account energy and momentum transfer between the matter and the evolving gravitational field itself. This is done through the covariant divergence $\nabla \cdot \mathcal{T}$, whose geometrical factors are constantly changing with time. And then, of course, these changes in the distribution of matter and momentum will act *back*, through the Einstein field equations, on the gravitational field, modifying it further.

If this sounds like a complicated problem, it really is. In principle, however, it should be only somewhat more complex than studying an evolving electromagnetic field in a charged plasma. All physics and most engineering students have thoroughly studied electromagnetics and Maxwell's field equations by the time they graduate from college. The gravitational field equations are very closely analogous to the electromagnetic ones, so we will use that analogy to explain evolving gravitational fields as much as possible. The best way to understand evolving electromagnetic fields is to study the generation of simple plane electromagnetic waves. The same is true of evolving gravitational fields. So, before we tackle gravitational collapse, we will spend some time deriving and using gravitational waves.

8.1 Plane Gravitational Waves

Just as the time-varying electromagnetic field can have a plane wave solution (see Section 6.6) so also can the time-varying gravitational field. Gravitational waves also can carry energy, angular momentum, and information away from a gravitating body whose mass distribution changes rapidly with time.¹

8.1.1 Basic Physics of Gravitational Waves

8.1.1.1 The Perturbation Metric $\mathcal{H}$

Unlike the electromagnetic field, however, the gravitational field equations are *not* linear; they contain nonlinear products of the Christoffel symbol (Γ) terms, which can make their solution very complicated. In the case of gravitational waves propagating through space, far from the source (*e.g.*, as they pass near the earth), we can get around this problem by assuming that the local metric deviates only very slightly from Minkowski's

$$g = \eta + \mathfrak{h} \quad (8.1)$$

where $\mathfrak{h}$ is a symmetric and very small "perturbation" on the Minkowski metric; each $\mathfrak{h}_{\alpha\beta} \ll 1$. In order to preserve the analogy to the electromagnetic gauge field equations, which use the (contravariant) vector $\mathbf{A}$ (equation (6.127)), we will normally deal with the contravariant tensor

¹ While some people use the term "gravity waves" for this phenomenon, we will refrain from doing so here. In terrestrial physics this latter term usually is reserved for a terrestrial phenomenon that is generated in the earth's atmosphere by buoyancy at the interface between two media such as the atmosphere and the ocean. *Gravitational* waves are giant oscillations (some the size of the earth's orbit about the sun), that travel around the universe and are generated by moving stars and black holes.

$$\mathcal{H} = g^{-1} \cdot \mathfrak{h} \cdot g^{-1} \quad (8.2)$$

$$= \begin{pmatrix} \mathfrak{h}_{00} & -\mathfrak{h}_{01} & -\mathfrak{h}_{02} & -\mathfrak{h}_{03} \\ -\mathfrak{h}_{01} & \mathfrak{h}_{11} & \mathfrak{h}_{12} & \mathfrak{h}_{13} \\ -\mathfrak{h}_{02} & \mathfrak{h}_{12} & \mathfrak{h}_{22} & \mathfrak{h}_{23} \\ -\mathfrak{h}_{03} & \mathfrak{h}_{13} & \mathfrak{h}_{23} & \mathfrak{h}_{33} \end{pmatrix} \quad (8.3)$$

How small do we expect the $\mathfrak{h}$'s to be? Gravitational wave detectors like the Laser Interferometer Space Antenna (LISA) and the Laser Interferometer Gravitational wave Observatory (LIGO) will be sensitive at the level of $h \sim 0.2\text{--}3 \times 10^{-21}$ – extraordinarily small compared to the values of 1 on the diagonal of $\boldsymbol{\eta}$ (see Fig. 8.3). Therefore, we can ignore all terms in Einstein's equations that are proportional to *two* factors of $\mathfrak{h}$ (*i.e.*, $\mathfrak{h}_{\alpha\beta} \mathfrak{h}_{\gamma\delta}$), because they will be at least as small as 10^{-41} .

8.1.1.2 The Gravitational Wave Equation and the Transverse-Traceless Gauge

Because the connection coefficients will be functions of the derivatives of the $\mathfrak{h}$'s, we can, therefore, neglect all products of the $\boldsymbol{\Gamma}$'s in Einstein's field equations (7.21) as well. The gravitational field equations, then, become

$$\begin{aligned} & \frac{1}{2} \left\{ (\boldsymbol{\eta} \cdot \boldsymbol{\nabla}) (\boldsymbol{\nabla} \cdot \mathcal{H}) + [(\boldsymbol{\eta} \cdot \boldsymbol{\nabla}) (\boldsymbol{\nabla} \cdot \mathcal{H})]^T \right. \\ & \left. - \boldsymbol{\nabla}^2 \mathcal{H} - (\boldsymbol{\eta} \cdot \boldsymbol{\nabla}) (\boldsymbol{\eta} \cdot \boldsymbol{\nabla})^T (\text{tr } \mathcal{H}) - \boldsymbol{\eta} [\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \cdot \mathcal{H}) - \boldsymbol{\nabla}^2 (\text{tr } \mathcal{H})] \right\} \\ & = 8\pi \frac{G}{c^4} \mathcal{T} \end{aligned} \quad (8.4)$$

where

$$\text{tr } \mathcal{H} = g : \mathcal{H}$$

is the trace (sum of the diagonal elements) of the matrix $\mathcal{H}$. Even with all the simplifications, this still looks complex. However, some of the strange-looking terms are easily understood, and will disappear anyway. The operator $\boldsymbol{\eta} \cdot \boldsymbol{\nabla}$ is a 1-form gradient instead of a vector gradient. And the dyadic product of two of them $(\boldsymbol{\eta} \cdot \boldsymbol{\nabla})(\boldsymbol{\eta} \cdot \boldsymbol{\nabla})^T$ produces a contravariant tensor from a simple scalar ($\text{tr } \mathcal{H}$).

Compare equation (8.4) with equation (6.125). Just like the latter equation, which had terms in $\boldsymbol{\nabla} \cdot \mathbf{A}$, there are several terms in $\boldsymbol{\nabla} \cdot \mathcal{H}$, and others in $(\text{tr } \mathcal{H})$, that we would like to get rid of. We can, in fact, play the same game of assuming a Lorenz gauge condition for each row/column in symmetric matrix $\mathcal{H}$

$$\boldsymbol{\nabla} \cdot \mathcal{H} = 0 \quad (8.5)$$

and choose $\mathcal{H}$ to be traceless as well

$$\text{tr } \mathcal{H} = 0 \quad (8.6)$$

This is called the “transverse-traceless” (TT gauge) or traceless-Lorenz gauge for gravitational waves. With these gauge assumptions, the field equation (8.4) becomes a simple wave equation with stress-energy sources

$$\boxed{\nabla^2 \mathcal{H} = -16\pi \frac{G}{c^4} \mathcal{T}} \quad (8.7)$$

and with $\mathcal{H}$ playing the role of a *tensor* potential for gravitational field waves, just as $\mathbf{A}$ was a vector potential for electromagnetic field waves in equation (6.131).

Are we allowed to make so many free gauge choices for the gravitational field? Yes. In fact, just as the Lorenz condition for electromagnetic waves left us with one additional free function to choose (Z in equations (6.86) and (6.87)), so equation (8.5) still leaves us with *four* additional choices (not just one). We know those four choices already as the *coordinate system*, and in this case one of those coordinate equations can be (8.6). Together, equations (8.5) and (8.6), plus three more coordinate conditions, make a total of eight conditions on the ten independent components of the symmetric tensor $\mathcal{H}$. *This leaves only two unspecified, these being the two possible polarizations of a gravitational wave.*

8.1.1.3 The Stress-Energy of Gravitational Waves

Just as the electromagnetic field contributes to the stress-energy of spacetime, even when there is no matter present, so also does the gravitational field. In the transverse-traceless gauge, the weak field stress-energy tensor has a particularly simple form

$$\mathcal{T}_{\text{GW}} = \frac{c^4}{32\pi G} (\boldsymbol{\eta} \cdot \boldsymbol{\nabla}) \mathcal{H} : (\boldsymbol{\eta} \cdot \boldsymbol{\nabla})^T \mathcal{H} \quad (8.8)$$

where the double-dot product is on the components of the $\mathcal{H}$ tensors (not on the components of $\boldsymbol{\nabla}$). That is, the elements of the covariant stress-energy tensor are

$$t_{\mu\nu \text{ GW}} = \frac{c^4}{32\pi G} \left(\frac{\partial \mathcal{H}}{\partial x^\mu} : \frac{\partial \mathcal{H}}{\partial x^\nu} \right) \quad (8.9)$$

with the contravariant tensor then given by $\mathcal{T}_{\text{GW}}^{\alpha\beta} = \sum_{\mu\nu} \eta^{\alpha\mu} \eta^{\beta\nu} t_{\mu\nu \text{ GW}}$

8.1.2 Plane Gravitational Waves in a Vacuum

In a vacuum equation (8.7) becomes

$$\nabla^2 \mathcal{H} = 0 \quad (8.10)$$

Since each component of $\mathcal{H}$ now satisfies its own wave equation in the TT gauge, we can choose each one to be represented by the real part of a complex sinusoidal function

$$\mathcal{H} = \Re \{ \mathbf{h} \exp [i(\mathbf{k} \cdot \mathbf{x})] \} \quad (8.11)$$

and we can examine the case when this planar wave propagates in the z direction

$$\mathbf{k} = \begin{pmatrix} k^w \\ 0 \\ 0 \\ k^z \end{pmatrix} \quad (8.12)$$

where $k^w \equiv \omega/c$, making $\mathbf{k}$ a null vector once again, as it was in Section 6.6.2. So the gravitational waves propagate at the speed of light.

8.1.2.1 Application of the Lorenz (Transverse) Condition

The Lorenz condition (equation (8.5)) tells us that the potential amplitudes and the propagation vector are orthogonal

$$\mathbf{k} \cdot \mathbf{h} = 0 \quad (8.13)$$

which leads to the following *amplitude* tensor in the right-hand side of equation (8.11)

$$\mathbf{h} = \begin{pmatrix} h_{ww} & h_{wx} & h_{wy} & h_{ww} k^w/k^z \\ h_{wx} & h_{xx} & h_{xy} & h_{wx} k^w/k^z \\ h_{wy} & h_{xy} & h_{yy} & h_{wy} k^w/k^z \\ h_{ww} k^w/k^z & h_{wx} k^w/k^z & h_{wy} k^w/k^z & h_{ww} k^w/k^z \end{pmatrix}$$

Note that the elements of $\mathbf{h}$ are *not* functions of space and time. All the spacetime dependence is in the exponential in equation (8.11).

8.1.2.2 Choice of a Coordinate System

Just as in the electromagnetic case, when we were free to choose $\mathbb{A}^w = 0$, we also are free to choose each $h_{w\beta}$ (elements in the first column of $\mathbf{h}$) to be zero. Formally, this means that we choose the components of $\mathbf{h}$ to be orthogonal to our four-velocity $\mathbf{U}$, whatever that might be

$$\mathbf{U} \cdot \mathbf{h} = 0 \quad (8.14)$$

(Note that these are conditions on $\mathbf{h}$, not on $\mathbf{U}$.) There actually are only three conditions in equation (8.14), $h_{ww} = h_{wx} = h_{wy} = 0$, because equation (8.13) already

sets $h_{wz} = h_{ww} k^w / k^z$, which is now zero. The final, eighth condition is equation (8.6). These conditions result in a gravitational wave amplitude tensor of

$$\mathbf{h} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_{xx} & h_{xy} & 0 \\ 0 & h_{xy} & -h_{xx} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (8.15)$$

As expected, there are only two functions, h_{xx} and h_{xy} , which represent the two gravitational wave polarizations.

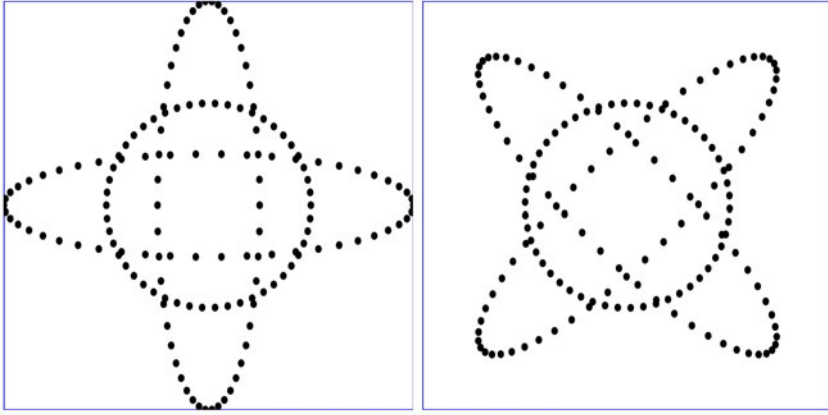


Fig. 8.1: Demonstration of the two gravitational wave polarizations, + and $\times$. *Left panel:* The + polarization ($h_{xx} \neq 0$, $h_{xy} = 0$) takes an initially circular ring of test masses and deforms it into alternating vertical and horizontal ellipses. The period of a single such cycle is the inverse of the gravitational wave frequency. *Right panel:* Same, but for the $\times$ polarization mode ($h_{xx} = 0$, $h_{xy} \neq 0$), which is rotated by 45° from the + mode. The orientation of the axes is, of course, arbitrary and defined by that of the detector itself; e.g., rotating the detector by 45° , 135° , 225° , or 315° will switch the identity of the + and $\times$ modes. Figure reprinted from Fig. 1 of [324], with permission of B. Sathyaprakash and the MPG.

8.1.2.3 Gravitational Wave Polarizations

Let us first examine the case when $h_{xx} \neq 0$ and $h_{xy} = 0$. The full metric (equation (8.1)) then is

$$\mathbf{g} = \Re \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 + h_{xx} \exp[i(\mathbf{k} \cdot \mathbf{x})] & 0 & 0 \\ 0 & 0 & 1 - h_{xx} \exp[i(\mathbf{k} \cdot \mathbf{x})] & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8.16)$$

So, a gravitational wave traveling along z compresses space ever so slightly in one direction in the x - y plane (here x) and stretches it in the orthogonal direction. And this squeezing and stretching alternates sinusoidally in space and time. This action is illustrated in the left panel of Fig. 8.1.

What about the case when $h_{xx} = 0$ and $h_{xy} \neq 0$? Of course, the full metric now becomes

$$\mathbf{g} = \Re \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & h_{xy} \exp[i(\mathbf{k} \cdot \mathbf{x})] & 0 \\ 0 & h_{xy} \exp[i(\mathbf{k} \cdot \mathbf{x})] & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8.17)$$

But, what does this mean in terms of the compressing and stretching of space by gravitational waves? We can easily see the answer if we rotate $\mathbf{g}$ about the z axis. Note that we need to deal *only* with the rotated amplitude tensor

$$\mathbf{h}' = \mathcal{L}^T \cdot \mathbf{h} \cdot \mathcal{L} \quad (8.18)$$

where

$$\mathcal{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta_p & -\sin \theta_p & 0 \\ 0 & \sin \theta_p & \cos \theta_p & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8.19)$$

because the rotation of the $\boldsymbol{\eta}$ part still gives $\boldsymbol{\eta}$

$$\begin{aligned} \mathbf{g}' &= \mathcal{L}^T \cdot \mathbf{g} \cdot \mathcal{L} \\ &= \mathcal{L}^T \cdot (\boldsymbol{\eta} + \mathbf{h}) \cdot \mathcal{L} \\ &= \boldsymbol{\eta} + \mathcal{L}^T \cdot \mathbf{h} \cdot \mathcal{L} \end{aligned}$$

So, setting $h_{xx} = 0$ in equation (8.15) and applying the transformation in equation (8.19), we arrive at an expression for the amplitude tensor in the new rotated frame

$$\mathbf{h}' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_{xy} \sin(2\theta_p) & h_{xy} \cos(2\theta_p) & 0 \\ 0 & h_{xy} \cos(2\theta_p) & -h_{xy} \sin(2\theta_p) & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (8.20)$$

Note that this is a bit different from electromagnetic waves, which had two polarization states that were perpendicular to each other. Instead, if we now we try to rotate by a full $\theta_p = 90^\circ$, we will flip the wave completely around and still have zeros on the diagonal. However, if we rotate *only* by $\theta_p = 45^\circ$, we obtain

$$\mathbf{h}' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_{xy} & 0 & 0 \\ 0 & 0 & -h_{xy} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

That is, this wave looks identical to the case when $h_{xy} = 0$, except that h_{xy} now has replaced h_{xx} . This second polarization simply represents a gravitational wave rotated by 45° (not 90°) about the z axis and is illustrated in the right panel of Fig. 8.1. The two polarizations ($\theta_p = 0^\circ$ and $\theta_p = 45^\circ$) are usually referred to as + (“plus”) and $\times$ (“cross”), because of the shape of their waves at phases $\mathbf{k} \cdot \mathbf{x} = 0^\circ$ and $\mathbf{k} \cdot \mathbf{x} = 90^\circ$ in the x - y plane when viewed from earth back toward the source.

8.1.2.4 Poynting Flux and Power

For a wave of general amplitude

$$h = (h_{xx}^2 + h_{xy}^2)^{1/2} \quad (8.21)$$

from equation (8.8) we will have an energy density (averaged over many wave oscillations) of

$$\epsilon^{GW} = \langle \mathcal{T}_{ww}^{GW} \rangle_t = \frac{c^2 \omega^2}{32\pi G} h^2 \quad (8.22)$$

and an energy (Poynting) flux of

$$\mathcal{S}^{GW} = c \langle \mathcal{T}_{wz}^{GW} \rangle_t = \frac{c^3 \omega^2}{32\pi G} h^2 \quad (8.23)$$

in the $+z$ direction. Note that in the above expressions, we get a factor of 2 from the double-dot product in equation (8.8), but we also get a factor of 1/2 from the time average of the square of the sinusoids.

8.1.3 Plane Gravitational Waves Generated by a Time-Varying Stress-Energy Distribution

We are now able to calculate the strength of a gravitational wave that is generated by a source, and we will follow essentially the same procedure as we used in Section 6.6.1.8 to calculate the strength of electromagnetic dipole radiation.

8.1.3.1 General Green's Function Quadrupole Solution

As in the electromagnetic case, the solution to the *inhomogeneous* wave equation (8.7) is the retarded potential

$$\mathcal{H}(x, t) = \frac{4G}{c^4} \int d^3x' \frac{\mathcal{T}(\mathbf{x}', t - r'/c)}{r'} \quad (8.24)$$

which is of very small amplitude (see equation (8.2)). Again we will ignore all w (time) components of this equation, because we will not need them, and we will expand $1/r'$ in powers of $\mathbf{x} \cdot \mathbf{x}'/r^3$. And we again will use a trick of mathematics and physics that converts the 4×4 tensor equation (8.24) into one for only the spatial components of $\mathcal{H}$, which we will write as the 3×3 matrix $\mathbf{H}$

$$\mathbf{H}(\mathbf{x}, t) = \frac{2G}{c^4 r} \frac{d^2(\mathbf{Q})}{dt^2} \quad (8.25)$$

where $\mathbf{Q}$ is a 3×3 tensor that is the time-dependent, retarded, *reduced quadrupole moment* of the *mass* distribution

$$\begin{aligned} \mathbf{Q} &= \frac{1}{c^2} \int \left(\mathbf{x} \mathbf{x}' - \frac{1}{3} \mathbf{I} r'^2 \right) \mathcal{T}^{ww}(\mathbf{x}', t - r'/c) d^3\mathbf{x} \\ &= \int \left(\mathbf{x} \mathbf{x}' - \frac{1}{3} \mathbf{I} r'^2 \right) \rho_m(\mathbf{x}', t - r'/c) d^3\mathbf{x} \end{aligned} \quad (8.26)$$

The derivation of equation (8.25) from (8.24) is similar to the derivation of equation (6.98) in Section 6.6.1. It again relies on Gauss's theorem, the chain rule for the gradient operator, and on conservation laws (here energy and momentum) generated by the field equations and their Bianchi identities.

8.1.3.2 Poynting Flux and Power Generated by a Simple Oscillating Quadrupole

For a simple sinusoidally-oscillating quadrupole

$$\mathbf{Q} = \Re \{ \mathbf{q} \exp[-i\omega(t - r/c)] \}$$

we will obtain an $\mathbf{H}$ of the form

$$\mathbf{H} = \Re \{ \mathbf{h} \exp[-i\omega(t - r/c)] \}$$

which gives, with equation (8.25), a gravitational wave amplitude tensor of

$$\mathbf{h} = \frac{2G\omega^2}{c^4 r} \mathbf{q}$$

However, as with the electromagnetic wave, if we choose the transverse-traceless gauge (equations (8.5) and (8.6)) in the observer's frame, we will need neither the w nor z components of equation (8.24). We will need only the projections on the sky-viewing plane of the perturbation amplitude $\mathbf{h}'$, and therefore of the quadrupole tensor amplitude $\mathbf{q}'$, in the *observer's* x and y directions. These generally will be smaller than the total amplitudes in the source frame, but will be related in the same manner

$$\mathbf{h}' = \frac{2G\omega^2}{c^4 r} \mathbf{q}' \quad (8.27)$$

where

$$\mathbf{q}' = \begin{pmatrix} q'_{\text{tr}} & q'_{xy} & 0 \\ q'_{xy} & -q'_{\text{tr}} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and $q'_{\text{tr}} \equiv (q'_{xx} + q'_{yy})/2$.

The computation of the components of the projected quadrupole tensor $\mathbf{q}'$ from the original tensor $\mathbf{q}$ is more difficult than that of the projected electric dipole in Section 6.6.1.8 and is described elsewhere [305, 325]. The result is a relation between the determinants of the two tensors

$$q' = q f(\Omega)$$

where Ω is the solid angle around the source, and q the total mass quadrupole amplitude in the source frame

$$q \equiv (q_{\text{tr}}^2 + q_{xy}^2)^{1/2}$$

From equations (8.23) and (8.27), the Poynting flux is then

$$\begin{aligned} \mathcal{S}^{GW} &= \frac{c^3 \omega^2}{32\pi G} h'^2 \\ &= \frac{G \omega^6}{8\pi c^5 r^2} q'^2 \\ &= \frac{G \omega^6}{8\pi c^5 r^2} q^2 f^2(\Omega) \end{aligned} \quad (8.28)$$

The integral of $f^2(\Omega)$ over solid angle is $4\pi/5$ [305, 325]; (*i.e.*, the quadrupole radiates into only one-fifth of the sky, compared to the dipole's two-thirds). So the total power radiated by an oscillating quadrupole gravitational wave source, in terms of its total quadrupole moment, is

$$\boxed{\mathcal{P}^{GW} = \frac{G \omega^6}{10 c^5} q^2} \quad (8.29)$$

and the *average* Poynting flux over the sky, at a given distance r from the source, is

$$\langle \mathcal{S}^{GW} \rangle_{\Omega} = \frac{G \omega^6}{40\pi c^5 r^2} q^2 \quad (8.30)$$

8.1.3.3 General Quadrupole Formulae

As in the electromagnetic dipole case, equations (8.28) and (8.29) represent only one Fourier component ω of a possible spectrum of oscillating quadrupoles in a source. So we can write general expressions for the Poynting flux (not averaged over 4π steradians) and for the total power as

$$\mathcal{S}^{GW} = \frac{G \omega^6}{4\pi c^5 r^2} \left\langle \left| \frac{d^3 Q}{dt^3} \right|^2 \right\rangle_t \quad (8.31)$$

$$\mathcal{P}^{GW} = \frac{G \omega^6}{5 c^5} \left\langle \left| \frac{d^3 Q}{dt^3} \right|^2 \right\rangle_{t \Omega} \quad (8.32)$$

where the factor of $1/2$ from sinusoidal averaging has been re-absorbed into the time average in these expression.

8.1.4 Example of a Gravitational Wave Source: A Close Binary System of Point Masses

Close binary systems, composed of two stars orbiting each other within a million Schwarzschild radii, are some of the simplest, and yet most important, sources of oscillating gravitational fields and, therefore, gravitational waves. Using the formulae derived in the previous section, we will be able to determine useful properties of such systems, including the power that they radiate, their lifetimes from now until the two stars merge, and the frequency and amplitude that their gravitational waves will produce in a detector on the earth. For the simple analysis below, we will assume that the two stars are *point masses* only and in simple (Newtonian) Keplerian orbits. We will ignore tidal and possible relativistic effects that complicate the evolution, especially when the stars get within a few stellar radii of each other and begin to merge. In the next section, however, we will examine some fully relativistic simulations that *do* take these complicating effects into account.

8.1.4.1 Binary Stars as Emitters of Powerful Gravitational Waves

As in Chapter 5, we will let m_1 and m_2 be the more and less massive of the two stars, respectively, let their distances from their joint center of mass be a_1 and a_2 ,

and let the total binary mass M and separation a be given by equations (5.8) and (5.29). Then the system has a static quadrupole moment of simply

$$q = m_1 a_1^2 + m_2 a_2^2 \quad (8.33)$$

Because of Keplerian orbital motion, however, the quadrupole moment varies at *twice* the Keplerian frequency

$$\omega = 2 \Omega_K = 2 \left(\frac{GM}{a^3} \right)^{1/2}$$

or, in cgs units,

$$f = \frac{\omega}{2\pi} = 2.2 \times 10^4 \text{ Hz} \left(\frac{M}{M_\odot} \right)^{-1} \left(\frac{a}{r_{SM}} \right)^{-3/2} \quad (8.34)$$

where

$$r_{SM} \equiv \frac{2GM}{c^2}$$

is the Schwarzschild radius computed using the total binary mass. Any binary system, therefore, will generate gravitational waves as it has a time-varying quadrupole moment.

We now can derive the gravitational wave power generated by the changing binary quadrupole moment, using equation (8.29), plus equation (5.32) for the reduced binary mass μ and equations (8.33) and (8.34) for q and ω . The result is

$$\begin{aligned} \mathcal{P}^{GW} &= \frac{32}{5} \frac{G^4}{c^5} \frac{\mu^2 M^3}{a^5} \\ &= \frac{c^5}{5G} \left(\frac{a}{r_{SM}} \right)^{-5} \left(\frac{\mu}{M} \right)^2 \end{aligned}$$

or, in cgs units,

$$\mathcal{P}^{GW} = 7.26 \times 10^{58} \text{ erg s}^{-1} \left(\frac{a}{r_{SM}} \right)^{-5} \left(\frac{\mu}{M} \right)^2 \quad (8.35)$$

This is potentially an enormous amount of power – up to $\frac{1}{5} \mu c^2$ radiated in only a light crossing time²

² The amount of power given by the simple expression $c^5/G = 3.63 \times 10^{59} \text{ erg s}^{-1}$ is the *maximum* luminosity that a gravitating object can produce, and it is independent of all other properties of that object (mass, size, *etc.*). In principle, a coalescing binary star can produce one-fifth of that power (nearly 10^{52} watts) in gravitational waves, or $>10^{10}$ times the *electromagnetic* bolometric luminosity of the brightest quasars! Of course, this power lasts for only a light-crossing time τ_{gM} and only occurs when the binary is composed of black holes and merges ($a \rightarrow r_{SM}$). In addi-

$$\tau_{gM} \equiv G M / c^3$$

Of course, if the binary separation is large compared to the Schwarzschild radius ($a \gg r_{SM}$) or the secondary star is much less massive than the primary ($m_2 \ll m_1$ or $\mu \ll M$), then the power emitted will be greatly reduced.

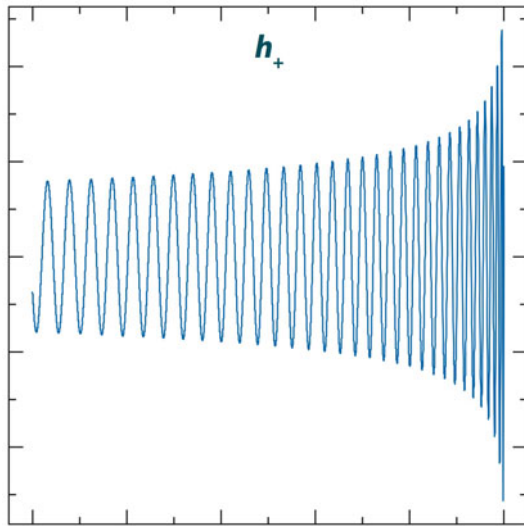


Fig. 8.2: The “chirp” gravitational wave signal from an inspiraling binary, found by solving eq. (8.38), with eq. (8.36) determining $a(t)$. Note: this is valid only for circular orbits of a test mass ($m_2/m_1 \rightarrow 0$) around a non-rotating black hole ($j_1 = 0$). More complex waveforms occur when these assumptions are relaxed, allowing the observer to deduce quite a bit of information about the binary source (see [326] for additional waveforms). Reprinted from Fig. 6b of reference [326], ©2009, by permission from Annual Reviews.

8.1.4.2 The Binary Gravitational Wave Coalescence Time

The loss that this power represents to the binary’s orbital energy

$$\mathcal{E}_{orb} = -\frac{G M \mu}{2 a}$$

does build up over time, however. We can calculate the time it takes for the separation to approach zero (or at least r_{SM}) by integrating the energy loss equation

tion, the power is reduced by a potentially small factor $(\mu/M)^2$. Nevertheless, it is this possibility of enormous luminosity in some coalescing black hole systems, plus the potential of probing the properties of the black holes themselves, that makes the effort to detect gravitational waves so compelling.

$$d\mathcal{E}_{orb}/dt = -\mathcal{P}^{GW} \text{ or}$$

$$\frac{da}{dt} = -\frac{64}{5} \frac{G^3}{c^5} \frac{\mu M^2}{a^3}$$

This has the solution

$$a(t) = r_{SM} \left[\frac{a^4}{r_{SM}^4} - \frac{16\mu}{5M} \frac{t}{\tau_{gM}} \right]^{1/4} \quad (8.36)$$

which can be solved for the time when the binary coalesces into a black hole ($a(t) \approx r_{SM}$)

$$\tau^{GW} = \tau_{gM} \frac{5M}{16\mu} \left[\frac{a^4}{r_{SM}^4} - 1 \right]$$

or, in familiar units,

$$\tau^{GW} = 4.7 \times 10^{10} \text{ yr} \left(\frac{P_b}{1 \text{ d}} \right)^{8/3} \left(\frac{M}{\mu} \right) \left(\frac{M}{M_\odot} \right)^{-5/3} \quad (8.37)$$

where we will continue to use the symbol $a \equiv a(0)$ (without time-dependence) as the binary separation at $t = 0$, and P_b is the binary period ($= 2/f$). Equation (8.37) is called the gravitational wave coalescence time and gets increasingly shorter as we make the separation a smaller. The frequency of the gravitational wave emitted (equation (8.34)) also increases as the binary separation decreases. The pitch of the signal is low at first, but increases as the separation decreases, and then accelerates to a very high pitch right before coalescence (see Fig. 8.2). This type of wave is called a “chirp”, because it would sound like the chirp of a bird, if we could hear the gravitational wave.

8.1.4.3 Observed Signal Strength

The gravitational wave form generated by a binary star can be computed from equation (8.25) in the observer’s (primed) coordinate system and from the projections $\mathbf{q}'$ of the binary’s quadrupole moment in equation (8.27). The wave strength varies with the inclination angle i of the binary, as viewed by the observer, as well as with time, and is given by [327]

$$h_+ \equiv H'_{xx} = \frac{2(G\mathcal{M})^{5/3} \omega^{2/3}}{c^4 r} (1 + \cos^2 i) \cos(\omega t) \quad (8.38)$$

$$h_\times \equiv H'_{xy} = \pm \frac{4(G\mathcal{M})^{5/3} \omega^{2/3}}{c^4 r} \cos i \sin(\omega t) \quad (8.39)$$

The quantity

$$\mathcal{M} \equiv \mu^{3/5} M^{2/5} \quad (8.40)$$

is called the “chirp mass” and is the only quantity with mass dimensions that is directly measurable using the gravitational wave amplitude and the source distance alone. (Other properties need the wave frequency and form.) The solution in equations (8.38) and (8.40) gives a characteristic wave amplitude of³

$$h = \left(\frac{32}{5}\right)^{1/2} 2^{2/3} \frac{GM}{c^2 r} \frac{G\mu}{c^2 a}$$

or

$$h \approx \left(\frac{a}{r_{SM}}\right)^{-1} \left(\frac{r}{r_{S\mu}}\right)^{-1} \quad (8.41)$$

where $r_{S\mu}$ is the Schwarzschild radius of the reduced mass

$$r_{S\mu} \equiv \frac{2G\mu}{c^2}$$

Equation (8.41) is a very nice, simple result. The amplitude h of a gravitational wave from a binary is determined simply by its separation and its distance, each in units of the total and reduced mass Schwarzschild radii, respectively.

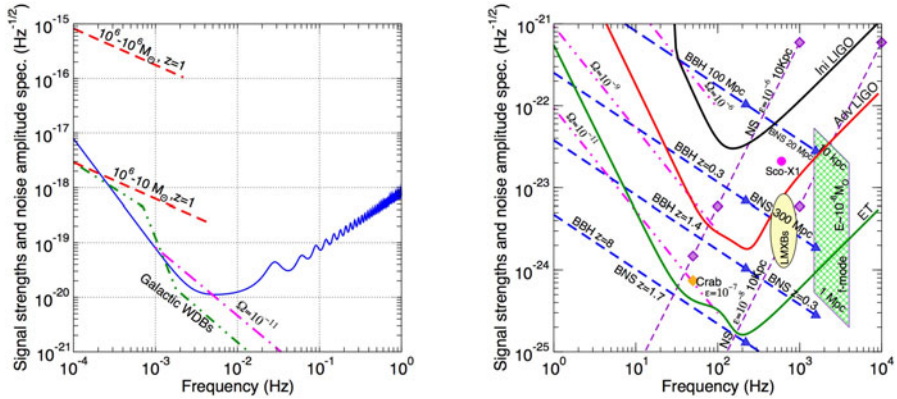


Fig. 8.3: Capabilities of several current and planned gravitational wave detectors in the signal strength vs. frequency plane. *Left panel:* The LISA mission (sensitivity curve in blue) will detect low-frequency emission; red dashed lines show the evolution in signal strength and frequency of a $10^6 M_{\odot}$ equal binary at $z = 1$ (EBBH; Table 8.1a) and a $10^6 M_{\odot}$ to $10 M_{\odot}$ EMRIBH at $z = 0.01$ (Table 8.1b); the green dashed-double-dot line shows the cumulative strength of the white dwarf binaries in the Galaxy. *Right panel:* Several current and future ground-based detectors (LIGO, advanced LIGO, and ET [the Einstein Telescope] in Europe) will detect high-frequency emission. Figure reprinted from Fig. 5 of [324], with permission of B. Sathyaprakash and the MPG.

³ This amplitude is very close to what we would naïvely obtain from equation (8.27) if we applied a sinusoidal averaging factor of 1/2 and a solid angle averaging factor of 1/5.

Table 8.1: Gravitational wave properties of different types of binary stars**a. Plunging ($a \approx 3r_{\text{SM}}$) EBBHs at cosmological distances ($r \sim 3 \text{ Gpc}$)**

$m_1/M_\odot$	$m_2/M_\odot$	f	P_b	τ^{GW}	h	$h f^{-1/2} (\text{Hz}^{-1/2})$
10^9	10^9	$2 \mu\text{Hz}$	10^{d}	10^{d}	5×10^{-15}	3×10^{-12}
10^8	10^8	$23 \mu\text{Hz}$	1^{d}	1^{d}	5×10^{-16}	1×10^{-13}
10^7	10^7	$230 \mu\text{Hz}$	$2^{\text{h}}.4$	$2^{\text{h}}.4$	5×10^{-17}	3×10^{-15}
10^6	10^6	2 mHz	14^{m}	14^{m}	5×10^{-18}	1×10^{-16}
10^5	10^5	23 mHz	86^{s}	86^{s}	5×10^{-19}	3×10^{-18}
10^4	10^4	230 mHz	$8^{\text{s}}.6$	$8^{\text{s}}.6$	5×10^{-20}	1×10^{-19}
10^3	10^3	2 Hz	860 ms	860 ms	5×10^{-21}	3×10^{-21}
100	100	23 Hz	86 ms	86 ms	5×10^{-22}	1×10^{-22}
10	10	230 Hz	8.6 ms	8.6 ms	5×10^{-23}	3×10^{-24}
10^6	10	4 mHz	$7^{\text{m}}.6$	5 mo	10^{-22}	2×10^{-21}

b. Plunging EMRIBHs at local extragalactic distances ($r \sim 30 \text{ Mpc}$)

$m_1/M_\odot$	$m_2/M_\odot$	f	P_b	τ^{GW}	h	$h f^{-1/2} (\text{Hz}^{-1/2})$
10^6	10^3	4 mHz	$7^{\text{m}}.6$	$1^{\text{d}}.4$	10^{-18}	2×10^{-17}
10^6	100	4 mHz	$7^{\text{m}}.6$	14^{d}	10^{-19}	2×10^{-18}
10^6	10	4 mHz	$7^{\text{m}}.6$	5 mo	10^{-20}	2×10^{-19}
10^6	1	4 mHz	$7^{\text{m}}.6$	4 yr	10^{-21}	2×10^{-20}

c. Typical compact binaries in our Galaxy ($r \sim 1 \text{ kpc}$)

$m_1/M_\odot$	$m_2/M_\odot$	a	f	P_b	τ^{GW}	h	$h f^{-1/2} (\text{Hz}^{-1/2})$
1	1	$1.5 \times 10^6 \text{ km}$	$100 \mu\text{Hz}$	6^{h}	1500 Myr	2×10^{-22}	2×10^{-20}
0.5	0.1	$1.5 \times 10^5 \text{ km}$	1 mHz	22^{m}	11 Myr	10^{-22}	3×10^{-21}

8.1.4.4 Binary Gravitational Wave Sources in the Galaxy and the Universe

Table 8.1 gives values of important input parameters (binary separation a , frequency $f = \omega/2\pi$, and period P_b) and the output properties of the gravitational wave (coalescence time τ^{GW} and amplitude h) for different types of possible binary star systems. These include equal-mass binary black holes (EBBHs), extreme mass-ratio inspiral black holes (EMRIBHs), and known types of neutron star and white dwarf binaries. **Fig. 8.3** shows where these binary sources lie in the $h f^{-1/2}$ vs. f diagram, along with the characteristics of the LIGO and LISA detectors. Sensitivities of LIGO ($h f^{-1/2} \sim 4 \times 10^{-23}$) and of LISA ($\sim 10^{-20}$) are good enough to easily detect many of these sources. The key to *seeing* an event, however, is not just its strength but also *how often* that event is expected to occur within the distance range of the device. For example, the current LIGO should be able to detect the merger of two $10 M_\odot$ black holes at 100 Mpc ($\sim 10^{-22}$). However, such mergers are expected to occur *only a few times in 10^6 yr* in a galaxy like the Milky Way, and in a volume of 100 Mpc in radius there are only $\leq 10^5$ such galaxies. So, we expect to see one of these fairly weak events only once in a decade or so. This is why it is so important to improve the sensitivity of instruments like these (*e.g.*, Advanced LIGO) – so they can begin to detect and observe gravitational wave sources of all types.

8.2 Nonlinear Gravitational Wave Sources and Numerical Relativity: The Merger of Black Hole Binary Systems

When two inspiraling stars become so close that the sizes of the stars are comparable to their separation, the simple point mass approximation used in the previous section is no longer correct. Tidal effects will distort the stars, perhaps allowing them to exchange matter or even merge. Additionally, if one or both of these stars is a neutron star or black hole, the assumptions that we made about the metric (Minkowski plus small perturbations $\mathcal{H}$ that we can compute using Kepler's and Newton's law of gravity equations (8.33) to (8.41)) are not valid either. In this case we need to include many more relativistic effects than our simple treatment above can handle.

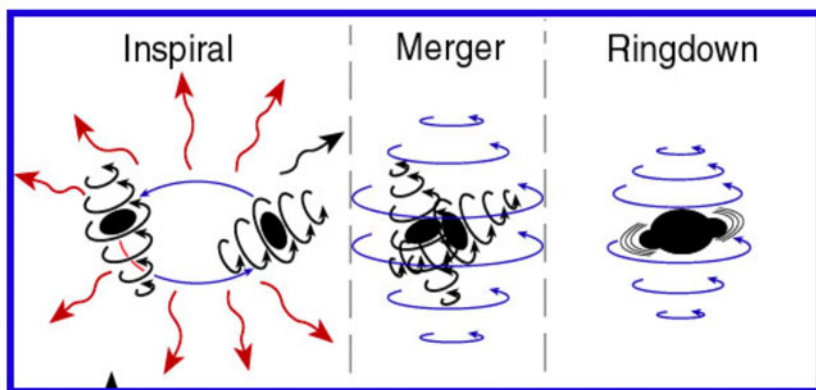


Fig. 8.4: Schematic diagram of the three phases of black hole mergers that are discussed in the text. Reproduced with kind permission of K. Thorne.

8.2.1 The Different Phases of a Binary Merger and Methods Used to Study Them

There are several approaches to solving Einstein's time-dependent equations:

- **Linear Perturbation Methods:** Begin with the standard nonlinear solutions (Schwarzschild, Kerr, *etc.*) perturb *those* metrics as we did with the Minkowski one in Section 8.1.1. This often can be performed with “analytic” methods (ones that use only calculus and algebra) as we did above.
- **“Post-Newtonian” Methods:** Perturb Einstein's equations around the low-gravity solution (Newton's theory) and add higher-order, nonlinear terms in the pertur-

bation metric ($\mathcal{H}^2$, $\mathcal{H}^3$, etc.). Often these equations can be reduced to a few ordinary differential ones, which can be solved on a small computer, so we sometimes refer to these generally as “semi-analytic” methods.

- **Full Numerical Simulations:** In this case one solves Einstein’s field equations on a grid of three spatial dimensions that evolves in time, or even in four dimensions. This is very similar to large plasma or weather simulations, but the equations are many more in number and much more difficult to handle. This area of study is called “numerical relativity”.

People in the field have found it convenient to break up the final approach and merger of a relativistic binary star into three distinct phases (see Fig. 8.4):

1. **“Inspirational Phase”:** This is similar to what we discussed in Section 8.1.4, where approximately sinusoidal gravitational waves are emitted by a distinct pair of stars with an ever-shrinking orbit.
2. **“Merger Phase”:** This begins when each star starts to feel the strong tidal forces of the other and continues into the time when the two coalesce and form a single, but very messy, black hole horizon.
3. **“Ringdown Phase”:** This final phase starts after a distinct, spheroidal black hole has formed, but its horizon has significant quadrupolar asymmetries that are being radiated away in a *decaying* sinusoidal-like wave.

The inspiral phase can be treated fairly well with analytic methods when the separation is larger than a few tens of Schwarzschild radii (as we did above) and with post-Newtonian methods inside this point. And the ringdown can be handled with linear perturbations of the Kerr metric. However, the only known method for dealing with the merger phase is performing full numerical simulations of Einstein’s equations. Fortunately, while the first two methods require a lot of mathematics to demonstrate, the numerical simulations (while even more complex to implement) are straightforward to demonstrate in a book of this type. So we will not discuss all the great work done by many relativists over the last several decades on the inspiral and ringdown phases, nor will we discuss any of the complex details of how numerical relativity is implemented. Instead, we will go straight to the recent, beautiful numerical results that solve the whole problem at once.

8.2.2 Simulating Colliding Black Holes on Supercomputers

For simplicity, we shall first deal with the collision of black holes *only*. This is the problem that most numerical relativists have decided to study, and it means that they need not deal with the stress, energy, and momentum of any matter in the system ($\mathcal{T} = 0$). The equations that they need to solve look deceptively simple

$$\mathcal{G} = 0 \tag{8.42}$$

Some numerical relativists, however, have taken the next step and re-introduced $\mathcal{T}$ in order to study the merger of two *neutron* stars as they form a black hole. We will deal with those more complex, and equally interesting, simulations later in Chapter 10.

Because $\mathcal{G}$ is symmetric, equation (8.42) consists of ten partial differential equations. However, the field constraints $\mathcal{G}_{00} = \mathcal{G}_{01} = \mathcal{G}_{02} = \mathcal{G}_{03} = 0$ are redundant with the other six equations (like the field constraints on the left side of Maxwell's equations (6.73) and (6.74) are redundant with the Ampère's and Faraday's laws on the right side). This leaves only six equations for the ten metric unknowns $g_{00}, g_{01}, \dots, g_{13}, g_{33}$. As before, the four remaining equations needed are obtained by choosing four gauge conditions that uniquely specify how the coordinates are to be laid out in space and time.

8.2.2.1 Early History of the Difficult Subject of Numerical Relativity: The 1970s to 2005

When this book was begun early in the first decade of the third millennium, numerical relativists had been working this problem for more than 30 years. However, after all this time they still were not able to evolve two colliding black holes more than one orbit before all the simulations crashed (*i.e.*, meaning that many numbers representing spacetime in the computer became infinite, forcing the computer to halt what it was doing because the results no longer could be trusted). Many problems had been solved up to this point, however:

- A consistent method for integrating the Einstein evolution equations forward in time, and satisfying the constraint equations implicitly with that evolution, had been developed in the early 1960s [318]. The ADM method uses the 3+1 split metric of equations (7.62) or (7.63) and allows for the lapse, shift, and spatial metric to be functions of space and time. Only the six independent components of $\mathbf{g}$ are evolved, however; α and β are determined by the choice of coordinates.
- Good methods also were developed for laying down coordinates in a constantly-changing geometry without creating *coordinate* infinities or “singularities” (such as one gets for longitude at the north and south poles). This was initially a big problem when black hole horizons formed, where Schwarzschild–Hilbert-like coordinates become infinite. These gauge choices are often referred to as “slicing” spacetime.
- Good methods had been developed to deal with the *physical* singularity that occurs at the center of a black hole. The values of the four-metric coefficients g become infinite⁴ there. So the center of the black hole was cut out of the grid or “excised”.
- Based on modern techniques used to perform plasma and weather simulations, good methods were developed for the stable time integration of the Einstein evolution equations. Without these methods, the computer simulation of the six Ein-

⁴ And the stress-energy $\mathcal{T}$ there is no longer zero, but infinite.

stein field equations would proceed correctly initially, but after a while would produce incorrect results that no longer satisfied the other four constraint equations. Eventually this also would develop infinities as well and the simulation had to be stopped. One of the more popular of these techniques is a reformulation of the ADM method developed by Masaru Shibata and Takashi Nakamura [328], and expanded upon by Thomas Baumgarte and Stuart Shapiro [329], called the BSSN method. The basic idea is to write the time-dependent Einstein field equations as set of first-order, symmetric and “hyperbolic” partial differential equations and then evolve them using the most advanced numerical hydrodynamic techniques.

- Furthermore, it turned out that initial conditions in numerical relativity are much more crucial than they are in plasma or weather simulations. It is not enough to simply set up an approximation to the starting model and allow the evolution to work out all errors in that approximation. For example, one has to take into account the gravitational waves that were emitted *before* the simulation starts. So much work was done developing these so-called “initial data” sets over the past few decades.
- Finally, over the decades much more powerful supercomputers were developed by the computing industry, allowing the time-dependent simulations to go from only two spatial dimensions to the three dimensions.

While all of these advances helped greatly in improving behavior of the numerical solutions, until 2005 there still were no simulations of black hole mergers possible. They still crashed after a short amount of evolution. Looking back, it appears that the main problem was that cutting out the singularity of the black hole produced errors in the (implicit) constraint equations that grew with time. Correctly handling these errors proved to be the key to solving the black hole merger problem.

8.2.2.2 Success!: 2005 and beyond

The first person to successfully solve these problems, and produce a full numerical simulation of merging black holes, was Frans Pretorius (then at Caltech) [330]. To deal with the constraint error problem he employed a gauge called “harmonic coordinates”, which chooses the coordinate system by solving yet another wave equation. This damps out the constraint errors instead of causing them to grow. Fig. 8.5 shows his first fully numerical results for the inspiral, merger, and ringdown of two equal-mass black holes for about 2.5 orbits. The final merged black hole ends up with a Kerr spin parameter of $j \approx 0.7$. Unlike regular astronomical observations, with a waveform like that in Fig. 8.5, we can *directly* measure the distance of the object: the period of the oscillations gives the mass of the binary, and the power of the wave predicted by these simulations, combined with the strength of the observed system in the gravitational wave detector, can be used to directly calculate its distance.

About four months after the Pretorius paper was initially circulated, two additional groups – one at The University of Texas at Brownsville, led by Manuela

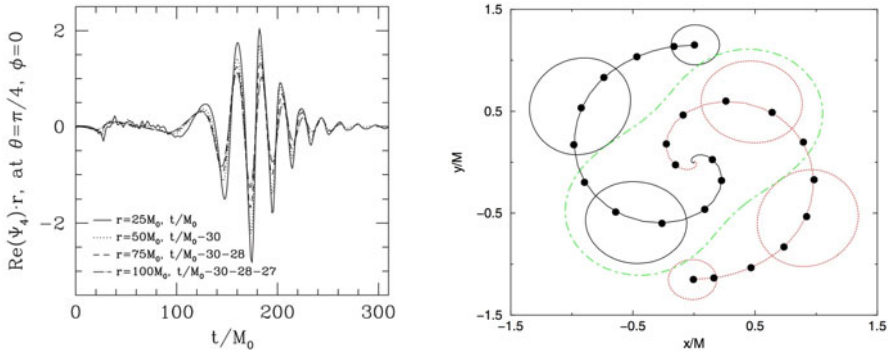


Fig. 8.5: First numerical simulations of gravitational waves generated by two merging Schwarzschild black holes of equal mass. *Left panel:* Pretorius’s first waveform simulation in 2005 [330] showing all three phases in Fig. 8.4. *Right panel:* Orbital evolution of a similar simulation by the UT Brownsville group in 2006 [332]. Filled circles show position of the two black holes every $2.5 \tau_g$, and black and red ovals show their “apparent” horizons (surfaces in the simulation where it is apparent that light cannot escape and which are always inside the true event horizon). Eventually, a *single common horizon* forms for both holes (green curve). The largest amplitude waves then are emitted, followed by a rapid ringdown as the merger product relaxes into a spherical state by emitting more gravitational radiation. Figures reprinted with permission; copyright (2005, 2006) by the American Physical Society (APS).

Campanelli and Carlos Lousto [331, 332], and one at NASA’s Goddard Spaceflight Center, led by John Baker and Joan Centrella [333] – also achieved success, but using quite a different method. They employed the standard BSSN evolutionary schemes and gauges, but did *not* excise the black hole’s central singularities out of the grid. Instead, they used an analytic approximation to the singularity called a “puncture”. The success of their efforts lay in properly advecting (dragging around) these punctures on the computational grid during the binary motion, *without* introducing large errors. Even though developed independently, when the two groups first compared results at a scientific meeting in mid 2006, they found that their results were remarkably similar to each other, and to those of Pretorius: about 3.5% of the rest mass of the final black hole is radiated as gravitational wave energy.

This sudden breakthrough in numerical relativity simulation, after decades of effort, is of tremendous importance. It allows theorists to investigate the predictions of general relativity for very strong gravitational events (*e.g.*, black hole mergers, *etc.*). When gravitational wave telescopes like LIGO and LISA detect such events, the results can be compared to the simulations to see if Einstein was correct. This has very important implications for our understanding of the fundamental physics of gravity. Such experimental confirmation (or even refutation) of the theory of general relativity is of Nobel Prize caliber. Furthermore, once we are certain that we know what the correct theory of gravity is, we will be able to turn the data analysis around and use it to determine the physics and astrophysics of black hole systems. This will begin a new age in the study of the universe.

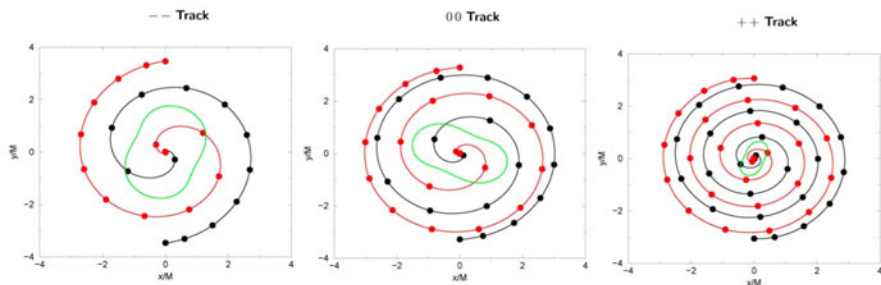


Fig. 8.6: Orbital evolution of spinning black hole systems with three different spin–orbit alignments. *Left panel:* Two holes with $j = -0.75$ (spinning in the opposite direction to the orbital angular momentum). *Center panel:* Same as Fig. 8.5 (i.e., two $j = 0$ holes), but simulated for a much longer time. *Right panel:* Two holes with $j = +0.75$ (spinning in the same direction as the orbital angular momentum). The behavior of each system depends on the total angular momentum (binary plus holes). The ++ simulation “hangs up” until it can radiate enough angular momentum to allow the merger to take place. The -- simulation plunges quickly to the common horizon stage, since there is even less angular momentum in the system than in the 00 case. Left and right panels reprinted with permission from reference [334]; copyright (2006) by the APS. Center panel kindly supplied by M. Campanelli; reproduced with permission.

8.2.3 Determining the Physics of Black Hole Systems by Comparing Simulations and Observations

Encoded in the gravitational wave train from two merging black holes will be a wealth of data about the properties of those holes: their masses, spins, orbital eccentricity, even the distance from the earth to the binary system. To illustrate this, consider one recent problem that was studied by the UT Brownsville group: the merger of two *spinning* Kerr black holes.

As we learned in Chapter 7, a black hole cannot have any more angular momentum than $J_{max} = G M^2/c$ (or $j = 1$). If the inspiraling binary has *more* than this amount of angular momentum, a single black hole cannot form until the excess $J_{tot} - J_{max}$ is radiated away in gravitational waves. This is called the “orbital hangup” effect.

Now, from Pretorius’s early work, there is no hangup when the two black holes are essentially non-spinning Schwarzschild ones, and the final single hole has $j \approx 0.7$. However, when the two black holes are *both* spinning, and when their spins are *aligned* with the *orbital* angular momentum vector, then we can have $J_{tot} > J_{max}$ and therefore orbital hangup.

Fig. 8.6 shows three different simulations, by the UT Brownsville group, of three different pairs of colliding black holes: one with each black hole having $j = -0.75$, one with each having $j = 0$, and one with $j = +0.75$ for each. The zero angular momentum case acts the same as Pretorius’s first simulation, merging in about two orbits. The negative angular momentum binary actually spirals together *faster* – in only one orbit; that is, there is less *total* angular momentum (including both spin and

orbital) in the binary that needs to be radiated away. However, in the positive angular momentum case, the merger takes more than three orbits to form the single black hole product. Because there is so much angular momentum in this latter system, the final single black hole cannot absorb all of the orbital part of the angular momentum, so the latter must be radiated away before the two holes can merge.

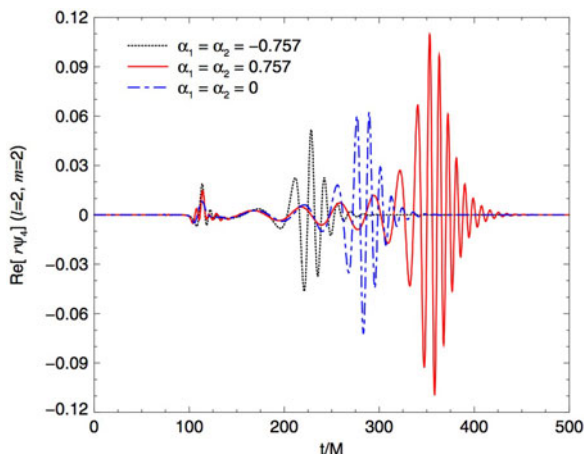


Fig. 8.7: Comparison of gravitational waveforms from the simulations in Fig. 8.6. Black curve shows the $--$ simulation, blue shows the 00 one, and red shows the $++$ one. The waveforms show progressively more cycles and a longer hangup time as the system angular momentum increases. The early spurious pulse is a “gauge wave” that is due to slight errors in the initial data [331, 333]. Reproduced from Fig. 1 of reference [335], with permission of IOP Publishing Ltd.

Fig. 8.7 displays the gravitational wave forms of the three different cases, showing how the key elements of the wave forms (number of final orbits, duration of the wave pulse, type of ringdown) all depend on how fast each hole is spinning and on the orientation of those spins.

8.2.4 Astrophysics of Merging Black Holes from Gravitational Wave Simulations: Black Hole Kick Velocities

8.2.4.1 Kicks from Merging Schwarzschild Holes

Even if the two black holes are not spinning, some interesting effects can occur if they do not have exactly the same mass. The NASA Goddard group simulated the merger of two holes with $m_1/m_2 = 1.5$. In this case, when the two holes finally merge, the last large gravitational wave pulse emitted is radiated in a specific direction, not isotropically. This wave carries quite a lot of energy (mass) and so produces a rocket effect or “recoil” [336].

The amount of kick is not large by relativity standards – about $50\text{--}100\text{ km s}^{-1}$ for mass ratios in the range of 1.5:1 to 10:1. However, the recoil is very important astrophysically, because any merged black hole, *of any size*, could have this sort of speed relative to its host galaxy. This implies that galaxies with masses less than $\sim 10^9 M_\odot$ (including the mass of their halos) will not be able to hold their merged black holes for more than a few tens of million of years. Therefore, many nearly-equal-mass merging binary black holes that form in small dwarf galaxies in the early universe may escape into the intergalactic medium. And those formed in globular clusters may escape into the host galaxy's outer *interstellar* medium.

Large mass ratios ($>10:1$) are immune to this effect, because the recoil is so small in that case. So, if a large black hole can be formed quickly with a small kick, and if it can be fed only with small black holes, then the large black hole probably will remain in the dwarf galaxy or globular cluster and not become part of an intergalactic population of black holes.

8.2.4.2 Kicks from Merging Kerr Holes

A much richer set of possibilities arises, however, when the merger of two Kerr black holes is considered. And that work advanced very far just in the two years after Pretorius's initial paper. As was well put by John Baker and his colleagues near the end of 2007, full numerical simulations had been performed for “non-spinning black holes with different mass ratios, ... black holes with equal masses and spins initially orthogonal to the orbital plane, ... spins initially parallel to the orbital plane, ... spins initially oriented at some angle between the orbital plane and the orbital angular momentum [vector], ... with unequal masses and spins initially either parallel to the orbital plane or oriented at some angle between the orbital plane and the orbital angular momentum [vector]” [337]. One very interesting result from these numerical simulation was that very large (but still sub-relativistic) kick velocities of hundreds to thousands of kilometers per second could result from black hole mergers. A velocity that high would eject the merger product (both merging holes) out of the galaxy and its halo into intergalactic space! The implication was that some, perhaps many, galaxies could have *no* black hole in their centers, *and* intergalactic space could be filled with runaway supermassive black holes of all sizes.

In order to make sense of all of the numerical results, these investigators and several others derived analytic forms for the kick velocities by expanding the Einstein field equations in powers of v^2/c^2 and $GM/(c^2 r)$ (*i.e.*, post-Newtonian expansions). The coefficients in these formulae then were fit to the less general numerical results (more accurate in the strong field cases) to arrive at general analytic formulae for the merger kick velocity *vector* as a function of parameters like binary mass ratio (q), initial spin vector components parallel and perpendicular to the orbital angular momentum vector ($j_1^\parallel, j_2^\parallel, j_1^\perp, j_2^\perp$), several orientations and difference angles of spins in the orbital plane, *etc.* The current best formulae have third and fifth order corrections with particular attention paid to the case when the black hole spins lie in the orbital plane, and they fit all the available numerical results to better than 5%

accuracy [338]. For example, the current estimate of the maximum recoil velocity (for a black hole binary of any mass) is $3680 \pm 130 \text{ km s}^{-1}$ [339], and this occurs when the two black hole spins lie in the orbital plane.

8.2.4.3 Are Galaxies Losing their Central Black Holes?

Such a gravitational recoil would send any newly-merged black hole far away into intergalactic space. The trip to escape its galactic host would take only a few million years, leaving the galaxy without its central supermassive object. After a billion years the hole would be several megaparsecs away from its original host. The universe could absorb an enormous amount of these errant wanderers without us noticing; they would be very dark, their number density would be not much larger than several times that of whole galaxies, and their mass density would be hundreds of times less than that of galaxies and their halos.

However, we *would* notice a large population of galaxies without central black holes, even if ones without holes eventually were to have their central regions replenished by merging with a galaxy that did have one. But, we do not see such a population of “hole-less” galaxies.

The most likely answer to this dilemma is that, while the above numerical simulations are certainly correct, they tell only part of the story. They solve a primarily *physical* problem, not an *astrophysical* one. That is, the simulations do not take into account how supermassive black holes interact with their stellar and gaseous environment prior to the merger. This becomes particularly important when one realizes that the two SMBHs are relatively minor constituents of their much more massive galaxies and halos. For some time now, it has been realized that the spins of two pre-merger black holes can be correlated through spin–orbit couplings with their surrounding star and gas distributions [340]. So far, astrophysical effects all seem to align, or anti-align, the black hole spins with their eventual *orbital* angular momentum vector. As a result, the actual recoil velocity of most astrophysical merging black hole binaries should be rather small: $\ll 70 \text{ km s}^{-1}$ [341]. Only very small stellar systems ($\ll 10^9 M_{\odot}$) then would lose their black holes during a merger.

It is possible that in the young universe there were large numbers of very low-mass galaxies (similar in mass to globular clusters) that temporarily lost their central black holes. But, eventually these merged with ones that did, building up to the galaxies we see today with escape speeds well above 70 km s^{-1} .

8.3 Basic Gravitational Collapse: Formation of a Black Hole Horizon in Collapsing Matter

The merging of two black holes to form a single, larger one is a demonstration of a fundamental property of these objects: when a single black hole encounters *any*

kind of matter, and merges with some of it, the result always is a black hole of larger mass. The original black hole always will grow or “accrete”; *a black hole cannot be torn apart or split into multiple black holes.*

This raises some interesting questions. A black hole merging with matter, or another black hole, creates a bigger one, but how does a black hole form in the first place from regular (non-gravitationally-collapsed) matter? And, when we observe a very massive black hole, how can we tell whether it formed in one single event or instead grew from a very small size by accreting a lot of stars and gas?

The answer to the first question, we believe, is the process of “gravitational collapse”, which we will briefly outline in the section below. Einstein’s equations, in fact, allow for matter to contract smoothly to such a dense state that a horizon will form in that matter before it collapses to infinite density. Once that horizon forms, the matter inside it cannot escape, and a black hole now exists.

The answer to the second question is complex and will be explored in Chapter 11. In fact, by simply *looking* at the black hole alone, one *cannot* tell how it formed and grew. A $10^6 M_\odot$ black hole with a spin of $j = 0.1$, say, that was formed by the collapse of a $10^6 M_\odot$ star will look identical to a million solar mass black hole (also with $j = 0.1$) that was formed over time by the merger of one thousand $10^3 M_\odot$ black holes. However, by studying the various astrophysical processes that lead to such black holes, and by observing the *environment* of such a black hole, we should be able to form a likely hypothesis of how it came to be and be able to test that hypothesis with future observations.

8.3.1 The Equations for the Structure and Evolution of Spherical, Adiabatic, Relativistic Stars and Gas Clouds

In order to demonstrate black hole formation, we shall choose a much simpler problem than usually occurs in nature: a ball (*i.e.*, star or cloud) of gas that is spherically symmetric, with structure that depends only on time t and on the radial distance r measured from the center of the object. We will not consider any motions that are not radial; that excludes rotation, convection, turbulence, jets, *etc.* Furthermore, we will assume that entropy in the gas star/cloud is neither generated nor lost – the so-called “adiabatic” approximation. So, we will exclude processes like nuclear burning, radiation losses, *etc.* However, our equations *will* include adiabatic compressional heating and expansive cooling. The gas also will have a pressure $p(r, t)$, a velocity $u(r, t)$, a rest mass density $\rho(r, t)$, and an internal energy density $\varepsilon(r, t)$. And p and ε will be related to each other through a thermodynamic equation of state, which we will assume to be polytropic. (More realistic equations of state will be discussed in the next chapter.) Finally, not only will we include the gravity of the gas in the evolution equations, we also will allow for that gravitational field to be *very* strong and use Einstein’s (spherically symmetric) field equations to describe it.

For these calculations, as we did in Chapter 5 (see Fig. 5.3), it is easiest to label each spherical shell in the ball of gas with the amount of *rest mass* that each encloses

$$m = \int_0^V \rho dV$$

where dV is the *proper* differential volume element of each thin spherical mass shell. We then will identify the velocity u as the speed with which the radius of a given mass shell expands or contracts

$$u \equiv \frac{\partial r}{\partial \tau}$$

τ is not a global time coordinate (unlike the Schwarzschild t). Rather, it ticks at a different rate in each mass shell, because each shell is moving (changing its radius) at a different rate and because each feels a different gravitational field from the mass shells *inside* their radius.

The May–White/Misner–Sharp equations of adiabatic, spherical stellar evolution are derived in Appendix C from Einstein’s field equations (7.19) and from the conservation laws generated by them (7.14). The equation of motion for each mass shell is

$$\xi \rho \frac{\partial u}{\partial \tau} = -W^2 \frac{\partial p}{\partial r} - \frac{G (\mathfrak{M} + 4\pi r^3 p / c^2)}{r^2} \rho \xi \quad (8.43)$$

where

$$\xi = 1 + \frac{(\varepsilon + p)}{\rho c^2} \equiv 1 + \frac{\mathfrak{h}}{c^2}$$

and $\mathfrak{h} \equiv h/\rho = (\varepsilon + p)/\rho$ is the specific enthalpy (*i.e.*, enthalpy per unit mass),

$$W = \left(1 + \frac{u^2}{c^2} - \frac{2G\mathfrak{M}}{c^2 r} \right)^{1/2} \quad (8.44)$$

is a generalized geometric/boost factor, and

$$\mathfrak{M} = \int_0^V W \left(\rho + \frac{\varepsilon}{c^2} \right) dV \quad (8.45)$$

is the amount of gravitating mass-energy enclosed by each mass shell. Also helpful will be a relation derived in Appendix C, which give us the derivative of $\mathfrak{M}$ with respect to time (equation (C.16))

$$\frac{\partial \mathfrak{M}}{\partial \tau} = -4\pi r^2 \frac{p}{c^2} u \quad (8.46)$$

Equation (8.43) is similar to the familiar *non*-relativistic equation of motion for a stellar mass shell

$$\rho \frac{\partial v}{\partial t} = -\frac{\partial p}{\partial r} - \frac{G m \rho}{r^2}$$

with some notable relativistic changes:

- The inertial and gravitational mass of the local gas are both enhanced by the specific enthalpy (ρ is replaced by $\xi \rho$).
- The velocity u is the r -component of the four-velocity $\mathbf{u}$; it is *not* the r -component of the three-velocity $\mathbf{v}$. That is, u ranges from $-\infty$ to $+\infty$, not from $-c$ to $+c$.
- The pressure gradient is enhanced by two geometrical factors of W .
- The gravitating mass $\mathfrak{M}$ includes not only internal energy (random plasma particle motions) but also bulk kinetic energy of the moving mass shells lying interior to each mass shell (note the factor of W in equation (8.45)).
- Finally, there is a correction term to the gravitating mass contributed by the *local pressure at r* ($4\pi r^2 p / c^2$), in addition to the internal energy integral in equation (8.45). This mass contribution will be very important in gravitational collapse.

In addition to the conservation of momentum (equation (8.43)), there also is the adiabatic conservation of energy equation

$$\frac{\partial(\varepsilon/\rho)}{\partial\tau} + p \frac{\partial(1/\rho)}{\partial\tau} = 0 \quad (8.47)$$

and the conservation of mass ($m = \text{constant}$) becomes an equation for the evolution of the rest mass density

$$\frac{\partial(\rho r^2)}{\partial\tau} = -\rho r^2 \frac{\partial u}{\partial r} \quad (8.48)$$

A word needs to be said about time. Each mass shell's clock can potentially click at a different rate, so we need to define a global (Schwarzschild–Hilbert-like) time t . The metric then will be given by

$$ds^2 = -\alpha^2 c^2 dt^2 + \frac{1}{W^2} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 \quad (8.49)$$

and the value for the lapse α can be found by integrating the pressure gradient *inward* from the surface of the star/cloud (see Appendix C). If we define $\mathcal{A}$, R , and U to be the values of the metric time coefficient α (in equation (8.49)), the shell radius r , and the velocity u on the outer surface of the *outermost* mass shell (*i.e.*, the outer surface of the star or cloud where $\mathfrak{M} = M$), then that integral is given by

$$\ln \frac{\alpha}{\mathcal{A}} = \int_r^R \frac{1}{\xi \rho c^2} \frac{\partial p}{\partial r} dr \quad (8.50)$$

where $\mathcal{A}$ is usually taken to be

$$\mathcal{A} = \frac{1 - \frac{2GM}{c^2 R}}{\left[1 + \frac{U^2}{c^2} - \frac{2GM}{c^2 R}\right]^{1/2}} \quad (8.51)$$

In the above equations for conservation of mass, momentum, and energy, we have hidden the fact that the coefficient of time α is different in different mass shells by defining the proper time derivative

$$\frac{\partial t}{\partial \tau} = \frac{1}{\alpha} \quad (8.52)$$

So, while equation (8.43) may look fairly simple to integrate, it turns out that it is not trivial to do so. If we want to perform a simulation and advance every mass shell according to the same clock (t), we really need to use equation (8.50) to obtain α at each time step, integrate the acceleration in time t

$$\frac{\partial u}{\partial t} = -\alpha \left[\frac{W^2}{\xi \rho} \frac{\partial p}{\partial r} + \frac{G (\mathfrak{M} + 4\pi r^3 p c^2)}{r^2} \right]$$

to get the radial velocity u of each mass shell and then integrate that velocity

$$\frac{\partial r}{\partial t} = \alpha u$$

to find out how each mass shell's radius varies with time.

8.3.2 Collapse of Pressure-Free Dust to a Black Hole

One of the simplest, and most instructive, types of gravitational collapse problems to study is the evolution of a ball of matter that has *zero pressure*. This is not quite as unrealistic a situation as it might seem. Collisions between interstellar dust grains, each with sizes of order 1–1000 nm but separations of several meters, are so infrequent that the pressure due to grain–grain interactions is essentially zero. (In fact, a dust cloud of $10^9 M_\odot$ could have a radius of order the Schwarzschild radius (20 AU) and, yet, the grain–grain distance still would be considerable larger than the grain size itself.) For this reason, the investigation of a gas with zero pressure is often referred to as the evolution of pressure-free dust.

8.3.2.1 Collapse of a Single Dust Shell

Setting $p = \varepsilon = 0$ in the spherical structure and evolution equations (8.43) to (8.48) and (8.50) to (8.52) yields a greatly simplified equation of motion for each mass shell

$$\frac{\partial u}{\partial \tau} = - \frac{G \mathfrak{M}}{r^2}$$

which looks exactly the same as the free fall equation in Newtonian gravity. In fact, for the case of pressure-free dust, it *is* exactly the same: equation (8.46) shows that, when $p = 0$, the gravitating mass $\mathfrak{M}$ is constant with time for each mass shell, even though it may differ for different mass shells. So, as a given mass shell expands or contracts, it will always see the same gravitational mass $\mathfrak{M}$ inside of it.

Let us now concentrate on only one of the mass shells in the star – one that begins with a radius r_0 at $t = 0$. Because the above equation for $\partial u / \partial \tau$ is identical to the Newtonian one (5.17), we can directly apply the same integration method to arrive at the following solutions that are analogous to equations (5.18), (5.19), and (5.20)

$$u = - \left(\frac{2G\mathfrak{M}}{r_0} \right)^{1/2} \left(\frac{r_0}{r} - 1 \right)^{1/2} \quad (8.53)$$

$$r = \frac{r_0}{2} (1 + \cos v) \quad (8.54)$$

$$\tau = \frac{\tau_0}{2} (v + \sin v) \quad (8.55)$$

where v is the collapse parameter (ranging from 0 to π) and

$$\tau_0 \equiv \left(\frac{r_0^3}{2G\mathfrak{M}} \right)^{1/2}$$

is the characteristic dynamical time for the mass shell we currently studying, and $\tau_{\text{ff}} = \pi \tau_0 / 2$ is the free-fall time (same as equation (5.21) in these coordinates). We also can plug equation (8.53) into equation (8.44) to get a very simple, familiar, and time-independent expression for the spatial geometric factor

$$W = \left(1 - \frac{2G\mathfrak{M}}{c^2 r_0} \right)^{1/2}$$

So, in the freely-falling coordinate system that we have chosen, each mass shell will maintain not only a constant $\mathfrak{M}$ but also a constant W , although their actual values will be different for each shell.

The proper velocity at each v is given by equation (5.22)

$$u = - \left(\frac{2G\mathfrak{M}}{r_0} \right)^{1/2} \tan \left(\frac{v}{2} \right)$$

starting at $u = 0$ and becoming *infinite* as $v \rightarrow \pi$. This means, of course, that the three-velocity remains finite and approaches the speed of light as $v \rightarrow \pi$

$$V^{\hat{r}} = \frac{u}{(1 + u^2/c^2)^{1/2}} = -c \frac{1}{\left[1 + \frac{r_0}{r_s} \cot^2 \left(\frac{v}{2} \right) \right]^{1/2}} \quad (8.56)$$

where $r_S = 2G\mathfrak{M}/c^2$ is the Schwarzschild radius for each mass shell $\mathfrak{M}$.

A quick look at equation (8.56) tells us that, unlike the free fall of a particle into a Schwarzschild black hole (equation (7.37)), the infall speed here does not reach $-c$ until we reach $r = 0$. In the earlier infall case in Chapter 7, the speed reached $-c$ when $r = r_S$. Why is this? Should not collapse of a dust shell be identical to the free fall of a particle into a black hole? The answer, actually, is yes; it *is* identical. What has changed here is not the physics but the coordinate system and, therefore, the way in which velocity is measured. In the Schwarzschild case in Section 7.4.3, distance and time were measured in global Schwarzschild–Hilbert coordinates – by an observer watching the black hole from a great distance. Here we (the observers) are *inside* the collapse, sitting on the surface of a particular mass shell. In that case we measure time differently; in fact, we fall into the horizon in a rather short (not infinite) amount of local time and continue to fall in until we reach $r = 0$. In Chapter 7, had we used a coordinate system that followed the falling particle, we would have obtained the same answer as in equation (8.56).

Equations (8.54) and (8.55) tell us that the shell collapses to zero radius at a time $\tau = \pi\tau_0/2$. So, at what time (before that) does it cross its own horizon? From equation (8.54) we can solve for the collapse parameter v when $r = r_S$

$$v_S = \cos^{-1} \left(\frac{2r_S}{r_0} - 1 \right) \quad (8.57)$$

and plug this into equation (8.55) to obtain the proper time when the shell collapses into its own horizon

$$\tau_S \equiv \tau_0 \left[\frac{1}{2} \cos^{-1} \left(\frac{2r_S}{r_0} - 1 \right) + \left(\frac{r_S}{r_0} \right)^{1/2} \left(1 - \frac{r_S}{r_0} \right)^{1/2} \right] \quad (8.58)$$

This time will be different for each mass shell, because τ_0 , r_S and r_0 are different for each mass shell. So this leaves the question of black hole formation by the entire cloud open. Which mass shell in the entire cloud forms an horizon first? Do they all form horizons at the same time? Does the horizon form in the center shell first and then grow outward? Does it form at some intermediate point in the dust cloud and move outward? Or does it form at the outer edge and envelope the entire cloud at once, immediately turning the entire object into a black hole? The answer will depend on what the initial distribution of matter $\rho(r, \tau = 0)$ was at the outset and how it evolves. To understand how a black hole forms, therefore, we need to consider a cloud with a particular initial density profile and look at its entire evolution all at once.

8.3.2.2 Collapse of an Entire Uniform Density Dust Cloud

In order to completely understand the collapse of an entire dust cloud to a black hole, we must consider all of the concentric mass shells together. We will begin by

Pressureless Dust Cloud Collapse

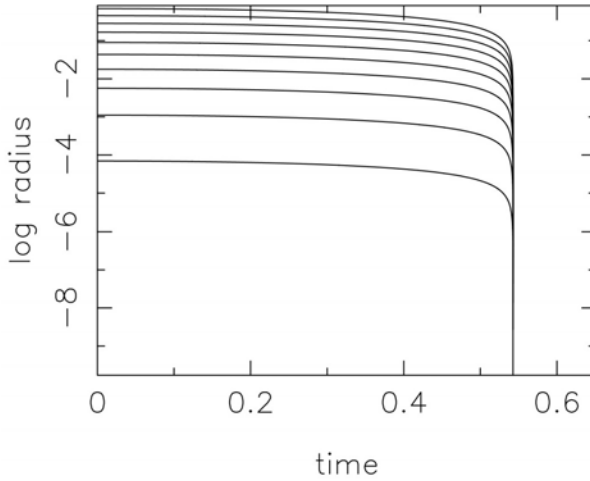


Fig. 8.8: Collapse of a uniform density, pressureless star. Curves show radius evolution of ten different mass shells having the following fractions of the stellar mass: $1e-12$, $4e-9$, $5e-7$, $1.7e-5$, $2e-4$, $2e-3$, 0.014 , 0.07 , 0.3 , and 1.0 . The evolution is computed using the relativistic eqs. (8.54) and (8.55), and each shell reaches zero radius at the same time. The evolution is quite different when the star has a realistic equation of state (Fig. 8.9).

studying a simply uniform density cloud with

$$\rho(r, \tau = 0) = \rho_0$$

Under our assumption of pressureless dust (p and ε both zero), the uniform density leads to a particularly simple expression for the (constant) gravitating mass interior to each shell (see equations (8.45), (C.5), and (C.13)). At $\tau = 0$ it is

$$\begin{aligned} \mathfrak{M} &= \int_0^V W \rho dV = \int_0^{r_0} W \rho 4\pi r^2 \frac{dr}{W} \\ &= \int_0^{r_0} 4\pi r^2 \rho dr = \frac{4\pi}{3} r_0^3 \rho_0 \end{aligned}$$

On the other hand, however, the expression for the enclosed *rest* mass is not as simple. At time $\tau = 0$ the rest mass is given by

$$m = \int_0^V \rho dV = \int_0^{r_0} \frac{4\pi r^2 \rho_0}{\left(1 - \frac{2G\mathfrak{M}}{c^2 r}\right)^{1/2}} dr$$

which can be integrated to obtain the ratio of gravitating mass $\mathfrak{M}$ to rest mass m inside each shell

$$\begin{aligned}\frac{\mathfrak{M}}{m} &= \frac{2}{3} \frac{\left(\frac{r_S}{r_0}\right)^{3/2}}{\left[\sin^{-1}\left(\frac{r_S}{r_0}\right)^{1/2} - \left(\frac{r_S}{r_0}\right)^{1/2} \left(1 - \frac{r_S}{r_0}\right)^{1/2}\right]} \\ &\approx \frac{1}{1 + \frac{3}{10} \frac{r_S}{r_0}} \quad (r_0 \ll r_S)\end{aligned}$$

In this case the gravitational mass is *less* than the rest mass, whereas we might have expected it to be larger. Why? The answer is related to our simple initial assumption of a zero pressure, zero velocity, and yet, partially-bound dust cloud. This assumes that, prior to $\tau = 0$, the binding mass released when the cloud became partially bound had already completely escaped the system. It did not go into kinetic energy of infall or into internal energy, because both of these are assumed to be initially zero. So, the initial binding energy of each shell in our uniform-density dust cloud must be negative

$$\mathfrak{M} - m = \frac{-\frac{3}{10} \frac{r_S}{r_0}}{\left(1 + \frac{3}{10} \frac{r_S}{r_0}\right)} m$$

The final strange and interesting property of a pressureless collapse is that, because $p = 0$, local time advances at the same rate everywhere in the cloud. The solution for the lapse function from equation (8.50) is then

$$\alpha = \mathcal{A}(\tau)$$

for *all* mass shells, independent of m (and also r_0). Now $\mathcal{A}(\tau)$ is not unity, and it is not constant with time; from equations (8.51) and (8.53) it is

$$\mathcal{A} = \frac{\left(1 - \frac{2GM}{c^2 R}\right)}{\left(1 - \frac{2GM}{c^2 R_0}\right)^{1/2}}$$

where R_0 is the initial cloud radius. By definition, global time t ticks at the same rate for all mass shells. However, we now have found that, in this case, *local* time τ *also* ticks at the same rate for all mass shells. So we can use either time coordinate to follow the global collapse. In fact, τ is not only more convenient mathematically (we do not have to integrate $\partial t / \partial \tau = \mathcal{A}(\tau)$), it also is more convenient physically: $t \rightarrow \infty$ as the cloud collapses to its *Schwarzschild* radius, whereas $\tau \rightarrow \frac{\pi}{2} \tau_0$ as the cloud collapses to *zero* radius! Such is the arbitrary nature of time in the realm of relativity.

Fig. 8.8 shows a plot of radius vs. time τ (equations (8.54) and (8.55)) for different shells in a uniform-density collapsing dust cloud. For each shell the value of the collapse parameter when an horizon forms is given by equation (8.57)

$$v_S = \cos^{-1} \left[\left(\frac{r_0}{c \tau_0} \right)^2 - 1 \right]$$

which is a monotonically *decreasing* function from the center of the cloud ($r_0 = 0$) to its surface ($r_0 = R_0$). So, in this uniform density case, a *black hole horizon forms first at the surface, immediately enveloping the whole cloud*. After that horizon forms, the entire cloud is now a black hole, even though it still is a dust cloud inside the horizon, and it still has some collapsing to do before it reaches $R = 0$.

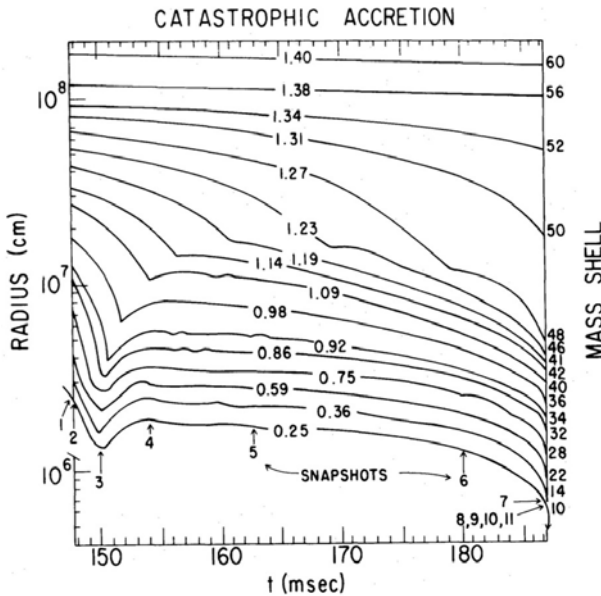


Fig. 8.9: Stellar collapse with a realistic equation of state. Unlike Fig. 8.8 the plot shows only the evolution after significant collapse has occurred. Note that the star does *not* proceed directly to the black hole stage; rather, it bounces at $t \approx 150$ ms. The inner core of the star then appears to reach a hydrostatic neutron star state for at least 20 ms. After another 20 ms, however, the relentless infall of the outer layers onto the core induces a second collapse, this time forming an horizon. Eventually the entire $1.4 M_\odot$ star will accrete into the new black hole. Reproduced from Fig. 1 of [342], by permission of the AAS.

8.3.3 More Realistic Collapse: Adiabatic Stars and Clouds with Pressure

As supercomputers became more powerful, groups began to study the May–White/Misner–Sharp equations for stellar structure and evolution in some detail. Two such

studies were done by Kenneth Van Riper of the University of Illinois in 1979 [342] and Stuart Shapiro and Saul Teukolsky of Cornell in 1980 [343]. Both of these investigations considered the collapse of stars with the simple polytropic equation of state (5.55). It was discovered that the polytropic index was very important to the behavior of the collapse, so several different values were studied, from the very stiff $n = 1$ (adiabatic index $\Gamma = 2$) to the softer values of $n = 3/2, 5/2, 3$, and 3.7 ($\Gamma = 5/3, 1.4, 4/3$, and 1.27). Sometimes the index was allowed to vary with density (say 3.7 at low density ($<10^{11} \text{ g cm}^{-3}$) and $5/2$ at high density ($>10^{13} \text{ g cm}^{-3}$)). This allowed one to build static initial models with $n = 3$, for example, and then start it collapsing by increasing n to 3.7 . Then, after the collapse ensued and the density increased dramatically in the core, n would begin to drop toward $4/3$ again.

There are two possible outcomes to a collapse of this nature:

1. The star's pressure and increasing adiabatic index Γ will be enough to eventually halt the collapse to a new, much more compact relativistic star of very high density.
2. The pressure and enthalpy terms in equation (8.43) will increase the gravitational force so much that they overwhelm the pressure *gradient*, and the star will continue to collapse to a black hole.

Pressure, therefore, can play different roles here: it can halt the collapse, or it can accelerate it, and it is important to understand when pressure prevents or assists black hole formation.

Fig. 8.9 shows an r vs. t plot from Van Riper's work for a $1.4 M_{\odot}$ star undergoing collapse in the above manner. In this case the abscissa is the global time coordinate t , not proper time τ , so the investigator had to integrate equations (8.50) and (8.52) in order to determine t for each mass shell. The first stages of the collapse last from $t = 0$ to $t = 150$ ms. Note that the core does not immediately form a black hole. Instead, it "bounces" and remains rather static for another several tens of milliseconds. The mass of the core that bounces is $\sim 1 M_{\odot}$ here.

During that prolonged period, the conservation of momentum is given approximately by the hydrostatic Tolman–Oppenheimer–Volkoff or TOV equation, which is equation (8.43) with $u \approx 0$

$$\frac{\partial p}{\partial r} = - \frac{1}{\left(1 - \frac{2G\mathfrak{M}}{c^2 r}\right)} \frac{G \left(\mathfrak{M} + 4\pi r^3 p/c^2\right) \xi \rho}{r^2} \quad (8.59)$$

This equation is so named because of the work done originally by Richard Tolman of Caltech and J. Robert Oppenheimer and George Volkoff of UC Berkeley in 1939 [344, 345]. There are two things to note about this equation that will affect the nature of the collapse. First, as p increases, the weight of the pressure's own gravity (through the enhanced gravitational mass $\mathfrak{M} + 4\pi r^3 p/c^2$ and the enhanced fluid mass density $\xi \rho = \rho + (\varepsilon + p)/c^2 = \rho + p\Gamma/(\Gamma - 1)c^2$) can overwhelm the pressure's own gradient. This is a relativistic effect only; in Newtonian stellar structure these pressure corrections to the gravitational force are absent (see Section 5.2.4). Secondly, if any part of the star should form an horizon (*i.e.*, $r \rightarrow 2G\mathfrak{M}/c^2$), all

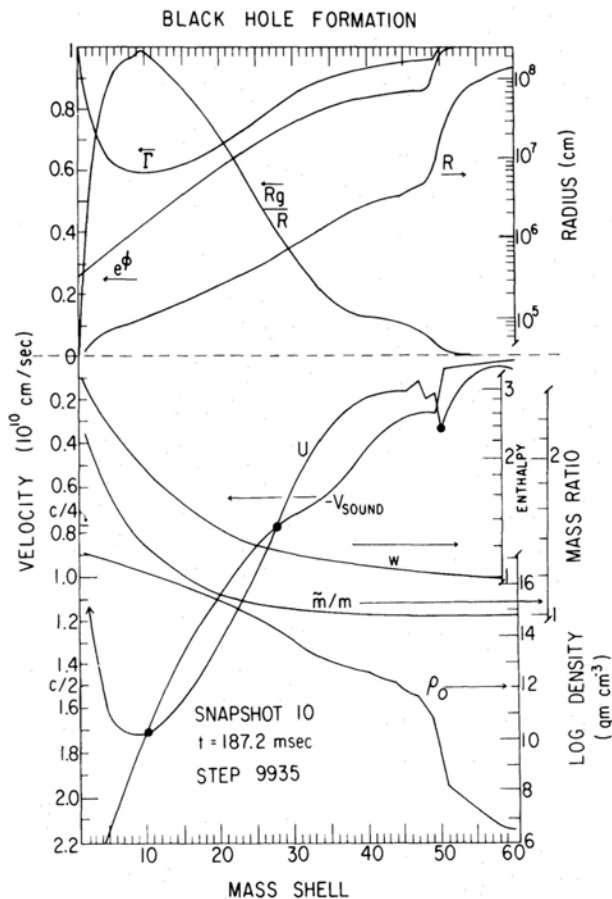


Fig. 8.10: State of the interior of the collapsed star in Fig. 8.9 at $t = 187.2$ ms when an horizon first forms. Top panel shows that it does so in the *interior* of the star (mass shell #9); in that shell $R_g/R = 1$ ($r_s/r = 1$ in our notation). Bottom panel shows various speeds, density, and two important quantities: $\tilde{m}/m$ ($\mathfrak{M}/m$ in our notation) and w (W in eq. (8.44)). Note how both $\mathfrak{M}$ and W are highly relativistic in the stellar core, but unity (non-relativistic) near its surface. Reproduced from Fig. 2 of [342], by permission of the AAS.

hope is lost. The gravitational force on the right-hand side of equation (8.59) becomes infinite, and no pressure gradient can overcome that.

We now are in a position to understand the rest of the evolution beyond $t = 170$ ms. If it were not for the additional $\sim 0.4 M_\odot$ of material raining down on the bounced core of $\sim 1 M_\odot$, the latter might have survived in the TOV static state. However, the weight of the additional material accreting onto the core continues to compress it, especially in the period $t \approx 170 - 185$ ms. By 185 ms another collapse has ensued, precisely because the enormous pressure p is now *causing* the collapse, rather than preventing it.

Then, at $t = 187.2 \text{ ms}$, an horizon forms with a radius of $r \approx 10^5 \text{ cm}$ and an enclosed mass of $\mathfrak{M} = 0.34 M_\odot$ (which has a rest mass of $m = 0.22 M_\odot$). Fig. 8.10 shows the conditions throughout the star at this exact instant. The horizon has formed in mass shell #9, where $r_S/r = 1$. In this case, because density (and pressure) were allowed to vary with radius in the initial star (rather than start with a uniform density as in the previous section), the horizon forms *inside* the star, not at its surface. Furthermore, because p and ε contribute here, the gravitational mass $\mathfrak{M}$ is greater than the rest mass m (0.34 vs. $0.22 M_\odot$). Only as we move outward in the star, and out of the relativistic potential, do we see $\mathfrak{M} \rightarrow m$.

After $t = 187.2 \text{ ms}$, the remaining material accretes into the small, nascent black hole, growing it into a $1.4 M_\odot$ one in a fraction of a second.

We see, then, that in the Einstein theory of gravity, a black hole can naturally form during the collapse of a stellar core, if the central density is high enough. However, whether the collapse actually does form a black hole, or simply forms a high-density (neutron) star, and what the final mass of that collapsed object is, will depend on many things: the equation of state of matter at high densities, the total mass that is collapsing, the time-dependent dynamics of that collapse and bounce, and finally the amount and nature of radiative energy released (which we have ignored here entirely). Even after more than 40 years of work on this subject, this still remains a very active field of astrophysical research. We will return to it and discuss the astrophysical implications of black hole formation in stars in Chapter 10.

Chapter 9

Nuts and Bolts of the Black Hole Engine: General Relativistic Mechanics

One of the most attractive, and also most daunting, features of astrophysics is that it brings together physics from many different fields and deals in the extremes of speed, gravity, temperature, and density. The deep gravitational potential of the black hole provides a stable engine block on which are hung all the key systems of the black hole engine. Many Schwarzschild radii away from the black hole lies the carburetion system. Fuel, in the form of gas clouds, or even whole stars, is tidally torn apart and dispersed into a smooth vapor of plasma flowing into the central regions of the engine. Within ten Schwarzschild radii lies the accretion disk combustion chamber, where the fuel releases its gravitational (not chemical or nuclear) energy, creating a power output greater than that of any other engine in the universe. Five exhaust systems carry angular momentum and released energy away from the combustion area: emitted light, viscous transport of angular momentum outward in a disk, a wind of thermal particles at speeds up to 10% the speed of light, winds and jets of nonthermal particles driven by a magnetic turbine to speeds up to 99% that of light, and even advection of plasma and heat into the black hole (an economical process that uses spent fuel to build up the engine block in both size and capability). Most of these exhaust systems are responsible for what we observe as macro- and microquasar activity. It is the job of astrophysicists to break the engine down into its key systems, and those into their sub-components, in order to understand how it operates, efficiently and unattended, for eons. The fact that there are billions of similar unattended engines all over the universe, operating in the same manner, makes this task all the more compelling.

In the previous chapters we spent a great deal of time working on the black hole engine block itself, and on how the gravitational and electromagnetic gauge fields generate conservation laws of energy, momentum, and charge. These conservation laws, operating within the gravitational field of the black hole and in the electromagnetic field of the plasma, are responsible for the inner workings of the engine components. This last chapter on black hole physics, therefore, will concentrate on the details of the conservation laws.¹ With this last set of designs in hand, we will

¹ Most of the material in the first section on conservation laws is based on MTW [305] and on a 2004 paper by the author [346]. That in the second section on equations of state is drawn from texts

be able to build black hole engines from the ground up and begin to understand how macro-, mini-, and microquasars operate.

9.1 Overview of General Relativistic Mechanics

Before discussing in detail the various fields and sub-fields of physics that are used to study black hole engines, it will be very helpful to give a brief overview of all of those areas – how they relate to one another and how they are derived. This global survey not only will help the reader understand the physical relationship between the chapters in this book, it also may help the more advanced reader decide which approach might be most useful in tackling a new problem in black hole research.

In a sense, all physics can be derived from a few simple laws that govern how matter interacts on a subatomic scale. The different fields of mechanics (particle, statistical, solid, fluid, *etc.*) simply use different levels of approximation to make the problem simpler and more manageable. Generally, for a small system (*e.g.*, a single atom or nucleus) very few approximations need to be made. For a larger system (*e.g.*, a supermassive black hole with stars and magnetized gas orbiting in its vicinity) many approximations probably will be necessary.

Fig. 9.1 shows the progression of approximation from quantum mechanics, through several important areas of classical mechanics used in the study of black holes, to the set of techniques that are the most common (general relativistic magnetohydrodynamics). Our overview will touch on each of these briefly.

9.1.1 Quantum Mechanics

Quantum mechanics is the most complete description of matter available. While it is possible to describe every physical system with it, in practice it is of course impossible. It is used primarily to describe systems from the subatomic to rather large, single organic molecules, plus certain properties of solids in which the wave properties of matter can be important.

In black hole physics quantum effects are important in black hole evaporation, information theory, and ultimately in the structure of spacetime on the subatomic level and how it interacts with matter there. This book will ignore all quantum properties of black holes, chiefly for two reasons: a full general relativistic description of quantum mechanics and the interaction of gravity and other fields on microscopic scales has not been fully worked out, and behaviors of black holes that require quantum mechanics to describe them are not (or have not yet been) observable by as-

on statistical and thermal physics [347, 348] and from Clayton's book on stellar evolution [349], with this author's own distillation and organization. Information in the remaining sections comes from the literature, with local citations therein.

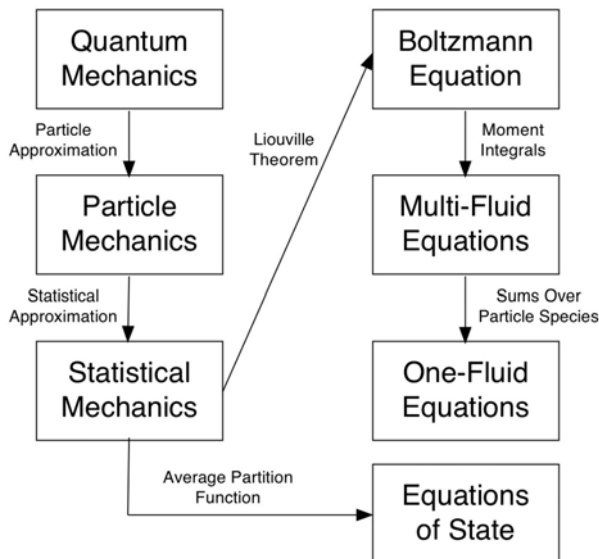


Fig. 9.1: Path through different fields of physics to reach the equations discussed in this chapter. The left-hand column is the same for both the one-fluid GRMHD and state equations. The path to the equations of state assumes that the plasma is in thermal and dynamical equilibrium. The path to one-fluid GRMHD, however, requires dealing with dynamically-evolving systems of particles of different types and then reducing that to a set of equations for the entire system.

tronomers. That is, so far there has been no need to use anything other than classical physics to describe observed black hole engines.

9.1.2 Particle Mechanics

The first approximation we can safely make, therefore, is that matter is composed of simple classical particles, rather than wavelike quanta. The particle approximation is used extensively in black hole astrophysics, not only in plasma flow around it, but also in the dynamics of whole stars trapped in its gravitational field.

9.1.2.1 Stellar Dynamics

In stellar dynamics each one of millions to many billions of stars in a cluster or galaxy is considered to be a single particle (of $\sim 10^{33}$ – 10^{35} g). The motion of each star is governed by the gravitational field created by the entire collection of stars, plus possibly one or more central black holes, with few or no stellar collisions. In four-dimensional relativistic language, the equation of motion for each star is very

simple (equation (6.65) with no forces)

$$\frac{d\mathbf{P}}{d\tau} = 0$$

where $\mathbf{P}$ is the particle four-momentum. That is, each star moves freely in the curved spacetime provided by all the objects in the stellar system. The gravitational “force” is hidden in the $d/d\tau$ total derivative. (See equations (7.31) and (7.32) for an example of this in Schwarzschild geometry.) Of course, many stellar systems can be described by Newtonian gravity, where the above equation of motion becomes

$$\frac{d\mathbf{V}}{dt} = -\nabla\psi$$

where $\mathbf{V}$ is the three-velocity of a given star, and $\psi(x, y, z)$ is the gravitational potential of the stellar system.

“N-body” simulations, which compute the motions of many stars, are employed extensively in the study of black hole formation and fueling. They are discussed a little more in Chapters 10 and 11.

9.1.2.2 Charged Particle Dynamics

A great many particle mechanics applications, of course, study large systems of single electron and ion particles. (In this case the mass of each particle is of order $\sim 10^{-27}$ – 10^{-24} g.) Unlike stellar dynamics, charged particle dynamics *does* have an important external force – electromagnetism. The four-dimensional relativistic equation of motion is again equation (6.65), but with the right-hand side being the Lorentz force in equation (6.116) (with the understood T superscript now dropped)

$$\frac{d\mathbf{P}}{d\tau} = \frac{q}{mc} \mathbf{P} \cdot \mathcal{F}$$

where q is the charge of a given particle, $\mathbf{P}/m$ the particle four-velocity, and $\mathcal{F}$ the Faraday tensor describing the electromagnetic field. (Again, any gravitational effects are hidden in the $d/d\tau$ derivative.)

Charged particle N-body simulations are sometimes used to study microscopic processes in relativistic jets and in very low-density accretion flows and winds near black holes. However, in this book we treat charged particles not as individual entities but as members of a large system of particles called a plasma.

9.1.3 Statistical Mechanics

The next level of approximation deals with the large number of particles in a statistical (probabilistic) manner, rather than trying to follow the motion of every particle.

Two such systems can be in the same thermodynamic state (same internal energy or same temperature) even though there is not a one-to-one correspondence between each particle in system #1 to those in system #2. The collection of all such possible systems in the same thermodynamic state is called an *ensemble*, and the fraction of systems in that ensemble in a specific state (*e.g.*, 50% of particle spins aligned) is the probability of that state occurring. Statistical mechanics is the science of calculating those probabilities and using that information to determine useful thermodynamic quantities of a plasma, such as internal energy, pressure, entropy, heat capacities, chemical potential, *etc.*

Like particle mechanics, statistical mechanics has had important applications in stellar dynamics. Before computers were powerful enough to perform large N-body simulations, the Fokker–Planck equation (which evolves the probability density function for star particles) was employed to study the evolution of globular clusters and galactic star systems. This is briefly discussed in Chapter 10. However, a still simpler statistical approach is taken in the case of studying the behavior of plasmas.

9.1.4 Kinetic Theory

The kinetic theory approximation brings us closer to standard fluid mechanics, but is more general and powerful than the latter while simpler than statistical mechanics. Kinetic theory considers a fluid or gas to be composed of many *systems* of particles, each occupying a small volume compared to the total size of the fluid being simulated but nevertheless still comprising a large number of particles. Each of these small systems is called a fluid element.² Only one function is of interest for each particle species in each fluid element: the number of particles at a given point in space with similar momenta in that space. Since this number depends on which three momentum components we choose to examine, as well as the three position components of each fluid element, plus time, the total distribution function depends on seven variables

$$f_i = f_i(\mathcal{X}, \mathcal{P}, t) = \frac{\partial^6 N_i}{\partial^3 \mathcal{X} \partial^3 \mathcal{P}}$$

where i represents a given particle species (electrons, protons, *etc.*). The six-dimensional space formed by coordinates $\mathcal{X}$ and $\mathcal{P}$ is called *phase space*, and f_i is simply the density of particles in a volume of that space.

The Liouville theorem of statistical mechanics states that f_i remains constant with time as it moves through phase space, unless it is affected by collisional processes. This theorem can be used to derive the Boltzmann equation for f_i

² Note that the collection of all these fluid elements is *not* an ensemble of different possible states. Instead, it is the entire fluid itself. In general, each fluid element is in a *different* thermodynamic state and has a different momentum in space.

$$\frac{\partial f_i}{\partial t} + \frac{\mathcal{P}}{m_i} \cdot \nabla f_i + \mathcal{F}_i \cdot \nabla_{\mathcal{P}} f_i = \left. \frac{\partial f_i}{\partial t} \right|_{\text{coll}} \quad (9.1)$$

where $\nabla_{\mathcal{P}}$ is the gradient in momentum space ($\partial/\partial\mathcal{P}^x$, $\partial/\partial\mathcal{P}^y$, $\partial/\partial\mathcal{P}^z$) and $\partial f_i/\partial t|_{\text{coll}}$ represents the rate at which particles are scattered into (or out of) that volume of phase space by collisions. Very often the Boltzmann equation is used for non-relativistic systems, so the force $\mathcal{F}_i$ includes both gravity and electromagnetism

$$\mathcal{F}_i = -m_i \nabla \psi + q_i \left(\mathbf{E} + \frac{\mathbf{v}_i}{c} \times \mathbf{B} \right)$$

(Note that we have used the script font ($\mathcal{X}$, $\mathcal{P}$, $\mathcal{V}$, $\mathcal{F}$), to denote quantities in six-dimensional phase space while regular bold math characters are used for average quantities in three-dimensional physical space.)

The general relativistic form of Boltzmann's equation, however, is actually a function of eight variables (the w component of momentum is added) with gravity now included in the ∇ four-gradient operator

$$\boxed{\frac{\mathbb{P}}{m_i} \cdot \nabla \aleph_i + \mathbb{F}_i \cdot \nabla_{\mathbb{P}} \aleph_i = \dot{\aleph}_{i, \text{coll}}} \quad (9.2)$$

The four-momentum gradient operator is now $\nabla_{\mathbb{P}} = (\partial/\partial\mathbb{P}^w, \partial/\partial\mathbb{P}^x, \partial/\partial\mathbb{P}^y, \partial/\partial\mathbb{P}^z)$ and the four-force is now simply

$$\mathbb{F}_i = \frac{q_i}{c} \frac{\mathbb{P}}{m_i} \cdot \mathcal{F} \quad (9.3)$$

The distribution function $\aleph_i$ is defined in eight-dimensional phase space

$$\aleph_i = \aleph_i(\mathbb{X}, \mathbb{P})$$

(In a similar manner we have used the blackboard font to denote quantities in a given volume in eight-dimensional phase space ($\mathbb{X}$, $\mathbb{P}$, $\mathbb{U}$, $\mathbb{F}$) and bold roman and calligraphic fonts for average vectors and tensors in physical four-dimensional space.)

Note, however, that not all momenta are allowed – only those that satisfy the conservation of four-momentum magnitude (equation (6.53))

$$\mathbb{P}^2 = -m_i^2 c^2$$

In four-dimensional momentum space this creates a limited three-dimensional region, called the “mass hyperboloid” or the “mass shell”, where only certain momenta are allowed. In terms of the density f_i in three-momentum space, then, density $\aleph_i$ in four-dimensional momentum space is given by

$$\aleph_i = m_i c f_i \frac{\delta(\mathbb{P}^{\hat{0}} - m_i \gamma_i c)}{\mathbb{P}^{\hat{0}}} \quad (9.4)$$

where $\gamma_i \equiv (1 - \mathcal{V}_i^2/c^2)^{-1/2}$. So, quantities like the three-volume density of particles of species i can be expressed as integrals over three-space quantities

$$\mathbf{n}_i \equiv \int_{\Omega_{\mathbb{P}}} \gamma_i \mathfrak{N}_i d^4\mathbb{P} = \int_{\mathcal{T}_{\mathbb{P}}} f_i d^3\mathcal{P}$$

where $\Omega_{\mathbb{P}}$ and $\mathcal{T}_{\mathbb{P}}$ are the four-dimensional and three-dimensional momentum volumes, respectively.

The great advantage of kinetic theory is the ability to evolve the distribution of particle momenta at every point in space. However, current computers can barely cope with the evolution of the three-dimensional simulations; accurate evolution in 6-dimensional phase space is quite out of the question at the present point in time. Therefore, a simpler approach than even kinetic theory is needed in order to simulate the great majority of plasma flows near black holes.

9.1.5 General Relativistic (Electro-)Magnetohydrodynamics

The nearly impossible equations of general relativistic kinetic theory can be made reasonably tractable by performing two moment integrals (over momentum space) and several summations over particle species. The integrals over momentum space generate a system of “multi-fluid” equations for each particle species i ; these are functions of $\mathbb{X}$ and t only (see Section D.1 in Appendix D). The various summations over all particle species further reduce the description of plasma flow to that of a *single* neutral fluid, with an electromagnetic field and with an internal current generated by net charge drift within the fluid (see Section D.2). In the process, we re-discover the conservation of charge (equation (6.115)) and of energy-momentum (equation (7.15)). And we also learn how to express the stress-energy tensor in primitive variables like the mass density, internal energy, and pressure

$$\rho \equiv \sum_i \mathbf{n}_i m_i \quad \varepsilon \equiv \sum_i \varepsilon_i \quad p \equiv \sum_i p_i$$

(where $\mathbf{n}_i$ is the space density of particles of species i) as well as how to compute the partial internal energies and pressures for each particle species from ρ and the local temperature T . We also are able to construct relatively simple evolution equations for ρ , the fluid four-velocity $\mathbf{U}$, ε , and four-current $\mathbf{J}$ (charge density ρ_q and three-current $\mathbf{J}$). The full set of equations is called general relativistic (electro-)magnetohydrodynamics or just GRMHD.

While Appendix D shows how a basic set of GRMHD conservation laws can be derived from the general relativistic Boltzmann equation, some physical processes that require a good treatment of the collision terms (*e.g.*, viscosity) are ignored in that derivation. In the next section we present a more complete version of these equations, without derivation. It is this set that we will need to discuss the inner workings of black hole engines.

9.2 The Conservation Laws of Relativistic Magnetohydrodynamics

In a fully dynamical situation, the purpose of the conservation laws is to determine the three appropriate components of the current for the three appropriate electromagnetic field equations (6.126), and the six appropriate components of the stress-energy tensor for the six appropriate gravitational field equations (7.21). Then the field equations are used to determine how the field components evolve. In a situation with a stationary metric, as will be the case for black hole engines, the conservation laws of energy and momentum will be used only to see how the fluid flows through the spacetime – essentially a study in weather prediction – but still with the possibility of an evolving electromagnetic field. Whatever the situation, we need to produce a full set of equations that uniquely determine all four non-redundant components of $\mathbf{J}$ and all ten of $\mathcal{T}$ in order to accomplish the above tasks.

As a result, it turns out that there are a total of *four* conservation laws, not two; two of these result in scalar equations and two result in four-vector equations, for a total of ten conservation law equations. All four can be derived from the relativistic Boltzmann equation, and we show how this is done in Appendix D. The first two are the conservation of rest mass and the conservation of charge. Both are derived from the conservation of particle number for each species (the zeroth moment of the Boltzmann equation), and are determined by weighting and summing over particle mass or particle charge, respectively. The second two are the conservation of energy-momentum and the conservation of four-current (commonly known as Ohm's Law). These are both derived from the conservation of particle energy-momentum for each species (the first moment of the Boltzmann equation), and also are determined by weighting and summing over particle mass or charge.³ We begin our discussion of fundamental physics by outlining the four conservation laws and determining what parts of the components of $\mathbf{J}$ and $\mathcal{T}$ we know and do not know. After that we will delve deeper into the physics necessary to fully close the equations so we have as many equations as unknowns for the state variables (*i.e.*, density, pressure, temperature, *etc.*).

9.2.1 Laws of Conservation of Rest Mass and Energy-Momentum

In order to count equations and unknowns, we will assign each conservation law to a particular state variable. This is actually what is done in numerical simulations also, but in analytic studies the equations that determine certain variables can be shuffled

³ It is less remarkable than it would seem that the same conservation laws demanded by the gauge theories can be derived independently from kinetic theory. After all, both Maxwell's and Einstein's field theories are for continuous media, not individual particles. There are corresponding field theories on an individual particle level that reduce to these when integrated over a system of particles. It is at that point of averaging that the particle distribution comes in, and the field theories make contact with the Boltzmann equation.

around. Nevertheless, this is a useful exercise to help understand physically what is going on.

9.2.1.1 The Law of Conservation of Rest Mass: Determining Rest Mass Density

In writing the conservation laws, all references to the multiple species of particles that make up the fluid are usually dropped. For example, one often defines an average particle mass rest m_0

$$m_0 \equiv \frac{1}{n} \sum_i n_i m_i$$

so the density can be written as

$$\rho \equiv m_0 n \quad (9.5)$$

where $n = \sum_i n_i$ is the total particle number density in the fluid. The conservation law for ρ , then, is the relativistic version of the continuity equation (5.63)

$$\boxed{\nabla \cdot (\rho \mathbf{U}) = 0} \quad (9.6)$$

where $\mathbf{U}$ is the four-velocity of equations (6.2), (7.29), or (7.51). Of course, because m_0 is a constant, it could be divided out of equation (9.6), leaving $\nabla \cdot (n \mathbf{U}) = 0$; so the conservation of rest mass is simply the conservation of particle number only. It is *not* the conservation of *total* mass. That is actually the conservation of energy, which is a component of the next conservation law.

9.2.1.2 The Law of Conservation of Energy-Momentum: Determining Temperature and Components of the Three-Velocity

The stress-energy tensor has the nice property of being linear. In order to include a new set of physical forces, one simply adds the stress-energy for those processes to the current set. There are three major stress-energy components that we will need to study black hole engines, and we have already encountered parts of two of them, the ideal fluid and electromagnetic components, in equations (6.68) and (6.121). The other is the radiation stress-energy tensor. The total conservation law is

$$\boxed{\nabla \cdot (\mathcal{T}_{\text{GAS}} + \mathcal{T}_{\text{RAD}} + \mathcal{T}_{\text{EM}}) = 0} \quad (9.7)$$

where $\mathcal{T}_{\text{GAS}}$ also includes non-ideal fluid properties. This conservation law will determine only four state variables: the temperature (from the energy conservation part) and the three spatial components of the four-velocity. The time component of the four-velocity can be found from the normalization (see equation (6.53))

$$U^2 = -c^2$$

We will consider each of these three components in turn.

Dynamics of a Perfect Gas

The stress-energy tensor for a perfect gas/fluid in any general relativistic frame is (equation (6.68))

$$\mathcal{T}_{\text{FL}} = \left[\rho + \frac{(\varepsilon_g + p_g)}{c^2} \right] \mathbf{U} \mathbf{U} + p_g \mathbf{g}$$

where $\mathbf{g}$ will be one of the metrics discussed in Chapter 7. In addition to $\mathbf{U}$, this introduces two new variables, the internal energy ε_g and the pressure p_g , which must be computed from algebraic relations called the equations of state

$$\boxed{\varepsilon_g = \varepsilon_g(\rho, T) \quad p_g = p_g(\rho, T)} \quad (9.8)$$

Because astrophysical fluids are almost always in the gaseous or plasma phase, we will use the subscript g to designate the fluid properties. As discussed in Section 5.2.2, ε_g and p_g in the local reference frame are often related by a proportionality constant n

$$\varepsilon_g = n p_g$$

resulting in a polytropic equation of state (5.55)

$$p_g = K_n \rho^{\frac{(n+1)}{n}} \quad (9.9)$$

Note that we have identified two kinds of equations that can determine the state variables: differential conservation laws and algebraic equations of state. The latter are self-contained and generally defined in the local reference frame of the gas. The differential nature of the former, however, means that they must be considered throughout space and time and requires that we supply boundary and initial conditions as well.

Heat Conduction and Viscosity: Deviations from a Perfect Gas

The above stress-energy tensor is sufficient to describe the fluid or gas as long as the mean free path of particles in the fluid is very short compared to the distance over which thermal and kinetic properties of the fluid change. However, if hot particles can travel long distances and deposit their heat in a cooler region of the fluid, then we must take this *heat conduction* into account by adding another component of $\mathcal{T}$

$$\mathcal{T}_{\text{COND}} = \frac{1}{c^2} [\mathbf{Q}_g \mathbf{U} + \mathbf{U} \mathbf{Q}_g] \quad (9.10)$$

where

$$\mathbf{Q}_g = -K_c [c^2 \mathcal{P} \cdot \nabla T + T \mathbf{U} \cdot \nabla \mathbf{U}] \quad (9.11)$$

is the heat conduction vector for the fluid [305]. The constant

$$K_c = K_c(\rho, T) \quad (9.12)$$

is potentially another equation of state, and we introduce a new tensor

$$\mathcal{P} = \frac{1}{c^2} \mathbf{U} \mathbf{U} + g \quad (9.13)$$

called the “projection tensor”. When $\mathcal{P}$ is dotted with another vector or tensor, it extracts only that part that is orthogonal to the four-velocity $\mathbf{U}$. That is, it extracts from that tensor the spatial part in a reference frame traveling with the four-velocity $\mathbf{U}$. Therefore, $\mathbf{Q}_g$ is a space-like vector for an observer moving with the velocity $\mathbf{U}$, because its first term is orthogonal to $\mathbf{U}$, and the second term (proportional to the acceleration $\mathbf{U} \cdot \nabla \mathbf{U}$) also is orthogonal to the four-velocity. In the rest frame of the fluid we have $\mathbf{U} = (c, 0, 0, 0)$, and the acceleration is zero. So $\mathbf{Q}_g$ has only three spatial components in that frame, and they are the classical heat flux

$$\mathbf{Q}_g = -K_c \nabla T$$

where we use the usual notation that the italic $\mathbf{Q}$ represents the three-vector heat flux and the roman $\mathbf{Q}$ the four-vector version. So, in the rest frame of the fluid, the stress-energy tensor has the matrix form

$$\mathcal{T}_{\text{COND}} = \frac{1}{c} \begin{pmatrix} 0 & Q_g^x & Q_g^y & Q_g^z \\ Q_g^x & 0 & 0 & 0 \\ Q_g^y & 0 & 0 & 0 \\ Q_g^z & 0 & 0 & 0 \end{pmatrix} \quad (9.14)$$

There are *no* energy or stress terms, only change in momentum density/heat flux terms.

Another related process that arises because of long particle mean free paths is viscosity; this transports momentum rather than energy. Two kinds of viscosity are recognized: shear and bulk. Shear viscosity transports momentum perpendicular to the fluid flow, and bulk viscosity does so parallel to the flow. The corresponding stress-energy component is

$$\mathcal{T}_{\text{VISC}} = -2\eta_{v,g} \Sigma - \zeta_{v,g} \Theta \mathcal{P} \quad (9.15)$$

The state functions

$$\eta_{v,g} = \eta_{v,g}(\rho, T) \quad \zeta_{v,g} = \zeta_{v,g}(\rho, T)$$

are the coefficients of shear and bulk viscosity, respectively. And Σ and Θ are the shear tensor and compression rate of the fluid

$$\begin{aligned} \Sigma &\equiv \frac{1}{2} [\mathcal{P} \cdot \nabla \mathbf{U} + (\mathcal{P} \cdot \nabla \mathbf{U})] - \frac{1}{3} \Theta \mathcal{P} \\ \Theta &\equiv \nabla \cdot \mathbf{U} \end{aligned}$$

The fact that viscosity transports only momentum does not mean that the energy content of the fluid remains unaffected. Indeed, the viscous terms cause a local “viscous heating” of the fluid

$$\rho \frac{dq}{d\tau} = 2\eta_{v,g} \Sigma : \Sigma + \zeta_{v,g} \Theta^2 \quad (9.16)$$

This heating can be understood in a very simple way. In the frame of the moving fluid, viscosity re-distributes *only* the three spatial components of the four-momentum, generally slowing down the flow in some region. The viscous component of the stress-energy tensor in this frame is

$$\mathcal{T}_{\text{VISC}} =$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2\eta_{v,g} \Sigma^{xx} - \zeta_{v,g} \Theta & -2\eta_{v,g} \Sigma^{xy} & -2\eta_{v,g} \Sigma^{xz} \\ 0 & -2\eta_{v,g} \Sigma^{yx} & -2\eta_{v,g} \Sigma^{yy} - \zeta_{v,g} \Theta & -2\eta_{v,g} \Sigma^{yz} \\ 0 & -2\eta_{v,g} \Sigma^{zx} & -2\eta_{v,g} \Sigma^{zy} & -2\eta_{v,g} \Sigma^{zz} - \zeta_{v,g} \Theta \end{pmatrix} \quad (9.17)$$

(Compare this with the stress-energy tensor for a *perfect* fluid in the same frame; equation (6.69).) While momentum is conserved, kinetic energy is not; there are no energy terms in the $\mathcal{T}^{ww}$ position above. But total energy-momentum of the fluid *must* be conserved (equation (9.7)), so there must be a transfer of energy from kinetic to heat.

Consider, for example, the *bulk* viscous process. It is very different from inviscid adiabatic compression of a perfect gas. Both also can slow down the flow. In the adiabatic case, the compression stores kinetic energy in internal energy (ε and p) of the fluid (equation (6.69)), leaving its heat content unchanged. This adiabatic compression often is followed by an adiabatic expansion, which then recovers all of the original kinetic energy of the flow. On the other hand, when $\Theta < 0$ in equation (9.17), the compression acts like a pressure also slowing the flow, but there is no corresponding Θ term in the $\mathcal{T}^{ww}$ position to store the kinetic energy loss. The

subsequent re-expansion of the flow does not recover the kinetic energy lost, so that energy goes into heat. A similar effect takes place with the shear viscosity.

Note that the heat conduction and viscous components of $\mathcal{T}$ cannot exist on their own. They are perturbations on the basic perfect fluid component and must be used in conjunction with the latter. Together the three properly account for the evolution of internal energy, heat, and momentum of the fluid, even if there are conductive and viscous processes acting in the fluid

$$\begin{aligned} \mathcal{T}_{\text{GAS}} \equiv \mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{COND}} + \mathcal{T}_{\text{VISC}} &= \left(\rho + \frac{\varepsilon_{\text{g}}}{c^2} \right) \mathbf{U} \mathbf{U} \\ &- 2\eta_{\text{v,g}} \boldsymbol{\Sigma} + (p_{\text{g}} - \zeta_{\text{v,g}} \Theta) \mathcal{P} + \frac{1}{c^2} [\mathbf{Q}_{\text{g}} \mathbf{U} + \mathbf{U} \mathbf{Q}_{\text{g}}] \end{aligned} \quad (9.18)$$

In the rest frame of the fluid, this total gas stress-energy tensor in matrix form is

$$\mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{COND}} + \mathcal{T}_{\text{VISC}} = \begin{pmatrix} \rho c^2 + \varepsilon_{\text{g}} & Q_{\text{g}}^x & & \\ Q_{\text{g}}^x & -2\eta_{\text{v,g}} \Sigma^{xx} - \zeta_{\text{v,g}} \Theta + p_{\text{g}} & & \\ Q_{\text{g}}^y & -2\eta_{\text{v,g}} \Sigma^{yx} & & \\ Q_{\text{g}}^z & -2\eta_{\text{v,g}} \Sigma^{zx} & & \\ & Q_{\text{g}}^y & Q_{\text{g}}^z & \\ -2\eta_{\text{v,g}} \Sigma^{xy} & & -2\eta_{\text{v,g}} \Sigma^{xz} & \\ -2\eta_{\text{v,g}} \Sigma^{yy} - \zeta_{\text{v,g}} \Theta + p_{\text{g}} & & -2\eta_{\text{v,g}} \Sigma^{yz} & \\ -2\eta_{\text{v,g}} \Sigma^{zy} & & -2\eta_{\text{v,g}} \Sigma^{zz} - \zeta_{\text{v,g}} \Theta + p_{\text{g}} & \end{pmatrix} \quad (9.19)$$

Radiation Dynamics

In many situations that we will study in the next few chapters, the fluid will be optically thick to radiation and both will be in thermodynamic equilibrium at the same temperature (*i.e.*, $T_{\text{r}} = T_{\text{g}} \equiv T$). In this case the photon gas will contribute to the fluid plasma pressure, energy density, heat conduction, and viscosity and will add stress-energy terms similar to those discussed in the above few paragraphs.

$$\begin{aligned} \rho &= \rho_{\text{g}} \\ p &= p_{\text{g}} + p_{\text{r}} \quad \varepsilon = \varepsilon_{\text{g}} + \varepsilon_{\text{r}} \\ \mathbf{Q} &= \mathbf{Q}_{\text{g}} + \mathbf{Q}_{\text{r}} \\ \eta_{\text{v}} &= \eta_{\text{v,g}} + \eta_{\text{v,r}} \quad \zeta_{\text{v}} = \zeta_{\text{v,g}} + \zeta_{\text{v,r}} \end{aligned}$$

Note that, while the photons do not contribute any *rest* mass, they will contribute inertia through the enthalpy ($h = \varepsilon + p$). Also, the radiation pressure and energy density are usually related as $\varepsilon_{\text{r}} = 3p_{\text{r}}$ and are only functions of temperature (see

Section 9.3.1). So the radiation part of the plasma acts as an $n = 3$ polytrope, or $\Gamma = 4/3$ gas.

In the case of radiation, the heat flux usually is computed from the radiative pressure and enthalpy, rather than the temperature (even though the former are functions of temperature only). In situations where the photon opacity is a function of frequency ν , $\partial \mathbf{Q}_r / \partial \nu$ also is and must be integrated over frequency to obtain the correct photon heat flux

$$\mathbf{Q}_r = \int_0^\infty \frac{\partial \mathbf{Q}_r(\nu)}{\partial \nu} d\nu$$

where

$$\frac{\partial \mathbf{Q}_r(\nu)}{\partial \nu} = -\frac{1}{\kappa(\nu) \rho c} \left\{ c^2 \mathcal{P} \cdot \nabla p_r(\nu) + [\varepsilon_r(\nu) + p_r(\nu)] \mathbf{U} \cdot \nabla \mathbf{U} \right\}$$

However, if we define two average inverse opacities of the plasma, which have units of area per unit mass,

$$\frac{1}{\bar{\kappa}_R} = \frac{\int_0^\infty \frac{1}{\kappa(\nu)} \frac{\partial p_r(\nu)}{\partial T} d\nu}{\int_0^\infty \frac{\partial p_r(\nu)}{\partial T} d\nu} \quad (9.20)$$

and

$$\frac{1}{\bar{\kappa}'_R} = \frac{\int_0^\infty \frac{1}{\kappa(\nu)} [\varepsilon_r(\nu) + p_r(\nu)] d\nu}{\int_0^\infty [\varepsilon_r(\nu) + p_r(\nu)] d\nu}$$

we can write the heat flux in a more compact form

$$\boxed{\mathbf{Q}_r = -\frac{c}{\bar{\kappa}_R \rho} [\mathcal{P} \cdot \nabla p_r] - \frac{1}{\bar{\kappa}'_R \rho c} [(\varepsilon_r + p_r) \mathbf{U} \cdot \nabla \mathbf{U}]} \quad (9.21)$$

where ε_r and p_r are the frequency-integrated radiation internal energy and pressure

$$\varepsilon_r = \int_0^\infty \varepsilon_r(\nu) d\nu \quad p_r = \int_0^\infty p_r(\nu) d\nu$$

The first opacity, $\bar{\kappa}_R$, is called the Rosseland mean opacity, and is used extensively in stellar interior investigations. The use of the temperature gradient in the kernel of the Rosseland mean comes from knowing that the photon energy density (per unit frequency or integrated) is a function only of temperature T and not density ρ . So, the gradient of p_r in equation (9.21) can be reduced simply to a gradient of T , making the derivative of $\varepsilon_r(\nu)$ with temperature of primary importance in the

average. The second opacity, $\bar{\kappa}'_R$ is similar to the Rosseland mean, without the temperature derivative, but is not used so extensively. It is needed only in relativistic situations where the second term in equation (9.21) is important.

The total radiation part of the stress-energy tensor (similar to equation (9.18)) is

$$\mathcal{T}_{\text{RAD}} = \frac{\varepsilon_r}{c^2} \mathbf{U} \mathbf{U} - 2\eta_{v,r} \boldsymbol{\Sigma} + (p_r - \zeta_{v,r} \Theta) \mathcal{P} + \frac{1}{c^2} [\mathbf{Q}_r \mathbf{U} + \mathbf{U} \mathbf{Q}_r] \quad (9.22)$$

In the rest frame of the fluid this becomes, in matrix form,

$$\mathcal{T}_{\text{RAD}} = \begin{pmatrix} \varepsilon_r & Q_r^x & & \\ Q_r^x & -2\eta_{v,r} \Sigma^{xx} - \zeta_{v,r} \Theta + p_r & & \\ Q_r^y & -2\eta_{v,r} \Sigma^{yx} & & \\ Q_r^z & -2\eta_{v,r} \Sigma^{zx} & & \\ & Q_r^y & Q_r^z & \\ -2\eta_{v,r} \Sigma^{xy} & -2\eta_{v,r} \Sigma^{xz} & & \\ -2\eta_{v,r} \Sigma^{yy} - \zeta_{v,r} \Theta + p_r & -2\eta_{v,r} \Sigma^{yz} & & \\ -2\eta_{v,r} \Sigma^{zy} & -2\eta_{v,r} \Sigma^{zz} - \zeta_{v,r} \Theta + p_r & & \end{pmatrix} \quad (9.23)$$

(compare with equation (9.19)) with the radiative heat flux in the fluid rest frame

$$\mathbf{Q}_r = -\frac{c}{3\bar{\kappa}_R \rho} \nabla \varepsilon_r \quad (9.24)$$

Electrodynamics

Because the electromagnetic field exerts forces on the plasma, it also must be taken into account in the conservation of energy and momentum. The stress-energy tensor term for the electromagnetic field was introduced for flat space in Chapter 6, equation (6.119). In the curved spacetime of a black hole it has the same form

$$\mathcal{T}_{\text{EM}} = \frac{1}{4\pi} [\mathcal{F} \cdot \mathbf{f} - \frac{1}{4} (\mathcal{F} : \mathbf{f}) \mathcal{I}] \cdot g^{-1}$$

where g is now the black hole metric, and the Faraday tensor and its 2-form are measured in the BL (Boyer–Lindquist) frame. However, the components of these tensors are often expressed in quantities from other frames, as pointed out in Section 7.7.2. That is, $\mathbf{E}$ and $\mathbf{H}$ are measured in the FIX frame, while $\mathbf{D}$ and $\mathbf{B}$ are measured in the OIS frame.

However, it is still very instructive to examine this tensor in the MOV frame (the frame of the fluid) and compare with the other components of $\mathcal{T}$ above that we also expressed in this frame:

$$\mathcal{T}_{\text{EM}} = \begin{pmatrix} \varepsilon_{\text{em}} & Q_{\text{em}}^x & -\frac{1}{4\pi}(E^{\hat{x}2} + B^{\hat{x}2}) + p_{\text{em}} & \\ Q_{\text{em}}^x & -\frac{1}{4\pi}(E^{\hat{y}}E^{\hat{x}} + B^{\hat{y}}B^{\hat{x}}) & -\frac{1}{4\pi}(E^{\hat{x}}E^{\hat{z}} + B^{\hat{x}}B^{\hat{z}}) & \\ Q_{\text{em}}^y & -\frac{1}{4\pi}(E^{\hat{y}}E^{\hat{x}} + B^{\hat{y}}B^{\hat{x}}) & -\frac{1}{4\pi}(E^{\hat{y}}E^{\hat{z}} + B^{\hat{y}}B^{\hat{z}}) & \\ Q_{\text{em}}^z & -\frac{1}{4\pi}(E^{\hat{x}}E^{\hat{z}} + B^{\hat{z}}B^{\hat{x}}) & -\frac{1}{4\pi}(E^{\hat{z}2} + B^{\hat{z}2}) + p_{\text{em}} & \end{pmatrix} \quad (9.25)$$

where the thermodynamic quantities

$$\varepsilon_{\text{em}} = p_{\text{em}} = \frac{1}{8\pi} (E^2 + B^2) \quad (9.26)$$

and the energy flux three-vector is the negative Poynting flux

$$\mathbf{Q}_{\text{em}} = -\mathbf{S} \equiv \frac{1}{4\pi} (\mathbf{E} \times \mathbf{B}) \quad (9.27)$$

The remaining terms are a viscous-like spatial tensor that is often referred to as the electromagnetic tension. Like viscosity, it transports momentum in a space-like direction only and, therefore, is responsible for the Joule heating losses by the electromagnetic field. Note that, since $\varepsilon_{\text{em}} = p_{\text{em}}$, the electromagnetic field acts like an $n = 1$ polytrope, or a $\Gamma = 2$ gas.

9.2.2 Laws of Conservation of Charge and Current

The conservation laws of charge and current are sometimes called “charge dynamics”, analogous to fluid dynamics. There is a one-to-one analogy between the conservation of rest mass and conservation of charge, and between the conservation of energy-momentum and conservation of current. The latter is often referred to a Ohm’s law.

9.2.2.1 The Law of Conservation of Charge: Determining Charge Density

The strict definition of charge density is

$$\rho_q \equiv \sum_i n_i q_i \quad (9.28)$$

where q_i is the charge of each particle of that species. Unlike the rest mass density, ρ_q can be zero everywhere for a neutral plasma, although on a microscopic level there can be charge separation and plasma charge oscillations (Langmuir waves).

The continuity equation for charge dynamics is equation (6.115)

$$\boxed{\nabla \cdot \mathbf{J} = 0} \quad (9.29)$$

regardless of reference frame, in curved and flat space alike. In the frame of the fluid, the four-current is given by equation (6.111). However, in a general frame, the four-current can be written as

$$\mathbf{J} = \rho_q \mathbf{U} + \mathfrak{J} \quad (9.30)$$

where ρ_q is the charge density in the fluid frame (“rest charge”) and

$$\mathfrak{J} \equiv \mathcal{P} \cdot \mathbf{J} \quad (9.31)$$

is the “spatial current”. $\mathfrak{J}$ is still a four-vector, but it is constrained to be orthogonal to $\mathbf{U}$

$$\mathbf{U} \cdot \mathfrak{J} = 0 \quad (9.32)$$

The conservation of charge determines only *one* component of $\mathbf{J}$ – the charge density ρ_q . In order to determine the other three components (*i.e.*, the spatial current), we will need three more equations.

9.2.2.2 The Law of Conservation of Beamed Current: Determining Components of the Three-Current

The charge dynamical counterpart to the conservation of energy-momentum is the conservation of beamed current or, as it is more popularly known, Ohm’s law. This law has several forms; we will begin with the fully time-dependent, general form and then specialize it to simpler and more familiar expressions.

The Generalized, Fully Relativistic Ohm’s Law

The version of Ohm’s law that includes the same physics as $\mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{COND}} + \mathcal{T}_{\text{RAD}} + \mathcal{T}_{\text{EM}}$, but for charge (rather than matter) dynamics, is

$$\boxed{\nabla \cdot \mathcal{C} = \frac{\omega_p^2}{4\pi} \left[\frac{1}{c} (\mathbf{U} + h_q \mathfrak{J}) \cdot \mathcal{F} - \eta_q (\rho_q \mathbf{U} + \mathfrak{J}) \right]} \quad (9.33)$$

where the charge-current tensor is given by

$$\mathcal{C} \equiv \left(\rho_q + \frac{\varepsilon_q + p_q}{c^2} \right) \mathbf{U} \mathbf{U} + \mathbf{U} \mathfrak{J}' + \mathfrak{J}' \mathbf{U} + p_q g$$

The above conservation law also is derived in Appendix D. In the rest frame of the fluid, the charge-current tensor has the form

$$\mathcal{C} = \begin{pmatrix} (\rho_q c^2 + \varepsilon_q) \mathfrak{J}'^x & \mathfrak{J}'^y & \mathfrak{J}'^z \\ \mathfrak{J}'^x & p_q & 0 & 0 \\ \mathfrak{J}'^y & 0 & p_q & 0 \\ \mathfrak{J}'^z & 0 & 0 & p_q \end{pmatrix} \quad (9.34)$$

This has the same structure as $\mathcal{T}_{\text{FL}} + \mathcal{T}_{\text{COND}}$. In addition, the physics associated with Lorentz force ($\mathcal{T}_{\text{EM}}$) also is included in the first term on the right-hand side of equation (9.33), *and* even $\mathcal{T}_{\text{RAD}}$ is implicitly included, since the direct contribution of photon particles to charge dynamics is zero, as they have zero charge. The conservation of beamed current has all the same physics as the conservation of energy-momentum, *except* for the spatial viscous terms (*i.e.*, $\mathcal{T}_{\text{VISC}}$). This is because the contribution of viscous collisional effects were not considered in the derivation of equation (D.14), not because such viscous effects on charge dynamics do not exist.

In addition to the charge density ρ_q , there are a number of new additional thermodynamic quantities that are introduced here, including the charge-weighted internal energy and pressure

$$\varepsilon_q \equiv \sum_i \frac{q_i}{m_i} \varepsilon_i \quad p_q \equiv \sum_i \frac{q_i}{m_i} p_i$$

with ε_i and p_i being the partial energy density and pressure for each particle species i . The “beamed spatial current” $\mathfrak{J}'$ is a Lorentz-enhanced version of the spatial current $\mathfrak{J}$. Both are given by summing charge drift over all particle species

$$\begin{aligned} \mathfrak{J}' &\equiv \sum_i q_i \int \gamma \mathbf{v} f(\mathbf{v}) d^3\mathbf{v} \\ \mathfrak{J} &\equiv \sum_i q_i \int \mathbf{v} f(\mathbf{v}) d^3\mathbf{v} \end{aligned}$$

where $f(\mathbf{v})$ is the distribution function of (three-vector) particle velocity $\mathbf{v}$. Note that the beamed spatial current has an extra particle Lorentz factor in the integral.

In order to “close” the equations (*i.e.*, not end up with more unknowns than equations and have to generate even higher-order moments of the relativistic Boltzmann equation), we will make the approximation that $\mathfrak{J}$ and $\mathfrak{J}'$ are parallel and differ only by a single average Lorentz current beaming factor

$$\boxed{\mathfrak{J}' = \gamma_q \mathfrak{J}} \quad (9.35)$$

Equation (9.33) then consists of four equations: the three spatial components of that equation determine the three spatial components of $\mathfrak{J}'$, and the time component of equation (9.33) that determines the current beaming factor γ_q . The fourth (time)

component of $\mathfrak{J}'$ is found from the orthogonality condition

$$\boxed{\mathfrak{J}' \cdot \mathbf{U} = 0} \quad (9.36)$$

which states that in the fluid frame $\mathfrak{J}'$ is a three-vector.

The sources and sinks of the beamed spatial current on the right-hand side of equation (9.33) are, respectively, the Lorentz effect ($\propto \mathbf{U} \cdot \mathcal{F}/c$), the Hall effect, and the resistive losses ($\propto \eta_q \mathbf{J}$). Their coefficients are the square of the plasma frequency

$$\omega_p^2 \equiv 4\pi \sum_i \frac{q_i^2 n_i}{m_i} \approx 4\pi \frac{n_e e^2}{m_e}$$

the Hall coefficient

$$h_q \equiv \frac{4\pi}{\omega_p^2 |\mathfrak{J}|} \sum_i \frac{q_i}{m_i} |\mathfrak{J}_i| \approx \frac{1}{n_e e}$$

and the resistivity

$$\eta_q \equiv 4\pi \frac{\nu_{\text{coll}}}{\omega_p^2}$$

where ν_{coll} is the particle collision frequency.

As these equations were derived only recently (by D. Meier [346]), most readers probably are not familiar with the fully-covariant, relativistic and time-dependent form of Ohm's law. Therefore, a little additional explanation will be given. The primary purpose of charge dynamics is to determine the components of the four-current $\mathbf{J}$ (i.e., ρ_q and $\mathfrak{J}$). In its full form, that would be done by solving equation (9.33) for the evolution of γ_q and $\mathfrak{J}'$, then computing $\mathfrak{J}$ from equation (9.35), and finally solving equation (9.29) for the evolution of ρ_q . That determines $\mathbf{J}$ for one time step. In the next time step, both ρ_q and $\mathfrak{J}$ will be needed in the right-hand side of equation (9.33) and the update of $\mathbf{J}$ begins again. Current and charge therefore are evolved in a manner just like momentum and mass density, using partial differential equations that are differentiated in space and integrated in time.

While the other quantities (ε_q , p_q , ω_p , h_q , and η_q) also must be determined, they can be found through algebraic equations of state. That is,

$$\varepsilon_q = \varepsilon_q(\rho, T) \quad p_q = p_q(\rho, T) \quad (9.37)$$

$$\omega_p = \omega_p(\rho, T) \quad h_q = h_q(\rho, T) \quad \eta_q = \eta_q(\rho, T) \quad (9.38)$$

These will be discussed more fully in the next section. To date, no fully relativistic simulations of the equations of charge dynamics have been performed. However, the terms in the tensor $\mathcal{C}$ may be very important for processes like reconnection and current sheet structure, where rapidly changing velocity and electromagnetic fields can affect current flow on short time and length scales.

One final comment on charge dynamics is useful, before we move on to simpler versions of equation (9.33). The *non*-relativistic version of that equation *is* well known and is often referred to as the “generalized Ohm’s law” (see, *e.g.*, [350])

$$\frac{\partial \mathbf{J}}{\partial t} + \nabla \cdot (\mathbf{V} \mathbf{J} + \mathbf{J} \mathbf{V} - \rho_q \mathbf{V} \mathbf{V}) + \nabla p_q = \frac{\omega_p^2}{4\pi} \left(\mathbf{E} + \frac{\mathbf{V}}{c} \times \mathbf{B} + h_q \frac{\mathbf{J}}{c} \times \mathbf{B} - \eta_q \mathbf{J} \right)$$

This is the non-relativistic limit of equation (9.33) and can be obtained by assuming a flat metric ($\mathbf{g} = \boldsymbol{\eta}$), slow bulk speed ($|\mathbf{V}| \ll c$), and slow particle thermal and drift velocities ($(\varepsilon_q + p_q) \ll \rho_q c^2$ and $\gamma_q = 1$). Note that $\mathbf{J}$ is a three-vector, representing the three spatial components of $\mathfrak{J}$.

The Static Ohm’s Law: Relativistic Hall MHD

For most macroscopic astrophysical situations, the time scale on which the evolution of $\mathcal{C}$ is so short ($2\pi/\omega_p$) that the left-hand side of equation (9.33) can be ignored. That is, $\omega_p^2/4\pi$ is so large that a steady state will hold on non-plasma time and length scales. We then can set the right-hand side to zero, yielding

$$(\mathbf{U} + h_q \mathfrak{J}) \cdot \frac{\mathcal{F}}{c} = \eta_q (\rho_q \mathbf{U} + \mathfrak{J}) \quad (9.39)$$

With this simplification, an interesting transformation of Ohm’s law has taken place. Not only has all time-dependence been eliminated, but also all terms in $\mathfrak{J}'$ also have disappeared, leaving only terms in the non-beamed spatial current $\mathfrak{J}$. Therefore, we no longer need the factor γ_q , so we can eliminate the time component of equation (9.33). The remaining spatial part, which will give us the three components of the spatial current $\mathfrak{J}$, can be projected out from equation (9.39) using the projection tensor $\mathcal{P}$. This gives us the following equations

$$\boxed{\frac{\mathbf{U}}{c} \cdot \mathcal{F} + h_q \frac{\mathfrak{J}}{c} \cdot \mathcal{F} \cdot \mathcal{P} = \eta_q \mathfrak{J}} \quad (9.40)$$

with the constraint (9.36) still giving the time component of $\mathfrak{J}$. Ohm’s law has now become an algebraic equation for the current $\mathfrak{J}$, which still must be used in conjunction with equation (9.29) to obtain ρ_q and then both plugged into Maxwell’s equations. In the rest frame of the fluid, equation (9.40) becomes (using 3+1 notation)

$$\mathbf{E} + h_q \frac{\mathbf{J}}{c} \times \mathbf{B} = \eta_q \mathbf{J}$$

Plasma studies that use the approximation $\omega_p^2 \rightarrow \infty$, but still retain the Hall effect term ($h_q \neq 0$) are called “Hall magnetohydrodynamics” or Hall MHD.

Relativistic Resistive Magnetohydrodynamics

A further often-used simplification of the problem is to ignore the Hall term in the above static Ohm's law. This is valid when

$$\frac{h_q |\mathfrak{J}| |\mathbf{B}| / c}{\eta_q |\mathfrak{J}|} \approx \frac{e B}{m_e c \nu_{\text{coll}}} = \frac{\nu_L}{\nu_{\text{coll}}} \ll 1$$

That is, when the Larmor frequency is low compared to the collision frequency, the Hall term will not be important. In this case, the many collisions that occur during a single Larmor orbit of the electron about the magnetic field affect the current much more than the current generated by the $\mathbf{J} \times \mathbf{B}$ force. In this case, equation (9.40) reduces to

$$\boxed{\frac{\mathbf{U}}{c} \cdot \mathcal{F} = \eta_q \mathfrak{J}} \quad (9.41)$$

Use of equation (9.41) to determine the three spatial components of $\mathfrak{J}$, along with the other conservation and Maxwell's equations, is called “resistive MHD”. In the fluid frame equation (9.41) reduces to $\mathbf{E} = \mathbf{J} \eta_q$ or, as we usually remember it, $V = IR$, where V is the voltage applied across a resistance of strength R , and I is the current that results (not current density $\mathbf{J}$).

Relativistic Ideal Magnetohydrodynamics: “Frozen-in” Magnetic Fields

One additional simplification is often used in astrophysics because the conductivity of astrophysical plasmas is usually very high. The conductivity can be estimated as follows:

$$\begin{aligned} \sigma_q &\equiv \frac{1}{\eta_q} = \frac{\omega_p^2}{4\pi \nu_{\text{coll}}} \approx \left(\frac{n_e e^2}{m_e} \right) \frac{1}{n_e (\pi r_e^2) V_{\text{th}}} \\ &\approx \frac{(2 k T)^{3/2}}{\pi e^2 m_e^{1/2}} \sim 2.1 \times 10^8 \text{ s}^{-1} T^{3/2} \end{aligned}$$

where $r_e = 2e^2/(m_e V_{\text{th}}^2)$ is the Coulomb electron radius, and $V_{\text{th}} = (kT/m_e)^{1/2}$ is the thermal velocity of the electrons in the plasma. For temperatures between 10^6 K and 10^{10} K, the conductivity of astrophysical plasmas is between 10^{17} and 10^{23} inverse seconds (in cgs units). Compare this to pure silver or copper (the most conductive of the terrestrial metals), which have conductivities of order $5 \times 10^{16} \text{ s}^{-1}$. Astrophysical plasmas, therefore, are at least as conductive as the best metals on earth and often much more so. Given the extreme lack of electrical resistance in astrophysical plasmas, one can safely ignore the resistivity term in equation (9.41), leading to the simple “ideal” Ohm's law expression

$$\boxed{\frac{1}{c} \mathbf{U} \cdot \mathcal{F} = 0} \quad (9.42)$$

or $\mathbf{E} = 0$ in the fluid frame. That is, the conductivity is so high that the electric field in the frame of the fluid is shorted out (essentially zero). An electric field can be measured only in a frame that does not follow the fluid, and that field is simply the emf associated with the moving magnetized fluid. In the non-relativistic limit, equation (9.42) in the fixed “laboratory” frame becomes

$$\mathbf{E} = -\frac{\mathbf{V}}{c} \times \mathbf{B}$$

which is the familiar ideal Ohm’s law expression. Simulations that use equation (9.42) along with the other conservation laws are called “ideal MHD” simulations.

Ideal MHD has quite a different character, both mathematically and physically, from the other styles of magnetohydrodynamics discussed above. Ohm’s law in the ideal case is *not* used to determine the current density $\mathfrak{J}$ directly. Instead, it is used to determine the electric field components in terms of the fluid velocity and *magnetic* field; these then are used in the other conservation laws and Maxwell’s equations. The current is determined, then, from the curl of the electric field in Faraday’s law.

Physically the ideal MHD assumption freezes the magnetic field into the fluid flow, forcing the former to advect with the fluid. Ideal MHD never displays magnetic reconnection and field restructuring. However, because *numerical* simulations always have some artificial numerical resistivity, due to the finite size of the grid spacing, they never can be truly ideal. In fact, usually this numerical resistivity is much greater than the true physical resistivity of plasmas in black hole accretion flows. So the artificial resistivity dominates the simulation, unless a larger physical η_q is assumed and equation (9.41) is used instead of (9.42).

To date, most astrophysical simulations use some form of the ideal MHD approximation (relativistic or not) and, if resistivity and reconnection are needed, these simulations use the artificial resistivity inherent in the numerical method rather than assuming a specific η_q . However, resistive MHD (using equation (9.41)) is much more common in solar plasma physics studies, but even in that case an anomalously large η_q is assumed rather than the 10^{-17} – 10^{-23} s values estimated above. However, those few astrophysical simulations that do use charged particle dynamics do not have to use an anomalous resistivity; simulating particle interactions directly appears to produce the proper physics, including reconnection, that occurs in astrophysical plasmas.

9.3 The Equations of State

Having presented the field equations and their conservation laws, our next task will be to discuss the equations of state that are needed to express certain thermodynamic quantities in those conservation laws. All of these will be algebraic functions of the

rest mass density ρ , temperature T , and in some cases the plasma composition. The composition of astrophysical gases is usually expressed in terms of the mass fractions of hydrogen X , helium Y , and other heavier elements (often called “metals”) Z . Unless the gas is exotic (*e.g.*, electron–positron), the sum of these quantities is normalized to unity

$$X + Y + Z = 1$$

For solar abundances, $X = 0.71$, $Y = 0.27$, and $Z = 0.02$. From the standard model of the early universe, the primordial abundances right after the big bang and before any stars formed are computed to be $X = 0.75$, $Y = 0.25$, and $Z = 4 \times 10^{-10}$. Except for the trace amount of lithium produced in the big bang, all elements heavier than helium were produced exclusively in the centers of stellar furnaces (see Section 10.1).

9.3.1 Pressure and Internal Energy of Thermal Gases

For any gas that can be characterized by a temperature, the particles that make up that gas are distributed in three-momentum (magnitude) $\mathcal{P}$ according to the following general distribution function

$$\frac{dn}{d\mathcal{P}} = \frac{g_s}{h^3} \frac{4\pi \mathcal{P}^2}{\exp \{[\mathcal{E}(\mathcal{P}) - \mu_{\text{chem}}] / kT\} \pm 1} \quad (9.43)$$

where $h = 6.62607 \times 10^{-27}$ erg s is Planck’s constant, g_s is an integer degeneracy factor that depends on the quantum spin of the particle,

$$\mathcal{E} = (\mathcal{P}^2 c^2 + m_0^2 c^4)^{1/2} \quad (9.44)$$

is the total energy per particle with m_0 its rest mass (see also equation (6.64) and Section 6.5.1.3), and μ_{chem} is the particle’s chemical potential that was introduced in Section 5.2.2. In relativistic situations, which we will encounter in this chapter, the rest mass (nuclear potential) energy should be included in μ_{chem} , in addition to the chemical potential energy.

The plus sign in equation (9.43) is used for half-spin particles, *i.e.* “fermions” (electrons, protons, neutrinos, *etc.*), while the minus sign is used for integer spin particles, *i.e.*, “bosons” (photons, gravitons, gluons, and other particles that carry the forces of nature).

From the distribution in equation (9.43) we can compute the total internal (kinetic) energy per unit volume for a given particle species i

$$\varepsilon_i = \int \mathcal{E}_K(\mathcal{P}) \frac{dn_i}{d\mathcal{P}} d\mathcal{P} \quad (9.45)$$

where $\mathcal{E}_K = \mathcal{E} - m_0 c^2$ and the corresponding pressure

$$p_i = \frac{1}{3} \int \mathcal{P} \mathcal{V} \frac{dn_i}{d\mathcal{P}} d\mathcal{P} \quad (9.46)$$

where $\mathcal{V}$ is the three-velocity of the particle (not volume).

With the above definition of kinetic energy (which is valid for all particles, including photons, *etc.*), we can rewrite equation (9.43) in terms of the derivative with respect to $\mathcal{E}_K$

$$\boxed{\begin{aligned} \frac{dn}{d\mathcal{E}_K} &= g_s \frac{4\pi}{h^3 c^3} [\mathcal{E}_K^2 + 2 \mathcal{E}_K m_0 c^2]^{1/2} \\ &\times \frac{\mathcal{E}_K + m_0 c^2}{\exp \{[(\mathcal{E}_K + m_0 c^2) - \mu_{\text{chem}}]/kT\} \pm 1} \end{aligned}} \quad (9.47)$$

which will be useful for deriving and plotting the special case distribution functions below.

9.3.1.1 The Non-Relativistic Ideal Gas: Tenuous, Warm Fermions

For fermions at not too high a density, the chemical potential μ_{chem} is negative and large, so we can ignore the $+1$ in the denominator of equation (9.47), and the spin factor $g_s = 2$. For “warm” (non-relativistic) particle motions, where $m_0 c^2 \gg \mathcal{E} - \mu_{\text{chem}} \gg kT$, we obtain the classical ideal gas law, with a Maxwellian particle distribution

$$\frac{dn}{d\mathcal{P}} = \frac{8\pi}{h^3} e^{(\mu_{\text{chem}} - m_0 c^2)/kT} \mathcal{P}^2 e^{-\mathcal{E}_K/kT}$$

In this case the kinetic energy of the particle has the Newtonian form

$$\mathcal{E}_K = \frac{\mathcal{P}^2}{2m_0}$$

which yields the Maxwell(–Boltzmann) kinetic energy distribution function

$$\boxed{\frac{dn}{d\mathcal{E}_K} = \frac{4\pi}{h^3} (2m_0)^{3/2} e^{(\mu_{\text{chem}} - m_0 c^2)/kT} \mathcal{E}_K^{1/2} e^{-\mathcal{E}_K/kT}} \quad (9.48)$$

Fig. 9.2 shows a log–log plot of this non-relativistic Maxwellian energy distribution and compares it to others. If we perform the integrals in equations (9.45) and (9.46) using the function in equation (9.48), and if we integrate that same distribution function itself to find $n(T)$, we can derive the familiar expressions for the internal energy and pressure

$$p_g = n k T \quad (9.49)$$

$$\varepsilon_g = \frac{3}{2} n k T \quad (9.50)$$

For this equation of state, the adiabatic index is $\Gamma = 5/3$, the polytropic index for the ideal gas is $n = 3/2$, and the specific heats are $C_p = \frac{5}{2}\mathcal{R}$ and $C_V = \frac{3}{2}\mathcal{R}$.

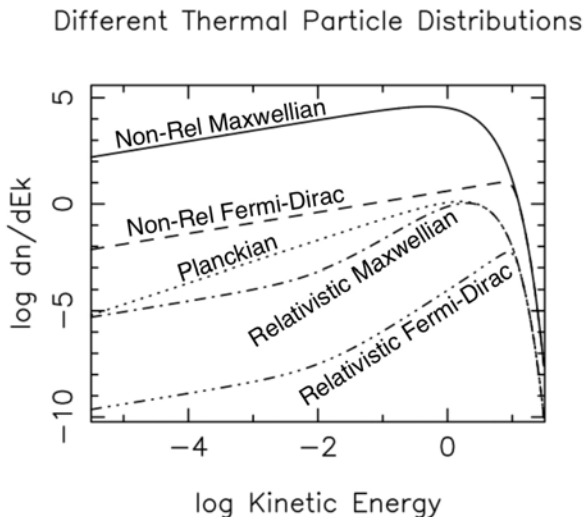


Fig. 9.2: Comparison of different thermal particle distributions discussed in the text, plotted as a function of $\log \mathcal{E}_K/kT$, with $kT = 1$ for all. Top two are the non-relativistic Maxwellian (eq. (9.48)) and Fermi–Dirac (eq. (9.62)) ones, each with $m_0 c^2/kT = 1000$; the latter has a Fermi energy of $\mathcal{E}_F/kT = 10$. The Planckian distribution (eq. (9.59)) is the dotted line in the middle of the figure. Bottom two are relativistic versions of the Maxwellian (eq. (9.55)) and Fermi–Dirac distributions, similar to those at top, but with a particle mass of $m_0/kT = 0.01$. For particle kinetic energies below this mass the distributions’ slopes assume their non-relativistic values.

For a general gas of arbitrary composition, equation (9.49) can be written as

$$p_g = \frac{\rho \mathcal{R} T}{\mu} \quad (9.51)$$

where μ is the “average molecular weight” of the gas (in units of grams per mole). For a neutral atomic hydrogen-only gas $\mu \approx 1$, and for a fully ionized hydrogen gas (protons plus electrons) $\mu = 0.5$. For a more general composition (non-zero X , Y , Z), the *neutral* gas will have an average molecular weight of

$$\mu_{\text{neutral}} = \frac{1}{X + 0.25 Y + 0.06 Z}$$

(Helium is about four times heavier than hydrogen, and the rest of the heavy elements contribute about sixteen times the weight of hydrogen per particle.) So, for solar abundances, $\mu_{\text{neutral}} \approx 1.28$. And, in the case of a fully-ionized gas, with arbitrary composition, the average molecular weight also must take into account the electron contribution to the pressure, so

$$\mu_{\text{ionized}} = \frac{1}{2X + 0.75Y + 0.56Z}$$

or approximately $\mu_{\text{ionized}} \approx 0.61$ for solar abundances. So a fully ionized solar-abundance plasma will have a pressure (including that of the electrons) of

$$p_g = 1.63 \rho \mathcal{R} T \quad (9.52)$$

which differs by only $\sim 20\%$ from that of pure ionized hydrogen.

For a fully ionized gas, the mean molecular weight *per electron* μ_e is often useful. For our simple (X, Y, Z) composition scheme, it is given by

$$\mu_e \approx \frac{1}{X + 0.5(Y + Z)} = \frac{2}{1 + X} \quad (9.53)$$

(which makes the approximation that each heavy atom has an equal number of protons and neutrons). The electron density then can be computed readily from the expression

$$n_e = \frac{N_A \rho}{\mu_e} \quad (9.54)$$

For a solar composition, $\mu_e \approx 1.17$, but for the central core of an evolved star (which has $X = 0$), $\mu_e \approx 2$.

9.3.1.2 The Relativistic Ideal Gas: Tenuous, Hot Fermions

When the gas temperature is high, but the particles still are not particularly dense (*i.e.*, μ_{chem} still large and negative or $\mathcal{E} - \mu_{\text{chem}} \gg kT \geq m_0 c^2$), many particles will be moving relativistically. In this case, the full expression for particle energy (9.44) must be used in the distribution function (9.47)

$$\mathcal{E}_K = (\mathcal{P}^2 c^2 + m_0^2 c^4)^{1/2} - m_0 c^2$$

This yields the “generalized Maxwellian” distribution

$$\frac{dn}{d\mathcal{P}} = \frac{8\pi}{h^3} \left\{ e^{(\mu_{\text{chem}} - m_0 c^2)/kT} \right\} \mathcal{P}^2 e^{-[(m_0^2 c^4 + \mathcal{P}^2 c^2)^{1/2} - m_0 c^2]/kT}$$

or, in kinetic energy,

$$\boxed{\frac{dn}{d\mathcal{E}_K} = \frac{8\pi}{h^3 c^3} \left\{ e^{(\mu_{\text{chem}} - m_0 c^2)/kT} \right\} (\mathcal{E}_K^2 + 2 \mathcal{E}_K m_0 c^2)^{1/2} \times (\mathcal{E}_K + m_0 c^2) e^{-\mathcal{E}_K/kT}} \quad (9.55)$$

which also is plotted in Fig. 9.2. When the kinetic energy is small compared to $m_0 c^2$, equation (9.55) reduces to equation (9.48). On the other hand, when the temperature is very high ($kT \gg m_0 c^2$), the distribution function is dominated by particles with relativistic kinetic energies ($\mathcal{E}_K \gg m_0 c^2$) and takes on the simple “relativistic Maxwellian” shape

$$\frac{dn}{d\mathcal{E}_K} = \frac{8\pi}{h^3 c^3} e^{(\mu_{\text{chem}} - m_0 c^2)/kT} \mathcal{E}_K^2 e^{-\mathcal{E}_K/kT} \quad (9.56)$$

In this case, the pressure (equation (9.46)) is still given by

$$\boxed{p_g = n k T} \quad (9.57)$$

but the internal energy (equation (9.45)) is now

$$\boxed{\varepsilon_g = 3 n k T} \quad (9.58)$$

That is, a relativistic thermal gas behaves like a $\Gamma = 4/3$ gas (polytropic index $n = 3$), rather than the $\Gamma = 5/3$ for a non-relativistic gas, and the specific heats are $C_p = 4\mathcal{R}$ and $C_V = 3\mathcal{R}$. This difference in Γ between relativistic and non-relativistic gases turns out to be a general rule for all kinds of gases and fluids, not just ideal thermal ones.

9.3.1.3 The Photon Gas: Hot Bosons

For photons, which always move relativistically, $\mu_{\text{chem}} = 0$ and $m_0 = 0$ (they have no chemical potential or mass), but $g_s = 2$ for two spin states (± 1), and the sign in equation (9.47) is negative, because photons are bosons⁴. They also have $\mathcal{E}_K = \mathcal{E} = \mathcal{P}c = h\nu$ (where ν is the photon frequency), so their *number* distribution function in kinetic energy space is

$$\boxed{\frac{dn}{d\mathcal{E}_K} = \frac{8\pi}{h^3 c^3} \frac{\mathcal{E}_K^2}{e^{\mathcal{E}_K/kT} - 1}} \quad (9.59)$$

⁴ While there has been some work in the field on boson stars, we will not be considering *cold* boson gases in this book. In the terrestrial environment, a fluid of cold ^4He atoms is essentially a cold boson gas: while helium is composed of two protons, two neutrons, and two electrons, in the ground state their spins all exactly cancel, leaving a single zero-spin particle that does not form molecules. The peculiar properties of liquid helium are due to the boson nature of these cold, non-relativistic atoms.

which is also plotted in Fig. 9.2. In frequency space, the photon *energy* distribution is the familiar Planckian spectral energy density

$$\mathcal{E}_K \frac{dn}{d\nu} = \frac{8\pi h \nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1}$$

and the intensity is given by

$$I(\nu) = \frac{c}{4\pi} \mathcal{E}_K \frac{dn}{d\nu} = B_\nu(T)$$

where $B_\nu(T)$ is the well-known Planck function that describes the intensity of photons, per unit frequency, when they are in thermal equilibrium at a temperature T

$$B_\nu(T) \equiv \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1} \quad (9.60)$$

Note that, at low energies ($\mathcal{E}_K \ll kT$), the energy density and intensity are proportional $\mathcal{E}_K^2$ (or ν^2) – the Rayleigh–Jeans spectrum – and at high energies ($\mathcal{E}_K \gg kT$), they are proportional to $\mathcal{E}_K^3 \exp(-\mathcal{E}_K/kT)$ – the Wien spectrum.

We again can integrate equations (9.45) and (9.46) to find that the pressure and internal energy of the photon gas vary as the fourth power of temperature⁵

$$\varepsilon_r = 3p_r = aT^4 \quad (9.61)$$

where $a = 8\pi^5 k^4/15 h^3 c^3 = 7.56577 \times 10^{-15} \text{ erg cm}^{-3} \text{ K}^{-4}$ is the radiation constant. So, as with the relativistic fermion gas, the relativistic boson gas has an adiabatic index of $\Gamma = 4/3$.

9.3.1.4 The Degenerate Gas: Dense Fermions

When the density of electrons, neutrons, *etc.* becomes so large that the Pauli exclusion principle becomes important, the chemical potential per particle climbs to a positive number. All the lower quantum levels are filled (only two fermions are allowed per level), and another particle can be added only at a higher energy. This peculiar distribution function usually occurs only in very dense and cold situations (*i.e.*, in the centers of degenerate stars like white dwarf and neutron stars or in conducting solid metals on the earth). We include the degenerate gas in this book because both white dwarf and neutron stars are important precursors for forming black holes, and because such dense matter also could occur in extreme accretion conditions where black holes consume tidally disrupted degenerate stars.

Because we cannot ignore the +1 in the denominator of equation (9.47) in this case, the particle distribution function is the full Fermi–Dirac one

⁵ Note that we also could have integrated n here and obtained equations of state like (9.49) and (9.50), *i.e.*, $\varepsilon_r = 3p_r \approx 669 nkT$, where n is the number density of photons, but such an equation of state is generally not used for radiation studies.

$$\frac{dn}{d\mathcal{E}_K} = \frac{8\pi}{h^3 c^3} [\mathcal{E}_K^2 + 2\mathcal{E}_K m_0 c^2]^{1/2} \frac{\mathcal{E}_K + m_0 c^2}{e^{[\mathcal{E}_K - (\mu_{\text{chem}} - m_0 c^2)]/kT} + 1} \quad (9.62)$$

The gas is highly degenerate if the temperature is low compared to a critical temperature, called the “Fermi temperature”

$$T_F \equiv \frac{\mu_{\text{chem}} - m_0 c^2}{k} = \frac{\mathcal{E}_F}{k} \quad (9.63)$$

where $\mathcal{E}_F = \mu_{\text{chem}} - m_0 c^2$ is the “Fermi energy”. The actual value of $\mathcal{E}_F$ will depend on whether the particles are electrons (as in white dwarf stars) or neutrons (as in neutron stars). Then the expression in the denominator has two asymptotic values: unity for $\mathcal{E}_K \ll \mathcal{E}_F$ and infinity for $\mathcal{E}_K \gg \mathcal{E}_F$. Equations (9.45) and (9.46) can be integrated again to obtain the pressure and internal energy density

$$p = \frac{8\pi}{3} \left(\frac{m_0 c}{h} \right)^3 m_0 c^2 \mathcal{P}(x) \quad (9.64)$$

and

$$\varepsilon = \frac{8\pi}{3} \left(\frac{m_0 c}{h} \right)^3 m_0 c^2 \mathcal{E}(x) \quad (9.65)$$

where $x \equiv \mathcal{E}_F / m_0 c^2$ and the normalized pressure and energy functions are

$$\begin{aligned} \mathcal{P}(x) &= x(2x^2 - 3)(x^2 + 1)^{1/2} + 3 \sinh^{-1} x \\ \mathcal{E}(x) &= 3x(2x^2 + 1)(x^2 + 1)^{1/2} - 8x^3 - 3 \sinh^{-1} x \end{aligned}$$

When $x \ll 1$ (non-relativistic Fermi–Dirac distribution) we have

$$\begin{aligned} \mathcal{P}(x) &\approx \frac{8}{5} x^5 - \frac{4}{7} x^7 \\ \mathcal{E}(x) &\approx \frac{12}{5} x^5 - \frac{3}{7} x^7 \end{aligned}$$

which gives $\varepsilon = \frac{3}{2} p$ as $x \rightarrow 0$, as expected for a non-relativistic gas. When $x \gg 1$ (relativistic Fermi–Dirac gas), we have

$$\begin{aligned} \mathcal{P}(x) &\approx 2x^4 - 2x^2 \\ \mathcal{E}(x) &\approx 6x^4 - 8x^3 - 6x^2 \end{aligned}$$

which gives the familiar $\varepsilon = 3p$ when $x \rightarrow \infty$, as expected. So, even in the degenerate case, the non-relativistic gas has $\Gamma = 5/3$, while the relativistic one has $\Gamma = 4/3$.

Handy expressions for the pressure for a degenerate *electron* gas are, for the non-relativistic and relativistic cases,

$$p_{e, NR} = 1.00 \times 10^{13} \text{ dyn cm}^{-2} \left(\frac{\rho}{\mu} \right)^{5/3} \quad (9.66)$$

$$p_{e, R} = 1.24 \times 10^{15} \text{ dyn cm}^{-2} \left(\frac{\rho}{\mu} \right)^{4/3} \quad (9.67)$$

and for a degenerate *neutron* gas

$$p_{n, NR} = 5.50 \times 10^9 \text{ dyn cm}^{-2} \rho^{5/3} \quad (9.68)$$

$$p_{n, R} = 1.24 \times 10^{15} \text{ dyn cm}^{-2} \rho^{4/3} \quad (9.69)$$

with ρ and μ in cgs units, and the standard $\varepsilon = p/(\Gamma - 1)$ giving the internal energy density for each. Note the similarity between the two different degenerate gases in the relativistic cases. The boundaries between the non-relativistic and relativistic cases are approximately $1.9 \times 10^6 \text{ g cm}^{-3}$ for the degenerate electron gas and $1.15 \times 10^{16} \text{ g cm}^{-3}$ for degenerate neutrons.

Fig. 9.3 shows the regions in the n - T plane where different thermal equations of state are important. Radiation pressure dominates the high-temperature, relatively low-density region, while degenerate equations of state dominate the opposite extremes.

9.3.2 Pressure and Internal Energy of Nonthermal Gases

In very tenuous plasmas with strong electromagnetic fields, such as near pulsars or in extragalactic radio jets and lobes, the high-energy tail of the particle distribution function does not cut off at a typical energy given by $\sim kT$. Instead, the non-interacting particles are distributed over many orders of magnitude in energy, extending to particle Lorentz factors as high as $\gamma \approx \mathcal{E}_K/m_0c^2 \sim 10^6$ or greater. These plasmas cannot be characterized by a single temperature and, so, are called “nonthermal” plasmas. The particle distribution function is typically a power law over a finite range in kinetic energy $\mathcal{E}_K$

$$\frac{dn}{d\mathcal{E}_K} = n \frac{1 - \beta}{\mathcal{E}_{K, \max}^{1-\beta} - \mathcal{E}_{K, \min}^{1-\beta}} \mathcal{E}_K^{-\beta} \quad \mathcal{E}_{K, \min} < \mathcal{E}_K < \mathcal{E}_{K, \max} \quad (9.70)$$

where the index of the power law $\beta > 0$, and n is the total density of nonthermal particles. If $\beta > 1$, then the distribution function is steep and dominated by low-energy particles, perhaps even a very low-energy thermal distribution. On the other hand, if $\beta < 1$, then the distribution is shallow, dominated by the high-energy end, and must be cut off more steeply beyond $\mathcal{E}_{K, \max}$. The particle energy density for this distribution is

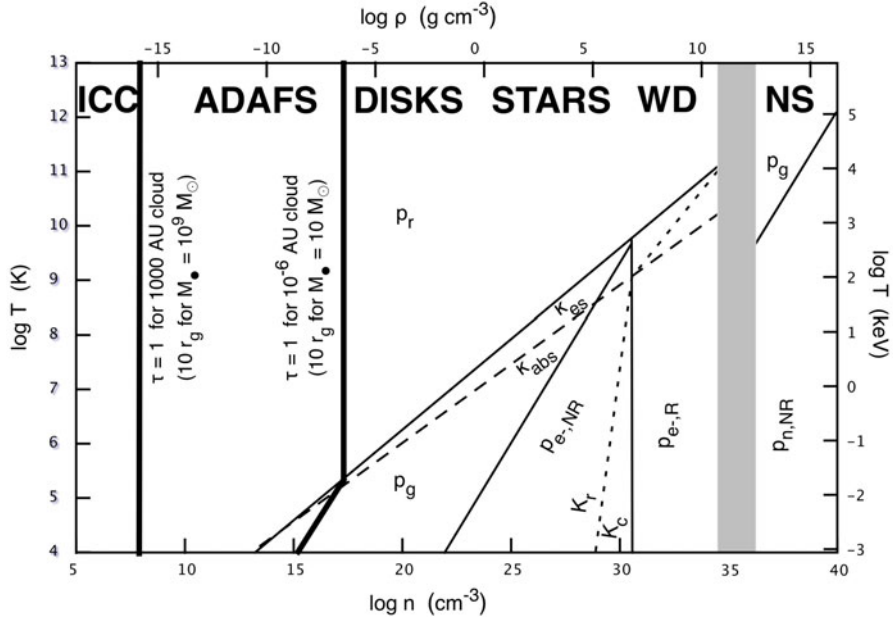


Fig. 9.3: The n - T plane for most equations of state discussed in Section 9.3. The thin solid lines divide the plane into regions where different pressure formulae dominate; from left-to-right they are: radiation (eq. (9.61)), ideal gas (eq. (9.49)), non-relativistic and relativistic electron degeneracy (eqs. (9.66) and (9.67)). The vertical gray patch is a rough boundary between white dwarf and neutron star physics and shows where plasmas are usually in dynamical collapse rather than hydrostatic equilibrium. To the right of it are the non-relativistic neutron degeneracy region (eq. (9.68)) and, above it, the region where hot neutrons act like an ideal gas. (Relativistic degenerate neutron pressure lies beyond the right-hand boundary for $\log n > 40.1$, and we have not included neutrino pressure in the diagram.) The dashed line divides the plane into regions where electron scattering (eq. (9.78)) and absorption (eq. (9.81)) dominate. The dotted line separates regions where heat transport is primarily by radiation (K_r) and by heat conduction (K_c). The two broad solid lines show where the assumption of an infinite medium breaks down for these equations of state: $\tau = 1$ for a blob with a half-height of $10 r_g$ for both a $10 M_\odot$ and a $10^9 M_\odot$ black hole. To the left of these boundaries the medium is optically thin to radiation, so pressure is due primarily to a hot ideal gas. Labels at the top show approximate density ranges of various astrophysical objects: interstellar cloud cores (ICC), ADAFs, Shakura–Sunyaev disks, stellar interiors, white dwarfs (WD), and neutron stars (NS).

$$\varepsilon = 3p = n \left(\frac{1-\beta}{2-\beta} \right) \left(\frac{\mathcal{E}_{K,\max}^{2-\beta} - \mathcal{E}_{K,\min}^{2-\beta}}{\mathcal{E}_{K,\max}^{1-\beta} - \mathcal{E}_{K,\min}^{1-\beta}} \right) \quad (9.71)$$

which, for $\mathcal{E}_{K,\max} \gg \mathcal{E}_{K,\min}$ and $\beta < 1$, is approximately

$$\varepsilon = 3p \approx n \left(\frac{1-\beta}{2-\beta} \right) \mathcal{E}_{K,\max}$$

Like the thermal relativistic gases, the nonthermal plasma is a $\Gamma = 4/3$ gas, and, for the conditions above, an average energy per particle of

$$\langle \mathcal{E}_K \rangle = \left(\frac{1 - \beta}{2 - \beta} \right) \mathcal{E}_{K, \max}$$

9.3.3 Thermal Conductivity

The term equation of state usually refers to expressions for pressure and internal energy as a function of local rest mass density and temperature. However, from the point of view of the conservation laws, any algebraic function of those variables that is needed to compute the conserved quantities is an equation of state. One such function is the thermal conductivity, needed in equation (9.10), which can be estimated from Frick's law of diffusion

$$K_c \approx \frac{1}{3} C_V \langle V_c \rangle \ell_c \quad (9.72)$$

where C_V is the specific heat at constant volume (see Section 5.2.2), $\langle V_c \rangle$ is the average velocity of particles that conduct heat, and ℓ_c is the mean free path of those particles. For a thermal plasma, electrons conduct the most heat, because of their low mass and high velocity, so

$$\begin{aligned} C_V &= \frac{3}{2} n_e k \\ \langle V_c \rangle &= (3 kT / m_e)^{1/2} \\ \ell_c &= \frac{1}{n_e \sigma_C} \end{aligned}$$

and the electron Coulomb cross-section

$$\sigma_C = \pi r_C^2$$

is determined from the impact parameter r_C (average closest approach) of Coulomb collisions of electrons with other ions. The closest approach of an electron to an ion of charge e occurs when the Coulomb energy equals the kinetic energy

$$\mathcal{E}_C = \frac{e^2}{r_C} = kT = \mathcal{E}_K$$

so

$$r_C = \frac{e^2}{kT}$$

Combining everything, equation (9.72) tells us that K_c depends only on the electron temperature

$$K_c \approx \frac{k}{2\pi e^4} \left(\frac{3}{m_e} \right)^{1/2} (kT)^{5/2}$$

or, in cgs units,

$$K_c \approx 5.38 \times 10^{18} \text{ erg cm}^{-1} \text{ s}^{-1} \text{ K}^{-1} \left(\frac{T}{4.0 \times 10^9 \text{ K}} \right)^{5/2} \quad (9.73)$$

How important is heat conduction for black hole engines? We can estimate this by comparing the heat flux to the energy flux created by gas motions (advection) at a speed V

$$\frac{Q_g}{V\varepsilon} \approx \frac{\langle V_c \rangle}{V} \frac{\ell_c}{R}$$

where $R \sim (10\text{--}100) r_g$ is a typical length scale in the accreting black hole system. For static stars, heat conduction is important since internal motions $V \ll \langle V_c \rangle$, and even slow diffusive processes will become important. But for accreting black holes, infall velocities can approach the sound speed. For an electron–proton plasma the ratio above becomes

$$\begin{aligned} \frac{Q_g}{V\varepsilon} &\approx \left(\frac{m_p}{m_e} \right)^{1/2} \frac{(kT)^2}{\pi e^4 n_e R} \\ &= 1.0 \left(\frac{n_e}{5.3 \times 10^{18} \text{ cm}^{-3}} \right)^{-1} \left(\frac{M}{10 M_\odot} \right)^{-1} \left(\frac{R}{10 r_g} \right)^{-1} \left(\frac{T}{4.0 \times 10^9 \text{ K}} \right)^2 \end{aligned}$$

So, conduction may be more important than advection of the thermal gas for an accreting $10 M_\odot$ black hole if $n_e < 10^{18} \text{ cm}^{-3}$, and for a $10^9 M_\odot$ black hole if $n_e < 10^{10} \text{ cm}^{-3}$. Generally, however, in accretion situations conduction is much less important than radiative transport, so the former often can be ignored, unless the accretion flow is radiatively inefficient.

9.3.4 Particle Viscosity

Viscosity can be analyzed in a manner similar to that used for thermal conductivity. The coefficients of viscosity are

$$\eta_{v,g} \approx \zeta_{v,g} \approx \rho \langle V_v \rangle \ell_v \quad (9.74)$$

where $\langle V_v \rangle$ is the average velocity of particles that carry three-momentum and $\ell_v = (kT)^2 m_p / \rho \pi c^4$ is their mean free path. For an electron–proton plasma, it is the

protons that carry the three-momentum, so $V_v = (3kT/m_p)^{1/2}$, and the magnitude of the viscosity coefficients is

$$\eta_{v,g} \approx \frac{1}{\pi e^4} (3m_p)^{1/2} (kT)^{5/2}$$

or, in cgs units,

$$\boxed{\eta_{v,g} \approx 3.0 \times 10^9 \text{ erg s cm}^{-3} \left(\frac{T}{4.0 \times 10^9 \text{ K}} \right)^{5/2}} \quad (9.75)$$

We again can estimate how important particle viscosity might be in an accretion situation by comparing the viscous stress components with the local pressure p

$$\begin{aligned} \frac{|\mathcal{T}_{\text{VISC}}|}{p} &\approx \frac{\eta_{v,g} \langle V_v \rangle / R}{n k T} \approx 3 \frac{\ell_v}{R} = 3 \frac{(kT)^2}{\pi e^4 n R} \\ &= 1.0 \left(\frac{n}{3.7 \times 10^{17} \text{ cm}^{-3}} \right)^{-1} \left(\frac{M}{10 M_\odot} \right)^{-1} \left(\frac{R}{10 r_g} \right)^{-1} \left(\frac{T}{4.0 \times 10^9 \text{ K}} \right)^2 \end{aligned}$$

This ratio is a factor of $(m_p/m_e)^{1/2}/3 \sim 14$ lower than the one above for conduction, so the criteria for viscosity to be important is somewhat more stringent than that for conduction. Nevertheless, in low accretion rate situations, it is possible that particle viscosity may be important as well.

9.3.5 Turbulent Viscosity

Turbulence is the random, chaotic, motion that often occurs in fluids on scales much smaller than the overall system size, but much larger than the distance between independent fluid particles or even between fluid elements. The phenomenon usually develops in fluids undergoing shear flow, unless the microscopic viscosity is strong enough to damp out any growing chaotic motions and turn them into heat. It also can occur in fluids that have a weak, but non-zero, magnetic field. Turbulence actually is produced by fluid motions and is not really a separate physical microscopic process. However, the motions are so complex, compared to regular *laminar* flow that statistical methods have been developed to treat chaotic fluid motions, just as such methods were developed to handle the mechanics of multiple particles in motion in the first place.

If the size scale of the turbulent eddies ℓ_t is much smaller than the overall size of the system, then one can use the *turbulent diffusion approximation* to define a turbulent viscosity

$$\eta_{v,t} \approx \zeta_{v,t} \approx \rho \langle V_t \rangle \ell_t$$

where $\langle V_t \rangle$ is the RMS velocity of the chaotic motions in the eddies. In the early 1970s the nature of turbulent flow in black hole accretion flows was largely unknown. So early investigators [351, 352] assumed that the RMS turbulent velocity was a fraction of the sound speed

$$\langle V_t \rangle \approx \alpha c_s$$

where the free parameter $\alpha \leq 1$. This “alpha model” of turbulence was quite successful in the early days of black hole accretion studies. See Chapter 12.

In addition to the obvious issues associated with choosing a diffusion approximation, treating turbulence as a viscous process has some other assumptions associated with it. Recall (Section 9.2.1) that having a viscosity implies that viscous dissipation of the shear exists (viscous heating). This is because the viscous part of the stress-energy tensor does not have its own energy density (ε_t is missing). In reality, however, turbulence *does* have an energy density, a pressure also, and heat flow as well, not just viscous-like properties. All of this physics is missing in this treatment, along with a model for how to convert turbulent energy into heat. Instead, the simple viscous approximation assumes that all mechanical energy lost due to viscosity immediately is converted into heat (equation (9.16)). While this works rather well in accretion models, it still should be remembered that turbulence can be much more complex than a simple *ad hoc* viscosity.

9.3.6 Radiative Opacity

The final equation of state needed in the conservation laws is the radiative opacity $\bar{\kappa}_R$ (equation (9.20)). Radiation opacity can be thought of as a thermal resistivity for photons, rather than for gas particles. In fact, radiative transport is often treated as an addition to the particle thermal conductivity

$$K = K_c + K_r \quad (9.76)$$

where

$$K_r = \frac{4}{3} \frac{a c}{\bar{\kappa}_R} \frac{T^3}{\rho}$$

(See equations (9.21) and (9.61).)

In the general, frequency-dependent, situation Frick’s law (equation (9.72)) can be re-written to give us an expression for the opacity

$$\kappa(\nu) = \frac{n \sigma_r(\nu)}{\rho} = \sigma_r(\nu) \frac{N_A}{\mu}$$

where $\sigma_r(\nu)$ is the photon cross-section for radiation scattering or absorption by a given type of particle, as a function of photon frequency. That is, Frick’s law

says that the opacity is equal to the cross-section divided by the average mass per particle. There are a variety of effects that can make matter opaque to radiation. In the temperature and density range in which we are interested they all involve an incoming photon colliding with an electron in some way, either directly or in conjunction with a positively charged ion.

9.3.6.1 Electron Scattering

Scattering by a free electron is the simplest process, and we will deal with that first. The electron scattering opacity is given by Frick's law as

$$\kappa_{\text{es}} = \sigma_{\text{KN}}(\nu) \frac{N_A(1 + X)}{2} \quad (9.77)$$

where we have substituted equation (9.53) for the mean molecular weight per electron, and $\sigma_{\text{KN}}(\nu)$ is the Klein–Nishina cross-section for general electron scattering

$$\sigma_{\text{KN}}(\nu) = \frac{3}{4} \sigma_{\text{T}} \left\{ \frac{2(1 + \theta_\nu)}{\theta_\nu^2} \left[\frac{1 + \theta_\nu}{1 + 2\theta_\nu} - \frac{\ln(1 + 2\theta_\nu)}{2\theta_\nu} \right] + \frac{\ln(1 + 2\theta_\nu)}{2\theta_\nu} - \frac{1 + 3\theta_\nu}{(1 + 2\theta_\nu)^2} \right\}$$

with $\theta_\nu \equiv h\nu/m_e c^2 \approx (T/5.9 \times 10^9 \text{ K})$ is the energy of the incoming photon in electron rest mass units, and σ_{T} is the Thomson cross-section

$$\begin{aligned} \sigma_{\text{T}} &= \frac{8\pi}{3} \left(\frac{e^2}{m_e c^2} \right)^2 \\ &= 6.65246 \times 10^{-25} \text{ cm}^2 \end{aligned}$$

For low photon energies ($10^4 \text{ K} < T < 10^7 \text{ K}$), the Klein–Nishina cross-section approaches the frequency-independent Thomson cross-section and we have, simply,

$$\boxed{\kappa_{\text{es}} = 0.400 \text{ cm}^2 \text{ g}^{-1} / \mu_e = 0.200 \text{ cm}^2 \text{ g}^{-1} (1 + X)} \quad (9.78)$$

or $\kappa_{\text{es}} \approx 0.34 \text{ cm}^2 \text{ g}^{-1}$ for solar abundances. For $T \gg 10^{11} \text{ K}$, the Klein–Nishina cross-section becomes very small, and other forms of opacity dominate.

In the intermediate range ($10^7 \text{ K} < T < 10^{11} \text{ K}$) the cross section is significant and the scattering is inelastic; *i.e.*, the photon loses some energy and the electron gains some energy. This is the well-known process of Compton scattering. On the other hand, if the electrons are hotter than the photons, Compton scattering can have the opposite effect: the “inverse Compton” process can cool the plasma and up-scatter the photons to higher energies. (See Section 9.4.3.)

9.3.6.2 Free–Free and Bound–Free Absorption

If a photon and an unbound electron collide near a positively charged ion, it is possible for the photon to be absorbed, rather than simply scattered. This process is called free–free absorption. The electron’s kinetic energy increases, and, when it eventually collides with another electron or ion, that extra energy will heat the plasma. Only much later may the inverse process (Bremsstrahlung emission), Section 9.4.1, or some other process, emit a photon again and convert that absorbed energy back into radiation. *Photo-absorption and photo-emission, therefore, are treated as separate heating and cooling processes, rather than two parts of a single scattering.*

A similar effect occurs if the electron is bound to a nucleus, but the incoming photon has enough energy to eject the electron from that nucleus and ionize it. The photon again is absorbed in the event, so this process is called bound–free absorption. The inverse process, recombination emission, also occurs separately from bound–free absorption, and need not involve the electron and ion that participated in the original ionization.

Because the free–free $\kappa_{\text{ff}}(\nu)$ and bound–free $\kappa_{\text{bf}}(\nu)$ opacities depend on the photon frequency ν , and because these are so important in stellar structure, their Rosseland means have been worked out and are well known. Generally, the Planck function $B_\nu(T)$ (equation (9.60)) is used to compute the Rosseland mean, since it differs from the photon energy density $\varepsilon_r(\nu)$ by only a constant factor

$$\bar{\kappa}_R = \frac{a c T^3}{\pi \int_0^\infty \frac{1}{\kappa(\nu)} \frac{\partial B_\nu}{\partial T} d\nu}$$

The resulting values for each process are

$$\bar{\kappa}_{R,\text{ff}} = 7.36 \times 10^{22} \text{ cm}^2 \text{ g}^{-1} (X + Y) \rho T^{-7/2} \bar{g}_{\text{ff}} / \mu_e \quad (9.79)$$

$$\bar{\kappa}_{R,\text{bf}} = 8.68 \times 10^{25} \text{ cm}^2 \text{ g}^{-1} f(T) Z \rho T^{-7/2} \bar{g}_{\text{bf}} / \mu_e \quad (9.80)$$

where $\bar{g}_{\text{ff}}$ and $\bar{g}_{\text{bf}}$ are “Gaunt factors” (correction factors for the emission spectrum not being exactly flat with frequency ν) which are generally of order unity and often ignored, and the function $f(T)$ is the fraction of heavy elements that are *not* ionized ($f(T) \rightarrow 0$ as $T \rightarrow \infty$). Note that the free–free opacity is dominated by electrons contributed by H and He, while the bound–free opacity is contributed mainly by partially-ionized heavy elements.

The total opacity due to both processes is, ignoring the Gaunt factors,

$$\bar{\kappa}_{R,\text{ff}} + \bar{\kappa}_{R,\text{bf}} \approx 3.68 \times 10^{22} \text{ cm}^2 \text{ g}^{-1} [X + Y + 1180 Z f(T)] (1 + X) \rho T^{-7/2}$$

or

$$\boxed{\bar{\kappa}_{R,\text{ff}} + \bar{\kappa}_{R,\text{bf}} \approx 1.55 \times 10^{24} \text{ cm}^2 \text{ g}^{-1} \rho T^{-7/2}} \quad (9.81)$$

for solar abundances and $f(T) \approx 1$.

9.3.6.3 Bound–Bound Absorption

At the temperatures in which we are interested ($>10^6$ K), bound–bound (line) absorption will not be important for energy balance, but it will be important in the cooler atmospheres for absorbing photon momentum. Line *emission* from partially-ionized heavy ions also will be important for interpreting the observed X-ray photon spectrum.

9.3.7 Radiative Heat Transport vs. Thermal Conduction

Which is more important, heat transport by conduction (Q_g) or by radiation (Q_r)? We can estimate this by computing the ratio

$$\frac{|Q_g|}{|Q_r|} = \frac{K_c}{K_r} = \frac{4 c T^3}{\frac{\partial \varepsilon}{\partial T} \langle V_c \rangle \ell_c \bar{\kappa}_R \rho}$$

Because the opacity can be either electron scattering or absorption, we must test both cases. They give, respectively, the following criteria for *radiative* heat transport to dominate over conduction

$$\begin{aligned} n &< 6.4 \times 10^{30} \text{ cm}^{-3} \left(\frac{T}{4 \times 10^9 \text{ K}} \right)^{1/2} & \bar{\kappa}_R &= \kappa_{\text{es}} \\ n &< 4.1 \times 10^{31} \text{ cm}^{-3} \left(\frac{T}{4 \times 10^9 \text{ K}} \right)^2 & \bar{\kappa}_R &= \bar{\kappa}_{R,\text{ff}} + \bar{\kappa}_{R,\text{bf}} \end{aligned}$$

Either way, these are enormous densities – much greater than the central density of the sun or other main sequence stars and approaching white dwarf densities. So, *for nearly all the applications of black hole accretion that we will encounter in this book, radiation will dominate conduction, and we can ignore the latter.*

Fig. 9.3 also shows region in the n – T plane where electron scattering vs. absorption and radiative heat transport vs. conduction dominate.

9.4 Optically Thin Radiative Emission

So far we have discussed only optically thick radiative transport of energy by diffusion (equation (9.24)). We can construct a heat per unit mass time derivative (see equation (5.50)) from this equation by defining

$$\dot{q}_r = \frac{1}{\rho} \nabla \cdot \mathbf{Q}_r = -\frac{c}{3\rho} \nabla \cdot \left[\frac{1}{\bar{\kappa}_{R,bf}} \nabla \varepsilon_r \right]$$

Since we assume here that the radiation and matter are thermal equilibrium ($\varepsilon_r = aT^4$), the local spectrum will be Planckian at this temperature.

In this subsection we consider the opposite case, where the photons are emitted locally (again at the local temperature when the emission is thermal), but they then immediately leave the area before being absorbed. That is, the radiation spectrum remains approximately that of the emission process and does not come into thermal equilibrium with the local gas. We discuss below two radiative processes that are important in high-temperature astrophysics, Bremsstrahlung and synchrotron emission, and we do consider the effects of inelastic (Compton) scattering of that radiation off electrons. Note that we will not consider the case where the absorption optical depth is of order unity, but we shall consider electron scattering optical depths of all values.

9.4.1 Bremsstrahlung (Free-Free Emission)

Bremsstrahlung is the inverse process of free-free absorption and occurs when an electron decelerates near another charged particle (ion, another electron, or a positron), emitting a photon to conserve energy. The cooling rate (per gram) for these processes is [353]

$$\dot{q}_{br}^- = \dot{q}_{ei}^- + \dot{q}_{ee}^- + \dot{q}_{\pm}^- \quad (9.82)$$

where

$$\begin{aligned} \dot{q}_{ei}^- &= 147 \text{ erg g}^{-1} \text{ s}^{-1} (n_e + n_+) \\ &\times \begin{cases} 1.00 \theta_e^{1/2} (1 + 1.781 \theta_e^{1.34}) & \theta_e < 1 \\ 1.41 \theta_e [\ln(1.123 \theta_e + 0.48) + 1.5] & \theta_e \geq 1 \end{cases} \end{aligned} \quad (9.83)$$

$$\begin{aligned} \dot{q}_{ee}^- &= 251 \text{ erg g}^{-1} \text{ s}^{-1} \frac{(n_e^2 + n_+^2)}{n_p} \\ &\times \begin{cases} 1.00 \theta_e^{1.5} (1 + 1.1 \theta_e + \theta_e^2 - 1.25 \theta_e^{2.5}) & \theta_e < 1 \\ 1.33 \theta_e [\ln(1.123 \theta_e) + 1.28] & \theta_e \geq 1 \end{cases} \end{aligned} \quad (9.84)$$

$$\begin{aligned} \dot{q}_{\pm}^- &= 337 \text{ erg g}^{-1} \text{ s}^{-1} \frac{n_e n_+}{n_p} \\ &\times \begin{cases} 1.00 \theta_e^{1/2} (1 + 1.7 \theta_e^{3/2}) & \theta_e < 1 \\ 1.99 \theta_e [\ln(1.123 \theta_e) + 1.24] & \theta_e \geq 1 \end{cases} \end{aligned} \quad (9.85)$$

represent the cooling due to electron-ion and positron-ion (9.83), electron-electron and positron-positron (9.84), and electron-positron (9.85) processes. Here $n_p = n_e - n_+$ is the number density of protons, n_+ is the number density of positrons,

and $\theta_e \equiv kT_e/m_e c^2$ is the dimensionless electron temperature. One can determine the ratio $n_+/n_p = (n_+/n_e)/(1 - n_+/n_e)$ needed to calculate some of these terms using the following expression

$$\frac{n_+}{n_e} = 6.4 \times 10^{-5} \left\{ 1 + \left[\frac{2\theta_e^2}{\ln(1.12\theta_e + 1.3)} \right] \right\} \times \begin{cases} 1.00 \theta_e^{3/2} & \exp(-2/\theta_e)(1 + 0.015\theta_e) \theta_e \ll 1 \\ 0.35 (\ln \theta_e)^3 (1 + 0.058/\theta_e)^{-1} & \theta_e \gg 1 \end{cases} \quad (9.86)$$

where α_f is the fine structure constant. We have found that this expression works well for $T_e \lesssim 2 \times 10^{11}$ K, but becomes infinite slightly above this temperature.

9.4.2 Synchrotron

Synchrotron radiation occurs when an electron decelerates in a magnetic field, again emitting a photon to conserve energy. The local synchrotron cooling rate is approximated as a sum of optically thick and thin emission

$$\dot{q}_s^- = \frac{2\pi kT_e}{3Hc^2} \nu_c^3 + \int_{\nu_c}^{\infty} \epsilon_s(\nu) d\nu \quad (9.87)$$

where H is the temperature scale height. The actual spectrum emitted by all the electrons depends on the electron distribution function. Here we consider a *thermal* electron distribution (non-relativistic or relativistic Maxwellian), so the radiation spectrum and its cooling rate will depend on the electron temperature T_e .

The critical frequency ν_c can be found by equating the optically thin and thick volume emissivities at that frequency

$$\epsilon_s(\nu_c) = \frac{2\pi}{H} \frac{\nu_c^2}{c^2} kT_e \quad (9.88)$$

and solving the above expression numerically. For an isotropic, Maxwellian distribution of electrons and positrons, the optically thin volume emissivity (for $\nu > \nu_c$) is [354, 353]

$$\epsilon_s(\nu, \vartheta) = 4.43 \times 10^{-30} 4\pi\nu (n_e + n_+) \frac{I(x_M/\sin \vartheta)}{K_2(1/\theta_e)} \text{ erg cm}^{-3} \text{ s}^{-1} \quad (9.89)$$

where $x_M = \nu/\nu_M$ is the normalized frequency (with $\nu_M = 6.27 \times 10^{18} B (kT_e)^2$ [cgs] being the critical *electron* frequency for a given temperature), ϑ is the angle between the observer and the magnetic field direction, K_2 is the modified Bessel function of the second kind of order 2, given by the integral

$$K_2(1/\theta_e) \equiv \frac{\theta_e^2}{3} \int_{1/\theta_e}^{\infty} (z^2 - 1/\theta_e^2)^{3/2} e^{-z} dz \quad (9.90)$$

and the electron-energy-integrated, unitless spectrum is given by [1]

$$I\left(\frac{x_M}{\sin \vartheta}\right) \equiv \frac{\sin \vartheta}{x_M} \int_{1/\theta_e}^{\infty} (z^2 - 1/\theta_e^2)^{3/2} e^{-z} F(x_M/z^2 \sin \vartheta) dz \quad (9.91)$$

where $F(x)$ is the normalized synchrotron spectrum for a single electron

$$F(x) = x \int_x^{\infty} K_{5/3}(\xi) d\xi$$

In the high-temperature limit ($\theta_e \gg 1$) equation (9.89) can be simplified and made tractable⁶

$$\epsilon_s(\nu) = 4.43 \times 10^{-30} 2\pi\nu (n_e + n_+) \frac{I'(x_M)}{\theta_e^2} \text{ erg cm}^{-3} \text{ s}^{-1} \quad (9.92)$$

That is, in the limit $1/\theta_e \rightarrow 0$, the integral in equation (9.90) becomes $K_2(1/\theta_e) \rightarrow 2\theta_e^2$, and the integral in equation (9.91) (when also averaged over angle ϑ) can be fit to the following expression [354]

$$I'(x_M) = \frac{4.0505}{x_M^{1/6}} \left(1 + \frac{0.40}{x_M^{1/4}} + \frac{0.5316}{x_M^{1/2}} \right) \exp(-1.8899 x_M^{1/3}) \quad (9.93)$$

with no more than 2.7% error over the range $0 < x_M < \infty$. Equations (9.92) and (9.93) represent the synchrotron cooling very well over the range in temperature where this type of emission is important, and, *when used together in equation (9.91)*, they avoid a cooling runaway that can produce the wrong physical results in MHD simulations [355].

9.4.3 Comptonization

In many of the high-temperature accretion situations that we shall investigate in this book the opacity due to bound-free and free-free absorption (equation (9.81)) will be largely unimportant. Instead, electron scattering (equation (9.78)) will dominate. In each of these photon scatterings, if the electron and photon energies are very different, an inelastic scattering (the Compton process) will occur, either raising or lowering the photon energy in the process. In most cases we will encounter, the pho-

⁶ For temperatures $\theta_e \ll 1$ the high-temperature approximation obviously is not correct. However, at that point Bremsstrahlung is the dominant emission mechanism and the error in retaining this value for $K_2(1/\theta_e)$ makes a negligible error in the overall cooling. (See Fig. 9.4.)

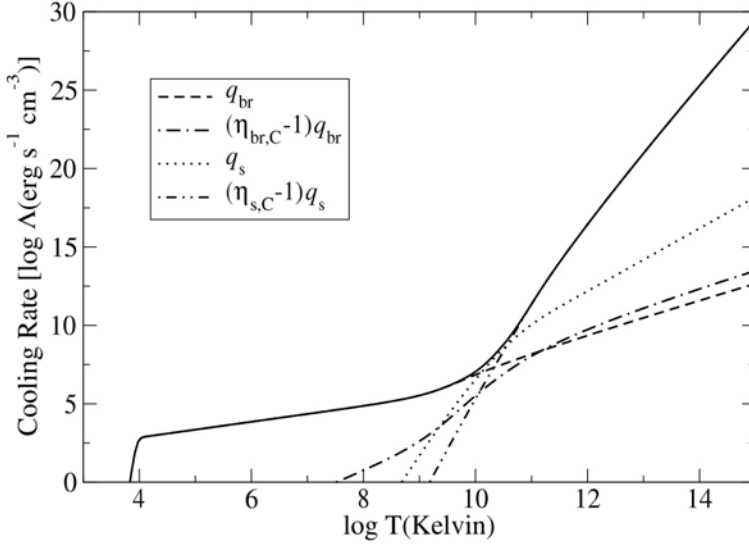


Fig. 9.4: Plot of a sample cooling function with $\rho = 10^{-10} \text{ g cm}^{-3}$, $B = 8380 \text{ G}$, and $H = 2.7 \times 10^7 \text{ cm}$. The total cooling function (*solid line*) and the following components are represented: Bremsstrahlung (*short-dashed*); Compton enhancement to Bremsstrahlung (*dot-long dash*); synchrotron (*dotted*); and the Compton enhancement to synchrotron (*dot-dot-dash*). Below $\sim 10^4 \text{ K}$ the free electron density drops off drastically and Bremsstrahlung is quenched. Reproduced from Fig. 1 of [355], by permission of the AAS.

tons will be cooler than the electrons, resulting in *raising* the energy of the photons and, therefore, cooling the hot, electron-scattering plasma in the process.

9.4.3.1 Total Compton Cooling

Since Comptonization does not emit photons itself, the total cooling including Comptonization is generally taken to be proportional to the base cooling rates of Bremsstrahlung and synchrotron (equations (9.82) and (9.87)) multiplied by their appropriate Compton-enhancement factor $\eta_C(\nu)$

$$\dot{q}^- = \eta_{\text{br},C} \dot{q}_{\text{br}}^- + \eta_{\text{s},C} \dot{q}_{\text{s}}^- \quad (9.94)$$

For the general Compton enhancement factor we use the expression from reference [353]

$$\eta_C(\nu) = e^y [1 - \gamma(j_m + 1, y + s)] + \eta_{C,\text{max}}(\nu) \gamma(j_m + 1, s) \quad (9.95)$$

where $\gamma(a, x) = [1/\Gamma(a)] \int_0^x t^{a-1} e^{-t} dt$ is the normalized lower incomplete gamma function (not the Lorentz factor γ) with parameters

$$A = 1 + 4\theta + 16\theta^2 \quad s = \tau_{\text{es}} + \tau_{\text{es}}^2$$

$$j_m = \frac{\ln[\eta_{C,\text{max}}(\nu)]}{\ln A} \quad \eta_{C,\text{max}}(\nu) = \frac{3kT_e}{h\nu}$$

The famous Compton parameter y is given by

$$y = s(A - 1) = 4(\theta_e + 4\theta_e^2)(\tau_{\text{es}} + \tau_{\text{es}}^2) \quad (9.96)$$

and τ_{es} is the electron scattering optical depth. Note that $\eta_C(\nu)$ is limited to a maximum value of $3kT/h\nu$, where h is Planck's constant. These expressions are most useful in numerical models and simulations.

In the Bremsstrahlung case $\eta_{\text{br},C}$ is found implicitly by integrating

$$\eta_{\text{br},C} \dot{q}_{\text{br}}^- = \int_{\nu_{\text{br}}}^{kT_e/h} \eta_C(\nu) \frac{d\dot{q}_{\text{br}}^-}{d\nu} d\nu \approx \int_{\nu_{\text{br}}}^{kT_e/h} \eta_C(\nu) \frac{\dot{q}_{\text{br}}^-}{(kT_e/h) - \nu_{\text{br}}} d\nu \quad (9.97)$$

over only the frequency range where Comptonization and Bremsstrahlung are important. (ν_{br} is the frequency where Bremsstrahlung becomes optically thick, and kT_e/h is that above which Bremsstrahlung falls off exponentially.) In the synchrotron case, the Compton enhancement factor is approximated by its value at the critical frequency below which the synchrotron emission becomes self-absorbed

$$\boxed{\eta_C(\nu) = \eta_C(\nu_c)} \quad (9.98)$$

Note: in reference [355], for example, while we implemented both enhancements, because synchrotron emission is dominant at temperatures where Comptonization becomes important, only $\eta_{s,C}$ was important in our simulations.

Fig. 9.4 shows a typical total cooling function plot as a function of temperature, assuming fixed values for ρ , H , and B . Note the temperature range over which each of the cooling processes is important. For $10^4 \text{ K} \lesssim T_e \lesssim 10^9 \text{ K}$, the dominant process is Bremsstrahlung. At higher temperatures Bremsstrahlung no longer dominates, even though its dependence with temperature steepens to be proportional to $T_e \log T_e$. Above 10^9 K the dominant cooling process is synchrotron, with Compton enhancement of synchrotron becoming important for temperatures somewhat above that. Comptonization of Bremsstrahlung, while included in this plot, is never particularly important.

9.4.3.2 The Emitted Comptonized Spectrum

After the initial emission of Bremsstrahlung or synchrotron photons, if the optical depth τ_{es} is high, these photons do not immediately leave the plasma. They remain within the hot medium, gaining energy from most of the inelastic scatterings they

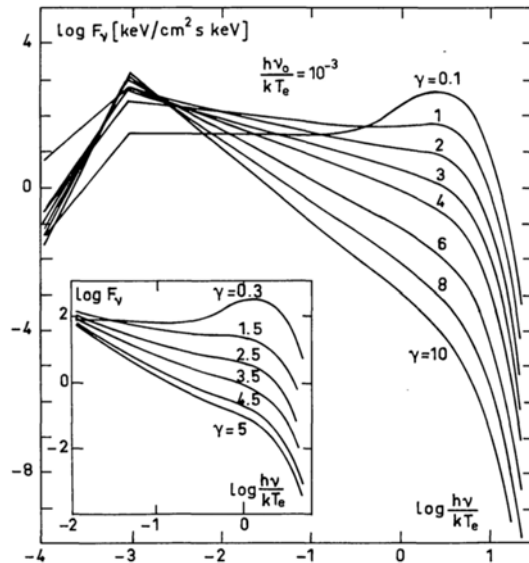


Fig. 9.5: Comptonized spectra for different values of Sunyaev and Titarchuk’s parameter $\gamma \approx y^{-1}$. A single-frequency photon spectrum of $\nu_0 = 10^{-3} kT_e/h$ is injected into the hot plasma. Here $\gamma = 0.1$ ($y \approx 10$) is Compton “thick” and produces a visible Wien peak; $\gamma = 10$ ($y \approx 0.1$) is Compton thin but still important, as it produces a power law spectrum up to the characteristic thermal frequency of kT_e/h . As long as the injected photons are much cooler than the hot electron scattering plasma, the nature of the output spectrum from Comptonization depends mainly on y , not on the nature of the input spectrum. Note that the axis labels are printed inside the plotting box, next to the axes. Credit: Fig. 5 of reference [356], reproduced with permission ©ESO.

encounter. Depending on how large the Compton y parameter is, this can drastically affect the final spectrum that emerges from the plasma. In 1980 Rashid A. Sunyaev and Lev G. Titarchuk, then at the Space Research Institute (IKI) in Moscow, studied the nature of this emerging spectrum in detail, for the case when the emitted photons are of much lower energy than the hot electrons themselves [356].

Fig. 9.5 shows a variety of normalized emergent spectra for different values of the y -parameter in this case. Photons are injected at the low-energy of $10^{-3} kT_e$, and their distribution is allowed to evolve until it comes into approximate equilibrium. Even for relatively “Compton thin” plasmas ($y \approx 0.1$), the emergent spectrum has significant photon energy at frequencies near the electron energy $h\nu \approx kT_e$. The power law shape of this ($\gamma = 10$) spectrum in Fig. 9.5 is similar to that seen in many binary X-ray sources and active galactic as well.

9.5 Useful Sets of Magnetohydrodynamic Equations for Solving Black Hole Astrophysical Problems

Despite the many approximations that were made in the derivation of the GRMHD equations (9.7), they still are enormously powerful and versatile, not to mention still extraordinarily complex. Here we present several different simplified sets of these, each of which represents more approximations made to the problem in order to make it more manageable. Each set of equations will be applicable in some situations but not in others. The trick, of course, is knowing what set is best for the job, without making too many more approximations and leaving out a crucial bit of physics that would have explained observed behavior of some black hole engines. Example sets include:

- *Adiabatic Ideal General Relativistic MHD (AIGRMHD)*: ignore processes like radiation, viscosity, and a complicated Ohm's law, but still consider flows in the strong gravity of a black hole
- *Force-Free Degenerate Electrodynamics (FFDE)*: ignore matter entirely and only treat the electromagnetic field, but still consider black hole gravity
- *Hydrodynamics (GRHD)*: ignore the electromagnetic field entirely and treat only the gas dynamics in black hole gravity
- *Special relativistic MHD (SRMHD)*: ignore gravity entirely but allow fast flow speeds, up to c
- *Non-Relativistic MHD (NRMHD)*: Newtonian MHD (as in Chapter 5), which is useful for accretion flows beyond, say, $10 r_S$ from the black hole

Below in this section we will treat these and others in turn. In that discussion we will not deal with the general evolving metrics of Einstein's field equations. At best we will consider only time-independent metrics like the Kerr black hole metric. And, sometime, we will leave general relativity behind entirely and discuss flow only in flat space, far from the black hole that created it.

9.5.1 Adiabatic Ideal Magnetohydrodynamics

9.5.1.1 Introduction

A large number of astrophysical simulations are adiabatic ideal magnetohydrodynamic ones, or a subset of ideal magnetohydrodynamics (including FFDE, GRHD, SRMHD, and NRMHD). The assumption that makes the simulation adiabatic is the use of an equation of state like (9.9), in which the local temperature and density are related, rather than independent. As shown in the last section, this still includes many types of gases. The main assumptions that make MHD ideal are that the time-dependent parts of Ohm's law (equation (9.33)), the Hall effect, and the resistivity of the plasma all can be set to zero. This reduces Ohm's law to the "frozen-in" condition (9.42). In the Kerr metric, using Komissarov's notation [322], Ohm's law can

be written as conditions on either the electric displacement or the electric field

$$\begin{aligned} \mathbf{D} &= -\frac{\mathbf{V}}{c} \times \mathbf{B} \\ \mathbf{E} &= -\alpha \left(\frac{\mathbf{V}}{c} + \boldsymbol{\beta} \right) \times \mathbf{B} \end{aligned}$$

where $\boldsymbol{\beta}$ is again the drift vector of the Kerr metric. When Ohm's law is substituted into $\mathcal{T}_{\text{EM}}$, we find that *the only field component needed to determine the evolution of an ideal MHD fluid is the magnetic induction $\mathbf{B}$* – hence the name “magnetohydrodynamics”.

Since Ohm's law is a relation between the magnetic induction and the electric field (and does not involve the current), we no longer can use it to determine the current in ideal MHD. Instead, we deduce what the current *should* be from Maxwell's inhomogeneous equation

$$\mathbf{J} = \frac{c}{4\pi} \nabla \cdot \mathcal{F} \quad (9.99)$$

In the stationary Kerr metric, these can be written as equations for the charge and current densities

$$\rho_q = \frac{1}{4\pi} \nabla \cdot \mathbf{D} \quad \mathbf{J} = \frac{c}{4\pi} \nabla \times \mathbf{H} - \frac{1}{4\pi} \frac{\partial \mathbf{D}}{\partial t}$$

With a *time* derivative on the right-hand side, this looks like it would be very difficult to determine the current in a computer simulation, since time derivatives usually need to be on the left-hand side. Fortunately, this turns out not to be a problem: the energy-momentum conservation law can be written in a way that involves *only* the electromagnetic field and not the charge or current densities. This is accomplished by computing the force using the stress-energy tensor form (equation (6.121)) rather than the Lorentz force form (equation (6.79)). Therefore, the current and charge densities (and magnetic field $\mathbf{H}$) need not be computed until *after* a simulation is finished.

9.5.1.2 The Adiabatic Ideal GRMHD Equations

As of this writing, there do not appear to be any scientific groups that have fully coupled general relativistic MHD and the evolution of Einstein's field equations into a giant simulation that forms a black hole *and* ejects a relativistic (possibly γ -ray burst) jet. However, at least two or three groups are working toward that end. On the other hand, there are several groups that have coupled GRMHD to a stationary Kerr black hole metric. These include, in approximate chronological order,

- Shinji Koide, formerly of Toyama University, Japan, and now at Kumamoto University; 1997 [357]
- Jean-Pierre De Villiers and John Hawley, the University of Virginia; 2003 [358]

- Charles Gammie, Jonathan McKinney and Gábor Tóth, University of Illinois; 2003 [359]
- Serguei Komissarov, Leeds University, UK; 2004 [360]
- Peter Anninos and Chris Fragile, and Jay Salmonson, Lawrence Livermore National Laboratory; 2005 [361]

These codes have been used to simulate disk-like accretion flows near black holes, as well as the ejection of magnetic jets by those disks. The equations of GRMHD used by these groups have several different forms, even though they solve the same physics. The reason for this wide variation is that each group has chosen to evaluate quantities like the density, velocity, and electromagnetic fields in different coordinate frames. Such is the freedom that general relativistic considerations offer. Writing down all these sets of equations would take many pages in this book and be of only cursory importance. Instead, we shall take one specific set and show how different fields of astrophysics arise from these equations. The reader may wish to look up some of the above authors' papers and compare the different approaches as an exercise.

The set of AIGRMHD equations that is most useful for an introductory discussion is that published by Koide [321]; we give them below using Komissarov's electrodynamics notation. The set of variables that forms the solution to these equations is called the *primitive variables*, and consists of the rest mass density, ρ , fluid three-velocity $\mathbf{V}$, pressure p , and magnetic induction $\mathbf{B}$. The full set of Koide's MHD equations that needs to be evolved in order to obtain the solution for the primitive variables is as follows: conservation of mass, momentum, and energy

$$\boxed{\frac{\partial \rho}{\partial t} = -\nabla \cdot [\alpha \varrho (\mathbf{V} + \beta c)]} \quad (9.100)$$

$$\boxed{\frac{\partial \mathfrak{P}}{\partial t} = -\nabla \cdot [\alpha (\mathbf{T} + \mathfrak{P} \beta c)] - \left(\varrho + \frac{\mathfrak{E}}{c^2} \right) c^2 \nabla \alpha - \mathfrak{P} \cdot \boldsymbol{\sigma}^{\text{KER}}} \quad (9.101)$$

$$\boxed{\frac{\partial \mathfrak{E}}{\partial t} = -\nabla \cdot [\alpha (c^2 \mathfrak{P} - \varrho c^2 \mathbf{V} + \mathfrak{E} \beta c)] - c^2 \mathfrak{P} \cdot \nabla \alpha - \mathbf{T} : \boldsymbol{\sigma}^{\text{KER}}} \quad (9.102)$$

(note that the divergence of the dyadic $\mathfrak{P}\beta$ forms an inner product with components of β , not $\mathfrak{P}$), along with Maxwell's homogeneous equations and Ohm's law

$$\boxed{\frac{\partial \mathbf{B}}{\partial t} = -c \nabla \times \mathbf{E}} \quad (9.103)$$

$$\boxed{\nabla \cdot \mathbf{B} = 0} \quad (9.104)$$

$$\boxed{\mathbf{E} = -\alpha \left(\frac{\mathbf{V}}{c} + \beta \right) \times \mathbf{B}} \quad (9.105)$$

The conserved variables are the densities of relativistic mass, total momentum, and total energy (minus rest mass energy)

$$\varrho = \gamma \rho \quad (9.106)$$

$$\mathfrak{P} = \gamma^2 \left(\rho + \frac{h}{c^2} \right) \mathbf{V} + \frac{1}{4\pi c} \mathbf{D} \times \mathbf{B} \quad (9.107)$$

$$\mathfrak{E} = \gamma(\gamma - 1) \rho c^2 + \gamma^2 h - p + \frac{1}{8\pi} (B^2 + D^2) \quad (9.108)$$

(written in the Fraktur font to emphasize their per-unit-volume nature) and the additional variables needed to compute them are the Lorentz factor, enthalpy, electric displacement, and equation of state

$$\gamma = \left(1 - \frac{\mathbf{V} \cdot \mathbf{V}}{c^2} \right)^{-1/2} \quad (9.109)$$

$$h = \frac{\Gamma}{\Gamma - 1} p \quad (9.110)$$

$$\mathbf{D} = \frac{1}{\alpha} \mathbf{E} + \boldsymbol{\beta} \times \mathbf{B} = -\frac{\mathbf{V}}{c} \times \mathbf{B} \quad (9.111)$$

$$p = K_{\Gamma} \rho^{\Gamma} \quad (9.112)$$

There also are some additional quantities that need to be computed in order to evolve the conserved variables: the three-dimensional stress and metric-shear tensors

$$\mathbf{T} = \left[p + \frac{1}{8\pi} (B^2 + D^2) \right] \mathbf{I} + \gamma^2 \left(\rho + \frac{h}{c^2} \right) \mathbf{V} \mathbf{V} - \frac{1}{4\pi} (\mathbf{B} \mathbf{B} + \mathbf{D} \mathbf{D}) \quad (9.113)$$

$$\boldsymbol{\sigma}^{\text{KER}} = g_{33}^{1/2} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1}{g_{11}^{1/2}} \frac{\partial \omega}{\partial x^1} & \frac{1}{g_{22}^{1/2}} \frac{\partial \omega}{\partial x^2} & 0 \end{pmatrix} \quad (9.114)$$

Finally, we have equations for the magnetic field and charge/current densities that can be used as *a posteriori* diagnostics

$$\mathbf{H} = \alpha (\mathbf{B} + \boldsymbol{\beta} \times \mathbf{D}) \quad (9.115)$$

$$\mathbf{J} = \frac{1}{4\pi} \left(c \nabla \times \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} \right) \quad (9.116)$$

$$\rho_q = \frac{c}{4\pi} \nabla \cdot \mathbf{D} \quad (9.117)$$

This is the most complex set of equations that we will encounter in this book, so it is important to discuss them a bit and note their main features before going on:

- The time and spatial derivatives ($\partial/\partial t$ and ∇) are performed in Boyer–Lindquist coordinates, using the 3+1 metric (equations (7.62) to (7.66)).

- However, the primitive and conserved variables generally are *not* evaluated in the Boyer–Lindquist frame. In fact, they are evaluated in a variety of frames, each chosen to make the equations look as similar to the Newtonian ones in Chapter 5 as possible:
 - orthonormal frame moving with the *fluid* and viewed locally near the black hole; scalars ρ , $\mathfrak{E}$, p , *etc.*
 - orthonormal frame moving with the *metric rotation* but not with the fluid, also viewed locally: scalars ϱ , ρ_q , γ ; vectors $\mathbf{V}$, $\mathfrak{P}$, $\mathbf{D}$, and $\mathbf{B}$; and tensors $\mathbf{T}$ and σ^{KER} . Note that $\mathbf{V}$ is a physical velocity (remains below c) and is used to construct the Lorentz factor γ .
 - orthonormal frame tied to Boyer–Lindquist coordinates and viewed from infinity; vectors $\mathbf{J}$, $\mathbf{E}$, and $\mathbf{H}$.
- The conservation laws (9.100) to (9.102) take into account motion by the fluid ($\mathbf{V}$) and by the metric (β), as well as time dilation near the black hole (α). They are written in the “conservative” form, with the time derivative of a conserved quantity on the left-hand side and the divergence of a vector or tensor (the “flux” of the conserved variable) on the right-hand side, along with only external forces.
- $-c^2 \nabla \alpha$ is the gravitational acceleration, and $-(\varrho + \mathfrak{E}/c^2) c^2 \nabla \alpha$ is the gravitational force density in the fluid equation of motion (9.101). This term reduces to $-\rho \nabla \psi$ in the non-relativistic limit.
- The $-c^2 \mathfrak{P} \cdot \nabla \alpha$ term in the energy equation (9.102) represents work done by, or on, the plasma as it plasma climbs out of, or falls into, the black hole’s gravitational potential well.
- Motion and stress across the metric shear σ^{KER} can transfer momentum and energy, respectively, in a manner similar to viscous shear.
- Pseudo forces, such as centrifugal and Coriolis forces, due to motion of the fluid and frame dragging of the Kerr metric are included implicitly in the tensor divergence in equation (9.101).
- Equation (9.104) is a *constraint* on the magnetic induction. If it is not enforced during a simulation, the results could become unphysical. Because it is so important, a variety of techniques have been developed to make sure that the magnetic induction remains divergence-free, even in the curved geometry of the black hole.

Equations (9.100) to (9.117) are the practical realization of the more formal and esoteric ones in (9.6), (9.7), (7.23), (7.24), and (9.42), and, unlike the latter, they explicitly show many of the effects of the black hole geometry. These equations allow the flow of magnetized plasma in the vicinity of black holes to be simulated in detail. There are, of course, other equivalent sets of equations that accomplish the same purpose.⁷ But, regardless of the implementation details, the procedure for simulating the plasma flow is much the same, and takes the following steps:

⁷ Gammie & McKinney and Komissarov, for example, use the four-velocity $\mathbf{U}$ (with normalization $\mathbf{U}^2 = -c^2$), rather than the three-velocity $\mathbf{V}$, and horizon-penetrating (not Boyer–Lindquist) coordinates. This removes the limit on $V < c$ ($\mathbf{U}$ can approach infinity) and allows the simulation to follow the flow into the black hole.

1. Simulations are performed in a manner very similar to those used to do weather prediction on the earth:
 - Three-dimensional space is divided into small cells, each of which can have a different value for the primitive variables.
 - The evolution of those variables through time is computed by taking a series of time steps, re-computing ρ , $\mathbf{V}$, p , and $\mathbf{B}$ at each time during using the above evolution equations.
2. The simulation begins with an initial model in which the primitive variables are known. These allow one to calculate both the initial values of the conserved variables ϱ_0 , $\mathfrak{P}_0$, and $\mathfrak{E}_0$ from equations (9.106) to (9.108) *and also the time derivatives of these (and $\mathbf{B}$) from equations (9.100) to (9.103).*
3. An update of the conserved variables to the next time step (ϱ_1 , $\mathfrak{P}_1$, $\mathfrak{E}_1$, and $\mathbf{B}_1$) is accomplished by computing, for example,

$$\varrho_1 = \varrho_0 - \Delta t \nabla \cdot [\alpha \varrho_0 (\mathbf{V}_0 + \beta_0 c)]$$

where Δt is the length of the time step chosen. This gives the conserved variables at time $t_0 + \Delta t$, but not the primitive variables (which are needed for the next update).

4. Finally, the new primitive variables (ρ_1 , $\mathbf{V}_1$, p_1 , and $\mathbf{B}_1$) are determined by inverting the nonlinear algebraic equations (9.106) to (9.115), *in each computational cell.*
5. Having new values for the primitive variables, we now can cycle back to step #2 and execute another time-step update.

Typical simulations take many tens to hundreds of thousands of small time steps or more. For relativistic flow, the primitive variable solution in each cell alone can take a large amount of time. And, if any one of the hundreds of thousands of cells fails to find a solution to the nonlinear primitive variable equations in one of the hundreds of thousands of time steps taken, the entire simulation halts or “crashes”.

Those who perform GRMHD simulations have overcome such difficulties and have used these large numerical codes to study black hole accretion and jet production since about the year 1997. And the field is continually growing with new codes being developed.

9.5.2 Standard Force-Free Electrodynamics

By definition, Maxwell’s equations are the ones of electrodynamics. And, by their very nature, time-varying electromagnetic fields are relativistic.⁸ Maxwell’s equations in the Kerr metric can be made to look like the equations of electrodynamics

⁸ Of course, these equations also give rise to the studies of electrostatics and magnetostatics, which are non-relativistic in a sense. But these are not of interest here.

in continuous media (equations (7.74) and (7.75)), using Komissarov’s “3+1” notation. Alternatively, by using equations (6.117) and (6.119), Maxwell’s equations can be written in a way that makes them look like the evolution of electromagnetic field energy and momentum on the left-hand side and Lorentz force on the right

$$\nabla \cdot \mathcal{T}_{\text{EM}} = -\frac{1}{c} \mathbf{J} \cdot \mathcal{F} \quad (9.118)$$

However, it is important to realize that equations (9.118) are *not* energy-momentum conservation laws in the classical sense. They are Maxwell’s inhomogeneous equations in disguise, and are fully valid whether we are dealing with matter or not.

9.5.2.1 The Equations of Standard Force-Free Degenerate Electrodynamics (FFDE)

The above point about the generality of equations (9.118) now becomes important, because only now do we actually consider the energy-momentum equation (9.7). In *force-free electrodynamics* we make the approximation that all forces, except electromagnetic ones, are negligible:

$$\mathcal{T}_{\text{GAS}} = \mathcal{T}_{\text{RAD}} = 0$$

so equation (9.7) becomes $\nabla \cdot \mathcal{T}_{\text{EM}} = 0$ or, from equation (9.118)

$$\mathbf{J} \cdot \mathcal{F} = 0 \quad (9.119)$$

which is called the “force-free condition”. In stationary metrics like the Kerr one, the force-free condition can be written in the “3+1” language as

$$\mathbf{J} \cdot \mathbf{E} = 0 \quad (9.120)$$

$$\rho_q \mathbf{E} + \frac{1}{c} \mathbf{J} \times \mathbf{B} = 0 \quad (9.121)$$

Note that equation (9.121) implies that the electromagnetic field is *degenerate*. That is,

$$\mathbf{E} \cdot \mathbf{B} = 0$$

and, by virtue of the relation between $\mathbf{E}$ and $\mathbf{D}$ (equation (9.111))

$$\mathbf{D} \cdot \mathbf{B} = 0$$

also.

This now gives us enough information to calculate the three-current $\mathbf{J}$ from the force-free condition and from Maxwell’s equations ((9.103), (9.104), (9.116), and (9.117)). Taking the *cross* product of the spatial force-free condition (9.121) with $\mathbf{B}$, substituting Maxwell’s equations, and realizing that $\partial(\mathbf{D} \cdot \mathbf{B})/\partial t = 0$, we find

that [362]

$$\mathbf{J} = \frac{c}{4\pi |\mathbf{B}|^2} [(\mathbf{B} \cdot \nabla \times \mathbf{H} - \mathbf{D} \cdot \nabla \times \mathbf{E}) \mathbf{B} + (\nabla \cdot \mathbf{D}) (\mathbf{E} \times \mathbf{B})] \quad (9.122)$$

The first term on the right-hand side is the current density parallel to the magnetic field ($\mathbf{J}_{\parallel}$) while the second term is that normal to the field ($\mathbf{J}_{\perp}$, the so-called “drift current”).

The complete set of force-free degenerate electrodynamics equations in Kerr spacetime, then, are the *time-dependent* Maxwell’s equations (repeated below)

$$\frac{\partial \mathbf{B}}{\partial t} = -c \nabla \times \mathbf{E} \quad \frac{\partial \mathbf{D}}{\partial t} = c \nabla \times \mathbf{H} - 4\pi \mathbf{J} \quad (9.123)$$

(which evolve $\mathbf{B}$ and $\mathbf{D}$), the relations between electric field and displacement and between the magnetic field and induction

$$\mathbf{E} = \alpha (\mathbf{D} - \boldsymbol{\beta} \times \mathbf{B}) \quad \mathbf{H} = \alpha (\mathbf{B} + \boldsymbol{\beta} \times \mathbf{D}) \quad (9.124)$$

(which give $\mathbf{E}$ and $\mathbf{H}$ in terms of known quantities), and expression (9.122) above for the current (which is also needed in the inhomogeneous equation (9.123)). When evolving equations (9.123) on a computer, Maxwell’s time-independent equations for $\mathbf{B}$ and $\mathbf{D}$ are applied as natural constraints on the magnetic induction and electric displacement, completing the set.

It is important to realize in all of these “3+1” equations that the vector operations must be performed using the three-metric of the Kerr geometry (equation (7.64)). Note, however, that this is a diagonal three-metric, at least, so the normal curvilinear formulae for divergence and curl can be used. In flat spacetime ($\alpha \rightarrow 1$ and $\boldsymbol{\beta} \rightarrow 0$) these simplify somewhat, since $\mathbf{D} \rightarrow \mathbf{E}$ and $\mathbf{H} \rightarrow \mathbf{B}$; and the vector operations assume their normal flat-space (*e.g.*, Cartesian) forms.

9.5.2.2 Difficulties with FFDE: The Need for Ohm’s Law

Simulations of pulsar and black hole magnetospheres that rely on sets of equations like (9.122) to (9.124) often encounter problems. As long as $B^2 - D^2 > 0$ the simulation proceeds reasonably smoothly and stably. However, when this condition is not satisfied, an exponential instability grows rapidly and destroys the simulation. This problem arises especially in “current sheets” – flat areas in space where the magnetic field points in one direction above the sheet and in a significantly different (often opposite) direction below the sheet. This produces a significant $\nabla \times \mathbf{H}$ and, by Ampère’s law, a large current in the sheet. A number of simplifying assumptions break down in current sheets, including those of force-free electrodynamics.

We shall call the condition $B^2 - D^2 > 0$ the *magnetospheric condition*, because it implies that a reference frame can be found where the electric displacement is zero,

leaving only magnetic field. Now, the equations of FFDE do not explicitly enforce the magnetospheric condition; this is why numerical evolution of those equations can often violate that condition, even when the initial conditions satisfy it. One method of explicitly enforcing the magnetospheric condition is to apply the ideal Ohm's law (9.42) as an additional constraint

$$\mathbf{U} \cdot \mathcal{F} = 0$$

or, in 3+1 language,

$$\mathbf{D} = -\frac{\mathbf{V}}{c} \times \mathbf{B}$$

and, alternately,

$$\mathbf{E} = -\alpha \left(\frac{\mathbf{V}}{c} + \boldsymbol{\beta} \right) \times \mathbf{B}$$

Since $D^2 = \frac{V^2}{c^2} \sin^2 \theta B^2$ (where θ is the angle between $\mathbf{B}$ and $\mathbf{V}$), then $D^2 < B^2$. (Whether this satisfies $D^2 < B^2$ in general is another issue and will be discussed below.)

However, if electrodynamics purports to ignore matter, then how do we interpret the matter velocity $\mathbf{V}$? First of all, we should realize that FFDE actually does not ignore the matter entirely, only the material *forces*. For example, we still have a charge density ρ_q and current density $\mathbf{J}$, which are responsible for generating the fields. Let us consider the frame where $\mathbf{V} = 0$ first. In this frame $\mathbf{D} = 0$ and we have only magnetic field – a stationary magnetosphere – with charges trapped in Larmor radii encircling the field lines. On the other hand, if we consider a frame in which $\mathbf{V} \neq 0$, we can divide the velocity into a part parallel to the magnetic field and one perpendicular to it. Now $\mathbf{V}_{\parallel} \times \mathbf{B} = 0$, so we can ignore the parallel part in equation (9.127). And, because the ideal Ohm's law states that matter and magnetic field are frozen together, $\mathbf{V}_{\perp} = \mathbf{V}_f$, where the latter is the *velocity of the magnetic field line*.

Now, it is customary in pulsar and black hole electrodynamic studies to measure the (angular) velocity of the magnetic field relative to an observer sitting at infinity. Therefore, we actually will take $\mathbf{V}_f$ to be measured in this frame (which will include the rotation of the metric and the time dilation)

$$\mathbf{V}_f \equiv \alpha (\mathbf{V}_{\perp} + \boldsymbol{\beta} c) \quad (9.125)$$

rather than defining it to be just $\mathbf{V}_{\perp}$. With this definition, the relations for $\mathbf{D}$ and $\mathbf{E}$ become

$$\mathbf{D} = -\frac{1}{\alpha c} (\mathbf{V}_f - \alpha \boldsymbol{\beta} c) \times \mathbf{B} \quad (9.126)$$

$$\mathbf{E} = -\frac{1}{c} \mathbf{V}_f \times \mathbf{B} \quad (9.127)$$

Because equation (9.126) refers to a magnetic field velocity, rather than a matter velocity, it also can be considered independent of the matter.

Will this ideal Ohm's law always satisfy the magnetosphere condition; that is, will we always have $D^2 < B^2$? The answer is *no*, at least not in the ergosphere. Except in the special case when $\mathbf{V}_f = \alpha\beta c$, the right-hand side of equation (9.126) will be non-zero. Then, a small lapse function always will guarantee a large coefficient of $|\mathbf{B}|$ on the right-hand side. So the force-free assumption cannot hold in general in ergospheres near black holes (where α is small and $|\beta| > 1$).

Introducing the magnetic field velocity also has some drawbacks. Consider the following modified version of FFDE, with Ohm's law applied to Faraday's law:

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{V}_f \cdot \nabla \mathbf{B} = \mathbf{B} \cdot \nabla \mathbf{V}_f - (\nabla \cdot \mathbf{V}_f) \mathbf{B} \quad (9.128)$$

This is the only equation that needs to be evolved; $\mathbf{E}$, $\mathbf{D}$, and $\mathbf{H}$ can be determined from Ohm's law or algebraic relations between the electromagnetic variables. However, in FFDE we have no equation to determine $\mathbf{V}_f$. We must *specify* the velocity for every magnetic field line and at every point in space. This is potentially an enormous number of free parameters in a general problem and, as such, would call any such solution into question. The use of a magnetic field velocity in electrodynamics (*i.e.*, without performing actual MHD), therefore, is usually limited to rather simple astrophysical applications. The main one is stationary electrodynamics – the study of rotating magnetospheres.

9.5.2.3 Stationary and Axisymmetric FFDE: The Grad–Schlüter–Shafranov Equation

Rotating objects like black holes and pulsars are often stationary ($\partial/\partial t = 0$) and axisymmetric ($\partial/\partial\phi = 0$), so there are a large number of systems that we can treat by assuming stationarity and axisymmetry. These assumptions lead to a single equation for the two-dimensional distribution of the electromagnetic fields in the (R, Z) -plane, which is the generalization of a famous equation derived to study the structure of laboratory magnetic fields by H. Grad, A. Schlüter, and Vitalii Dmitrievich Shafranov in the late 1950s [363, 364, 365]. Appendix E derives this equation for the Kerr metric from the equations of FFDE in the previous subsections, and from the application of Ohm's law. We reproduce it here

$$\boxed{\nabla \cdot \left\{ \frac{\alpha}{R^2} \left[1 - \frac{R^2}{\alpha^2 c^2} (\Omega_f - \omega)^2 \right] \nabla \Psi \right\} + \frac{(\Omega_f - \omega)}{\alpha c^2} \frac{d\Omega_f}{d\Psi} (\nabla \Psi)^2 + \frac{1}{2\alpha R^2 c^2} \frac{dI^2}{d\Psi} = 0} \quad (9.129)$$

where ω , $R = \Sigma \sin\theta/\rho$, and α are the angular rotation rate, cylindrical radial coordinate, and lapse function of the Kerr metric (see Sections 7.5 and 7.7);

$$\Psi(r, \theta) = \frac{1}{2\pi} \int \mathbf{B} \cdot d\mathbf{S} = R A_\phi \quad (9.130)$$

is the magnetic flux function (the “solution” to the problem, which we discuss below); ∇ is the two-gradient in the (R, Z) -plane; $\Omega_f = V_f/R$ is the *angular velocity* of the magnetic field (which *could* vary from field line to field line, *i.e.*, $\Omega_f = \Omega_f(\Psi)$); and $I = I(\Psi)$ is the poloidal current integrated up to a given field line Ψ . The GSS equation is a “cross-field” equation, which determines the relationship between adjacent magnetic field lines. It is used to construct steady, axisymmetric magnetospheres of a variety of objects.

The flux function Ψ is a scalar potential that allows us to determine the electric field

$$\mathbf{E} = -\frac{\Omega_f}{c} \nabla \Psi \quad (9.131)$$

and the magnetic induction

$$\mathbf{B} = \frac{1}{R} \nabla \Psi \times \mathbf{e}_{\hat{\phi}} - \frac{I}{\alpha R c} \mathbf{e}_{\hat{\phi}} \quad (9.132)$$

as well as the electric displacement

$$\mathbf{D} = -\frac{(\Omega_f - \omega)}{\alpha c} \nabla \Psi = \frac{(\Omega_f - \omega)}{\alpha \Omega_f} \mathbf{E} \quad (9.133)$$

and the magnetic field

$$\begin{aligned} \mathbf{H} &= \frac{\alpha}{R} \left[1 + \frac{R^2}{\alpha^2 c^2} \omega (\Omega_f - \omega) \right] \nabla \Psi \times \mathbf{e}_{\hat{\phi}} - \frac{I}{R c} \mathbf{e}_{\hat{\phi}} \\ &= \alpha \left\{ \left[1 + \frac{R^2}{\alpha^2 c^2} \omega (\Omega_f - \omega) \right] \mathbf{B}_p + \mathbf{B}_\phi \right\} \end{aligned} \quad (9.134)$$

where $\mathbf{B}_p = B_R \mathbf{e}_{\hat{R}} + B_Z \mathbf{e}_{\hat{Z}}$ and $\mathbf{B}_\phi$ are the poloidal and toroidal components of the magnetic induction, respectively. The current density can be decomposed into a component that depends on the total current parallel to the magnetic field plus a component representing charge dragged around by the rotation of that field

$$\mathbf{J} = -\frac{1}{4\pi} \frac{dI}{d\Psi} \mathbf{B} + \rho_q \Omega_f R \mathbf{e}_{\hat{\phi}} \quad (9.135)$$

Comparison of equation (9.135) with (9.122) shows that the steady-state assumption comes at a price. The perpendicular part of the current can be determined from the equation for $\mathbf{D}$ and from the rotation rate. However, while in regular FFDE we obtained the parallel current from the time-varying electromagnetic fields ($\nabla \times \mathbf{H}$, $\nabla \times \mathbf{E}$, and Maxwell’s equations), we now must specify the total current along $\mathbf{B}$ (*i.e.*, I becomes another free parameter). Making the function $I(\Psi)$ arbitrary has caused many problems and headaches in trying to solve the structure of pulsar

magnetospheres. These problems were solved only recently (since 1999) when sophisticated methods were developed to determine $I(\Psi)$ as part of a time-dependent FFDE solution.

9.5.2.4 The Pulsar Equation

The standard equation used to solve the structure of *pulsar* magnetospheres is a further simplification of equation (9.129) to flat space. We set the Kerr parameters $\omega = 0$ and $\alpha = 1$, assume that all field lines rotate at the same rate as the pulsar ($d\Omega_f/d\Psi = 0$), and replace R and Z with

$$\begin{aligned} x &\equiv \frac{R}{R_L} \\ y &\equiv \frac{Z}{R_L} \end{aligned}$$

where the *light cylinder radius* is given by

$$R_L = \frac{c}{\Omega_f} \quad (9.136)$$

Equation (9.129) then becomes the well-known *pulsar equation*

$$\boxed{(1 - x^2) \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} \right) - \frac{1 + x^2}{x} \frac{\partial \Psi}{\partial x} = -R_L^2 I \frac{dI}{d\Psi}} \quad (9.137)$$

which has one singular surface at $x = 1$ ($R = R_L$). Note that, when $x \ll 1$ (the non-relativistic limit), the pulsar equation reduces to the standard laboratory Grad–Schlüter–Shafranov equation for tokamaks (without plasma pressure)

$$x \frac{\partial}{\partial x} \left(\frac{1}{x} \frac{\partial \Psi}{\partial x} \right) + \frac{\partial^2 \Psi}{\partial y^2} = -R_0^2 I \frac{dI}{d\Psi}$$

where we have switched the scale length from R_L to the size R_0 of the laboratory equipment being studied.

9.5.3 An Alternate Form for FFDE: GRMHD with No Material Forces

A quite different, but equally effective, set of equations for force-free electrodynamics can be obtained by applying the force-free condition (9.119) to the right-hand side of equation (9.118) and forming the relativistic conservation laws for the electromagnetic field. This is equivalent to beginning with the equations for GRMHD

in Subsection 9.5.1 and then setting all terms dealing with the plasma (ρ , p , *etc.*) to zero. The result is

$$\boxed{\frac{\partial \mathfrak{P}}{\partial t} = -\nabla \cdot [\alpha (\mathbf{T} + \mathfrak{P} \beta c)] - \mathfrak{E} \nabla \alpha - \mathfrak{P} \cdot \boldsymbol{\sigma}^{\text{KER}}} \quad (9.138)$$

$$\boxed{\frac{\partial \mathfrak{E}}{\partial t} = -\nabla \cdot [\alpha (c^2 \mathfrak{P} + \mathfrak{E} \beta c)] - c^2 \mathfrak{P} \cdot \nabla \alpha - \mathbf{T} : \boldsymbol{\sigma}^{\text{KER}}} \quad (9.139)$$

where the field momentum and internal energy densities, and three-dimensional electromagnetic stress tensor, are given by

$$\begin{aligned} \mathfrak{P} &= \frac{1}{4\pi c} (\mathbf{D} \times \mathbf{B}) \\ \mathfrak{E} &= \frac{1}{8\pi} (B^2 + D^2) \\ \mathbf{T} &= \frac{1}{8\pi} (B^2 + D^2) \mathbf{I} - \frac{1}{4\pi} (\mathbf{B}\mathbf{B} + \mathbf{D}\mathbf{D}) \end{aligned}$$

(with $\mathbf{I}$ again being the three-dimensional identity tensor). Equations (9.138) and (9.139) are now true conservation laws for the electromagnetic field, under the assumption of force-free electrodynamics.

In addition to the conservation laws, we still will need Maxwell's *homogeneous* equations to evolve the magnetic induction $\mathbf{B}$, along with an equation to give us the electric displacement $\mathbf{D}$. Finally, as discussed above, *even in FFDE* we will need Ohm's law. This adds the following to complete the set

$$\begin{aligned} \frac{\partial \mathbf{B}}{\partial t} &= -c \nabla \times \mathbf{E} \\ \nabla \cdot \mathbf{B} &= 0 \\ \mathbf{E} &= -\frac{1}{c} \mathbf{V}_f \times \mathbf{B} \\ \mathbf{D} &= \frac{1}{\alpha} \mathbf{E} + \beta \times \mathbf{B} = -\frac{1}{\alpha c} (\mathbf{V}_f - \alpha \beta c) \times \mathbf{B} \end{aligned}$$

By comparing with the equations in Subsection 9.5.1, we arrive at a very important conclusion: *proper force-free electrodynamics (with Ohm's law added) is simply a limiting case of general relativistic magnetohydrodynamics, with the matter forces ignored.*

It would be highly advantageous, therefore, if GRMHD codes could be developed that could operate not only in the regime where the matter forces are important, but also when they become negligible. In that way both MHD plasma flow *and* the electrodynamics of magnetospheres could be treated at the same time, should they occur simultaneously in different parts of a simulation. Such a code also would naturally and self-consistently determine the magnetic field velocity $\mathbf{V}_f$ from the MHD equations, without the need to specify it as a set of free parameters.

9.5.4 General Relativistic Hydrodynamics (GRHD)

Many investigators often ignore the electromagnetic field entirely. This is the opposite extreme to FFDE. GRHD uses the same general relativistic fluid equations (9.100) to (9.102) with hydrodynamic versions of the momentum/energy densities and stress tensor:

$$\begin{aligned}\mathfrak{P} &= \gamma^2 \left(\rho + \frac{h}{c^2} \right) V \\ \mathfrak{E} &= \gamma(\gamma - 1) \rho c^2 + \gamma^2 h - p \\ \mathbf{T} &= p \mathbf{I} + \gamma^2 \left(\rho + \frac{h}{c^2} \right) \mathbf{V} \mathbf{V}\end{aligned}$$

Because GRMHD codes emphasize plasma flow, virtually all of these can be run in a GRHD mode, simply by setting $\mathbf{D} = \mathbf{E} = \mathbf{B} = \mathbf{H} = 0$ in the simulation.

9.5.5 Special Relativistic Magnetohydrodynamics (SRMHD)

As shown in Chapters 2 and 3, even very far from the strong gravity of a black hole, jets travel at relativistic speeds. Therefore, there is a need for equations that may not contain the curved black hole metric, but still can handle plasma flow at speeds near that of light. Many groups around the world now study the special relativistic magnetohydrodynamics (or hydrodynamics) of jet flows, and the variation amongst their approaches to SRMHD is much less than in the case of GRMHD. One of the most popular sets of equations is found by simply reducing the Kerr geometric terms to Minkowski ones ($\alpha = 1, \beta = 0$) in the above GRMHD equations (9.100) to (9.105). The evolution equations are much simpler (and in a fully conservative form, with no external forces on the right-hand sides)

$$\boxed{\frac{\partial \varrho}{\partial t} = -\nabla \cdot (\varrho \mathbf{V})} \quad (9.140)$$

$$\boxed{\frac{\partial \mathfrak{P}}{\partial t} = -\nabla \cdot \mathbf{T}} \quad (9.141)$$

$$\boxed{\frac{\partial \mathfrak{E}}{\partial t} = -\nabla \cdot (c^2 \mathfrak{P} - \varrho c^2 \mathbf{V})} \quad (9.142)$$

$$\boxed{\frac{\partial \mathbf{B}}{\partial t} = -c \nabla \times \mathbf{E}} \quad (9.143)$$

$$\boxed{\nabla \cdot \mathbf{B} = 0} \quad (9.144)$$

$$\boxed{\mathbf{D} = \mathbf{E} = -\frac{\mathbf{V}}{c} \times \mathbf{B}} \quad (9.145)$$

But, because the conserved variables and three-dimensional stress tensor $\mathbf{T}$ were defined in a local orthonormal frame, all quantities in the above equations still are exactly the same ones in equations (9.106) to (9.113), except that the electric displacement $\mathbf{D}$ is now replaced by the electric field $\mathbf{E}$. Finally, the inhomogeneous Maxwell (diagnostic) equations now reflect that $\mathbf{H} = \mathbf{B}$ also

$$\mathbf{J} = \frac{1}{4\pi} \left(c \nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} \right) \quad (9.146)$$

$$\rho_q = \frac{c}{4\pi} \nabla \cdot \mathbf{E} \quad (9.147)$$

Chapter 15 deals with jet acceleration and propagation, and makes extensive use of the SRMHD equations.

9.5.6 Stationary, Axisymmetric SRMHD in Newtonian Gravity

In the discussion of electrodynamics we learned that a popular way to study black hole and pulsar magnetospheres (and tokamaks in the laboratory) is to simplify the equations further by assuming that the magnetosphere structure is stationary ($\partial/\partial t = 0$) and axisymmetric ($\partial/\partial\phi = 0$). This casts all fields in terms of a single flux function Ψ and gives us only one partial differential equation to solve instead of many. The resulting GSS equation describes the force balance *across* magnetic field lines; everything else (fields, currents, *etc.*) follows from that one solution.

The same trick (assuming stationarity and axisymmetry) is also very useful for MHD problems. However, because the plasma flow now is included, the equations are somewhat more complex than in FFDE. Nevertheless, when these assumptions are applied to magnetohydrodynamic problems, they *do* reduce the large set of equations to only *two* coupled ones:

- an equation *along* the magnetic field lines that describes how the plasma accelerates or decelerates, and
- a cross-field equation that can be considered to be a generalized GSS force balance equation.

In electrodynamics we ignored the matter, so we did not need the first equation. However, plasma flow is so crucial to MHD that an equation *parallel* to the field lines now becomes as important as one perpendicular to them. Below we discuss the main aspects of ideal, relativistic, stationary, axisymmetric MHD in a Newtonian gravitational field and the two main equations that govern it.

9.5.6.1 The Equation(s) along the Magnetic Field

Appendix F derives the basic properties of this simplified form of MHD. Note that we still can speak of a flux function (equation (9.130) above), and the magnetic and

electric fields still can be derived from that potential in the same manner as was done in Section 9.5.2. However, the equations now give us a solution for the current along the magnetic field (I) and, hence, also $B_{\hat{\phi}}$. Some authors choose to not use Ψ to represent the solution at all, instead opting to express the equations in terms of the poloidal velocity $\mathbf{V}_p = V_R \mathbf{e}_{\hat{R}} + V_Z \mathbf{e}_{\hat{Z}}$ or a dimensionless Mach number.

The stationary and axisymmetric assumptions imply that several physical quantities are conserved along each field line, although the value of that quantity can differ from field line to field line. Three of these conserved quantities are:

- Ω_f – the field angular velocity. As in FFDE each field line rotates like a stiff wire, only now the shape of that wire is determined by plasma forces as well as electromagnetic.
- $\mathbf{B} \cdot d\mathbf{S}_p$ – the local magnetic flux.
- $4\pi \rho \gamma \mathbf{V} \cdot d\mathbf{S}_p$ – the local mass flux, where $\gamma = (1 - \mathbf{V} \cdot \mathbf{V}/c^2)^{-1/2}$ is the standard Lorentz factor.

where $d\mathbf{S}_p$ is a small poloidal area vector. Together these three conservation laws imply that the plasma velocity and magnetic field are interrelated in a special manner (“frozen” together)

$$\mathbf{V} = \frac{k}{4\pi \gamma \rho} \mathbf{B} + R \Omega_f \mathbf{e}_{\hat{\phi}} \quad (9.148)$$

The constant

$$k \equiv \frac{4\pi \gamma \rho V_p}{B_p} \quad (9.149)$$

is a key parameter of the problem and often called “the ratio of matter flux to magnetic flux” or, more simply, the “mass-loading parameter”. (Note that k is *not* dimensionless.) In studying an axisymmetric, stationary MHD problem, one usually chooses different values for k to see how the results vary with small or large amounts of matter loading a given field line. When $k \rightarrow 0$ the equations approach those of FFDE, with the plasma velocity $\mathbf{V} = \mathbf{V}_f = R \Omega_f \mathbf{e}_{\hat{\phi}}$.

Three other quantities also are conserved along a field line:

- the angular momentum per unit mass (“specific” angular momentum, including that of the magnetic field)

$$\ell = R \left[\gamma (1 + \mathfrak{h}/c^2) V_{\hat{\phi}} - \frac{B_{\hat{\phi}}}{k} \right] \quad (9.150)$$

where

$$\mathfrak{h} \equiv \frac{h}{\rho} = \frac{\Gamma}{\Gamma - 1} \frac{p}{\rho}$$

is the enthalpy per unit mass.

- the thermal adiabatic coefficient K_Γ (see equation (9.112)).
- the specific energy (the Bernoulli constant)

$$\boxed{Be = (\epsilon - 1)c^2 + \epsilon\psi} \quad (9.151)$$

where ϵc^2 is the total specific energy of the plasma (including rest mass)

$$\epsilon = \gamma(1 + \mathfrak{h}/c^2) - R\Omega_f \frac{B_{\hat{\phi}}}{k c^2}$$

whereas $(\epsilon - 1)c^2$ includes only kinetic, thermal, and electromagnetic energy.

These conserved quantities arise directly from equations (9.100) to (9.117).

A *single* equation of motion along each field line (*i.e.*, the conservation of energy or “Bernoulli” equation) then is found by combining equations (9.148), (9.150), and (9.151) to give the poloidal plasma velocity as a function of the strength and shape of that field line and in terms of the *five* model parameters Ω_f , k , ℓ , K_Γ , and Be .

9.5.6.2 The Equation(s) across the Magnetic Field

As in FFDE the cross-field equation is a force balance equation, but now the plasma forces also must be taken into account. Once again, the inhomogeneous Maxwell’s equations ((9.116) and (9.117)) and Ohm’s law play a crucial role here, determining the current and charge densities (in terms of $\mathbf{B}$) that are to be used in the Lorentz force. The cross-field equation is formally the component of the time-independent momentum equation (9.101) that is normal to the magnetic field and in the poloidal plane. In the cylindrical coordinates that we are using, this is written as

$$\begin{aligned} \frac{B_{\hat{Z}}}{R} \frac{\partial(R\mathbf{T}_{\hat{R}\hat{R}})}{\partial R} - \frac{B_{\hat{R}}}{R} \frac{\partial(R\mathbf{T}_{\hat{R}\hat{Z}})}{\partial R} + B_{\hat{Z}} \frac{\partial\mathbf{T}_{\hat{R}\hat{Z}}}{\partial Z} - B_{\hat{R}} \frac{\partial\mathbf{T}_{\hat{Z}\hat{Z}}}{\partial Z} - B_{\hat{Z}} \frac{\mathbf{T}_{\hat{\phi}\hat{\phi}}}{R} \\ + \left(\gamma\rho + \frac{\mathfrak{E}}{c^2} \right) \left(B_{\hat{Z}} \frac{\partial\psi}{\partial R} - B_{\hat{R}} \frac{\partial\psi}{\partial Z} \right) = 0 \end{aligned}$$

with the components of $\mathbf{T}$ given by equation (9.113). However, if it is known that $\mathbf{B}_p$ will be primarily along $\mathbf{e}_{\hat{Z}}$ (axial, as in a jet) or primarily along $\mathbf{e}_{\hat{R}}$ (radial, as in a wind), then we simply can use the component of equation (9.101) that is normal to the primary direction (*i.e.*, R or Z , respectively) in place of the true cross-field equation. As in FFDE, the cross-field equation determines the strength and shape of the field lines in terms of the field rotation rate and spacetime geometry. In the MHD case the plasma flow velocity plays a significant role in this as well. Together the cross-field and Bernoulli equations determine the R and Z components of the velocity and, therefore, density ρ and the poloidal components of the magnetic induction as well. $V_{\hat{\phi}}$ and $B_{\hat{\phi}}$ follow from equations (9.148) and (9.150).

Stationary, axisymmetric, ideal MHD has been used extensively to study jet production in black hole engines. We will look at how this in detail in Chapter 15.

9.5.7 Non-Relativistic Dynamics

Even in the study of plasma flows around black holes, sometimes the flow speeds and gravitational forces are not extremely high. For example, except near the last stable orbit, the speeds of accretion flows toward black holes remain well below c , and the gravitational field remains approximately Newtonian. It is *very* common, therefore, to study a black hole accretion problem in Newtonian mechanics first in order to understand the basic physics. After that, then the relativistic corrections can be added to see what additional effects they may produce. Working with non-relativistic flow also gives us the opportunity to add in all the stress-energy tensor terms in equation (9.7) and still work with equations of complexity similar to SRMHD. Therefore, in this subsection we also will include the effects of viscosity, conduction, radiation transport, and even nuclear burning, but all under the assumption of low flow speeds and gravity.

9.5.7.1 Non-Relativistic Dynamics with Total Energy Conservation

To derive the equations of non-relativistic dynamics, we start with equation (9.7) but this time include all of its components ($\mathcal{T}_{\text{GAS}}$, $\mathcal{T}_{\text{RAD}}$, and $\mathcal{T}_{\text{EM}}$), expand the appropriate terms in V/c , and drop everything of order V^2/c^2 or higher. We also ignore any metric frame dragging ($\beta = 0$) and recognize that in the Newtonian limit the lapse function gradient becomes $-c^2 \nabla \alpha = -\nabla \psi$ and Einstein's field equations for ψ become Newton's (equation (5.42))

$$\nabla^2 \psi = 4\pi G \rho$$

The result of this exercise is the following set

$$\boxed{\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{V})} \quad (9.152)$$

$$\boxed{\frac{\partial(\rho \mathbf{V})}{\partial t} = -\nabla \cdot \mathbf{T} - \rho \nabla \psi} \quad (9.153)$$

$$\boxed{\frac{\partial \mathfrak{E}}{\partial t} = -\nabla \cdot \mathbf{Q} + \rho \dot{q}_n} \quad (9.154)$$

$$\boxed{\frac{\partial \mathbf{B}}{\partial t} = -\nabla \cdot (\mathbf{B} \mathbf{V}) + \mathbf{B} \cdot \nabla \mathbf{V}} \quad (9.155)$$

where the total momentum density has been reduced to simply $\rho \mathbf{V}$ (the momentum of the electromagnetic field not being significant), the total energy density now has an additional gravitational term (from the $-c^2 \mathfrak{P} \cdot \nabla \alpha$ in equation (9.102))

$$\mathfrak{E} = \frac{1}{2} \rho V^2 + \varepsilon + \frac{B^2}{8\pi} + \rho \psi \quad (9.156)$$

and $\dot{q}_n(\rho, T)$ is the local nuclear energy generation rate, with T being the thermal temperature. (Note that we are no longer assuming adiabatic flow.)

The three-dimensional stress tensor in equation (9.153) now has dynamical, viscous, and electromagnetic parts

$$\mathbf{T} = \left[p + \frac{B^2}{8\pi} - \zeta_v(\nabla \cdot \mathbf{V}) \right] \mathbf{I} + \rho \mathbf{V} \mathbf{V} - 2\eta_v \boldsymbol{\sigma}_{\text{sh}} - \frac{1}{4\pi} \mathbf{B} \mathbf{B} \quad (9.157)$$

where $D^2 = E^2 \leq (V^2/c^2)B^2$ (equation (9.105)) now can be neglected, and the total pressure, total internal energy density, and shear tensor are given by

$$\begin{aligned} p &= p_g + p_r & \varepsilon &= \varepsilon_g + \varepsilon_r \\ \boldsymbol{\sigma}_{\text{sh}} &= \frac{1}{2} [\nabla \mathbf{V} + (\nabla \mathbf{V})] - \frac{1}{3} (\nabla \cdot \mathbf{V}) \mathbf{I} \end{aligned}$$

Components of the viscosity coefficients could include gas, radiation, and even turbulent effects. For example, in Cartesian coordinates the shear tensor is

$$\sigma_{\text{sh}}^{ij} = \frac{1}{2} \left(\frac{\partial V^i}{\partial x^j} + \frac{\partial V^j}{\partial x^i} \right) - \frac{1}{3} (\nabla \cdot \mathbf{V}) \delta^{ij}$$

The total energy *flux* vector in equation (9.154) is given by

$$\mathbf{Q} = \left(\frac{1}{2} \rho V^2 + p + \varepsilon + \rho \psi \right) \mathbf{V} + \mathbf{Q}_{\text{em}} + \mathbf{Q}_r + \mathbf{Q}_g \quad (9.158)$$

with the electromagnetic, radiative and heat energy fluxes given by

$$\mathbf{Q}_{\text{em}} = \frac{c}{4\pi} \mathbf{E} \times \mathbf{B} \quad (9.159)$$

$$\mathbf{Q}_r = -\frac{c}{3\bar{\kappa}_R \rho} \nabla \varepsilon_r \quad \mathbf{Q}_g = -K_c \nabla T \quad (9.160)$$

Note that not only is the plasma not adiabatic any more, but we cannot even write $\varepsilon = p/(\Gamma - 1)$, since the gas and radiation have different indices Γ .

9.5.7.2 Replacing Total Energy Conservation with the Internal Energy Equation(s)

Sometimes (particularly when doing analytic calculations rather than full numerical simulations) it is easier to work with the *internal* energy evolution equation rather than the total energy one (9.156)

$$\frac{\partial \varepsilon}{\partial t} = -\nabla \cdot (\varepsilon \mathbf{V}) - p \nabla \cdot \mathbf{V} + 2\eta_v \boldsymbol{\sigma}_{\text{sh}} : \boldsymbol{\sigma}_{\text{sh}} + \rho \dot{q}_n - \nabla \cdot (\mathbf{Q}_r + \mathbf{Q}_g)$$

(9.161)

This is similar to that in equation (5.65), but with heating by viscous friction and energy losses/gains due to radiation and conduction added in.

Furthermore, if the ion, electron, and photon gases are not in thermal equilibrium, the internal equation must be split into three separate ones for each particle species

$$\frac{\partial \varepsilon_i}{\partial t} = -\nabla \cdot (\varepsilon_i \mathbf{V}) - p_i \nabla \cdot \mathbf{V} + 2\eta_{v,i} \boldsymbol{\sigma}_{sh} : \boldsymbol{\sigma}_{sh} + \rho \dot{q}_{n,i} - \nu_t (\varepsilon_i - \varepsilon_e) \quad (9.162)$$

$$\begin{aligned} \frac{\partial \varepsilon_e}{\partial t} = & -\nabla \cdot (\varepsilon_e \mathbf{V}) - p_e \nabla \cdot \mathbf{V} + 2\eta_{v,e} \boldsymbol{\sigma}_{sh} : \boldsymbol{\sigma}_{sh} + \rho \dot{q}_{n,e} + \nu_t (\varepsilon_i - \varepsilon_e) \\ & - \nabla \cdot \mathbf{Q}_g - c \bar{\kappa}_P (aT_e^4 - \varepsilon_r) \end{aligned} \quad (9.163)$$

$$\frac{\partial \varepsilon_r}{\partial t} = -\nabla \cdot (\varepsilon_r \mathbf{V}) - p_r \nabla \cdot \mathbf{V} - \nabla \cdot \mathbf{Q}_r + c \bar{\kappa}_P (aT_e^4 - \varepsilon_r) \quad (9.164)$$

When $T_i \neq T_e \neq T_r$, the internal energies and pressures are computed separately

$$\begin{aligned} \varepsilon_i &= \frac{3}{2} n_i k T_i & p_i &= n_i k T_i \\ \varepsilon_e &= \frac{3}{2} n_e k T_e & p_e &= n_e k T_e \\ \varepsilon_r &= a T_r^4 & p_e &= \frac{1}{3} a T_r^4 \end{aligned}$$

These energy equations can be understood by dividing them into terms that affect the particle distribution directly and those that affect it due to interactions with the other species. Each species is directly affected by advection and $pd\mathcal{V}$ work on the distribution; and the ion and electron distributions can be affected by viscous heating and nuclear burning. Electrons and photons also are affected directly by transport of their respective particle species (conduction or radiative transfer). Electrons also lose and gain energy by interacting with ions (via Coulomb collisions and other processes) and with photons (via emission and absorption of radiation).

Equations (9.162) to (9.164), of course, sum to produce the single equation (9.161). However, by splitting energy conservation into its individual parts, we see more clearly how it works. For example, in some cases most thermal energy may reside in the ions, and most energy losses may be through radiation. However, the manner in which this loss happens is to first transfer internal energy to the electrons (the ν_t terms), then radiate that energy locally (the aT_e^4 term), and finally have the photons carry that energy away from the local area ($\mathbf{Q}$ term). Until the particles reach thermal equilibrium, equation (9.161) will not give the correct results for these processes.

9.5.7.3 Subsets of Non-Relativistic Dynamics

Several simpler sets of equations can be derived from the above set of non-relativistic dynamics equations (9.152) to (9.160) by setting certain terms to zero.

Rather than presenting all of these sets, we will mention them briefly below and do more with them later:

- *Accretion Disk Structure Equations.* These have no nuclear burning ($\dot{q}_n = 0$), and the gravitational potential is that of a point mass ($\psi = -GM/r$). Traditionally, they also assume no electron conduction ($\mathbf{Q}_g = 0$), but this process may, in fact, be important in the treatment of low-density accretion flows and coronae. There are a couple of different subsets of accretion disk equations:
 - **Adiabatic accretion flows: non-relativistic MHD (NRMHD).** By also assuming adiabatic ($\mathbf{Q}_r = \mathbf{Q}_g = 0$) and inviscid ($\eta_v = \zeta_v = 0$) flow, we obtain the *non-relativistic* versions of equations (9.140) to (9.145). This system of equations is often used to study the structure of non-radiative plasma around black holes (Chapter 12). Such a flow is often unstable to MHD turbulence (see Section 9.6.3), which will lead to angular momentum transport away from the black hole, allowing the orbiting plasma to spiral inward and accrete onto the black hole.
 - **Viscous and radiative accretion flows (NRHD).** Because accretion produces heat and radiation, adiabatic flow normally is not a good approximation. However, with viscosity, radiation, *and* MHD, accretion studies can become quite complicated. In order to not have to deal with MHD turbulence also, many investigators have made the *turbulent viscosity* approximation (Section 9.3.5). In this case one ignores the magnetic field ($\mathbf{B} = 0$) and uses a turbulent viscosity ($\eta_{v,t}$) to perform the angular momentum transport and heating.
- *Stellar Structure and Dynamics Equations.* Processes like electron conduction and nuclear burning are usually very important in stellar interiors, as is a solution for the gravitational potential at every point in the star. Traditionally, viscous and magnetic forces are ignored ($\eta_v = \zeta_v = 0$; $\mathbf{B} = 0$). (However, as the study of rotating stars becomes more important, the transfer of angular momentum and production of magnetic field within them eventually must be taken into account.) Stellar evolution is briefly discussed in the context of black hole formation in Chapter 10. Relativistic stellar collapse was discussed in Section 8.3.3.

9.6 Waves and Instabilities in the Fluid Plasma

In Chapters 6 and 8 we showed how waves can propagate in electromagnetic and gravitational fields. Here we also show how waves propagate in a fluid medium, magnetized or not. In addition, under certain conditions, waves do not simply propagate; they also can damp or grow. The growth of wave modes indicates the presence of an instability, and the frequency of the mode is an indication of its growth rate. Unlike the electromagnetic field (but like the general relativistic gravitational field) the equations of fluid dynamics are *nonlinear*. So, while we can analyze an unstable situation using the linearized equations, growing modes eventually will become so

large that the small perturbation approximation will not be valid. In order to see how an instability fully develops, therefore, we eventually will have to solve the full non-linear equations for that situation, perhaps by running a large computer simulation.

Below we discuss waves in some detail using the non-relativistic MHD equations. For the relativistic case we only present the results and then comment briefly on the differences between the relativistic and more familiar non-relativistic cases. Most of the information in this section comes from articles by [366, 367, 368], and from unpublished work and discussions with my colleague Masanori Nakamura.

9.6.1 General Features of Wave Analysis

Before analyzing the MHD equations for waves and instabilities, we briefly will review the techniques of wave analysis.

9.6.1.1 Small Amplitude Sinusoidal Perturbations

While waves, of course, can be any type of variation, here we will restrict ourselves to only small amplitude variations, *i.e.*

$$\rho = \rho_0 + \delta\rho'$$

where $\delta\rho' \ll \rho_0$. We then can *linearize* the nonlinear equations of MHD by substituting $\rho_0 + \delta\rho'$ for ρ , $p_0 + \delta p'$ for p , *etc.* in those equations, and then ignoring expressions that involve the product of two or more small perturbations (*e.g.*, $\delta\rho' \delta p'$, *etc.*). Furthermore, since *any* linear variation can be decomposed into its Fourier components

$$\delta\rho'(\mathbf{X}, t) = \int \delta\rho(\mathbf{k}, \omega) \exp[i(\mathbf{k} \cdot \mathbf{X} - \omega t)] d^3k d\omega$$

we can learn everything we need to know about waves from simply considering a small *sinusoidal* perturbation $\delta\rho(\mathbf{k}, \omega)$. Note that $\delta\rho$ is *constant* with time t and uniform in space $\mathbf{X}$; however, $\delta\rho'$ is not.

9.6.1.2 Angular Frequency and Propagation Vector

In beginning physics discussions waves often are characterized by a frequency ν and a wavelength λ . Here ω is the angular frequency

$$\omega = 2\pi \nu$$

(in units of radians per unit time) and $\mathbf{k}$ is the propagation or *wave number* vector

$$\mathbf{k} = \frac{2\pi}{\lambda} \mathbf{e}_{\hat{k}}$$

(in units of radians per unit length). $\mathbf{e}_{\hat{k}} = \mathbf{k}/k$ is the direction of propagation of the wave. In Chapters 6 and 8 we assumed that the propagation direction was along the Z axis only ($\mathbf{e}_{\hat{k}} = \mathbf{e}_{\hat{z}}$); however, a single wave can propagate in many or all directions. So $\mathbf{k}$ should be thought of as having many directions as well as different magnitudes (*i.e.*, different wavelengths). Different waves also can have different frequencies ω , of course.

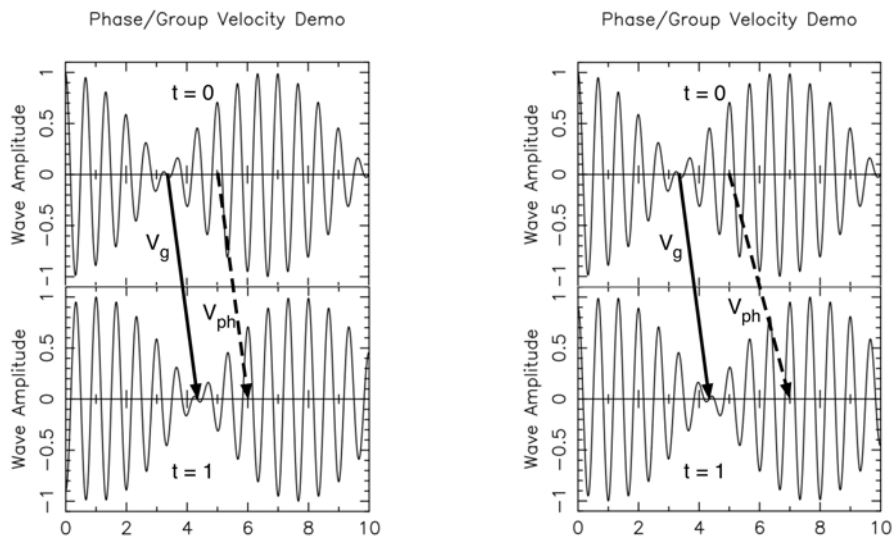


Fig. 9.6: Demonstration of the difference between phase and group velocity with plane wave amplitudes plotted as a function of distance (horizontal axis) and time (top and bottom plots). Both left and right diagrams show the propagation of two long-wavelength packets, each with ten short-wavelength waves, from $t = 0$ to $t = 1$. Also in both diagrams, the group velocity is the same: $V_g = 1$ (solid arrow that follows the null between two packets). At left the phase velocity is the same as V_g ($V_{ph} = 1$; dashed arrow that follows the crests of individual short-wavelength waves), causing the entire pattern to propagate at the same speed. At right the phase velocity is *twice* the group speed ($V_{ph} = 2$), causing the individual waves to propagate twice as fast as the packets themselves. Since information is encoded as modulations of a high-frequency carrier wave, that information travels at the group, not phase, velocity.

9.6.1.3 Phase and Group Velocity: The Dispersion Relation

However, waves in a given field or medium cannot have arbitrary values of ω and k . Each wavelength can have only a small number of possible frequencies (often only one). And different wavelengths, in general, will be associated with different frequencies. Fig. 9.6(left) plots a simple $\omega(k)$ relation. The ratio of angular fre-

quency to wave number has units of velocity, and there are two useful types of wave velocities to consider: the *phase velocity*

$$V_{ph} = \frac{\omega}{k}$$

and the *group velocity*

$$V_g = \frac{d\omega}{dk}$$

The phase velocity is the rate of advance of a single wave crest (or valley) along $e_{\hat{k}}$. However, waves also often travel in groups or *packets* (see Fig. 9.6(right)), and the speed of the packet (V_g) is not necessarily equal to V_{ph} . If $V_g \neq V_{ph}$, the wave packet could contract or *steepen* in the direction of propagation and comprise fewer total actual wave crests. On the other hand, the packet could spread out or *disperse* over more wave crests. The $\omega(k)$ relation tells us exactly how waves of a given type disperse (or steepen), so not surprisingly it is called the *dispersion relation*. The goal of wave analysis of plasma flows is to compute the dispersion relation for different situations and determine its physical implications for the flow.

9.6.2 Non-Damped Waves

We first will consider waves that oscillate only, neither damping nor growing.

9.6.2.1 Sound Waves in Non-Relativistic Hydrodynamic Flow

The simplest type of non-damped wave is the sound wave that occurs in non-magnetized fluids. We will use this as an example of how to derive a dispersion relation and the propagation speed. The exercise also will illustrate several concepts of flow with speeds slower, or greater, than the wave propagation speed.

The Dispersion Relation for Sound Waves

When the flow of a fluid gas or plasma is non-relativistic, adiabatic ($\dot{q}_n = 0$, $\mathbf{Q}_r = \mathbf{Q}_g = 0$), inviscid ($\eta_v = \zeta_v = 0$), not in a gravitational field ($\psi = 0$), and non-magnetized ($\mathbf{B} = 0$), equations (9.152) to (9.154) become those of non-relativistic hydrodynamics (see also Section 5.2):

$$\frac{\partial \rho}{\partial t} + \mathbf{V} \cdot \nabla \rho + \rho \nabla \cdot \mathbf{V} = 0 \quad (9.165)$$

$$\rho \frac{\partial \mathbf{V}}{\partial t} + \rho \mathbf{V} \cdot \nabla \mathbf{V} = -\nabla p \quad (9.166)$$

$$p = K_T \rho^\Gamma \quad (9.167)$$

We now will apply small-amplitude perturbations on the hydrodynamic flow, assuming that the equilibrium gas has zero velocity ($\mathbf{V}_0 = 0$), but arbitrary equilibrium density ρ_0 and pressure p_0 . The oscillating density, velocity, and pressure then can be written as

$$\rho = \rho_0 + \delta\rho \exp[i(\mathbf{k} \cdot \mathbf{X} - \omega t)] \quad (9.168)$$

$$\mathbf{V} = \delta\mathbf{V} \exp[i(\mathbf{k} \cdot \mathbf{X} - \omega t)] \quad (9.169)$$

$$p = p_0 + \delta p \exp[i(\mathbf{k} \cdot \mathbf{X} - \omega t)] \quad (9.170)$$

where ρ_0 , p_0 , $\delta\rho$, $\delta\mathbf{V}$, and δp all are constant in time and uniform in space. Inserting equations (9.168) to (9.170) into the nonlinear hydrodynamic equations, ignoring terms that contain products of the perturbation coefficients, and dividing through by $i \exp[i(\mathbf{k} \cdot \mathbf{X} - \omega t)]$, we obtain the *linearized* hydrodynamic equations

$$-\omega \delta\rho + \rho_0 \mathbf{k} \cdot \delta\mathbf{V} = 0 \quad (9.171)$$

$$-\omega \rho_0 \delta\mathbf{V} = -\mathbf{k} \delta p \quad (9.172)$$

$$\delta p = \left(\frac{\Gamma p_0}{\rho_0} \right) \delta\rho \quad (9.173)$$

Let us look at the linearized equations more closely. In principle, perturbations in velocity *could* have a component parallel to *and perpendicular* to the propagation vector $\mathbf{k}$. However, equation (9.172) tells us that $\mathbf{k} \times \delta\mathbf{V} = 0$ while $\mathbf{k} \cdot \delta\mathbf{V} = k^2 \delta p / (\omega \rho_0)$. So sound waves have *only a longitudinal variation* (in the direction of propagation); there are no transverse sound waves. Combining equations (9.171) to (9.173) and dividing by $\delta\rho$, we obtain the simple dispersion relation

$$\omega^2 = c_s^2 k^2 \quad (9.174)$$

where

$$c_s = \left(\frac{\Gamma p_0}{\rho_0} \right)^{1/2} \quad (9.175)$$

is the *adiabatic speed of sound*. Note that c_s depends only on the local equilibrium conditions in the ambient medium; it is independent of the amplitude of the small disturbances. It also does not depend on the wave frequency or propagation vector, so adiabatic sound waves are non-dispersive

$$V_g = V_{ph} = \pm c_s$$

Fig. 9.7(left) and 9.8(left) show simple phase and group *polar* diagrams for a sound (HD) wave. The wave speed in any direction is the sound speed c_s , both for individual wave crests and for wave groups.

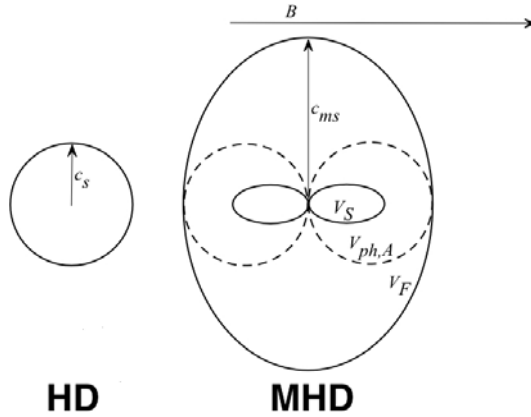


Fig. 9.7: Plots in phase velocity space (“phase polar diagrams”) showing the phase propagation speed of common waves in two dimensions. The magnitude and direction of a wave can be determined by drawing a vector from the origin of the diagram to the appropriate curve for a given mode. *Left:* Sound waves travel in all directions at c_s . *Right:* Three MHD modes are present; fast and slow magneto-acoustic (solid outer and inner ovals, respectively) and Alfvén (dashed “figure 8”). Parallel to the field B , fast and Alfvén waves propagate at the Alfvén speed, and slow waves propagate at the sound speed. Normal to the field, slow and Alfvén waves do not propagate at all, and fast waves travel at the magnetosound speed c_{ms} .

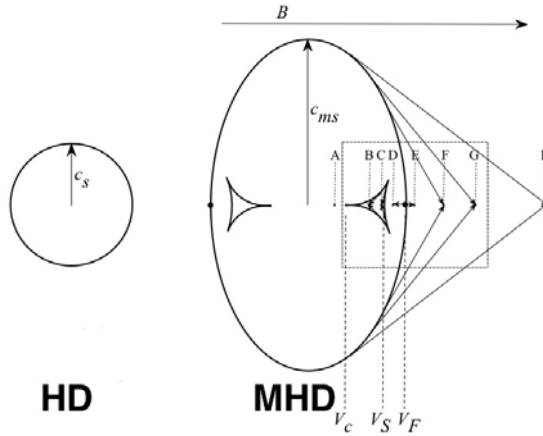


Fig. 9.8: Similar to Fig. 9.7, but plotted in group velocity space for the group propagation speed of common waves. *Left:* For sound (HD) waves, $V_g = V_{ph} = c_s$. *Right:* MHD waves, however, can be complex. The group speed of fast waves is on the outer oval; while our schematic depiction here looks similar to the MHD phase polar, it is not. Only the extremes (parallel and perpendicular to the field) are the same. (Compare equations (9.187) and (9.189).) The Alfvén group speed is $\pm V_A$, along the magnetic field direction only (two dots along the oval). The slow magneto-acoustic wave polar shows both a slow mode propagating along the field direction and one that spreads out normal to the field (the cusp wave). The characteristic analysis of the group polar diagrams, shown in the small box, is expanded and discussed further in Fig. 9.10.

The speed of a sound wave, however, *does* depend on the character of the energy equation. If, for example, the gas were held at a constant temperature by an outside heat bath (*i.e.*, $\Gamma = 1$), then we would have obtained a slightly different answer for the group and phase velocities

$$V_g = V_{ph} = \pm a$$

where

$$a = \left(\frac{p_0}{\rho_0} \right)^{1/2}$$

is the *isothermal sound speed*.

Subsonic (Elliptic) Flow: Modification of the Polar Diagram by Motion

When the fluid is in motion ($V_0 \neq 0$), the situation changes. In physical space the waves propagate faster in the direction of motion than against it. The wave front remains spherical, but its centroid travels downstream at the flow speed (see the moving polar diagram in Fig. 9.9a). As long as $|\mathbf{V}| < c_s$, however, waves still can propagate in all directions from the disturbance. The structure of that flow, therefore, is completely causally *connected* (sonically) and often called *elliptic* (instead of subsonic) flow, because of the mathematical and physical character of the hydrodynamic equations. The structure of elliptic flow is determined by conditions on all boundaries of a region, because information from all boundaries can propagate throughout that region.

Supersonic (Hyperbolic) Flow: Mach Cones, Characteristics, and Sonic Causality

The Mach cone. When $|\mathbf{V}| \geq c_s$, the wave front still is spherical, but its centroid now travels faster than the wave propagation speed (see Fig. 9.9b). This leaves the upstream part of the flow unaffected by the disturbance. In fact, in supersonic flow the propagating sound wave traces a *Mach cone* in physical space with an opening half-angle equal to the *Mach angle*

$$\alpha_M = \sin^{-1} \left(\frac{c_s}{|\mathbf{V}|} \right) \quad (9.176)$$

The Mach cone in physical space is similar to the light cone in spacetime, with the speed of light replaced by the sound speed and our exact speed of travel toward the future c replaced by a general velocity $\mathbf{V}$. In terms of sound waves the causal properties of Mach cones also are similar to those of light cones: when the flow speed is equal to (or greater than) the speed of sound, the downstream flow is sonically

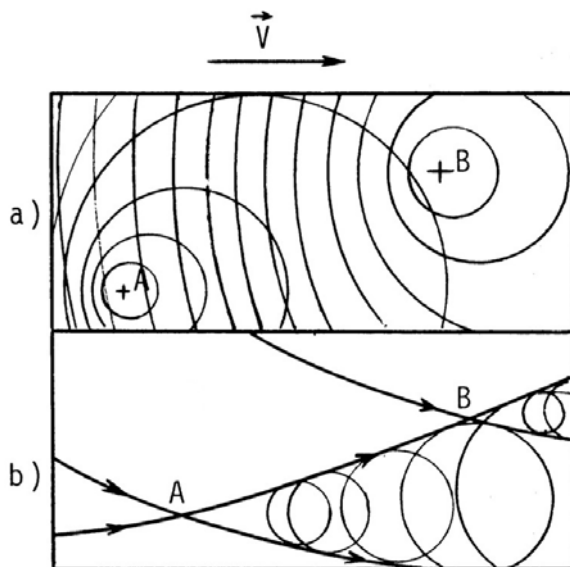


Fig. 9.9: A moving polar diagram, showing how characteristics are generated in hydrodynamics. (a) When points A and B, in a moving fluid, travel at subsonic speeds, the HD group polars of waves generated at those points remain spherical, but their centers move at subsonic speeds and still can transmit waves in any direction. (b) When points A and B move with speeds above the sound speed, the polars still remain spherical, but their waves can propagate only in the forward sound (or “Mach”) cone, delineated by the two vector-like curves. These curves are called “characteristics”, and they appear only when the fluid motion is supersonic. Reproduced from Fig. 1a and 1b of [366], by permission of J. Wiley and Sons, Ltd.

disconnected from the upstream flow. That is, waves, shocks, explosions or other hydrodynamic phenomena at points A or B in Fig. 9.9b can affect only the *sonic future* downstream (inside the forward Mach cone), and events at points A and B can be affected only by disturbances in the *sonic past* upstream (inside the backward Mach cone).

Characteristics and critical surfaces. The lines beginning at the origin of a disturbance in supersonic flow, and tangent to the edge of the Mach cone are called *characteristics*.⁹ Often they are drawn like vectors, showing a direction of information propagation as well as the orientation of the light cone in the flow. Mach cones and characteristics exist *only in supersonic flow*.

When characteristics *do* exist, the flow depends only on boundary conditions upstream (in the sonic past) and is often called *hyperbolic flow* because of this property. Since characteristics lie at the Mach angle, the component of the flow velocity per-

⁹ Note that the term “characteristic” has a specific meaning in the study of hydrodynamic flow; it does not refer simply to a general property of fluids. Instead, it refers to a boundary in space that delineates the limit of the propagation of waves.

pendicular to a characteristic is simply the appropriate wave phase velocity, and that flow velocity is directed inward into the future of the Mach cone. This inward flow velocity exactly cancels any outward wave propagation speed, leading to a stalling of the sound wave at the characteristic location.

Often, in astrophysics, flows are accelerated or decelerated by gravitational or dynamical forces from subsonic to supersonic speeds. The physical boundary between the elliptic and hyperbolic flow regions in such a flow is called a *critical surface*. That is, the critical surface is the set of points in space where characteristics suddenly appear (transition to hyperbolic flow) or disappear (transition to elliptic flow) along streamlines. Critical surfaces play an important role in the physics of accretion, wind and jet flows in black hole engines (see Chapters 12 to 15).

9.6.2.2 Waves in Non-Relativistic Ideal Magnetohydrodynamic Flow

Adding a finite magnetic field to the hydrodynamics complicates the wave analysis, not only because there are more variables, but also because the magnetic field adds a preferred direction to the wave propagation (along the field lines). If we ignore gravity, the non-relativistic equations of ideal, adiabatic MHD derived from equations (9.152) to (9.155) are

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0 \quad (9.177)$$

$$\rho \frac{\partial \mathbf{V}}{\partial t} + \rho \mathbf{V} \cdot \nabla \mathbf{V} = -\nabla \left(p + \frac{B^2}{8\pi} \right) + \frac{1}{4\pi} \mathbf{B} \cdot \nabla \mathbf{B} \quad (9.178)$$

$$p = K_\Gamma \rho^\Gamma \quad (9.179)$$

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{V} \cdot \nabla \mathbf{B} = \mathbf{B} \cdot \nabla \mathbf{V} - (\nabla \cdot \mathbf{V}) \mathbf{B} \quad (9.180)$$

Linearizing these equations about an equilibrium zero velocity flow yields a more complex dispersion relation than equation (9.174)

$$[\omega^2 - (\mathbf{k} \cdot \mathbf{V}_A)^2] [\omega^4 - k^2 \omega^2 c_{ms}^2 + (\mathbf{k} \cdot \mathbf{V}_A)^2 k^2 c_s^2] = 0 \quad (9.181)$$

Alfvén Waves

The first factor in equation (9.181) comes from considering *transverse* waves ($\mathbf{k} \cdot \delta \mathbf{V} = 0$) and is the Alfvén wave branch, with $\mathbf{V}_A$ being the Alfvén velocity *vector*

$$\mathbf{V}_A \equiv \frac{\mathbf{B}}{(4\pi \rho_0)^{1/2}} \quad (9.182)$$

The Alfvén wave has a non-zero frequency only if $\mathbf{k} \cdot \mathbf{V}_A \neq 0$, *i.e.*, only if $\mathbf{k}$ has a component along the magnetic field. Alfvén waves, therefore, behave like transverse

waves on a taut string, with magnetic tension ($\mathbf{B} \cdot \nabla \mathbf{B}/4\pi$) creating the restoring force. This and other phenomena are named after Hannes Alfvén, who received half the 1970 Nobel prize in physics for his work in magnetohydrodynamics and plasma physics.

Phase Velocity of Alfvén Waves. From equation (9.181) the dispersion relation for Alfvén waves is

$$\omega = \pm \mathbf{k} \cdot \mathbf{V}_A = \pm k V_A \cos \theta$$

which gives us the Alfvén phase velocity

$$\boxed{V_{phA} = \pm V_A \cos \theta} \quad (9.183)$$

where $\cos \theta \equiv \mathbf{k} \cdot \mathbf{V}_A / (k V_A)$ is the cosine of the angle between the propagation direction of the wave and the magnetic field direction, and $V_A = |\mathbf{V}_A|$ is the magnitude of the Alfvén velocity. Fig. 9.7(right) shows the *phase* velocity polar diagram for the Alfvén wave, along with the magneto-acoustic waves (discussed below).

Group Velocity of Alfvén Waves. In general, the group velocities of MHD waves differ from their phase velocities. That is, MHD waves tend to be dispersive. Because the direction of $\mathbf{k}$ is now important, we must use a more complex (vector) definition of group velocity than simply $d\omega/dk$

$$d\omega \equiv \mathbf{V}_g \cdot d\mathbf{k} \quad (9.184)$$

However, the group velocity of the Alfvén wave is very simple. If we have a uniform magnetic field, then the Alfvén dispersion relation can be written as

$$d\omega = \pm \mathbf{V}_A \cdot d\mathbf{k}$$

which tells us that

$$\boxed{V_{gA} = \pm V_A} \quad (9.185)$$

That is, the Alfvén group speed is simply the Alfvén speed, either parallel or anti-parallel to the magnetic field. These are represented by two points (dots) along the magnetic field direction in the MHD *group* (velocity) polar diagram in Fig. 9.8(right), which is often called the Friedrichs diagram¹⁰. The group polar is very useful in graphically determining the angle and direction of flow characteristics (see below).

¹⁰ Named after the mathematician Kurt O. Friedrichs, whose name also is linked to the well-known Courant–Friedrichs–Lewy condition used in the numerical solution of the partial differential equations in fluid flow.

Magneto-acoustic Waves

The second factor in equation (9.181) is the longitudinal (or magneto-acoustic) branch ($\mathbf{k} \times \delta \mathbf{V} = 0$), with c_{ms} being the magneto-acoustic or *magnetosound* speed

$$c_{ms} \equiv (V_A^2 + c_s^2)^{1/2} \quad (9.186)$$

Note that c_{ms} is the maximum value of any MHD wave speed.

Phase Velocity of Magneto-acoustic Waves. The magneto-acoustic waves have two branches

$$\omega^2 = c_{ms}^2 \left\{ \frac{1}{2} \left[1 \pm \left(1 - \frac{4 c_s^2 V_A^2 \cos^2 \theta}{c_{ms}^4} \right)^{1/2} \right] \right\} k^2$$

The positive root is called the “fast MHD mode”, while the negative root is called the “slow MHD mode”. The phase velocities of the two branches are

$$V_F^2 = \frac{1}{2} \left[c_{ms}^2 + (c_{ms}^4 - 4 c_s^2 V_A^2 \cos^2 \theta)^{1/2} \right] \quad (9.187)$$

$$V_S^2 = \frac{1}{2} \left[c_{ms}^2 - (c_{ms}^4 - 4 c_s^2 V_A^2 \cos^2 \theta)^{1/2} \right] \quad (9.188)$$

It is interesting to consider some limiting cases of the magneto-acoustic waves. Along the field direction, if $V_A > c_s$, then we have

$$V_F(\theta = 0) = \pm V_A \quad V_S(\theta = 0) = \pm c_s$$

(The reverse is true if $c_s > V_A$.) *Normal* to the magnetic field direction we have

$$V_F(\theta = \pi/2) = \pm c_{ms} \quad V_S(\theta = \pi/2) = 0$$

regardless of whether magnetic or plasma pressure forces dominate. We see, therefore, that the fast magnetosonic waves are the “ion-acoustic” waves of plasma physics. The slow waves, on the other hand, are plain sound waves when propagating along the magnetic field and standing waves when propagating perpendicular to the field direction. The phase polar diagrams for the slow and fast magneto-acoustic waves also are shown in Fig. 9.7(right).

Group Velocity of Magneto-acoustic Waves. The group velocities of the two magneto-acoustic branches are more complex than that of the Alfvén wave. Using equation (9.184) for the definition of the magneto-acoustic group velocity, we find that

$$\mathbf{V}_{g,F} = V_F \mathbf{e}_{\hat{k}} - \frac{c_s^2 V_A^2 \sin \theta \cos \theta}{V_F (c_{ms}^4 - 4 c_s^2 V_A^2 \cos^2 \theta)^{1/2}} \mathbf{e}_{\hat{\perp}} \quad (9.189)$$

$$\mathbf{V}_{g,S} = V_S \mathbf{e}_{\hat{k}} - \frac{c_s^2 V_A^2 \sin \theta \cos \theta}{V_S (c_{ms}^4 - 4 c_s^2 V_A^2 \cos^2 \theta)^{1/2}} \mathbf{e}_{\hat{\perp}} \quad (9.190)$$

where $\mathbf{e}_{\hat{\perp}}$ is the *unit* vector normal to both $\mathbf{k}$ and $\mathbf{B}$; $\mathbf{e}_{\hat{k}} = \mathbf{k}/k$ is the unit vector in the $\mathbf{k}$ direction; and V_F and V_S are always positive in these equations. Note that the component of the group velocity *along* the propagation direction is simply the phase velocity. However, an MHD wave packet also spreads *transversely* to the propagation direction and normal to the magnetic field.

Let us again look at some limiting cases, this time for the group velocity:

$$\begin{aligned} V_{gF}(\theta = 0) &= V_F(\theta = 0) = \pm V_A & V_{gS}(\theta = 0) &= V_S(\theta = 0) = \pm c_s \\ V_{gF}(\theta = \pi/2) &= V_F(\theta = \pi/2) = \pm c_{ms} & V_{gS}(\theta = \pi/2) &= \pm V_c \end{aligned}$$

where V_c is the “cusp” or tube velocity

$$V_c \equiv \frac{c_s V_A}{c_{ms}} \quad (9.191)$$

In the first three of these limiting cases the group speed matches the phase speed, so there is no dispersion. However, in the case of the standing slow mode wave, a wave packet spreads out normal to $\mathbf{k}$ and $\mathbf{B}$ and, as a result, transports information sideways at the cusp velocity. So, while V_c is not the phase speed of any wave (magneto-acoustic, Alfvén, or otherwise), it nevertheless is very important in MHD flows. In particular, because information can be propagated at the cusp speed, albeit transversely to the propagation direction, the boundary between regions where the flow speed is below the cusp speed ($|\mathbf{V}| < V_c$) and above that speed is also a “critical surface” of the flow, where the flow transitions between elliptic and hyperbolic.

Fig. 9.8(right) also shows the Friedrichs diagrams for the magneto-acoustic waves. A fast wave group advances at about the same speed as the wave crests, with a little dispersion near (but not at) $\theta = 0$. While the head of a slow wave packet advances at $V_S = c_s$ along the magnetic field, the tail of the slow packet lags at an even slower velocity (V_c) and propagates sideways.

Characteristics in an Accelerating MHD Flow

When a magnetized fluid is in very rapid motion, characteristics can exist for each type of wave possible: slow, Alfvén, and fast. The group polar in Fig. 9.8(right), which is expanded for clarity in Fig. 9.10, can be used to determine the shape, and direction, of the characteristics at each stage in the acceleration.

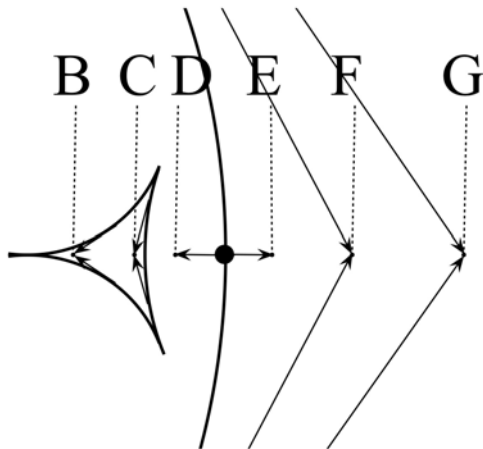


Fig. 9.10: Graphical demonstration of characteristic construction from a group polar diagram. The characteristics for a particular wave mode, and at a given point (A–H) in the fluid flow, are tangent to the group polar for that mode and pass through that point. See Fig. 9.8 and the text for more details and a full description of the technique.

The procedure is as follows [369, 370]: choose a flow velocity along the magnetic field direction and locate that point in the group polar diagram, (*e.g.*, point B in the super-cusp/sub-slow region in Fig. 9.10). Since each characteristic is normal to the flow of information at that point, a line that is tangent to one of the polar curves, and that passes through that velocity locus, forms a characteristic for that flow. And the direction of information flow is from the tangent point to the velocity locus (here to point B). In the case of Alfvén waves (*e.g.*, points D and E), characteristics *always* exist (any line passing through a point is a tangent). Alfvén characteristics lie either parallel or anti-parallel to the magnetic field lines, and there are no “Alfvén cones”. As a result, generally only the magneto-acoustic characteristics are shown on flow diagrams.

Fig. 9.11 shows the characteristics for a flow that is magnetically-dominated ($V_A \gg c_s$) and that accelerates linearly along the magnetic field direction, from below the cusp speed to well above the fast magnetosonic speed. The different flow velocity points in Fig. 9.8(right) and Fig. 9.10 also are shown. We see that the flow diagram is divided by three critical surfaces (CS, SMS, and FMS) into four regions of alternating elliptic and hyperbolic flow:

- For any point where $V < V_c$ (*e.g.*, point A) there are *no* lines that can be drawn from the velocity locus that are tangent to a wave polar curve; so no characteristics exist in this (elliptic) flow region (*cf.*, Fig. 9.7(right)).
- In the region where $V_c < V < V_S$ (points B and C) characteristics do exist, so the cusp surface (CS) is a critical surface. Note that the slow-mode characteristics and their Mach cones are directed *upstream*. That is, the slow mode group velocity has a component *opposite* to the direction of fluid flow, as well as nor-

mal to it. When $V = V_S$ the characteristics disappear, so the slow magnetosonic surface (SMS) also is a critical surface. Note that, right before the SMS, there is a discontinuity in the direction of the slow-mode characteristics that occurs when they jump from the sideways-directed slow-mode branch to the forward-directed one (point C).

- In the region $V_S < V < V_F$ (e.g., points D and E in Figs. 9.7(right) and 9.10) no magnetosonic characteristics exist, and the Alfvén characteristics (shown in Fig. 9.11) only change sign at the Alfvén surface (AS); they do not appear or disappear there. Therefore, *the Alfvén surface is not a critical surface* – a well-known property of MHD flow. However, the Alfvén surface is a *separatrix surface*, s-named because it separates information propagating against the flow from that propagating with the flow. In the general MHD case two additional separatrix surfaces exist, all of which will be very important in the acceleration and collimation of jets (Section 14.1). The entire trans-Alfvénic region is in elliptical flow and connected by Alfvén or fast magnetosonic waves.
- Finally, in the super-fast region, when $V_F < V$ (e.g., points F, G, and H in Fig. 9.8(right) and 9.10), characteristics form again. This time they and their Mach cones point downstream and become more parallel to the velocity as the fluid accelerates. The flow is, once again, hyperbolic, and now *all* information flow is downstream, so the fast magnetosonic surface (FMS) forms the third critical surface. *The super-fast flow is causally disconnected from the sub-fast part*: no disturbance in the former can affect the latter (except an electromagnetic one that travels at the speed of light!). The sub-fast flow, on the other hand, is fully causally connected, by either slow, Alfvén, or fast MHD waves.

9.6.2.3 Relativistic HD and MHD Waves in the Frame of the Fluid

We now will present the relativistic versions of HD and MHD waves in the frame of the moving fluid (*i.e.*, the frame where $\mathbf{V} = 0$). In the discussion on jets in Chapter 15 we will need to transform these to the rest frame of the black hole.

The Relativistic MHD Wave Phase Velocities

Instead of the non-relativistic MHD equations (9.177) to (9.180), we now use the relativistic ones (9.140) to (9.145) and perform the same type of wave analysis in the rest frame of the fluid (*i.e.*, $\mathbf{V} = 0$). The results give us the relativistic Alfvén wave speed

$$V_{A,R} = c \frac{B}{[4\pi (\rho c^2 + \varepsilon + p + B^2/4\pi)]^{1/2}} \quad (9.192)$$

and the relativistic forms for the fast and slow magneto-acoustic speeds

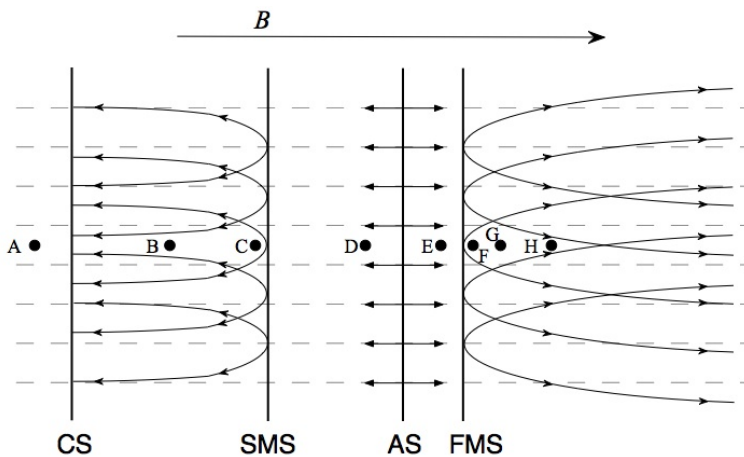


Fig. 9.11: Linearly accelerating flow along the magnetic field line B and the characteristics that are generated therein. Refer to Fig. 9.8 for locations of the points A–H in the MHD group polar diagram and to Fig. 9.10 for why characteristics have a particular directionality. Slow mode characteristics appear at the cusp surface (CS) and disappear at the slow magnetosonic surface (SMS). The Alfvén ones change sign at the Alfvén surface (AS), and fast mode characteristics appear at the fast magnetosonic surface (FMS).

$$V_{ph,R}^2 = \frac{c_{ms,R}^2}{2} \left\{ \left(1 + \frac{V_{c,R}^2}{c^2} \cos^2 \theta \right) \pm \left[\left(1 + \frac{V_{c,R}^2}{c^2} \cos^2 \theta \right)^2 - 4 \frac{V_{c,R}^2}{c_{ms,R}^2} \cos^2 \theta \right]^{1/2} \right\} \quad (9.193)$$

where the relativistic cusp, adiabatic sound, and magnetosound speeds are

$$V_{c,R} = \frac{c_{s,R} V_{A,R}}{c_{ms,R}} \quad (9.194)$$

$$c_{s,R} = c \left(\frac{\Gamma p}{\rho c^2 + \varepsilon + p} \right)^{1/2} \quad (9.195)$$

$$c_{ms,R} = \left[V_{A,R}^2 + c_{s,R}^2 \left(1 - \frac{V_{A,R}^2}{c^2} \right) \right]^{1/2} \quad (9.196)$$

Note that, while the Alfvén speed can approach the speed of light (as $B \rightarrow \infty$), in general the relativistic sound speed can approach only $c_{s,R} \rightarrow (1/3)^{1/2} c \approx 0.577 c$ (as $p \rightarrow \infty$ for a relativistic $\Gamma = 4/3$ plasma). Of course, when $B^2/4\pi\rho c^2 \ll 1$ and $(\varepsilon + p)/\rho c^2 \ll 1$, equations (9.192) to (9.196) reduce to the non-relativistic forms (9.182), (9.187) & (9.188), (9.191), (9.175), and (9.186), respectively.

Our four limiting cases look the same as before, but with the relativistic expressions for the characteristic speeds replacing their non-relativistic counterparts.

The Relativistic MHD Wave Group Velocities

Like the non-relativistic Alfvén wave, the relativistic one also is non-dispersive, so $V_{gA,R} = V_{phA,R}$ as before. The relativistic group velocities for the magneto-acoustic waves are now

$$V_{g,R} = V_{ph,R} \mathbf{e}_{\hat{k}} \mp \frac{V_{c,R}^2}{V_{ph,R}^2} \frac{\left(1 - \frac{V_{ph,R}^2}{c^2}\right) \sin \theta \cos \theta}{\left[\left(1 + \frac{V_{c,R}^2}{c^2} \cos^2 \theta\right)^2 - 4 \frac{V_{c,R}^2}{c_{ms,R}^2} \cos^2 \theta\right]^{1/2}} \mathbf{e}_{\perp} \quad (9.197)$$

and again the limiting cases are the same as in the non-relativistic case, but with the relativistic expressions substituted.

Note that the cusp speed plays an even more important role in relativistic MHD than in the non-relativistic version. In addition to determining the group velocity normal to the propagation direction, it also determines the relativistic corrections to the non-relativistic expressions for the magneto-acoustic phase and group velocities.

9.6.3 Instabilities in Non-Relativistic, Shearing HD and MHD Flow

Instabilities in the plasma that orbits black holes are ultimately responsible for powering black hole engines. The instabilities arise because of shear flow between different radii in the orbiting disk of plasma. They grow to full turbulence, and the turbulent viscosity heats the plasma and transports angular momentum outward, allowing plasma to sink slowly toward the black hole and eventually accrete onto it. Here we introduce some of the main instabilities that can occur in rotating, shearing fluids. These concepts will be important in Chapter 12 when we discuss the astrophysics of disk accretion. All of this discussion will use the non-relativistic equations of HD and MHD.

9.6.3.1 The Hydrodynamic Rotational Shearing Instability: Sound Waves Gone Bad

Beginning early in the 20th century, many laboratory experiments were performed with rotating fluids. The standard setup involved two co-axial cylinders made of a sturdy, but transparent, material (*e.g.*, acrylic or Plexiglas most recently), with fluid

in the space between them. Friction between the fluid and the cylinder walls causes this fluid to rotate with the inner and outer cylinders. If the two cylinders rotate at different angular rates, a *shearing* fluid flow develops (called Couette flow in honor of the French physicist Maurice Couette, who studied fluid dynamics in the late 19th and early 20th centuries). Understanding the behavior and instability of waves in shearing flows was an important step in understanding the behavior of orbiting plasma about black holes.

In Couette flow we deal with the adiabatic *hydrodynamic* equations (9.165) to (9.167) only and write them in cylindrical geometry. The equilibrium flow is a rotating one, with shear in the radial direction only

$$\mathbf{V} = R \Omega(R) e_{\hat{\phi}} + \delta \mathbf{V} \exp [i(k_{\hat{z}} Z - \omega t)] \quad (9.198)$$

Because of this shear, there no longer is any advantage to performing the analysis in the fluid frame; so it will be done in the rest frame of the laboratory. Note that we also consider waves that propagate along the cylinder axis (Z) only. The process of linearizing the fluid equations proceeds as in the simple sound wave case before, but the presence of rotating shear flow now modifies the dispersion relation

$$\omega^4 - (k^2 c_s^2 + \kappa^2) \omega^2 + k^2 c_s^2 \kappa^2 = 0 \quad (9.199)$$

where

$$\kappa \equiv \Omega \left[2 \frac{d \ln (R^2 \Omega)}{d \ln R} \right]^{1/2} \quad (9.200)$$

is the *epicyclic frequency*. (The epicyclic frequency is the rate at which fluid seems to rotate around an observer who sits in the fluid frame at some specific radius R .) Note that when $\kappa = 0$ (i.e., $\Omega = 0$ or $\Omega \propto R^{-2}$), equation (9.199) reduces to the simple dispersion relation for sound waves.

The new dispersion relation has two branches, sound waves and epicyclic waves

$$\begin{aligned} \omega_+ &= \pm k c_s \\ \omega_- &= \pm \kappa \end{aligned}$$

Because k and c_s are always real, these are the same oscillatory sound waves as we studied in the previous subsection. Furthermore, as long as

$$\boxed{\frac{d(R^2 \Omega)}{dR} > 0} \quad (9.201)$$

κ is real, and the epicyclic waves also remain oscillatory.

However, if inequality (9.201) is violated, then κ is imaginary, and there are modes for $\delta \rho'$, $\delta \mathbf{V}'$, and $\delta p'$ whose magnitudes grow with time at the rate

$$e^{-i\omega t} = e^{|\kappa| t}$$

In only a short time $|\kappa|^{-1}$ the perturbations become so large that they dominate the equilibrium flow and the system undergoes a dramatic change.¹¹ This change is called an *instability*.

Inequality (9.201) is the famous Rayleigh criterion for stability of a rotating fluid: angular momentum $R V_{\phi} = R^2 \Omega$ must increase outward for that flow to remain stable. The physical interpretation of this instability is as follows. The angular momentum shear gradient term κ^2 arises because of the Coriolis force. When angular momentum *increases* outward, the Coriolis force is a restoring force that pushes back into place any small disturbances, creating an oscillating small-amplitude wave. However, when angular momentum *decreases* outward, the Coriolis force changes sign, continually enhancing and growing any small perturbations in the flow. The energy that drives the growth comes from the differential rotation in the fluid.

The linearized equations and dispersion relation tell us nothing about the long-term behavior of an unstable system. They only tell us under what conditions it is stable or not. To determine the nonlinear evolution of an unstable system, one must perform the actual experiment and follow it for $t \gg |\kappa|^{-1}$ (or simulate the long-term flow in a computer using the nonlinear fluid equations). In the case of Rayleigh unstable Couette flow, the fluid becomes turbulent and transfers the excess angular momentum at small radii outward. This mixing produces a constant angular momentum fluid ($\Omega \propto R^{-2}$ or *vortex*) which is “marginally stable” ($\kappa = 0$).

The Rayleigh criterion is a good way to demonstrate the onset of a simple instability in sheared rotating flow mathematically and show how the frequency of an oscillating wave can become imaginary and produce a growing mode. However, this particular instability itself is of limited use in astrophysics. The reason is that virtually all plasma disks orbiting around black holes are in Keplerian rotation ($\Omega = (GM/R^3)^{1/2}$) or slower. In this case $\kappa = \Omega$ is a real quantity. *Gaseous Keplerian disks, therefore, are stable to the hydrodynamic shearing instability.* The realization of this led theorists to look elsewhere for the instability that might cause accretion and power black hole engines.

9.6.3.2 The Magneto-Rotational Shearing Instability (MRI): MHD Waves Gone Bad

The case that is relevant to black hole engines is rotational, *magnetized* fluid flow. As with the hydrodynamic shearing instability, we shall consider waves propagating parallel to the rotation axis and perturb about rotating equilibrium flow (equation (9.198)); but this time we will use the MHD equations (9.177) to (9.180). The resulting dispersion relation is much more complicated, but very important

¹¹ Modes also exist that *damp* at the rate $\exp(-|\kappa| t)$ in this case, but as with oscillating waves, damped waves do not produce a dramatic change in the character of the fluid flow.

$$\begin{aligned}
& \omega^6 - \omega^4 \left[\kappa^2 + k^2 c_{ms}^2 + (\mathbf{k} \cdot \mathbf{V}_A)^2 \right] \\
& + \omega^2 \left[(\mathbf{k} \cdot \mathbf{V}_A)^2 \left(k^2 c_s^2 + k^2 c_{ms}^2 + \frac{d(\Omega^2)}{d \ln R} \right) + \kappa^2 k^2 (c_s^2 + V_{A\hat{\phi}}^2) \right] \\
& - (\mathbf{k} \cdot \mathbf{V}_A)^2 k^2 c_s^2 \left[(\mathbf{k} \cdot \mathbf{V}_A)^2 + \frac{d(\Omega^2)}{d \ln R} \right] = 0 \quad (9.202)
\end{aligned}$$

where $V_{A\hat{\phi}}$ is the Alfvén speed computed using $B_{\hat{\phi}}$ only. Again, in the absence of rotation ($\kappa = \Omega = 0$) this dispersion relation reduces to the one for MHD waves (9.181). Equation (9.202) contains some of the secrets of black hole engines: the rotating, torsional Alfvén waves that are believed to accelerate and collimate jets, and the rotational shearing instabilities that explain why matter in Keplerian orbits accretes toward black holes and provides the “combustion” that powers their engines.

The Weak Magnetic Field (Slow Mode) MRI

Let us consider first the weak field limit of the dispersion relation (9.202), where $c_s/V_A \rightarrow \infty$ and $k^2 c_s^2/\Omega^2 \rightarrow \infty$. One might think that the magnetic field would be unimportant in this limit. However, this is decidedly *not* the case. Even an extremely weak, but non-zero, magnetic field will have important consequences for rotating flow. Keeping only terms proportional to $k^2 c_s^2$ in equation (9.202) and then dividing by that factor, we obtain a simpler, weak field dispersion relation

$$\omega^4 - \omega^2 \left[\kappa^2 + 2(\mathbf{k} \cdot \mathbf{V}_A)^2 \right] + (\mathbf{k} \cdot \mathbf{V}_A)^2 \left[(\mathbf{k} \cdot \mathbf{V}_A)^2 + \frac{d(\Omega^2)}{d \ln R} \right] = 0 \quad (9.203)$$

The roots of this relation have the form

$$\omega^2 = \frac{1}{2} \left[\mathcal{B} \pm (\mathcal{B}^2 - 4\mathcal{C})^{1/2} \right]$$

where $\mathcal{B} = 2(\mathbf{k} \cdot \mathbf{V}_A)^2 + \kappa^2$ and

$$\mathcal{C} = (\mathbf{k} \cdot \mathbf{V}_A)^2 \left[(\mathbf{k} \cdot \mathbf{V}_A)^2 + \frac{d(\Omega^2)}{d \ln R} \right]$$

In the weak field limit, the MHD waves divide into a slow mode (negative root) and an Alfvén wave (positive root). It is the slow mode that is the dangerous one as far as instabilities are concerned, because $\mathcal{B} - (\mathcal{B}^2 - 4\mathcal{C})^{1/2}$ can be negative, yielding an imaginary ω and rapidly-growing, unstable modes. Therefore, the stability criterion must be that this expression must remain positive, which reduces to simply $\mathcal{C} > 0$, or

$$\boxed{\frac{d(\Omega^2)}{d \ln R} > -(\mathbf{k} \cdot \mathbf{V}_A)^2 = -\left(\frac{2\pi}{\lambda} V_A \cos \theta\right)^2} \quad (9.204)$$

If this stability criterion is not satisfied, rotating weakly-magnetized flow is unstable. This process is called the magneto-rotational instability or the MRI.

Compare this criterion with inequality (9.201). There are two differences. First, in (9.204) it is the gradient in *angular velocity*, not angular momentum, that causes the instability. All Keplerian gaseous disks, including those around black holes, have rotational shear with angular velocity decreasing outward, so all such disks are potentially unstable. Secondly, the right-hand side of inequality (9.204) is not zero, so there is a *threshold* for the magneto-rotational instability. This can be put in terms of the wavelength of a given slow-mode wave perturbation. If that perturbation's wavelength is greater than a minimum threshold value

$$\lambda > \lambda_{min} \equiv 2\pi V_A \cos \theta \left[-\frac{d(\Omega^2)}{d \ln R} \right]^{-1/2}$$

then that slow-mode wave is *unstable*. The range of unstable wavelengths *increases* as the inward gradient of Ω increases and as the magnetic field strength *decreases*. So, *the weaker the magnetic field, the more unstable the magnetized rotational flow becomes*, because even the shorter wavelengths now are unstable. Therefore, for reasonable rotational shear (as in Keplerian flow) and for a weak magnetic field, the small threshold does not inhibit the instability from growing.

Being unstable is not, by itself, a big problem if the growth rate of the instability is slow. We can find the maximum growth rate from equation (9.203) by noting that ω^2 is parabolic in $(\mathbf{k} \cdot \mathbf{V}_A)^2$, with a *negative* minimum at

$$\omega^2 = -\frac{1}{16 \Omega^2} \left(\frac{d(\Omega^2)}{d \ln R} \right)^2$$

This gives a maximum (imaginary) growth rate, and mode for which this occurs, of

$$\omega_{max} = i \frac{1}{2} \left| \frac{d\Omega}{d \ln R} \right| \quad (9.205)$$

$$(\mathbf{k} \cdot \mathbf{V}_A)_{max}^2 = -\frac{1}{16} \left(\frac{d \ln(R^8 \Omega^2)}{d \ln R} \right) \left(\frac{d(\Omega^2)}{d \ln R} \right) \quad (9.206)$$

For any rotating magnetized gas whose angular rotation rate varies as a reasonable power of the cylindrical radius, $\omega_{max} \sim \Omega$. So, in fact, this is a very rapidly growing instability and, therefore, a very important one.

What is it about weakly-magnetized gas that is so unstable? Steven A. Balbus and John F. Hawley (both at University of Virginia), who discovered the instability, describe it in terms of two orbiting gas blobs connected by a single magnetic field line in tension (see Fig. 9.12). If the leading blob moves inward, it rotates faster, pulling harder on the trailing one. In normal linear motion this would not be a problem, but in orbital motion this tug transfers angular momentum from the leading blob to the trailing one. This loss of angular momentum causes the leading blob to fall to a still

lower orbit and *speed up*, while the gain of angular momentum causes the trailing blob to drift outward to a higher orbit where it *slows down*. This *increases* the velocity difference between the two blobs, as well as the strength of the magnetic field line, causing the process to accelerate.

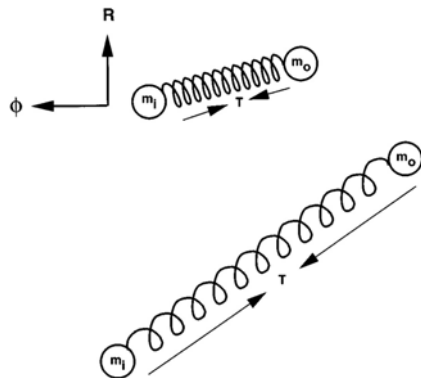


Fig. 9.12: Graphical demonstration of the magneto-rotational instability (MRI) and how it results from magnetic tension between two fluid elements in orbit about a central mass (well below the bottom of this page). The orbital motion is toward the left, in the ϕ direction. Transfer of angular momentum outward, due to the magnetic tension (top spring system), causes the left blob to fall toward the gravitating object, where it orbits even faster (bottom spring system). This instability continues until the spring (magnetic tension) breaks, and the subsequent nonlinear motions develop into turbulence. Reprinted from Fig. 16 of reference [367] with permission; copyright (1998) by the APS.

The behavior is almost identical to two spacecraft in orbit about the earth that are connected by a tether line. A “tether-rotational” instability will pull the line taut in a time comparable to an orbital time; however, if the tidal forces do not snap the tether, the “instability” will saturate, and the two spacecraft will become locked together in the same orbit.

The Strong Magnetic Field (Alfvén Mode) MRI

When will the magnetic field be so strong that the MRI turbulence will shut down? As with the tether-rotational instability, the MRI should do so when the magnetic field strength becomes strong enough to resist the shear. However, because the stability criterion (9.204) was derived in the weak field limit, it may not give accurate results when the field is strong. So, we must look at the MRI in the limit that the magnetic field is strong and develop a complementary dispersion relation.

The strong field limit assumes the opposite condition to the weak field limit, namely $k^2 c_s^2 \rightarrow 0$. Applying this assumption to equation (9.202) and now dividing by ω^2 , we obtain the strong field dispersion relation

$$\omega^4 - \omega^2 \left[\kappa^2 + k^2 V_A^2 + (\mathbf{k} \cdot \mathbf{V}_A)^2 \right] + \left\{ (\mathbf{k} \cdot \mathbf{V}_A)^2 \left[k^2 V_A^2 + \frac{d(\Omega^2)}{d \ln R} \right] + \kappa^2 k^2 V_{A\hat{\phi}}^2 \right\} = 0 \quad (9.207)$$

Again the roots have the form $\omega^2 = [b \pm (b^2 - 4c)^{1/2}]/2$ where $b = \kappa^2 + k^2 V_A^2 + (\mathbf{k} \cdot \mathbf{V}_A)^2$ and

$$c = (\mathbf{k} \cdot \mathbf{V}_A)^2 \left[k^2 V_A^2 + \frac{d(\Omega^2)}{d \ln R} \right] + \kappa^2 k^2 V_{A\hat{\phi}}^2$$

This time, however, the positive root is the fast magneto-acoustic wave, and the negative root is the Alfvén (not the slow mode) wave. The slow mode is now essentially the sound wave, which we ignore when $c_s^2/V_A^2 \rightarrow 0$.

The stability criterion again is simply $c > 0$. However, now we need to be careful about the direction of the magnetic field, mainly because of the term with the azimuthal factor $V_{A\hat{\phi}}$. Here we will consider magnetic fields with no radial component ($B_{\hat{R}} = 0$)

$$\mathbf{B} = B \cos \theta \mathbf{e}_{\hat{z}} + B \sin \theta \mathbf{e}_{\hat{\phi}}$$

Using the definition (9.200) of κ , the stability criterion in the strong field case becomes

$$k^2 V_A^2 \cos^2 \theta + \frac{d(\Omega^2)}{d \ln R} + 4 \Omega^2 \sin^2 \theta > 0 \quad (9.208)$$

For waves traveling perpendicular to the strong magnetic field (fast magneto-acoustic waves with $\theta = \pi/2$) this inequality is simply the Rayleigh criterion (9.201), and the wave modes are $\omega = (\pm \kappa, \pm k V_A)$ – epicyclic and magnetosonic waves. So, for all rotating flows that are Rayleigh stable (like disks of gas orbiting black holes) all the waves perpendicular to the strong magnetic field are oscillatory.

For waves traveling parallel to the magnetic field (Alfvén waves in this case, with $\theta = 0$) the stability criterion turns out to be the same as the weak field MRI criterion

$$\boxed{\frac{d(\Omega^2)}{d \ln R} > - \left(\frac{2\pi}{\lambda} V_A \right)^2} \quad (9.209)$$

So, *equation (9.204) can be used to study the onset of the MRI in the weak field situation and its shut down when the field becomes strong.*

What does this mean for the possible shut down of the MRI as the magnetic field becomes strong? Again, we can write the condition for *instability* (9.209) as

$$\lambda > \lambda_{min} \equiv \frac{2\pi V_A}{\left[-\frac{d(\Omega^2)}{d \ln R} \right]^{1/2}}$$

Now, for any system of finite size H , all wave modes will be shorter than that size (*i.e.*, $\lambda < H$). So, when the magnetic field becomes so strong that

$$V_A = \frac{H}{2\pi} \left[-\frac{d(\Omega^2)}{d \ln R} \right]^{1/2} \quad (9.210)$$

then no more waves in the system can be unstable. It now becomes marginally stable to the MRI. H turns out to be the disk scale height, and equation (9.210) can be used to determine the magnetic field strength above which the MRI turns off.

Part III

Astrophysics of Black Hole Engines

Chapter 10

Assembling the Engine Block: Formation of Black Holes in the Universe

In Chapter 8 we discussed the basic *physical* processes of how normal matter can collapse to a radius so small that it forms an horizon and, therefore, becomes a black hole. That chapter also showed that black holes can merge and can grow by accreting material. Indeed, once formed, black holes cannot break apart; they only can grow.

However, that was only the beginning of the story. Where, when, and how black holes form throughout the universe depends critically on complex environmental matters, including the formation, evolution, and death processes of stars, what their heavy metal abundances are, and, ultimately, on the formation and evolution of the galaxies themselves.¹ Each of these latter areas of *astrophysics* are actually entire fields of study, in which many astronomers and astrophysicists spend their lives working on a few aspects of these vast, and only partially-understood, subjects. We cannot hope to treat them well even in a single book, let alone in one chapter. However, we shall introduce enough of the basic concepts below to properly discuss the current ideas on how we believe black holes of many sizes (stellar mass, intermediate mass, massive, or supermassive) have formed and continue to do so.

Much of this story still remains to be written, and each year brings new surprising discoveries and revelations. So, what is discussed below may have to be modified or even rewritten several times in the next decade or two before it becomes accepted as consensus by most astrophysicists. For example, it is possible that the formation of black holes in our three principal mass ranges may be largely unrelated and inherently different. However, there also is a hint from current observational and theoretical work that, at least in some circumstances if not all, the formation of black holes of stellar mass may be related to that of intermediate mass and possibly even to the assembly and growth of the supermassive ones.

¹ Note that we will concern ourselves here only with black hole objects of the type that we believe to exist in observed astrophysical systems (*i.e.*, $1\text{--}10^{10} M_{\odot}$ in mass), and not with more exotic, and hypothetical, microscopic or hypermassive black holes, the latter of galaxy mass or larger.

10.1 The Formation of Neutron Stars and the Synthesis of Heavy Elements

Just as the discussion of the observations of stellar-mass black holes cannot be separated from that of neutron stars, so the stories of how both of these compact stellar remnants are produced also are intertwined. The key to determining what kinds of normal stars produce neutron stars and stellar black holes is in understanding which stars produce elements heavier than helium (often called “metals”) and in understanding how those elements are dispersed throughout the Galaxy in powerful stellar explosions called supernovae. Moreover, whether a neutron star or black hole is produced in the center of a star depends on the rate at which that parent star loses mass from its outer layers. This mass loss, in turn, depends on how much heavy element material was in the original gas that contracted to form that star in the first place. So, the astrophysical processes by which black holes of stellar mass form are closely linked to the nucleosynthesis of heavy elements – from the big bang to the present day. Compared to other astrophysical fields, this one is rather well understood.

10.1.1 Big Bang Nucleosynthesis

The story of the cosmic metal abundance began about one second after the big bang, when the temperature of the universe was cool enough (a few billion kelvins) for regular nuclear reactions to take place. (Before that, neutrons and protons were in “nuclear statistical equilibrium”, changing from one to the other and back again, but maintaining a reasonably constant neutron-to-proton ratio.) Hydrogen (protons) and free neutrons cooked into only a few key elements, before the expanding primordial soup “froze out” when the temperature dropped below 400 million K (about 1000 seconds after the big bang). All the free neutrons at that time then were used up in creating these new elements, with the final mixture having the following abundances by *mass*:

^1H	(protons)	75%
^4He	(α particles)	25%
^2H	(deuterium)	0.003%
^3He	(helium-3)	0.001%
^7Li	(lithium-7)	$4 \times 10^{-8} \%$

with no other elements being produced. The big bang, then, produced a cosmic metal abundance of $Z \approx 4 \times 10^{-10}$, whereas the metal abundance in the sun, for example, is $Z_{\odot} \approx 0.02$ (50 million times higher). And it produced a helium abundance of $Y \approx 0.25$, whereas the solar value is a little higher, $Y_{\odot} \approx 0.27$. Therefore, virtually all the heavy metals, and some of the current helium, in the solar system must have been manufactured after the big bang freeze-out – presumably in stellar interiors.

10.1.2 Stellar Evolution Nucleosynthesis for Different Star Masses

Compared to the big bang, stellar interiors are a much better environment in which to cook the elements that make up our world. Stars contract to a dense and hot state, rather than expanding to cool and low-density conditions like the early universe.

Stellar structure and evolution is calculated using a subset of equations (9.152) to (9.157) (with $\eta_v = \zeta_v = 0$; $\mathbf{B} = 0$; see Section 9.5.7). These include the radiation contribution to pressure and internal energy and also radiative heat transport from the stellar interior to the surface. In more evolved stars, which tend to be much denser, the equations also include the effects of electrical conduction of heat. One further complication is added to sophisticated models, which was not treated in Chapter 9: all the equations of state also are functions of the composition of the gas, and that changes with time as the star synthesizes more heavy elements in its interior as time goes on. The stellar structure equations themselves, which are generally one-dimensional, actually are fairly simple to solve. However, keeping track of the abundances of ^1H , ^2H , ^3He , ^4He , *etc.* (all the way up to iron and beyond in some cases), and computing all the proper nuclear reaction networks can be very complex. After more than a half of a century working in this field, many groups of astrophysicists have tackled these problems independently and generally agree on the results below.

How much helium and heavy metal material is produced and dispersed by a given star depends critically on the mass the star has ($M_\star$) when it is first born. There are six major mass ranges: brown dwarfs ($0.01 M_\odot < M_\star < 0.085 M_\odot$), low-mass stars ($0.085 M_\odot < M_\star < 0.4 M_\odot$), solar-type stars ($0.4 M_\odot < M_\star < 7 M_\odot$), moderate-mass stars ($7 M_\odot < M_\star < 9 M_\odot$), intermediate-mass stars ($9 M_\odot < M_\star < 11 M_\odot$), and massive stars ($11 M_\odot < M_\star < 100 M_\odot$).²

10.1.2.1 Brown Dwarfs ($0.01 M_\odot < M_\star < 0.085 M_\odot$): No Element Synthesis at All

These stars never get hotter in their interiors than one or two million kelvins and, so, never start nucleosynthetic reactions. Like the planets, they are not responsible for the production of any elements in the universe. However, as the low-end of the stellar initial mass spectrum, brown dwarfs represent a large percentage of the total number of stars in the universe. And they serve as missing links between the cool M dwarf stars and giant gas planets like Jupiter and Saturn.

² The sizes of the mass ranges are fairly accurate, but their relative boundaries depend critically on the exact physics used in one's stellar evolution code. The discussion below of the mass ranges and the behavior of the stars within them were synthesized from (1) lecture notes of Harding E. (Gene) Smith, of the University of California San Diego, for the lower masses ($M_\star < 9 M_\odot$) [371] and (2) from reference [372] plus numerous discussions with Stan Woosley, of the University of California Santa Cruz, for the higher masses.

10.1.2.2 Low-Mass Stars ($0.085 M_{\odot} < M_{\star} < 0.4 M_{\odot}$): Slow Burning, Unproductive Furnaces

These are basically the fainter M dwarf stars (M3 and higher). They *do* begin hydrogen burning via the proton–proton chain reaction ($10^7 K$) and do produce helium in their interiors, albeit at a very slow rate. Their luminosities drop steeply with decreasing stellar mass

$$L_{\star} \approx 0.04 L_{\odot} \left(\frac{M_{\star}}{0.4 M_{\odot}} \right)^{3.5}$$

which gives them extremely long lifetimes

$$\tau_{\star} \approx 10^{10} \text{ yr} \left(\frac{M_{\star}/M_{\odot}}{L_{\star}/L_{\odot}} \right) = 10^{11} \text{ yr} \left(\frac{M_{\star}}{0.4 M_{\odot}} \right)^{-2.5}$$

These are much longer than the current age of the universe – 1.37×10^{10} yr. Therefore, *none* of these stars (not even any that may have formed shortly after the big bang) have had a chance to die and disperse around the Galaxy any elements that they have synthesized.

Eventually, after a time of about $\tau_{\star}$ years, these stars will deplete the hydrogen in their central cores and burn it on the surface of a small helium “ash” core in their centers. The core will contract, the hydrogen envelope will expand, and the star will become a red giant. However, the temperatures in the helium core will never become hot enough (several hundred million K) to start the “triple- α ” reaction chain and burn helium to carbon. So low-mass stars will never contribute to heavy element synthesis at all, no matter how long we wait. And most of the helium that they synthesize will be locked up in the small white dwarf remnants that they leave.

10.1.2.3 Solar-Type Stars ($0.4 M_{\odot} < M_{\star} < 7 M_{\odot}$): Major Producers of Helium

This category covers a wide range of spectral types: most B dwarfs (B4 to B9); all A, F, G, and K dwarfs; and the brighter M dwarfs (M0 to M2). These range in lifetime from 3×10^8 yr to 10^{11} yr. The smaller mass ones (0.4 – $1.2 M_{\odot}$; G, K, and M0 to M2 dwarfs) continue to burn hydrogen by the p–p cycle, but the more massive ones (1.2 – $7 M_{\odot}$; F, A, and B4 to B9 dwarfs) have interiors hot enough to use the carbon–nitrogen–oxygen (CNO) cycle.

Stars with a mass similar to or greater than that of the sun (*i.e.*, $M_{\star} > 0.8 M_{\odot}$) that were created when the Galaxy first formed have already depleted the hydrogen in their cores and have become red giants. Their helium cores have contracted and become hot enough ($T > 200 \times 10^6 K$) for the triple- α reaction chain to ignite. Helium burning in the core causes the star to settle into a new stable “helium-burning main sequence phase”, but this lasts only 10–20% as long as the original hydrogen-

burning main sequence phase. During this period not only do we have ${}^4\text{He}$ burning to ${}^{12}\text{C}$, but we also have the carbon combining with free helium ions to produce ${}^{16}\text{O}$. The carbon and oxygen ash collects in the center of the core, inside what is now the helium mantle.

Finally, after the helium in the center is exhausted, helium burning, like hydrogen burning, occurs only in a shell around the C/O core. There are, then two burning shells (H and He) that separate three different regions in the star: C/O core, He mantle, and H envelope. The cooling carbon–oxygen core contracts, causing the hydrogen envelope to expand, initiating a variety of processes (winds, thermal pulses, bipolar outflows [jets], *etc.*) that cause loss of some or most of the hydrogen envelope and helium mantle. This mass loss creates an observable “planetary nebula” around the star’s location and leaves a C/O white dwarf in its center with a mass up to $\sim 1.0 M_{\odot}$, depending on how massive the original star was.

10.1.2.4 Moderate-Mass Stars ($7 M_{\odot} < M_{\star} < 9 M_{\odot}$): Producers of Carbon, Nitrogen, and Oxygen

These comprise a narrow range of spectral type ($\sim\text{B3}$), with lifetimes of only 30–300 Myr. The evolution of moderate-mass stars is fairly complex and, like most stars with masses significantly greater than the sun, still a little uncertain. The bottom line for nucleosynthesis is that moderate-mass stars also tend to leave C/O white dwarfs, but they eject significant amounts of carbon, nitrogen, and oxygen into the Galaxy. They thus contribute significantly more heavy elements to the interstellar medium than the solar-type stars.

The following is a description of the current best results in this area of study so far. If single moderate-mass stars were not to lose any mass during their evolution, they would develop very degenerate C/O cores. (That is, the equation of state at the center would be given by equation (9.67).) Then, as the core approached the $1.4 M_{\odot}$ Chandrasekhar mass limit (equation (5.75) with $\mu_e = 2$), carbon would ignite and completely explode the star as an enormous nuclear bomb. Degenerate carbon burning would be explosive for two reasons. First, because the core is highly degenerate, the increase in temperature caused by the nuclear burning does *not* increase the internal core pressure, so instead of the core expanding and cooling, its temperature continues to rise and the carbon burning becomes even more vigorous. Second, by the time the core temperature exceeds the Fermi temperature (which then *does* allow it to expand), the burning of carbon to heavier elements has released so much energy that the core is now unbound (exploding) and its expansion tears the star apart.

However, theorists have determined that this stellar doomsday scenario does not occur for stars in this mass range – for two additional reasons. First, a deep convection zone develops in the star, reaching down into the C/O core and up into the H-burning shell. This reduces the mass of the C/O core so it never approaches $1.4 M_{\odot}$ and thus never ignites carbon explosively. Finally, the H- and He-burning shells become quite unstable, pulsating dramatically. These pulses eject the envelope, leaving only the core, which becomes a C/O white dwarf.

However, the above processes have important consequences for heavy element dispersal throughout the Galaxy. The carbon dredged up from the C/O core by the unstable H-burning shell undergoes some further nuclear processing in that shell. Carbon combined with hydrogen creates nitrogen and oxygen, as well as more helium, in the other nuclear burning chain – the CNO cycle. So, even if the $7\text{--}9\ M_{\odot}$ stars start out with no metals, they still would disperse a significant amount of CNO products into the Galaxy when they become planetary nebulae. Indeed, the moderate-mass stars may explain much of the amount of carbon, nitrogen, and oxygen in the universe today, but not the majority of the heavy elements.

10.1.2.5 Intermediate-Mass Stars ($9\ M_{\odot} < M_{\star} < 11\ M_{\odot}$): Possible Producers of Molybdenum and Ruthenium – and Some Supernovae and Neutron Stars

This narrow range of mass (not to be confused with intermediate mass black holes) consists entirely of $\sim B2$ stars. Near the end of their short 30-Myr lifetimes they also produce H- and He-burning shells, with C/O ash collecting in the center. However, the deep convection zone does not develop in stars with $M_{\star} > 9\ M_{\odot}$. The C/O cores of these stars then grow so rapidly that they become hot enough to ignite carbon burning when the core is *not* degenerate. Unlike the degenerate type, nondegenerate carbon burning is not explosive, and it proceeds to create a new type of stellar core not seen in lower-mass stars – a degenerate core of even heavier elements.

Carbon has several burning products, the most abundant of which are neon and magnesium. These collect in the center, along with the oxygen that was produced by the earlier helium burning, to form a degenerate ONeMg core that steadily grows in mass until most of the carbon has been used up. Eventually, there are a total of *three* burning shells – hydrogen, helium, and carbon – surrounding the contracting ONeMg core. The star again ascends the red giant branch and becomes quite large – a red supergiant.

Unlike stars with less than $9\ M_{\odot}$, however, the stellar death in this mass range is violent. Instead of gentle outflows or even wild thermal pulsations, the intermediate-mass star explodes in a fraction of a second, ejecting the carbon–helium–hydrogen mantle and envelope, *leaving a neutron star pulsar*. The basic mechanism for the explosion is *not* thermonuclear, but rather the release of gravitational energy during a violent implosion of the core. It is, in fact, the classical model for all *core-collapse* supernovae, and we discuss the main steps of this process below.

Core Implosion

The degenerate core (in this case composed of ONeMg) grows in mass until it becomes nearly the Chandrasekhar limit of $1.4\ M_{\odot}$ in those elements alone. When the central density reaches $\sim 5 \times 10^9\ \text{g cm}^{-3}$, the neon begins to burn in an unusual way: it absorbs free electrons. This burning is quite rapid, but not explosive. Instead,

the loss of so many electrons by the degenerate gas (which *holds up* the star) is so great, and the increase in pressure due to the neon-burning so slight, that the core *collapses*.

As the collapse ensues, the tremendous pressures and temperatures created by the compression cause the O, Ne, and Mg nuclei to collide and *dissociate* into neutrons and protons: all the long work of building up those elements is destroyed in an instant. Each proton then captures an electron under these extreme conditions, becoming a neutron instead and turning the original ONeMg core into a simple neutron gas. This further loss of essentially *all* electrons causes the collapse to accelerate to free fall until the pressure created by the neutron gas becomes important. However, this does not happen until the density becomes greater than $10^{13} \text{ g cm}^{-3}$ – nearly that of atomic nuclei themselves ($\sim 4 \times 10^{14} \text{ g cm}^{-3}$).

Core Bounce

At this point the equation of state stiffens to that of a degenerate free *neutron*, not electron, gas (equation (9.68)), and the collapse is abruptly halted. A “bounce” shock forms, like the one we saw in Fig. 8.9. In the early days of supernova work (the 1960s), it was thought that this bounce directly, and dynamically, caused the star to explode. However, this is not the case. Instead, the pressure and temperature in the shock are so great that they dissociate all remaining Ne, O, and Mg nuclei falling through the shock into individual nucleons. This dissociation is an *endothermic* reaction that absorbs so much of the energy in the shock that the latter stalls and becomes an *accretion shock*, rather than an explosive shock. The remaining electrons beneath that accretion shock then combine with the remaining protons to leave only a warm neutron gas deep in the core. This new *neutron star* then cools and contracts rapidly (in only a few seconds) by emitting neutrinos, not photons.

The Race between Neutrino Heating and Infall

It is this neutrino emission that is, indirectly, responsible for the explosion. The neutrons deep in the core are very opaque to neutrinos, so that radiation leaks from a “neutrinosphere” in a manner very similar to photons leaking from the photosphere of a normal star. Above the neutrinosphere the material becomes increasingly optically thin, so eventually most neutrinos escape the core and carry away most of the energy that will bind the *cold* neutron star ($\frac{3}{5} GM^2/R \sim 3 \times 10^{53} \text{ erg}$). However, a small amount of this radiation is absorbed by the protons and neutrons just behind the stalled accretion shock, creating a hot proton-rich bubble that can “revive” the shock and possibly drive an explosion. Whether or not this happens depends on who wins in an important timed race: the remaining infalling 8–10 $M_{\odot}$ of Mg/Ne/O/C/He/H *above* the shock, or the neutrinos that are heating the protons and neutrons *beneath* the shock. If the neutrinos win the race and heat the bubble enough, the bounce shock will be revived and expand outward in a neutrino-heated

proton/neutron wind. On the other hand, if the infalling material wins the race and gets there before a hot wind develops, the central neutron star will grow to several solar masses and be crushed into a black hole.

Explosion!

In the case of the intermediate-mass stars, the neutrino heating wins. Enough energy is deposited beneath the shock before the relatively light and “low” density mantle can crush the neutron star. The proton/neutron wind pushes the shock outward, driving away everything above the collapsing core. If this does not sound like an explosion, it really is not from the point of view of the central core. The neutrino heating and ejection takes place over 100–200 ms, which is a *long* time compared to the dynamical time of the central core (<1 ms). To the nascent neutron star, this appears as a steady outflowing wind. However, to the outer stellar mantle and envelope, with dynamical times of many seconds to hours, and to observers on earth, this ejection looks like an explosion. There have been suggestions that the Crab supernova, seen by the Chinese in 1054, was created by the core collapse of a $\sim 10 M_{\odot}$ star.

Nucleosynthesis

As far as nucleosynthesis of elements lighter than iron is concerned, the intermediate-mass stars enhance only the helium abundance and little else. However, these stars may be important for creating some intermediate elements beyond the iron peak, such as molybdenum and ruthenium. But this still leaves the vast majority of the elements heavier than oxygen unexplained, and we have already examined stars up to $11 M_{\odot}$ – about 99% of all stars in the universe. We are rapidly running out of stars to explain how the elements are synthesized.

10.1.2.6 Massive Stars ($11 M_{\odot} < M_{\star} < 100 M_{\odot}$): Producers of Almost Everything Else

Fortunately, there *is* a stellar mass range that is responsible for making most of the elements, most of the existing neutron stars, *and* all of the stellar-mass black holes (SBHs). These are the massive stars (high-mass B stars [B0, B1] and all O stars), with lifetimes ranging from a few tens of Myr down to only 3 Myr for the most massive O stars. Unfortunately, after more than forty years of trying to make sense of these rare but important monsters, we still do not understand fully how they accomplish all of these tasks. To be sure, we *do* understand how they cook the heavy elements and, in almost all cases, how they do so in order to create the abundances of those elements that we measure in our Galaxy and others. However, we do not understand exactly how these stars explode, dispersing those elements. And we do

not fully understand which ones leave neutron star remnants and which ones form black holes.

Non-Degenerate Neon Burning through Iron Core Collapse

A massive star is defined here as one in which the ONeMg core reaches the ignition temperature for neon when the core is still *non*-degenerate. This then triggers the development of additional neon, oxygen, and silicon burning shells, in an increasingly rapid succession, and the formation of a growing degenerate *iron* core. The entire sequence develops roughly in the following manner:

- H-burning to helium lasts for 3–30 Myr; after H is exhausted in the core, helium ignites and, from now on, H-burning occurs in a shell around the helium core.
- He-burning to carbon and oxygen lasts for another 0.5–3 Myr; after He is exhausted in the core, carbon ignites; He-burning persists in a shell around the carbon–oxygen core.
- C-burning to neon and magnesium lasts for 1–2 *thousand* years; after C is exhausted in the center, neon ignites in the oxygen–neon–magnesium core; a C-burning shell is now added.
- Ne-burning to oxygen, magnesium, and silicon lasts only about 1 *year*; after Ne is exhausted in the center, oxygen ignites in the oxygen–magnesium–silicon core; a Ne-burning shell is added as well.
- O-burning to sulfur and silicon lasts only about a year as well; after central O is exhausted, silicon ignites in sulfur–silicon core; an O-burning shell still remains.
- Si-burning to iron and nickel lasts *only a few days to weeks*; after central Si is exhausted, a Si-burning shell persists, and the iron–nickel ashes collect in the center to form a degenerate iron core.

The “iron peak” elements generally are stable to both fusion and fission, so they never ignite. (Elements below the iron peak fuse toward iron, and those above the peak tend to fission toward iron.) The iron core then simply grows rapidly (in only a few hours!) to the Chandrasekhar mass limit, reaching the same state as the ONeMg core reached at the end of the intermediate-mass stars’ lives.

The subsequent evolution of the iron core, then, is precisely the same as that of the core in the intermediate-mass stars: it suddenly and violently collapses in a fraction of a second, forming a warm neutron star core at its center. That core cools via neutrino emission, and those neutrinos have the potential of heating the proton/neutron gas behind the shock front.

Explosion?

However, the final outcome of the core collapse in massive stars is not as clear as it was in the case of intermediate-mass stars. There is ample observational evidence

that many, if not most, stars in this mass range explode, particularly the more numerous (11–21 $M_{\odot}$) ones:

- The heavy element abundances created in the mantles and envelopes of these stars during all the different burning phases closely match the abundances of elements that are observed in our Galaxy and others. This can be the case *only* if those elements are ejected in some sort of outflow so they can enrich the Galactic interstellar medium.
- Moreover, the additional elements synthesized in the explosive shock wave by rapid neutron capture (the so-called “r-process”) *also* match cosmic abundances.
- Furthermore, the large number of pulsars and accreting neutron stars discovered by X-ray astronomers is consistent with most of these stars leaving $\sim 1.4 M_{\odot}$ neutron star remnants instead of black holes.
- Finally, the large historical and on-going observations of supernovae (Types Ib/c, II) that occur in the presence of very young stars (in our Galaxy and in others) is striking evidence that massive O and B stars do, in fact, explode.

Yet, despite our certainty that it is the collapse of the iron core that triggers the explosion of massive stars, after four decades of effort by hundreds of astrophysicists in this field, it still is not understood how that collapsing core channels the binding energy it releases into a supernova explosion. That is to say, to date *for all spherical numerical simulations of the process* (when the physics of degenerate iron, neutrons, neutrinos, relativistic hydrodynamics, *etc.* are properly taken into account) *no explosions take place inside the computer models for this mass range.*

What is the problem? There are $10^{53.5}$ ergs of binding energy released in the collapse, and supernovae need only 1/300th of this to explode. The problem is the same one discussed in the last subsection on intermediate-mass stars: the competitive race between the infalling $\sim 10\text{--}20 M_{\odot}$ of Si/Ne/Mg/O/C/He/H that is trying to crush the core into a black hole and the neutrino heating of the nucleons behind the stalled shock that is trying to revive it into an outflowing wind. So far, in the computer models, the infall wins for the massive stars, and *only black holes form* – in disagreement with the observations.

Possible Solutions to the Supernova Problem

Solutions with Neutrino Heating. A great many supernova theorists still believe that neutrino heating causes explosions, but that some of the simplifying assumptions in the computer models (sphericity, no rotation, no magnetic fields, *etc.*) are preventing the simulations from heating as rapidly as real stars do. Here are some ideas that have been, or are being, tried in an attempt to solve this problem:

- *Non-rotational, non-magnetic effects.* In the 1990s, it was recognized that convection should naturally develop in the heated proton/neutron bubble, leading to greater heating. However, this does not create enough additional heating in two-dimensional simulations to drive an explosion. In the early part of the following decade, it also was shown that the bouncing core could develop global sloshing

oscillations called the “standing accretion shock instability” or SASI. Because this prolongs heating by neutrinos, it may allow the hot bubble to win over the infalling mantle and create an explosion. A full treatment in 3-D probably will be necessary to settle this question. The introduction of 3-D effects into the (unrelated) field of Type Ia supernovae has solved significant problems and also might do so in the field of *core-collapse* supernovae.

- *Rotational and two-dimensional effects.* While rotation alone can slow the collapse of the mantle onto the neutron star, it also would slow the collapse of the core itself even more, thereby actually reducing the neutrino luminosity and heating. It does not appear that these effects alone can solve the problem.
- *Magnetic effects in 2-D and 3-D.* Even moderately strong magnetic fields could help enhance the neutrino heating. Begelman [275] has shown (in another context) that pockets of magnetic bubbles in a fluid could allow radiation to escape at a high rate, even greater than the Eddington limit in some cases. In the case of the new neutrino-radiating neutron star, it is not necessary to approach the neutrino Eddington limit in order to create an explosion – only enhance the neutrino luminosity by a factor of a few in order to increase the heating. Thus, neutrinos streaming out of magnetized bubbles could contribute significantly to the neutrino heating and, hence, the proton/neutron wind.

Explosion Mechanisms Other than Neutrino Heating. Taking the view that perhaps neutrino heating is *not* the main explosion mechanism, several investigators have suggested more exotic, but quite plausible, alternate mechanisms:

- *Acoustic effects.* Recently Adam Burrows, then at the University of Arizona and Steward Observatory, suggested that the core develops large-amplitude motions that generate strong sound waves. These waves then propagate to the stalled shock and deposit their energy there as heat, driving the proton/neutron wind in that manner. This mechanism is reminiscent of an early mechanism for heating the solar corona via acoustic waves, driven by motions in the solar atmosphere. (At present, the heating of the solar corona is now believed to be accomplished by Alfvén waves, which can travel great distances before dissipating as heat.) This is a new possible mechanism that only has begun to be explored in two-dimensional simulations. Like other concepts, it will certainly evolve as the field moves to three-dimensional simulations and adds magnetic fields and rotation.
- *Rotation and magnetic effects.* In 1970, James M. LeBlanc and James R. Wilson of Lawrence Radiation Laboratory (now LLNL) simulated the collapse of a rapidly-rotating iron core that *also* began with a weak magnetic field [373]. The resulting collapse created a magnetic explosion that exited the rotational poles of the core, with enough energy to eject an overlying stellar mantle and envelope. Subsequent work by this author and his colleagues [575] showed that this mechanism *cannot* be the reason that core-collapse supernovae explode. (The jet emanating from the neutron star core would overproduce r-process elements by a factor of 1000.) However, LeBlanc and Wilson’s calculation nevertheless opened up the possibility that the interaction of rotation and magnetic field could conspire to eject some supernova envelopes.

In 2002 J. Craig Wheeler, of the University of Texas at Austin, D. Meier, and James Wilson spearheaded an effort to understand the role of magnetic fields and rotation in supernova explosions [374]. Such early work on magnetic processes is comparable in level to the initial work performed on neutrino transport and heating in the 1960s and 1970s. Regardless of whether or not supernovae turn out to be powered by MHD effects, such studies will be very important to the overall understanding of supernovae and pulsar formation.

The basic mechanism of an MHD explosion proposed by Wheeler *et al.* is similar to the LeBlanc and Wilson one, but occurs well *outside* the proto-neutron star (PNS), where the material is still mainly iron. A warm, rapidly-rotating, magnetized PNS is assumed to form right after the core collapses, with a radius of ~ 50 km. Then, the PNS cools to a radius of ~ 10 km in a time of about one second. During this cooling, the PNS releases about four times more binding energy ($\sim 5 \times 10^{53}$ erg) than was released in the collapse itself, with most of that energy escaping as neutrinos. However, some of that binding energy will be converted into rotational energy of the final neutron star (NS):

$$\mathcal{E}_{\text{rot,NS}} = \frac{1}{2} I_{\text{NS}} \Omega_{\text{NS}}^2 \approx 2.4 \times 10^{52} \text{ erg} \left(\frac{P_{\text{NS}}}{1 \text{ ms}} \right)^{-2}$$

where $I_{\text{NS}} = f_{\text{NS}} M R_{\text{NS}}^2$ is the NS moment of inertia, and the geometric factor f_{NS} is a function of the internal density profile of the star. If angular momentum during this cooling is conserved ($I_{\text{NS}} \Omega_{\text{NS}} = I_{\text{PNS}} \Omega_{\text{PNS}}$), and if the density profiles of the neutron star and proto-neutron star are similar ($f_{\text{NS}} \approx f_{\text{PNS}}$), then the rotation periods of the neutron star and the PNS will be related by the square of their radii

$$P_{\text{NS}} \approx (R_{\text{NS}}/R_{\text{PNS}})^2 = P_{\text{PNS}}/25$$

However, we do *not* expect angular momentum to be conserved; instead, the combination of rotation and strong magnetic field should then drive a broad MHD bipolar outflow along the rotational poles (see Chapter 15), with the amount of energy liberated in that outflow of

$$\begin{aligned} \mathcal{E}_{\text{MHD}} &= \epsilon_{\text{MHD}} \mathcal{E}_{\text{rot,NS}} \\ &\approx 1.5 \times 10^{49} \text{ erg } \epsilon_{\text{MHD}} \left(\frac{P_{\text{PNS}}}{1 \text{ s}} \right)^{-2} \end{aligned}$$

We have assumed a $1.5 M_{\odot}$ NS, expressed the energy in terms of the original PNS rotational period, and applied an efficiency of conversion of rotational to outflow energy of ϵ_{MHD} . The outflow velocity will be of order the escape speed from the core surface ($\sim 0.5 c$), so the amount of iron in the jet will be

$$M_{\text{jet}} \approx \frac{\mathcal{E}_{\text{MHD}}}{\frac{1}{2} V_{\text{j}}^2} \sim 6.7 \times 10^{-5} M_{\odot} \epsilon_{\text{MHD}} \left(\frac{P_{\text{PNS}}}{1 \text{ s}} \right)^{-2}$$

The constraints on the supernova mechanism are that 10^{50-52} erg must be produced, but (in order to match the cosmic iron abundance) no more than $\sim 0.01 M_{\odot}$ of the iron core can be ejected into the envelope which will be ejected into the Galaxy. (The latter condition doubles the lower limit on the range of allowed proto-neutron stellar periods.) This produces a fairly narrow range of proto-neutron star rotational periods that would allow this mechanism to work:

$$80 \text{ ms } \epsilon_{\text{MHD}}^{1/2} < P_{\text{PNS}} < 400 \text{ ms } \epsilon_{\text{MHD}}^{1/2}$$

This range is consistent with the simulations of core-collapse in rotating stars by Chris L. Fryer and Alexander Heger at Lick Observatory in 2000 and in more recent years, lending significant credence to the MHD explosion mechanism [375]. For reference, Woosley’s assumption (see Section 10.2 below) that 10^{51} erg is deposited in each of his supernova simulations corresponds to a proto-neutron-star period of ~ 120 ms in this model.

Black hole formation also occurs in this mechanism in two possible ways. In the absence of a working neutrino-heating process, for very slowly-rotating cores ($P_{\text{PNS}} \gg 10$ ms), the bipolar outflow will not be energetic enough to eject the mantle, resulting in mantle “fallback” and black hole formation, similar to how black holes are formed in current simulations. However, for very rapidly-rotating cores ($P_{\text{PNS}} \ll 2$ ms), the bipolar jet will be produced in less than a dynamical time, be particularly energetic, *highly* collimated, and will “punch through” the mantle, rather than couple well to it. Too rapid a rotation of the core, therefore, also could lead to the direct formation of a black hole.

A team led by Alexander Heger of the Los Alamos National Laboratory has followed the evolution of massive stars *with* rotation and magnetic fields [376]. They find neutron core periods ranging from 15 ms for stars with mass $\sim 12 M_{\odot}$, down to 3 ms or less for $35 M_{\odot}$ stars. Given the uncertainties, this is tantalizing close to what would be needed to turn this mass range into MHD supernovae, even *without* a working neutrino heating process.

There is an important point to remember here. In order to compare theoretical pulsar periods with observed ones in young pulsars, one *must* take into account the production of an MHD jet by the spinning neutron core. The production of this jet in the first few hundred milliseconds will sap much of the core’s rotational energy, turning a 5 ms period into 10 or 20 ms. Whether supernovae have an element of MHD power or not, *the origin of pulsar periods must take MHD processes into account in the collapse and post-collapse phase.*

The Author’s Personal Synthesis of the Supernova and Neutron Star Production Problem

Despite the immensely complex nuclear physics that has gone into supernova simulations over the past several decades, the dynamics of the simulations are still rather simplified. Only recently have two-dimensional simulations become a serious part

of the field, and these have not yet considered the effects of rotation and magnetic fields in any significant manner. Three-dimensional simulations are even further in the future. So, given the large number of interesting physical processes that could enhance the neutrino heating, it is premature to rule out this heating as the main supernova explosion mechanism.

On the other hand, the neutrino heating mechanism has a lot to overcome. Observations strongly indicate that stars in the mass range $11\text{--}21\text{ M}_{\odot}$ synthesize and disperse the heavy elements *and* create pulsars and other neutron star remnants. Unless a *very good* fix for neutrino heating is discovered, this mechanism will have even greater problems ejecting $20\text{ M}_{\odot}$ envelopes than it has ejecting $10\text{ M}_{\odot}$ ones. To this author, there seems to be at least an even chance that something like the MHD bipolar outflow mechanism not only will be required, but will naturally occur in the simulations once multidimensional motions and magnetic fields are taken into account properly. If so, then the mechanism might help explain many other properties of pulsars and magnetars (the origin of their magnetic fields, kick velocities, *etc.*).

It would be esthetically pleasing if it turned out that core-collapse supernovae also form MHD bipolar outflows during stellar death. Such an event is very common in stars both more massive and less massive than the $11\text{--}21\text{ M}_{\odot}$ mass range. There is not much dispute these days that stars *more* massive than this create such jets when they form black holes in a γ -ray burst explosion. And, it also is becoming apparent that the planetary nebulae produced when lower-mass stars die *also* are bipolar outflows. If core-collapsing stars also produced bipolar jets when they died, *all* stellar death involving contraction or collapse would be linked by a simple mechanism involving angular momentum conservation and rapid spindown due to magnetic jet production.

10.2 Formation of Stellar-mass Black Holes in *Single Massive Stars*

The theory of formation of true black holes of stellar mass (like those in Cygnus X-1 or GRS 1915+105) is simply an extension of the theory of the formation of neutron stars in core-collapse supernovae. Indeed, as mentioned in the last section, it currently is easier *theoretically* to form black holes than neutron stars in supercomputer simulations. However, neither the mass nor the number of those black holes formed in supercomputer models agree with actual observations.

There are strong reasons to believe that neutron stars are made in the lighter massive (early B and late O) stars, while black holes themselves are made in the heavier massive (early O) stars. The greater numbers of neutron stars *vs.* black holes (discussed in Chapter 3) correspond to the greater numbers and higher death rate of the lighter massive stars. Black holes and heavy O stars both are equally rare. Furthermore, whatever the explosion mechanism, it must become increasingly difficult to eject stellar mantles of higher mass (*e.g.*, $20\text{--}99\text{ M}_{\odot}$) than those of lighter mas-

sive stars ($10\text{--}20\text{ M}_{\odot}$). Therefore, there must be some demarcation mass $M_{\star\rightarrow\bullet}$, above which massive stars make black holes and below which they leave neutron star remnants.

10.2.1 Mass Loss and the Production of Wolf–Rayet Stars

This simple picture, however, is complicated by stellar wind mass loss in massive stars. The stellar winds emanating from O and B stars are very strong ($10^{-6}\text{--}10^{-4}\text{ M}_{\odot}\text{ yr}^{-1}$). So, even in the short 3–30 Myr lives of these stars, enough mass can be lost that a $20\text{ M}_{\odot}$ star can shrink to a $15\text{ M}_{\odot}$ one, a $30\text{ M}_{\odot}$ star can become a $10\text{ M}_{\odot}$ one, and a $60\text{ M}_{\odot}$ star can be reduced to only $8\text{ M}_{\odot}$, even before the end of helium burning. If the stars are in binary systems, the mass loss can be even greater. An O star primary can lose its entire H envelope to a lesser companion in the first red giant phase (H shell and He core burning). The remaining helium star then will lose even more mass via hot stellar winds. Indications are that all stars above $\sim 20\text{ M}_{\odot}$ that are in binaries may be reduced to only $\sim 4\text{ M}_{\odot}$ objects before they undergo core collapse.

This is believed to be the origin of the Wolf–Rayet stars. They probably were originally O stars in binary systems that lost much of their outer envelopes and mantles. The major types are as follows:

- *WN stars* (N for nitrogen): These are stars that have lost most or all of their hydrogen envelope, exposing the helium-burning central regions. When these stars explode, they will appear as Type IIb supernovae (strong helium and weak hydrogen lines detected).
- *WC stars* (C for carbon): These have lost most or all of the helium mantle as well, exposing the carbon-burning regions. They produce Type Ib supernovae.
- *WO stars* (O for oxygen): These have lost much of the carbon mantle, exposing the oxygen-rich central regions. WO stars may be in the advanced stages of carbon-burning and, therefore, have only hundreds of years left before their cores collapse. They are believed to produce Type Ic supernovae.

For single stars, where the mass loss is less severe, the star will explode in the red supergiant (RSG) phase with an extended hydrogen envelope. These explosions will appear as one of the Type II supernova varieties (P, L, n); see Chapter 3.

The exact mass loss mechanisms and rates are still uncertain. Therefore, it is quite possible that the mid-range massive stars (say, $21\text{--}50\text{ M}_{\odot}$) make black holes, but those beginning with even more mass (say, $50\text{--}80\text{ M}_{\odot}$) leave neutron stars, because the latter have shrunk to only $\sim 4\text{ M}_{\odot}$ helium stars. As the reader can see, while much has been learned in the last few decades, this still is a very active field.

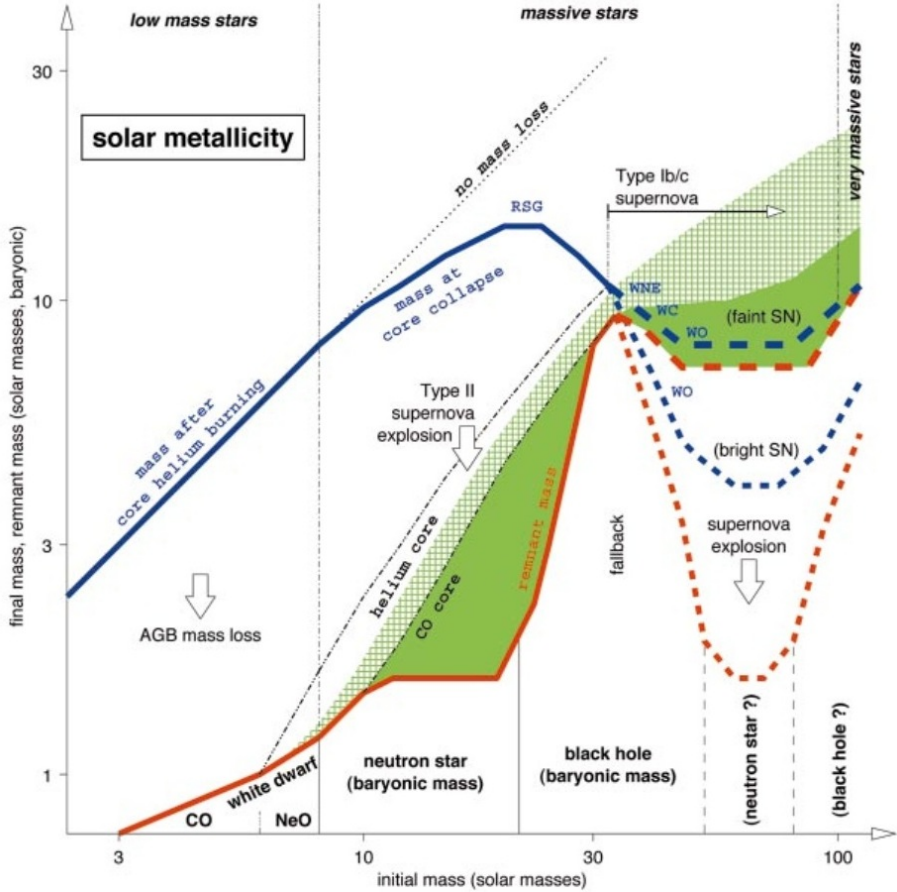


Fig. 10.1: Final stellar and remnant mass at end-of-life vs. initial mass on the main sequence, for stars that begin with a solar-level metallicity. On the vertical axis, the thick blue line gives the final stellar mass just before end-of-life, while the thick red line gives the final remnant mass after envelope ejection, plotted as a function of initial stellar mass on the horizontal axis. The latter curve shows that low-, solar-, and moderate-mass stars ($\lesssim 9 M_{\odot}$) form white dwarfs of increasing mass. Type II supernova explosions and neutron star remnants of $\sim 1.4 M_{\odot}$ are created beyond that (up to at least $\sim 21 M_{\odot}$), and black hole remnants as large as $10 M_{\odot}$ are created by stars with $\lesssim 35 M_{\odot}$. The fate of stars that begin with masses above $\sim 35 M_{\odot}$ is not fully understood yet and depends critically on the effects of strong mass loss in the Wolf–Rayet stage. (Note: the label “WNE”, along the dashed blue line, refers to the type WN star discussed in the text.) While all of these are expected to end as a Type Ib/c supernova, the character of the remnant (black hole or neutron star) is quite uncertain. Moderate mass loss (dashed lines) still leaves black holes for all massive stars, while heavy mass loss can reduce the original star mass so much that it leaves only a neutron star when it explodes (dotted lines). These results are based on numerous simulations of the final stages of stellar evolution, reviewed in [372]. Reprinted from Fig. 16 in that review with permission; copyright (2002) by the APS.

10.2.2 *Current Best Theoretical Estimates for Black Hole Formation*

With the above caveats, Woosley, Heger, and Thomas A. Weaver have undertaken the large task of investigating black hole formation in stars. Because the stellar explosion mechanism is still uncertain, these investigators have had to *assume* that enough energy is deposited behind the core-collapse accretion shock to keep all stars from forming black holes. The position of the energy deposition in the collapsing core was chosen to be consistent with both neutrino physics and with the nucleosynthesis arguments above. The amount of energy deposited was equivalent to that of a typical supernova: $\sim 10^{51}$ ergs.

Fig. 10.1 shows a summary of their results from a very large number of studies. The abscissa gives the initial mass M_* , while the ordinate shows the mass of the entire star *after* He-burning and mass loss (blue line) and also the mass of the compact stellar remnant left behind after death (red line). Also shown is the mass of the He mantle and that of the C/O core and heavier element material (all of which are ejected in a supernova explosion). In stars with masses less than $M_* \leq 21 M_\odot$ the injected 10^{51} ergs is sufficient to explode the stars, and stellar wind mass loss is also of minor importance. These stars all explode as red supergiants and leave neutron star/pulsar remnants.

In the mass range $21 M_\odot \leq M_* \leq 35 M_\odot$ the stars *partially* explode, ejecting their rather diminished H envelopes and some of their C/O cores. However, as M_* increases, more and more of the heavy C/O cores fall back onto the newly-born neutron star, crushing it into a black hole that accretes all the fallback material. The typical black hole mass of stars in this range will be about $M_\bullet \approx 3\text{--}4 M_\odot$, dominated more by stars in the lower portion of the range. In the upper part of this range ($30\text{--}35 M_\odot$), the entire C/O core ends up in the black hole, contributing little to nucleosynthesis but immediately forming a black hole of somewhat larger mass ($8\text{--}9 M_\odot$).

For $M_* \gtrsim 35 M_\odot$ the wind mass loss is enough to completely remove the H envelope and more, producing an *isolated* Wolf-Rayet star without the need for mass transfer to a binary companion. The effects of the uncertain mass loss rates in the Wolf-Rayet stage now are quite apparent in Fig. 10.1. For low mass-loss rates, black holes continue to be formed by all stars larger than $\sim 21 M_\odot$. However, if the higher estimates of Wolf-Rayet mass-loss rates are correct, then in the mass range $50\text{--}80 M_\odot$, the remaining small $4 M_\odot$ He star may be able to make only a neutron star.

To be sure, these results are rather preliminary. The details certainly will change as the core-collapse explosion mechanism is better understood. They also will become richer and more varied as two-dimensional and three-dimensional effects are taken into account. nevertheless, this general picture is consistent with observations of element abundances, supernova rates, pulsar populations, and black holes. Even with current uncertainties, this story is likely to be a good approximation to the truth.

10.3 Formation of Neutron Stars and Stellar-mass Black Holes in *Binary* Star Systems

Because many neutron stars and black holes have been found in *close* binary systems, the study of how *two* stars evolve together, and affect each others' lives, has been of great interest in the last three decades. The transfer of mass between the two stars can significantly change how they go through their lives, and change how and when they die. There are generally four types of mass transfer recognized, in order of increasing relative binary separation, and all of these, except Case D, refer to stages early in the binary evolution, *before* the primary (more massive star) dies:

- *Case A*: In Case A mass is transferred from the primary to the secondary star when the former is still on the main sequence. Since the primary is more massive, it generally will be larger than the secondary as well.
- *Case B*: In Case B mass transfer does not occur until after H shell and He core burning begin (in the first red giant stage). It always will be the primary star that expands into a red giant first, as the more massive star evolves faster.
- *Case C*: In Case C mass transfer begins only after He shell burning begins (in the supergiant phase).
- *Case D*: Mass is not transferred before the primary star dies.

Case A corresponds to very close binary systems in which both stars share a common envelope. Case D is the opposite extreme, in which stars in wide binaries evolve essentially independently in the same manner as single stars. Most stars that form X-ray binaries are believed to go through Case B mass transfer – close enough to be affected by mass transfer in the giant phase, but not in the main sequence phase.

10.3.1 Formation of High-Mass X-Ray Binaries and Binary Relativistic Stars

The standard evolutionary scenario for forming a high-mass X-ray binary is shown in Fig. 10.2. The binary begins with at least one massive star and one of moderate mass or higher. After a few million years, the massive star exhausts H in its core, expands into a red giant, transfers its entire envelope to the secondary, and becomes a WN star. The secondary, if not already a massive star, now becomes one as the primary has donated $\gtrsim 6 M_{\odot}$ of hydrogen to the smaller star. Being near the end of its life, however, the WN star soon explodes, forming a neutron star or black hole, depending on exactly how massive that helium star is (see previous section and Fig. 10.2). This leaves a compact object orbiting a massive O or B star.

Because the original secondary is now a massive star, its powerful wind alone will be enough to kindle X-ray emission as the compact object orbits through, and accretes from, that wind. Furthermore, when the secondary itself expands into a red giant, the mass transfer will be so rapid that the X-rays may be quite soft, like a

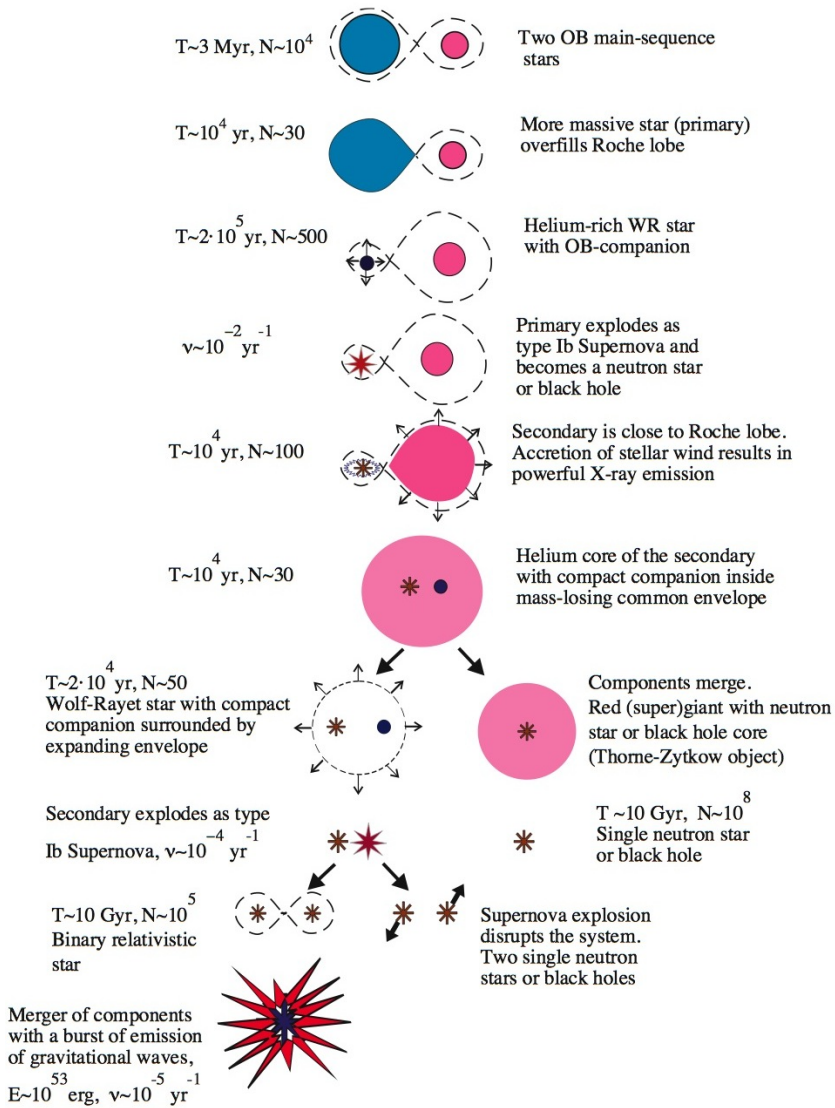


Fig. 10.2: Evolution of high-mass binary stars and the various observable phenomena that they can create. Initial binary formation with both stars on the main sequence begins at top, leading to several final fates at bottom. The first compact object (black hole or neutron star) is formed in the fourth stage, possibly producing a long-duration GRB in the process. The fifth stage shows a typical HMXB (like Cyg X-1 or Cen X-3). Between the fifth and sixth stage there is another that is not shown, in which the secondary fills its Roche lobe and begins rapid mass transfer; this should produce a highly super-Eddington object (super-soft X-ray source or “Cyg X-1 on steroids”), the model for which we discuss in Section 16.2.4. After significant angular momentum loss, the binary shrinks inside a common envelope and ejects most of the latter’s mass. This very compact binary can lead to a second supernova, a relativistic binary (BH+BH, BH+NS, or NS+NS), a gravitational wave-producing merger, or possibly a short-duration GRB if a neutron star is involved in that merger. Reprinted from Fig. 4 of [327], with permission of K. Postnov and the MPG.

ULX source, or quenched altogether, as in SS433. The details of how both slowly- and rapidly-accreting stars appear are discussed in Chapters 12 and 13, respectively.

Without some additional evolutionary steps, however, the above scenario for forming HMXBs will not work for other types of close binary systems. In fact, there are *many* classes of close binaries that have orbital separations much smaller than the radius of a red giant or supergiant and yet have one or even two remnants of evolved stars (white dwarfs, neutron stars, black holes). This indicates that one or both of these remnants' parent stars went through giant or supergiant phases.

The key to explaining how these binary orbits become so small actually lies in this paradox: at some point in the past the two stars must have shared a common giant or supergiant envelope when the binary had a much larger separation. That envelope then absorbed the energy and angular momentum of the wider binary, allowing it to shrink as the heated envelope expanded and was ejected, carrying away that energy and angular momentum.

There are several methods for estimating how the binary evolves in this common envelope phase. The simplest was proposed in 2005 by Gijs Nelemans and Christopher Tout of the Institute of Astronomy in Cambridge, UK [377]. The loss of angular momentum per unit time is given approximately by the amount of mass lost, assuming that the angular momentum per unit mass is approximately uniform

$$\frac{\Delta J_B}{J_B} = 1.5 \frac{\Delta M_{\text{Env}}}{M} \equiv 1.5 \delta m$$

where the factor of 1.5 was determined from more detailed simulations. That is, if much of the mass of the original binary is lost with the envelope ($\Delta M_{\text{Env}} < 0$), then the binary can shrink to an orbit whose angular momentum is only

$$J_B = [GM a (1 - e^2)]^{1/2} \mu$$

where M , μ , a , and e are the usual binary orbital parameters (total and reduced mass, semi-major axis, and eccentricity). If we denote with primes their new values, assume circular orbits ($e = e' = 0$), assume that most of the mass is lost by the primary only (*i.e.*, $M'_2 \approx M_2$), and assume that the binary is *initially* dominated by the primary ($M \approx M_1$, $\mu \approx M_2$), then the new binary separation and period are

$$a' \approx a (1 - 1.5 \delta m)^2$$

$$P' \approx P \frac{(1 - 1.5 \delta m)^3}{1 - \delta m}$$

That is, if much of the original binary mass is lost with the envelope ejection (*i.e.*, $\delta m \rightarrow 2/3$), then the binary can shrink to a very small separation and period.

Fig. 10.2 also shows a possible *future* common envelope evolution of an HMXB after the secondary becomes a red giant and engulfs the compact object. It can lead to a very close binary composed of two neutron stars, two black holes, or one of each.

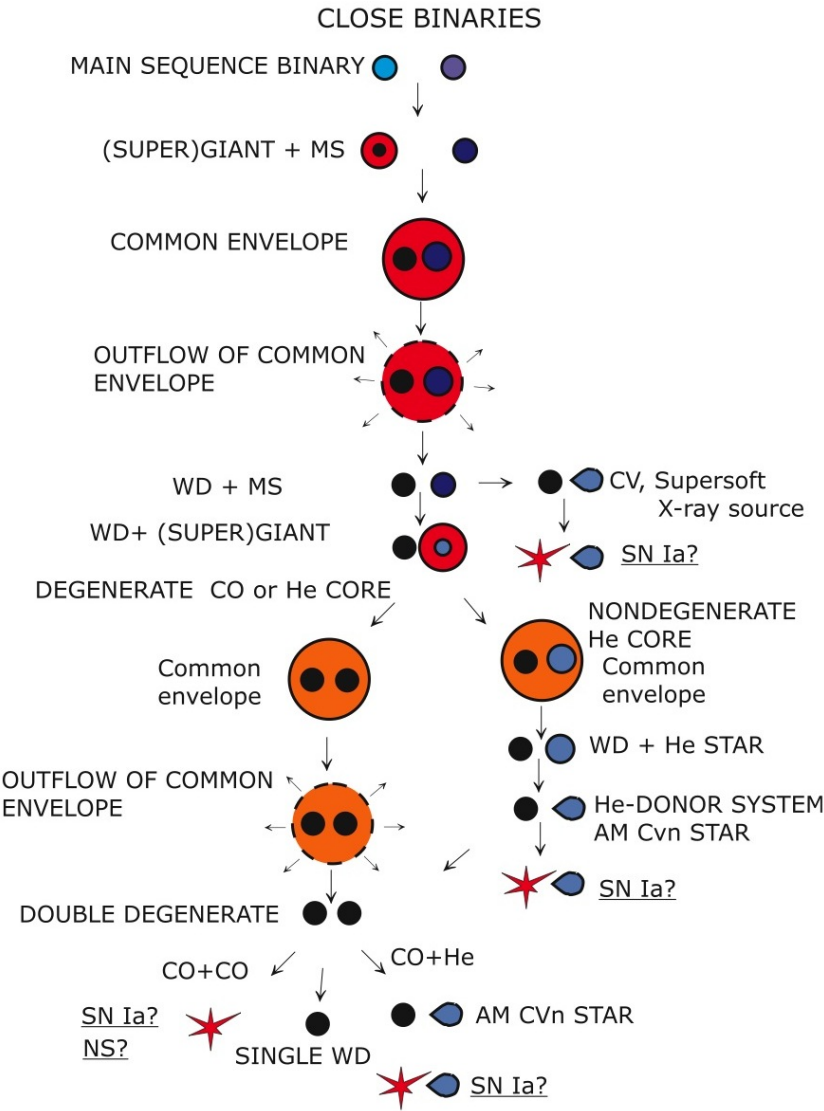


Fig. 10.3: Formation of close binary star systems that can lead to formation of neutron stars or black holes (bottom half of Table 10.1). As in Fig. 10.2 common-envelope evolution is an important process in shrinking the binary orbit and producing a very close system. None of these paths, however, lead to the standard LMXB system (low-mass main sequence star with a compact object). Reprinted from Fig. 5 of [327], with permission of K. Postnov and the MPG.

10.3.2 Formation of Close Binary Systems that May Form Neutron Stars and Black Holes

In Fig. 10.3 we show how a lower-mass binary, composed of a moderate to heavy mass star and a solar-type star, can produce cataclysmic variable binaries or, alternatively, “doubly-degenerate” (DD) binaries. The former are composed of one white dwarf and one main sequence or giant star that transfers matter onto the white dwarf. The latter type of binary is composed of *two* white dwarfs of varying types. During the evolution of the binary there can be one or two common envelope stages. The first occurs when the primary leaves the main sequence, leaving a white dwarf, neutron star, or black hole that can accrete from its main sequence companion. A second common-envelope phase can occur when the secondary leaves the main sequence, leaving a compact star accreting from a small helium star (the low-mass version of a Wolf–Rayet star).

Many of these final close binary systems have periods of less than half a day or so. This is so close that they will merge through gravitational radiation in less than the age of the Galaxy (see equation (8.37)) and will begin to deplete their current numbers. In a steady state, the current merger and birth rates of these binaries will be equal. Table 10.1 shows estimates of these rates (from Neleman’s work) for all possible close binaries involving white dwarfs, neutron stars, and black holes, where in line #6 the values 2.5×10^{-2} and $1.1 \times 10^{-2} \text{ yr}^{-1}$ mean one WD+WD binary born currently in our Galaxy every ~ 40 yr and one currently merging every ~ 90 yr, *etc.* While these numbers are uncertain by factors of 3 or sometimes even 10, the BH+BH merger rate is potentially even more uncertain. Here it has been assumed that it is as easy to get a BH+BH binary close enough to merge via gravitational waves as it is for NS+BH binaries to do so. Tables like this one are useful for estimating the rates at which LIGO and LISA will be able to detect gravitational waves from systems like these.

Table 10.1: Estimated birth and death rates for different types of close binaries

Binary type	Current birth rate (yr^{-1})	Current death rate (yr^{-1})	Final product
HMXB (NS)		(S)	NS/BH
HMXB (BH)		(S)	BH
LMXB (NS)		(C)	ms pulsar (NS)
LMXB (BH)		(C)	BH
CVs		(S)	Type Ia SN
WD + WD	2.5×10^{-2}	1.1×10^{-2} (M)	Type Ia SN, NS?
WD + NS	2.4×10^{-4}	1.4×10^{-4} (M)	NS, BH?
WD + BH	8.2×10^{-5}	1.9×10^{-6} (M)	BH
NS + NS	5.7×10^{-5}	2.4×10^{-5} (M)	BH
NS + BH	2.6×10^{-5}	2.9×10^{-6} (M)	BH
BH + BH	1.6×10^{-4}	$10^{-5}?$ (M)	BH

Table 10.1 also shows potential final products of these close binaries after either (1) the secondary star is consumed totally by the primary's remnant (C), (2) the binary becomes unbound after a supernova explosion (S), or (3) the compact objects merge (M). In the HMXB case, Fig. 10.2 shows some possible outcomes, leaving either a single neutron star or black hole.

10.3.3 Formation of Low-Mass X-Ray Binaries and Microquasars

Note that none of the above binary evolution scenarios produce an LMXB system (solar-type main sequence star plus compact object). Microquasars (low-mass star plus high-mass black hole) are particularly hard to form. While several models have been suggested (rapid common envelope evolution of two intermediate-mass stars, three-body interactions in a globular cluster that result in a main sequence star with a black hole), the formation of LMXBs is still an unsolved problem. At present, there is no good consensus on how these reasonably common neutron star and black hole binaries are produced.

If the dominant mechanism occurs in the galactic field (as opposed to in globular clusters), then the evolutionary path is likely to involve massive binary star evolution. After all, the formation of a neutron star or black hole *requires* that a massive star be formed at some stage in that evolution. Rapid angular momentum evolution in a super-Eddington wind phase, *after* a black hole has formed in the fifth stage of Fig. 10.2 and *before* the common envelope phase (sixth stage), seems the best way to grow a high-mass stellar black hole while shrinking the mass of the companion main sequence star, as well as the binary orbit itself.

On the other hand, if the compact companion is a neutron star, instead of a black hole, it can accept the accreting matter at only the Eddington rate ($\sim 10^{18} \text{ g s}^{-1}$ or $\sim 10^{-8} M_{\odot} \text{ yr}^{-1}$); the remainder of the enormous amount of mass being transferred will be driven from the system in a strong wind. Since the super-Eddington phase should last for much less than 10^8 yr , the neutron star has a very good chance of surviving to the LMXB stage and not being driven to collapse to a black hole.

10.3.4 Formation of Neutron Stars and Black Holes via Accretion and Mergers in Close Binary Systems

In LMXBs that undergo Case B mass transfer, the primary's remnant tends to consume its lower-mass companion entirely, but not before being spun up by that accretion. Neutron stars generally are *not* pushed over the black hole mass limit by that accretion, but they *are* spun up to millisecond periods, leaving an isolated rapidly-spinning (recycled) pulsar. Black holes are only spun up moderately as they consume their lower-mass companions, ultimately leaving a low j isolated black hole of several solar masses.

In Case C mass transfer, however, the LMXB likely will go into a common envelope phase when the secondary becomes a red supergiant itself, creating the WD+NS or WD+BH systems shown near the middle of Table 10.1. When a WD+BH system merges, it certainly will leave a single black hole. However, unless the neutron star in a WD+NS system is particularly massive and collapses to a black hole, the final product still will be a single, although more massive, neutron star.

When the remaining *compact* binaries (NS+NS, NS+BH, BH+BH) merge, all will leave a single black hole that is rotating moderately to very rapidly ($j \rightarrow 1$). The latter two cases obviously produce single black holes, but the NS+NS case merits more discussion. The NS+NS and NS+BH compact binaries both are potentially short-duration γ -ray burst progenitors. Given the much higher expected merger rate of the former, neutron star mergers have been studied lately with very detailed numerical relativity simulation codes. Such codes are *more* complex than the black hole merger codes discussed in Chapter 8: not only do they need to solve Einstein's equations properly, they also need to (1) solve the matter conservation laws properly, (2) properly use the matter field to generate gravity (the *inhomogeneous* Einstein field equations), and (3) form horizons within the matter stably in order to follow the late-time evolution as most of the material in the neutron stars eventually disappears into the newly-formed black hole. One of these super-codes has been developed by Mark Miller, of the Jet Propulsion Laboratory in Pasadena, California – a significant extension of the NASA “grand challenge” neutron star code he developed in the 1990s.

Fig. 10.4 shows the merger of two orbiting neutron stars and the formation of a new black hole as simulated by Miller [378, 379]. The evolution is complex, but not unlike the merger of two galaxies (although on a scale 18 orders of magnitude smaller). When the orbit becomes small enough, the two neutron stars disrupt each other, forming a rotating peanut-shaped mass and long tidal tails. The black hole forms as the peanut shrinks and becomes more spherical, leaving an accretion disk composed of iron, other heavy elements, protons, and neutrons surrounding the black hole. It can take a fairly long time for this disk to accrete into the hole (0.5–2 s, or hundreds of dynamical times). If the relativistic jet that this accreting black hole produces points toward the earth, or nearly so, it will be observed as a short-duration γ -ray burst. The merger of two neutron stars also was studied in the early 2000s [380, 381, 382]. Early indications were that the peanut may form a rapidly and *differentially* rotating super neutron star (one that is massive enough to collapse to a black hole, but does not because of the rotation) for much longer than a few dynamical times. The effect is similar to the orbital hangup encountered in BH+BH mergers (Section 8.2.3) when the spins of the two black holes are aligned: there is too much angular momentum for the system to collapse immediately to a single black hole. How the double-neutron-star system exactly evolves through this similar phase will have important consequences for the production of gravitational radiation from merging neutron stars.

There are, however, some forces working against this scenario. We will learn in Chapter 12 that magnetic stresses and turbulence will build up in only a few rotation times, even if the initial magnetic field is weak. Not only will this reduce or eliminate

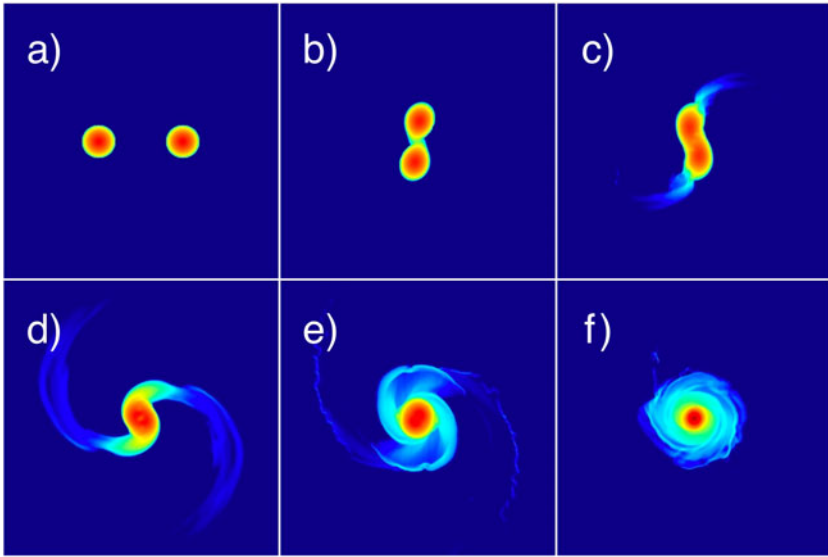


Fig. 10.4: Fully general relativistic evolution of the merger of two neutron stars of equal mass [378, 379]. (Compare with the black hole mergers in Figs. 8.5 and 8.6.) The stages are as follows: (a) initial binary; (b) first contact; (c) peanut-shaped single merged product forms; (d) and (e) the overmassive neutron star “rings down” by emitting gravitational waves *and* shedding mass and angular momentum in tidal tails; (f) an horizon has formed in the center of the merger product. Much later (seconds or so) nearly all the material will accrete into the new black hole because of angular momentum transfer in the accretion disk by the magneto-rotational instability (MRI; Section 9.6.3). Figure generously supplied by M. Miller and reprinted with his kind permission.

radiating quadrupole moments (which would generate strong gravitational waves), it also may transfer angular momentum outward rapidly, causing the two neutron stars to collapse to a black hole in only a few rotation times. The simulation codes to study the effects of electromagnetic stresses on merging neutron stars still are being built. A better understanding of these processes should be possible in the early 21st century. Observations of short-duration GRBs with LIGO then may help determine how important magneto-rotational-turbulent processes are in merging NS+NS systems and others that produce hyper-critical accreting black holes.

10.3.5 Cataclysmic Variable Binaries: Possibly Type Ia Carbon-Detonation Supernovae, Definitely not Type Ib/c Core-Collapse Supernovae

Cataclysmic variable or CV binaries are “singly-degenerate” close binary systems, composed of a non-degenerate (main sequence or giant) star transferring matter onto a white dwarf. They are s-named because they form the class that includes rather ex-

plosive objects: novae, recurrent novae, dwarf novae, nova-like variables, and symbiotic stars. The accretion rates are generally not that large (10^{-10} – $10^{-8} M_{\odot} \text{ yr}^{-1}$). However, even at those low rates, a half a solar mass of material or more can be transferred onto the white dwarf in less than the age of the universe, pushing the white dwarf to the Chandrasekhar mass limit of $1.4 M_{\odot}$.

Many of these accreting white dwarfs will be carbon–oxygen ones. And even helium white dwarfs are expected to be hybrid ones (C/O cores, He mantles) as they approach the Chandrasekhar limit. (Massive He white dwarfs should behave like cores of solar-type and moderate-mass stars, igniting helium and leaving carbon and oxygen ash in their centers.) And, as was briefly mentioned in the discussion on moderate-mass stars (Section 10.1.2), degenerate C/O cores that grow to the Chandrasekhar limit *do not undergo core collapse*. They should explosively ignite carbon in their interiors and obliterate the entire white dwarf. The picture of how this happens has been worked out by astrophysicists like J. Craig Wheeler at The University of Texas, Alexi Khokhlov of the University of Chicago, and Peter Höflich of Florida State University and their teams [383, 384].³

The explosion is quite complex, not a simple thermonuclear bomb. It begins as a carbon deflagration (subsonic, but rapid burning). This occurs at rather high densities ($10^{6-7} \text{ g cm}^{-3}$) but less than those needed to initiate core collapse via electron capture on heavy ions ($10^{9-10} \text{ g cm}^{-3}$, as occurs in intermediate-mass and massive star cores). The release of so much nuclear energy in the deflagration does expand the white dwarf, leaving pockets of burned and unburned carbon fuel inside that star. At some specific point in the white dwarf, the unburned fuel ignites explosively, sending a supersonic burning front (“detonation wave”) through the star. This process appears to make iron and radioactive nickel in the right amounts to explain both the amount of iron in the Galaxy and the light curve of Type Ia supernovae (as being powered by the decay of radioactive nickel).

The bottom line is that *accreting degenerate white dwarfs avoid core collapse and do not form neutron stars*. Instead, they explode as massive thermonuclear carbon bombs – a process quite different from the core-collapse-driven supernovae (Types Ib & c, and all types II) that make neutron stars and black holes. Type Ia supernovae light curves are so similar to one another that their *absolute* luminosity can be calibrated by measuring how their brightness rises and falls. This makes them very good “standard candles” for measuring distances to galaxies at high redshift. Their use in this capacity has led to a deeper understanding of the evolution of the universe as a whole, and the role that dark matter and dark energy play in the process.

There is, however, a problem with CV systems specifically as progenitors of Type Ia supernovae. In these systems it is *hydrogen* that accretes onto the white dwarf. Periodically this hydrogen burns into helium in a nova flash, but much of the burning products are ejected in the nova explosion. As a result, the white dwarf may never grow much in mass by this accretion process. Moreover, Type Ia supernovae are famous for showing *no* hydrogen in their ejecta. This means that the white dwarf,

³ The author had the privilege of working with Professor Wheeler and Dr. T. Mazurek on some early work in this area [385].

if it does explode eventually, must do so in such a way that none of the low-mass (hydrogen-rich) companion ends up in the explosion. This constraint is a tall order, and it makes it very difficult to argue that CV binaries are immediate progenitors of all Type Ia supernovae.

10.3.6 Double-Degenerate Binaries: Some Type Ia Supernovae, an Occasional Neutron Star, and Many Low-Mass Helium Stars

As with CV binaries, we are interested in doubly-degenerate binaries (those with *two* white dwarfs) because they do *not* produce neutron stars – at least not directly – even though we might have thought that they should. Nelemans and his group studied the evolution of many different binary star systems around the Galaxy that were expected to produce WD+WD systems [386]. They found the final distribution of the possible different types shown in Table 10.2. Furthermore, nearly half of these binaries (48%) will merge in less than 10^{10} yr and of those, about 1/12th (*i.e.*, $\sim 4\%$ of all the above systems) will have a combined binary mass larger than the Chandrasekhar mass. That is, 96% of all doubly-degenerate binaries will *not* undergo core collapse or explode.

Table 10.2: Estimated fraction of different types of doubly-degenerate binaries

DD type	Relative fraction of all DDs
He WD + He WD	53%
C/O WD + C/O WD	25%
C/O WD + He WD	20%
ONeMg WD + other WD	<2%

Of the remaining 4% that *will* merge to form white dwarfs approaching the Chandrasekhar mass, most likely will produce Type Ia supernovae (as discussed above), and rather cleanly at that. Because white dwarfs have *no* hydrogen, the merger of two of them, and subsequent carbon deflagration/detonation, would easily explain the absence of hydrogen emission lines in Type Ia supernovae ejecta. However, if the WD+WD merger rates per galaxy in Table 10.1 are correct, and if only 1/12th of those have combined masses $>1.4 M_{\odot}$, then we would expect a Type Ia supernova from merging white dwarfs at most once every 1000 years. This is 1/10th of the actual observed rate in galaxies like our own (one every 100 years). So it would appear that most Type Ia supernovae must come from other sources.

This creates a dilemma: CV binaries may have trouble exploding and, if they do, have trouble keeping their ejecta from being contaminated by hydrogen; and there do not seem to be enough doubly-degenerate binaries to create all the Type Ia SN either. Where, then, do Type Ia's come from? Paul Groot (private communication)

has a suggestion. The merger of a doubly-degenerate binary whose total mass is greater than the Chandrasekhar mass limit is the best *theoretical* model for Type Ia progenitors. Therefore, perhaps we are underestimating the number of these objects that are created periodically in the Galaxy. This is possible if we have *overestimated* the lifetime of such binaries and/or greatly *underestimated* the number of doubly-degenerates that exist currently. The birth and death rates depend on both of these estimates. If, for example, many such binaries are hidden inside common-envelope stars, which appear single but are not, and if those have a very short lifetime in the common-envelope phase, then the production rate may be much higher than has been estimated. Also, very short period doubly-degenerate binaries with high inclinations also may have been missed. One might argue that these have rather short lifetimes and so should be rare, but that is just what is needed to match the Type Ia SN rate.

In a rare instance, the merger of a massive ONeMg white dwarf with another white dwarf might trigger a core collapse, in a manner similar to the collapse of cores of intermediate-mass stars. This *would* result in a Type Ib/c supernova (no hydrogen lines) of some sort and *would* leave a neutron star or magnetar. If Neleman's rate estimates are correct, however, that would be a rare occurrence – once every 100,000 years or more.

The great majority of merging white dwarfs will *not* form neutron stars, as their total binary mass is below $1.4 M_{\odot}$. They will produce interesting objects, however: low-mass helium stars, in particular (which cannot form from interstellar gas by themselves) and fairly massive ($\geq 1.0 M_{\odot}$) C/O and hybrid white dwarfs. As interesting as these are, however, we see that they are not part of the story of black hole and neutron star formation. So we now will leave the subject.

10.4 Formation of Intermediate Mass Black Holes

The results for forming black holes from single stars can be summed up as follows: black holes of stellar mass are produced by massive stars with initial main sequence mass of $M_{\star} \gtrsim 21 M_{\odot}$; however, because of severe mass loss due to stellar winds, the mass of this remnant appears to never get above $M_{\bullet} \sim 10 M_{\odot}$, even when the parent star's mass exceeds $100 M_{\odot}$. This statement is true so long as the mass loss dominates the evolution of the star.

However, there are at least two circumstances when we believe the extreme wind mass loss can be avoided, allowing much larger ($100\text{--}10^4 M_{\odot}$) black holes to form. One case occurs in the early universe, shortly after the big bang, when the first stars are forming. Such “Population III” stars were created out of extremely metal-poor gas ($Z \sim 4 \times 10^{-10}$; see Section 10.1.1) and would not have had the absorption lines of carbon, nitrogen, oxygen, and iron that contribute so strongly to the radiation pressure force driving the powerful stellar winds in present-day massive stars.

The other case can occur nowadays in young star clusters, even when the gas forming those stars is metal-rich (Population I). When large numbers of O and B

stars collide and merge very quickly in the centers of those clusters, they can form a very large “superstar” that can evolve quickly to core collapse *before* the stellar winds have a chance to strip this monster of its outer envelope and mantle.

We will consider the Population III case first and then deal with the Population I monster star case after that.

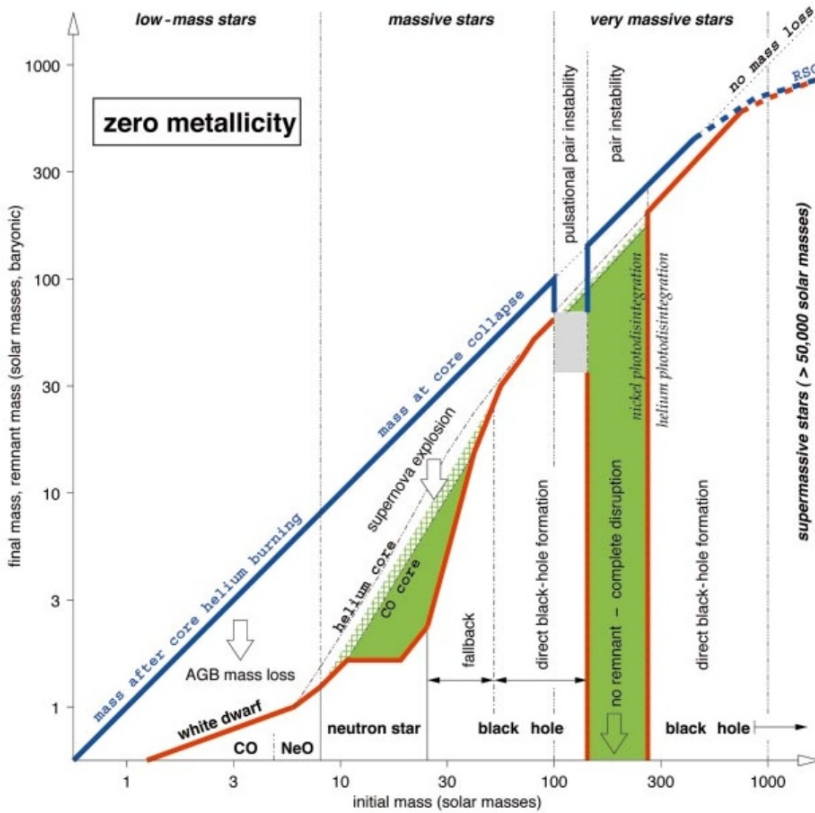


Fig. 10.5: Final stellar and remnant mass at end-of-life vs. initial mass on the main sequence, for stars that begin with a *zero metallicity* (Population III stars). Notation is same as in Fig. 10.1. Results for stars up to $\sim 35 M_{\odot}$ are similar to those in the high metallicity case, including the leaving of neutron star and black hole remnants with masses similar to the case in Fig. 10.1. Above $35 M_{\odot}$ initial mass, however, there is very little mass loss, until several hundred solar masses or more. Stars up to $1000 M_{\odot}$ or more can survive at their initial mass until they explode or form a black hole. In the range $140 < M_{*}/M_{\odot} < 260$ stars are obliterated in a pair-instability supernova explosion; the rest form a black hole almost as massive as the original star itself. Zero metallicity stars, therefore, can produce a reasonable quantity of IMBHs. The results are based on numerous simulations of the final stages of stellar evolution, reviewed in [372]. Reprinted from Fig. 12 in that review with permission; copyright (2002) by the APS.

10.4.1 Formation of IMBHs in the First Stars Created after the Big Bang

Fig. 10.5 shows another plot of remnant mass vs. stellar mass from the Woosley, Heger, and Weaver review article. It is similar to Fig. 10.1, except for one key difference. Fig. 10.5 shows the results of stellar evolution when the stars are formed out of gas with essentially *no* metals. The difference in results is quite striking: while stars above $\sim 20 M_{\odot}$ lose most of their envelopes and mantles when the metal abundance is high (Fig. 10.1), *stars with no metals remain virtually intact beyond masses of $1000 M_{\odot}$!*

10.4.1.1 Recap of $M_{\star} < 100 M_{\odot}$ for Population III Stars

Except for slight differences in the two figures, the results up to $M_{\star} \sim 50 M_{\odot}$ for zero-metallicity stars are very similar to those for stars with solar metallicity. Stars up to $\sim 8 M_{\odot}$ make white dwarfs, ones up to $\sim 25 M_{\odot}$ make neutron stars, and ones up to $\sim 50 M_{\odot}$ make black holes by fallback of much of the non-ejected C/O core onto the new neutron star. In the range $M_{\star} \sim 50\text{--}100 M_{\odot}$, however, the process of black hole formation is more dramatic: the collapse of the central iron core produces such a large neutron core that it collapses directly to a black hole rather than bouncing. Then the entire helium mantle collapses into it, forming a black hole of $30\text{--}50 M_{\odot}$.

10.4.1.2 Death of Very Massive ($100 M_{\odot} < M_{\star} < 1000 M_{\odot}$) and Supermassive ($M_{\star} > 1000 M_{\odot}$) Population III Stars

Above $M_{\star} \sim 100 M_{\odot}$ the cores of these very massive stars (VMSs) are subject to the pair instability. In this process, the temperature becomes high enough that $e^{\pm}$ pairs are created profusely in the C/O core. This use of thermal energy to create the mass of electrons and positrons causes the core to contract quickly, which then ignites the heavy elements in the core (oxygen or silicon) in a vigorous conflagration that becomes increasingly explosive for stars with larger $M_{\star}$. In the range $100 M_{\odot} < M_{\star} < 140 M_{\odot}$, these stars pulsate wildly, ejecting some of their helium mantle; but they do eventually collapse, leaving black holes of $40\text{--}50 M_{\odot}$ in size.

For even more massive Population III stars ($140 M_{\odot} < M_{\star} < 260 M_{\odot}$), the thermonuclear, heavy-element burning becomes explosive, obliterating even these monster stars. *No* black hole remnant is left by these massive explosions, but there is considerable nucleosynthetic enhancement of the interstellar medium in metals. The relative abundances of these heavy elements are rather close to those we see in the sun, with a few peculiar exceptions: there is a deficit of elements with odd atomic number (numbers of protons) and no elements heavier than zinc ($A = 30$) are produced. In addition, it is notable that up to $57 M_{\odot}$ of radioactive nickel is dis-

persed in each explosion. Because observed supernova light curves are powered by the decay of this nickel, pair-instability supernovae may be the brightest thermonuclear explosions in the universe. It has been suggested that the recent supernova SN 2006gy may have been of a star with mass greater than $150 M_{\odot}$, although that star must have been a massive Population I star, not a Population III one, as the redshift of its parent galaxy (NGC 1260) is only 0.019.

For zero-metallicity stars with $M_{\star} > 260 M_{\odot}$, the vigorous thermonuclear reactions, while still present, become ineffective against the weight of the large helium mantle and core ($>133 M_{\odot}$) crushing the core. Formally, it is the photodisintegration instability (not electron capture or the increased mass of thermal energy itself) that triggers the collapse toward a black hole in this mass range. Collisions of heavy nuclei in the core (and, in the most massive objects, collisions of helium nuclei) dissociates them into protons and neutrons. The loss of this heat from the internal core pressure is not made up by the pressure increase due to having many more particles available from the dissociation. So, the core collapses catastrophically. Any increase in thermal pressure during this collapse is never great enough to halt it and, in the late stages, that pressure energy only makes matters worse (see Section 8.3.3).

So Population III stars have the potential for forming black holes in two mass ranges: stellar mass ($M_{\bullet} \sim 3\text{--}50 M_{\odot}$) and intermediate mass ($M_{\bullet} > 133 M_{\odot}$). These sometimes are referred to as Population III.2 and III.1 stars, respectively. Of course, very massive and supermassive stars need to actually form out of low metallicity gas in order for the black holes in the intermediate mass range to be produced. We now will consider that prospect.

10.4.1.3 Formation of Very Massive and Supermassive Population III Stars

Up until now we have not needed to discuss the complex process of star formation. This is because, in the previous sections, we have dealt with present-day (Population I) stars. Because there are many observations of such stars, in young clusters as well as in the Galactic field, we know that the typical stellar mass is approximately $\sim 1 M_{\odot}$ and that the initial distribution of stars with mass is given approximately by an empirical relation called the *initial mass function* (IMF), such as the one derived by John Scalo of the University of Texas at Austin [387]:

$$\begin{aligned} \frac{dn}{d \ln m} &= m_{\star}^{-0.2 \pm 0.3} & 0.1 < m_{\star} < 1.0 \\ &= m_{\star}^{-1.7 \pm 0.5} & 1.0 < m_{\star} < 10 \\ &= 0.40 m_{\star}^{-1.3 \pm 0.5} & 10 < m_{\star} < 100 \end{aligned} \quad (10.1)$$

where $m_{\star} \equiv M_{\star}/M_{\odot}$. This IMF is plotted with the solid line in Fig. 10.6.

As of this writing, however, *no* Population III stars have been observed, let alone several young clusters or galaxies of such objects. Indeed, no Population III stars exist locally any more (not since perhaps 13–14 billion years ago), and current telescopes are not yet powerful enough to observe the first stellar systems at redshifts of

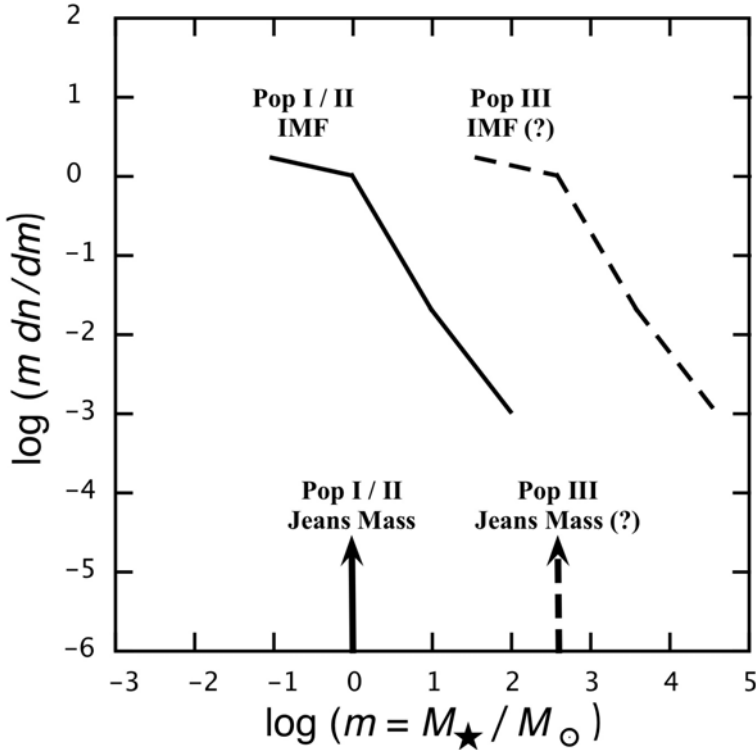


Fig. 10.6: The IMF for Pop I and II stars (solid curve), given in eq. (10.1), and a suggested one for Pop III stars in the early universe, based on the temperature and Jeans mass in equation (10.4) (dashed line). Whatever the actual Pop III IMF, it should have a feature at the predominant Jeans mass at that time, allowing much more massive stars (and, therefore, black holes) to form.

>10. Furthermore, we do not yet know enough to accurately estimate the IMF for Population III stars, so we will have to use a simple procedure based on our knowledge of Population I stars. Of course, this is only an educated guess, and it may change significantly as more is learned about star formation in the early universe.

The Properties of Population I Stars Formed at the Present Time

It is well known that a large mass of neutral hydrogen (H I) gas of density n atoms per cubic centimeter and temperature T will fragment into gravitationally contracting clumps whose radius is larger than the Jeans radius (named after Sir James Hopwood Jeans, who was a British astrophysicist and mathematician in the early 20th century)

$$R_J \simeq \frac{1}{2} \frac{GM_J \mu}{\mathcal{R}T} \quad (10.2)$$

where M_J is the mass enclosed within the radius R_J and μ is the mean molecular weight of the gas ($\mu \sim 1.6$) in units of grams per mole.⁴ Assuming that the ambient density outside the collapsing star is

$$\rho_a \approx \rho(R_J) = \frac{2}{\pi} \left(\frac{\mathcal{R}T}{G\mu} \right)^3 \frac{1}{M_J^2}$$

we can solve for the Jeans mass itself (that which is collapsing inside radius R_J)

$$M_J \simeq \left(\frac{2}{\pi n \mu m_p} \right)^{1/2} \left(\frac{\mathcal{R}T}{G\mu} \right)^{3/2} \quad (10.3)$$

where n is the atomic number density and m_p is the mass of the proton. For typical dense molecular cloud conditions in our Galaxy ($n \sim 3 \times 10^4 \text{ cm}^{-3}$, $T \sim 10 \text{ K}$), the Jeans mass is $\sim 0.97 M_\odot$, in very good agreement with the observed mass of a typical star.

This characteristic mass is labeled in Fig. 10.6 and appears to be a rough indicator of what the actual stellar mass distribution might be for newly-forming stars in our Galaxy. The number of stars more massive than M_J falls off rapidly (but not exponentially) with increasing mass. And, while the IMF flattens out below M_J , the number of lower-mass stars formed is still substantial. This may be due to additional fragmentation of some collapsing clouds into smaller clumps as the density n increases.

An Estimate of the Properties of Population III Stars Formed in the Early Universe

In the early universe, the first *dark matter objects* (not gaseous stars!) to fragment and collapse from the expansion had a mass of

$$M_{\text{gal}} \approx 1.45 \times 10^5 M_\odot \left(\frac{\Omega_m h^2}{0.13} \right)^{-1/2}$$

where $\Omega_m \approx 0.26$ is the fraction of matter (dark and luminous) in the universe and $h \approx 0.71$ is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Of course, an entire spectrum of these objects will form throughout the young universe, with a few reaching $10^{6-7} M_\odot$. Their physical sizes will be on the order of 10–30 pc. These objects are not stars, of course, but rather dark matter *mini-halos*, from which the first galaxies soon will form.

About one-fifth of the matter in these protogalaxies is made up of baryons (*i.e.*, hydrogen and helium mainly). This gas has virtually no metals and, therefore, no strong metal emission lines to rapidly cool that gas. Nevertheless, the emission lines

⁴ The derivation of this particular form of Jeans radius and mass uses a simple singular isothermal sphere (equation (5.78)), truncated at $r = R_J$, to determine the coefficient of $1/2$.

of hydrogen and helium are enough to promote some weak cooling. Simulations by three different cosmology groups in the late 1990s (Tom Abel and his colleagues, then at the National Center for Supercomputing Applications in Illinois [388]; Fumitaka Nakamura of Niigata University with Masayuki Umemura of Tsukuba University [389, 390]; and Volker Bromm and his colleagues, then at Yale University [391, 392]), all obtained essentially the same answer on how this baryonic gas would evolve. The gas falls toward the center of the small protogalaxy, settling into a rotating disk with a density of $n \sim 10^{3-4} \text{ cm}^{-3}$ and a temperature of 200–300 K. Unlike the dark matter in the mini-halo, this central disk of gas *can* form stars.

The Jeans mass in this primordial disk, however, is much larger than the Jeans mass in molecular clouds in our Galaxy

$$M_J \approx 380 M_\odot \left(\frac{n}{3 \times 10^3} \right)^{-1/2} \left(\frac{T}{250 \text{ K}} \right)^{3/2} \quad (10.4)$$

Because the typical Population III star should be about 400 solar masses, it is reasonable to conclude that the first stars, with essentially no metals, were primarily very massive and supermassive stars. In fact, if we assume that the IMF for Population III stars was similar in shape to that of the Population I stars, but scaled by the Jeans mass, then the distribution of zero-metallicity stars formed in these early protogalaxies might have looked something like the dashed line in Fig. 10.6.

Nucleosynthesis and IMBHs from Population III Stars

There are two things to notice about this heavy IMF for Population III stars. First, while the Jeans mass is $\sim 400 M_\odot$, and these stars tend to collapse directly to black holes and not perform much nucleosynthesis, we still expect significant numbers of stars in the range 100–260 $M_\odot$. As noted earlier, these stars will explode as pair production supernovae and dirty the interstellar, and intergalactic, gas with metals like carbon, oxygen, neon, magnesium, silicon, iron and nickel. After mixing with the interstellar medium, this will provide enough heavy elements ($Z \sim 10^{-4}$ or higher) with radiatively-cooling emission lines that the formation of the next generation “Population II” stars in future globular clusters and galactic bulges will occur at much cooler temperatures, allowing stars of much smaller mass to form than could do so in the zero-metallicity Population III phase.

Secondly, if 10% of the gas forms Population III stars in a protogalaxy of about $10^5 M_\odot$ (80% of which is dark matter), then the total mass of stars expected to form is only $1/5 \times 0.1 \times 10^5 M_\odot = 2000 M_\odot$. So, given the IMF we have developed here, only a few stars in the 200 $M_\odot$ range will form, and occasionally one as large as $\sim 600 M_\odot$ will appear as well. These numbers, of course, will be proportionately larger for newly-formed dark matter mini-halos of $10^{6-7} M_\odot$ total mass.

In summary, then, the first (Population III) stars should produce IMBHs roughly in the range 130–500 $M_\odot$ in addition to enriching the primordial gas in the mini-halos to $Z \sim 10^{-4}$.

10.4.2 Formation of IMBHs in Dense Star Clusters in the Present Universe

The other possible IMBH formation site is the center of dense star clusters. One often immediately thinks of globular clusters when this is mentioned, but actually a more promising site is the center of dense, *young* clusters, which may have many massive O and B stars. The physics of stellar dynamics is basically the same in any dense star cluster, however, so we shall present some *simple* stellar dynamics principles before discussing the different cluster IMBH formation scenarios.

10.4.2.1 Simple Stellar Dynamics: Whole Stars as Single “Gas” Particles

A large system of stars (*i.e.*, star clusters or galaxies) behaves in a manner similar to a fluid or gas. Indeed, the interior of these objects is sometimes called the “stellar fluid” or “stellar gas”. To lowest order, this “gas” is collisionless: that is, scattering of individual stars with the entire, gravitating system is most important; individual star–star interactions are not. However, as the space density of stars rises, two-body interactions (from casual encounters to close near-misses) become important in determining how the internal structure of the cluster or galaxy is re-arranged. At still higher stellar densities, three-body encounters (*e.g.*, a binary interacting with another star) become crucial, and even four-body (binary + binary) interactions are of interest.

The stellar fluid has its own set of physical laws that govern how it behaves. As with atomic and molecular particles, stars are governed by a Boltzmann-type equation; it is called the Fokker–Planck equation and includes a collision term and has gravity as the main external force. One can determine the distribution function of stars in phase space (number of stars per unit spatial volume and per unit volume in velocity space). If the distribution function is regular, one can characterize it with simple parameters like *density* and *temperature*.

Furthermore, like molecular, atomic, and nuclear particles, reactions can occur that modify the behavior of the stellar fluid. Binary star systems can form and, through their interactions with single stars, can alter the energy content of cluster. Also, under certain circumstances, those stars can merge, forming a more massive star and changing the composition of the cluster. Finally, the stellar fluid can collapse gravitationally and even “bounce”. These latter processes all quite possibly play roles in forming intermediate mass black holes.

The Isothermal Sphere Polytrope: A Stable Model of a Star Cluster

Using the “stellar gas” model for star clusters with 10^{5-7} stars, we can write the radial structure of a spherical cluster using the stellar structure equation (5.67) of hydrostatic equilibrium

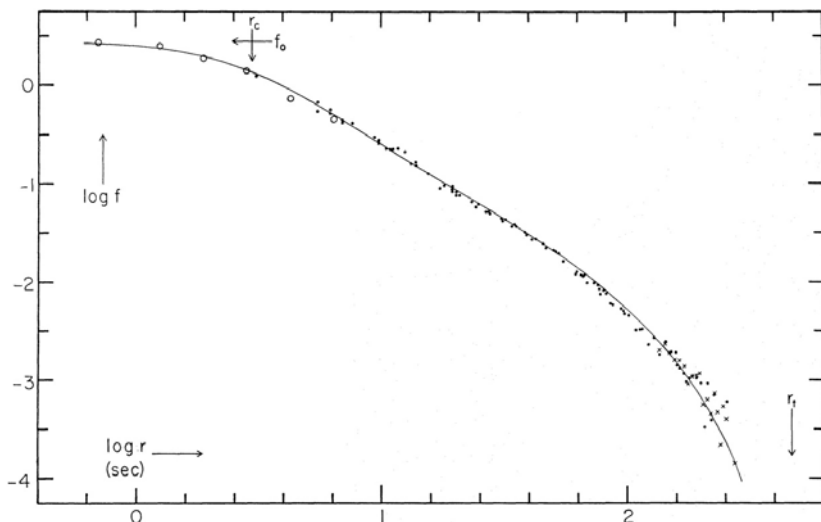


Fig. 10.7: The profile of a King isothermal sphere star cluster model, applied to actual data from the elliptical galaxy NGC 3379. The core r_c and tidal r_t radii also are shown. In addition to fitting a galaxy profile, King also used his model in this paper to explain globular clusters M15 and NGC 5053 in our Galaxy. Reproduced from Fig. 4 in [393], by permission of the AAS.

$$\frac{dp}{dr} = -\rho \frac{Gm(r)}{r^2}$$

The pressure equation of state is given by

$$p = n kT = \rho \sigma_V^2 \quad (10.5)$$

where n and ρ are the number and mass densities of stars, T is the “temperature” of the system of stars, and $\sigma_V(r)$ is the average random stellar velocity (“velocity dispersion”), all at radius r . Equation (10.5) shows that σ_V^2 plays the role of a temperature in stellar dynamics. (Note that we are assuming that the star velocities are in “thermal” equilibrium, with a Maxwellian distribution function (equation (9.48)) in velocity space.) The mass $m(r)$ in the hydrostatic equilibrium equation is the usual mass interior to radius r

$$m(r) = \int_0^r 4\pi r^2 \rho dr$$

Now, let us examine the situation where *all* stars’ velocities in the cluster can be described by a *single* temperature. That is, σ_V is independent of r . This is called the *isothermal* velocity distribution and was introduced in Section 5.2.4 during the discussion of polytropes. The structure is given by equation (5.77) with $K_\infty = \sigma_V^2$; so the solution for density is given by equation (5.78)

$$\rho = \frac{\sigma_V^2}{2\pi G} \frac{1}{r^2} \quad (10.6)$$

and the mass interior to radius r is

$$m_{SIS}(r) = \frac{2\sigma_V^2 r}{G} \quad (10.7)$$

This is the SIS, or singular isothermal sphere, solution ($n = \infty$ polytrope), and has the problems of infinite central density (although $m_{SIS}(r) \rightarrow 0$ as $r \rightarrow 0$) and infinite mass as $r \rightarrow \infty$.

As noted by Ivan King of the University of California, Berkeley, neither of these fit observations of globular clusters very well: most observed clusters appear to have a flatter density profile inside some core radius r_c , and they also appear to have been stripped of their outer stars beyond some “tidal radius” r_t . This suggests two modifications to equation (10.6):

- Modify the singular isothermal distribution to be of the form

$$\rho = \frac{\sigma_V^2}{2\pi G} \frac{1}{(r^2 + r_c^2)}$$

so it flattens inside the core radius.

- Assume that, beyond the tidal radius, the cluster mass does not increase much. Then $m_K(r) \approx M_{\text{clst}}$ for $r > r_t$. Then equation (5.67) can be integrated to obtain the density structure for an isothermal atmosphere

$$\rho = \rho_t \exp \left[-\frac{G M_{\text{clst}}}{r_t \sigma_V^2} \left(1 - \frac{r_t}{r} \right) \right]$$

For $r \gg r_t$, ρ rises exponentially with decreasing r from an external low uniform value ($\rho_e = \rho_t \exp [-G M_{\text{clst}} / r_t \sigma_V^2]$) to ρ_t at $r = r_t$.

The above suggested changes are a simple description of the somewhat more detailed models of clusters developed by King [393]. His models (see Fig. 10.7) can be characterized by a core or “King” radius interior to which the density is approximately constant, and a tidal radius, exterior to which the density drops rapidly. For $r_c < r < r_t$, however, the structure follows equation (10.6). In all cases, while they are not exactly isothermal, the temperature throughout a King model does not vary much, so they are still thought of as isothermal stellar distributions. King models of varying compactness ratios $C \equiv \log(r_t/r_c)$ describe many globular clusters quite well.

Note that, when a cluster core exists, well inside the core radius the density is $\rho \approx \sigma_V^2 / (2\pi G r_c^2)$, so that inside the core the mass interior to radius r is only

$$m_K(r) \approx \frac{2\sigma_V^2}{3G r_c^2} r^3 \quad (10.8)$$

At the core radius, then, the mass interior to r_c is only about one-third of what it would be if the distribution had been a true singular isothermal sphere ($m_K(r_c) = 2\sigma_V^2/(3G)$). Therefore, when a core exists, there is a *mass deficit* of

$$M_{def} \equiv m_{SIS}(r_c) - m_K(r_c) = \frac{4}{3} \frac{\sigma_V^2 r_c}{G} \quad (10.9)$$

from what would be in a singular core to what actually is in the uniform density core.

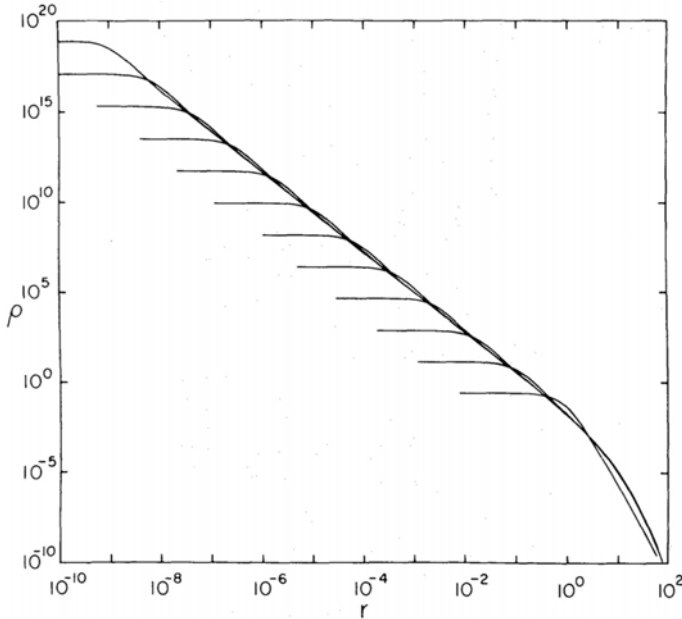


Fig. 10.8: Evolution of a King model core radius during core collapse, from simulations by Haldan Cohn when he was at Harvard Center for Astrophysics [394]. Note that an approximate $\rho \propto r^{-2}$ profile is preserved throughout the event. While this collapse can proceed to an extremely high central density, the mass enclosed in the core is proportional to r_c , which goes to zero as the core shrinks. As a result, core collapse in a cluster of equal-mass stars is not thought to form a central black hole of any significant mass. Reproduced from Fig. 1 in [394], by permission of the AAS.

The Gravo-Thermal Catastrophe and Classical Star Cluster Core Collapse

While isothermal spheres like King models are dynamically stable, in principle, they can evolve slowly to an unstable configuration on many thermal time scales. The core radius r_c will then shrink to a very small value on a thermal time scale, leaving a structure that is approximately the singular isothermal distribution in equation

(10.6; see Fig. 10.8) [394]. (The actual distribution is closer to $\rho \propto r^{-2.21}$.) If all the stars in the cluster have approximately the same mass, this process is called “classical” core collapse.

In the stellar gas approximation for stellar dynamics described above, the gravothermal catastrophe that leads to cluster core collapse can be described as undergoing the following steps, which were first described by Donald Lynden-Bell (University of Cambridge) and Roger Wood (Royal Greenwich Observatory) in Britain in 1968 [395]:

1. The initially isothermal cluster begins to lose stars from its central regions through evaporation (loss of stars scattered into energies $\mathcal{E} \gg \frac{1}{2} \langle m_\star \rangle \sigma_V^2$, making them unbound by the cluster core).
2. This loss of stars causes a loss of “pressure” in the stellar gas, so the cluster core shrinks and heats up to increase the pressure. The velocity distribution now is no longer isothermal.
3. The outwardly-decreasing temperature gradient in the stellar gas causes a diffusion of stellar energy outward; this is identical to the process of heat conduction (with whole stars replacing atomic particles), and is driven by two-body, star–star interactions.
4. The conduction of heat outward causes the core to lose more pressure and shrink more, while the cluster halo gains energy from the evaporating stars and expands. Because the cluster is a gravitating body, however, the orbits of the stars in the shrinking core take on faster (hotter) velocities, while the orbits of stars in the expanding halo become slower (cooler).
5. This process then continues to cycle through the above steps, running away: the cluster core continues to contract and the halo absorbing the energy conducted out of the core continues to expand. The entire cluster, therefore, evolves like a red giant, with a shrinking but heating core and an expanding, cooling envelope.

The characteristic time scale for cluster core collapse is the two-body relaxation time in the cluster core

$$\begin{aligned} \tau_{\text{rlx}} &= \left(\frac{r_c^3}{G M_{\text{clst}}} \right)^{1/2} \frac{M_{\text{clst}}}{\langle m_\star \rangle 8 \ln \Lambda} \\ &= 120 \text{ Myr} \left(\frac{r_c}{\text{pc}} \right)^{3/2} \left(\frac{M_{\text{clst}}}{10^5 M_\odot} \right)^{1/2} \left(\frac{\langle m_\star \rangle}{0.5 M_\odot} \right)^{-1} \left(\frac{\ln \Lambda}{10} \right)^{-1} \quad (10.10) \end{aligned}$$

where $\ln \Lambda$ is the Coulomb logarithm that takes into account collisions from close near misses to distant casual encounters. (Here $\ln \Lambda \approx \ln(0.1 M_c / \langle m_\star \rangle) \approx 9\text{--}14$.) Numerical and analytic models of clusters where all stars have the same mass indicate that a truncated isothermal sphere (King model) will undergo core collapse after about $16 \tau_{\text{rlx}}$. This is a very long time (2 Gyr for typical cluster conditions and long even for very compact star clusters with $\tau_{\text{rlx}} < 25 \text{ Myr}$), but it still is much less than the age of the Galaxy. About one-fifth of all Galactic globular clusters have collapsed cores, so the process is not all that rare.

Halting Cluster Core Collapse: “Binary Burning”. Not only does the cluster core shrink like a red giant core, the termination of this collapse also mimics the red giant process as well. When the central cluster density and temperature (which are increasing due to the collapse) become high enough, binaries will form in the core. Repeated scatterings of single stars by these binaries will be *inelastic*, giving up energy to the single star and binding the binary system more tightly.⁵ The release of binding energy to the stellar gas by binaries is analogous to nuclear burning in the core of a red giant. Specifically, in the actual red giant case it is helium burning to more tightly bound carbon nuclei that halts the shrinkage of the red giant core. In the case of the star cluster, it is the “binary burning” process that heats the core to the point where the energy generated and injected into the core by inelastic binary scatterings equals that lost to the halo by the conduction process discussed above.

No IMBH Formation from Classical Core Collapse. Initially, it was thought that the process of cluster core collapse described above would lead to the formation of an intermediate mass black hole. However, the mass within the cluster core actually *decreases* as the collapse ensues ($M_c \propto r_c^{0.79}$). Furthermore, the halting of core collapses by binary burning also mitigates the uncontrolled growth of the cluster central density. If black holes do exist at the centers of globular and other star clusters, they probably are not formed by classical core collapse.

Core Collapse in Clusters with a Distribution of Stellar Masses

The situation changes significantly when we consider star clusters that have a range of stellar masses. Heavier stars will sink to the cluster center faster than the lighter stars, through a process called dynamical friction: two-body encounters between unequal stars will result in more energy being imparted to the less massive star and more energy removed from the more massive star’s orbit. Objects with mass m_* will segregate toward the center of cluster in a time

$$\begin{aligned} \tau_{\text{seg}} &\approx \frac{\langle m_* \rangle}{m_*} \tau_{\text{rlx}} \\ &= 6 \text{ Myr} \left(\frac{r_c}{\text{pc}} \right)^{3/2} \left(\frac{M_{\text{clst}}}{10^5 M_\odot} \right)^{1/2} \left(\frac{m_*}{10 M_\odot} \right)^{-1} \left(\frac{\ln A}{10} \right)^{-1} \quad (10.11) \end{aligned}$$

The process of mass segregation is another form of core collapse, but it occurs on a much shorter time scale. In fact, the first dynamically-formed binaries appear at $\tau_{\text{cc}} = 0.2 \tau_{\text{rlx}}$, which is a few τ_{seg} for $m_* \sim 10 M_\odot$ objects. (These are ~ 20 times heavier than the typical $\sim 0.5 M_\odot$ star in an older globular cluster.) These heavier objects can be stellar ($10 M_\odot$) mass black holes or massive stars ($> 11 M_\odot$) before they explode as supernovae. We will finish out the remainder of this chapter section

⁵ This process is essentially the gravitational slingshot process that is used by NASA to propel spacecraft to the outer planets on fast orbits. Jupiter and the sun form the binary that gives up some of its binding energy, and the spacecraft is the third body in the interaction.

on intermediate mass black hole formation by examining each of these cases in the next two sub-subsections.

10.4.2.2 Formation of IMBHs from Stellar-mass Black Holes in Star Clusters

Stellar-mass black holes should be formed after the first ~ 5 Myr of the cluster's life. And after a time τ_{seg} (another 10–20 Myr) these black holes should sink into the center of the cluster, forming a collapsed core composed primarily of stellar-mass black holes. If this cloud of black holes is in equilibrium with the rest of the cluster stars, the density should be similar to that of a collapsed cluster core ($\rho \propto r^{-2.21}$). In any case, BH+BH binaries also will form and stabilize the collapsing central black hole cloud.

Evolution of the Central Black Hole Cloud

In 2006 Ryan O'Leary and his colleagues at Northwestern University performed a detailed study of the further evolution of this type of black hole cloud [396]. As before, collisions of binaries with single black holes will cause the binaries to harden (become more tightly bound). In addition, the encounters not only will heat the cluster core, they also will eject many of the single black holes (and some of the binaries) from the star cluster. This occurs simply because sometimes the energy imparted to the single star can be greater than the cluster's escape energy.

Nevertheless, some black hole binaries will coalesce and form larger black holes. *If* these holes remain in the cluster, they will form binaries with other black holes in the core and continue their growth. In this manner black holes of size 100–6000 $M_{\odot}$ could form in the cluster center, and the probability of forming such intermediate mass black holes would increase for clusters that are more massive and more centrally condensed (large r_t/r_c concentration ratio).

The Gravitational Wave Recoil Problem

However, for most star clusters, *this process will not work*, because the cluster gravity is not sufficient to hold the merging BH+BH binaries. As shown in Section 8.2.4, merging black hole binaries with a modest mass ratio will produce a gravitational wave recoil of 50–100 km s^{-1} . (This numerical relativity result was obtained after O'Leary and his colleagues studied the black hole cluster core problem.) The velocity dispersion can be computed from the core radius as

$$\sigma_V = \left(\frac{1}{3} \frac{G M_{\text{clst}}}{r_c} \right)^{1/2}$$

$$= 12 \text{ km s}^{-1} \left(\frac{r_c}{\text{pc}} \right)^{-1/2} \left(\frac{M_{\text{clst}}}{10^5 M_\odot} \right)^{1/2}$$

So, for most globular clusters, the gravitational wave recoil will be a factor of 4–8 times faster than the cluster’s escape speed. As a result, *all* merging binaries with two 5–10 $M_\odot$ black holes will eject the merger product (new black hole) from the cluster. Globular clusters, therefore, cannot build up black holes of intermediate mass by merging many 10 $M_\odot$ ones.

The only way to utilize the black holes being formed by stellar evolution is to create a large seed black hole in the center – more than 10 times more massive than the typical stellar-mass hole (*i.e.*, $\sim 100 M_\odot$). And the formation of this “anchor” black hole must be done *before* significant numbers of stellar-mass black holes are formed. In this manner the anchor can consume most of the smaller holes and keep them from being ejected from the system.

10.4.2.3 Formation of IMBHs from Colliding Massive Stars in Star Clusters

There *is* one very promising process for forming intermediate mass black holes in star clusters, and it does so on a very short time scale. This process was first introduced by Simon Portegies Zwart of the University of Amsterdam and Stephen McMillan of Drexel University [272]. However, in order for the mechanism to work, the cluster must be *very compact* ($\tau_{\text{rlx}} \leq 25 \text{ Myr}$) and very young ($t < \tau_{\text{seg}} \approx 0.2 \tau_{\text{rlx}} \leq 5 \text{ Myr}$). In such clusters, stars that normally would form black holes ($m > 21 M_\odot$) have not yet done so and will sink to the center of the cluster in a rapid mass-segregation core collapse, as described above. Unlike stellar-mass black holes (only $\sim 30 \text{ km}$ in radius), these O and B stars are nearly a million times larger. As binaries containing these massive stars harden and merge, they will form *more* massive stars *with essentially no gravitational wave recoil*. Subsequent collisions, either through additional binary formation or through direct hits, will lead to runaway growth of a very massive star (VMS) of order 0.1% of the cluster mass (*i.e.*, $10^{2-4} M_\odot$ for clusters that are $10^{5-7} M_\odot$ in mass).

Marc Freitag of the University of Heidelberg and his colleagues studied this mechanism in detail and found that a runaway growth of a VMS happens for *all* star clusters that are compact and young [397, 398]. The growth is very robust, occurring even if the velocity dispersion is as high as 1000 km s^{-1} . The growth of the VMS is, in fact, so rapid that it gains mass on a time scale shorter than the VMS’s own thermal time scale! That is, between successive collisions with new massive stars, the VMS will have not have time to come into thermal equilibrium (although it will be able to come into hydrostatic equilibrium). The central VMS, therefore, becomes heavier and heavier, but cannot get rid of the heat building up in its center due to its own gravitational contraction and due to hot O and B stars merging in its interior. For black hole formation this is good news, because there will not be time for the VMS to form the strong stellar winds that normally reduce such stars to only a few solar masses in a few hundred thousand years, even if it has relatively normal metal

abundances. On the other hand, it is not clear how a growing, compressionally-heated superstar will evolve. If it is anything like the low-metallicity VMSs (see Section 10.4.1), it will collapse directly to an intermediate mass black hole once its mass exceeds $250\text{--}300\text{ }M_{\odot}$.

The formation of an IMBH in compact, young clusters and then the subsequent evolution of those black-hole-anchored clusters, and their possible inspiral into the centers of their host galaxies, has the potential for explaining many current puzzles in black hole research. That is, in principle, it could provide formation scenarios for *all four* possible IMBH candidates discussed in Chapter 4:

- *ULXs associated with star-forming regions.* The brightest, and hence the most likely to be of order $1000\text{ }M_{\odot}$, ULX sources could indeed be recently-formed IMBHs in the centers of young, compact star clusters.
- *IMBHs in globular cluster centers.* If any globular cluster is confirmed to have a dark IMBH at its center, the best explanation for its existence would be that it is an ancient remnant IMBH, formed when the cluster was very young – 10 billion years ago as the Galaxy was forming, either from a large Population III star or from a compact Population II star cluster when it was young.
- *The possible IMBH in a young star cluster near the Galactic center black hole.* Like the ULXs, this is associated with young stars and so, again, fits the Portegies Zwart and McMillan formation scenario. This object would be a somewhat older version of a ULX cluster that had been partially stripped of most of its stars as it spiraled into the Galactic center.
- *Low-luminosity X-ray sources in the Galactic bulge.* If any of these are found to be high-mass objects, a possible explanation would be that they also are tidally stripped compact star clusters that are accreting gas from the ISM as they spiral in toward the Galactic center through the process of dynamical friction.

The mass-segregated core-collapse mechanism appears to be one of the most exciting developments in this field. Its implications have only begun to be understood and could reach into the area of supermassive black hole formation as well (see below). However, despite the very interesting possibilities, more work needs to be done, especially on the evolution of the nonthermal-equilibrium VMS to see exactly how large an IMBH is formed. Over the next decade or so, we should be able to see just how robust these promising ideas are.

10.5 Formation of Supermassive Black Holes at the Centers of Galaxies

The problem of how the most prominent and well-studied of all black holes formed is an extremely complex one. Ultimately, the creation and buildup of black holes of $10^5\text{--}10^6\text{ }M_{\odot}$ in the centers of galaxies is *intimately* tied to the following:

- cosmology;

- the formation and buildup of the galaxies themselves;
- the formation and evolution of Population III, II, and even I stars;
- the behavior of interacting, colliding, and merging galaxies, how they restructure themselves after those collisions, and how they transport material to the black hole(s) in the center;
- and, of course, general relativity.

There is scarcely an area of astrophysics that does *not* play a role – with the possible exception of planet formation. (But supermassive black hole formation and growth certainly will have a profound *back* reaction on the formation and even health of the poor planetary systems that get caught up in the feeding and nurturing of the monster.)

Even the term “formation” is in danger of losing its meaning here. While there are some suggested scenarios for creating supermassive black holes in short-lived events *before* the galaxies formed, the most favored and likely scenarios instead “build up” these objects by accreting various types of matter (gas, stars, neutron stars, other black holes) into a seed black hole of modest mass. Distinguishing between *formation*, *growth*, and *feeding* can be difficult, perhaps even impossible. Nevertheless, because this chapter is about the “formation” of black holes, and the next one is about the “feeding” and fueling of those engines, we will make an attempt here to deal mostly with processes in which large amounts of mass are added to a black hole in a galaxy center in a time much less than the age of that galaxy. It may help the reader to think of supermassive black hole formation in the same sense as galaxy formation. Galaxies and SMBHs both are *assembled* over a fairly long period of time by, nevertheless, very specific and deliberate processes that result in the final products appearing very similar despite being manufactured in different corners of the universe.

10.5.1 Broad Observational Constraints on SMBH Formation

Before discussing some specific SMBH formation scenarios, we should orient ourselves by developing a broad overview of the field, and by understanding the important observational constraints on how the events might have taken place.

10.5.1.1 The Black Hole–Galaxy Bulge Connection

By far the most important observations telling us something about how black holes form in galactic centers are the $M_{\bullet}-\sigma$ and $M_{\bullet}-M_{\text{bulge}}$ relations discussed in Section 2.5. Somehow, the supermassive black hole knows in what type of galaxy it will form (or has formed). And when many galaxies merge into a single one, those relations are preserved. In fact, the $M-\sigma$ relation is so strong that it would seem that the formation of SMBHs *and* galactic bulges may occur together in a single process.

10.5.1.2 Elliptical Galaxies Form Early

Elliptical galaxies (ones that are mostly bulge and have little or no disk of gas and/or stars) appear to be in place by a redshift of ≥ 2 (3 Gyr after the big bang). In fact, the high redshift ($z \sim 6$) quasars recently discovered, with black hole masses of $\sim 10^9 M_\odot$ (and therefore host bulge masses of $\sim 10^{12} M_\odot$) must be in elliptical galaxies that formed before 600 Myr after the big bang. While these may be the exception rather than the rule, the SMBH process must allow the formation of *some* very massive galaxies and black holes very shortly after the first stars formed.

10.5.1.3 Existence of Dwarf Spheroidal Galaxies with Supermassive Black Holes

The extension of the $M_\bullet$ – σ and $M_\bullet$ – M_{bulge} relations to very low mass galaxies is a very important result. (See Section 4.3.2.) These galaxies have $\sigma_V < 40 \text{ km s}^{-1}$ (and therefore spheroidal bulge masses of $M_{\text{bulge}} < 10^8 M_\odot$), which is considerably *less* than the typical binary gravitational wave recoil velocity range of 50–100 km s^{-1} . Yet, these dwarf galaxies also have black holes of mass $M_\bullet \leq 10^5 M_\odot$. We therefore can draw the following conclusion: gravitational wave recoil during the merger of binary black holes does *not* evaporate all black holes from dwarf spheroidal galaxies early in their history. This conclusion can be extended up to galaxies with velocity dispersions $\sigma_V < 100 \text{ km s}^{-1}$ ($M_{\text{bulge}} < 10^{10} M_\odot$) as well. Note that we are not concluding that black holes were not built up early by merging many smaller black holes. However, it does mean that the merging of black hole binaries, if they did occur, did not eject the merged product from the system. This can mean only one of two things: either BH+BH mergers did not occur at all in the formation of a dwarf spheroidal galaxy (which is rather unlikely) or the mergers mostly occurred with large mass ratios (>5) so that the recoil velocity was kept well under 40 km s^{-1} . That is, dwarf spheroidal galaxies must have formed an unusually large anchor black hole (say $>1000 M_\odot$) that was considerably more massive than the more typical Population III black holes of $\sim 200 M_\odot$ that existed at the time.

10.5.1.4 Accounting for the Many Types of Macroquasars

A complete theory not only should explain how galaxies and their SMBHs formed, it also should account for the major types of active galaxies that have been observed – eventually in their proper numbers. This should include not only the most numerous classes (QSOs and Seyfert galaxies) but also the other important classes (radio galaxies and QSRs; ULIRGs and HyLIRGs) and eventually the nearly quiescent objects (LLAGN and inactive galactic nuclei). While such a complete theory is still many years away, such thinking still should guide current research. *A model that predicts no radio galaxies and quasars with jets, for example, may not be worth pursuing.*

10.5.2 The Current, Sketchy View of Galaxy and SMBH Formation

10.5.2.1 Overview of Major Scenarios

In a broad sense, SMBH formation scenarios can be divided into three general classes, depending on whether the black hole is a cause or an effect of galaxy formation:

1. *Black hole forms before the galactic bulge.* In this scenario, a black hole of great mass forms directly (out of the dark matter itself, or from Population III stars, or from the metal-free primordial gas), and the galaxy forms around the black hole. That is: SMBH $\rightarrow$ Bulge.
2. *Galactic bulge forms first.* In this scenario the galaxy not only goes through the dark matter halo and Population III phases first, it also enters the Population II phase and forms the galactic bulge of stars. After that, an SMBH then forms in the center of that bulge as gas clouds, stars, and other large objects sink to the center and collect into a black hole. In this case: Bulge $\rightarrow$ SMBH.
3. *Bulge and black hole formation are closely linked.* In this third, hybrid scenario, the processes that create and build the galactic bulge also create and build the central black hole. The difference from the other two scenarios may seem subtle, but actually are important. In this case there is a master process creating both bulge and black hole simultaneously. That is: Master Process $\rightarrow$ Bulge; and Master Process $\rightarrow$ SMBH.

In the discussion below of the currently favored scenarios we will see some aspects of all of these. However, in the most plausible scenario, Type #3 will play a dominant role.

10.5.2.2 Setting the Stage: The Formation of Pre-Bulge Mini-Halos

In spite of the fact that there are *no* observations of the Population III era, or any clear remnants from those stars (unless they *are* the SMBHs themselves), as pointed out earlier there nevertheless is some agreement on what events led up to the formation of objects we see today. As discussed in Section 10.4.1, mini-halos of dark matter, $10^{5-7} M_{\odot}$ in mass, form out of the expanding universe at redshifts of $z \sim 20$. The most dense of these objects collapse quickly, with a few Population III stars in their cores. Some of these stars explode and enrich not only the mini-halo but also the surrounding *intergalactic* medium with newly-formed metals, increasing the metallicity to $Z \sim 10^{-3}$ – 10^{-4} . Some of the Population III stars collapse and form black holes of masses 130–500 $M_{\odot}$ and possibly larger.

10.5.2.3 Formation of the Bulge Stars and the Initial Central Black Hole

The Metagalactic Environment of Mini-Halos

These mini-halos are not isolated. In fact, there are perhaps thousands of them clustered in a larger-scale density enhancement that eventually will collect together to form a galaxy-sized object of perhaps $10^{8-11} M_{\odot}$. As the mini-halos begin to fall together and collide, a number of things happen:

- Population II stars form from the metal-enriched gas. Because this gas can cool more efficiently than the metal-poor Population III gas, the typical mass of Population II stars is expected to be closer to the current Population I Jeans mass (*i.e.*, $1-2 M_{\odot}$).
- The intermediate mass black holes in each mini-halo begin to settle at the center of the forming galaxy and form BH+BH binaries.
- Stars and gas in the merging halos also accrete onto the central black holes, growing them even more in mass.

The evolution of this cluster of black holes in the center of the forming spheroidal galaxy is an important step in the formation of the central black hole. This is true no matter how that black hole will ultimately gain most of its mass (by accreting gas, stars, or other black holes). As discussed above, an unusually large black hole must have acted as a seed and anchor, catching and merging with other smaller black holes that existed at the time. To be sure, many smaller ($\sim 200 M_{\odot}$) black holes may have been ejected as the mini-halos began to coalesce. As we have emphasized, any small binary composed of two $100-200 M_{\odot}$ black holes in systems with $\sigma_V < 50-100 \text{ km s}^{-1}$ would have ejected itself through gravitational wave recoil when it merged to form a single black hole. In addition, however, collisions of those binaries *before they merged* with other black holes could have ejected the latter in the binary hardening process as well.

Recapture of Ejected Black Holes by a Larger Galaxy

Are the black holes ejected by binary hardening and mergers really lost from the centers of galaxies forever? The answer is no in many cases. As the merger process continues, with larger and larger galaxies building up, the $100-200 M_{\odot}$ black holes that were ejected by ancient mini-halos could be “recaptured” by these larger child galaxies. Larger galaxies ($> 10^7 M_{\odot}$) would have velocity dispersions of $\sigma_V > 100 \text{ km s}^{-1}$, and so could capture the rather slowly-moving black holes. Since an object traveling at $\sim 100 \text{ km s}^{-1}$ for 100 Myr travels only 10 kpc, it is likely to be captured by a large galaxy at some point in its travels, perhaps the one containing its original mini-halo or perhaps another nearby large galaxy. Once captured, these black holes would quickly sink to the center of the galaxy, where mergers could take place without ejecting them out of the galaxy. So, given a large enough galaxy,

black holes ejected in the past could be harvested and used to build up a black hole in that galaxy's center.

Formation of the Seed Anchor Black Hole

Note that this “black hole recapture” process does not change our conclusion that a large seed anchor black hole would be necessary to start the SMBH formation process, especially for the formation of black holes in the center of dwarf spheroidal galaxies. Spheroidal galaxies that have remained dwarfs all the way to the present time have never been large enough to recapture rogue black holes that were ejected from ancient mini-halos and have been roaming intergalactic space. So, if they still appear to have a central SMBH, that central black hole must have formed in spite of its galaxy not being able to recapture any black holes that it previously ejected.

The creation of a central black hole in dwarf spheroids, then, probably depended on the existence of a large seed and anchor black hole – something around $1000 M_{\odot}$. There are three possible sources of such a seed, each of which may have been used by different forming galaxies in the early universe:

1. The remnant of an unusually large Population III star (*i.e.*, the most massive stars from the Population III IMF). See Section 10.4.1.
2. Delayed formation of Population III stars until most of the mini-halo gas has collected in the center to form a super central quasi-star.
3. Formation of a compact Population II star *cluster* that then experienced mass segregation core collapse and runaway star collisions in its center. See Section 10.4.2.

At the end of this phase, we now have an object that looks more like an elliptical galaxy than the original mini-halos. It is composed of a large dark matter halo (formed from the merger of the many mini-halos), a bulge of Population II stars (formed during the many collisions of the enriched baryonic gas), and containing a central massive black hole that has absorbed most of the smaller Population III remnants, as well as other sorts of galactic refuse. The buildup of galaxy and black hole has only begun, however, and these dwarf (10^8 – $10^{10} M_{\odot}$) galaxies are destined to undergo many more mergers with other similar objects.

10.5.2.4 Buildup of the Bulge and Central Black Hole: Establishing the $M_{\bullet}$ – σ and $M_{\bullet}$ – M_{bulge} Relations

There are many different theories for how the correlations described in Section 2.5 occur. Here we will discuss the most popular class: the black hole grows until it cuts off its own fuel source. In this scenario, then, there is a master process, and it is the evolution of gas inside the forming galaxy. This process feeds and grows the central black hole as well as the bulge star population. In the next subsection we will relate this scenario to recent theoretical ideas on black hole formation and star formation.

Formation of Black Hole and Bulge in Simple Numerical Simulations

In the year 2000 Guinevere Kauffmann and Martin Haehnelt of the Max Planck Institute for Astrophysics in Garching, Germany, published models for the simultaneous formation and growth of galactic bulges and central black holes [399]. The basic process involved the merging of dark matter halos, with stars and black hole attached, which triggered more star formation and the accretion of matter into the central hole. While the models were fairly simple, they accounted for a surprising number of properties of galaxies, including the $M_\bullet$ – σ and $M_\bullet$ – M_{bulge} relations as well as the Faber–Jackson M_{bulge} – σ relation.

The basic elements of the model were as follows:

- During the collision and merger of two galaxies, the dark matter halos were assumed to merge without incident and the two central black holes to coalesce without incident.
- Collisions of the cold gas mixed in with these halos, however, caused (Population II) stars to form out of that gas.
- A fraction of that gas was assumed to accrete into the central black hole and cause it to grow.
- Both the fraction of cold gas going into stars and the fraction of gas going into the black hole were limited in the same way, and that limitation depended on the mass of the galaxy through the velocity dispersion σ_V as

$$\dot{M}_{\text{stars, acc}} \propto \frac{M_{\text{gas}}}{1 + \left(\frac{280 \text{ km s}^{-1}}{\sigma_V} \right)^2}$$

So, when the spheroidal bulge grew to be above $\sim 3 \times 10^{11} M_\odot$ ($\sigma_V > 280 \text{ km s}^{-1}$), the cold gas was converted into stars (and black hole growth) at a fraction that did not depend on total galaxy mass (σ_V). However, in smaller galaxies, the fraction of gas that was converted into stars decreased with decreasing galaxy mass. The rationale for this was that smaller galaxies were more likely to lose gas by processes like supernova heating of the ISM, a central quasar, and other energetic galactic processes.

While these simulations had a few normalizing parameters (the fraction of gas going into black hole growth was $f_\bullet \sim 3\text{--}4\%$ with the remainder $f_\star \sim 96\text{--}97\%$ going into stars), they did fit the $M_\bullet$ – σ_V , $M_\bullet$ – M_{bulge} , and M_{bulge} – σ_V relations surprisingly well, including predicting the *scatter* in the relations.⁶ The main conclusion that can be drawn here is that a model that simultaneously builds both the stellar population and the black hole mass out of the cold gas in merging galaxies

⁶ Note that the fraction of gas going into the black hole ($f_\bullet$) is far greater than the typical ratio of $M_\bullet/M_{\text{bulge}} \sim 0.0013$. Therefore, in these simulations the bulge and black hole growth occurred mostly in the merger of the two original bulges and black holes, respectively, and not because of star formation or gas accretion. This is a rather stochastic merger scenario, with only a little feedback being provided by σ_V in the equation above.

can explain many of the fundamental properties of galaxies and black holes, and their correlations.

Super-Eddington Winds from the Central Quasar as the Limitation on Black Hole Growth

The above concepts were made more quantitative by work done by Andrew King of the University of Leicester [400]. His model is rather specific, but has broader implications. After a galaxy merger, some of the cool gas falls into the center of the forming spheroidal bulge and accretes onto the black hole. This continues until the black hole, generally radiating at the Eddington luminosity and driving a strong wind outflow (see Section 13.1.2), becomes so powerful (*i.e.*, so massive) that its wind drives the cooling gas out of the nucleus, shutting off the black hole's own fuel supply.

King found that, for an isothermal model for the galactic bulge (equation (10.6)) the actual speed of the wind's shock wave interacting with the infalling gas will be

$$V_{\text{sh}}^2 = \frac{G L_{\text{Edd}}}{2 f_g \sigma_V^2 c}$$

where $f_g = 0.04/0.22 \approx 0.18$ is the fraction of gas compared with the dark matter in a typical galaxy. When $V_{\text{sh}} \geq \sigma_V$, the super-Eddington wind will blow the cold gas out of the galaxy. Inserting the Eddington luminosity from equation (1.1), we find that the gas will be expelled when

$$\begin{aligned} M_{\bullet} \geq M_{\bullet, \text{max}} &\equiv \frac{f_g \kappa_{\text{es}}}{2\pi G^2} \sigma_V^4 \\ &\approx 1.5 \times 10^8 M_{\odot} \left(\frac{\sigma_V}{200 \text{ km s}^{-1}} \right)^4 \end{aligned} \quad (10.12)$$

This is very close to the observed $M_{\bullet}$ - σ relation (equation (2.14)). It not only gets the power of σ_V correct, it also obtains the correct coefficient (*i.e.*, the correct normalization of the $M_{\bullet}$ - σ relation).

While this result was derived from a specific model, it is actually much more general. It says, essentially, that the black hole will grow until the wind momentum it produces sweeps up and expels virtually *all* the gas in the spheroidal bulge core. This not only stops accretion into the black hole, it also stops formation of more Population II stars. So *any* process that both forms stars and fuels the central black hole will be shut off when $M_{\bullet} = M_{\bullet, \text{max}}$.

10.5.3 Unsolved Problems and the Possible Role of IMBHs Formed in Compact Star Clusters

We now discuss some unsolved problems with the above scenario for SMBH and galaxy formation and offer some suggested solutions that involve *intermediate* mass black holes that form in the centers of compact star clusters.

10.5.3.1 Transport of Gas to the Nucleus and its Accretion onto the Central SMBH

In the above scenarios the gas that feeds the black hole is assumed to simply fall into the nucleus and accrete onto the supermassive black hole there. However, transport of gas into the nucleus always has been a very difficult problem to solve physically. Gas will settle in a disk, even if it does fall into the nucleus, and that disk will be 0.1–1.0 kiloparsecs ($\sim 3 \times 10^{20}$ – 3×10^{21} cm) in size or so. This is many orders of magnitude larger than the accretion region of a central SMBH (100–1000 r_S , or $\sim 3 \times 10^{16}$ – 3×10^{17} cm). How does the gas get close enough to the black hole to feed it? And what keeps the gas from forming stars along the way? Here we explore the distinct and exciting possibility that the gas actually *does* form stars and is locked up in compact star clusters that each are anchored by an IMBH. As these clusters spiral in toward the center through dynamical friction, they each bring in $\sim 10^6 M_\odot$ of gas each (in stars) plus a small amount in the IMBH itself ($\sim 10^3 M_\odot$). A scenario like this for feeding the black hole was first proposed by Toshikazu Ebisuzaki of the Riken Center in Wako, Japan, and his colleagues [401]. Not only does it present a physical mechanism for getting gas into the vicinity of the black hole, it also naturally explains why the fraction of gas going into stars and that going into the black hole have a constant ratio.

To analyze this situation, we once again return to the dynamical friction equation (10.11) that describes the mass segregated core collapse. Now, however, the inspiraling “star” will be a single compact star cluster, and the “cluster” in that equation now will be the entire galaxy itself. The same equation can be used, but rewritten as

$$\tau_{\text{seg}} \simeq 16 \text{ Myr} \left(\frac{r}{300 \text{ pc}} \right)^{3/2} \left(\frac{M_{\text{bulge}}}{10^9 M_\odot} \right)^{1/2} \left(\frac{M_{\text{clst}}}{10^6 M_\odot} \right)^{-1} \left(\frac{\ln A}{20} \right)^{-1}$$

This is still significantly longer than a crossing (dynamical) time

$$\begin{aligned} \tau_{\text{cross}} &\simeq \left(\frac{r^3}{G M_{\text{bulge}}} \right)^{1/2} \\ &= 2.5 \text{ Myr} \left(\frac{r}{300 \text{ pc}} \right)^{3/2} \left(\frac{M_{\text{bulge}}}{10^9 M_\odot} \right)^{-1/2} \end{aligned}$$

How does the time for compact star clusters to spiral into the nucleus compare with the typical time between mergers? In order to build up a $10^{12} M_{\odot}$ galaxy from many $10^5 M_{\odot}$ mini-halos by a redshift of 2 (3 Gyr after the big bang), we need $\sim \ln 10^7 / \ln 2 \sim 23$ merging stages (where each merging stage is assumed to be the coalescence of two equal-mass galaxies from the previous stage). The mean time between each stage then is $3000/23 \sim 130$ Myr. So, in building up the average giant elliptical galaxy, there is generally plenty of time for many of the compact star clusters with an IMBH in their core to sink from several hundred parsecs out into the nucleus and interact with the black hole.

However, if we want to form a giant elliptical by $z = 6$ (600 Myr after the big bang), the mean time between merger stages must be very short (< 25 Myr). This suggests that impacts of partially-merged mini-halo sub-dwarf-galaxies may occur *before* some compact star clusters would sink to the center of the forming galaxy. Cluster-cluster interactions in the nucleus, IMBH binaries, and the like all would be possible, resulting in perhaps some $10^{2-4} M_{\odot}$ black holes ejected into the inter-galactic medium (by the binary hardening process) before the forming nucleus settled down. Such a “near-monolithic” collapse of many mini-halos into a single galaxy is reminiscent of the early ideas of galaxy formation before the merger concept became popular (*e.g.*, Richard Larson’s simulations in the early-to-mid 1970s). Technically, such a rapid multi-mini-halo collision still would be a “merger”, but it would occur on a time scale similar to the collapse time of an entire $10^{12} M_{\odot}$ galaxy – only a few hundred million years.

10.5.3.2 Assisting the Hardening and Coalescence of Supermassive Black Hole Binaries

After several merging stages, but long before the final elliptical galaxy has been assembled, subsequent mergers will be between substantial-sized objects, *e.g.*, two $10^9 M_{\odot}$ bulges, each having a $\sim 10^{10} M_{\odot}$ dark matter halo and a central black hole of $\sim 10^6 M_{\odot}$. (Note that the central black hole is now significantly greater than any hole that can form in a compact star cluster and be brought in by inspiral.) After the two galaxies merge in a “major” (nearly equal-mass) merger, the two black holes will spiral into the center of the new galaxy in a dynamical friction time (here ~ 100 Myr), find each other, and form a supermassive black hole binary. The subsequent evolution of this extraordinary binary system is still one of the (several) unsolved problems of astrophysics. A very good review of this area of study was published on line in 2005 by David Merritt of Rochester Institute of Technology and Miloš Milosavljević [402].

The Final Parsec Problem

The main problem is very simple to state. After the supermassive binary shrinks (through dynamical friction) to a semi-major axis of

$$\begin{aligned}
 a_{\text{hard}} &\equiv \frac{G \mu}{4 \sigma_V^2} \\
 &= 0.3 \text{ pc} \left(\frac{\mu}{10^6 M_\odot} \right) \left(\frac{\sigma_V}{60 \text{ km s}^{-1}} \right)^{-2}
 \end{aligned}$$

(where μ is the reduced mass of the binary). Here “hard” means that the binding energy of the binary $G M \mu / 2a$ is a factor of 4 greater than its kinetic energy in the center of the galaxy ($\frac{1}{2} M \sigma_V^2$). At this point the binary will appear as a point mass to the surrounding system of bulge stars, and it will cease to shrink and harden. Furthermore, gravitational radiation is not yet a factor for a binary of this size. From equation (8.37) the time for it to coalesce due to gravitational radiation is

$$\tau^{GW} = 2 \times 10^{17} \text{ yr} \left(\frac{\mu}{10^6 M_\odot} \right)^3 \left(\frac{M}{10^6 M_\odot} \right)^{-4} \left(\frac{\sigma_V}{60 \text{ km s}^{-1}} \right)^{-8}$$

much longer than the age of the universe by a factor of 10 million.

How small does the binary need to be in order to coalesce from gravitational radiation in only a billion years? Using equation (8.37) again we can solve for the separation of such a binary and find that

$$a_{GW} = 0.0025 \text{ pc} \left(\frac{\tau^{GW}}{\text{Gyr}} \right)^{1/4} \left(\frac{\mu}{10^6 M_\odot} \right)^{1/4} \left(\frac{M}{10^6 M_\odot} \right)$$

or, in terms of the binary mass ratio,

$$a_{GW} = 0.0025 \text{ pc} \left(\frac{\tau^{GW}}{\text{Gyr}} \right)^{1/4} \left(\frac{\mu}{M} \right)^{1/4} \left(\frac{M}{10^6 M_\odot} \right)^{5/4}$$

It is significantly smaller than a_{hard} by two orders of magnitude! This is called the “final parsec” problem. There does not seem to be a good way to shrink the binary any more after it has shrunk to a_{hard} . Without any additional interaction with fairly massive objects, it may remain a supermassive black hole binary, never shrink to a_{GW} , and never coalesce.

In order for the BH+BH binary to be able to coalesce by gravitational radiation in 1 Gyr, it must decrease its separation by a factor of ~ 100 , or *increase* its current binding energy

$$\begin{aligned}
 \mathcal{E}_b &= \frac{G M \mu}{2 a_{\text{hard}}} = 2 M \sigma_V^2 \\
 &= 1.43 \times 10^{53} \text{ erg} \left(\frac{M}{10^6 M_\odot} \right) \left(\frac{\sigma_V}{60 \text{ km s}^{-1}} \right)^2
 \end{aligned}$$

by a factor of 100. Hardening such a binary is possible if it can interact with about 100 times more mass than it has (*i.e.*, $10^8 M_\odot$) and impart velocities to those stars and other objects that are $V > \sigma_V$. Any star on an orbit that comes near the BH+BH binary will indeed help to harden the system. However, once those stars have done

their job, they will be ejected from the galactic core, if not from the galaxy altogether, and no longer of use in helping to harden the supermassive binary further. This problem is called the “depletion of the loss cone” problem and will be discussed in more detail in Section 11.1.3. One of the most hopeful methods of solving the final parsec problem (and getting the binary black hole to coalesce) is to refill the loss cone by scattering more stars into orbits that intersect the binary. So far, it has not been clear that this refilling can be accomplished in a short time.

A Possible IMBH/Compact Cluster Solution

Here, again, compact star clusters of mass $\geq 10^6 M_\odot$, anchored by an IMBH, may be able to help, especially if there are many of them, as would be the case in a massive star burst of the kind that occurs in ULIRGs or HyLIRGs:

- Casual encounters with the central binary will send the clusters back out into the galactic bulge, where they again will sink toward the center through dynamical friction. Each cluster could potentially visit the BH+BH binary several times.
- As the clusters inspiral toward the center each time, they will “stir the galactic pot” and possibly refill the loss cone, potentially many times.
- Intimate encounters with the BH+BH binary would likely tidally disrupt the compact star cluster, making each of its stars, and its IMBH, available one last time for hardening the binary.

All of this is rather speculative at the moment; detailed simulations of merging galaxies with many compact star clusters have not been done yet to test this idea.

The Possible Role of a Future Merger and Third SMBH

If stellar dynamical processes *within* the galaxy do not complete the hardening and gravitational-wave merger, the BH+BH binary will remain intact until the galaxy encounters its *next* merger partner(s). This will initiate more harassment of the binary by more stars, compact star clusters, IMBHs, *and* a third supermassive black hole. The binary will eventually harden, but not before ejecting a considerable mass of black holes and other things from that galaxy. Whether this will occur on a short enough (<100 Myr) time scale or not, however, remains the central question.

10.5.4 The Author’s Personal Perspective on Formation of SMBHs

Despite the general perception by some that we do not know how SMBHs form, there actually is a fair amount of agreement in the cosmology community on a rough set of steps that may have accomplished this task. Some of the outstanding issues are:

- How is a large seed black hole formed early on so that coalescing black holes can be kept from ejecting themselves into the IGM via the process of gravitational wave recoil? Is such an anchor hole necessary, or is there another way to avoid recoil ejection in dwarf spheroidal galaxies?
- How does gas migrate to the center of the galaxy to feed the forming black hole there in a short amount of time? Is there some credence to the IMBH/compact star cluster idea that the gas that feeds central black holes is actually locked up in stars in those clusters and trucked in via dynamical friction?
- What is the overall role of compact star clusters formed in galaxy collisions and mergers, and how do their constituent IMBHs contribute to the overall process of SMBH formation?
- What is the fate of the SMBH+SMBH binary formed in a major merger? Does it coalesce quickly? What is the main process for causing this to happen? Can it be accomplished by the newly-merged galaxy itself, or must it wait for more mergers to complete the hardening process? Or is there *no* good coalescence process, and SMBH binaries stay hung up for more than a Hubble time?

Some of these questions will have better answers as these issues are explored theoretically in more detail. However, it will not be known how nature really accomplishes these tasks until observations by instruments like LISA, SIM, and JWST (and others that are planned) can tell us the properties of galactic nuclei, supermassive black hole binaries, the prevalence of gravitational radiation from them, and the properties of Population III stellar systems in the early universe. Given how rapidly this field is advancing, this part of the black hole formation story may need to be rewritten a decade or so from now.

Chapter 11

Fueling and Carburetion: Breakup and Accretion of Stars and Gas Clouds

In Chapter 7 we learned that black holes store energy in two different forms: (1) gravitational mass, which is expressed by the black hole mass parameter $M_{\bullet}$ (usually given in units of solar masses), and (2) rotational kinetic energy, which is embodied in the normalized angular momentum parameter

$$j \equiv \frac{J}{(GM_{\bullet}^2/c)} \quad (11.1)$$

(with j lying in the range $-1 \leq j \leq 1$). And in the next few chapters we will learn that black holes can *produce* energy in two different ways: (1) by converting matter falling toward the black hole into heat, radiation, and wind outflow (mechanical) energy (a process known as *accretion*), and (2) by converting the rotational energy of the black hole and of the matter falling toward it into particle, radiation, and jet outflow energy. Ultimately, the source of this rotational energy also is the accretion process, occurring either when the black hole first formed or as it takes on more matter from the gas and stars around it. *Black holes, therefore, can have two different kinds of fuel – mass and angular momentum – and they are both obtained during accretion.* In this chapter we discuss ways in which they might acquire this fuel in different situations.

11.1 Accretion of Matter

Accretion is the chief source of energy generated in the universe. It releases far more than that released by nuclear reactions in the center of stars or even in the thermonuclear explosions of stars. And it produces many, many orders of magnitude more energy than the variety of chemical reactions we rely upon in our daily lives (fossil fuel fires and engines; gunpowder, dynamite, and C4 explosions; *etc.*).

Examples of the power produced by the release of gravitational energy are readily apparent in our daily lives. Falling water is used by nature to erode mountains

and excavate canyons, and by man to power engines ranging from small woodland grain mills to large hydroelectric plants. Evidence of enormous releases of energy by astronomical accretion events is seen in the craters on the moon and on other satellites and planets in the solar system, the Barringer Meteor Crater in Arizona in the USA, the Tunguska impact of a meteorite or comet fragment in Siberia on 30 June 1908, and the series of explosions seen on Jupiter as pieces of Comet Shoemaker-Levy 9 impacted the planet on 16–22 July 1994. Yet, while terribly destructive by human standards, these events are insignificant when compared to accretion events onto neutron stars and black holes. The former events release a mere 0.000004% of the accreting matter's rest mass (mc^2) energy. On the other hand, as shown in Chapter 7, accretion onto black holes can release many percent or, under the right conditions, *even more than 100%*, of that energy – some of that coming from the rotational energy of the black hole itself.

11.1.1 Accretion Basics

For astrophysical engines powered by accretion onto a compact object of mass M the most important parameter describing the energy release is the accretion rate

$$\dot{M} \equiv \frac{dM}{dt} \quad (11.2)$$

i.e., the rate at which the mass of the central accreting compact object grows with time. For our purposes the compact object is a neutron star or black hole, but it also could be a moon, planet, or more normal star as well. Other important parameters are M and the radius of the compact object R .

11.1.1.1 Accretion Efficiency

For a star or planet, in which general relativity is not important, if the accretion rate is steady, the total energy released per unit time is

$$L_{\text{acc}} = \frac{GM}{R} \dot{M} \equiv \epsilon_{\text{acc}} \dot{M} c^2 \quad (11.3)$$

where the efficiency of accretion $\epsilon_{\text{acc}} \equiv GM/Rc^2 = r_g/R$ is a measure of how much of the rest energy of the accreting matter is released. For accretion onto the earth and sun, the energy released is not very significant: $\epsilon_{\text{acc}} \approx 7 \times 10^{-10}$ and 2×10^{-6} , respectively. Even for accretion onto the very compact white dwarf stars, the efficiency is still low ($\epsilon_{\text{acc}} \approx 7 \times 10^{-5}$). These are all much smaller than the efficiency of nuclear burning, which is about 7×10^{-3} (0.7%). However, for a $1.4 M_{\odot}$ neutron star, 10 km in radius, the efficiency of accretion is about 20% – thirty times

greater than that of nuclear burning.¹ So, for black holes and neutron stars accretion power wins big over nuclear burning.

11.1.1.2 The Eddington Accretion Rate

When $L_{\text{acc}} \geq L_{\text{Edd}}$ (equation (1.1)) the outward radiation pressure force on the electrons in the ionized accreting plasma will exceed the gravitational force on the accreting matter. This can happen when the accretion rate exceeds the Eddington rate

$$\dot{M}_{\text{Edd}} \equiv \frac{4\pi GM}{\epsilon_{\text{acc}} c \kappa_{\text{es}}} = 1.39 \times 10^{18} \text{ g s}^{-1} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right)^{-1} \left(\frac{M}{M_{\odot}} \right) \quad (11.4)$$

or for very massive accreting objects we can write

$$\dot{M}_{\text{Edd}} = 22.0 M_{\odot} \text{ yr}^{-1} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right)^{-1} \left(\frac{M}{10^9 M_{\odot}} \right)$$

Note that $\dot{M}_{\text{Edd}}$ is defined with the accretion efficiency included. When $\dot{M} \leq \dot{M}_{\text{Edd}}$ (many accretion situations, but not all), the accretion luminosity, relative to the Eddington luminosity, and the accretion rate, relative to $\dot{M}_{\text{Edd}}$ are equal

$$\dot{m} \equiv \frac{\dot{M}}{\dot{M}_{\text{Edd}}} = \frac{L_{\text{acc}}}{L_{\text{Edd}}} = \ell_{\text{rad}}$$

where ℓ_{rad} is the Eddington ratio (equation (2.17)). Since ϵ_{acc} appears within $\dot{M}_{\text{Edd}}$, the former rarely appears in most formulae that describe accretion.

When $\dot{M} > \dot{M}_{\text{Edd}}$ there is no guarantee that all of the matter accreting toward the compact object will actually make it onto the object's surface (or, if a black hole, into the horizon). Some (perhaps much) of the accreting matter may be blown outward by the radiation pressure force in a wind or other type of outflow. Often the Eddington accretion rate is taken to be a maximum possible rate at which plasma can accrete onto a compact object of mass M .

11.1.2 Generic Bondi–Hoyle–Lyttleton Accretion: Black Holes in the ISM

One of the most important aspects of accretion theory is a method of estimating the accretion rate $\dot{M}$ that was devised by Sir Herman Bondi, Sir Fred Hoyle, and Raymond Lyttleton of the University of Cambridge, UK, during the 1940s and early

¹ The actual general relativistic value from Chapter 7, equation (7.35) is $\epsilon = 1 - (1 - 2GM/Rc^2)^{1/2} = 23\%$.

1950s [403, 404, 405]. Originally developed for stars accreting gas, the method can be modified to handle black holes accreting whole stars or large gas clouds by allowing the black hole to play the role of the central star and the stars or gas clouds to play the role of individual gas particles.

11.1.2.1 Gas Accretion by a Stationary Star

Consider a central star of mass M immersed in a uniform-density cloud of hydrogen gas of temperature T in the interstellar medium (see Fig. 11.1a). We then can define a distance from the central object inside of which the average gas particle velocity $V_p = (3kT/m_H)^{1/2}$ is smaller than the escape velocity $V_{\text{esc}} = (2GM/R)^{1/2}$ from the central object. This distance is called the Bondi radius

$$R_B^{\text{GAS}} \equiv \frac{2GM}{V_p^2} = \frac{2}{3} \frac{GMm_H}{kT} = 5.36 \times 10^{16} \text{ cm} \left(\frac{M}{M_\odot} \right) \left(\frac{T}{20 \text{ K}} \right)^{-1} \quad (11.5)$$

When gas particles enter the region interior to this radius, they fall toward the central object and eventually accrete onto it. Using the Bondi radius, we can calculate the total mass that crosses into this region with time by noting that the gas will expand toward the central star at about the average particle velocity (*i.e.*, at about the sound speed)

$$\begin{aligned} \dot{M}_B^{\text{GAS}} &= \frac{4\pi R_B^2 n m_H V_p}{6} \\ &= 7.0 \times 10^{16} \text{ g s}^{-1} \left(\frac{M}{M_\odot} \right)^2 \left(\frac{T}{20 \text{ K}} \right)^{-3/2} \left(\frac{n}{100 \text{ cm}^{-3}} \right) \quad (11.6) \end{aligned}$$

The factor of 1/6 occurs because only half the particles in the radial direction stream toward the central object, and the radial direction is only one of three possible dimensions in which particles can travel. This rate is called the Bondi accretion rate and, for a star of 1–10 $M_\odot$ immersed in the dense interstellar medium (with $n \sim 100 \text{ cm}^{-3}$ and $T \sim 20 \text{ K}$), it has values of $10^{15-19} \text{ g s}^{-1}$. *Just sitting at rest in a dense region of the interstellar medium a star can double its mass in less than a billion years.*

11.1.2.2 Accretion by a Moving Star

At the other extreme a star might be moving through the gas at a (supersonic) velocity V_0 that is so fast that the average particle velocity V_p is unimportant (see Fig. 11.1b). Then, relative to the accreting star, the gas particles simply follow independent orbits with eccentricity

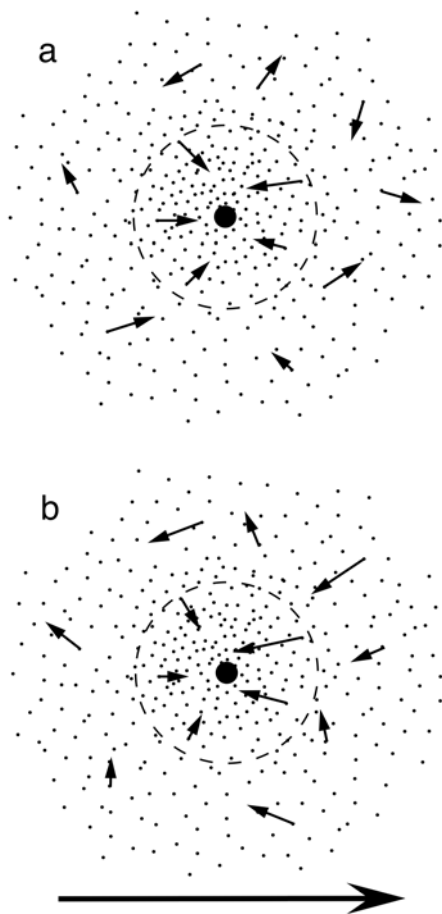


Fig. 11.1: Schematic pictures of stationary and moving accreting stars immersed in an ambient gas. (a) accretion by a stationary star; gas particle motions are random in this case. (b) accretion by a star moving toward the right; relative to the star gas particles have a net motion to the left. Inside each Bondi radius (dashed circle) gravity overcomes the motions of the particles, causing the gas to flow toward the star.

$$e = \left[1 + \frac{2\varepsilon\ell^2}{(GM)^2} \right]^{1/2}$$

where $\varepsilon = \frac{1}{2}V_0^2 - \frac{GM}{r}$ is the particle energy per unit mass at a distance r , and $\ell = rV_0 \sin \theta$ is the particle's angular momentum per unit mass, all relative to the moving star. Particles become bound to the star if $e < 1$ or $\varepsilon < 0$. The Bondi radius for collisionless particles, then, is

$$\begin{aligned}
 R_B^{PART} &= \frac{2GM}{V_0^2} \\
 &= 5.41 \times 10^{16} \text{cm} \left(\frac{M}{M_\odot} \right) \left(\frac{V_0}{7.0 \times 10^4 \text{cm s}^{-1}} \right)^{-2} \quad (11.7)
 \end{aligned}$$

and the amount of matter crossing into the Bondi sphere is only that on the side facing into the wind, or²

$$\begin{aligned}
 \dot{M}_B^{PART} &= \pi R_B^2 n m_H V_0 \\
 &= 1.08 \times 10^{17} \text{g s}^{-1} \left(\frac{M}{M_\odot} \right)^2 \left(\frac{V_0}{7.0 \times 10^4 \text{cm s}^{-1}} \right)^{-3} \left(\frac{n}{100 \text{cm}^{-3}} \right) \quad (11.8)
 \end{aligned}$$

This estimate of the accretion rate is similar to that in equation (11.6), with V_p replaced by V_0 . Of course, equation (11.8) is not the final accretion rate onto the central star unless particle–particle collisions do eventually become important in the bound gas, allowing it to dissipate its orbital (internal) energy and sink toward the star.

The similarity of equations (11.6) and (11.8) prompted Bondi and Hoyle to suggest a general estimate of

$$\boxed{\dot{M}_B \approx 10^{17} \text{g s}^{-1} \left(\frac{M}{M_\odot} \right)^2 \left(\frac{V_p + V_0}{7.0 \times 10^4 \text{cm s}^{-1}} \right)^{-3} \left(\frac{n}{100 \text{cm}^{-3}} \right)} \quad (11.9)$$

While this is certainly an oversimplification, it is a reasonable order-of-magnitude estimate, good to a factor of two or so, and we will use it to estimate the accretion rate onto isolated neutron stars and stellar-mass black holes traveling through the interstellar medium. The actual rate will depend not only on the details of gas dynamical flow, but also in how fast the particles collide and redistribute their velocities inside the Bondi radius.

11.1.3 Accretion of Stars onto Massive Black Holes in AGN and Globular Clusters

The situation in the centers of many galaxies, AGN, and quasars is very similar to the Bondi problem, except that (1) the star is replaced by a supermassive ($M_\bullet \approx 10^6\text{--}10^8 M_\odot$) black hole and (2) the gas particles are replaced by whole stars (of $\sim 2 \times 10^{33}$ g instead of 1.7×10^{-24} g). The velocity dispersion of the stars in a galactic center can be several hundred km s^{-1} or higher and the inner density of the stellar “gas” can be as high as a thousand or more stars per cubic parsec. We will

² Note that our expression for $\dot{M}_B^{PART}$ in equation (11.8) differs significantly from the incorrect one in [406] because we use $R_B = 2r_g (c/V_0)^2$ instead of $2r_g (c/V_0)$. This produces a V_0^{-3} dependence instead of an inverse linear one.

find that the accretion of stars alone from the inner part of the galactic bulge can explain the rates of accretion seen in “inactive” or low-luminosity AGN, but not in the more active Seyfert galaxies and quasars.

11.1.3.1 The Structure of Star Clusters, Stellar Bulges, and Spheroidal Galaxies

Before we discuss how stars can feed black holes at the centers of spheroidal galaxies (or spheroidal bulges), we must briefly discuss the structure of spheroidal stellar systems. In general, these systems are tri-axial, not spherical; and they may be not only egg-shaped (prolate elliptical) or yo-yo-shaped (oblate spheroidal), which are bi-axial, but sometimes they may be further shortened in a third dimension (tri-axial). In this sub-section, however, we will keep things simple and treat all spheroidal stellar systems as having spherical symmetry – generally isothermal spheres (or $n = \infty$ polytropes) to be precise. The stellar density of such a sphere was discussed in Sections 5.2.4 and 10.4.1, and is given by

$$n = \frac{\sigma_V^2}{2\pi G \langle m_\star \rangle} \frac{1}{r^2} \quad (11.10)$$

where $\langle m_\star \rangle$ is the average star mass in the stellar system, and σ_V is its velocity dispersion. “Isothermal” means that σ_V can be consider constant throughout the system of stars.

A famous relation between σ_V and the mass of a spheroidal galaxy was discovered in 1976 by Sandra Faber and Robert Jackson of UC Santa Cruz: the galaxy mass is proportional to the fourth power of the velocity dispersion [407]. For computational purposes we will use a theoretical Faber–Jackson relation that is consistent with our theoretical $M_\bullet$ – σ relation (equation (10.12)) and the empirical M_{bulge} – $M_\bullet$ relation (equation (2.12))

$$\sigma_V \approx 108 \text{ km s}^{-1} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/4} \quad (11.11)$$

Taking $\langle m_\star \rangle$ to be about $1 M_\odot$ the density of any singular isothermal spheroidal system then is

$$n = 44 \text{ pc}^{-3} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/2} \left(\frac{\langle m_\star \rangle}{M_\odot} \right)^{-1} \left(\frac{r}{100 \text{ pc}} \right)^{-2} \quad (11.12)$$

For a “core” galaxy, the stellar density profile will be similar to a singular isothermal sphere profile for radii greater than the galaxy core radius; for radii inside the core radius, the profile flattens. For our theoretical treatment, therefore, we shall take a core galaxy’s profile to be

$$\boxed{\begin{aligned} n^{\text{CORE}}(r) &= n_c & r < r_c \\ &= n_c \left(\frac{r_c}{r}\right)^2 & r_c \leq r \end{aligned}} \quad (11.13)$$

where the core radius (from equations (2.12) and (2.15)) is

$$r_c = 48 \text{ pc} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{1/2} \quad (11.14)$$

and the core density is

$$n_c = n(r_c) = 190 \text{ pc}^{-3} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{-1/2} \left(\frac{\langle m_{\star} \rangle}{M_{\odot}} \right)^{-1} \quad (11.15)$$

Core galaxies, therefore, do not reach the high central densities predicted by equation (11.12). Finally, even though “power-law” galaxies do not have a core, we still can write equation (11.12) as

$$\boxed{n^{\text{SIS}}(r) = n_c \left(\frac{r_c}{r}\right)^2} \quad (11.16)$$

for all r .

With equations (11.11) and (11.14), we can use equations (10.8) and (10.9) to compute the core mass and the mass deficit

$$\boxed{M_{\text{core}} = 0.0087 M_{\text{bulge}} = 6.7 M_{\bullet}} \quad (11.17)$$

$$\boxed{M_{\text{def}} = 0.0174 M_{\text{bulge}} = 13 M_{\bullet}} \quad (11.18)$$

Because the mass deficit is so small compared to the total bulge mass, we simply can use equation (10.7) for the total spheroidal mass, whether or not the structure has a core. This gives us a very simple mass–radius relation for all of these model spheroidal galaxies:

$$\boxed{r_{\text{bulge}} = 1.8 \text{ kpc} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{1/2}} \quad (11.19)$$

Beyond r_{bulge} we will assume that the bulge is truncated, like a King model, so any stars outside that radius do not contribute appreciably to the total bulge mass.

11.1.3.2 Delivery and Carburetion of the Stellar Fuel

The accretion of stars is essentially Bondi accretion, but instead of the Bondi radius, investigators usually use the radius of influence of the black hole

$$r_h = \frac{G M_\bullet}{\sigma_V^2} \approx 4.8 \text{ pc} \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right)^{1/2} = 4.8 \text{ pc} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/2} \quad (11.20)$$

which is simply one-half of the Bondi radius and which scales the same way as r_c and r_{bulge} . (This is about 10% of the core radius, which is a good approximation to the data for M87 in Fig. 1.2.)

The region inside r_h is often referred to as the “cusp”. Not surprisingly, if the galaxy has a power-law profile, the amount of stellar mass in the cusp is of order the black hole mass itself

$$m_{\text{SIS}}(r_h) = 2 M_\bullet \quad (11.21)$$

However, if the galaxy has a core structure, and if $r_h \approx 0.1 r_c$, even if the stellar density profile in the cusp also is an isothermal sphere, the mass of the stars there is only $2 M_\bullet (r_h/r_c)^2 \approx 0.02 M_\bullet$. So, for the remainder of this subsection we will deal only with black holes in power-law galaxies, since these have more stellar fuel within the cusp and, therefore, a better chance of powering the central black hole engine with those stars.

Two-Body Relaxation: The Nearly Collisionless Stellar Gas

If stars in the cusp were to collide frequently, like a dense ideal gas, then every star that entered the Bondi radius would enter the black hole, and we would predict the very large accretion rate of

$$\begin{aligned} \dot{M}_{B\bullet}^* &= \pi r_h^2 n(r_h) \langle m_\star \rangle \sigma_V \\ &= 150 M_\odot \text{ yr}^{-1} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{3/4} \end{aligned} \quad (11.22)$$

(cf. equation (11.8)). However, stars do *not* collide frequently, as we can see by comparing the mean free path of stars in this region

$$\lambda_\star = \frac{1}{n \Sigma_\star}$$

with the size of the region r_h , where the “Coulomb” cross-section for star scattering

$$\Sigma_\star = 2\pi r_{\text{coul}}^2 \ln A$$

r_{coul} is the average star–star impact parameter when an encounter occurs

$$r_{\text{coul}} = \frac{2G \langle m_\star \rangle}{\sigma_V^2} = 2.3 \times 10^{12} \text{ cm} \left(\frac{\langle m_\star \rangle}{M_\odot} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{-1/2}$$

and $\ln A$ is the Coulomb logarithmic factor

$$\ln A = \ln(r_{max}/r_{min}) = \ln\left(\frac{1}{n r_{coul}^3}\right)^{1/3} \approx 8-12$$

We will take $\ln A$ to be about 10. The mean free path for star–star encounters, then, is

$$\lambda_\star = 1.5 \times 10^5 \text{ pc} \left(\frac{\langle m_\star \rangle}{M_\odot}\right)^{-1} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot}\right)^{3/2} \left(\frac{\ln A}{10}\right)^{-1} \quad (11.23)$$

which is much greater than the black hole radius of influence in equation (11.20). To illustrate just how ineffective stellar encounters are, consider the two-body relaxation time (equation (10.10)) for this situation ($M_{\text{clst}} = (r_h/r_{\text{bulge}}) M_{\text{bulge}}$, $r = r_h$, $\langle m_\star \rangle = 1 M_\odot$)

$$\tau_{\text{rlx}}(r_h) = 10 \text{ Gyr} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot}\right)^{5/4} \left(\frac{\langle m_\star \rangle}{M_\odot}\right)^{-1} \left(\frac{\ln A}{10}\right)^{-1} \quad (11.24)$$

which is nearly the present age of the universe.

The stellar “gas”, therefore, is nearly collisionless, with stars following simple orbits and almost no re-distribution of star velocities by star–star encounters. In AGN, therefore, even though stars within the Bondi radius are bound to the central black hole, they will remain trapped in their orbits about that hole and sink toward it only on very long time scales.

In order to get a better, and more realistic, estimate of the accretion rate of stellar fuel onto a galaxy’s central black hole than equation (11.22), we need to understand the processes of tidal disruption of stars and the concept of the loss cone.

Carburetion of the Fuel: Tidal Disruption of Stars by the Black Hole

Stars like white dwarfs, neutron stars, and stellar-mass black holes can plunge directly into a supermassive black hole’s horizon *as whole entities*, without being disrupted. This process is one type of accretion, and it will generate gravitational waves and grow the black hole over time. However, such accretion will not produce the electromagnetic radiation and jets that we see in active galactic nuclei. For that we need to accrete tenuous ionized plasma toward the black hole, not dense “bricks”. One source of ionized plasma is the *tidal disruption* of less dense stars as they approach too close to the central black hole.

If a star of radius $R_\star$ lies at a distance r from the central black hole, the tidal force exerted on that star by the hole is $\sim GM_\bullet M_\star R_\star / r^3$, while the gravitational force of the star on its own outer layers is $GM_\star^2 / R_\star^2$. The tidal force of the black hole, therefore, begins to exceed the star’s gravitational force inside the “tidal radius” or

“Roche limit”³

$$\begin{aligned} r_T^* &= R_\star \left(\frac{M_\bullet}{M_\star} \right)^{1/3} \\ &= 1.6 \times 10^{13} \text{ cm} \left(\frac{R_\star}{R_\odot} \right) \left(\frac{M_\star}{M_\odot} \right)^{-1/3} \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right)^{1/3} \end{aligned} \quad (11.25)$$

or inside about 5.3×10^{-6} pc for a solar-type star approaching the central supermassive black hole in a $10^{10} M_\odot$ galactic bulge. ($R_\odot = 6.955 \times 10^{10}$ cm is the average solar radius.) When this occurs, the star is shredded back into ionized plasma and forms an accretion disk around the black hole. (How such an accretion disk produces the observed radiation and jets is discussed in Chapters 12 to 15.) In 1994 we witnessed a similar phenomenon when Comet Shoemaker-Levy 9 first approached Jupiter and was tidally disrupted, with its remains strewn along an elliptical orbital path.

Solar-type stars ($R_\star = R_\odot$, $M_\star = M_\odot$) are good sources of plasma until the black hole becomes too massive, and its horizon radius ($r_S = 2GM_\bullet/c^2$ for a non-rotating black hole) exceeds R_T^* . This happens when $M_\bullet$ exceeds the value

$$M_{T\bullet} = 1.1 \times 10^8 M_\odot \left(\frac{R_\star}{R_\odot} \right)^{3/2} \left(\frac{M_\star}{M_\odot} \right)^{-1/2} \quad (11.26)$$

Many supermassive black holes in AGN (those in quasars and radio galaxies) are much more massive than $10^8 M_\odot$, and so *must be swallowing stars like the sun whole*, just like more compact objects. However, red giant and supergiant stars have $R_\star = 15\text{--}150 R_\odot$, so even black holes with masses as great as

$$M_\bullet = (0.6\text{--}20) \times 10^{10} M_\odot$$

still can accrete a substantial amount of ionized plasma from disrupted giant and supergiant stars. This includes all currently known AGN and quasars.

The closest approach of a stellar orbit to a central black hole r_P is called the “peribarathron”.⁴ If a star travels on an orbit such that $r_P \leq r_T^*$, then it will be tidally disrupted during its peribarathron passage and will accrete onto the black hole as a gaseous disk. We will call any orbit for which $r_P = r_T^*$ a “tidal” orbit. In principle, tidal orbits can be any elliptical shape. However, because stars in galactic nuclei tend to be, on the average, a distance $\sim n^{-1/3} \approx 0.04$ pc or farther from the black hole, any orbits approaching within 10^{13} cm (10^{-5} pc) must be *highly* elliptical (comet-

³ This and other tidal phenomena are named after the French mathematician Edouard A. Roche of the university at Montpellier who, in 1848, proposed that Saturn’s rings were formed by the tidal breakup of a former moon of that planet.

⁴ The Greek word $\beta\alpha\rho\alpha\theta\rho\omicron\nu$ (analogous to “helion” for orbits about the sun) was first used for supermassive black hole orbits in 1977 by Peter Young, Greg Shields, and J. Craig Wheeler, of The University of Texas at Austin [408]. In ancient Athens the Barathron was a supposedly bottomless pit into which incorrigible criminals were thrown and from which there was no return.

like) orbits, with an eccentricity very near unity ($1 - r_T^*/r < e \approx 1$). If we define the angle $\theta_{\ell c} = 0$ to describe a “head-on” collision with the black hole, then any star that is at a distance r from the hole and is heading toward it at an angle less than the ratio of the semi-minor and semi-major axes of the tidal orbit will be disrupted and accreted as gas. This angle is called the “loss-cone angle” and is given by

$$\theta_{\ell c} \approx (1 - e_{crit}^2)^{1/2} \approx \left(\frac{2r_T^*}{r} \right)^{1/2} \quad (11.27)$$

The loss cone is actually a conical region in velocity space (V_r , V_θ , V_ϕ) with an opening angle of $2\theta_{\ell c}$, which is different for different distances from the black hole. Stars at greater distances have narrower loss cones. If the star’s velocity vector is directed inward, and lies within the loss cone in velocity space, then that star will be accreted.

This type of accretion will be short-lived, however, because there are few star–star collisions to refill the loss cone as it is depleted by tidal disruption. In fact, in no more time than it takes for all stars on doomed orbits to make it to the center

$$\tau_{cross} \approx \frac{r_h}{\sigma_V} = 4.4 \times 10^4 \text{ yr} \left(\frac{M_{bulge}}{10^{10} M_\odot} \right)^{1/4} \quad (11.28)$$

all potential accreters will be gone; this is essentially instantaneous on cosmic time scales.

11.1.3.3 Realistic Estimates of the Tidal Disruption Accretion Rate onto a Central Black Hole

Determining the actual accretion rate onto a black hole in the center of a star cluster is a difficult task. These days it requires numerical simulations. One of the first *analytic* estimates was given by J.M. Anthony Danby and G.L. Camm of the University of Manchester, UK, in 1957 [409]. They reasoned that after the loss cone were depleted, the accretion rate would be reduced by a factor of $\sim \theta_{\ell c}^2$ to $\dot{M}_{B\bullet}^* (r_T^*/r_h)$ or, with our Galaxy and black hole model,

$$\dot{M}_{DC\bullet}^* = 2 \times 10^{-4} M_\odot \text{ yr}^{-1} \left(\frac{M_{bulge}}{10^{10} M_\odot} \right)^{7/12}$$

for solar-type stars. Other analytic estimates by Juhan Frank and Martin Rees in 1976 [410], Alan P. Lightman and Stuart L. Shapiro in 1977 [411], and Haldan Cohn and Russell M. Kulsrud in 1978 [412] all produce similar low levels of accretion rate, although their variations with black hole/galaxy mass differ, depending on the assumptions made.

Some of the best estimates of stellar accretion rate so far have been determined by David Merritt of the Rochester Institute of Technology in New York. He found a rate that actually *decreases* (albeit slowly) with black hole/galaxy mass [413]

$$\begin{aligned}
 \dot{M}_{M_\bullet}^* &= 7.1 \times 10^{-4} \text{ M}_\odot \text{ yr}^{-1} \left(\frac{\sigma_V}{70 \text{ km s}^{-1}} \right)^{7/2} \left(\frac{M_\bullet}{10^6 \text{ M}_\odot} \right)^{-1} \\
 &= 2.5 \times 10^{-4} \text{ M}_\odot \text{ yr}^{-1} \left(\frac{M_{\text{bulge}}}{10^{10} \text{ M}_\odot} \right)^{-1/8}
 \end{aligned}
 \tag{11.29}$$

when we use an M – σ relation of $M_{\text{bulge}} \propto \sigma_V^4$ and a core structure for the stellar distribution. For galaxies with a singular isothermal sphere stellar distribution in their centers, the accretion rate will be about 100 times higher. These simulations take into account the contribution of star orbits outside r_h and flatter density profiles that develop for larger black holes – effects not included in the earlier analytic estimates.

Table 11.1 compares equation (11.29) (upper half of table) with the presumed accretion rates for various classes of active galaxy (lower half) for the different black hole masses in (columns 3–6). In the upper half, predicted accretion rates (in solar masses per year) are given for the models discussed in Sections 11.1.3 and 11.1.4. In the lower part of the table, column 2 gives typical Eddington ratios (L/L_{Edd}) for each AGN type at lower left, which convert to the accretion rates (again in $\text{M}_\odot \text{ yr}^{-1}$) for the different black hole engine masses given in the remaining columns.

Table 11.2 shows the relative accretion contribution from tidal breakup of various types of stars and clouds (but not of star clusters) that might exist in a typical galactic nucleus (column 1). Columns 2 and 3 give each object’s mass and size in solar units; column 4 gives the maximum black hole mass that can fully tidally disrupt each object; column 5 gives the number of objects in a typical Seyfert galaxy, relative to the number of solar-type stars; and column 6 gives the estimated accretion rate contribution of each type (proportional to the product of object mass, size, and relative fraction), divided by that contributed by solar-type stars alone. Blue supergiants contribute little, because their numbers tend to be much smaller than predicted from a Salpeter IMF ($f_\star \sim 0.02$), and white dwarfs contribute nothing, as they are not tidally disrupted by black holes in the $>10^6 \text{ M}_\odot$ range. However, red giants and supergiants may be responsible for perhaps one-third of the accretion in a typical LLAGN or Seyfert powered by stellar tidal disruption.

The low accretion rates from tidal disruption of stars in a core galaxy (**Table 11.1**) are sufficient to explain *only* quiescent galaxies. In most cases low-luminosity AGN may be explainable solely by tidal disruption of stars *if* the central stellar density structure is cusp-like (singular isothermal sphere). However, if the theoretical decrease of stellar consumption rate with black hole/galaxy mass (equation (11.29)) is valid, it appears that even a high-density stellar cusp may not be able to power LLAGN with black holes much above $10^8 \text{ M}_\odot$. Tidal disruption in a strong stellar cusp, however, may indeed power Seyfert galaxies with black holes of $10^{6-7} \text{ M}_\odot$ and even the lower-mass narrow-line Seyfert 1 galaxies. However, from what we now know about the central structure of galaxies, *it appears that simple tidal disruption of stars is incapable of powering the higher-luminosity, higher-mass active Seyfert galaxies and quasars* (numbers in bold in the table).

Table 11.1: Comparison of accretion rate by tidal disruption of stars and rates typical of various types of AGN for different mass central black holes*

Type of tidal disruption		$10^6 M_\odot$	$10^7 M_\odot$	$10^8 M_\odot$	$10^9 M_\odot^{**}$
<i>Of stars in core galaxy (§11.1.3)</i>		3.5×10^{-4}	2.6×10^{-4}	1.9×10^{-4}	~ 0
<i>Of stars in cusp galaxy (§11.1.3)</i>		3.5×10^{-2}	2.6×10^{-2}	1.9×10^{-2}	~ 0
<i>Of newborn star clusters (§11.1.4)</i>		7×10^{-2}	0.12	0.21	10^{-2}
Type of AGN	Typical Eddington ratio (L/L_{Edd})	$10^6 M_\odot$	$10^7 M_\odot$	$10^8 M_\odot$	$10^9 M_\odot$
<i>Quiescent galaxy</i>	$\leq 10^{-4}$	$\leq 2.2 \times 10^{-6}$	$\leq 2.2 \times 10^{-5}$	$\leq 2.2 \times 10^{-4}$	$\leq 2.2 \times 10^{-3}$
<i>LLAGN</i>	$\leq 10^{-2}$	$\leq 2.2 \times 10^{-4}$	$\leq 2.2 \times 10^{-3}$	$\leq 2.2 \times 10^{-2}$	≤ 0.22
<i>Seyfert/quasar</i>	≤ 0.1	$\leq 2.2 \times 10^{-3}$	$\leq 2.2 \times 10^{-2}$	≤ 0.22	≤ 2.2
<i>NLSy1/BAL QSO</i>	≤ 1	$\leq 2.2 \times 10^{-2}$	≤ 0.22	≤ 2.2	≤ 22

*In $M_\odot \text{ yr}^{-1}$; Eddington rates assume 10% accretion efficiency

**Above $\sim 10^8 M_\odot$ (equation (11.26)) solar-type stars are not tidally disrupted and the fueling rate is dominated by the few giants and supergiants that can get to the black hole (see Table 11.2).

Table 11.2: Long-term black hole growth by different types of stars and clouds

Star type	$M_{\star}/M_{\odot}$	$R_{\star}/R_{\odot}$	$M_{T\bullet}/M_{\odot}$	$f_{\star}$	$\dot{M}_{avg\bullet}^{\star}/\dot{M}_{avg\bullet}^{\odot}$
<i>White dwarf</i>	0.5	0.01	2.3×10^5	~ 1	~ 0
<i>Dwarf</i>	1	1.0	1.6×10^8	~ 1	1.0
<i>Red giant</i>	1	15	9×10^9	0.01	0.15
<i>Red supergiant</i>	1	150	3×10^{11}	0.002	0.3
<i>Blue supergiant</i>	20	15	2×10^9	≤ 0.0002	≤ 0.06
<i>Molecular cloud</i> *	10^4	1.3×10^8	2.5×10^{18}	—	3.7

*For active giant spiral (Seyfert) galaxies

It is currently believed that the solution to this problem of powering the most luminous engines in the universe lies with the other component of accretion fuel available in the host galaxies – the interstellar gas itself. Indeed, interstellar gas accretion, combined with stellar disruption, is also probably important in powering many of the lower-luminosity objects in [Table 11.1](#).

11.1.4 Accretion of Interstellar Gas onto Massive Black Holes in AGN: Tidal Disruption of Newborn Star Clusters

Each galaxy, even an elliptical, has enough gas in its interstellar medium to potentially feed a central AGN: spirals have a large amount of gas in their disks and merging or forming galaxies can have more gas than stars collected inside their massive dark matter halos. Therefore, we now shall consider the accretion of gas from the interstellar medium into the vicinity of a central supermassive black hole in a galaxy. This is a very large and complex field of study, so we will begin to only scratch the surface of this topic here. To keep the discussion from becoming too extensive, we shall deal mainly with galactic bulges of moderately large size ($M_{\text{bulge}} \sim 10^{10-11} M_{\odot}$) that contain black holes of Seyfert-type mass ($\sim 10^7-8 M_{\odot}$). And we shall concentrate primarily on one particular series of gas accretion processes (among many such possible series) that appears very promising. This series of steps will carry gas from the spiral disk into the nucleus, lock that gas into bound clusters of newly-formed stars, and carry those stars to within the black hole's sphere of influence. This chain of events can supply up to $\sim 0.2-0.4 M_{\odot} \text{ yr}^{-1}$ to within r_h . Therefore, together with the process of tidal disruption discussed in the previous subsection, these processes can account for all but the most massive black holes ($> 10^9 M_{\odot}$) and the most rapidly accreting objects (bright quasars and BAL QSOs) in [Table 11.1](#).

There are three major steps in this chain, each of which works to remove angular momentum of the interstellar gas or clusters:

1. Funneling of gas into the galactic nucleus (to within ~ 1 kpc) by a galactic stellar bar.

2. Breakup of that nuclear gas into clouds and then dissipation of those clouds' angular momentum by the process of dynamical friction.
3. Conversion of those gas clouds into star clusters and then further dynamical friction inspiral of those clusters into the radius of influence of the black hole, where their young stars then contribute to the tidal disruption rate.

While the bar dissipation step has been discussed in the literature for almost three decades now, the two dynamical friction steps are new, proposed herein by the author, and therefore have not yet been subject to the rigors of peer review nor to the benefits of multiple investigations by different authors. Nevertheless, this author thought it important to suggest in this book at least one plausible mechanism for fueling supermassive black holes that is both connected with the current theoretical thinking on the subject and consistent with the current understanding of interstellar gas and star formation. I am grateful to Jorge Pineda, Konstantinos Tassis, and Thangasamy Velusamy (all at JPL), for their lengthy discussions on the subject of molecular clouds and star formation.

11.1.4.1 STEP 1: Funneling of Atomic Gas into a Galactic Nucleus by a Large-Scale Stellar Bar

Summary of the Bar Accretion Process

At the present time the most favored mechanism for getting interstellar gas into the nucleus of a galaxy, where it could potentially fuel a black hole, is the action of a large stellar bar that dominates much of a barred spiral galaxy's structure. The problem with interstellar gas (even in ellipticals) is that it orbits the galaxy at large radii – many kiloparsecs. The gas must somehow shed 99.99% or more of its angular momentum in order to get from 10 kpc down to less than 1 pc from the black hole where the gas can be utilized as fuel.

A galactic bar can funnel the gas to inside a kiloparsec or less (a loss of 99% of the angular momentum), but a different mechanism is needed to remove the last 0.99%. In order to understand how a bar funnels gas into a galactic nucleus, we first must understand a little about galactic bars themselves (see Fig. 11.2). In the spheroidal potential of a central galactic bulge orbits of stars about the galactic center can be oriented in any direction and have a variety of shapes and sizes. In a barred spiral potential, however, the orbits are preferentially along the length of the bar (Fig. 11.2a); these are called x_1 orbits. There also exist x_2 orbits that are perpendicular to the bar, which have smaller orbital semi-major axes, giving the bar its smaller extent perpendicular to its length.⁵ Stars exist in both x_1 and x_2 orbits, but because stellar collisions essentially never occur, both populations of stars fly right

⁵ There also are x_2 orbits with semi-major axes larger than the bar width, but they are more rare and they must lie outside two regions called the inner Linblad resonances (isolated regions at $x = 0$, $y \approx \pm 6.5$ in Fig. 11.2a).

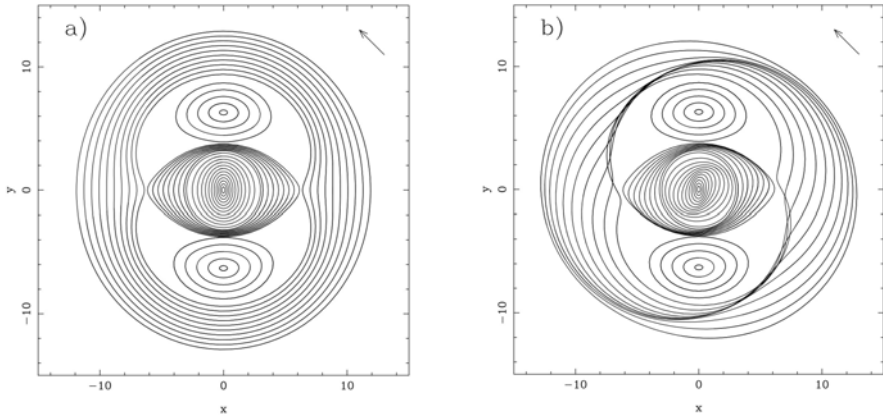


Fig. 11.2: STEP 1: Accretion of gas into the central 1 kpc in a spiral disk galaxy. (a): Locus of stellar orbits in a disk galaxy (rotation is counter-clockwise); orbits avoid the outer Linblad resonances (isolated regions at $[0, \pm 6.5]$) and have different sized semi-major axes, depending on whether they are oriented along x or y . This creates the horizontal bar shown. (b): Same as (a) with gas cloud orbits added; gas collides and produces spiral shocks, which funnel material into the bar (*i.e.*, into the inner 4 kpc of the galaxy). If a central mass concentration exists, a secondary trailing spiral within the bar can funnel gas into the central kiloparsec. However, more than two more orders of magnitude further decrease in radius are needed to get the gas into the sphere of influence of the central black hole. Figures from Fig. 78a and 79b of [414]; originals kindly supplied by F. Combes; reproduced with permission.

by each other, affected only by the global gravitational potential of the star, gas, and dark matter components.

Interstellar atomic gas behaves quite differently from stars. The gas also must exist preferentially in either x_1 or x_2 orbits, but these two streams of gas cannot pass through each other. The streams collide, shock, and lose angular momentum in the process. This leads to inwardly-directed gas streams that spiral toward the galactic center (see Fig. 11.2b).⁶ With this process inflow rates of a few solar masses per year or more into the inner 1 kpc can be realized – plenty to fuel an AGN, if that gas can make it another factor of several hundred in distance to the black hole.

It is important to note that the gas *must remain atomic* in this mechanism. Once it begins to condense into compact molecular clouds, those clouds will be nearly as collisionless as the stars themselves; they then will *not* form angular-momentum dissipating shocks.

⁶ If a second, smaller bar also exists in the same galaxy – well inside the larger bar (as is sometimes the case) – a second stage of inflow can be set up (see center of Fig. 11.2b). Hence, this double-bar mechanism is often called the “bars within bars” process.

Problems with the Bar Accretion Mechanism

Despite the popularity and theoretical success of the bar mechanism, there is scant evidence for it in astronomical observations. Initial studies of disk galaxies [415, 416] found no difference in the incidence of bars in Seyferts compared with normal galaxies. However, more recent studies [417, 418] suggest that there may be an excess of bars in AGN (70–80% in Seyferts vs. 55–60% in normal spirals), but this is only a 2σ result.

Surprisingly, the one observable feature that *does* correlate with AGN activity is the existence of a “nuclear star-formation ring” – a narrow annulus of newly-formed stars around the nucleus, perhaps a few hundred parsecs in radius [416]. Such rings also are strong tracers of starburst activity in the galactic center. The biggest surprise of these rings, however, is that they *cannot* exist in the presence of a strong bar! The asymmetric potential of a strong bar would tear the ring apart.

There are other phenomenological properties that do not favor a strong bar as a feeding mechanism. AGN occur preferentially in early type galaxies (ellipticals, S0s, and spirals up to Sbc); they are much less common in Sc–Sd galaxies. Yet, atomic gas is not very common in early type galaxies; in such galaxies atomic gas is dominated by the molecular gas component by a factor of 10 or more. Furthermore, in the early spiral galaxies the atomic gas occurs mainly in the outer regions of those galaxies, while it is the molecular gas that completely dominates the central galactic regions. As pointed out earlier, when the atomic gas condenses into molecular clouds, that gas is lost to the bar mechanism. There must be a different method of getting all of this molecular gas into the nucleus so it can fuel an AGN.

Possible Ways out of the Bar Accretion Problem

In her review of the fueling and evolution of AGN [419], Shardha Jogee, then at the Space Telescope Science Institute, lists several possible methods for getting gas from ~ 1 kpc down to parsec scales. They are as follows:

- Dynamical friction of H I clumps of gas that condense from the general inward bar-driven flow. We find below that this is an important process, but the “clumps” evolve significantly – into star clusters.
- Tidal disruption of the H I clumps near the black hole. Below we show that the tidal field of the galaxy itself is more important, and clumps are disrupted too far from the hole to be accreted in the gaseous state.
- Bars within bars, going deep into the nucleus. Besides a few secondary bars in some galaxies, however, there is as yet no strong observational evidence for this process.
- Tidal disruption of stellar clusters and their stars. This mechanism also features prominently in our discussion below.

To these four possibilities this author would like to add a fifth:

- Magnetic braking of the H I clumps. This mechanism was suggested to the author by Konstantinos Tassis, of JPL.

Such clumps, though only partially ionized, nevertheless may be supported internally by magnetic fields, which could be tied to the strong interstellar magnetic field. Their angular momentum loss would be described mathematically by one of the MDAF (magnetically-dominated accretion flow) solutions discussed in Section 12.3 – perhaps the same sub-Alfvénic solution used to describe accretion very close to the black hole. This process can potentially remove angular momentum faster than dynamical friction, but will depend critically on how much the cold gas is ionized (which will, in general, be much higher near the nucleus, where supernova explosions and AGN create a strong field of ionizing cosmic rays).

Three of the above processes figure prominently in the discussion below. The bars-within-bars process does not seem to be consistent with the observed existence of star-formation rings, nor are secondary bars common in AGN; so it is discounted here. The fifth, magnetic, process is not discussed in detail in this chapter; but it appears to be the one remaining process that may be able to drive accretion in the most massive and luminous objects – ones that have a lot of gas, have a lot of potential cosmic ray activity, require the highest accretion rates, and yet have too massive a black hole to tidally disrupt most of the stars.

11.1.4.2 STEP 1.5: The Physics of Dynamical Friction Accretion of Mass-Losing Atomic Gas Clumps, Molecular Clouds, and Star Clusters

Before we discuss the actual inspiral of clumps, clouds, and clusters, we first must outline the physics of dynamical friction accretion of these objects in a spherical galactic gravitational potential.

Structure of Hydrogen Molecular Clouds and Atomic Gas Clumps

As in Section 11.1.3, we first will discuss briefly the structure of two additional (gaseous) types of spherical objects that will be important in dynamical friction accretion in a galactic nucleus.

Molecular Clouds. Molecular clouds are the factories in which stars are formed. They have a mass–radius relation rather similar to that of star clusters (equation (11.19)) [420]

$$R_{cld} = 6 \text{ pc} \left(\frac{M_{cld}}{10^5 M_{\odot}} \right)^{1/2.3}$$

For simplicity we shall use an $R^{1/2}$ scaling, which is reasonable for the mass range we will be investigating and easier to work with theoretically

$$R_{cld} = 6 \text{ pc} \left(\frac{M_{cld}}{10^5 M_{\odot}} \right)^{1/2} \quad (11.30)$$

However, before stars form in their interiors, the molecular gas is quite cold (10–20 K) – much “colder” than the stellar systems of similar mass and radius discussed in the previous subsection, whose velocity dispersion “temperatures” are of the order

$$T \approx \frac{\sigma_V^2}{R} = 4500 \text{ K} \left(\frac{M_{\text{bulge}}}{10^5 M_{\odot}} \right)^{1/2}$$

If molecular clouds and star clusters have a similar mass–radius relation, but the former are far colder, what keeps molecular clouds from collapsing in a free-fall time ($\tau \sim 1.3 \times 10^5 \text{ yr} (M_{cld}/10^5 M_{\odot})^{1/2} (R_{cld}/6 \text{ pc})^{3/2}$)? Molecular clouds appear to have lifetimes about 200 times longer than that.

The answer is that molecular clouds are supported mainly by magnetic pressure and/or turbulence – processes that mimic the effects of a gas temperature hundreds of times hotter than the actual thermal temperature. In many molecular clouds the MHD turbulence appears to diminish toward the center. One-third of molecular clouds appear to be collapsing, presumably in the process creating molecular cloud “cores” that will eventually form stars. One-third are static and apparently inactive currently. And one-third are expanding, the latter possibly from radiation pressure from newly-formed stars in their centers [421]. It is estimated that the total lifetime of a giant molecular cloud, from when it first produces detectable CO emission until the gas is dispersed by radiation pressure from its newly-formed stars, is on the order of $27 \pm 12 \text{ Myr}$ [422].

Atomic Hydrogen Clumps and a Tentative Universal Mass–Radius Relation.

Less is known about self-gravitating clumps of un-ionized atomic hydrogen (H I). In particular, there is no observed mass–radius relation of which this author is aware. However, we have seen that stellar systems and molecular clouds follow an $R \propto M^{1/2}$ relation, and stars more massive than $1.7 M_{\odot}$ also follow a similar scaling [423]. Shardha Jogee and her colleagues discovered H I clumps in NGC 2782, with the largest being $\sim 10^8 M_{\odot}$ in mass and ~ 3 arcseconds (or $\sim 500 \text{ pc}$) in apparent diameter. If we assume an $R \propto M^{1/2}$ relation for H I clumps, this gives a radius of 80 pc for a clump of $10^7 M_{\odot}$ – only 33% larger than the radius of a similar mass star cluster or molecular cloud from equations (11.19) and (11.30). Therefore, for the analysis below, we shall assume that *all* three types of objects (clusters, clouds, and clumps) have the same universal mass–radius relation

$$R_{cl} = 60 \text{ pc} \left(\frac{M_{cl}}{10^7 M_{\odot}} \right)^{1/2} \quad (11.31)$$

This is a fairly remarkable result and may not be a complete coincidence. Again, the internal temperatures implied by this relation for atomic gas clumps are in the thousands of kelvins, but the clumps have $T < 100 \text{ K}$. So, for H I clumps to be

gravitationally bound and stable, they also must be supported internally by magnetic or turbulent pressure.

The Accretion Rate of Infalling Clumps, Clouds, and Clusters when they Lose Mass

Anticipating that some atomic gas clumps, molecular clouds, or star clusters may lose some of their material through tidal stripping as they spiral toward the black hole, we will briefly outline the simple accretion physics of mass-losing objects. In particular, in a steady state, how does the mass accretion rate of these objects vary with radius r ?

We can calculate this function if we know how the mass of a typical object changes ($M(r)$) as they all sink inward and if we assume that the total *number* of objects remains constant. The conservation of object number is similar to that for the conservation of mass (equation (5.46)). In a spherically symmetric galaxy the conservation of object number is

$$\frac{dn}{dt} = -\frac{n}{r^2} \frac{d}{dr} (r^2 V_r)$$

In a steady state we can replace dn/dt with the advective derivative $V_r dn/dr$ and rearrange the equation to read $d \ln n/dr = -d \ln(r^2 V_r)/dr$ or

$$n \propto \frac{1}{r^2 V_r} \quad (11.32)$$

Now, the mass accretion rate is given by

$$\dot{M} = -4\pi r^2 n(r) M(r) V_r$$

where $\rho(r) = n(r) M(r)$ is the local mass density of the accreting objects. So, substituting proportionality (11.32) into the above equation, we find that the factor of $r^2 V_r$ falls out, and we now have, simply,

$$\dot{M}(r) \propto M(r) \quad (11.33)$$

Therefore, when the mass of the objects does not change with radius (*i.e.*, $M(r) = \text{constant}$), we of course recover the mass conservation law ($\dot{M}(r) = \text{constant}$). However, when $M(r)$ varies as, say, a power in r ($M(r) \propto r^m$), we find that the accretion rate also varies with that same power ($\dot{M}(r) \propto r^m$). So, knowing how object mass varies with r tells us not only how object radius varies with r (*i.e.*, $R(r) \propto r^{m/2}$ for an $R \propto M^{1/2}$ mass–radius relation), but also how the effective mass accretion rate of those objects varies with r .

The Physical Environment in which Infalling Clumps, Clouds, and Clusters Evolve

Finally, before we discuss the actual evolution of objects that infall into the galactic nucleus, we also must examine the environment and physical processes that affect the evolution of these objects.

The Galactic Environment and Clump/Cloud Distribution. As these objects spiral inward by the dynamical friction process, they will be subject potentially to both tidal and collisional forces in the central spheroidal bulge of the spiral, S0, or moderate-sized elliptical galaxy. We begin our discussion of these processes by first noting that, in the galactic nucleus (<1 kpc), the gravitational potential of the central bulge will be dominated by that of the stars and gas there, not by the small amount of dark matter that inhabits this central region. We therefore define the parameter η_g to be the ratio of gas mass to stellar mass only in this bulge

$$M_g \equiv \eta_g M_{\text{bulge}} \quad (11.34)$$

While $0 < \eta_g < \infty$, we generally expect the galaxy to be dominated by the stellar mass, *i.e.*,

$$\eta_g \leq 1$$

unless the galaxy is in a gas-dominated LIRG or ULIRG phase.⁷ To keep the analysis simple, we will assume that η_g is uniform throughout the outer regions of the bulge where the gas clumps form. Therefore, the average local density of clumps must be

$$\begin{aligned} n_{\text{clmp}} &= \eta_g \frac{\sigma_V^2}{2\pi G M_{\text{clmp}}} \frac{1}{r^2} \\ &= 4.4 \times 10^{-8} \text{ pc}^{-3} \eta_g \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/2} \left(\frac{M_{\text{clmp}}}{10^7 M_\odot} \right)^{-1} \left(\frac{r}{1 \text{ kpc}} \right)^{-2} \end{aligned} \quad (11.35)$$

(*cf.* equation (11.10)). Note that, for this simple analysis, we assume that all clumps in the bulge at a given radius r have the same mass $M_{\text{clmp}}(r)$, which itself may be a function of r .

⁷ Note that we now *specifically* define M_{bulge} to be the mass of the stars only, which is consistent with its usage earlier in this chapter. Note also that formally η_g is not the same as the mass *fraction* parameter which is often expressed as μ_g or f_g

$$f_g \equiv \frac{M_g}{M_g + M_{\text{bulge}}} = \frac{\eta_g}{\eta_g + 1}$$

and is *always* less than unity. However, since we are assuming that the bulge is dominated by the stellar mass, we have $f_g \approx \eta_g$; so the analysis here is *not* valid for $\eta_g \gg 1$.

Clump Collisions vs. Tidal Disruption. Since a single gas clump has a radius given by equation (11.31), or one clump in a box of volume $(2 R_{\text{clmp0}})^3 = 1.7 \times 10^6 \text{ pc}^3 (M_{\text{clmp0}}/10^7 M_\odot)^{3/2}$, clumps will begin to collide at a distance from the galactic center where equation (11.35) equals this number density, or

$$r_{C,\text{clmp}} = 487 \text{ pc } \eta_g^{1/2} \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/4} \left(\frac{M_{\text{clmp0}}}{10^7 M_\odot} \right)^{1/4} \quad (11.36)$$

where the subscript C signifies the collisional process and the subscript clmp0 indicates the initial properties of the H I clumps before tidal stripping begins. We now want to compare this process with the one in which clumps are torn apart by the tidal forces of the bulge mass *interior* to the radius r . To calculate the tidal radius, we use equation (11.25) with the following replacements: $M_\bullet \rightarrow M_{\text{bulge}}(r_T)$, $R_\star \rightarrow R_{\text{clmp0}}$, $M_\star \rightarrow M_{\text{clmp0}}$, and we use equation (10.7) for $M_{\text{bulge}}(r)$. The result is

$$r_{T,\text{clmp}} = 344 \text{ pc } \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/4} \left(\frac{M_{\text{clmp0}}}{10^7 M_\odot} \right)^{1/4} \quad (11.37)$$

Somewhat surprisingly, the collisional and tidal disruption radii have the same mass scalings. So, normally, when $\eta_g \lesssim 0.5$ we have $r_{T,\text{clmp}} > r_{C,\text{clmp}}$, and *the clumps will be tidally disrupted a few hundred parsecs from the black hole, before they begin to collide*. However, if $\eta_g > 0.5$, as might be the case in a LIRG, ULIRG or other star-forming galaxy, the clumps will begin colliding at $r \approx r_{C,\text{clmp}}$ – before they are tidally disrupted. Such collisions should promote a significant amount of star formation in gas-rich galaxies a half a kiloparsec or more from the center.

As stated above, we will concentrate here on the $\eta_g < 1$ situation – when the clumps are tidally disrupted before they begin to collide with each other.

The Dynamical Friction Process. Dynamical friction is the same process used in Section 10.4.2 to segregate heavier stars toward the center of a compact star cluster. Here we use the same equation (10.11), with the following substitutions: $M_\star \rightarrow M_{\text{clmp0}}$ and $M_{\text{clst}} \rightarrow M_{\text{bulge}}(r) = M_{\text{SIS}}(r)$.

The most difficult choice here is a value for $\ln A$, which could make a factor of 2 or more difference in the estimated dynamical friction time. A is a measure of the number of times an object (like a clump) interacts with pieces of the bulge that are similar in size to the clump. A liberal estimate might be $A \approx M_{\text{bulge}}(r)/M_{\text{clmp0}}$, but a clump on an inspiraling orbit might see only

$$A \approx \frac{R_{\text{bulge}}(r)}{R_{\text{clmp0}}}$$

We will choose the latter more conservative estimate here, with a rough estimate of $\ln A \approx \ln(344 \text{ pc}/60 \text{ pc}) \approx 2$.

These substitutions then give us a dynamical friction time for the inspiral of H I clumps of⁸

$$\tau_{\text{df}} = \tau_{\text{seg}} = 80 \text{ Myr} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{1/4} \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{-1} \left(\frac{\ln A}{2} \right)^{-1} \left(\frac{r}{600 \text{ pc}} \right)^2 \quad (11.38)$$

Note that equations (11.36) to (11.38) take into account the decrease of bulge mass with decreasing radius r ($M_{\text{bulge}} \propto r$).

11.1.4.3 STEP 2: Dynamical Friction Accretion of Atomic Gas Clumps to a Few Hundred Parsecs

We now are in a position to describe the second step in the gas accretion process described at the beginning of this subsection – breakup of the bar-fueled gas into gas clouds and then accretion of those clouds toward the black hole by the process of dynamical friction. Refer to Fig. 11.3 for a graphical depiction of this second step.

Evolution of the First Wave of Accreting Gas Clumps

The initial evolution of the ensemble of H I gas clumps is rather disappointing, as far as black hole fueling is concerned. While Jogee *et al.* considered clumps of $10^8 M_{\odot}$, we will be more conservative again and allow for the possibility that the large clumps they observed may not be typical, or may be composed of $10^7 M_{\odot}$ pieces. Equation (11.38) predicts that it will take the better part of 100 Myr for these clumps to inspiral from 600 pc (where gas is injected by the dynamical bar process) down to ~ 340 pc, where they then begin to be tidally disrupted. The problem is not with the long time scale; the clumps are supported by magnetic fields and turbulence and in no danger of collapsing and forming stars. Instead, the problem is with their total disruption.

Once inside $r_{\text{T,clmp}}$ the clump will continue to have its outer layers stripped by the bulge's tidal forces. Survival of each clump depends on how it responds internally to this stripping, *i.e.*, on how its radius changes as it loses mass. If, for example, the central core of the clump remains tightly bound, then it may survive as it sinks toward the black hole, even though its outer layers may be heavily stripped. However, for a gas clump it is more likely that, by losing its outer layers, a variety of processes (cosmic rays, UV radiation) will alter quickly its internal energy to be like any other interstellar H I clump of a similar mass. That is, we expect a gas clump to evolve along the mass–radius relation as it loses mass

⁸ This time scale is about a factor of 2 shorter than that of Jogee *et al.* [419], chiefly because of the factor of $\ln A$.

$$R_{\text{clmp}}(r) = 60 \text{ pc} \left(\frac{M_{\text{clmp}}(r)}{10^7 M_{\odot}} \right)^{1/2} \quad (11.39)$$

Assuming that the clump is continually stripped as it spirals inward (*i.e.*, $r_{\text{T,clmp}}(r) = r$), we can recompute equation (11.37), include the process of continuous stripping (equation (11.39)) for radii interior to $r_{\text{T,clmp}}$ to arrive at the simple expressions

$$\begin{aligned} M_{\text{clmp}}(r) &= M_{\text{clmp}}(r_{\text{T,clmp}}) \left(\frac{r}{r_{\text{T,clmp}}} \right)^4 \\ &= 10^7 M_{\odot} \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right) \left(\frac{r}{r_{\text{T,clmp}}} \right)^4 \end{aligned} \quad (11.40)$$

$$\begin{aligned} R_{\text{clmp}}(r) &= R_{\text{clmp}}(r_{\text{T,clmp}}) \left(\frac{r}{r_{\text{T,clmp}}} \right)^2 \\ &= 60 \text{ pc} \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{1/2} \left(\frac{r}{r_{\text{T,clmp}}} \right)^2 \end{aligned} \quad (11.41)$$

for radii interior to $r_{\text{T,clmp}}$. So, for example, a $10^7 M_{\odot}$ clump would be tidally stripped to only a few solar masses by the time it has sunk to $r \sim r_{\text{h}}$.

A worse problem, however, is that the dynamical friction time for such a clump to spiral inward *increases* as the inverse square of the decreasing radius r

$$\tau_{\text{df,clmp}}(r) = \tau_{\text{df}}(r_{\text{T,clmp}}) \left(\frac{r}{r_{\text{T,clmp}}} \right)^{-2} \quad (11.42)$$

for $r < r_{\text{T,clmp}}$ and where the dynamical friction time at the clump tidal radius (combination of equations (11.37) and (11.38)) is

$$\tau_{\text{df}}(r_{\text{T,clmp}}) = 26 \text{ Myr} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{3/4} \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{-1/2} \left(\frac{\ln A}{2} \right)^{-1}$$

However, the time for the continually-stripped clumps to spiral inward is not 26 Myr; it is essentially infinite. Equation (11.42) tells us that at 10 pc, for example, not only are the residual clump cores down to a few solar masses (at best), but the time for those residuals to sink further inward from 10 pc is 31 Gyr – longer than the age of the universe!

So, *dynamical friction of H I gas clumps alone cannot feed a central black hole*. On the face of it, the first two suggestions for solving the bar accretion problem do not seem to work.

The Buildup of a Dynamically Important Gas Ring at the Clump Tidal Radius

Gas, however, continues to rain down to $r_{\text{T,clmp}}$ at the rate

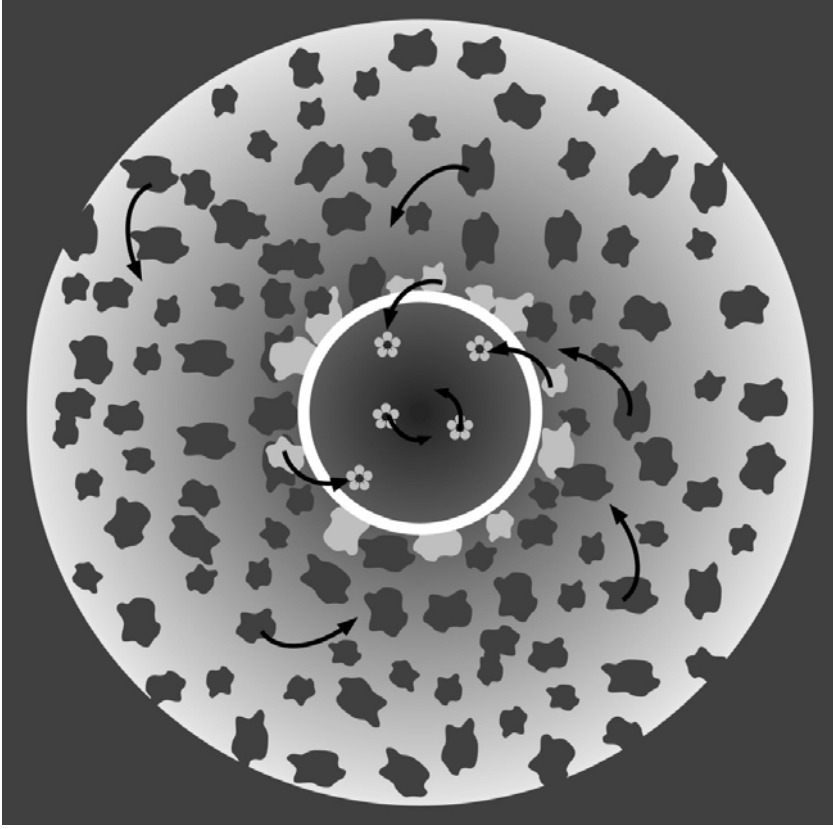


Fig. 11.3: STEP 2: Dynamical friction infall of H I clumps, their tidal disruption, and the creation of a star-formation ring. Inside 1 kpc H I clumps (dark gray blobs) inspiral by dynamical friction with the stellar background (light-to-medium gray). Tidal disruption of these clumps halts the infall process, forming a dense H I ring. Collision of new clumps with the H I ring triggers star formation, similar to a spiral shock, forming clusters of stars that inspiral inside that ring.

$$\begin{aligned}
 \dot{M}(r_{T,\text{clmp}}) &= 4\pi r_{T,\text{clmp}}^2 n_{\text{clmp}}(r_{T,\text{clmp}}) M_{\text{clmp}0} V_{\text{df}}(r_{T,\text{clmp}}) \\
 &= 73 M_{\odot} \text{ yr}^{-1} \eta_g \left(\frac{M_{\text{clmp}0}}{10^7 M_{\odot}} \right)^{3/4} \left(\frac{\ln \Lambda}{2} \right)
 \end{aligned} \tag{11.43}$$

independent of the bulge mass. Here $V_{\text{df}} \approx r/\tau_{\text{df}}$ is the inward drift velocity of a clump. Of course, before we conclude that all this gas collects at $r_{T,\text{clmp}}$, we need to check to see how much gas is drained inward by dynamical friction on the clumps.

The answer is, not much. If we assume that, as they drift inward, the *number* of clumps is conserved (but not their mass), then we find that when a steady state builds up ($r^2 V_{\text{df}} dn_{\text{clmp}}/dr = -n_{\text{clmp}} d(r^2 V_{\text{df}})/dr$), the number density of clumps actually increases rapidly inward ($n_{\text{clmp}} \propto r^{-5}$), although the mass density contributed by the clumps increases much more slowly ($\rho_{\text{clmp}} \propto r^{-1}$). The

important quantity to look at is the accretion rate itself inside $r_{T, \text{clmp}}$

$$\begin{aligned}\dot{M}(r) &= 4\pi r^2 \rho_{\text{clmp}}(r) V_{\text{df}}(r) \\ &= \dot{M}(r_{T, \text{clmp}}) \left(\frac{r}{r_{T, \text{clmp}}} \right)^4\end{aligned}\quad (11.44)$$

So, only half-way inside $r_{T, \text{clmp}}$ the accretion rate of H I gas has dropped to less than 10% of that at $r_{T, \text{clmp}}$. Interstellar matter, therefore, will build up rather quickly only a few hundred parsecs from the black hole.

At some point the interstellar density of built-up gas will become so high that it will affect the inspiraling clumps, compressing them via dynamical pressure ($\rho_{\text{gas}} \sigma_V^2 > p_{\text{clmp}}$, where p_{clmp} is the effective internal pressure in the clump due to magnetic fields, turbulence, *etc.*). An *underestimate* of the rate at which gas builds up just inside $r_{T, \text{clmp}}$ is

$$\dot{\rho}_{\text{gas}} > \frac{\rho_{\text{clmp}}(r_{T, \text{clmp}})}{\tau_{\text{df}}}$$

So, an *overestimate* of the time it takes to build up enough clump debris to begin compressing newly entering clumps is

$$\begin{aligned}\tau_{\text{gas buildup}} &= \frac{\rho_{\text{gas}}}{\dot{\rho}_{\text{gas}}} < \frac{p_{\text{clmp}}}{\rho_{\text{clmp}}(r_{T, \text{clmp}}) \sigma_V^2} \tau_{\text{df}} \\ &= 0.85 \text{ Myr } \eta_g^{-1} \left(\frac{M_{\text{clmp0}}}{10^7 M_\odot} \right)^{1/4} \left(\frac{\ln A}{2} \right)^{-1}\end{aligned}\quad (11.45)$$

Even for small η_g , the time for gas to build up to a dynamically important level still is only a few million years, significantly shorter than the dynamical friction time there τ_{df} . Therefore, when there is significant gas accretion by the bar accretion and clump dynamical friction processes, a shock-like gas barrier will quickly form at a few hundred parsecs from the black hole. This will begin to compress the H I clumps in a “ring shock” that is comparable to a galactic spiral shock. The gas then will emerge inside this shock, no longer as condensed clumps of magnetically supported H I, but as collapsing *molecular* clouds in the process of forming new stars.

Molecular Cloud and Star Formation at $r_{T, \text{clmp}}$: The “Star-Formation Ring”

The presence of a ring shock only a few hundred parsecs from the center of a galactic nucleus should be a common phenomenon in gas-rich galaxies. This star-forming ring brings several advantages to the theory:

- It explains the presence of a star-formation ring observed in many active galaxies at such a radius.

- The ring shock will create new compact star clusters, which then can continue the dynamical friction inspiral process and which are much less susceptible to tidal effects. The interstellar gas, therefore, can continue to accrete toward the black hole again, *but now locked up in newly-formed star clusters*.

Tidal disruption of the H I clumps, therefore, actually explains why a star-formation ring should exist and why it is associated with active galaxies: it promotes the formation of new, heavy star clusters whose dynamical friction inspiral can feed the central black hole. The dynamical friction accretion concept, therefore, is saved by the subsequent molecular cloud and star formation.

As we shall see, once an H I clump has been converted into a molecular cloud, the process of star formation is so fast (a few million years) that it can be considered instantaneous here. That is, the transition of H I clump to molecular cloud and then to young globular-like cluster all occurs within the vicinity of $r_{T,clmp}$. We therefore shall ignore the short, though interesting, molecular cloud and star formation phases.

11.1.4.4 STEP 3: Dynamical Friction Accretion of Newly-Formed Star Clusters Deep in a Galactic Nucleus

We now examine the third step of interstellar gas accretion: inspiral of that gas in the form of young star clusters and eventual tidal disruption of those stars by the central black hole.⁹ Refer to Fig. 11.4 for a graphical depiction of this third and final step.

The most difficult part of the remaining analysis is determining the mass and number of star clusters formed at $r_{T,clmp}$. As a conservative estimate, and for illustrative purposes, we shall take the number of star clusters to be the same as the number of currently inspiraling clumps, but their mass will be only $\sim 30\%$ of the clump mass (*i.e.*, $M_{clst0} \sim 3 \times 10^6 M_{\odot}$). So, we assume a 70% mass loss from the clumps due to unprocessed H I gas and radiation-pressure-ejected molecular gas. Stars lost from the molecular cloud/star cluster complexes due to tidal stripping are treated below.

Inspiral of Tidally Unstripped Star Clusters

The new clusters would correspond to young globular cluster-like objects and also would be subject to tidal stripping at some point in the galactic center. Since such clusters can be modeled as singular isothermal spheres, their radii will be

$$R_{clst0} = 32 \text{ pc} \left(\frac{M_{clst0}}{3 \times 10^6 M_{\odot}} \right)^{1/2}$$

⁹ Note that it is this third step, and only this one, that relates directly to the work of Ebisuzaki *et al.* [401]. My work reported herein on the entire process (STEPS 1–3) was done much later than the Ebisuzaki *et al.* work, but independently of it.

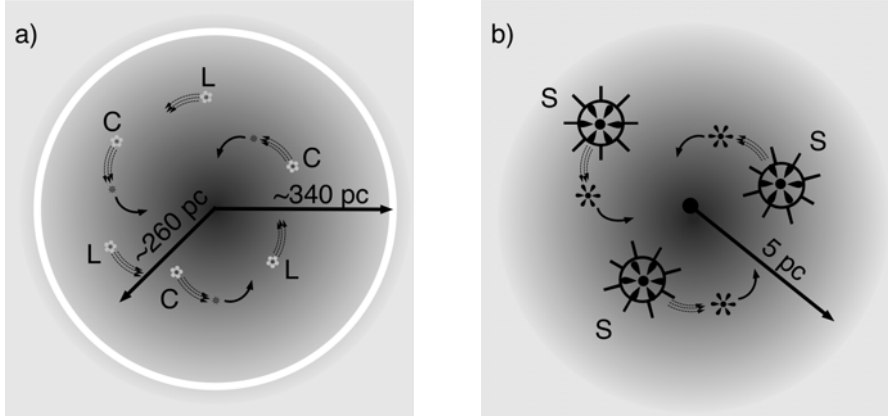


Fig. 11.4: STEP 3: Inspiral of star clusters via dynamical friction. (a) Inside the star-formation ring star clusters inspiral initially without much tidal stripping. Soon, however, the clusters begin to be stripped of stars, which become new members of the bulge. Loose clusters (L) are fully disrupted, contributing nothing to the further accretion of stars toward the nucleus. Compact clusters (C), however, continue to spiral inward, retaining some fraction of their original stellar mass. (b) At the radius of influence of the central SMBH ($r_h \sim 5$ pc), uncollapsed (“core”) clusters (not shown) are fully disrupted, even if they were compact initially. Only those with an inner cusp or singular isothermal sphere (S) structure survive inside r_h . Eventually they are also stripped of most of their stars – still at a radius much larger where a single star would be tidally disrupted by the SMBH itself. However, the dynamical friction interactions of the cluster with the cusp stars should serve to repopulate the loss cone during the inspiral.

so $\ln A$ will be a little larger, ~ 3 instead of ~ 2 . Because this is the same mass–radius relation as for clumps and molecular clouds¹⁰, the tidal radius for such a cluster in the galactic bulge still is given by equation (11.37)

$$r_{T,clst} = 257 \text{ pc} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{1/4} \left(\frac{M_{clst0}}{3 \times 10^6 M_{\odot}} \right)^{1/4} \quad (11.46)$$

so the newly-formed cluster will experience an initial phase of dynamical friction from $r_{T,clmp}$ down to $r_{T,clst}$, *without tidal stripping* (chiefly because it has a smaller mass and radius than its parent). The dynamical friction time for this initial phase is moderately long

¹⁰ The close correspondence between the mass–radius relation for molecular clouds and for star clusters – in not only scaling but also absolute value – may have a physical origin, at least for smaller star cluster masses ($< 10^8 M_{\odot}$). A new star cluster may inherit its velocity dispersion from the MHD turbulence that existed in its parent molecular cloud, thus preserving the size and structure of the object even though it undergoes a radical phase change.

$$\tau_{\text{df,clst}}(r_{\text{T,clmp}}) = 56 \text{ Myr} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{3/4} \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{1/2} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{-1} \left(\frac{\ln A}{3} \right)^{-1} \quad (11.47)$$

Note that this time scale will dominate the remaining inspiral process, so it represents the time between the formation of the new star clusters and the tidal breakup of those stars by the black hole to produce accretion luminosity. The length of this dynamical friction time scale is significant for a number of reasons:

- It is significantly longer than the estimated time for a molecular cloud to be born, form stars, and then expel its remaining molecular gas (12–20 Myr). It also is longer than the lifetimes of the massive, O and early B stars that would dominate the cluster's initial luminosity (3–30 Myr; see Section 10.1.2). Therefore, the hot blue stars in all such clusters will have died in supernova explosions before the clusters sink appreciably inside $r_{\text{T,clmp}}$. *The star-formation ring, therefore, will appear as a ring not because the stars are in static orbits, but because they are in clusters that are spiraling inward on a time scale significantly longer than the O/B star lifetime.*
- 50 Myr or so has been suggested as the lag time between a *nuclear* starburst in a gas-rich galaxy and the onset of subsequent AGN activity [424]. In the present model this lag is equal to the time it takes newly-formed star clusters to sink from the ring shock radius to the vicinity of the black hole. In making this comparison, it is important to restrict starburst age measurements to the nuclear ring itself and not to the entire star-forming galaxy.

Because there is no tidal stripping yet of the cluster, the aggregate accretion rate of these clusters does not change with radius between $r_{\text{T,clmp}}$ and $r_{\text{T,clst}}$

$$\dot{M}_{\text{clst}} = 23 M_{\odot} \text{ yr}^{-1} \eta_g \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{-1/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right) \left(\frac{\ln A}{3} \right) \quad (11.48)$$

Inspiral of Tidally Stripped Star Clusters Inside $r_{\text{T,clst}}$

Inside its tidal radius (equation (11.46)) the cluster will begin to be tidally stripped as well. However, the internal velocity dispersion of its stars likely will be unaffected by this process on much less than the cluster two-body relaxation time (~ 120 Gyr from equation (10.10)!). And, unlike H I gas clumps or molecular clouds, σ_v also will be unaffected by cosmic rays, UV light, and other galactic center effects, save cluster–cluster collisions themselves. On the other hand, the cluster *will* dynamically relax in a few dynamical times and, so, reacquire an isothermal structure, albeit with a different mass–radius normalization. Therefore, a cluster being stripped of its outermost stars, but maintaining the same σ_{clst} , will evolve along a linear mass–radius relation

$$R_{\text{clst}}(r) = \frac{M_{\text{clst}}(r)}{M_{\text{clst0}}} R_{\text{clst0}} \quad (11.49)$$

rather than the square root curve that would occur if the cluster were to relax to a new *thermal* configuration as its mass changes (*cf.* equation (11.39)). Thus, for the same cluster mass $M_{\text{clst}}(r)$ (which will be less than M_{clst0}) the tidally stripped cluster will have a more compact structure. If we once again recalculate equation (11.46) for each radius r , set $r_{\text{T,clst}}(r) = r$, and use proportionality (11.33), we arrive at the following evolution equations for the cluster properties and for the aggregate mass accretion rate:

$$\begin{aligned} M_{\text{clst}}(r) &= M_{\text{clst0}} \left(\frac{r}{r_{\text{T,clst}}} \right) \\ R_{\text{clst}}(r) &= R_{\text{clst0}} \left(\frac{r}{r_{\text{T,clst}}} \right) \\ \dot{M}_{\text{clst}}(r) &= \dot{M}_{\text{clst0}} \left(\frac{r}{r_{\text{T,clst}}} \right) \end{aligned} \quad (11.50)$$

That is, the cluster loses mass and shrinks in radius, and the accretion rate decreases, only linearly with decreasing distance from the black hole. In addition, the dynamical friction time decreases linearly with r as well

$$\tau_{\text{df,clst}}(r) = \tau_{\text{df}}(r_{\text{T,clst}}) \left(\frac{r}{r_{\text{T,clst}}} \right) \quad (11.51)$$

where

$$\tau_{\text{df}}(r_{\text{T,clst}}) = 31 \text{ Myr} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{-1/2} \left(\frac{\ln A}{3} \right)^{-1}$$

Unlike a tidally stripped gas clump, a significant portion of each compact star cluster survives even down to the black hole radius of influence (equation (11.20))

$$r_{\text{h}} = 4.8 \text{ pc} \left(\frac{M_{\bullet}}{1.3 \times 10^7 M_{\odot}} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{-1/2}$$

or

$$\frac{r_{\text{h}}}{r_{\text{T,clst}}} = 0.019 \left(\frac{M_{\bullet}}{1.3 \times 10^7 M_{\odot}} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{-3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{-1/4} \quad (11.52)$$

The properties of the stripped cluster remnants as they enter r_{h} are, then,

$$\begin{aligned}
M_{\text{clst}}(r_h) &= 5.9 \times 10^4 M_\odot \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{-3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_\odot} \right)^{3/4} \\
R_{\text{clst}}(r_h) &= 0.60 \text{ pc} \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{-3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_\odot} \right)^{1/4} \\
\tau_{\text{df,clst}}(r_h) &= 1.1 \times 10^4 \text{ yr} \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{1/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_\odot} \right)^{-1} \\
&\quad \times \left(\frac{\ln A}{3} \right)^{-1} \\
\dot{M}_{\text{clst}}(r_h) &= 0.43 M_\odot \text{ yr}^{-1} \eta_g \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right) \left(\frac{M_{\text{bulge}}}{10^{10} M_\odot} \right)^{-3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_\odot} \right)^{3/4} \\
&\quad \times \left(\frac{\ln A}{3} \right) \left(\frac{M_{\text{clmp0}}}{10^7 M_\odot} \right)^{-1/4}
\end{aligned} \tag{11.53}$$

$\dot{M}_{\text{clst}}(r_h)$ may be an underestimate of the amount of stellar matter that this process brings to the SMBH. In addition to continually adding stars inside r_h , there will be field and cluster stars scattered into the loss cone outside r_h by the dynamical friction process and by tidal stripping. We have not taken these processes into account here.

Inspiral of Star Cluster Remnants Inside r_h

Compact star clusters, therefore, appear to be able to venture within a few parsecs of a central supermassive black hole. However, a cluster's days are also numbered: it will continue to spiral toward the black hole inside r_h but will not survive long enough to deliver its stars to within a *stellar* tidal radius (equation (11.25), only 10^{-5} – 10^{-4} pc from the black hole), where those stars could be disrupted and converted into gas immediately.

The details of the final evolution of the star cluster depend on its central density structure. Does it have a core or cusp (collapsed) structure? Does it have a central *intermediate* mass black hole? Below we consider clusters with, and without, collapsed cores, but we will assume that both types have a central IMBH of mass

$$M_{\bullet, \text{clst}} = 4100 M_\odot \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_\odot} \right)$$

Clusters with Collapsed Cores. In this case the cluster core remains a singular isothermal sphere down to the IMBH, so the cluster mass–radius evolution continues to be $M_{\text{clst}} \propto R_{\text{clst}}$. However, the central gravitational potential inside r_h is dominated by the SMBH, not the bulge, so the cluster tidal radius is now given by equation (11.25)

$$r_T^{\text{clst}} = 1.6 \times 10^{13} \text{ cm} \left(\frac{R_{\text{clst}}(r)}{R_\odot} \right) \left(\frac{M_{\text{clst}}(r)}{M_\odot} \right)^{-1/3} \left(\frac{M_\bullet}{1.3 \times 10^7 M_\odot} \right)^{1/3} \tag{11.54}$$

As with the derivation of equations (11.50), we still assume that σ_{clst} will remain the same as the cluster is stripped, and we will set $r_{\text{T,clst}}(r) = r$. But solving equation (11.54) for $R_{\text{clst}}(r)$ now produces faster cluster stripping with decreasing radius r :

$$\dot{M}_{\text{clst}}(r) \propto M_{\text{clst}}(r) \propto R_{\text{clst}}(r) \propto r^{3/2}$$

The cluster eventually will be stripped down to its own IMBH's radius of influence

$$R_{\text{h}} = 0.085 \text{ pc} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{1/2}$$

with only about twice the IMBH's mass in stars remaining (equation (11.21)) or only 14% of what it had when it entered r_{h} . This happens at a “final” distance from the central SMBH of only

$$r_{\text{f}} = 1.3 \text{ pc} \left(\frac{M_{\text{bulge}}}{10^{10} M_{\odot}} \right)^{3/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{-1/4}$$

So, if the cluster has a singular isothermal sphere structure, its central compact region will survive down to about $0.27 r_{\text{h}}$. But, the end result will be that the young cluster will donate most of its remaining stellar mass inside what is essentially the black hole Bondi radius. If $\eta_{\text{g}} \sim 0.3\text{--}0.4$, this tidal disruption of newborn star cluster remnants inside r_{h} could increase by a factor of 10–1000 the amount of accretion expected from tidal disruption of field stars alone, if this process results in repopulating the loss cone. (See [Table 11.1](#).) This is enough to explain most active objects except for the mightiest of quasars and very bright Eddington-accreting AGN.

Clusters with Uncollapsed Core Structure. If the cluster is not collapsed, it will have a *flat* density distribution inside its core radius of

$$R_{\text{c}} = 0.85 \text{ pc} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)$$

This is $\gtrsim R_{\text{clst}}(r_{\text{h}})$, so core clusters that have not collapsed by the time they reach r_{h} will be *uniform density* spheres inside the SMBH's Bondi radius. Such a sphere evolves as $M_{\text{clst}} \propto R_{\text{clst}}^3$, so equation (11.54) in this case gives

$$\dot{M}_{\text{clst}}(r) \propto M_{\text{clst}}(r) \propto R_{\text{clst}}(r)^3 \propto r^{\infty}$$

So, if the original star cluster has an uncollapsed core structure, that cluster will be stripped of essentially all its stars at r_{h} ($M_{\text{clst}}(r_{\text{h}})$ in equation (11.53)), where the tidal force of the black hole itself becomes dominant.

Summary and Comments on the Dynamical Friction Accretion Process

The mechanism described above accretes interstellar gas onto a central black hole by converting it to young, compact star clusters that can efficiently, and quickly, spiral into the galactic center via dynamical friction, even though they are stripped of some stars along the way. This conversion of gas to stars is needed, and predicted, because the dynamical friction process for H I clumps stalls a few hundred parsecs from the black hole where the clumps are completely tidally disrupted by the galactic gravitational potential. The gaseous debris collects there, eventually forming a star-forming ring shock. This creates compact star clusters that, though partially tidally stripped, can spiral inward into the sphere of influence of the black hole. Actual feeding of the hole is by the same process as discussed in the previous subsection: tidal disruption of these new stars.

A particular advantage of this proposed mechanism is that it naturally explains the presence of a star-formation ring a few hundred parsecs from the black hole and even predicts the time lag of ~ 50 Myr between the onset of this star formation and the beginning of activity in the AGN. The lag in this model is due to the time it takes new clusters to spiral in by dynamical friction to the vicinity of the black hole.

The scenario also has some disadvantages, however. Because tidal disruption of stars cannot efficiently fuel black holes with masses above $\sim 10^8 M_\odot$ (most stars are swallowed whole and not disrupted), neither this nor the classical disruption feeding mechanism of the previous subsection can fuel the bright quasars. Yet another, even more powerful, gas fueling mechanism must exist. In addition, this mechanism is only being proposed here for the first time; it has not been vetted in the open literature yet. After peer review and study by others it might change or even be shown to not be viable, for yet unforeseen reasons. For example, one point that bears further investigation is, how reasonable are the assumed typical masses of H I clumps and compact clusters? Also, we have assumed that all new stellar mass entering r_h will be accreted in a short period of time. How realistic is that assumption? Will the process of dynamical friction aid or hinder the scattering of these new stars into the loss cone? What happens to the stars stripped from the cluster, both outside and inside r_h ? Can these also fuel the black hole and further increase the predicted accretion rate?

Many of these questions will need to be investigated numerically with codes that can evolve star clusters spiraling inward in a large galactic nucleus – a capability that does not exist yet.

11.1.5 Accretion in Stellar-mass Binary Systems

In binary star systems black holes and neutron stars accrete by the process of mass transfer, in which only a small portion of the companion star is dumped onto the compact object during many orbits. At first, these systems would seem to be quite different physically from accreting black holes in the interstellar medium or in the

centers of AGN. However, analysis of HMXBs and LMXBs actually uses physics that is very similar to that used for isolated black holes and AGN. The steady HMXB and LMXB systems accrete via Roche lobe overflow, which acts similarly to a stellar tidal disruption, although usually on a time scale much longer than the dynamical one. And the steady HMXBs that accrete a portion of a companion O/B stellar wind behave similarly to the moving black hole we treated at the beginning of this chapter.

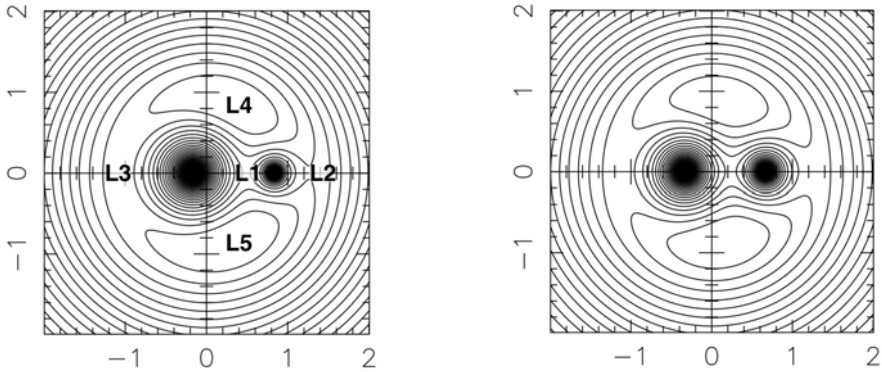


Fig. 11.5: Two examples of Roche equipotentials in the orbital (equatorial or X - Y) plane of binary star systems (contours of φ (eq. (11.55))). *Left:* $q \equiv M_{*2}/M_{*1} = 0.2$. *Right:* $q = 0.5$. Distances are in units of the binary separation, with the origin at the center of mass and primary and secondary stars at, respectively, positions $(-a_1, 0)$ and $(a_2, 0)$.

11.1.5.1 Gravity in Binary Systems

As shown in Chapter 5, the gravitational potential outside a *single* star of mass M_* in Newtonian gravity is $\psi = -G M_*/|\mathbf{r} - \mathbf{r}_0|$, where $|\mathbf{r} - \mathbf{r}_0|$ is the absolute value of the vector difference between the point at which we are evaluating the potential $\mathbf{r}$ and position of the center of the star $\mathbf{r}_0$. For two stars M_{*1} and M_{*2} in a circular orbit about each other, the *effective* potential is the more complicated Roche potential

$$\varphi = -\frac{G M_{*1}}{|\mathbf{r} - \mathbf{r}_1|} - \frac{G M_{*2}}{|\mathbf{r} - \mathbf{r}_2|} - \frac{1}{2} |\boldsymbol{\omega} \times \mathbf{r}|^2 \quad (11.55)$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the positions of the two stars' centers and $\boldsymbol{\omega}$ is the angular velocity vector of the binary system. As in Section 5.1.3 we will take the origin of this coordinate system to be at the binary's center-of-mass point $\mathbf{r}_0$, so that $|\boldsymbol{\omega} \times \mathbf{r}| = \omega r$ in the orbital plane. Note that this potential is evaluated in a frame that is co-moving with the rotating binary. That is, the binary remains stationary while the rest of the universe rotates in a direction opposite to the orbital motion. This co-moving frame is not an inertial system, so equation (11.55) includes the centrifugal pseudo-force per unit mass $-|\boldsymbol{\omega} \times \mathbf{r}|^2/2$.

Far from the binary system and in the equatorial plane, the potential surfaces are nested circles centered on the center of mass. On the scale of the binary system, on the other hand, the potential is determined by the gravitational field of both stars. And very close to either star it will be dominated only by that one star. Fig. 11.5 shows a contour plot of the Roche potential in two different binaries, along with the five principal Lagrangian points. L4 and L5 are the stable “Trojan” points: matter placed there will tend to remain in orbit at those locations. L1, L2, and L3 are unstable points: matter at the inner Lagrangian point L1 will tend to fall toward one of the stars; matter at L2 or L3 tends to fall toward the binary or be flung outward.

In plots like those in Fig. 11.5 there is one figure-eight shaped contour (not shown) that encloses both stars and passes through the L1 point. This is called the *Roche surface*, and each separate egg-shaped portion of it is called the *Roche lobe* for the star that it encloses. The Roche lobe represents the boundary between the region where the gravity of one of the binary star components dominates and the region where the gravitational effects of the double star system begin. If we define the mass ratio $q \equiv M_{\star 2}/M_{\star 1}$, then the radius of the Roche lobe for the star $M_{\star 1}$ is given by an approximation derived by Peter Eggleton of University of Cambridge, UK [425]

$$R_{\text{Roch}} = a \frac{0.49 q^{2/3}}{0.60 q^{2/3} + \ln(1 + q^{1/3})} \quad (11.56)$$

where a is the binary separation between the two stars. Equation (11.56) is good to 1% over the entire possible range $0 < q < \infty$.

11.1.5.2 Mass Transfer by Roche Lobe Overflow

For most binary systems the radius of the star or planet is much smaller than R_L . For example, for the earth–moon system with $q = 0.012$ and $a = 7.7 \times 10^{10}$ cm, the moon’s Roche radius is about 8.3×10^9 cm – 47 times larger than its equatorial radius. The moon will not begin to fill its Roche lobe until its orbit decays to close to the Roche limit of ~ 1.5 earth radii. Such a system, with neither component filling its Roche lobe, is called a *detached* binary system.

For some binary star systems, however, one or both of the stars may be as large as their Roche lobe sizes. This can happen in two ways: (1) when a star expands into a giant phase because of stellar evolution, or (2) when the binary evolves to a smaller separation, decreasing the sizes of the lobes around the stars. If one star fills its Roche lobe, the system is called *semi-detached*; when both fill their Roche lobes, it is said to be a *contact binary*.

Roche lobe overflow in an X-ray binary is a form of tidal disruption, with the black hole or neutron star slowly shredding the companion over a given time scale and the outer stellar layers being drawn toward the compact object through the inner Lagrangian point (see Fig. 11.5). The rate at which this occurs depends on how the companion star readjusts its structure as it loses mass to the compact object. If it

cannot maintain even hydrostatic (dynamical) equilibrium, then the star is disrupted on a dynamical time

$$\begin{aligned}\tau_{\text{dyn}}^* &= \left(\frac{r_T^3}{GM_\bullet} \right)^{1/2} = \left(\frac{R_\star^3}{GM_\star} \right)^{1/2} \\ &= 0.44 \text{ hr} \left(\frac{M_\star}{M_\odot} \right)^{-1/2} \left(\frac{R_\star}{R_\odot} \right)^{3/2}\end{aligned}\quad (11.57)$$

and the accretion rate is given by

$$\dot{M}_{\text{dyn}\bullet}^* \approx \frac{M_\star}{\tau_{\text{dyn}}^*} = 1.3 \times 10^{30} \text{ g s}^{-1} \left(\frac{M_\star}{M_\odot} \right)^{3/2} \left(\frac{R_\star}{R_\odot} \right)^{-3/2} \quad (11.58)$$

On the other hand, if, after transferring some material to the compact object, the companion *can* readjust to maintain hydrostatic equilibrium (which is generally the case in binary systems), but cannot readjust its temperature structure before the next transfer of matter takes place, then the mass transfer takes place on the *thermal time scale* (i.e., the time it takes to radiate the gravitational energy of the star)

$$\tau_{\text{th}\bullet}^* = \frac{GM_\star^2}{R_\star L_\star} = 3.1 \times 10^7 \text{ yr} \left(\frac{M_\star}{M_\odot} \right)^2 \left(\frac{R_\star}{R_\odot} \right)^{-1} \left(\frac{L_\star}{L_\odot} \right)^{-1} \quad (11.59)$$

However, if the companion star can readjust both its dynamical *and* thermal structure, then the mass transfer takes place on the *secular* or *nuclear time scale* (i.e., the time it takes to convert the star's core (we assume $f_{\text{core}} \approx 10\%$ of the entire star), into energy at a nuclear burning efficiency $\epsilon_{\text{nuc}} \approx 0.007$)

$$\begin{aligned}\tau_{\text{nuc}\bullet}^* &= \frac{\epsilon_{\text{nuc}} f_{\text{core}} M_\star c^2}{L_\star} \\ &= 1.0 \times 10^{10} \text{ yr} \left(\frac{M_\star}{M_\odot} \right) \left(\frac{L_\star}{L_\odot} \right)^{-1} \left(\frac{\epsilon_{\text{nuc}}}{0.007} \right) \left(\frac{f_{\text{core}}}{0.1} \right)\end{aligned}\quad (11.60)$$

In order to compare the accretion rates on these two time scales we will need the mass–luminosity and mass–radius relations for stars. The first is given approximately by

$$\frac{L_\star}{L_\odot} \approx \left(\frac{M_\star}{M_\odot} \right)^{3.7} \quad (11.61)$$

for stars between $0.43 M_\odot < M_\star < 46 M_\odot$ (and then $L_\star \approx L_{\text{Edd}}$ above that range). The mass–radius relation is, approximately,

$$\frac{R_\star}{R_\odot} \approx \left(\frac{M_\star}{M_\odot} \right)^{0.8} \quad (11.62)$$

Combining these with equations (11.59) and (11.60), we obtain two estimates for the mass transfer rate. For the thermal case the mass transfer rate onto the compact object can be rather high

$$\dot{M}_{\text{th}\bullet}^* \approx \frac{M_\star}{\tau_{\text{th}}^*} = 2.0 \times 10^{18} \text{ g s}^{-1} \left(\frac{M_\star}{M_\odot} \right)^{3.5} \quad (11.63)$$

near or exceeding the Eddington limit for 1–10 $M_\odot$ compact objects, while for the secular case the mass transfer rate is much slower

$$\dot{M}_{\text{nuc}\bullet}^* \approx 6.3 \times 10^{15} \text{ g s}^{-1} \left(\frac{M_\star}{M_\odot} \right)^{3.7} \left(\frac{\epsilon_{\text{nuc}}}{0.007} \right)^{-1} \left(\frac{f_{\text{core}}}{0.1} \right)^{-1} \quad (11.64)$$

only of order 0.005 $\dot{M}_{\text{Edd}\bullet}$ or less.

11.1.5.3 Wind Accretion in HMXBs

In cases where the binary is a high-mass one, mass transfer can take place even if the companion star does not fill its Roche lobe. Hot O and B stars drive strong stellar winds from their surfaces during all three wind stages of their lives (main sequence, luminous blue variable, and Wolf–Rayet star). The neutron star or black hole compact object can intercept this wind and accrete a portion of it.

In the main sequence and W–R phases wind mass-loss rates are typically $\dot{M}_{\text{wind}\star} \approx -10^{-6}$ to $-10^{-5} M_\odot \text{ yr}^{-1}$, with velocities of $\sim 1000 \text{ km s}^{-1}$. (In the LBV phase mass loss is higher, $\approx -10^{-5}$ to $-10^{-4} M_\odot \text{ yr}^{-1}$, and velocities are an order of magnitude lower.) HMXB systems usually have one of these massive stars and a neutron star compact object (or occasionally a $\sim 10 M_\odot$ black hole). With binary periods of one day to a few weeks and masses of 10–40 $M_\odot$, the binary separation is typically 3×10^{11} to $3 \times 10^{12} \text{ cm}$. This is of order the radius of an O/B star (10^{12} cm), but somewhat larger than that of a very hot W–R star ($2 \times 10^{11} \text{ cm}$). Using the law of conservation of mass for a constant velocity stellar wind

$$\dot{M}_{\text{wind}\star} = -4\pi r^2 V_w m_H n_w$$

we then can estimate the wind density near the compact object ($r = a$) to be $n_w \sim 3 \times 10^9$ to $3 \times 10^{12} \text{ cm}^{-3}$.

This now gives us enough information to estimate the rate at which the compact object accretes material from the wind. With $V_w \sim 10^8 \text{ cm s}^{-1}$ and $n_w \sim 10^{11} \text{ cm}^{-3}$, we can re-scale the Bondi equation (11.8) to obtain

$$\dot{M}_{\text{wind}\bullet}^* \approx 3.7 \times 10^{16} \text{ g s}^{-1} \left(\frac{M_\bullet}{M_\odot} \right)^2 \left(\frac{V_w}{10^8 \text{ km s}^{-1}} \right)^{-3} \left(\frac{n_w}{10^{11} \text{ cm}^{-3}} \right) \quad (11.65)$$

For long-period, weak-wind systems this rate can be as low as 10^{15} g s^{-1} , and for short-period, strong-wind systems as high as 10^{18} g s^{-1} (*i.e.*, between $0.001 < \dot{M}_\bullet / \dot{M}_{\text{Edd}\bullet} < 1$).

The distinction between wind accreting and Roche lobe accreting HMXB systems is often difficult to determine when the star and compact object are very close. Indeed, as the massive companion begins to fill its Roche lobe, the stellar wind becomes more focused toward the compact object and wind accretion begins to look more like Roche lobe overflow. It is believed that Cygnus X-1 is in this hybrid state.

11.1.6 Accretion in Collapsing Supernova Cores

Stars that begin their lives with masses in the range $11\text{--}21 M_\odot$ develop iron cores of $\sim 1.5 M_\odot$ that collapse to proto-neutron stars (see Section 10.1.2). Because it happens in a fraction of a second, we usually do not think of this collapse as a slowly evolving accretion event. Yet, in fact, it does happen “slowly”. The time scale for the core to collapse is, of course, its own dynamical time

$$\tau_{\text{dyn,core}} = \frac{R_{\text{core}}^{3/2}}{(GM_{\text{core}})^{1/2}} = 0.37 \text{ s} \left(\frac{M_{\text{core}}}{1.5 M_\odot} \right)^{-1/2} \left(\frac{R_{\text{core}}}{3000 \text{ km}} \right)^{3/2} \quad (11.66)$$

Once a, say, $0.5 M_\odot$ proto-neutron star forms, however, the dynamical time scale for that object is

$$\tau_{\text{dyn,PNS}} = 1.4 \text{ ms} \left(\frac{M_{\text{PNS}}}{0.5 M_\odot} \right)^{-1/2} \left(\frac{R_{\text{PNS}}}{50 \text{ km}} \right)^{3/2} \quad (11.67)$$

which is hundreds of times shorter than the core-collapse time. Therefore, the accretion onto this new compact object, which is governed by how fast the core collapses, occurs on a time scale much longer than the accreting object’s dynamical time.

On the other hand, the accretion rate onto the proto-neutron star is still enormous (equation (11.58))

$$\dot{M}_{\text{dyn}\bullet}^{\text{core}} \approx \frac{M_{\text{core}}}{\tau_{\text{dyn,core}}} = 8.2 \times 10^{33} \text{ g s}^{-1} \left(\frac{M_{\text{core}}}{1.5 M_\odot} \right)^{3/2} \left(\frac{R_{\text{core}}}{3000 \text{ km}} \right)^{-3/2} \quad (11.68)$$

which is $1.7 \times 10^{16} \dot{M}_{\text{Edd}\bullet}$! This large accretion rate is called *hypercritical* or *hyper-Eddington*, but, in fact, the Eddington limit for electron scattering of photons does not really apply here. There are few electrons and few photons to interact in this case. Instead, the dominant energy loss is via neutrino emission, and the object is *not* super-Eddington for neutrino absorption or scattering. The story is approximately the same when the collapse forms a black hole instead of a proto-neutron star.

11.2 Accretion of Angular Momentum: Spinup of the Engine Flywheel

It is important to understand the distinction between matter and angular momentum accretion. In terms of the black hole metric, matter accretion increases the *irreducible* mass of the black hole. Energy is released as the material falls toward the black hole; but, to high astrophysical accuracy, once that matter enters the hole no more binding energy of that matter can be extracted from it. It is this type of accretion that is believed to power the infrared, optical, and X-ray luminosity of black hole systems.

On the other hand, angular momentum accretion increases the *reducible* mass of the black hole, *i.e.*, its rotational energy. Of course, the rotational energy of the accreting matter can be tapped during the process of accretion. Additionally, however, even after the matter enters the black hole, *its rotational energy still can be tapped at a much later time*. The black hole acts like a large flywheel, storing rotational energy for future use. While the accretion of matter is usually needed to tap this rotational energy, we will find that it does not have to be highly efficient at producing observable thermal emission to cause the black hole to produce strong jet outflows. The theory of how black holes acquire and lose angular momentum, therefore, is an important part of the science of black hole engines. Unfortunately, many aspects of it are still poorly understood.

11.2.1 Angular Momentum Accretion in Binary Systems

Most studies of angular momentum evolution in binaries are concerned with the evolution of the *total* angular momentum of the two stars orbiting about each other, and how that affects the separation between the two (see Section 10.3). We do not deal with that here. Instead, we wish only to determine the rate at which angular momentum is deposited onto the compact object, affecting how fast it spins. In the next chapter we will find that, during the accretion process, almost all of the angular momentum in the accretion flow is transferred outward by turbulent and magnetic viscosity processes. The angular momentum per unit mass left in the accreting matter is, at most, that of the innermost stable circular orbit, which we obtain from equation (7.40). Multiplying by the accretion rate, we obtain the total angular momentum deposit rate onto a Schwarzschild hole at $r = r_{\text{isco}} = 3 r_{\text{S}}$ of

$$j^{\text{SCH}} \leq \sqrt{3} \dot{M}_{\bullet} r_{\text{S}} c = 2.1 \times 10^{36} \text{ g cm}^2 \text{ s}^{-2} \left(\frac{\dot{M}_{\bullet}}{\dot{M}_{\text{Edd}}} \right) \left(\frac{M_{\bullet}}{10 M_{\odot}} \right)^2 \left(\frac{\epsilon_{\text{acc}}}{0.1} \right)^{-1}$$

The time to spin the black hole up to a speed close to maximal is, therefore,

$$\tau_{\text{spinup}}^{\text{SCH}} = \frac{GM_{\bullet}^2/c}{\dot{J}_{\text{SCH}}} = 1.33 \times 10^7 \text{ yr} \left(\frac{\dot{M}_{\bullet}}{\dot{M}_{\text{Edd}}} \right)^{-1} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right) \quad (11.69)$$

This is the maximum rate at which a black hole can be spun up by accretion.

For neutron stars the calculation is similar, but we will use Newtonian formulae at the surface of the star at $r = R_{\text{NS}} = 10 \text{ km}$

$$\begin{aligned} \dot{J}_{\text{NS}} &\leq \dot{M}_{\bullet} (GM_{\bullet} R_{\text{NS}})^{1/2} \\ &= 1.6 \times 10^{34} \text{ g cm}^2 \text{ s}^{-2} \left(\frac{\dot{M}_{\bullet}}{\dot{M}_{\text{Edd}}} \right) \left(\frac{M_{\bullet}}{M_{\odot}} \right)^{3/2} \left(\frac{R_{\text{NS}}}{10 \text{ km}} \right)^{1/2} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right)^{-1} \end{aligned}$$

giving a neutron star spinup time of

$$\begin{aligned} \tau_{\text{spinup,NS}} &= \frac{I_{\text{NS}} \omega_{\text{NS}}}{\dot{J}_{\text{NS}}} = \frac{\frac{2}{5} M_{\bullet} R_{\text{NS}}^2 (GM_{\bullet}/R_{\text{NS}}^3)^{1/2}}{\dot{J}_{\text{NS}}} \\ &= 1.82 \times 10^7 \text{ yr} \left(\frac{\dot{M}_{\bullet}}{\dot{M}_{\text{Edd}}} \right)^{-1} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right) \end{aligned} \quad (11.70)$$

similar to equation (11.69). If accretion even at the low rate of $0.01 \dot{M}_{\text{Edd},\bullet}$ can be maintained for a long period of time, neutron stars and black holes can be spun up to significant rotational speeds in a time shorter than the age of the Galaxy. In the process, of course, the compact object's mass must increase by about $\Delta M_{\bullet} = \dot{M}_{\bullet} \tau_{\text{spinup}} \sim 0.35 M_{\odot}$ in the neutron star case and $\sim 3.5 M_{\odot}$ in the black hole case. The companion, therefore, must start out sufficiently massive to supply this amount of matter to the compact object.

11.2.2 Angular Momentum Accretion onto Massive Black Holes in AGN and Globular Clusters

11.2.2.1 Tidal Disruption of Stars and Molecular Clouds

If all orbits of stars and gas clouds accreted onto the central black hole in active galactic nuclei and globular clusters had parallel angular momentum vectors, then equation (11.69) would give the spinup time for supermassive black holes as well. All accretion events would add angular momentum with the same sense. However, in a purely spherical distribution of stars and clouds, there is an equal chance of the stellar orbit having an angular momentum vector in *any* direction. Over long periods of time, therefore, the accretion events should add *no* angular momentum, or $\dot{J}_{\bullet} = 0$. The normalized angular momentum, therefore, should decrease as the

black hole gains mass according to the relation

$$j = j_0 \left(1 + \frac{\dot{M}_\bullet}{M_{\bullet 0}} t \right)^{-2}$$

If the black hole doubles its mass, for example, its relative spin rate should decrease by a factor of four. The *spindown* time scale $-j_0/\dot{j}$, therefore, is $\tau_{spindown} \approx M_{\bullet 0}/2\dot{M}_\bullet$, which is similar to the *spinup* time scale in the binary case. This problem was first pointed out by Young, Shields, and Wheeler [408]. It suggests that massive and supermassive black holes should all be spinning slowly after a few tens of millions of years of accretion by tidal disruption of stars.

However, if the inner galaxy core is rotating, or the gas clouds have collected in a disk around the black hole, then the accretion is not purely spherical. In [Table 11.2](#) we estimated the relative contributions due to star and gas accretion. For molecular cloud accretion, if the material collects in a dusty torus and if that material can be efficiently funneled to the nucleus, then the material will have a large net angular momentum (parallel to the rotation axis of the torus), and it will accrete at about $0.6 \dot{M}_{\text{Edd}}$. This alone will spin up the black hole in a time comparable to equation (11.69). There are many examples of active galaxies in which the axis of the radio jet is perpendicular to the plane of the dusty torus or gas disk imaged by the Hubble Space Telescope or other means (M51, NGC 4258, NGC 4261), indicating that the molecular gas does have net angular momentum and that it is likely to be affecting the central engine. Note that these disks are often not aligned with the plane of the host spiral galaxy.

It also is possible that the inner galaxy core rotates as well. This rotation could be a remnant of the original formation of the galaxy or be caused by a recent merger with another galaxy.

11.2.2.2 Galaxy Mergers

A galaxy merger will have several effects on the central cores of the two coalescing star systems. First, the interaction of individual stars with the global asymmetric potential of the merging systems will cause some to be scattered into the loss cones of the two black holes at a rate much higher than the nominal rate discussed earlier in this chapter. One can expect the amount of material available for accretion to rise substantially and remain at a high rate through the merger process, lasting at least several tens of millions of years.

Second, early in the merger, the material from each core will be in orbit about the other one, thereby providing fuel for the other's black hole that has net angular momentum. Furthermore, as they approach each other more closely, the two galaxy cores will tidally interact and transfer orbital angular momentum to each other, spinning up each core in the process. The stars accreting onto the respective black holes in this supermassive binary system, of course, still will be on highly elliptical orbits, so some will be tidally disrupted into retrograde orbits about the holes. However,

over the duration of the merger, more stars will be accreted on prograde than retrograde orbits, spinning up the central black hole in each merging galactic center in the process. The net result should be two rapidly spinning black holes with essentially parallel spin. The time scale for this again will be given by equation (11.69), because there is far more angular momentum available in each galaxy core than is needed to spin up the central black hole to fast rotation rates.

However, Priyamvada Natarajan and James Pringle, of the University of Cambridge, UK, have shown that spin–orbit coupling between each black hole and its accretion disk causes the holes to precess and come into alignment with their respective rotating accretion flows on an even faster time scale [340]. The time for this to occur is much shorter than the accretion time by a factor of $j (r_S/R_{\text{warp}})^{1/2}$, where R_{warp} is the distance from the black hole out to where its rotating metric can drag the accreting material into a disk aligned with the black hole spin, creating a warp in the disk. With a model for the accretion disk, Natarajan and Pringle arrive at the very short alignment time of

$$\begin{aligned}\tau_{\text{align}\bullet} &= j \frac{M_\bullet}{\dot{M}_\bullet} \left(\frac{r_S}{R_{\text{warp}}} \right)^{1/2} \\ &= 2.5 \times 10^4 \text{ yr } j^{11/16} \left(\frac{\dot{M}_\bullet}{\dot{M}_{\text{Edd}}} \right)^{-7/8} \left(\frac{M_\bullet}{10^9 M_\odot} \right)^{-1/16} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right)^{7/8} \quad (11.71)\end{aligned}$$

As with the accretion time, $\tau_{\text{align}\bullet}$ scales strongly with the accretion efficiency and with the inverse of the accretion rate, but very weakly with black hole mass. Holes of any rotation rate will become aligned with the spin axis of the galaxy core or accretion torus in only a few tens of thousands of years. Then, if they are spinning slowly, they then will be spun up to rapid rotation rates on a time scale comparable to that in equation (11.69).

11.2.2.3 Black Hole Mergers

Eventually, the two binary black holes themselves will coalesce, forming a new, larger black hole at the center of the new, larger galaxy. What effect will this have on the final black hole spin? Scott Hughes and Roger Blandford, when at Caltech, considered this process in 1995. They found that a more rapid rotation (spinup) results from such mergers only if the two black holes are of similar mass (*i.e.*, within less than a factor of two of each other), or if the larger of the two already rotates rapidly *and* the smaller collides with it in a nearly prograde orbit [426]. They believed that these conditions occur only in rare cases, and so concluded that black hole mergers usually will spin *down* the central black hole. However, this analysis did not take into account the effect discussed above: the rapid alignment and spinup of the black holes by their respective galaxy cores. We now expect these processes to contribute to the alignment of both black hole spins with their orbital angular mo-

mentum, making such events much less rare. Will this change the conclusions about black hole mergers?

The answer is probably yes, but the degree to which it will is not yet clear. There is plenty of time during a galaxy merger for the two black holes to become aligned before they themselves coalesce. Alignment will occur in a few tens of thousands of years and spinup in a few tens of millions, while the galaxy merger will take a hundred million or more. On the other hand, the degree to which this effect is important will decrease as the mass ratio of the two galaxies (and hence of their respective black holes) decreases, just as Hughes and Blandford found for the pure black hole – black hole merger case. That is, a dwarf elliptical merging with a giant cD galaxy probably will not have much of an effect on the spin of the larger galaxy's central core and, therefore, not on the central black hole. A strong effect, therefore, still will require the two merging objects to be of similar mass, but the range of mass ratio may widen (from $q > 0.5$ to a much smaller value). Even in a 1:10 merger, for example, the mass of the smaller central galaxy core still might be a significant fraction of the larger black hole mass and, therefore, still may have a strong influence on the latter.

11.2.3 Angular Momentum Accretion in Collapsed Supernova Cores

We also can use the Wheeler, Meier, and Wilson paper [374] (discussed in Chapter 10) to understand the physics of collapsing, rotating iron cores and black holes in supernovae interiors. The rotation period of a proto-neutron star can be found from the law of conservation of angular momentum $I_{\text{PNS}}\omega_{\text{PNS}} = I_{\text{mantle}}\omega_{\text{mantle}}$ or

$$P_{\text{PNS}} = \left(\frac{R_{\text{PNS}}}{R_{\text{core}}} \right)^2 P_{\text{core}} = 28 \text{ ms} \left(\frac{R_{\text{PNS}}}{50 \text{ km}} \right)^2 \left(\frac{R_{\text{core}}}{3000 \text{ km}} \right)^{-2} \left(\frac{P_{\text{core}}}{100 \text{ s}} \right)$$

After the initial collapse, the proto-neutron star continues to contract quasi-statically as it cools by neutrino emission. Again, angular momentum conservation gives the final new neutron star's rotation period

$$P_{\text{NS}} = 1.1 \text{ ms} \left(\frac{R_{\text{NS}}}{10 \text{ km}} \right)^2 \left(\frac{R_{\text{core}}}{3000 \text{ km}} \right)^{-2} \left(\frac{P_{\text{core}}}{100 \text{ s}} \right) \quad (11.72)$$

So, if the initial iron core, which is roughly half the size of the earth, has a rotational period of about two minutes, the final neutron star can be born with a high rotation rate. This will be important for the formation of jets.

A similar statement can be made if the iron core collapses all the way to a black hole. This is believed to occur for stars that begin with masses greater than $30 M_{\odot}$ and, late in their lives, collect $4\text{--}10 M_{\odot}$ of iron in the core alone. With conservation of angular momentum again, the final reduced angular momentum of the black hole

will be approximately

$$j = \frac{I_{\text{core}} \omega_{\text{core}}}{GM_{\bullet}^2 c} = 1.0 \left(\frac{M_{\bullet}}{5 M_{\odot}} \right)^{-1} \left(\frac{R_{\text{core}}}{10^4 \text{km}} \right)^2 \left(\frac{P_{\text{core}}}{100 \text{s}} \right)^{-1} \quad (11.73)$$

Core rotation periods of a couple of minutes, therefore, are expected to lead to the formation of rapidly rotating black holes as well.

In 2003 Alexander Heger of the Los Alamos National Laboratory and his colleagues studied the evolution of rotating massive stars up to and beyond core collapse. They found the pre-supernova core to have a radius of ~ 2200 km and a period of about 17 s; after collapse, it produced a neutron star about 12 km in radius [427]. Equation (11.72) predicts a pulsar period of 0.5 ms in this case, which is close to the period that they obtained of 0.6 ms. The assumption leading to equations (11.72) and (11.73), namely conservation of angular momentum during collapse and accretion of the iron core, appears to be a reasonably valid one.

Chapter 12

The Combustion Chamber: Energy Generation by Gravitational Accretion

Bondi accretion theory is excellent for estimating the general properties of accreting black holes, particularly their accretion rate, luminosity, and long-term variability (*e.g.*, over weeks and months for X-ray binaries and several millions of years for quasars). However, it cannot tell us about the spectral properties and short-term variability. In order to determine whether an accreting black hole will be an X-ray source or an optical–UV source, for example, and whether it may vary on time scales of minutes, seconds or shorter, we must study the flow of the accreting gas itself well inside the tidal radius R_T . Only by investigating the accretion process itself will we be able to explain observations of black hole systems and fully understand these engines.

In some ways the task of understanding luminous accretion is simpler than understanding fueling and carburetion. Interior to the Bondi radius R_B the gravitational field is dominated by the black hole, and interior to R_T the fuel is all in gaseous form. Furthermore, for a variety of reasons, it is a good approximation to assume that the fuel, particularly the dominant light elements of hydrogen and helium, is fully ionized. This is because the temperature of the gas and radiation field near black holes tends to be very high, sometimes as high as the virial temperature (*i.e.*, the temperature when the pressure p is of order the gravitational energy density of the gas $GM\rho/r$)

$$\begin{aligned} T_V &\equiv \frac{GM\mu}{r\mathcal{R}} \\ &= 5.4 \times 10^{12} \text{ K} \left(\frac{r}{r_S} \right)^{-1} \end{aligned} \quad (12.1)$$

where μ here is the molecular weight of the gas, and r_S is again the Schwarzschild radius. But this basic simplicity is more than offset by the complexity of the hydrodynamics, radiative and atomic physics, and electrodynamics of hot plasma spiraling toward a black hole. Note that in this chapter we shall drop the “bullet” subscript, as it will be clear that the mass M refers to the accreting black hole (or neutron star); we do not need to distinguish between star, cloud, galaxy, and black hole.

In principle, we could simulate everything (equations (9.6) and (9.7), plus Maxwell's equations with (9.29) and some form of Ohm's law) on a computer, although that task is still daunting and only being partially addressed at the moment. The standard methods of treating accreting black holes is similar to that used to study stellar structure in the 1930s and 40s before the advent of powerful computers. There are four important stages in this process:

1. *Physical Structure: The Basis for Determining the Energy Generation.* First, the equations of hydrodynamics are solved with rather simple assumptions about the thermal properties of the gas (*e.g.*, steady state, equilibrium, *etc.*). In the case of stellar structure this takes the form of the assumption of hydrostatic equilibrium: nothing varies in time and all gas velocities are assumed to be zero. For black holes, there must be accretion, so zero velocities are out of the question. Steady-state inflow is a more reasonable approximation, but we need to allow for rapid flow speeds. Generally the inflow can be of two types: the gas can be in spherical free fall (rapid inflow speeds), or it can be stuck in orbit in a disk about the black hole with only a slow trickle of accreting material inching toward the black hole (slow inflow speeds). (Of course, there will be intermediate types of accretion as well.) Once we have a good idea of the physical structure of the flow, we can use that structure to compute how energy is generated.
2. *Thermal Structure: Computing the Continuum Emission Spectrum.* The balance between heating by viscous friction or compression and cooling by radiation or conduction then will tell us the run of temperature inside the gas. This thermal structure lets us calculate how much radiation is emitted at different temperatures, and the physical structure determined earlier tells us how that radiation diffuses throughout the accretion flow. The temperature of the radiation at places where the matter becomes optically thin – the *photosphere* – will tell us if the object will appear cool (red), hot (blue), or extremely hot (UV or X-ray source).
3. *Atmospheric Studies: Absorption/Emission Lines and Re-processing.* The atmosphere is the material that lies between the photosphere (where the continuum emission spectrum is formed) and the observer. It can scatter and absorb the continuum radiation, dramatically altering the spectrum. It even can add additional heat to the radiation, if part of the atmosphere is a corona whose temperature is much hotter than that of the photospheric radiation itself. In some cases the accretion flow is so optically thick that much of the radiation falls into the black hole before it can leak out. In other cases, the gas is so tenuous that it essentially is *all* atmosphere, allowing us to see deep into the accretion flow, very close to the black hole. Detailed analysis of the observed spectrum alone can tell us a great deal about the structure and operation of the engine.
4. *Stability Studies: Rapid Time Dependence.* Just as major changes in the accretion flow (due to intermittent disruption of stars and gas clouds) can cause the engine output to be time-dependent, so minor variations in the accretion flow near the black hole also can cause variations in the source luminosity. Even a steady rain of accretion from external sources can have its own instabilities. These can lead to flickering, oscillations, and even violent outbursts or a complete change in state of the accretion flow. The main purpose of linear stability studies is to perform

a reality check on the structural results above. (Structures determined under the assumption of a steady state sometimes prove to be unstable, rather than stable, equilibria.) If a structure has been found to be unstable, it means that the steady solution that we have found is in doubt. However, even if a stability analysis tells us that the flow is unstable, this may not mean disaster. The nonlinear evolution in the system could be minor (*e.g.*, turbulence, convection) or major (an explosion or a state change). We will not know the outcome of an instability until we actually perform time-dependent simulations of that unstable flow.

In the last three decades, a large number of analytic and semi-analytic (solving simple structure equations on a computer) studies have been made for accretion flows. While this has led to a better understanding of black hole engines, many questions and problems still remain. In short, accretion is not as well understood as stellar structure. Recently numerical simulations using powerful supercomputers have begun to answer some of these questions. However, most of these simulations are still in Stage 1 above, dealing in detail only with the dynamics and not much with the thermal and radiative properties, and the instabilities those behaviors bring.

One aspect of astrophysics is largely untouched in accretion studies: the evolution of the composition of the accreting gas. Temperatures can reach enormous values in the flow (equation (12.1)). However, generally, the gas densities are so low that nuclear reactions and processes still proceed more slowly than the time it takes the material to fall into the black hole. On the other hand, this is not true of accretion processes that occur in collapsing stellar cores. Densities there range from $10^{11-15} \text{ g cm}^{-3}$ and higher. Iron dissociates to make nucleons, and neutrons may combine to form more exotic kinds of matter before forming or entering a black hole. A full treatment of this problem will combine stellar evolution theory with accretion theory *and* probably with electrodynamics as well.

12.1 Spherical Accretion and Winds

We first will consider accretion toward a black hole when the infalling material has *no* angular momentum. The infall then is purely radial. Furthermore, because a true black hole has no solid surface, the inflow velocity then must achieve the speed of light at the horizon (Section 7.4.3). While pure spherical accretion does not appear to be very common, studying it first will help introduce some important concepts that will be important later in the study of disk accretion and wind outflows as well.

12.1.1 Physical Structure

Stage 1 in the analysis of spherical accretion is, of course, a determination of the physical structure of the flow. The non-relativistic equations are the easiest to ana-

lyze initially, but some relativistic accretion studies also have been done and those yield considerable insight into the influence of the black hole metric.

12.1.1.1 Non-Relativistic Spherical Accretion and Winds

Let us consider a gas falling radially in a steady state ($\partial/\partial t = 0$) in a Newtonian gravitational potential. If the flow is polytropic, the non-relativistic fluid equations (5.63, 5.64, and 5.55) for conservation of mass, momentum, and energy are

$$\begin{aligned} \frac{1}{r^2} \frac{d(\rho V r^2)}{dr} &= 0 \\ \rho V \frac{dV}{dr} &= -\frac{dp}{dr} - \frac{GM\rho}{r^2} \\ p &= A \rho^\Gamma \end{aligned}$$

where V is the radial velocity component. After a little manipulation, these equations can be combined into a single equation

$$\frac{d \ln V}{d \ln r} = - \left[\frac{GM/r - 2c_s^2}{V^2 - c_s^2} \right] \quad (12.2)$$

where $c_s = (\Gamma p/\rho)^{1/2}$ is the local polytropic sound speed in the gas. Equation (12.2) is called the wind/accretion equation and was studied for the sun by Eugene N. Parker of the University of Chicago. The equation is valid whether V is positive (wind outflow) or negative (accretion).

The wind equation has a host of solutions both physical and unphysical (see Fig. 12.1):

1. Wind Solutions

- *Breeze solution* (labeled IV): the reverse of the settling solution; flow accelerates outward but eventually slows to zero speed as $r \rightarrow \infty$.
- *Wind solution* (V): flow accelerates outward, becomes supersonic (also passing through the sonic point), and then asymptotically approaches a constant supersonic velocity.

2. Accretion Solutions

- *Settling solution* (IV): flow accelerates inward but then slows to zero speed near the accreting object.
- *Supersonic solution* (VI): flow accelerates inward, reaches the sound speed c_s at the “sonic point”, and then is in near free fall interior to that radius.

Solutions I–III are either unphysical (double-valued in r) or do not satisfy typical boundary conditions for winds or accretion flows.

The solutions that pass through the sonic point are called regular solutions, because both the numerator and denominator in equation (12.2) must be zero at that

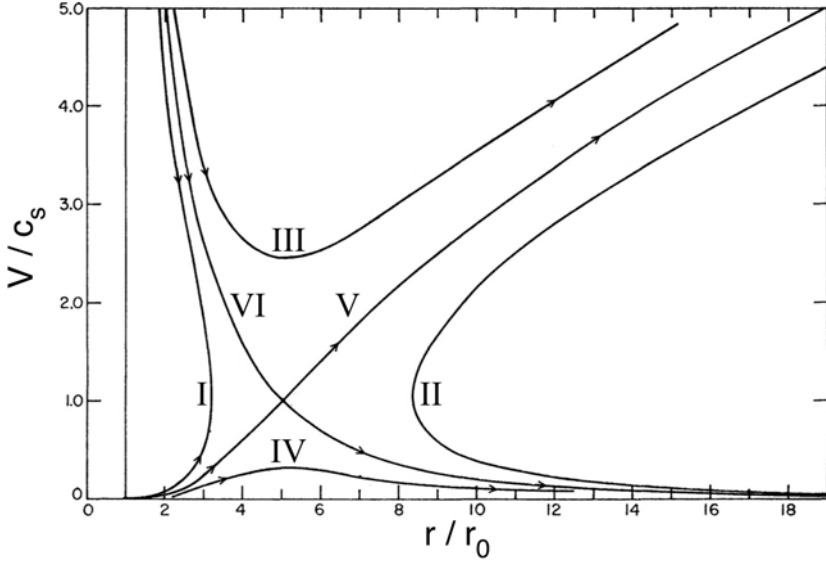


Fig. 12.1: Solutions of the Parker wind/accretion eq. (12.2): velocity in units of the sound speed vs. radius in units of an arbitrary r_0 (usually the stellar surface). Solutions I and II are double-valued in radius and, therefore, unphysical. Solution III is entirely supersonic (not relevant for an accelerating wind) and IV is a “breeze” solution, which never becomes a supersonic flow and asymptotically reaches zero velocity far from the star. Solution V is the critical wind solution that accelerates from slow to supersonic speeds. Solution VI is most useful as a spherical accretion, rather than wind, model. Inflow begins at subsonic speeds, passes through the sonic point, and accretes onto the star at supersonic speeds. (Note that eight years earlier Bondi [405] had constructed a very similar diagram; yet, as he was only interested in accretion, he did not discuss the equally valid wind solution.) Adapted from Fig. 4d in [428]. Reproduced by permission of the AAS.

radius, if we want the acceleration (left-hand side of equation (12.2)) to be finite. Setting both of these to zero allows us to solve for the radius of the sonic point and for the flow velocity there:

$$\begin{aligned} V &= \pm c_s \\ r_s &= \frac{GM}{2c_s^2} \end{aligned} \quad (12.3)$$

If we define the accretion rate to be the integration constant of the continuity equation above

$$\dot{M} = 4\pi r^2 \rho V \quad (12.4)$$

then we can solve completely for r_s in terms of constants¹

¹ Note that r_s is the radius of the sonic point, not r_S , which is the Schwarzschild radius.

$$r_s = \frac{1}{2} \left[(GM)^{\Gamma+1} (A\Gamma)^{-2} \left(\frac{\dot{M}}{\pi} \right)^{-2(\Gamma-1)} \right]^{\frac{1}{5-3\Gamma}}$$

From this equation we see that three types of accretion flow are interesting:

1. *Isothermal flow* ($\Gamma = 1$). The gas radiates most of its thermal energy and remains at a constant temperature. Here

$$r_s = \frac{GM}{2A} = \frac{GM}{2\mathcal{R}T}$$

2. *Adiabatic relativistic gas flow* ($\Gamma = 4/3$). The sonic radius is at

$$r_s = \frac{1}{2} (GM)^{7/3} \left(\frac{4A}{3} \right)^{-2} \left(\frac{\dot{M}}{\pi} \right)^{-2/3}$$

This has a similar behavior to the isothermal case, but the sonic radius now also is affected by the accretion rate: the lower the accretion rate the further out the sonic radius is, and the greater the region of supersonic flow.

3. *Adiabatic non-relativistic gas flow* ($\Gamma = 5/3$). In this special case the sonic radius is either zero or infinite, depending on the relative values of the entropy A , the accretion rate $\dot{M}$, and the central object mass M . In the high entropy, high accretion rate, low stellar mass case, the “zero” r_s means that the accreting gas never becomes supersonic. In the opposite case (infinite r_s), the accretion velocity is *always* supersonic. These extreme situations never will be achieved in practice, however, because some emission keeps $\Gamma < 5/3$, leaving a finite exponent in the expression for r_s .

Singular and Critical Surfaces of an Accelerating Hydrodynamic Flow

Solutions to the wind/accretion equation that have a vanishing denominator in equation (12.2) are called *singular* solutions, because they have the potential for an infinite (or nearly-infinite) acceleration if the numerator does not vanish also. The sonic point, therefore, often is referred to as a singular point because of this property. And, because the spherically symmetric flow is actually three-dimensional, there exists a spherical singular *surface*, which is the locus of all the singular (sonic) points. It is important to note that the singular surface, therefore, is a mathematical, not physical, concept, because its existence depends on the nature of an equation used to describe the flow. Usually, all singular solutions are ignored, and the regularity condition (found by setting the numerator of equation (12.2) to zero) is applied to obtain a physically regular solution that passes smoothly through the singular surface.

In Section 9.6.2 we discussed the concepts of flow characteristics and critical surfaces. Because the flow transitions from subsonic to supersonic at the sonic surface, the latter is also a *critical* surface. In the early days of spherical accretion and

winds, it was not immediately obvious that singular and critical surfaces did not necessarily coincide, so often the older astrophysical literature used the two terms interchangeably.² However, singular and critical surfaces are not the same thing; the former is a mathematical concept, while the latter is a physical one. They do *not* coincide in general, therefore, unless the flow streamlines are normal to surfaces of assumed symmetry in the mathematical analysis of the flow. In the case of spherical accretion, the flow is radial and, therefore, normal to the spherical surfaces of symmetry. The latter condition is satisfied, so the sonic surface, therefore, is both a singular and critical surface. This degeneracy will persist for the spherical winds of Chapter 13, but will be lifted for the MHD jets of Chapter 15.

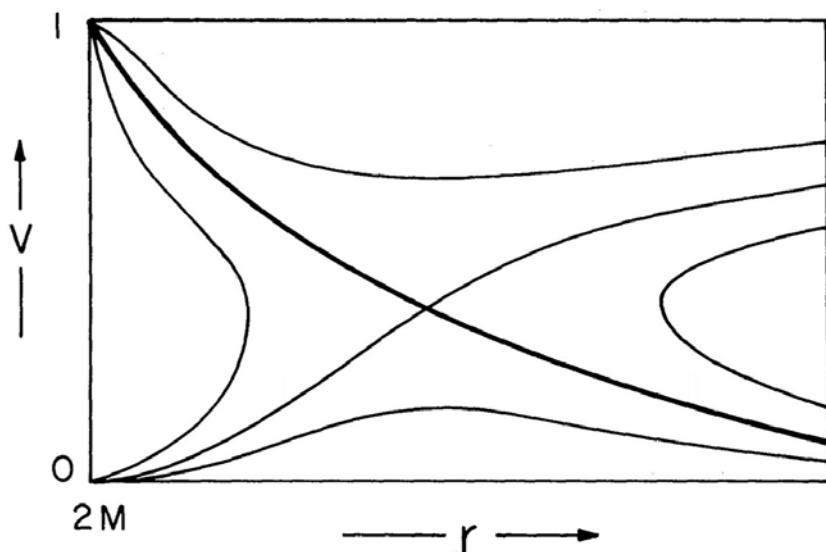


Fig. 12.2: Solutions to the relativistic spherical wind/accretion eq. (12.5) for a Schwarzschild black hole. The topological structure of the solutions is the same as in the non-relativistic case (unphysical, supersonic, breeze, and critical). Note that the stellar surface is now the Schwarzschild radius ($2GM/c^2$, with $c = G = 1$), and the inflow velocity at the horizon is unity (the speed of light). Adapted from Fig. 1 of [429]. Reproduced by permission of J. Wiley and Sons, Ltd.

12.1.1.2 Relativistic Spherical Accretion and Winds

Spherical accretion near a Schwarzschild black hole is similar to non-relativistic accretion but with some new features. This was studied by Kip Thorne, the late

² This ambiguity of terminology could confuse the diligent reader who may look up some of the original papers in order to obtain further information. The terminology that we use in this book was first introduced rather recently by [370] and then extended and used by several authors after the year 2000, particularly by [368].

Richard Flammang, and Anna Żytkow (of Caltech) in the early 1980s [430]. The wind/accretion equation for relativistic polytropic flow is

$$\frac{d \ln V}{d \ln r} = - \left[\frac{\left(1 - \frac{c_s^2}{c^2}\right) \frac{GM}{r} \epsilon^2 - \left(1 - \frac{V^2}{c^2}\right) 2 c_s^2}{V^2 - c_s^2} \right] \quad (12.5)$$

where

$$\epsilon = \left[\left(1 - \frac{2GM}{c^2 r}\right) / \left(1 - \frac{V^2}{c^2}\right) \right]^{1/2}$$

is the total energy of the accreting material (including mass) at infinity per unit mass. [In the non-relativistic case $\epsilon \rightarrow 1$ and $V^2/c^2, c_s^2/c^2 \rightarrow 0$, so equation (12.5) reduces to (12.2).] We again can solve for the radius of the sonic point, now in the relativistic case, as

$$r_s = \frac{r_g}{2} \left(\frac{c^2}{c_s^2} + 3 \right) \quad (12.6)$$

which reduces to equation (12.3) when $c_s \ll c$. As we found out in Chapter 9, the maximum sound speed is that of a highly relativistic gas ($p = \frac{1}{3} \rho c^2$, $\Gamma = 1$)

$$c_s \leq \frac{c}{\sqrt{3}}$$

So, plugging this into equation (12.6) we find out that the sonic radius is always outside the photon orbit. That is,

$$r_s \geq 3 r_g$$

Fig. 12.2 shows the wind/accretion solutions of equation (12.5). They have the same critical (sonic point) character as the non-relativistic flow, but with one important difference. The supersonic accretion solution reaches the speed of light *exactly* at the black hole horizon ($r = 2r_g$) and it *must* go supersonic outside the photon orbit at $r_{\text{ph,SH}} = 3r_g$.

12.1.2 Thermal Structure and Accretion Luminosity: The Trapping Radius

Because spherical accretion is not of great interest in the study of black holes, we will not spend much time on the details of the thermal structure and spectrum of the flow. However, there are a few important concepts worth mentioning. One is the “trapping radius”, interior to which the advection of photons toward the black hole

($|V| \varepsilon_r$) exceeds the rate at which they diffuse out ($-\frac{c}{\kappa \rho} \frac{d\varepsilon_r}{dr} \approx \frac{c \varepsilon_r}{\kappa \rho r}$). The radius interior to which the photons are trapped, therefore, is

$$r_{\text{tr}} = \frac{c}{\kappa \rho |V|}$$

where κ is the radiative opacity (Section 9.2.1), which is usually taken to be that of electron scattering for hot accretion flows. If we use the conservation of mass (equation (12.4)), we find that

$$r_{\text{tr}} = \frac{\kappa_{\text{es}} \dot{M}}{4\pi c} = \frac{\dot{M}}{\epsilon_{\text{acc}} \dot{M}_{\text{Edd}}} r_g$$

In other words, *the photons become trapped in the accreting hot gas when the accretion rate approximately exceeds the Eddington rate* (equation (11.4)). We then can estimate the accretion luminosity in this high accretion rate situation by assuming that only the accretion energy outside the trapping radius escapes the black hole system (the rest being dragged into the black hole before it can radiate)

$$L \approx \frac{GM \dot{M}}{r_{\text{tr}}} = L_{\text{Edd}}$$

This rough estimate turns out to be very accurate. As $\dot{M}/\dot{M}_{\text{Edd}} \rightarrow \infty$ the luminosity approaches L_{Edd} and does not exceed it for pure spherical accretion. The efficiency $\epsilon_{\text{acc,sph}} = L/\dot{M}c^2$ for spherical accretion then *decreases* with accretion rate

$$\epsilon_{\text{acc,sph}} = \left(\frac{\dot{M}}{\epsilon_{\text{acc}} \dot{M}_{\text{Edd}}} \right)^{-1} = \frac{r_g}{r_{\text{tr}}}$$

12.1.3 Bondi Accretion vs. Begelman Accretion

12.1.3.1 Bondi Accretion

In Chapter 11 we learned that a star immersed in a uniform gas of density ρ_∞ and sound speed $c_{s\infty}$ will accrete at the Bondi rate of

$$\dot{M}_{\text{Bnd}} = \frac{8\pi}{3} \frac{G^2 M^2 \rho_\infty}{c_{s\infty}^3}$$

This implies an accretion efficiency of

$$\epsilon_{\text{acc,Bnd}} = 3 \frac{c_{s\infty}^2}{c^2} \left(\frac{1}{2} \frac{c_{s\infty}}{c} \frac{\lambda_\infty}{r_g} \right)$$

where $\lambda_\infty = 1/\kappa\rho_\infty$ is the mean free path of a photon at infinity (well away from the black hole). We can define a parameter

$$\chi \equiv \frac{c_{s\infty} \lambda_\infty}{c r_g}$$

that tells us if photons leak out of the infalling gas before or after they are scattered. In the Bondi case, $\chi \ll 1$, the photons are scattered or absorbed *before* they leak out, so the accretion should proceed at the Bondi rate $\dot{M}_{\text{Bnd}}$.

12.1.3.2 Begelman Accretion

However, when photons leak out faster than they are scattered ($\chi \gg 1$), then the efficiency can *rise* to the maximum for spherical accretion found by Mitchell Begelman

$$\epsilon_{\text{acc,Beg}} = \frac{3 c_{s\infty}^2}{c^2}$$

This kind of flow is *non*-adiabatic, with a significant amount of heat lost due to radiation. This is what gives it different characteristics from Bondi accretion. The heat loss lowers the pressure, allowing the accretion to rise to the Begelman rate

$$\dot{M}_{\text{Beg}} = \frac{\dot{M}_{\text{Edd}}}{\epsilon_{\text{acc,Beg}}} = \frac{1}{3} \frac{c^2}{c_{s\infty}^2} \dot{M}_{\text{Edd}}$$

which is much higher than the Bondi rate. The maximum *efficiency* (and minimum accretion rate $\dot{M}_{\text{Beg}} = \dot{M}_{\text{Edd}}$) occurs when the external gas is highly relativistic ($\epsilon_{\text{acc,Beg}} = 1$). That is, a black hole immersed in a large cloud of optically thin, highly-relativistic plasma (with $c_{s\infty} = c/\sqrt{3}$) will accrete spherically at the Eddington rate and produce one Eddington luminosity. Of course, this result assumes that the black hole remains in the plasma for long enough for the accretion flow to set up and reach a steady state.

12.2 Classical Turbulent Accretion Disk Theory

Disk accretion is believed to be, by far, the dominant mode of accretion in black hole engines. Any time the accreting gas has more angular momentum per unit mass than that of gas particles at the last stable orbit, the gas will orbit the black hole rather than falling into it directly. For a Schwarzschild black hole, with $r_{\text{isco}} = 6 GM/c^2$ and orbital speed $V_{\text{isco}} = c/2$, the angular momentum per unit mass is only

$$\ell_{\text{isco}} = \frac{2\sqrt{3}GM}{c} = 1.5 \times 10^{16} \text{ cm}^2 \text{ s}^{-1} \left(\frac{M}{M_{\odot}} \right)$$

At the Bondi radius, the gas would have to be moving at less than 0.3 cm s^{-1} in order to have less angular momentum than at ℓ_{isco} . Even gas with a rotational velocity of only 1% of the sound speed would circularize at a radius of

$$r_c = 7.3 \times 10^7 r_g \left(\frac{c_s}{7 \times 10^4 \text{ cm s}^{-1}} \right)^{-1} \left(\frac{V_{\text{rot}}}{0.01 c_s} \right)$$

– very far from the black hole. So, for the rest of this book we will concentrate on disk accretion and on how to get rid of the angular momentum so the gas can fall into the black hole.

12.2.1 Physical Structure

To first order, all disk accretion has the same structure:

- The gas is in stable orbits about the black hole, whose gravitational field totally dominates over that of the disk itself.
- At each cylindrical radius R , the half-height of the disk is thin ($H \ll R$) or “slim” ($H < R$).
- Inflow occurs toward the black hole and normal to the orbital motion, with a velocity slower than free fall ($-V_R < V_{\text{ff}}$).
- The inward drift of matter also brings in angular momentum, which is exactly balanced by a transport of angular momentum *outward* via some kind of viscous torque.
- The mass accretion rate is assumed to be constant with disk radius.

The continuity equation and three components of the momentum equation, therefore, define a steady state of orbital equilibrium radially (in R), hydrostatic equilibrium vertically (in Z), and equilibrium angular momentum transport (the ϕ component), with steady accretion flow. The non-relativistic equations for this situation can be derived from equations (5.63) and (5.64), expressed in cylindrical coordinates

$$\begin{aligned} \frac{1}{R} \frac{d(\rho V_R R)}{dR} &= 0 \\ V_{\phi}^2 &= \frac{GM R^2}{(R^2 + Z^2)^{3/2}} \approx \frac{GM}{R} \\ \frac{dp}{dZ} &= -\frac{GM Z}{(R^2 + Z^2)^{3/2}} \rho \approx \frac{GM Z}{R^3} \rho \\ \rho \frac{V_R}{R} \frac{d(RV_{\phi})}{dR} &= -\frac{1}{R^2} \frac{d(R^2 t_{\phi R})}{dR} \end{aligned}$$

where $t_{\phi R}$ is the ϕR component of the (spatial) viscous stress tensor (equation (9.15)) and the approximations assume a thin or slim disk.

Let us now define the accretion rate and total torque as the vertical integrals at radius R of the mass flow and stress per unit volume

$$\dot{M} = -2\pi R \int_{-H}^H \rho V_R dZ$$

$$W_{\phi R} = \int_{-H}^H t_{\phi R} dZ$$

and use the “one-zone” approximation³ that $dp/dZ \approx p/H$. We also integrate the angular momentum conservation equation in R , taking advantage of $\dot{M}$ being constant. The results are the disk physical structure equations

$$\dot{M} = -4\pi R H \rho V_R \quad (12.7)$$

$$V_\phi = \left(\frac{GM}{R} \right)^{1/2} \quad (12.8)$$

$$c_s^2 = \frac{p}{\rho} = \frac{GM H^2}{R^3} = V_\phi^2 \left(\frac{H^2}{R^2} \right) \quad (12.9)$$

$$2 t_{\phi R} H \approx W_{\phi R} = \frac{\dot{M}}{2\pi} \left(\frac{GM}{R^3} \right)^{1/2} \left[1 - \left(\frac{R}{r_{\text{isco}}} \right)^{-1/2} \right] \quad (12.10)$$

Equation (12.10) applies a boundary condition that the torque $W_{\phi R}$ is zero at $R = r_{\text{isco}}$. This was first applied by Nikolay Shakura (Sternberg Astronomical Institute, Moscow) and Rashid Sunyaev (then at the Academy of Sciences, Moscow), who first worked out the theory of disk accretion that we present here [351]. Recently there has been much discussion about whether or not this is an appropriate boundary condition to apply. Clearly, the torque should be zero at some radius, but r_{isco} may not be the correct one. After all, orbits between r_{mb} and r_{isco} ($= 4r_g$ and $6r_g$, respectively, for a Schwarzschild black hole) do exist; they simply are unstable and destined to spiral toward the black hole. In fact, photons can exert a torque (photon viscosity) down to at least the photon orbit radius at $3r_g$. Nevertheless, we will continue to use r_{isco} as the standard zero-torque radius in this book. In practice this choice will not matter much, anyway. This is one question that should be answered eventually by relativistic numerical simulations of disk accretion.

³ The one-zone approximation is a popular method of analyzing stellar structure and obtaining useful estimates of central densities, temperatures, *etc.* We will make extensive use of it in this chapter when we want to obtain analytic models of accretion flows.

12.2.1.1 Energy Conservation, Frictional Heating, Disk Luminosity, and the Optically Thick Spectrum

Equation (9.16) showed that viscous torque produces a frictional heating locally at a rate per unit volume of

$$\rho \frac{dq}{dt} = -2 t_{\phi R} \Sigma_{\phi R}$$

where the ϕR component of the shear tensor here is

$$\Sigma_{\phi R} = \frac{1}{2} R \frac{d(V_{\phi}/R)}{dR} = -\frac{3}{4} (GM/R^3)^{1/2}$$

If we integrate this energy generation rate over the disk height, half will emerge above the disk and half below, so that the energy generated per unit area on the disk surface is

$$\begin{aligned} Q_{\text{acc}} &= \frac{1}{2} \int_{-H}^H \rho \frac{dq}{dt} dZ = -W_{\phi R} \Sigma_{\phi R} \\ &= \frac{3\dot{M}}{8\pi} \frac{GM}{R^3} \left[1 - \left(\frac{R}{r_{\text{isco}}} \right)^{-1/2} \right] \end{aligned}$$

Now, if we integrate this heating *again*, this time over the disk surface, from r_{isco} to infinity, and take into account both the upper and lower disk surfaces, we find the total luminosity of the disk to be the simple expression

$$L_{\text{acc}} = 2 \int_{r_{\text{isco}}}^{\infty} Q_{\text{acc}} 2\pi R dR = \frac{GM\dot{M}}{r_{\text{isco}}}$$

That is, in this accretion model, the accreting matter releases *all* of its binding energy as viscous heating as it sinks down to r_{isco} . Therefore, even though it would appear less violent than free fall, disk accretion can be as, or more, efficient than spherical accretion.

If the disk is optically thick and radiates like a black body at each radius, we can immediately determine the run of effective temperature with radius on the surface of the disk by setting $Q_{\text{acc}} = \sigma T_*^4$

$$T_* = 1.5 \times 10^7 \text{ K} \times \left(\frac{M}{M_{\odot}} \right)^{-1/4} \left(\frac{\dot{M}}{\dot{M}_{\text{Edd}}} \right)^{1/4} \left(\frac{R}{6r_g} \right)^{-3/4} \left[1 - \left(\frac{R}{r_{\text{isco}}} \right)^{-1/2} \right]^{1/4}$$

(12.11)

which has a maximum at $R = (49/6) r_g = 8.2 r_g$ when $r_{\text{isco}} = 6 r_g$. This is a famous and very useful result. It says that, because the engine is a thin disk, we

can see deep into the center of the engine, very near r_{isco} . Solar-mass black holes ($M = 10 M_{\odot}$) should be X-ray sources (10^7 K) while supermassive black holes ($M = 10^9 M_{\odot}$) should be ultraviolet sources (10^5 K). This is indeed the case, at least for the thermal emission, but there is much more to the story.

From this run of temperature, then the total radiation spectrum emitted by a disk of this type can be found by integrating the emitted flux over the upper and lower disk surfaces

$$I_{\nu} = 4\pi \int_{R_{\text{in}}}^{R_{\text{out}}} F_{\nu} [T_{*}(R)] R dR \quad (12.12)$$

If the effective optical depth to absorption τ_{*} is high, then the local spectrum will be Planckian (equation (9.60)), and the effective radiation temperature can be written as a scaling law from equation (12.11)

$$T_{*} = T_{*,\text{in}} \left(\frac{R}{R_{\text{in}}} \right)^{-3/4}$$

If we choose inverse temperature z as our new integration variable in equation (12.12), with $R = R_{\text{in}} (z/z_{\text{in}})^{4/3}$ and $z_{\text{in}} = h\nu/kT_{*,\text{in}}$, we can integrate all the black body spectra over radius for low frequencies ($\nu \ll kT_{*,\text{in}}/h$), holding ν fixed (we are integrating over R , not ν), and obtain

$$I_{\nu} = \frac{32 \Gamma(8/3) \zeta(8/3) h R_{\text{in}}^2}{3 c^2} \left(\frac{kT_{*,\text{in}}}{h} \right)^{8/3} \nu^{1/3} \quad (12.13)$$

where Γ and ζ are the standard Gamma and Zeta functions. For $\nu \gg kT_{*,\text{in}}/h$, the spectrum has an exponential cutoff, $I_{\nu} \propto f(x_{\text{in}}) \exp(-h\nu/kT_{*,\text{in}})$, with f being a slowly-varying function of x_{in} .

Equation (12.13) is a famous result. It says that the spectrum of a thermally-emitting, thin accretion disk around a black hole should rise slowly with frequency and then cut off at the temperature of the inner disk, near the innermost stable circular orbit. This thermal component has been seen in many accreting black holes as, for example, the “big blue bump” in quasars or the cool emission in the “soft state” of X-ray binaries (*e.g.*, Fig. 16.7). However, this thermal emission does not dominate in all accreting black hole systems. Often most of the local emission is produced instead by a corona-like inflow, perhaps one hundred times hotter, that lies above the disk or even replaces it entirely.

12.2.1.2 The α -Model for Accretion Torque

The above basic structure equations are still missing some key information. The most important is a physical model for the torque – and an equation for $t_{\phi R}(\rho, T)$. A solution for this problem was proposed by Shakura and Sunyaev (SS) in their 1973 paper, and this solution remained the standard for accretion disk theory for

almost three decades. Noting that particle viscosity is much too small to account for accretion rates inferred from observations of X-ray sources, they proposed that angular momentum transport in accretion disks probably occurred either by magnetic torques ($t_{\phi R}^{\text{mag}} = B_\phi B_R / 4\pi$) or by turbulent viscosity ($t_{\phi R}^{\text{turb}} = \langle \rho V_R V_\phi \rangle$). Both of these, they reasoned, would be limited in strength by the local dynamics, specifically the local pressure. If the magnetic field pressure greatly exceeded the local thermal pressure, the field lines would bulge out of the disk into the corona (due to the Parker instability) and be less effective in coupling to the matter in the disk and transferring angular momentum. Therefore, we expect

$$t_{\phi R}^{\text{mag}} = \frac{B_\phi B_R}{4\pi} \approx \frac{B^2}{4\pi} < p \quad (12.14)$$

On the other hand, if the turbulence were more important, then it likely would be subsonic: supersonic turbulence would tend to produce internal shock heating in the disk and quickly slow to subsonic. So, we would expect

$$t_{\phi R}^{\text{turb}} \approx -2\eta_{v,t} \Sigma_{\phi R} = \langle \rho V_\phi V_R \rangle \approx \rho V_{\text{turb}}^2 < \rho c_s^2 = p \quad (12.15)$$

Therefore, Shakura and Sunyaev proposed, a reasonable model for the torque is to assume that it is proportional to the pressure

$$\boxed{t_{\phi R} = \alpha p} \quad (12.16)$$

with $\alpha \leq 1$. Equation (12.16) is the “alpha” model expression for the torque that transports angular momentum in accretion disks. It remained the best model for that process until the mid–late 1990s, when magnetohydrodynamic computer simulations of magnetized accretion disks showed more precisely how magnetic fields and turbulence transport angular momentum.

Even without detailed simulations, we can estimate quantities like the magnetic field and the turbulent velocity from equation (12.16). Using equations (12.14) and (12.15) we have

$$B \approx (4\pi \alpha p)^{1/2} \quad (12.17)$$

$$V_{\text{turb}} \approx \alpha^{1/2} c_s \quad (12.18)$$

assuming fully developed turbulence, where $B_\phi \sim B_Z \sim B_R$ and $V_{\text{turb}, \phi} \sim V_{\text{turb}, Z} \sim V_{\text{turb}, R}$.

Furthermore, with equation (12.16) giving us the force per unit volume that transports angular momentum, we now can use the structure equations to obtain a more useful expression for the accretion velocity

$$V_R = -V_\phi \alpha \frac{H^2}{R^2} \left[1 - \left(\frac{R}{r_{\text{isco}}} \right)^{-1/2} \right]^{-1} \quad (12.19)$$

If the disk is very thin ($H \ll R$) and $\alpha \ll 1$, then the accretion drift velocity V_R will be much smaller than the orbital and free-fall velocities. On the other hand, if the disk is fairly thick ($H \sim R$) and the torque very strong ($\alpha \sim 1$), then the inflow will be nearly in free fall and will behave more like the spherical accretion we studied Section 12.1.

12.2.1.3 Estimate of the Vertical Disk Magnetic Field: The Disk–Jet Connection

By its very nature, analytic turbulent accretion disk theory does not produce a well-determined value for each magnetic field component. It produces only an estimate of the average total magnetic field strength locally, without specifying the direction in which that field is pointing (equation (12.17)). Indeed, because of the turbulence, the field direction may change randomly from point to point inside the disk. However, we will find in Chapter 15 that an estimate of the vertical component (B_Z) of the field will be needed in order to estimate the power of any MHD jet being launched from the disk.

The standard approach to this dilemma is to adopt a suggestion by Mario Livio (of the Space Telescope Science Institute) and Gordon Ogilvie and James Pringle (both of the University of Cambridge, UK) [431], which was, in turn, used by this author to develop a general picture of jets from accreting black hole systems [432]. Livio and his colleagues argued that, in an accretion disk, the vertical component of the magnetic field should be smaller than the total magnetic field by the ratio of the disk height to the radius where that height is measured, raised to some power

$$B_Z \approx B \left(\frac{H(R)}{R} \right)^p \quad (12.20)$$

with $p = 1$ being the most favored value. Because B_Z is so important for estimating the power of jet outflows from accretion disks, in the models below we also will include the approximate estimate of B_Z above and other components as needed.

12.2.2 Thermal Structure and Continuum Emission Spectrum

Virtually *all* accretion disk models use the same physical structure that was introduced in the above subsections. The differences between the different disk models occur mostly in the type of energy conservation equation used and, to a lesser extent, in the equation of state and laws of radiative emissivity that are used. Which energy equation terms to use in which situations is still a matter of some controversy. It is likely that a good answer to this question will not be found until numerical simulations become sophisticated enough to perform complex energy generation and transport.

In the next several subsections we shall treat a number of different types of accretion disks, each one characterized by a different energy equation and a different physical situation (optically thick flow, optically thin flow, *etc.*). All of these then will be combined into a comprehensive theory for accretion flows onto black holes for different accretion rates and black hole masses.

12.2.2.1 Thermal Disk Structure #1: Optically Thick, Geometrically Thin, Radiative Disks

We will begin with the “standard” energy equation that was proposed by Shakura and Sunyaev, in which we assume (1) that *all* heat generated by viscous friction in the disk is radiated away locally at each radius in the disk and (2) that the disk plasma is optically thick, and radiation diffuses vertically upward in the Z direction. That is, we can set the local viscous heating rate equal to the radiative losses

$$Q_{\text{acc}} = Q_{\text{r}} \equiv -\frac{a c}{\kappa \rho} \frac{d(T^4)}{dZ} \approx \frac{a c T_c^4}{\kappa \rho H} \quad (12.21)$$

where T_c is the midplane (central) disk temperature at each radius R . With expressions for the pressure $p(\rho, T)$ and the opacity $\kappa(\rho, T)$, the disk structure equations close, and a complete solution for all the variables then can be found.

The Three Radial Regions of a Standard Optically Thick Disk

Based on what type of plasma pressure and what type of opacity dominates, Shakura and Sunyaev identified three important *annular* regions in disk radius, whose names indicate their relative distance from the black hole: the “outer”, “middle”, and “inner” disk regions. These regions are somewhat analogous to different spherical shells in stellar structure, but generally these regions are all visible to outside observers simultaneously.

The detailed properties of each region are as follows:

1. *The “Outer” Disk Region* ($p = p_g$, $\kappa = \kappa_{\text{ff/bf}}$). For moderately hot temperatures, the equations of state are dominated by the (fully ionized) plasma pressure (equation (9.51)) and free–free/bound–free absorption (equation (9.81)). For solar abundances, the molecular weights are $\mu \approx 0.61$ and $\mu_e \approx 1.17$; so we have

$$p = 1.64 \rho \mathcal{R} T \quad (12.22)$$

$$\kappa = 1.55 \times 10^{24} \text{ cm}^2 \text{ g}^{-1} \rho T_e^{-7/2} \quad (12.23)$$

Equations (12.7), (12.9), (12.10), (12.16), and (12.21) to (12.23) each are simple algebraic relations that have only one term on each side of each equation, with the seven unknowns ρ_c , T_c , p_c , κ_c , $t_{\phi R, c}$, H , and V_R , where the c subscript indicates the central midplane value at a given cylindrical radius R . We also introduce two

new, vertically-integrated quantities of optical depth

$$\tau_c = \int_0^H \kappa \rho dZ \approx \kappa_c \rho H \quad (12.24)$$

and the disk *surface* density

$$\Sigma = \int_{-H}^H \rho dZ \approx 2 \rho H \quad (12.25)$$

The above equations can be solved easily by hand to give several extraordinarily useful equations for the disk structure

ρ_c	=	17	g cm ⁻³	$\alpha^{-7/10}$	$m^{-7/10}$	$\dot{m}^{11/20}$	$x^{-15/8}$	$f^{11/20}$
T_c	=	9.3×10^7	K	$\alpha^{-1/5}$	$m^{-1/5}$	$\dot{m}^{3/10}$	$x^{-3/4}$	$f^{3/10}$
κ_c	=	3.4×10^{-3}	cm ² g ⁻¹			$\dot{m}^{-1/2}$	$x^{3/4}$	$f^{-1/2}$
p_c	=	2.2×10^{17}	dyn cm ⁻²	$\alpha^{-9/10}$	$m^{-9/10}$	$\dot{m}^{17/20}$	$x^{-21/8}$	$f^{17/20}$
H	=	8.2×10^3	cm	$\alpha^{-1/10}$	$m^{9/10}$	$\dot{m}^{3/20}$	$x^{9/8}$	$f^{3/20}$
$\tau_{c,\text{ff/bf}}$	=	470		$\alpha^{-4/5}$	$m^{1/5}$	$\dot{m}^{1/5}$		$f^{1/5}$
V_R	=	-1.04×10^6	cm s ⁻¹	$\alpha^{4/5}$	$m^{-1/5}$	$\dot{m}^{3/10}$	$x^{-1/4}$	$f^{-7/10}$
Σ	=	2.8×10^5	g cm ⁻²	$\alpha^{-4/5}$	$m^{1/5}$	$\dot{m}^{7/10}$	$x^{-3/4}$	$f^{7/10}$
B_c	=	1.66×10^9	G	$\alpha^{1/20}$	$m^{-9/20}$	$\dot{m}^{17/40}$	$x^{-21/16}$	$f^{17/40}$
B_Z	≈	1.53×10^7	G	$\alpha^{-1/20}$	$m^{-11/20}$	$\dot{m}^{23/40}$	$x^{-19/16}$	$f^{23/40}$
V_{turb}	=	1.13×10^8	cm s ⁻¹	$\alpha^{2/5}$	$m^{-1/10}$	$\dot{m}^{3/20}$	$x^{-3/8}$	$f^{3/20}$
$L_{\text{acc,O}}$	=	2.4×10^{38}	erg s ⁻¹		m	$\dot{m}$		

(12.26)

where we have saved space by defining a measure of proximity to the ISCO as

$$f \equiv \left(1 - x^{-1/2}\right)$$

This will be our standard way of quoting results for accretion disk structure, but it can appear strange and cumbersome to people not used to this type of notation. So, we will take a little time to explain it below:

- The equations tell us the structure as a function of radius $x \equiv R/r_{\text{isco}}$ only; in this notation R is given by

$$R = 8.9 \times 10^5 \text{ cm } m x$$

x can have a minimum of 1 (the radius of the innermost stable circular orbit around the black hole) out to essentially infinity. None of the quantities have any *vertical* dependence in height Z ; they are either central midplane values (ρ_c , T_c , *etc.*), parameters at each radius (H), vertically-integrated quantities (τ , Σ), or variables that are believed to vary slowly with Z (V_R).

- The remaining parameters are the viscosity parameter α , the mass of the black hole, scaled to the mass of the sun

$$m \equiv \frac{M}{M_{\odot}} \quad (12.27)$$

and the accretion rate, scaled to the Eddington rate

$$\dot{m} \equiv \frac{\dot{M}}{\dot{M}_{\text{Edd}}} \quad (12.28)$$

The Eddington accretion rate is given by equation (11.4). *Note that the definition of the Eddington rate assumes a 10% efficiency of energy release*, so the actual Eddington accretion rate is 10 times that we would have computed had we assumed 100% efficiency.⁴ For these parameters, we expect values in the ranges $0.01 < \alpha < 1.0$, $1 < m < 10^{10}$, and $10^{-5} < \dot{m} < 100$. Not all sets of equations we present will be valid for all ranges of the parameters.

- Each result in equations (12.26) contains a numerical coefficient, in cgs units. That value represents the conditions near r_{isco} for a viscosity parameter of $\alpha = 1$, a black hole mass of $M = M_{\odot}$, and an accretion rate of $\dot{M} = \dot{M}_{\text{Edd}}$.
- The expressions also give factors of powers of the three parameters and functions of the radius. Unlike numerical simulations, which generally deal with only one black hole mass and one accretion rate, expressions like those above immediately give results for *all* types of black holes in all types of accretion situations, and do so for many positions in the accretion disk. As long as the assumptions made in deriving the equations remain valid, these results are enormously powerful in their breadth of application.
- The careful reader will notice that the central absorption opacity in equations (12.26) is $\kappa_c = 0.0034 \text{ cm}^2 \text{ g}^{-1}$, which is much smaller than the electron scattering opacity of $\kappa_{\text{es}} = 0.40/\mu_e \text{ cm}^2 \text{ g}^{-1}$ (equation (9.78)). This particular set of equations, therefore, cannot be valid all the way down to r_{isco} for any but the very lowest accretion rates. In fact, they are valid only when $\kappa_{\text{ff/bf}} \geq \kappa_{\text{es}}$, *i.e.* the region in the n - T plane (Fig. 9.3) below the dashed line. (So, all outer disk regions lie in the triangle, labeled “ p_g ” and “ κ_{abs} ”, in the lower center of that figure.) In a standard accretion disk the absorption opacity dominates when

$$R \geq R_{\text{O-M}} \approx 2830 r_g \dot{m}^{2/3} \left(1 - x_{\text{O-M}}^{-1/2}\right)^{2/3} \quad (12.29)$$

where $R_{\text{O-M}}$ is the cylindrical radius of the boundary between the “outer” region and the “middle” region just inside it. Note that equation (12.29) actually is a double-valued function of x , because of the factor involving $(1 - R_{\text{O-M}}/r_{\text{isco}})$, and that function has an extremum at $x = 16/9$. So,

⁴ Ten percent is representative of the accretion efficiencies of Schwarzschild and Kerr black holes (5.7–40%) computed from equation (7.54) and the values of r_{isco} given in Table 7.1.

as long as

$$\dot{m} \geq \dot{m}_M \equiv 9.3 \times 10^{-4}$$

there will be an “outer”–“middle” boundary R_{O-M} , and there will be a range in radius where the opacity in equation (12.21) will be dominated by electron scattering rather than absorption. The outer radius of this range will be the usual “outer”–“middle” boundary ($R_{O-M} \approx 2830 r_g \dot{m}^{2/3}$), while the inner radius will be close to r_{isco} , where the disk density, temperature, and thickness begin to decrease rapidly.

2. *The “Middle” Disk Region* ($p = p_g$, $\kappa = \kappa_{\text{es}}$). We also can derive the structure equations for the case when electron scattering dominates, simply by replacing equation (12.23) with

$$\kappa = 0.40/\mu_e \text{ cm}^2 \text{ g}^{-1} \approx 0.34 \text{ cm}^2 \text{ g}^{-1} \quad (12.30)$$

The structure then takes a slightly different form

ρ_c	=	4.3	g cm^{-3}	$\alpha^{-7/10}$	$m^{-7/10}$	$\dot{m}^{2/5}$	$x^{-33/20}$	$f^{2/5}$
T_c	=	2.4×10^8	K	$\alpha^{-1/5}$	$m^{-1/5}$	$\dot{m}^{2/5}$	$x^{-9/10}$	$f^{2/5}$
p_c	=	1.37×10^{17}	dyn cm^{-2}	$\alpha^{-9/10}$	$m^{-9/10}$	$\dot{m}^{4/5}$	$x^{-51/20}$	$f^{4/5}$
H	=	1.30×10^4	cm	$\alpha^{-1/10}$	$m^{9/10}$	$\dot{m}^{1/5}$	$x^{21/20}$	$f^{1/5}$
$\tau_{c,\text{es}}$	=	1.90×10^4		$\alpha^{-4/5}$	$m^{1/5}$	$\dot{m}^{3/5}$	$x^{-3/5}$	$f^{3/5}$
$\tau_{c,*}$	=	186		$\alpha^{-4/5}$	$m^{1/5}$	$\dot{m}^{1/10}$	$x^{3/20}$	$f^{1/10}$
V_R	=	-2.6×10^6	cm s^{-1}	$\alpha^{4/5}$	$m^{-1/5}$	$\dot{m}^{2/5}$	$x^{-2/5}$	$f^{-3/5}$
Σ	=	1.11×10^5	g cm^{-2}	$\alpha^{-4/5}$	$m^{1/5}$	$\dot{m}^{3/5}$	$x^{-3/5}$	$f^{3/5}$
B_c	=	1.31×10^9	G	$\alpha^{1/20}$	$m^{-9/20}$	$\dot{m}^{2/5}$	$x^{-51/40}$	$f^{2/5}$
B_Z	$\approx$	1.92×10^7	G	$\alpha^{-1/20}$	$m^{-11/20}$	$\dot{m}^{3/5}$	$x^{-49/40}$	$f^{3/5}$
V_{turb}	=	1.79×10^8	cm s^{-1}	$\alpha^{2/5}$	$m^{-1/10}$	$\dot{m}^{1/5}$	$x^{-9/20}$	$f^{1/5}$
$L_{\text{acc,M}}$	=	2.4×10^{38}	erg s^{-1}		m	$\dot{m}$		

(12.31)

Some more points are worth noting here:

- While we have changed the type of opacity in this “middle” region of the disk, the structure is remarkably similar to the “outer” region. The α and m dependencies are exactly the same, and most of the dependencies on $\dot{m}$ and on radius are very similar. Density, temperature, and pressure in both cases drop as $\sim R^{-3/2}$, $\sim R^{-1}$, and $R^{-5/2}$, respectively. The disk height H increases approximately linearly with radius, giving the disk a “flared” appearance.
- The main difference between the two models is in the optical depth: $\tau_{c,\text{ff/bf}}$ in the “outer” region has essentially *no* dependence on R , while $\tau_{c,\text{es}}$ decreases fairly rapidly with increasing radius. However, we have now introduced a new expression for optical depth

$$\tau_* \equiv (\tau_{\text{es}} \tau_{\text{ff/bf}})^{1/2} \quad (12.32)$$

which represents the *effective absorption* depth in the predominantly electron-scattering disk interior. That is, because of the scattering, a photon can random walk around in the disk interior, greatly increasing its time in the medium and increasing the probability that it will be absorbed. Equation (12.32) gives the effective optical depth to absorption when $\tau_{\text{es}} > 1$.

- The “middle” region equations are valid down to a much smaller radius than $R_{\text{O-M}}$, but they also break down when radiation pressure (equation (9.61)) begins to exceed the gas pressure, *i.e.* the region above the thin solid line in the center of the n - T plane in Fig. 9.3. (That is, all “middle” disk regions occupy the relatively narrow wedge between the dashed line in that figure and the solid line just above it.) In a standard accretion disk radiation pressure dominates inside the radius

$$R_{\text{M-I}} \approx 190 r_g (\alpha m)^{2/21} \dot{m}^{16/21} \left(1 - x_{\text{M-I}}^{-1/2}\right)^{16/21} \quad (12.33)$$

This also is a double-valued function for $R_{\text{M-I}}$, and has a solution as long as

$$\dot{m} \geq \dot{m}_{\text{I}} \equiv 0.091 (\alpha m)^{-1/8} \quad (12.34)$$

When $\dot{m} = \dot{m}_{\text{I}}$ a radiation-pressure-dominated region begins to appear at $x = 1.907$ ($R = 11.442 r_g$); and $R_{\text{M-I}}$ grows in size as $\dot{m}$ exceeds this value.

3. The “Inner” Disk Region ($p = p_r$, $\kappa = \kappa_{\text{es}}$). At accretion rates above $\dot{m}_{\text{I}}$ the central disk region near the black hole is dominated by radiation pressure and electron scattering. This actually is to be expected. The Eddington limit depends on radiation pressure and electron scattering being important, so these two physical processes should dominate as one approaches $L_{\text{acc}} = L_{\text{Edd}}$. Replacing p_g with p_r produces a significantly different disk structure than in the “middle” and “outer” regions:

ρ_c	$= 2.3 \times 10^{-5} \text{ g cm}^{-3}$	α^{-1}	m^{-1}	$\dot{m}^{-2}$	$x^{3/2}$	f^{-2}	(12.35)
T_c	$= 3.1 \times 10^7 \text{ K}$	$\alpha^{-1/4}$	$m^{-1/4}$		$x^{-3/8}$		
p_c	$= 2.4 \times 10^{15} \text{ dyn cm}^{-2}$	α^{-1}	m^{-1}		$x^{-3/2}$		
H	$= 7.4 \times 10^5 \text{ cm}$		m	$\dot{m}$		f	
$\tau_{c,\text{es}}$	$= 5.9$	α^{-1}		$\dot{m}^{-1}$	$x^{3/2}$	f^{-1}	
$\tau_{c,*}$	$= 4.6 \times 10^{-3}$	$\alpha^{-17/16}$	$m^{-1/16}$	$\dot{m}^{-2}$	$x^{93/32}$	f^{-2}	
V_R	$= -8.4 \times 10^9 \text{ cm s}^{-1}$	α		$\dot{m}^2$	$x^{-5/2}$	f	
Σ	$= 35 \text{ g cm}^{-2}$	α^{-1}		$\dot{m}^{-1}$	$x^{3/2}$	f^{-1}	
B_c	$= 1.74 \times 10^8 \text{ G}$		$m^{-1/2}$		$x^{-3/4}$		
B_Z	$\approx 1.45 \times 10^8 \text{ G}$		$m^{-1/2}$	$\dot{m}$	$x^{-7/4}$	f	
V_{turb}	$= 1.02 \times 10^{10} \text{ cm s}^{-1}$	$\alpha^{1/2}$		$\dot{m}$	$x^{-3/2}$	f	
$L_{\text{acc,I}}$	$= 2.4 \times 10^{38} \text{ erg s}^{-1}$		m	$\dot{m}$			

There are a number of interesting and strange properties of the flow in the “inner”, radiation-pressure-dominated, region:

- The effective optical depth to free–free/bound–free absorption is generally small. This means that photons are emitted in this region by the Bremsstrahlung process, but they are not re-absorbed. Instead, they rattle around in the electron-scattering disk interior and have their spectrum further modified by Compton scattering.
- Outside the ISCO the disk height is *constant with radius*, unlike the “middle” and “outer” regions. And, when $\dot{m} \geq 1$, we have $H \geq R$ near r_{isco} . This means that the assumption of a thin disk breaks down precisely when the accretion rate becomes super-Eddington. Furthermore, when $\dot{m} \geq 1$, the inward accretion velocity V_R approaches the free fall speed $V_{\text{ff}} = c/\sqrt{6} = 10^{10} \text{ cm s}^{-1}$. So, when the accretion rate becomes super-Eddington, the disk accretion gives way to spherical accretion very close to the black hole.
- Finally, the behavior of the disk surface density with increasing accretion rate $\dot{M}$ is counter-intuitive. Σ *decreases* as $\dot{m}$ increases. As we shall see below, this is a sign of a secular (viscous) instability.

The above three sets of equations comprise a self-consistent theory for optically thick disk accretion. They satisfy all the assumptions made in their derivation⁵ – thin disk, cooling by advective accretion flow is negligible, cooling by radiative heat loss is dominant, electron scattering or free–free/bound–free absorption dominates the opacity (in the appropriate regions), and gas or radiation pressure also is dominant.

Fig. 12.3 shows the standard disk model for $m = 10$, $\dot{m} = 0.1$, and $\alpha = 0.03$. Because we have assumed that one process only is active at any one point in space, the transitions between disk regions appear sharp; in reality, they are smooth, of course. Note how the disk height is approximately proportional to disk radius in the “middle” and “outer” regions.

Stability of Standard Disks

A full treatment of the stability of disk accretion flows is beyond the scope of this book. Instead, we simply will state the criteria for disk stability and use them to see if our structures above are indeed stable, as we have assumed. It will help the reader, however, to understand a little of how these stability criteria are derived. The general idea in a stability analysis is to see how the flow behaves when small amplitude waves are sent through the structure in all directions (R , Z , and even ϕ). To do this we return to the time-dependent equations (5.63) and (5.64), linearize them so we can look at small amplitude waves, and then finally solve for the “dispersion” relation, the relation between the waves’ frequency and wavelength. There are generally three types of waves – oscillating ones, damped ones, and growing ones – and the dispersion relation will tell us what kinds exist for a particular flow. Oscillating and

⁵ Except for one: the “inner” region will turn out to be unstable and not in a steady state.

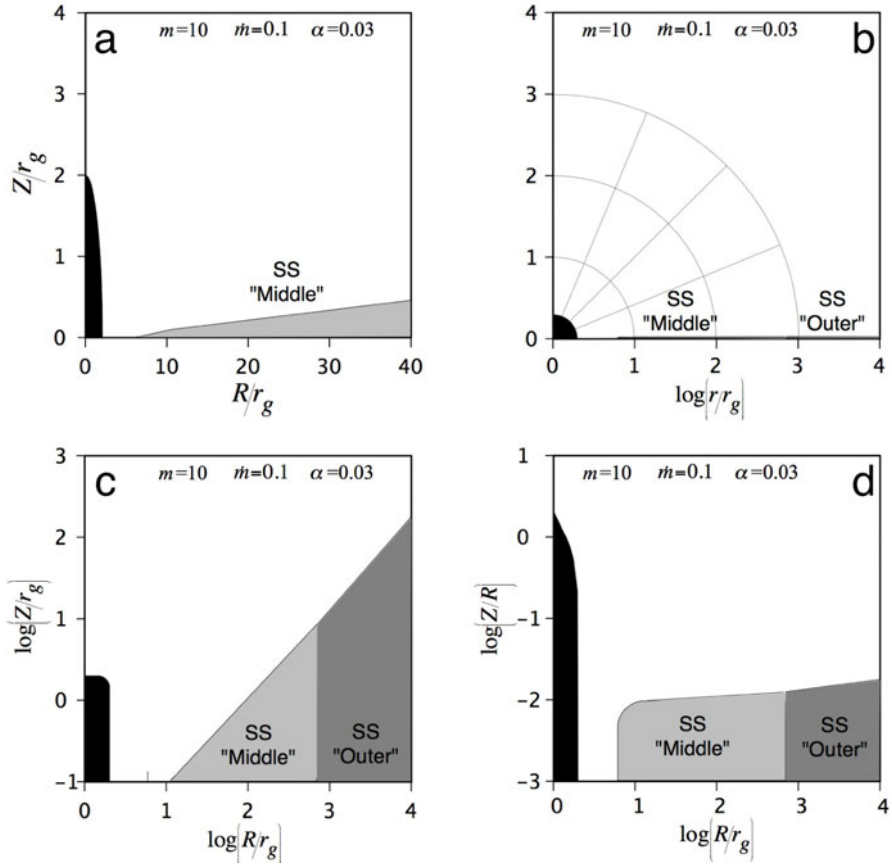


Fig. 12.3: Four different views of the same accretion disk around a $10 M_{\odot}$ black hole, with a fairly high $\dot{m} = 0.1$ and a typical viscosity parameter of $\alpha = 0.03$. (a): Linear cylindrical coordinate view, vertical axis stretched by a factor of 10; (equal axial scales would make the disk too thin to see clearly); black hole is quarter-ellipse at left; we see only the inner $40 r_g$ of the Shakura–Sunyaev “middle” disk region; “outer” region lies at $R > 675 r_g$. (b): View in $(\log r, \theta)$ space, with coordinate lines shown for emphasis; black hole is spherical in this system, and we now can see “outer” region, but the disk is only $\lesssim 0.015$ radians in half-thickness. (c): View in $(\log R, \log Z)$ space, with the vertical axis shifted up by a factor of 10; both the “middle” and “outer” disk structures are now visible, but the ISCO region lies below the abscissa; another factor of 10 shift upward would reveal the ISCO region, but move much of even the thin “outer” disk region above the upper boundary. (d): View that plots $(\log Z/R, \log R)$, scaling the vertical height of each disk annulus by its radius; here rays of constant θ are horizontal lines; an upward shift of the vertical axis by three orders of magnitude is easily accommodated, while still allowing both geometrically thin and very thick structure to be viewed over several orders of magnitude in cylindrical radius. View (a) allows a simple physical interpretation of the structure of a single model, while view (d) allows models of many different types to be compared with one set of axes.

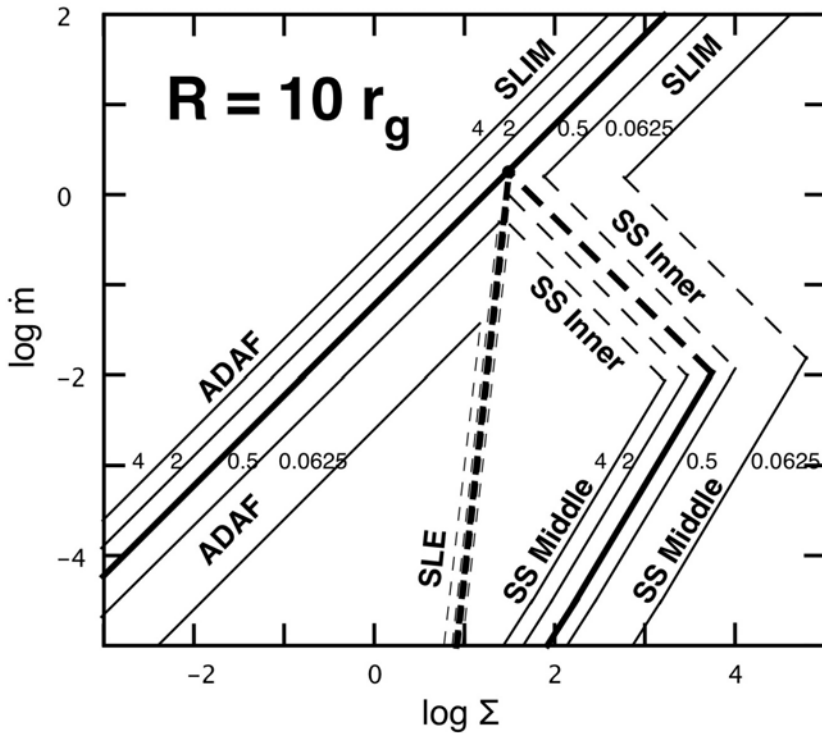


Fig. 12.4: The Σ - $\dot{m}$ plane at a disk radius of $R = 10 r_g$, built from our disk models in eqs. (12.31), (12.35), (12.47), and (12.52). The critical viscosity parameter is found by equating $\dot{m}(\Sigma)$ in the latter three equations, resulting in the intersecting thick curves for $\alpha = \alpha_{\text{crit}}$. The remaining curves (labeled 4, 2, 0.5, and 0.0625) are for different multiples of that critical value. The SLE models (short dashed lines in the center) are thermally unstable and should give way to either a hot ADAF or a cool standard (SS) thin disk (or both.) The SS “Inner” region (long dashed lines) is both thermally and secularly unstable (inverted $\dot{m}$ vs. Σ curve). The S-shaped curves can lead to limit-cycle behavior (see text and Fig. 12.5). At this radius ($R = 10 r_g$) the SS “Outer”-“Middle” boundary lies at $\log \dot{m} = -4.53$. Note, as stated in the text, that ADAFs and slim disks both have the same physical structure; only their thermal structures differ.

damped waves pose no problem for a structure. Oscillating ones (like sound waves) simply will propagate around the structure, and eventually, all waves probably will be damped at some level, turning the wave energy into heat. Growing wave modes, on the other hand, pose a potential serious problem, if the growth rate is the secular time scale of the structure or shorter. For disks the secular time scale is the accretion time

$$\tau_{\text{acc}} \sim R/V_R \quad (12.36)$$

so if waves grow on accretion, thermal or dynamical times, then the assumptions of steady accretion, thermal balance, or dynamical equilibrium that we made above were not correct. The disk structure that we have computed, then, is unstable.

In Chapters 8 and 9 we performed a wave analysis for gravitational and magnetohydrodynamic waves. In Section 9.6.3 we found that certain MHD wave modes were unstable, if rotational shear were present and the magnetic field was weak. Like this case, a stability analysis leads to simple, and often intuitive, inequalities that tell us when the structure is *stable*. In the case of disks, there are two principal types of disk instabilities that are common: secular/viscous and thermal.

Viscous Stability. In 1974, scarcely a year after the standard disk equations were published, Alan P. Lightman and Douglas M. Eardley, then at Caltech, showed that the “inner” disk region was *unstable* on the accretion time scale [433]. A numerical simulation of this instability showed that when annular regions of the disk got denser than the average, the accretion tended to slow down, making these regions even denser. Likewise, less dense regions accreted *faster*, turning them into voids.

The reason for this instability is that, as pointed out earlier, the disk surface density Σ varies *inversely* with accretion rate $\dot{m}$ (see equations (12.35)) – a behavior that is quite different from that in the “middle” and “outer” regions. In other words, the viscous stress

$$W_{\phi R} \propto \dot{M} \propto \Sigma^{-1}$$

decreases as the disk surface density *increases*. As elegant as equations (12.35) appear to be, they do not describe a stable model: they are wrong. Fig. 12.4 shows the $\dot{m}$ vs. Σ curve for the model in Fig. 12.3, evaluated at a single radius ($R = 10 r_g$). The “inner” region is plotted with a dashed line to indicate that it is unstable. Note that it is the only region that has surface density decreasing as $\dot{M}$ increases. The general criterion for secular stability of an accretion flow, then, is

$$\boxed{\frac{d \ln \Sigma}{d \ln \dot{M}} > 0} \quad (12.37)$$

The question now arises, if equations (12.35) are not the correct structure for a radiation-pressure-dominated disk, what is? We will address this in the discussion of the numerical models for thick disks below.

Thermal Stability. In 1976 Shakura and Sunyaev wrote another comprehensive paper on accretion disks, this time on their overall stability [434]. They showed that the Lightman–Eardley instability was only one branch of a spectrum of instabilities in the “inner” region. And it was not the most destructive one. A thermal instability, acting on the time scale $\tau_{\text{th}} = p H / (T - 1) Q_r$, grows much faster. Hot parts of the “inner” region will become even hotter, and cool parts even cooler.

The reason for this instability is a mismatch between the viscous heating and radiative cooling rates. The heating rate is given by

$$Q^+ = Q_{\text{acc}} = -2\alpha p, \Sigma_{\phi R} H = \frac{1}{2} \left(\frac{GM}{R^3} \right)^{1/2} \alpha a T_c^4 H$$

while the cooling rate is

$$Q^- = Q_r \approx \frac{2ac}{\kappa_{\text{es}} \Sigma} T_c^4$$

At a given radius R , the surface density Σ is constant on the (short) thermal time scale, and hydrostatic equilibrium ($p \propto H^2 \rho \propto H \Sigma \propto H$) is maintained. Therefore, heating varies with temperature as $Q^+ \propto p^2 \propto T_c^8$, while cooling varies as $Q^- \propto T_c^4$. This is an unstable situation: only one solution exists for T_c at any one point in the disk and, if the temperature increases slightly from that value, heating will be *faster* than cooling.

The general criterion for thermal stability is

$$\boxed{\frac{d \ln Q^+}{d \ln T} < \frac{d \ln Q^-}{d \ln T}} \quad (12.38)$$

so that, if the gas becomes hotter than the equilibrium temperature, it will cool faster than it will heat; and, conversely, it will heat up if it goes below the equilibrium temperature. Since $8 > 4$, however, the “inner” region of the standard disk model is thermally unstable. This is entirely the fault of the pressure being dominated by radiation and the assumption that the viscous stress is proportional to that pressure. For the “middle” and “outer” regions, the cooling remains radiative and still varies as $Q^- \propto T^4$. However, because the pressure is dominated by gas pressure, the heating varies as follows: $Q^+ \propto p H \propto \rho T_c H \propto \Sigma T \propto T$. Since $1 < 4$, the “middle” and “outer” regions are thermally stable according to equation (12.38).

Numerical Models of Unstable Standard Disks

Unstable situations often lead to some time-dependent state, such as turbulence, cyclic behavior, or explosion. Sometimes they lead to a completely different equilibrium structure, such as collapse from a compact stellar core to a neutron star or black hole. The best technique for determining what the final outcome is of an instability is to follow the *nonlinear* behavior to its conclusion, and that usually can be done only through numerical modeling or simulation.

In the 1990s two groups, one in Japan (Fumio Honma and Shoji Kato of Kyoto University and Ryoji Matsumoto of Chiba University) and one in Europe (Ewa Szuszkiewicz of the University of Szczecin in Poland and John C. Miller of SISSA in Trieste, Italy), performed numerical modeling of unstable accretion disks that had a substantial “inner” region. These time-dependent models improved on the Lightman–Eardley models by including the evolution not only of the surface density, radial velocity, and angular momentum (the secular evolution), but also of the entropy of the plasma at each radius in the disk (the thermal evolution). These numeri-

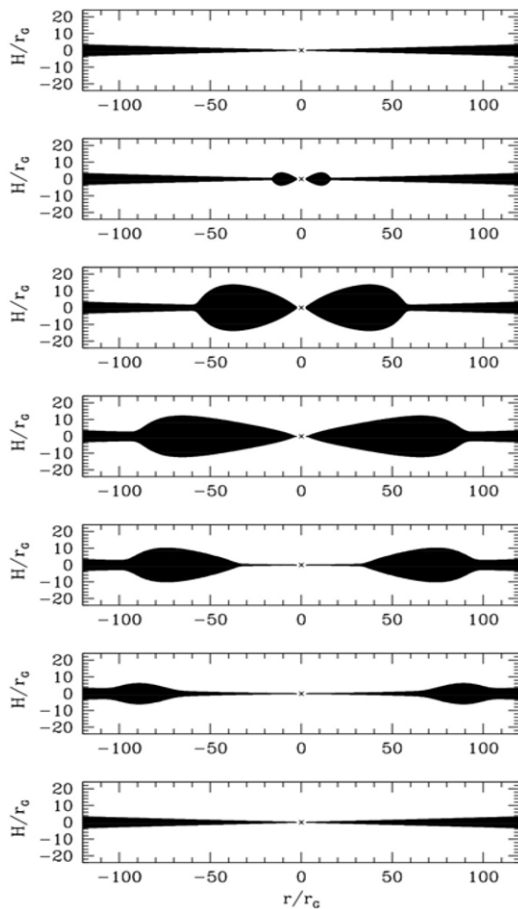


Fig. 12.5: Example of both the secular and thermal instabilities in the “inner” region of a standard α -model accretion disk. Parameters assumed are $M = 10 M_{\odot}$, $\dot{m} = 0.06$, and $\alpha = 0.1$. The bloating of the accretion disk in the vertical direction is due to the thermal instability, while the overall limit-cycle behavior (from top to bottom and then back to top again) is produced by the secular instability itself. From top-to-bottom the panels occur at the following relative times: 0s, 2s, 8s, 16s, 18s, 22s, and 787s; *i.e.*, the draining of the bloated torus (in which the object is in a temporary ADAF-like substate) occurs in the first ~ 30 seconds or so, with the recovery of the thin disk structure taking the remaining ~ 12.5 minutes of the cycle. (The entire cycle time is several times longer than we would have predicted from the analytic theory in eqs. (12.33) and (12.35), but still within the errors in the analytic theory.) We might expect a strong jet to be ejected during the ADAF-like period (2–16 s or $< 2\%$ duty cycle) while the source would remain in a relatively radio-quiet substate for $> 98\%$ of the cycle. Reproduced from Fig. 1 in [238], by permission of J. Wiley and Sons, Ltd.

cal models were not as sophisticated as the two-dimensional and three-dimensional general relativistic ones of colliding black holes or neutron stars, discussed in Chapters 8 and 10, or of MHD accretion disks discussed below. They assumed only one-dimensional flow in the disk radial R direction, symmetry in the azimuthal ϕ direction, and used simple one-zone equations in the vertical direction to obtain the disk height and radiative transfer. And, rather than use a full Schwarzschild background metric, the investigators used Newtonian dynamics and a “pseudo-relativistic” gravitational potential that mimicked the behavior of a black hole

$$\psi = \frac{GM}{r - r_S} \quad (12.39)$$

which was suggested as an alternative to real general relativistic dynamics by Bodhan Paczyński and Paul Wiita in 1980 when they both were at the University of Warsaw [435]. However, these numerical models are excellent for demonstrating the thermal instability of standard disks that we discussed here and for showing what the nonlinear result of that instability should be.

Fig. 12.5 shows the evolution of the “inner” region of an otherwise standard Shakura–Sunyaev disk. The accretion rate in this model is $\dot{m} = 0.06$, $\alpha = 0.1$, and the black hole mass is $10 M_\odot$. We can see clearly that the thermal instability dominates, but the viscous one does have some effect also. When the model begins to evolve, the disk begins to heat up near the black hole, with the entire “inner” region doubling in temperature (to $\sim 3 \times 10^7$ K) and bloating into a hot torus (Panel 2). As the evolution proceeds, the disk surface density Σ drops in the bloated region, indicating that the viscous instability is also at work (Panels 3–4): the hot, low-density part of the disk has begun to accrete faster, while matter piles up just outside of that region, where the accretion rate has now slowed. Eventually (Panels 5–6), the innermost regions of the disk cool and become thin again, and this cooling wave propagates outward until the disk appears thin everywhere again (Panel 7). After enough material accretes into the innermost region again, trying to set up a steady radiation-pressure-dominated disk, the cycle begins again. The total cycle time in this model is about 790 s (~ 13 min).

Unstable Evolution on More than One Time Scale

This limit cycle behavior in the theoretical models is remarkably similar to that seen in the X-ray source GRS 1915+105 when it is in a high-luminosity, oscillatory state (Section 3.5.2). The cycle time there is 10–20 min, and the black hole mass is about $14 M_\odot$. While GRS 1915+105 also has many other behaviors at other times, it seems quite possible that sometimes it develops a secularly and thermally unstable “inner” region whose viscous stress is given approximately by αp_r . Furthermore, for some reason, each of these cycles is accompanied by the ejection of a strong jet as well. We will discuss this more in Chapter 16 after we learn more about jet production.

A similar limit cycle behavior also has been seen in the radio Seyfert galaxy 3C 120 (see Section 3.5.2), which has a cycle time of about 0.5–1 year and a black

hole mass of $m \approx 3.1 \times 10^7$. As with GRS 1915+105, the accretion rate in 3C 120 appears to be near the Eddington limit ($\dot{m} \sim 0.27$). And, furthermore, 3C 120 ejects a jet during every cycle as well. It seems plausible, therefore, that these two sources, differing in mass by a factor of two million, nevertheless are experiencing the same type of instability in their disk “inner” regions.

Unfortunately, things do not appear as simple as the application of a single formula. The main problem is predicting the length of each cycle. Theoretically, the duration of each full cycle should be the time it takes for the disk to fill up the entire radiation-pressure-dominated “inner” region so the innermost part can begin to heat up again. The time for this to occur is the accretion time at the “middle”–“inner” boundary R_{M-I}

$$\tau_{\text{acc}} = 7.0 \text{ s } \alpha^{-2/3} m^{4/3} \dot{m}^{2/3} \quad (12.40)$$

For $\alpha = 0.03$ and $\dot{m} = 0.3$, this gives $\tau_{\text{acc}} \approx 35 \text{ s } m^{4/3}$. So, for $m = 14$, our predicted formula gives a cycle time of $\tau_{\text{cyc}}^{\text{pred}} \sim 1200 \text{ s}$, which is close to the $\sim 1000 \text{ s}$ is observed. (This is why the numerical models gave such good agreement with the observed cycle time for GRS 1915+105.) However, for 3C 120, the formula gives a cycle time of $\sim 11,000 \text{ yr}$ – four orders of magnitude longer than the actual cycle time for this object.

What could be the problem? The solution may lie in recognizing that another, shorter, time scale is at work here – the thermal time scale. This is given by dividing the thermal energy content per unit area ($\sim 3pH$) by the viscous heating rate Q_{acc}

$$\tau_{\text{th}}^{\text{max}} = 0.0148 \text{ s } \alpha^{-6/7} m^{8/7} \dot{m}^{8/7} \quad (12.41)$$

For $\alpha = 0.03$ and $\dot{m} = 0.3$, this gives $\tau_{\text{th}}^{\text{max}} \approx 0.076 \text{ s } m^{8/7}$. So, for $m = 14$, the thermal time is only 1.5 s , which may represent rapid variations in the GRS 1915+105 X-ray emission (QPOs?), but not the jet outburst cycle time. On the other hand, for 3C 120, $\tau_{\text{th}}^{\text{max}} \sim 0.88 \text{ yr}$, which is very close to the observed $0.75 \pm 0.25 \text{ yr}$ cycle time. In other words, the black hole in 3C 120 may be so large that each half-year or so we observe quasi-periodic variations in the accretion flow on a “very short” thermal time scale. Thermal instabilities could occur throughout the “inner” region, which extends down to the ISCO ($x = 1$), where

$$\tau_{\text{th}}^{\text{min}} = 1.47 \times 10^{-4} \text{ s } \alpha^{-1} m \quad (12.42)$$

So, for 1915+105, the thermal instabilities could produce variations in the range $0.7\text{--}15 \text{ Hz}$, while 3C 120’s variations would be on time scales of 2 days to 0.9 yr .

Later in the chapter we shall discuss more sophisticated simulations of MHD-turbulent accretion flows, in which the angular momentum transport occurs naturally in the flow, rather than having to assume an α . Unfortunately, even after more than 20 years, these simulations are only beginning to be sophisticated enough to study the viscous mechanism when radiation pressure dominates. Much more progress is needed in this area, and this field is still wide open for new results to be discovered.

12.2.2.2 Thermal Disk Structure #2: Optically Thin, Geometrically Thick, Two-Temperature Radiative Disks

Standard, if sometimes unstable, disks appear to be reasonable models for X-ray sources when they are in a relatively rare high-luminosity state, like GRS 1915+105. However, they do not explain the relatively common, very hot (~ 100 keV), steep-spectrum sources like Cygnus X-1 displays in its hard state. The first attempt to rectify this deficiency unfortunately resulted in another thermally unstable model. Nevertheless, it set the stage for the development of more successful models and, therefore, remains one of the pioneering and important efforts in accretion theory.

The Two-Temperature Energy Equation

In 1975, based on the Cyg X-1 soft and hard state X-ray spectra, Kip Thorne and Richard Price suggested that the source contained two emitting regions: (1) a cool, optically thick (*to absorption*), geometrically thin disk like that of the SS models and (2) a hot, optically thin, geometrically thick torus inside about $50 r_g$ [436]. The following year, Stuart Shapiro of Cornell University, along with Lightman and Eardley, had developed a disk model for the “inner” region in which the electron temperature T_e and ion temperature T_i were *unequal* [437]. Two-temperature plasmas are not uncommon, even in less extreme solar system and terrestrial situations, so the proposal was quite plausible. The ions were assumed to be heated by the viscous processes and, therefore, quite hot – up to the virial temperature (equation (12.1)), which can be 10^{10-12} K. Electrons, on the other hand, cool very efficiently when they become relativistic ($T \geq 5.9 \times 10^9$ K ≈ 511 keV), particularly by the inverse Compton effect. So their temperature should remain near or below this cooler value. As long as the thermal coupling between hot ions and the cooling electrons is weak, the plasma can remain a two-temperature one. Weak thermal coupling requires low densities, which occur naturally when the disk becomes very hot and bloated into an optically thin, gas-pressure-dominated torus, rather than a geometrically thin, optically thick radiation-pressure-dominated disk.

The equations for this “SLE” two-temperature disk model are similar to those for the “middle” and “outer” regions, except that there are two cooling processes:

1. Coulomb collisions between the ions and electrons, which cool the ions and transfer heat to the electrons. The ion cooling rate is then

$$Q_i^- = \frac{3}{2} \frac{\rho H \nu_E k}{m_p} (T_i - T_e) \quad (12.43)$$

where

$$\nu_E = 2.4 \times 10^{21} \ln \Lambda \rho T_e^{-3/2} \quad (12.44)$$

(all quantities in ν_E are in CGS units) is the ion–electron collision rate. As in previous chapters (see Sections 10.4.2, 11.1.3, and 11.1.4),

$$\ln A = \ln \left(\frac{1}{n r_{coul}^3} \right)^{1/3} = 40\text{--}50$$

is the Coulomb logarithm for typical optically thin accretion flows, n is the particle number density, and $r_{coul} = e^2/kT$ is the ion–electron impact radius. We will take $\ln A \approx 45$.

2. Radiative inverse Compton cooling of the electrons at the rate

$$Q_e^- = (\rho H \kappa_{es} U_r c) \left(\frac{4kT_e}{m_e c^2} \right) \quad (12.45)$$

The first factor in equation (12.45) is the maximum photon energy flux out of the disk, and the second is the fraction of that lost to the inverse Compton process.

Equilibrium between heating and cooling occurs when $Q_e^+ = Q_i^- = Q_e^-$.

The Source of the Cooling Photons and the Resulting Photon Spectrum

The key to the SLE model is in the source of the soft photons that are Comptonized, which then cools the plasma. They are proposed to come from an *external* photon source (*e.g.*, the cooler “middle” and “outer” regions of the disk). The soft photons do not come from the emission within the plasma itself (*i.e.*, not from Bremsstrahlung or synchrotron). This means that, in equilibrium, the photon energy density in equation (12.45) is related to the electron scattering optical depth τ_{es} (rather than to the effective absorption optical depth τ_*) as

$$U_r = \frac{Q_r}{c} \max(1, \tau_{es}) = \frac{Q_e^-}{c} \max(1, \tau_{es})$$

Equation (12.45) then becomes, simply

$$y \equiv 4 \theta_e \max(\tau_{es}, \tau_{es}^2) = 1 \quad \theta_e \equiv \frac{kT_e}{m_e c^2} \quad (12.46)$$

The parameter y is the well-known Compton parameter. A value of $y \approx 1$ indicates that Comptonization is unsaturated, and the spectrum is a steep power law with an exponential cutoff $\propto \exp(-h\nu/kT_e)$. Assuming $T_i \gg T_e$ and $\tau_{es} > 1$, equations (12.7), (12.9), (12.10), (12.16), (12.22), (12.43), and (12.46) have the solution

$$\begin{array}{llllllll}
\rho_c & = & 2.1 \times 10^{-5} & \text{g cm}^{-3} & \alpha^{2/3} & m^{-1} & \dot{m}^{-1/3} & x^{-1} & f^{-1/3} \\
T_{\text{ic}} & = & 7.4 \times 10^{11} & \text{K} & \alpha^{-10/9} & & \dot{m}^{8/9} & x^{-4/3} & f^{8/9} \\
T_{\text{ec}} & = & 5.5 \times 10^7 & \text{K} & \alpha^{-2/9} & & \dot{m}^{-2/9} & x^{1/3} & f^{-2/9} \\
p_c & = & 2.1 \times 10^{15} & \text{dyn cm}^{-2} & \alpha^{-4/9} & m^{-1} & \dot{m}^{5/9} & x^{-7/3} & f^{5/9} \\
H & = & 7.3 \times 10^5 & \text{cm} & \alpha^{-5/9} & m & \dot{m}^{4/9} & x^{5/6} & f^{4/9} \\
\tau_{\text{c,es}} & = & 5.2 & & \alpha^{1/9} & & \dot{m}^{1/9} & x^{-1/6} & f^{1/9} \\
V_R & = & -8.2 \times 10^9 & \text{cm s}^{-1} & \alpha^{-1/9} & & \dot{m}^{8/9} & x^{-5/6} & f^{-1/9} \\
\Sigma & = & 30 & \text{g cm}^{-2} & \alpha^{1/9} & & \dot{m}^{1/9} & x^{-1/6} & f^{1/9} \\
B_c & = & 1.62 \times 10^8 & \text{G} & \alpha^{5/18} & m^{-1/2} & \dot{m}^{5/18} & x^{-7/6} & f^{5/18} \\
B_Z & \approx & 1.33 \times 10^8 & \text{G} & \alpha^{-5/18} & m^{-1/2} & \dot{m}^{13/18} & x^{-4/3} & f^{13/18} \\
V_{\text{turb}} & = & 1.00 \times 10^{10} & \text{cm s}^{-1} & \alpha^{-1/18} & & \dot{m}^{4/9} & x^{-2/3} & f^{4/9}
\end{array} \tag{12.47}$$

Note that here we assume $\tau_{\text{es}} > 1$, whereas Shapiro, Lightman, and Eardley assumed $\tau_{\text{es}} < 1$. Low electron scattering optical depths occur only for very low accretion rates ($\dot{m} < 3.8 \times 10^{-7} \alpha^{-1} x^{3/2}$). The SLE solution has a number of nice properties:

- The electron temperature T_e is near 10^9 K, especially for $\alpha = 0.01$ and $\dot{m} = 0.01$, and it varies slowly with the radius x .
- The assumption of unsaturated Comptonization ($y = 1$) yields a self-consistent model. This gives credence to the idea that the power law spectra of X-ray sources may indeed come from soft photons that are re-processed by a hot, optically thin Comptonizing cloud.
- As with the one-temperature, thin disk model, H/R approaches unity as $\dot{m} \rightarrow 1$

$$\frac{H}{R} = 0.82 \alpha^{-5/9} \dot{m}^{4/9} x^{-1/6} f^{4/9}$$

However, H is nearly proportional to R , as in the “middle” and “outer” regions, so we do not have the strange increase in H/R with decreasing R as we had in the unstable “inner” region.

- While the SLE model originally was sought as a solution to the viscously unstable “inner” region, in fact the SLE solution is valid formally for *any* $\dot{m} < 1$.

Unfortunately, while the SLE model is secularly stable, it turns out to be *thermally unstable* and not very useful for modeling black hole engines. This is the reason we have not highlighted equations (12.47) with a bounding box.

Stability

The Shapiro–Lightman–Eardley disk model is stable to viscous flow (the Lightman–Eardley instability) since $d \ln \Sigma / d \ln \dot{m} = 1/9 > 0$. However, in 1976 James Pringle of the University of Cambridge, UK, showed that the SLE solution is thermally *unstable* and, therefore, cannot exist [438]. The argument proceeds as before

with the Shakura–Sunyaev “inner” region. The heating function at a given R , assuming Σ constant again, varies as

$$Q^+ = Q_{\text{acc}} \propto p H \propto T_i \Sigma \propto T_i$$

The variation of Q_i^- with T_i is a little more difficult to determine. From equation (12.43) we see that, if $T_i \gg T_e$

$$Q_i^- \propto \rho T_e^{-3/2} T_i$$

Now, hydrostatic equilibrium gives us $\rho \propto H^{-1} \propto T_i^{-1/2}$; so, equating (12.43) and (12.45) gives us $T_e \propto T_i^{1/5}$. (Since U_r is an *external* source of soft photons, it is not affected by any change in the thermal content of the “inner” part of the disk itself; therefore, U_r remains constant in this analysis.) So, this gives us the dependence of Q_i^- on T_i of

$$Q_i^- \propto T_i^{-1/2} T_i^{-3/10} T_i \propto T_i^{1/5}$$

Since

$$1 = \frac{d \ln Q^+}{d \ln T_i} > \frac{d \ln Q^-}{d \ln T_i} = \frac{1}{5}$$

the solution is, unfortunately, thermally *unstable*. Such a disk could form only for a short thermal time, but not for the weeks at a time that the hot power law X-ray spectrum remains in sources like Cyg X-1 and others.

Fig. 12.4 shows the SLE disk model in the Σ – $\dot{m}$ plane as a dashed line because it is thermally unstable, even though it is secularly stable.

12.2.2.3 Thermal Disk Structure #3: Optically Thick, Geometrically Thick (“Slim”), Advectively-Cooled Disks for Super-Eddington Flows

Highly super-Eddington accretion rates are possible only for true black holes: rather than radiating away the photons produced above the Eddington limit (and thereby producing a strong radiation-driven outflow), black holes can *swallow* the photons. However, they cannot simply be radiated into the hole; they must be dragged or “advected” into the hole along with the flow of matter. The first authors to treat this kind of advection-dominated accretion flow (or ADAF) were Mitchell Begelman and D. Meier, in 1982 [439]. They are similar to super-Eddington spherical accretion flows, but with rotation, shear, and viscous dissipation that produces heating. In 1988 Marek Abramowicz, Bożena Czerny (Copernicus Institute in Warsaw), Jean Pierre Lasota (Paris Observatory), and Ewa Szuszkiewicz developed this field into an analytic one, like Shakura and Sunyaev’s, and gave these objects the name by which they now are known: “slim” disks. Our development below will closely follow the analytic treatment of Abramowicz and his colleagues.

Radiative losses in the “inner” region (if it were stable) are given by equation (12.21)

$$Q_r = 3.1 \times 10^{33} \text{ erg s}^{-1} \text{ cm}^{-2} m^{-1} \dot{m} x^{-3} f$$

while the *advective* “losses” at each disk radius, due to heat advecting inward, are

$$\begin{aligned} Q_{\text{adv}} &= \int_{-H}^H dZ \frac{\rho V_R}{\Gamma - 1} \frac{d(p/\rho)}{dR} \approx \frac{\dot{M}}{2\pi R^2} \frac{p}{\rho} \\ &= 2.7 \times 10^{33} \text{ erg s}^{-1} \text{ cm}^{-2} m^{-1} \dot{m}^3 x^{-5} f^2 \end{aligned} \quad (12.48)$$

Because the advective flux varies steeply with accretion rate ($Q_{\text{adv}} \propto \dot{m}^3$), while the radiative flux varies much more slowly ($Q_r \propto \dot{m}$), normally radiative losses dominate. However, at the accretion rate

$$\dot{m} = 1.1 x f^2$$

the advective and radiative losses become comparable. That is, as the Eddington accretion rate is approached and then exceeded, advection of photon thermal energy toward the black hole becomes equal to, and then dominates over, the radiative losses. Just when that radiation is needed to drive out the excess amount of matter streaming toward the hole and overcome the excess photon production, those photons are dragged into the black hole, where they cannot be used to drive a strong wind.

In the study of super-Eddington accretion flows, an important point must be emphasized. The flow is sufficiently close to spherical accretion, although there may be a very geometrically thick disk-like flow, that there is great doubt that near the innermost stable circular orbit the torque suddenly goes to zero, and the disk suddenly becomes geometrically thin. Such an assumption was suspect even for the standard disks, and probably quite incorrect for very geometrically thick flows. Therefore, we will drop all factors of f for geometrically thick flows and consider only the factors of α , m , $\dot{m}$, and x in the structure equations.

Physical and Thermal Structure

The slim disk solution actually is another region of the accretion disk, inside the “inner” region. It exists only for $1.1 x < \dot{m}$, or for disk radii smaller than the trapping radius

$$R < R_{\text{tr}} \equiv 8.2 \times 10^5 \text{ cm } m \dot{m} \quad (12.49)$$

This is essentially the same trapping radius discussed in the spherical accretion section above. Interior to this radius, photons are dragged toward the black hole faster than they can diffuse out of the flow. Outside R_{tr} we have, potentially, all the other standard disk regions (“inner”, “middle”, and “outer”).

In the case of advection-dominated accretion flow, the energy equation in the slim disk region now becomes

$$Q_{\text{acc}} = Q_{\text{adv}} \quad (12.50)$$

replacing equation (12.21) in the standard disk models and equations (12.43) and (12.45) in the SLE model. Along with hydrostatic equilibrium equation (12.9), we can derive a very simple expression for the disk height

$$H = \left(\frac{3}{4}\right)^{1/2} R \quad (12.51)$$

Therefore for an advective accretion disk, $H \leq R$. And, for this high-luminosity, radiation-pressure-dominated flow, the opacity is given by $\kappa = \kappa_{\text{es}}$. The rest of the physical structure equations then follow in a manner similar to the solutions for the standard α -model disks

ρ_c	$= 2.1 \times 10^{-5}$	g cm^{-3}	α^{-1}	m^{-1}	$\dot{m}$	$x^{-3/2}$	(12.52)
p_c	$= 2.3 \times 10^{15}$	dyn cm^{-2}	α^{-1}	m^{-1}	$\dot{m}$	$x^{-5/2}$	
H	$= 7.7 \times 10^5$	cm		m		x	
$\tau_{c,\text{es}}$	$= 5.5$		α^{-1}		$\dot{m}$	$x^{-1/2}$	
V_R	$= -9.2 \times 10^9$	cm s^{-1}	α			$x^{-1/2}$	
Σ	$= 32$	g cm^{-2}	α^{-1}		$\dot{m}$	$x^{-1/2}$	
B_c	$= 1.70 \times 10^8$	G		$m^{-1/2}$	$\dot{m}^{1/2}$	$x^{-5/4}$	
B_Z	$\approx 1.47 \times 10^8$	G		$m^{-1/2}$	$\dot{m}^{1/2}$	$x^{-5/4}$	
V_{turb}	$= 1.05 \times 10^{10}$	cm s^{-1}	$\alpha^{1/2}$			$x^{-1/2}$	

and the thermal structure follows immediately from setting the pressure to $p = p_r$

T_c	$= 3.1 \times 10^7$	K	$\alpha^{-1/4}$	$m^{-1/4}$	$\dot{m}^{1/4}$	$x^{-5/8}$	(12.53)
$\tau_{c,*}$	$= 5.7 \times 10^{-3}$		$\alpha^{-17/16}$	$m^{-1/16}$	$\dot{m}^{17/16}$	$x^{-5/32}$	
$L_{\text{acc,slm}}$	$= 2.1 \times 10^{38}$	erg s^{-1}		m			

Note that, unlike the thin disk accretion models above, the slim disk luminosity remains at the Eddington limit, just like the spherical accretion models in Section 12.1.2. Most of the photons generated are not radiated outward but instead are dragged into the black hole. Otherwise, the *structure* of this advective, super-Eddington solution is rather similar to the sub-Eddington solution for the “inner” disk region (equations (12.35)). In fact, while the $\dot{m}$ and x dependencies are different from the “inner” region equations, the α and m dependencies *are the same*, and even the coefficients are almost identical. This means that at the Eddington accretion rate, and near the inner edge of the disk ($\dot{m} = x = 1$), the “inner” region solution and the slim disk solutions *match*. Therefore, *the slim disk solution is a super-Eddington accretion flow interior to the radiative “inner” disk solution*. While the slim disk solution is formally stable (because advective energy losses dominate), the “inner”

disk solution (now *exterior* to the slim disk flow) still is limit-cycle unstable because radiative losses dominate there. Therefore, the slim disk itself may be slaved to the unsteady behavior of the “inner” disk region, appearing and disappearing with the ebb and flow of the unstable region outside it.

When Does Advection of the Magnetic Field Become Important?

In all the above subsections, we used equation (12.17) to derive a representative magnetic field strength from the pressure. However, this may not be the full story in the case of advection-dominated flow. In such a flow the gas is compressed and dragged inward faster than it can radiate its energy and change its internal energy content. Advection can have a similar effect on each magnetic field component, if the inflow is so rapid that magnetic flux is conserved. For B_ϕ and B_R the magnetic induction equation (9.155) give us, for an advective flow,

$$\begin{aligned} V_R B_\phi H &= \text{constant} \\ B_R R H &= \text{constant} \end{aligned}$$

With $H \propto R \propto x$, these relations predict individual relations for B_ϕ and B_R

$$\begin{aligned} B_\phi &\propto m^{-1/2} \dot{m}^{1/2} x^{-1/2} \\ B_R &\propto m^{-1/2} \dot{m}^{1/2} x^{-2} \end{aligned}$$

that are not the $B_\phi \propto B_R \propto x^{-5/4}$ that we get from a fully turbulent structure.

In the advection-dominated flow described here, is the magnetic field advection-dominated in addition to the internal energy? The answer lies in which process is stronger: turbulence, or advective flow? In all accretion flows described so far (equations (12.26), (12.31), (12.35), and (12.47)), the ratio $|V_R|/V_{\text{turb}}$ is less than unity if $\alpha \leq 1$, $m \geq 1$, $\dot{m} \leq 1$, and $x \geq 1$. So, turbulence dominated there, and the assumption $B_\phi \sim B_R$ was acceptable. In the present advective case, we see that $|V_R|/V_{\text{turb}} \approx \alpha^{1/2} \leq 1$ also, even in the presence of the super-Eddington ($\dot{m} > 1$) accretion that is assumed here. So, even for these internal energy advection-dominated accretion flows, the turbulent turnover (and rotational shear) is so fast that it can create new B_ϕ from B_R in the turbulence (and new B_R from B_ϕ) and bring the magnetic field components into approximate equipartition. Only if we had $|V_R| \gg V_{\text{turb}}$ would we need to take into account advection of magnetic flux.

Viscous and Thermal Stability

The first thing we note about the slim disk solutions is that they are viscously stable ($\Sigma \propto \dot{m}$). Fig. 12.4 connects these super-Eddington solutions to the Shakura–Sunyaev standard disk ones and shows that they form a well-known S-shaped curve. S-shaped curves for equilibrium structures often create an unstable behavior differ-

ent from that described for the thermal limit cycle in the previous section. If the accretion rate were in the unstable range ($0.08 < \dot{m} < 1$), then material would collect at the outer edge of the unstable region: at that point, therefore, Σ will increase with time and $\dot{m}$ will *decrease*. Suddenly, when Σ exceeds the value at the right-most turning point (here $\log \Sigma \approx 4.6$), the solution will *jump* to the upper stable branch and start accreting *rapidly*. As the reservoir of material is depleted, $\dot{m}$ and Σ will decrease down the stable branch until the disk reaches the left-most turning point (here at $\log \Sigma \approx 3$). The surface density then will increase once again along the unstable branch, and the cycle will repeat. Early numerical models by Lightman and Eardley, which did not include the possibility of a thermal instability, behaved in this manner. A viscous instability like this one, which is not supplanted by a faster thermal instability, may be responsible for many X-ray nova outbursts (see below). These outbursts appear to reach super-Eddington accretion X-ray luminosities for a short period of time and then return to a quiescent state.

Slim disks are also thermally stable. The heating varies with temperature as

$$Q^+ = Q_{\text{acc}} \propto p H \propto p^2 \propto T^8$$

(The last step above assumes hydrostatic equilibrium, equation (12.9)). For the cooling, we cannot assume $\dot{M}$ constant in the advection-dominated case. (We can do so only on time scales longer than the viscous time.) So we have instead

$$Q^- \propto H V_R p \propto p H^3 \propto p^4 \propto T^{16}$$

Since $d \ln Q^+ / d \ln T < d \ln Q^- / d \ln T$, the above slim disk solutions are thermally stable.

With the addition of the advection-dominated, slim disks we now have a fairly complete description of accretion at high accretion rates (say, $\dot{m} > 0.1$): the standard, Shakura–Sunyaev disks (with a limit cycle governing the “inner” radiation-pressure-dominated region) for $0.2 \lesssim \dot{m} < 1$, and the slim disks describing the accretion for $\dot{m} > 1$. These models display properties similar to those of X-ray sources in the very high/intermediate (and unstable) state, high/soft state, and the transient super-Eddington state. What is still missing, however, is a description of accretion at low accretion rates ($\dot{m} < 0.1$) that is not only stable, but also hot and optically thin, as the observations suggest.

12.2.2.4 Thermal Disk Structure #4: Optically Thin, Geometrically Thick, Advectively-Cooled (“ADAF”) Disks for Very *Sub-Eddington* Flows

The solution to producing a successful accretion disk theory at low accretion rates was first suggested in 1977 by Setsuo Ichimaru of the University of Tokyo [440]. He proposed to stabilize the two-temperature SLE disk solution with advective cooling, and this suggestion proved highly successful. Ichimaru’s paper pre-dated, by almost 20 years, the later work on such flows that became popular for low accretion rates in the mid-1990s, and even pre-dated the Begelman and Meier 1982 paper on ad-

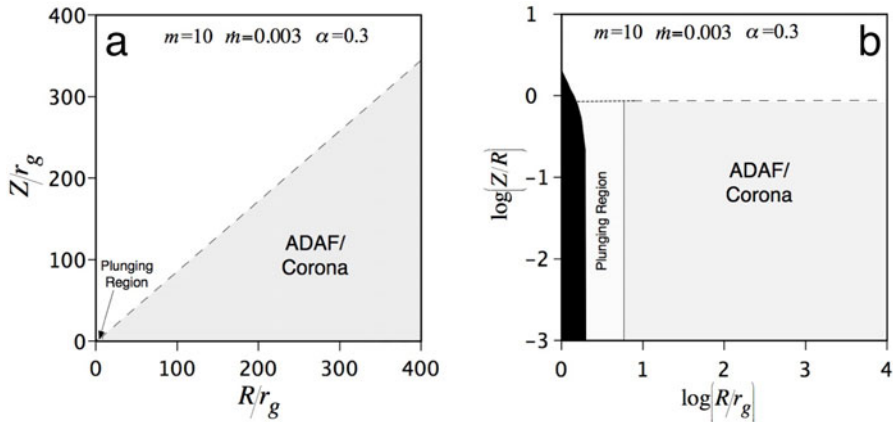


Fig. 12.6: Two different views of the same advective accretion flow around a $10 M_{\odot}$ black hole, with a very low $\dot{m} = 0.003$ and a typical viscosity parameter of $\alpha = 0.3$. (a): Linear cylindrical coordinate view over $400 r_g$ in radius; equal axial scaling is now reasonable, because height of inflow is of order its radius; however, black hole and plunging region now are almost too small to be seen; Fig. 12.3a occupies only the inner 0.1% of the area of this figure. (b): View in $(\log Z/R, \log R)$ space; same axes as Fig. 12.3d allow us to directly compare both the geometrically thin Shakura–Sunyaev disks and thick advective disks directly, over four orders of magnitude in radius. Note that the physical structure of a “slim” disk (*e.g.*, $\dot{m} = 3.0$) will be very similar (see eqs. (12.52)), but the inflow will be very optically thick (to both electron scattering and even absorption) and highly radiation-pressure-dominated (compare eqs. (12.53) and (12.58)).

vective flow at high accretion rates by 5 years [439]. However, an advective solution to low-luminosity accretion did not become widely accepted until well *after* the high-luminosity, slim disk advective flows had been well studied and thoroughly discussed in the literature.

The concept of advective flow solving the SLE stability problem was considerably expanded and placed into the broader accretion picture by Ramesh Narayan and Insu Yi of Harvard University, and by the many students and postdoctoral fellows that followed [441]. They identified when the advective flow was valid, put considerable effort into calculating detailed spectra and comparing that with observational data, and widely promoted the model and its capabilities. Our discussion below follows their approach, and to some extent that in a 1998 review by Narayan, Rohan Mahadevan, and Eliot Quataert, all at Harvard at the time [442]. These low-luminosity advective flows have become known as “advection-dominated accretion flows” or ADAFs, and this term is applied *exclusively* to ones with $\dot{m} \ll 1$. On the other hand, the (equally advection-dominated) flows with $\dot{m} > 1$ that were discussed above always are referred to as “slim disks”, because this is the term used historically and because it helps distinguish high-luminosity advection from low-luminosity advection.

Physical Structure

The physical structure of all advection-dominated flows is independent of each flow's thermal structure. See Fig. 12.6 for sample views of an advective disk at very low accretion rate ($m = 10$, $\dot{m} = 0.003$, $\alpha = 0.3$). Therefore, expressions for most quantities (*i.e.*, ρ_c , p_c , H , $\tau_{c,es}$, V_R , Σ , B_c , and V_{turb}) in equations (12.52) remain the same for these low accretion rate flows. That is, in order to compute ρ_c , for example, we simply set $\dot{m}$ to a number well below unity, rather than above unity in those equations.

The Valid Accretion Range and Efficiency for ADAFs

The thermal structure of ADAFs is quite different from that of slim disks, however, and this is due chiefly to the lower density ρ_c for lower $\dot{m}$. The low density decreases the opacity, creating very optically thin accretion that is no longer dominated by radiation pressure or electron scattering. Instead, we return to gas pressure and free-free/bound-free opacity domination (equations (12.22) and (12.23)). The ADAFs retain a key feature of the SLE solution: the two-temperature thermal structure. Again, the ions are assumed to be heated by viscous processes and cooled by infrequent collisions with electrons. And the electrons are assumed to be cooled by thermal synchrotron or Bremsstrahlung emission, which, in turn, may be enhanced by Comptonization of *locally*-produced photons.

Valid Accretion Rates for ADAFs. For what accretion rates is the ADAF solution valid? Determining this involves comparing the advective cooling time to the competing cooling time. In the case of the slim disks, the competing cooling time was the radiative time, and we obtained $\dot{m} > 1$ as the valid range for slim disks. In the case of hot, two-temperature disks, the competing cooling time is the ion-electron coupling time ν_E^{-1} from equation (12.44). Setting this equal to the advection time scale (equation (12.36)), we find that advection dominates over ion collisional heat losses when

$$\dot{m} < \frac{45}{\ln \Lambda} \alpha^2 \left(\frac{T_e}{3.6 \times 10^9 \text{ K}} \right)^{3/2}$$

Detailed numerical models give a result close to

$$\dot{m} < \dot{m}_A \equiv \alpha^2 \quad (12.54)$$

So, advective cooling of the ions dominates only for low accretion rates, which is consistent with observations of X-ray sources in a low/hard state.

Equation (12.54) actually allows us to estimate the viscosity parameter α itself. Cygnus X-1 makes a transition between its high/soft state and its low/hard state when it is accreting at about 10% the Eddington luminosity, *i.e.*, when $\dot{m} = 0.1$. Equation (12.54) then tells us that $\alpha \approx 0.3$, which is considerably larger than the

~ 0.01 values obtained from both observations and simulations of thin disks. We will discuss why this might be the case when we discuss the numerical simulations in Section 12.4.

Radiative Accretion Efficiency for ADAFs. According to equations (12.22) and (12.52) the ion temperature must remain approximately virial ($\propto R^{-1}$). However, a number of arguments⁶ indicate that the electron temperature must remain approximately constant at

$$T_e \approx 10^9 \text{ K} \quad (12.55)$$

or $\theta_e \sim 0.17$. This means that the ADAF becomes a two-temperature structure for

$$R < R_A = 7.3 \times 10^8 \text{ cm } m \quad (12.56)$$

or $x < x_A = 820$. An estimate of the accretion efficiency of an ADAF then can be computed from the amount of energy transferred to electrons by ion collisions divided by that lost down the black hole

$$\epsilon_{\text{acc,adf}} = \frac{L_{ie}}{\dot{M} c^2} \approx 2.8 \frac{\dot{m}}{\dot{m}_A} \epsilon_{\text{acc}} \quad (12.57)$$

where $\epsilon_{\text{acc}} = 0.1$ is our chosen accretion efficiency for standard Shakura–Sunyaev α -model disks, and

$$L_{\text{acc,adf}} = \int_{r_{\text{isco}}}^{R_A} 2 \pi Q_i^- R dR \approx 3.5 \times 10^{38} \text{ erg s}^{-1} \alpha^{-2} m \dot{m}^2$$

This is a famous result: *the accretion luminosity of an ADAF is not proportional to the accretion rate; it is proportional to that rate squared*. Combined with relation (12.54) this implies that an inefficient ADAF dominates the accretion luminosity for low $\dot{m}$ ($< \alpha^2$) while an efficient Shakura–Sunyaev-type thin accretion disk dominates for high $\dot{m}$ ($> \alpha^2$).

At this point it is fair to ask, “If only a fraction $\dot{m}/\alpha^2$ of the accretion luminosity is radiated, what happens to the rest of the binding energy liberated in the accretion flow?” In an ADAF the answer is that the energy remains in two non-radiative forms: thermal energy of the hot ions and kinetic energy of the inflow. *These energies are then absorbed as mass by the black hole when the inflow enters the horizon*. This is a consequence of the advection process dominating the ion–electron energy exchange rate.

⁶ For example, derivation of relation (12.54) from the inequality preceding it implies $\theta_e \sim 0.6$. And equating the ion–electron, Coulomb coupling rate (equation (12.43)) with an approximate expression for synchrotron cooling

$$Q_{\text{sync}} \approx 3.2 \times 10^{10} \text{ erg cm}^{-2} \text{ s}^{-1} \rho H p_{\text{mag}} \theta_e$$

gives $\theta_e \sim 0.18 \alpha^{-2/5}$.

Two-Temperature Thermal Structure

Equations (12.22), (12.23), and (12.55) give us the new thermal structure of a low-accretion-rate ADAF

$$\begin{array}{llll}
 T_{c,i} & = & 8.2 \times 10^{11} \text{ K} & x^{-1} \\
 T_{c,e} & = & 10^9 \text{ K} & \\
 \tau_{c,ff/bf} & = & 3.3 \times 10^{-11} & \alpha^{-2} m^{-1} \dot{m}^2 x^{-2} \\
 L_{acc,adf} & = & 3.5 \times 10^{38} \text{ erg s}^{-1} & \alpha^{-2} m \dot{m}^2
 \end{array} \tag{12.58}$$

There are several important points to note from equations (12.58):

- We are back to using $\tau_{ff/bf}$ rather than τ_* , because the electron scattering optical depth τ_{es} generally is less than unity in the ADAF. Also, because the plasma is optically thin to electron scattering and to free–free/bound–free absorption, the temperature of the radiation emitted is generally the same as the central electron temperature $T_{c,e}$.
- The central “disk” ion temperature is essentially the virial temperature – all the way down to the last stable orbit and beyond. This extremely hot ($\leq 10^{12}$ K) plasma is needed to keep the flow geometrically thick ($H \sim R$) against the tremendous tidal squeezing forces of the black hole. If the ion cooling losses ever were to become more efficient than Coulomb scattering inside some radius, then the entire ADAF structure would collapse inside that radius.
- Equations (12.52) and (12.58) show that the ADAF remains optically thin to both scattering and absorption ($\tau_{es} < 1$ and $\tau_{ff/bf} \ll 1$). The electron scattering depth becomes of order unity only very near the last stable orbit. Photons are never absorbed, only possibly Compton scattered by the free electrons.

In an ADAF model T_e is independent of black hole mass. Therefore, the hot, optically thin, Comptonized X-ray spectrum should be similar in both X-ray binaries and in quasars. To first order, this is indeed the case. The ADAF model, therefore, successfully explains the main properties of black holes accreting at rates well below the Eddington limit.

Stability

Because their physical structure equations are identical, the ADAFs, like the slim disks, are viscously stable. We have plotted the Σ – $\dot{m}$ line for ADAFs and slim disks in Fig. 12.4.

ADAFs also are thermally stable. The ion heating goes as

$$Q_i^+ = Q_{acc} \propto p H \propto \rho T_i H \propto T_i$$

while the ion cooling goes as

$$Q_i^- \propto H V_R p \propto p H^3 \propto p^2 H^2 \propto \rho^2 T_i^2 H^2 \propto T_i^2$$

Since $1 = d \ln Q_i^+ / d \ln T_i < d \ln Q_i^- / d \ln T_i = 2$, the ions in the ADAFs are thermally stable.

It also is interesting to see if the electrons are stable as well. Changes in T_e do not affect the disk structure, so the electron heating has the following electron temperature dependence

$$Q_e^+ \propto \nu_E \propto T_e^{-3/2}$$

while the synchrotron cooling varies as

$$Q_e^- \propto T_e$$

So, since $-3/2 < 1$, the electrons are thermally stable also.

Hot, geometrically thick, optically thin, advection-dominated accretion flows, therefore, provide a mode of accretion at low rates ($\dot{m} < 0.1$) that is an alternative to standard disks. At low rates, therefore, nature has a choice in how it allows matter to accrete onto black holes, and it always seems to choose the hot, optically thin solution. Precisely why this is the case is still not understood.

A Significant Unresolved Issue in Advection-Dominated Accretion Flow Theory

We have glossed over a serious issue that is not yet well understood in the theory of ADAFs: how does an ADAF flow develop in the central regions of a low-accretion-rate disk inflow? Our previous discussion may give the impression that, in a low/hard X-ray binary source, the ADAF structure (equations (12.52) and (12.58)) extends from the black hole all the way out to the Roche lobe where the inflow is beginning to form. This is clearly not the case. The geometrically thick ADAF is believed to occupy only the inner few thousand Schwarzschild radii of the accretion flow – about the radius of a white dwarf or of the earth. Beyond that the accretion flow is believed to be regular Shakura–Sunyaev geometrically thin, optically thick accretion.

The means by which the cool geometrically thin disk turns into a geometrically thick, but optically thin, accretion flow inside some radius is poorly understood. A process of “evaporation” of the thin disk is believed to be at work. Most investigators simply assume that the accretion rate in the evaporated corona (ADAF) actually increases with decreasing radius until the underlying thin disk has completely dissipated. Then, interior to that point, which is called the evaporation radius R_{ev}

$$\begin{aligned} \dot{m}^{\text{ADAF}}(R) &= \dot{m} (R/R_{\text{ev}})^{-n} & R > R_{\text{ev}} \\ &= \dot{m} & R \leq R_{\text{ev}} \end{aligned} \quad (12.59)$$

only an ADAF with an accretion rate $\dot{m}$, uniform with radius, remains. The exponent in equation (12.59) is believed to be between $1 < n < 2$. This simple model, however, does not have any predictive power. For example, one cannot predict how

the evaporation radius varies with the general external accretion rate $\dot{m}$ or with other parameters of the inflow.

A few groups, however, have investigated the physics of this process more deeply than this simple model. Fredrich Meyer, Bifang Liu, and Emmi Meyer-Hofmeister (and their colleagues at the Max Planck Institute for Astrophysics in Garching, Germany and Kyoto University in Japan) developed a quantitative model for the evaporation process [443, 444, 445, 446]. (A similar idea also was suggested by Bożena Czerny, Agata Różańska, and Agnieszka Janiuk, of the Nicolaus Copernicus Center in Warsaw [447].) The evaporation model works as follows:

- First, the outer geometrically thin Shakura–Sunyaev disk develops a corona, similar to that of the sun, and by similar processes. Initially (*i.e.*, at large disk radii), the corona plays very little role in the disk accretion process.
- Further inward, however, the corona itself becomes denser and begins to act like an accretion flow itself, but *above* the cool thin disk: shear in the corona, along with a magnetic/turbulent-type of viscosity, generates heat and transports angular momentum outward. The corona acts like a very tenuous ADAF.
- Because of the inefficient heat transfer between ions and electrons in the corona, the heat generated by the coronal viscous processes can be used either to (1) drive an outflowing thermal wind or (2) to heat the surface of the accretion disk below through the process of conduction, with perhaps 50% of the coronal heating going into each outlet.
- The heating of the disk surface can have two effects: the denser parts of the disk atmosphere will simply radiate this heat away, while the less dense layers will expand upward into the corona and become part of it. This begins the evaporation process.
- The addition of more material into the corona then accelerates the mass lost from the disk into that corona. When the evaporation rate equals the accretion rate $\dot{M}$, the thin disk has disappeared, leaving a corona only, *i.e.*, an ADAF.

Evaporation theory holds some promise for developing complete models of accretion flows of the X-ray hard state (when the accretion rate is much smaller than $\dot{M}_{\text{Edd}}$), from the outer thin accretion disk all the way to the black hole. However, like most accretion theory, evaporation theory is still in its infancy, and only a few groups have begun to study it in a serious manner. Being able to include it in full MHD simulations of disks, for example, is still many years away, and there are many other processes to be added as well before the accretion picture is complete.

12.2.2.5 Numerical Models of Classical Accretion Disks

The analytic models above made a lot of assumptions in order to obtain an analytic solution. If we were to solve more detailed equations, which take both gas and radiation pressure into account, as well as several different kinds of opacity, *etc.*, how well would our analytic results compare?

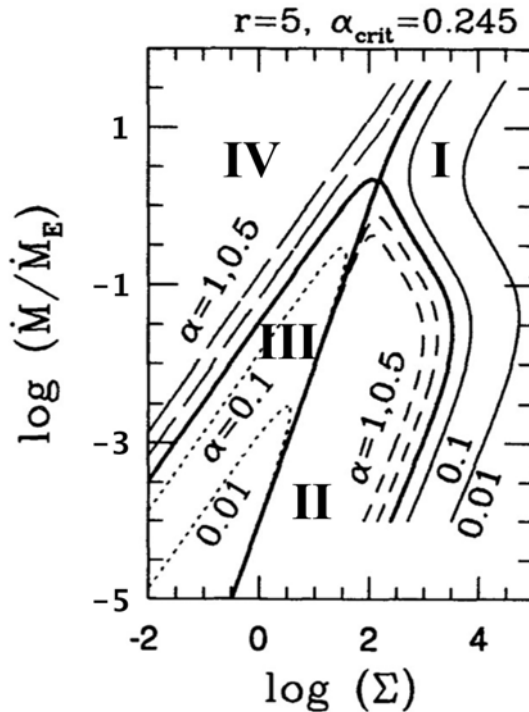


Fig. 12.7: Comparable figure to Fig. 12.4, but computed using more accurate numerical models of the accretion flows. Parameters are the same as in Fig. 12.4 (but $r = R/r_S = R/(2r_g)$ here). Also, labels on these curves are for *absolute* (rather than relative) values for the α parameter. Adapted from Fig. 1a in [448]; the vertical axis has been reduced by a factor of ten here to take into account the inclusion of $\epsilon_{\text{acc}} = 0.1$ in our definition of $\dot{M}_{\text{Edd}}$ (eq. (11.4)). Reproduced by permission of the AAS.

In 1995, a much more careful, numerical analysis was done of classical accretion theory by Xingming Chen, of Chalmers University in Sweden, and the advection-dominated and slim disk teams [448]. They solved the equilibrium equations of accretion disk structure, but did not take into account time-dependence. (The stability, or lack thereof, was determined after the fact with stability relations.) Fig. 12.7 shows the Σ - $\dot{m}$ plane that was produced by Chen and his colleagues. The parameters are the same as in Fig. 12.4. Again, we can see how all the different disk models fit together in a single theory. On the right of the diagram is the familiar S-curve formed by the standard and slim-supercritical disks (equations (12.26), (12.31), (12.35), and (12.52)). On the left is the familiar $\Sigma \propto \dot{m}$ line that describes *all* advection-dominated accretion flows (low and high luminosity; again, equations (12.52)). The lower portion of the diagram is bisected by the (thermally unstable) SLE solutions (equations (12.47)). Because the SLE models are weak functions of α ($\Sigma \propto \alpha^{1/9}$), all SLE models lie essentially along a thin line. The diagram can be divided into four regions, marked on the plot:

- I. The standard/slim disk, S-shaped region, which includes slim super-Eddington models having low α .
- II. The standard disk/SLE, hump-shaped region.
- III. The ADAF/SLE, delta-shaped region.
- IV. The completely advection-dominated region, which includes slim super-Eddington models having high α .

Nature appears to avoid the lower central portion of the diagram (regions II and III). Region II has $\alpha > 1$, which is rather unlikely. The II–III boundary is thermally unstable. And, region III contains ADAF models with low α , something that is avoided but unclear as to why. This diagram prompted Chen and his colleagues to point out that there appears to be a critical viscosity parameter α_{crit} (which is a function of position in the disk) that appears to control what types of solutions are realized in nature. The solutions along the thick solid lines all have $\alpha = \alpha_{\text{crit}}$. If $\alpha < \alpha_{\text{crit}}$, then nature chooses the geometrically thin standard disks solutions in region I. If $\alpha > \alpha_{\text{crit}}$, then nature chooses the geometrically thick advection-dominated flows in region IV.

The fact that the solutions that nature chooses differ by the value of the parameter α could be a cause or an effect of the different accretion states that an accreting black hole can enter. If it is a cause, then one model for a state transition might be the sudden change in the viscosity in the disk. If it is an effect, then the differing viscosity strengths could be due simply to a fundamental difference in the way thin and thick disks behave. Later we argue that this behavior is an effect, not a cause. (That is, high viscosity is to be expected for thick disks and low for thin disks.) In this case, there must be another mechanism that drives the change in accretion state.

Comparing Figs. 12.7 and 12.4 also shows the need for more sophisticated models and simulations than the ones in the algebraic equations (12.31), (12.35), (12.47), and (12.52). The α_{crit} values derived with Chen's models have reasonable values (0.2–0.4), while the α_{crit} value in our Fig. 12.4 is rather high (~ 1.5). Given that the results in these equations are accurate to about a factor of two or three, this is to be expected. However, this is not good enough to do the kind of analysis that Chen's group did, hence the need for more accurate numerical models that include better expressions for the cooling, equations of state, gravity, and fluid flow.

The Accretion Rate Model for Disk State Transitions

While advection-dominated solutions can exist for all accretion rates, the gas-pressure-dominated one can exist only for low accretion rates ($\dot{m} < \dot{m}_A$). This prompted Ann Esin (now at Harvey Mudd College) and her colleagues at Harvard to suggest a unified model for disk accretion in which the *primary* driving parameter is the accretion rate $\dot{m}$. Fig. 12.8, taken from one of her papers [449], shows a schematic representation of the disk structure in the five different states identified in Chapter 3.

At very low accretion rates ($\dot{m} \ll 1$), binary X-ray sources are in a quiescent state, and at higher rates, but still $\dot{m} < 0.1$, we identify them as being in the low/hard

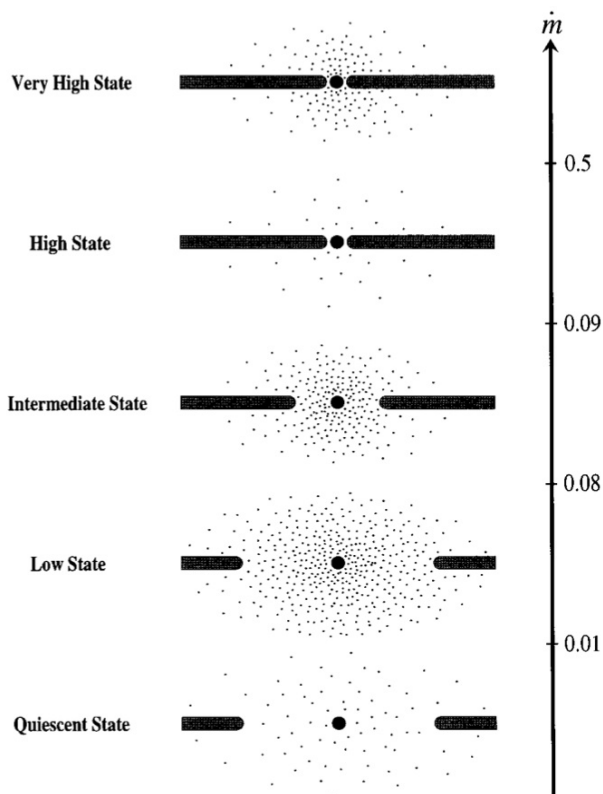


Fig. 12.8: Schematic representation of different observed accretion states in the ADAF and standard disk theories. (We present a more detailed set of scenarios in Figs. 16.4, 16.6, 16.8, 16.9, 16.11, and 16.12.) The tenuous speckled gas represents the ADAF/corona component, while the horizontal dense solid line represents the geometrically thin standard Shakura–Sunyaev disk models. The prominence of the ADAF/corona increases from the quiescent state ($\dot{m} \lesssim 0.003$) through the intermediate state ($\dot{m} = \dot{m}_A \sim 0.1$, where $\dot{m}_A$ is the critical accretion rate between ADAF and thin disk dominance). The high ($\dot{m} \sim 0.3$) and very high ($\dot{m} \rightarrow 1$) states are dominated by the cool, thin disk. Reproduced from Fig. 1 in [449], by permission of the AAS.

state. These states are primarily in the ADAF mode. Even if a cool disk is also a solution, for some (still unknown) reason nature chooses the ADAF over the standard flow, at least for radii near the black hole.

When $\dot{m} > 0.1$ Coulomb coupling to electrons dominates ion cooling over advective cooling, and the disk *must* become cool (a few keV) and thin. This situation describes the high/soft state. However, it does *not* explain the very hard γ -ray component that is actually more prominent in the high/soft state than in the low/hard state. When $\dot{m} > \dot{m}_{M-I}$, the “inner” disk becomes radiation-pressure-dominated and unstable to limit-cyclic behavior.

Important Result: No Stable Cool Disk Exists for Supermassive Black Holes

For standard accretion disks, we found that an unstable “inner” region does not exist until the accretion rate exceeds $\dot{m} > \dot{m}_I \propto (\alpha m)^{-1/8}$. The coefficient of $\dot{m}_I$, however, is not well determined from our analytic theory. So, we turn to Chen’s numerical models for a better, more accurate estimate. Fitting to his results, we find

$$\dot{m}_I \simeq 0.25 \left(\frac{\alpha}{0.03} m \right)^{-1/8} \quad (12.60)$$

(See reference [432].) We use $\alpha \sim 0.03$ here because this argument applies mainly to the thin, standard disk solutions; but, of course, the value of α is uncertain also.

We want to compare this to the accretion rate below which a low/hard state (ADAF) becomes important, but the value of $\dot{m}_A$ is equally uncertain. In our discussion above we use a value of $\dot{m}_A \sim 0.1$, while observational results put it in the range 0.05–0.08 (see Section 16.1.3), with an RMS value of ~ 0.06 . Putting this all together, and solving for the critical mass where $\dot{m}_I$ drops below $\dot{m}_A$, we find that $M_{\text{crit}} \sim 10^{5 \pm 1} M_\odot$ – at the upper IMBH range, but with quite a bit of uncertainty due to the large exponent on the factor $(\dot{m}_I/\dot{m}_A)^8$ used to determine M_{crit} .

So, depending on the mass of the black hole, the accretion disk can have two different sets of available accretion states:

1. For $M \lesssim 10^5 M_\odot$, the disk can have *four* possible structures:
 - $\dot{m} < 0.06$: ADAF (low/hard state).
 - $0.06 < \dot{m} < \dot{m}_I$: cool, stable standard disk, all gas-pressure-dominated (high/soft state).
 - $\dot{m}_I < \dot{m} < 1$: unstable, radiation-pressure-dominated “inner” region (unstable state).
 - $1 < \dot{m}$: super-Eddington accretion in a slim disk.
2. For $M \gtrsim 10^5$ there are only *three* possible structures:
 - $\dot{m} < 0.06$: ADAF.
 - $0.06 < \dot{m} < 1$: unstable state.
 - $1 < \dot{m}$: super-Eddington accretion in a slim disk.

The stable, cool, constant $\dot{M}$ “middle” region should never be seen in accreting supermassive black holes. It cannot occur until the accretion rate drops below $\dot{m}_I$, but because $\dot{m}_I < \dot{m}_A \approx 0.06$, at this point the disk then will become an ADAF. Therefore, *all rapidly-accreting supermassive black holes (Seyferts, radio galaxies, quasars) should be in a limit-cyclic-unstable state.* In the n – T plane in Fig. 9.3 we see this loss of the “middle” region in the lower-left portion of the diagram, where the dashed and thin solid lines cross.

12.3 Magnetically-Dominated Accretion Disk Theory

12.3.1 *The Case for Strong Magnetic Fields in the Center of Low-Accretion Rate Flow*

Classical turbulent accretion disk theory has been successful in explaining a number of observed properties of accreting black holes (of both stellar and super mass): the hard X-ray emission in the low state, the soft emission in the high state, and unstable behavior in the very high state. Nevertheless, there are a number of deficiencies and criticisms of the theory that are not addressed in this nice picture:

- *The myth of the radiatively inefficient hard state.* The classical turbulent accretion theory asserts that whenever the flow enters a hot, hard X-ray state, the transfer of thermal energy between ions and electrons is inefficient, leading to a hot, optically thin plasma that advects most of its enthalpy into the black hole. However, there is good *observational* evidence that energy often *is* imparted efficiently to electrons, even when a black hole is in a hard X-ray state. Sources at the top right of the FBG diagram (Fig. 3.14) are quite hard and yet quite luminous. As a concrete example, Cygnus X-1 produces $\sim 90\%$ or more as much bolometric luminosity in the hard state as in its soft state. This is hardly “radiatively inefficient” accretion by any standard. The only true inefficient (possibly collisionless plasma) state might be the very low/hard or “off” state, of which the Galactic center black hole Sgr A* is a good example.
- *The difficulty of two-temperature flow.* There also is some reason to be suspicious of the inefficient hard state from a theoretical point of view. The model is critically dependent on the ions remaining much hotter in the central portion of the accretion flow so the disk can remain geometrically thick and optically thin. However, in this state each ion has 1000 or more times as much energy as each electron, and the plasma is still quite magnetically turbulent. If a significant amount of that thermal energy is transferred by *any* means to the electrons (*e.g.*, MHD or plasma waves), the electrons will radiate that energy via relativistic cooling so rapidly that the entire ADAF could quickly collapse into a geometrically much thinner and cooler structure. Indeed, while the solar corona shows some two-temperature structure, *ion and electron temperatures differ by a factor of only a few*. At this time of writing there have been no calculations that definitely show that alternate means of energy transfer between ions and electrons exist. However, the existence of highly luminous hard states in many X-ray sources raises the distinct possibility that efficient transfer *does* occur somehow, and yet the hard state does not completely collapse to a soft thin disk.
- *The X-ray power fluctuation spectrum.* As discussed in Chapter 3, many sources in the hard X-ray state (particularly those that produce jets) display low-frequency, bandwidth-limited noise plus a strong low-frequency quasi-periodic oscillation. For black holes of order $10 M_{\odot}$, these occur in the narrow frequency range of 0.1–10 Hz, which corresponds to the Keplerian frequency at a radius of several tens to a thousand r_g . The problem is not with the existence of low-frequency

QPOs. (They do not have to be generated far from the black hole.⁷) Instead, the fact that these sources in the hard state have bandwidth-limited fluctuation spectra is a *severe* problem. ADAFs are quite turbulent and quite optically thin; theoretically this sort of structure should extend all the way down to the black hole ($\leq r_{\text{isco}}$; equations (12.52)). We therefore should see considerable noise in the 10–1000 Hz range, but we do not. Instead, the noise is limited to a band below 10 Hz. *The inner accretion flow is remarkably quiet at frequencies above the QPO.*

- *Mildly-relativistic jets.* We will see in the next chapter that ADAF-like flows can be the source of winds and jets. However, simulations of radiatively-inefficient accretion flows usually produce relativistic jet outflows – exactly the opposite of the mildly-relativistic jet flows that are observed in the X-ray hard states. Relativistic jets are more characteristic of the transition to the soft state (see Section 3.5.2). Our understanding of how jets are launched, accelerated, and collimated from accretion flows toward black holes, therefore, is still very limited.

There is at least one way in which these deficiencies in the theory of the X-ray hard state can be addressed, which is still reasonably consistent with all the work that has been done so far on classical accretion theory: inside some cylindrical radius R_0 (which we will define below), the magnetically-turbulent ADAF begins to shut down and give way to a strongly-magnetized, magnetosphere-like structure that extends from R_0 *toward*, and into, the black hole. Strong magnetic fields would explain the lack of magneto-rotational turbulence for frequencies above ~ 10 Hz. (The MRI turbulence requires a weak field and will shut down in a strong field; see Section 9.6.3.) Furthermore, a rotating magnetosphere inside a few hundred r_g or so could explain the presence of a strong QPO and the cutoff in noise above that frequency. Finally, the strong field of a magnetosphere can support a geometrically thick, optically thin plasma structure without the need for extremely hot, virial ion temperatures. This plasma could remain around 10^{9-10} K – all the way to the black hole – without collapsing into a geometrically thin, optically thick, cool disk like the standard Shakura–Sunyaev disks above. Ions and electrons could remain coupled energetically, and the electrons could radiate that energy efficiently at hard X-ray temperatures without turning into a soft X-ray structure.

For these reasons, this author has introduced the concept of magnetically-dominated accretion flows or MDAFs, and this section of the chapter will be devoted to introducing them. The reader should be cautioned that these accretion models are only a few years old: they do not have the decades of effort by many people that the above classical accretion models have. The models below, therefore, should be considered representative of theoretical concepts that may explain the more luminous X-ray hard states more fully than current turbulent ADAF models. However, as more is understood theoretically about this state and its power fluctuation spectrum and jets, the details and understanding of magnetically dominant accretion flows may change significantly in the next several years. As this understanding increases,

⁷ There are a variety of ways in which *low* frequency oscillations might be generated near the last stable orbit, even though the Keplerian frequency there is high.

we will be able to apply these ideas to the soft accretion states as well and produce a much more sophisticated theory of black hole engines.

12.3.2 Magnetically-Advective (Transitional) Accretion Flow

How might an ADAF transition from a magnetically turbulent inflow to one that has a strong magnetic field, primarily radial magnetic field that funnels plasma into the black hole? Here we explore the possibility that (see Fig. 12.9)

- interior to some radius R_0 , the ion/electron temperature becomes so hot that the electrons cool rapidly via relativistic processes, particularly synchrotron radiation and the Compton effect
- the cooling of the ions inside this radius causes the distended ADAF to partially collapse toward the disk midplane, decreasing gas pressure but increasing the magnetic pressure because of the vertical compression
- the accretion flow also stretches (advects) the magnetic field in the radial direction toward the black hole, making the radial component B_R more important than B_ϕ or B_Z
- this eventually causes the plasma $\beta_p (= 8\pi p_g/B^2)$ to decrease below unity, inside some radius R_1 , turning off the turbulence and creating a distended magnetosphere with the plasma trapped in the magnetic field lines

In the disk region $R_1 < R < R_0$, the flow will be in a *transitional* state between ADAF and MDAF. In the region $R < R_1$ the flow will be in a truly magnetically-dominated state.

Note that, because transitional flow explicitly gives an estimate for B_Z , as well as B_R and B_ϕ , the approximation for the former in turbulent accretion theory (equation (12.20)) is no longer needed.

12.3.2.1 Equations for the Physical Structure of Transitional Flow

With the above prescription, we can write down the equations that govern the physical structure of the magnetically-advective accretion flow. Transitional flow is a turbulent accretion flow just like the classical accretion flow models. We therefore can use most of the disk structure equations ((12.7) to (12.10), (12.14), and (12.16)). However, we will no longer assume that all components of the magnetic field have the same strength. Instead, we will allow them to advect with the accretion inflow, using the standard conservation of magnetic flux in cylindrical coordinates (equation (9.155))

$$B_R \propto R^{-1} H^{-1} \quad (12.61)$$

$$B_\phi \propto V_R^{-1} H^{-1} \quad (12.62)$$

$$B_Z \propto V_R^{-1} R^{-1} \quad (12.63)$$

where H is the usual half-thickness of the accretion flow in the Z direction. Because B_R and B_ϕ now scale with R in potentially different ways, α no longer necessarily will be constant with radius. Indeed, with equations (12.7) and (12.9) the ratio of the magnetic stress to the pressure increases with decreasing radius

$$\alpha(R) \equiv \frac{t_{\phi R}}{p} \propto \frac{B_\phi B_R}{p} \propto \frac{R^3}{H^3}$$

which must *increase* with decreasing R because transitional flow is collapsing from the ADAF structure (which has $H \sim R$).

12.3.2.2 Equation for the Thermal Structure

The cooling by relativistic electrons is rapid only above $\sim 10^9$ K. Below this temperature cooling is by Bremsstrahlung only and, therefore, not efficient as long as the free-free/bound-free optical depth of the plasma remains small. So this radiation is not expected to cool the plasma all the way down to low ($\sim 10^7$ K) temperatures and form a geometrically thin, optically thick disk. We therefore will retain the simple energy equation (12.55) for the plasma temperature, which assumes that cooling results in an electron temperature of $\sim 10^9$ K.

We will *further* assume that, at some radius $x_0 = 820 \vartheta^{-1}$, or

$$R_0 = 7.3 \times 10^8 \text{ cm } m \vartheta^{-1} = R_A \vartheta^{-1} \quad (12.64)$$

the ion temperature will saturate to a finite multiple ϑ of the electron temperature

$$\vartheta \equiv \frac{T_i}{T_e} \geq 1 \quad (12.65)$$

and cooling will affect even the ions themselves. Interior to R_0 the ion temperature will no longer be virial, and the hot plasma will not be able to support the geometrically thick accretion flow. The ADAF then will begin to collapse vertically, lose gas pressure dominance, and become magnetically-dominated instead.

Note that, if $\vartheta = 1$, then $R_0 = R_A$, where the two-temperature ADAF (equations (12.58)) begins. On the other hand, if $\vartheta \geq 820$, the transition to an MDAF occurs inside the last stable orbit, ensuring that an ADAF structure exists down to the vicinity of the black hole. In our magnetically-dominated accretion flow models we shall assume that $1 < \vartheta < 820$, *i.e.*, that an MDAF exists somewhere outside the ISCO and interior to the two-temperature ADAF.

12.3.2.3 Scaling Law Structure Equations for Transitional Flow

We will form solutions for the accretion flow structure in the transitional and MDAF flow regions in a manner a little different from that used for the classical accretion

flow models. We will derive radial scaling laws interior to the radius R_0 that match the ADAF flow at R_0 but have the scalings appropriate for transitional flow. We then will follow with a similar procedure for the MDAF structure inside R_1 . This allows us to match ADAF, transitional flow, and MDAF accretion structure consistently at the boundary radii. With the above prescription for physical and thermal structure, the scaling law solution for transitional flow is then

$$\begin{aligned}
 \rho &= \rho_0 (R/R_0)^{-3/2} & \Sigma &= \Sigma_0 \\
 T_e &= T_{e,0} & B_R &= B_{R0} (R/R_0)^{-5/2} \\
 T_i &= T_{i,0} & B_\phi &= B_{\phi 0} (R/R_0)^{-1/2} \\
 p &= p_0 (R/R_0)^{-3/2} & B_Z &= B_{Z0} \\
 H &= H_0 (R/R_0)^{3/2} & V_{AR} &= V_{AR0} (R/R_0)^{-7/4} \\
 \tau_{es} &= \tau_{es0} & V_{turb} &= V_{turb0} \\
 \tau_{ff/bf} &= \tau_{ff/bf0} (R/R_0)^{-3/2} & \alpha(R) &= \alpha (R/R_0)^{-3/2} \\
 V_R &= V_{R0} (R/R_0)^{-1}
 \end{aligned}$$

where the subscript 0 refers to the ADAF conditions at R_0 , $\alpha(R)$ refers to the ratio of the magnetic stress to the pressure, and α (≈ 0.3) is still the viscosity parameter used in the outer ADAF part of the disk.

Is the magnetically-advective assumption valid? Does the inflow velocity V_R now dominate over the turbulent velocity V_R ? The answer is yes. In the ADAF case, the ratio $V_R/V_{turb} = \alpha^{1/2} H/R \approx 0.47$. However, in the above transitional flow, V_{turb} remains constant as R decreases, while V_R increases as R^{-1} toward the black hole. So transitional flow becomes increasingly magnetically-advective inward, justifying the use of equations (12.61) to (12.63).

Plugging in the ADAF conditions at R_0 , we can now write down the full scaling laws in terms of the parameters of the flow

ρ_c	$=$	2.1×10^{-5}	g cm^{-3}	α^{-1}	m^{-1}	$\dot{m}$	$x^{-3/2}$	
$T_{c,e}$	$=$	10^9	K					
$T_{c,i}$	$=$	10^9	K				ϑ	
p_c	$=$	2.8×10^{12}	dyn cm^{-2}	α^{-1}	m^{-1}	$\dot{m}$	ϑ	$x^{-3/2}$
H	$=$	2.7×10^4	cm		m		$\vartheta^{1/2}$	$x^{3/2}$
$\tau_{c,es}$	$=$	0.19		α^{-1}		$\dot{m}$	$\vartheta^{1/2}$	
$\tau_{c,ff/bf}$	$=$	1.18×10^{-12}		α^{-2}	m^{-1}	$\dot{m}^2$	$\vartheta^{1/2}$	$x^{-3/2}$
V_R	$=$	-2.6×10^{11}	cm s^{-1}	α			$\vartheta^{-1/2}$	x^{-1}
Σ	$=$	1.12	g cm^{-2}	α^{-1}		$\dot{m}$	$\vartheta^{1/2}$	
B_R	$=$	7.4×10^{11}	G		$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{-5/4}$	$x^{-5/2}$
B_ϕ	$=$	1.11×10^6	G		$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{3/4}$	$x^{-1/2}$
B_Z	$=$	3.4×10^4	G		$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{5/4}$	
V_{turb}	$=$	3.7×10^8	cm s^{-1}	$\alpha^{1/2}$			$\vartheta^{1/2}$	
V_{AR}	$=$	4.6×10^{13}	cm s^{-1}	$\alpha^{1/2}$			$\vartheta^{-5/4}$	$x^{-7/4}$
$\alpha(x)$	$=$	2.4×10^4		α			$\vartheta^{-3/2}$	$x^{-3/2}$
$\beta_p(x)$	$=$	1.27×10^{-10}		α^{-1}			$\vartheta^{7/2}$	$x^{7/2}$

(12.66)

where $\alpha(\approx 0.3)$ is the standard viscosity constant in the outer ADAF, while $\alpha(x)$ shows how the *local* value of the ratio of magnetic stress to the pressure varies with radius.

With $H \propto R^{3/2}$, the radiating, turbulent flow is now collapsing quickly as it accretes inward. And, with $B_R \propto R^{-5/2}$, $\alpha(R) \propto R^{-3/2}$, and the Alfvén speed $V_{AR} \propto R^{-7/4}$ all rapidly increasing inward, the magnetic field is rapidly becoming dominant. Both the plasma β_p and the effective viscosity parameter $\alpha(R)$ become of order unity at approximately the same radius

$$R_1 = 7.3 \times 10^8 \text{ cm } \alpha^{2/3} m \vartheta^{-1} \quad (12.67)$$

or $x_1 = 820 \alpha^{2/3} \vartheta^{-1}$. The ratio that tells us how large this transition flow region can be is

$$\frac{R_0}{R_1} = \alpha^{-2/3}$$

For the ADAF standard $\alpha \approx 0.3$, the region is quite modest (only a factor of two or so in radius). ADAFs are not very far from having a strong magnetic field turning off their turbulence, and transitional flow accomplishes just that.

12.3.2.4 Stability of Transitional Flow

The transitional flow solution is both viscously and thermally stable. Viscous stability occurs because $d \ln \Sigma / d \ln \dot{m} > 0$. For determining thermal stability, we note that, as with the ADAF solution, Compton or thermal synchrotron cooling varies as $Q^- \propto T$. The heating, on the other hand, varies as $Q^+ \propto t_{\phi R} H \propto B_R B_\phi H \propto 1/(V_R H)$. Now, at a given radius V_R varies only as H^2 (equation (12.19)), and H varies as $T^{1/2}$ in hydrostatic equilibrium (equation (12.9)). Therefore, $Q^+ \propto T^{-3/2}$. So, since $d \ln Q^+ / d \ln T < d \ln Q^- / d \ln T$, the transitional flow is also thermally stable.

12.3.3 Magnetically-Dominated Accretion Flows (MDAFs)

Interior to R_1 the magnetic field dominates the dynamic forces, causing the MRI turbulence to shut off. The accretion flow then will take place on inwardly-directed magnetic field lines. This kind of accretion is similar to plasma flow along a pulsar (or earth's) magnetic field lines, but the currents that create those fields reside in the ADAF/transitional portion of the accretion flow, *not in the black hole*.

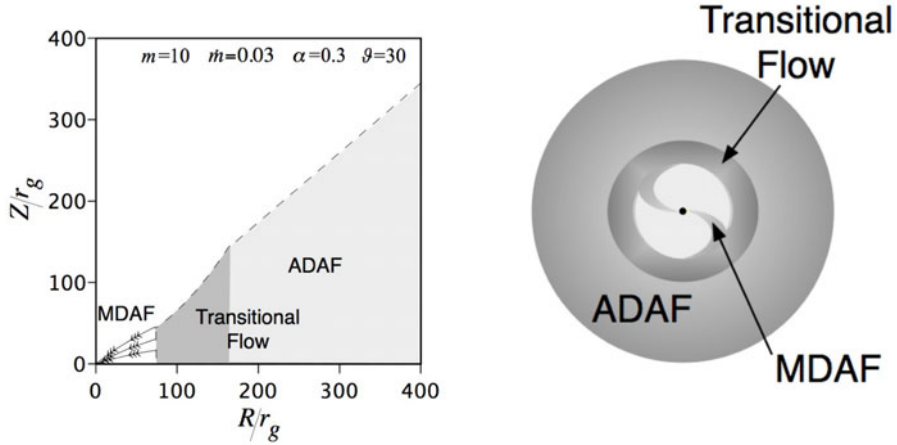


Fig. 12.9: Structure of a magnetically-dominated accretion flow (MDAF) interior to an ADAF. *Left:* Similar diagram to Fig. 12.6a, but with 10 times higher accretion rate; an MDAF with $\vartheta = 30$ (suggested in Section 16.2.4) now exists interior to the ADAF. Note the transitional flow region, where cooling leads to a $\beta_p < 1$ region (inwardly-facing magnetosphere) and accretion flow along strong magnetic field lines. *Right:* View of the MDAF/ADAF inflow from the north pole. Note cooler transitional flow and inspiral of accreting matter along strong MDAF field lines. The open inspiral here represents a sub-Alfvénic solution for the MDAF (eq. (12.79)), where $B_\phi (\propto V_\phi \propto R)$ rapidly decreases with radius. Rotation of the MDAF will be at approximately $\Omega_f = \Omega_K$ in the transitional flow region; this could generate a sharply-peaked QPO at that (or twice) Ω_f .

12.3.3.1 Physical Structure of Magnetically-Dominated Accretion Flow

To analyze the accretion flow in the presence of strong field lines, we will return to the basic non-relativistic MHD equations ((9.152) to (9.155), with no viscosity or nuclear burning) and express them as conservation laws in cylindrical coordinates. First, there is mass conservation

$$-4\pi \rho V_R R H = \dot{M} \quad (12.68)$$

radial momentum/energy conservation (where we now shall ignore thermal forces because of the magnetic dominance)

$$-4\pi R H \left\{ \rho V_R \left[\frac{1}{2} (V_R^2 + V_\phi^2) - \frac{GM}{(R^2 + H^2)^{1/2}} \right] + V_R \frac{B_\phi^2 + B_Z^2}{4\pi} - V_\phi \frac{B_R B_\phi}{4\pi} \right\} = \dot{\mathcal{E}} \quad (12.69)$$

angular momentum conservation

$$-4\pi R^2 H \left(\rho V_R V_\phi - \frac{B_R B_\phi}{4\pi} \right) = \dot{\mathcal{L}} \quad (12.70)$$

and magnetic flux conservation

$$B_R R H = \Psi_R \quad (12.71)$$

$$B_\phi V_R H = (\omega - \Omega_f) \Psi_R \quad (12.72)$$

$$B_Z V_R R = \Psi_Z \quad (12.73)$$

where $\dot{M}$, $\dot{\mathcal{E}}$, and $\dot{\mathcal{L}}$ are the mass, energy, and angular momentum accretion rate constants, and Ψ_R and Ψ_Z are the radial and vertical magnetic flux constants. Ω_f is the magnetic *field* rotation rate, which is constant in space and time for these time-independent equations. And $\omega = V_\phi/R$ is the *plasma* rotation rate, which is a function of R . Equation (12.72) is noteworthy as the “isorotation law” that equates the vertical frozen-in electric field component ($c E_Z = -(\mathbf{V} \times \mathbf{B})_Z$) to the electromotive force generated by the rotating radial magnetic field ($R \Omega_f B_R$).

These six equations are functions of the seven unknowns ρ , V_R , V_ϕ , B_R , B_ϕ , B_Z , and H . They replace the classical turbulent accretion flow equations (12.7), (12.8), (12.10), (12.16), and (12.17). (The analogy to equation (12.9) is missing and will be discussed below.) The main difference between these two sets is that thermal pressure forces have been ignored in equations (12.69) to (12.73) and magnetic forces have been emphasized. Our goal will be to extract scaling laws for the seven unknowns in a manner similar to those we presented for classical turbulent accretion theory in Section 12.2.

We will approach this problem in three stages: (1) first solve for the angular velocity ω and azimuthal magnetic field component B_ϕ , which are strongly coupled in this highly magnetized flow; (2) develop a technique for handling the complex energy equation (12.69); and then (3) decide upon a seventh equation that closes the set and allows us to solve for scaling laws for all seven unknowns.

Stage 1: Angular Momentum Transport by Strong Magnetic Fields

The process of accretion in the presence of strong magnetic fields is quite different from that in the weak field case. In the latter, the process involves magneto-rotationally-induced turbulence, while in the strong field case plasma flows along stiff magnetic field lines (mainly B_R and B_ϕ) that transport angular momentum outward from the inner portion of the MDAF accretion flow back up into the turbulent flow near R_1 . Fortunately, the angular momentum transport equations have analytic wind solutions (singular plus sub- and super-Alfvénic), making accretion flow structure nearly as easy to study as the weak field/turbulent case.

Equations (12.68) and (12.70) to (12.72) can be combined *algebraically* to obtain a general solution for both the plasma angular velocity and the azimuthal magnetic field component

$$\omega = \frac{\Psi_R^2 \Omega_f R + \dot{\mathcal{L}} V_R H}{\Psi_R^2 R + \dot{M} R^2 V_R H} \quad (12.74)$$

$$B_\phi = \frac{-\Psi_R \dot{M} \Omega_f R^2 + \dot{\mathcal{L}} \Psi_R}{\Psi_R^2 R + \dot{M} R^2 V_R H} \quad (12.75)$$

The character of the accretion flow interior to R_1 (*i.e.*, in the MDAF), therefore, depends critically on the radial magnetic flux Ψ_R and on the rate of angular momentum accretion $\dot{\mathcal{L}}$ in that region. We will now examine several possible solutions for the azimuthal velocity and magnetic field.

The Singular MDAF Solution. Because $\dot{M}$ is positive when V_R is negative, while Ψ_R^2 is always positive, it is easy to see that the denominators of equations (12.74) and (12.75) will be zero when

$$\Psi_R^2 R = -\dot{M} R^2 V_R H$$

Substituting equation (12.68) for $\dot{M}$ and (12.71) for Ψ_R , we see that this singular point is the radial *Alfvén* point – where the accretion inflow (or outflow) equals the Alfvén speed

$$V_R = \pm V_{AR}$$

where

$$V_{AR} \equiv \frac{B_R}{(4\pi \rho)^{1/2}}$$

is the Alfvén speed using only the radial magnetic field component. In order to keep ω and B_ϕ finite at the Alfvén singular point, the numerators also must be zero, which means that the Alfvén radius R_{Alf} must be given by

$$R_{\text{Alf}} = \left(\frac{\dot{\mathcal{L}}}{\dot{M} \Omega_f} \right)^{1/2}$$

In this *singular* MDAF accretion solution, therefore, we must have $-V_R < V_{AR}$ for $R > R_{\text{Alf}}$ and then $-V_R > V_{AR}$ for $R < R_{\text{Alf}}$.

Sub-Alfvénic Solutions. In our treatment of MDAFs, however, we shall reject the singular solution above and, instead, choose one where the inflow speed is always sub-Alfvénic, *i.e.*, one where $V_R^2 \ll V_{AR}^2$. The reasons for making this choice are

1. The radial magnetic field in the transition region ($R_1 < R < R_0$) grows stronger inward, yielding an Alfvén velocity that grows faster than either V_R or even the Keplerian velocity V_K ; by the time the transitional flow gets to R_1 , V_{AR} is significantly greater than the radial inflow velocity, and it rivals the Keplerian rotational velocity; see equations (12.66).

2. Inside R_1 the inflow velocity V_R should never exceed the free-fall speed; indeed, it might be slower than that at some points, depending on the character of the azimuthal field; so, the ratio V_R/V_{AR} always should remain less than unity. However, as the inflow becomes more magnetically-dominated, the radial Alfvén velocity might become larger than the free-fall speed, or at least remain comparable.

These arguments will break down, of course, very near the black hole horizon, where relativistic effects (ignored here) become important. Critical and separatrix surfaces (the light cylinder [analogous to the Alfvén point] and the causality limit [the fast magnetosonic separatrix surface FMSS; see Section 15.1.3]) then *must* form in the flow before it crosses the horizon. (See Section 14.3.) We conclude that a sub-Alfvénic flow should be maintained during most of the accretion process, and that an (Alfvén) point should not occur until very close to the black hole horizon – well inside the MDAF structure being considered here.

In the sub-Alfvénic case, therefore, the denominators of equations (12.74) and (12.75) always will be dominated by the $\Psi_R^2 R$ term, so that

$$\omega = \Omega_f + \frac{\dot{\mathcal{L}} V_R H}{\Psi_R^2 R} \quad (12.76)$$

$$B_\phi = -\frac{\dot{M} \Omega_f}{\Psi_R} R + \frac{\dot{\mathcal{L}}}{\Psi_R R} \quad (12.77)$$

If we define a dimensionless angular momentum accretion rate

$$\dot{\ell}(R) \equiv \frac{\dot{\mathcal{L}}}{\dot{M} \Omega_f R^2} \quad (12.78)$$

then the sub-Alfvénic assumption reduces the number of possible solutions to only three, based on the value of $\dot{\ell}$

- $\dot{\ell}(R) < 1$ (low angular momentum accretion rate). This solution gives, simply

$$\boxed{\omega = \Omega_f \quad B_\phi = -\frac{\dot{M} \Omega_f}{\Psi_R} R} \quad (12.79)$$

In this case, the plasma rotates with the rigidly-rotating magnetic field, which becomes increasingly dominated by B_R , while B_ϕ diminishes in strength as the matter flows inward along the increasingly radial field lines. (See Fig. 12.9.) This will be our primary MDAF solution used in the models below.

- $1 < \dot{\ell}(R) < V_{AR}^2/V_R^2$ (moderate angular momentum accretion rate). This solution gives

$$\omega = \Omega_f \quad B_\phi = \frac{\dot{\mathcal{L}}}{\Psi_R} R^{-1}$$

So the plasma still rotates with the field, but now B_ϕ *increases* in strength inward.

- $V_{AR}^2/V_R^2 < \dot{\ell}(R)$ (high angular momentum accretion rate). The angular velocity of the plasma now depends on the accretion flow velocity and could increase inward

$$\omega = \frac{\dot{\mathcal{L}}}{\Psi_R^2} V_R H R^{-1} \quad B_\phi = \frac{\dot{\mathcal{L}}}{\Psi_R} R^{-1}$$

Super-Alfvénic Solutions. While we will be rejecting solutions that are singular or super-Alfvénic, it is nevertheless instructive to write down the latter as well. Super-Alfvénic solutions occur when the denominators of equations (12.74) and (12.75) are dominated by their second term

$$\omega = \frac{\Psi_R^2 \Omega_f}{\dot{M} R V_R H} + \frac{\dot{\mathcal{L}}}{\dot{M} R^2} \quad (12.80)$$

$$B_\phi = -\frac{\Omega_f \Psi_R}{V_R H} + \frac{\dot{\mathcal{L}} \Psi_R}{\dot{M} R^2 V_R H} \quad (12.81)$$

This generates another three possible solutions based on the value of $\dot{\ell}$:

- $\dot{\ell}(R) < V_{AR}^2/V_R^2$ (low angular momentum accretion rate). This solution gives

$$\omega = \frac{\Psi_R^2 \Omega_f}{\dot{M}} \frac{1}{V_R H} R^{-1} \quad B_\phi = -\Omega_f \Psi_R \frac{1}{V_R H}$$

If $V_R H$ is constant with radius, as is sometimes the case, the azimuthal velocity ($V_\phi = R\omega$) and B_ϕ both will be constant with radius.

- $V_{AR}^2/V_R^2 < \dot{\ell}(R) < 1$ (moderate angular momentum accretion rate). This solution gives

$$\omega = \frac{\dot{\mathcal{L}}}{\dot{M}} R^{-2} \quad B_\phi = -\Omega_f \Psi_R \frac{1}{V_R H}$$

The velocity field now assumes a vortex character – uniform angular momentum per unit mass ($R^2 \omega$).

- $1 < \dot{\ell}(R)$ (high angular momentum accretion rate). The azimuthal magnetic field now increases inward rapidly, in response to the rapidly twisting by the rapidly-rotating plasma vortex.

$$\omega = \frac{\dot{\mathcal{L}}}{\dot{M}} R^{-2} \quad B_\phi = \frac{\dot{\mathcal{L}} \Psi_R}{\dot{M}} \frac{1}{V_R H} R^{-2}$$

In the first of these super-Alfvénic solutions, the stress component $t_{\phi R} = B_R B_\phi / 4\pi$ is strong enough to cause some angular momentum transport outward in the MDAF plasma. The other two solutions have a familiar “vortex” solution ($\omega \propto R^{-2}$), where

the angular momentum inflow is so rapid that none can be transferred outward in an inflow time (*i.e.*, $\dot{M} \omega R^2 = \dot{\mathcal{L}}$).

Stage 2: Techniques for Extracting Scaling Laws from the Complicated Energy Equation

The complexity of the energy equation (12.69) makes it difficult to generate simple scaling laws like those of Shakura and Sunyaev's thin disk solutions. There are really two problems: (1) the large number of terms (we prefer only two terms be important in order to generate a simple scaling with R) and (2) the constant $\dot{\mathcal{E}}$ on the right-hand side (which adds an effective eighth term equal to $\dot{\mathcal{E}}/4\pi R H$).

We shall eliminate the $\dot{\mathcal{E}}$ problem by noting that $\dot{\mathcal{E}}$ is a true constant of the motion inside R_1 . Therefore, while the value of $\dot{\mathcal{E}}$ at that radius may be comparable to some of the terms in equation (12.69) (when evaluated at R_1), $\dot{\mathcal{E}}$ will be much smaller than those same terms well inside the MDAF region (*i.e.*, for $R \ll R_1$). So we can set $\dot{\mathcal{E}} \approx 0$ for $R \ll R_1$, reducing the energy equation to

$$\rho V_R \left[\frac{1}{2} (V_R^2 + V_\phi^2) - \frac{GM}{R} \right] + V_R \frac{B_\phi^2 + B_z^2}{4\pi} - V_\phi \frac{B_R B_\phi}{4\pi} \approx 0 \quad (12.82)$$

The problem with having too many terms is solved in a similar manner: we shall keep only those terms with the steepest *negative* exponent of R (the scaling radius). For example, if two terms have an R^{-3} dependence, and if that is the steepest in the energy equation, then all terms with a shallower dependence (*i.e.*, $R^{-5/2}$, R^{-2} , *etc.*) are discarded. At least *two* terms in equation (12.82) must possess the same steepest negative R dependence in order for the scaling law to be valid. This makes the energy equation useful in the development of scaling laws, even though its character might change as different terms become dominant.

Stage 3: The Seventh Relation Necessary to Close the Set of Equations

The final task is to determine a seventh relation amongst the variables in order to solve for the seven unknowns. In the turbulent disk case vertical hydrostatic equilibrium in the disk (equation (12.9)) served as the additional relation (along with radiative transport equations to determine the plasma temperature and pressure). Therefore, one obvious choice for a seventh relation might be vertical hydromagnetic equilibrium (hydrostatic equilibrium with the magnetic pressure replacing the plasma pressure)

$$\frac{B_R^2}{8\pi} = \frac{GM}{R^3} H^2 \rho \quad (12.83)$$

or, $V_{AR} = V_{\text{esc}} (H/R)^2$. This does indeed close equations (12.68) to (12.73) and give a set of scaling solutions for the flow. However, this relation probably is not

correct for a magnetosphere that is anchored in the nearby transition region. Observations of strong magnetic field structures on the sun, for example, show that they do not stably support plasma against gravitational forces. Instead, coronal loops, for example, bulge out of the solar photosphere into its upper atmosphere. Much of the plasma trapped in the loop's field lines drains back onto the solar surface, leaving a magnetic field that is in force-free equilibrium. This is essentially the "Parker instability": when the magnetic field dominates the dynamics, the plasma goes along for the ride, and its gravitational weight is of little or no consequence. So, the assumption of the field supporting the weight of the plasma above is wrong. Instead, we should assume a force-free magnetic field.

In order to solve the structure of a true force-free field, we must solve the GSS cross-field equation for the magnetic potential Ψ and then extract the R and Z components by computing the poloidal field $\mathbf{B}_p$. This task is beyond the scope of our current discussion, although Chapter 14 deals with the structure of black hole magnetospheres in some detail.

However, we currently are looking only for laws that scale with radius R , perhaps over several decades in radius. Now, in order for the poloidal magnetic field to remain force-free over several orders of magnitude in radius, both B_R and B_Z must scale with R in the same manner. That is, a simple scaling condition that represents a force-free magnetic field solution is

$$B_Z \propto B_R \quad (12.84)$$

rather than magnetic hydrostatic equilibrium. We therefore will use proportionality (12.84) as our seventh relation.

For our chosen MDAF solution we find below that, indeed, V_{AR} is much larger than hydromagnetic equilibrium would predict (*i.e.*, $V_{AR} \approx 2V_{\text{esc}} \alpha^{-1/3}$, rather than H^2/R^2 times this). However, interior to a third radius R_2 , the magnetized structure will again become as thick as its radius ($H \sim R$). In that case equation (12.83) will replace flux conservation as the equation determining B_R .

12.3.3.2 An MDAF Example

As a single example of the several MDAF solutions discussed above we shall choose one likely possibility that could eventually help explain phenomena like quasi-periodic oscillations and jets that are seen in the X-ray hard state (see Chapter 3). This solution is the sub-Alfvénic, low-angular-momentum one ($\dot{\mathcal{L}} < \dot{M} \Omega_f R^2$).⁸ As with the transitional flow, we will derive scaling laws that satisfy the equations and then use a starting radius (here R_1) to find the coefficients of the scaling laws.

The scaling laws that satisfy all seven physical structure equations, plus the electron temperature condition (12.55), as well as the slim disk condition ($H < R$), are

⁸ We will take a look at the other super-Alfvénic flows when we discuss numerical simulations of MDAFs below.

$$\begin{aligned}
\rho &= \rho_1 & B_R &= B_{R1} (R/R_1)^{-1/2} \\
T &= T_1 & B_\phi &= B_{\phi 1} (R/R_1) \\
p &= p_1 & B_Z &= B_{Z1} (R/R_1)^{-1/2} \\
H &= H_1 (R/R_1)^{-1/2} & V_{\text{turb}} &= 0 \\
\tau_{\text{es}} &= \tau_{\text{es}1} (R/R_1)^{-1/2} & V_{\text{AR}} &= V_{\text{AR}1} (R/R_1)^{-1/2} \\
\tau_{\text{ff/bf}} &= \tau_{\text{ff/bf}1} (R/R_1)^{-1/2} & \alpha(R) &= \alpha (R/R_1)^{1/2} \\
V_R &= V_{R1} (R/R_1)^{-1/2} & \beta_p(R) &= \beta_{p1} (R/R_1) \\
V_\phi &= V_{\phi 1} (R/R_1) & \nu_f &= \Omega_f/2\pi = V_{\phi 1}/(2\pi R_1) \\
\Sigma &= \Sigma_1 (R/R_1)^{-1/2}
\end{aligned}$$

With equation (12.67) for R_1 , we can evaluate each of equations (12.66) at that radius, find the coefficients in the above scaling laws and obtain the structure equations for our chosen MDAF solution

ρ_c	$= 8.9 \times 10^{-10}$	g cm ⁻³	α^{-2}	m^{-1}	$\dot{m}$	$\vartheta^{3/2}$	
$T_{c,e}$	$= 10^9$	K					
$T_{c,i}$	$= 10^9$	K				ϑ	
p_c	$= 1.20 \times 10^8$	dyn cm ⁻²	α^{-2}	m^{-1}	$\dot{m}$	$\vartheta^{5/2}$	
H	$= 1.81 \times 10^{10}$	cm	$\alpha^{4/3}$	m		$\vartheta^{-3/2}$	$x^{-1/2}$
$\tau_{c,es}$	$= 5.5$		$\alpha^{-2/3}$		$\dot{m}$		$x^{-1/2}$
$\tau_{c,ff/bf}$	$= 1.42 \times 10^{-15}$		$\alpha^{-8/3}$	m^{-1}	$\dot{m}^2$	$\vartheta^{3/2}$	$x^{-1/2}$
V_R	$= -9.2 \times 10^9$	cm s ⁻¹	$\alpha^{2/3}$				$x^{-1/2}$
V_ϕ	$= 5.2 \times 10^5$	cm s ⁻¹	α^{-1}			$\vartheta^{3/2}$	x
Σ	$= 32$	g cm ⁻²	$\alpha^{-2/3}$		$\dot{m}$		$x^{-1/2}$
B_R	$= 1.11 \times 10^6$	G	$\alpha^{-4/3}$	$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{3/4}$	$x^{-1/2}$
B_ϕ	$= 47$	G	α^{-1}	$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{9/4}$	x
B_Z	$= 9.6 \times 10^5$	G	$\alpha^{1/3}$	$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{3/4}$	$x^{-1/2}$
V_{turb}	$= 0$						
V_{AR}	$= 1.05 \times 10^{10}$	cm s ⁻¹	$\alpha^{-1/3}$				$x^{-1/2}$
$\alpha(x)$	$= 0.035$		$\alpha^{-1/3}$			$\vartheta^{1/2}$	$x^{1/2}$
$\beta_p(x)$	$= 2.4 \times 10^{-3}$		$\alpha^{2/3}$			ϑ	x
ν_f	$= 0.093$	Hz	α^{-1}	m^{-1}		$\vartheta^{3/2}$	

(12.85)

As a consistency check, we see that the radial velocity remains less than the Alfvén speed V_{AR} (at a constant $\sim 26\%$ of it), and both scale with radius in the same way as the free-fall speed ($R^{-1/2}$). There are several things to note about this MDAF solution:

- Plugging these expressions into the energy equation (12.82), we find that the most important terms in the radial energy balance are radial kinetic energy, gravitational potential energy, and B_Z magnetic pressure. That is, the plasma is in near free-fall toward the black hole, with partial support by magnetic pressure.
- However, the plasma is *not* in Keplerian rotation, or even near it. Instead, the inner magnetized accretion flow rotates uniformly at a rate ν_f revolutions per second and has a magnetic field that becomes increasingly radial inward. V_ϕ is

unimportant in the inner regions of the MDAF inflow. That is, the accretion flow begins as a spiral inflow that eventually becomes almost purely radial.

- The disk half-thickness H actually *increases* with decreasing radius R , leading to a uniform density throughout this MDAF region.
- Finally, the magnetic stress transferring angular momentum ($t_{\phi R} \propto B_R B_\phi$) *decreases* with decreasing radius. That is, most of the angular momentum loss of the plasma is out near the radius R_1 ; well interior to that radius, the matter has rather little angular momentum.

Equations (12.85) remain valid with *decreasing* radius until the thin disk condition is no longer met ($H \geq R$). For simplicity, we will use our slim disk condition in equation (12.51) to mark this boundary which then lies at the radius

$$R_2 = 7.3 \times 10^8 \text{ cm } \alpha^{8/9} m \vartheta^{-1} \quad (12.86)$$

or

$$x_2 = 820 \alpha^{8/9} \vartheta^{-1}$$

Interior to this radius the accretion flow is, like an ADAF, geometrically thick and optically thin. However, unlike an ADAF this distension is due not to hot ions in a two-temperature plasma but rather to a force-free magnetic field – similar to the magnetospheres of the earth, sun, or pulsar, but directed inward from the transitional flow near R_1 toward the black hole.

12.3.3.3 The Accretion “End Game”: Radial Flow into the Black Hole

Note that, at $R = R_2$, the ratio of magnetic pressure force on the plasma in the Z direction (mainly from the magnetic field R component) to the tidal gravitational force on that plasma has *fallen* to a value of order unity. That is, equation (12.83) is now satisfied. From here on in toward the black hole, this equation will become our new condition that determines the R magnetic field strength, not flux conservation. But we will continue to assume a magnetosphere where $B_Z \propto B_R$ in order to compute the magnetic field Z component.

The final MDAF flow into the black hole, therefore, is similar to that of the original ADAF, but the structure is now supported in the vertical direction, not by a very high thermal pressure (with temperature approaching 10^{12} K), but rather by a high magnetic pressure (primarily from B_R). In addition, the magnetosphere is rigidly rotating; this continues the previous MDAF structure from that at R_2 down into the black hole itself:

$$\begin{aligned} \rho &= \rho_2 (R/R_2)^{-3/2} & B_R &= B_{R2} (R/R_2)^{-5/4} \\ T &= T_2 & B_\phi &= B_{\phi2} (R/R_2) \end{aligned}$$

$$\begin{aligned}
p &= p_2 (R/R_2)^{-3/2} & B_Z &= B_{Z2} (R/R_2)^{-5/4} \\
H &= H_2 (R/R_2) & V_{\text{turb}} &= 0 \\
\tau_{\text{es}} &= \tau_{\text{es}2} (R/R_2)^{-1/2} & V_{\text{AR}} &= V_{\text{AR}2} (R/R_2)^{-1/2} \\
\tau_{\text{ff/bf}} &= \tau_{\text{ff/bf}2} (R/R_2)^{-2} & \alpha(R) &= \alpha_2 (R/R_2)^{5/4} \\
V_R &= V_{R2} (R/R_2)^{-1/2} & \beta_p(R) &= \beta_{p2} (R/R_2) \\
V_\phi &= V_{\phi 2} (R/R_2) & \nu_f &= \nu_{f2} \\
\Sigma &= \Sigma_2 (R/R_2)^{-1/2}
\end{aligned}$$

As in the previous MDAF example above, the angular momentum transport further decreases toward the black hole ($\alpha(R) \rightarrow 0$). Nevertheless, the magnetic field still becomes increasingly dominant ($\beta \rightarrow 0$).

Matching this flow onto the MDAF structure at R_2 , the final “end game” scaling laws, then, are

ρ_c	=	2.1×10^{-5}	g cm^{-3}	$\alpha^{-2/3}$	m^{-1}	$\dot{m}$	$x^{-3/2}$	
$T_{c,e}$	=	10^9	K					
$T_{c,i}$	=	10^9	K				ϑ	
p_c	=	2.8×10^{12}	dyn cm^{-2}	$\alpha^{-2/3}$	m^{-1}	$\dot{m}$	ϑ	$x^{-3/2}$
H	=	7.7×10^5	cm		m			x
$\tau_{c,\text{es}}$	=	5.5		$\alpha^{-2/3}$		$\dot{m}$		$x^{-1/2}$
$\tau_{c,\text{ff/bf}}$	=	3.3×10^{-11}		$\alpha^{-4/3}$	m^{-1}	$\dot{m}^2$		x^{-2}
V_R	=	-9.2×10^9	cm s^{-1}	$\alpha^{2/3}$				$x^{-1/2}$
V_ϕ	=	5.2×10^5	cm s^{-1}	α^{-1}			$\vartheta^{3/2}$	x
Σ	=	32	g cm^{-2}	$\alpha^{-2/3}$		$\dot{m}$		$x^{-1/2}$
B_R	=	1.70×10^8	G	$\alpha^{-2/3}$	$m^{-1/2}$	$\dot{m}^{1/2}$		$x^{-5/4}$
B_ϕ	=	47	G	α^{-1}	$m^{-1/2}$	$\dot{m}^{1/2}$	$\vartheta^{9/4}$	x
B_Z	=	1.47×10^8	G	α	$m^{-1/2}$	$\dot{m}^{1/2}$		$x^{-5/4}$
V_{turb}	=	0						
V_{AR}	=	1.05×10^{10}	cm s^{-1}	$\alpha^{-1/3}$				$x^{-1/2}$
$\alpha(x)$	=	2.3×10^{-4}		α^{-1}			$\vartheta^{5/4}$	$x^{5/4}$
$\beta_p(x)$	=	2.4×10^{-3}		$\alpha^{2/3}$			ϑ	x
ν_f	=	0.093	Hz	α^{-1}	m^{-1}		$\vartheta^{3/2}$	
$L_{\text{acc,mdf}}$	=	3.0×10^{38}	erg s^{-1}	α^{-2}	m	$\dot{m}^2$		

(12.87)

Comparison of the MDAF solution in the previous section to the end-game (equations (12.85) and (12.87)) shows that many quantities (τ_{es} , V_R , V_ϕ , Σ , B_ϕ , V_{AR} , $\alpha(x)$, and ν_f) scale with radius in the same manner in both of these portions of the flow. Only ρ , p , H , $\tau_{\text{ff/bf}}$, B_R , and $\beta_p(x)$ change their character at $R = R_2$.

Comparison of end-game flow with the ADAF structure equations (12.52) and (12.58) is even more striking. Virtually all structural quantities are the same as the ADAF ones, save for a factor of α to some power and for the expected MDAF/ADAF differences (p , V_R , and B_ϕ). *This means that the MDAF end-game flow could potentially produce a similar observational signature and spectrum to that predicted for the ADAF, but with the strong magnetic advantages of the MDAF*

(rigidly rotating structure and bandwidth-limited noise). Even the predicted accretion luminosity $L_{\text{acc,mdf}}$ and efficiency

$$\epsilon_{\text{acc,mdf}} = 2.4 \left(\frac{\dot{m}}{\alpha^2} \right) \epsilon_{\text{acc}} \quad (12.88)$$

are nearly identical to that of the MDAF.

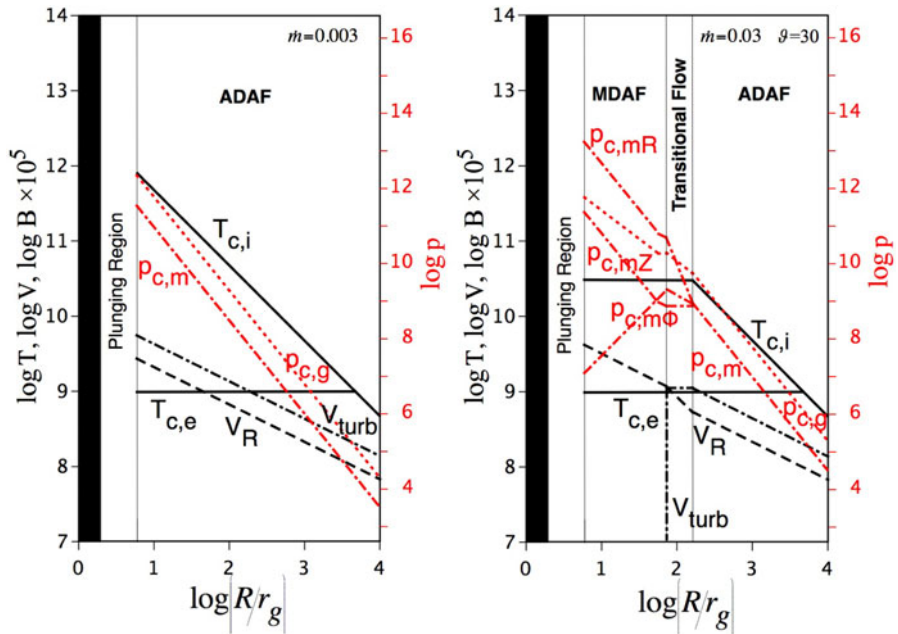


Fig. 12.10: ADAF vs. MDAF. Profiles of temperature, velocity, and pressure for the models in Figs. 12.6 and 12.9; in both cases the black hole mass is $10 M_{\odot}$, and the viscosity parameter is $\alpha = 0.3$. Black curves refer to the left ordinate, red to the right; the pressure scale is compressed by a factor of two and shifted slightly so that the magnetic pressure curves double as magnetic field curves ($\times 10^5$) on the left. That is, $p_{c,m} = 2 \times 10^{12} \text{ dyn cm}^{-2}$ gives magnetic pressure at right and $B_c \approx 8 \times 10^6 \text{ G}$ at left. *Left panel:* a pure ADAF with $\dot{m} = 0.003$. The two-temperature thermal structure begins inside of $R_A \approx 5000 r_g$, with the midplane central electron temperature $T_{c,e}$ remaining roughly constant inward toward the black hole; the central ion temperature $T_{c,i}$, however, stays virial in this region, supporting the ADAF structure in the vertical Z direction. Turbulent and radial velocity have a $R^{-1/2}$ Keplerian profile, while magnetic fields have a steep $R^{-5/4}$ one. *Right panel:* A combination ADAF/MDAF structure with $\dot{m} = 0.03$ and $\vartheta = 30$. The two-temperature ADAF is truncated at $R_0 = R_A/\vartheta \approx 160 r_g$, with transitional flow ending at $\sim 70 r_g$. Inside this point is the MDAF region, where there is *no* turbulence (only a rotating magnetosphere), B_{ϕ} ($p_{c,m\phi}$) drops linearly with radius, and the structure is supported vertically mainly by radial magnetic pressure $p_{c,mR}$.

Advantages of and Challenges to MDAF Theory

Fig. 12.10 shows the profile of the ADAF+MDAF accretion flow model for the X-ray hard state, along with plots of some important variables. The greatest advantage of the MDAF+ADAF model, as opposed to the ADAF-only one, is the strong magnetic field region inside R_1 . Strong fields near the black hole may explain the lack of high-frequency noise (turbulence) in some hard state sources and the presence of a strong (possibly rotational) quasi-periodic oscillation.

The greatest challenge to the above MDAF+ADAF model is that it does not yet explain the observed decrease in “truncation radius” R_0 with $\dot{m}$. On the other hand, this behavior is, in fact, predicted by evaporation theory (see the discussion near the end of Section 12.2.2). Therefore, a combination of evaporation theory to deal with the interplay between the ADAF and thin disk, plus MDAF theory to describe the innermost accretion flow near the black hole, may provide a fruitful avenue of approach toward building a full model of accretion in the X-ray hard state.

12.4 Magnetohydrodynamic Numerical Simulations of Accretion Flows

The words “simulation” and “modeling” are often used interchangeably in engineering and science. However, here we make a clear distinction between the two. By numerical “modeling” we mean solving on a computer a limited set of physical equations (*e.g.*, the α -model turbulent accretion disks we discussed in Section 12.2.2 above), but without all the approximations that we made in order to extract analytic scaling laws. (So, we set $p = p_g + p_r$, rather than choosing specific regions where $p \approx p_g$ or $p \approx p_r$.) By numerical “simulation”, on the other hand, we mean full two- or three-dimensional, time-dependent evolution of the flow velocity, magnetic, thermal and radiation fields, including details of the turbulence, shear, magnetic dynamo, angular momentum transport, *etc.* Models typically take perhaps a few minutes or so on a desktop computer to solve. Simulations, however, can take days or even weeks, on the world’s faster supercomputers, and require rather sophisticated software to create images and movies in order to understand the results.

12.4.1 The Current State of MRI Simulations

12.4.1.1 The Power and Promise of MRI Simulations

True simulations of accretion disks really began with the (re-)discovery of the magneto-rotational instability (MRI, see Section 9.6.3 and [367]) by Balbus and Hawley, and with the realization that the magnetic turbulence produced by the MRI

can transfer a considerable amount of angular momentum in accretion flows. That is, the turbulence is probably the “viscosity” mechanism that astrophysicists had been searching for. This allowed one to dispense with the assumption that $t_{\phi R} = \alpha p$ and, instead, simulate the details of MHD turbulent angular momentum transport from first principles. *The advent of these simulations was as important to the study of black hole engines as the discovery of nuclear reactions was to stellar structure.* Before understanding nuclear reactions, the stellar core was assumed to be a point source of luminosity; afterward, astrophysics began to understand the life cycles of stars, their birth, death, and contribution to the chemical makeup of the universe. The ability to understand the details of energy generation and angular momentum transport in accretion flows will have a similar impact on black hole physics.

However, it also has been said by some that everything accomplished before the MRI simulations had been started (*i.e.*, the α -model work) is wrong and will have to be completely re-done. This is decidedly *untrue*. The old work certainly will have to be updated, and some results may change, but there is far too much agreement with current observations of accreting systems (as we discussed above) to consider all analytic and semi-analytic accretion theory wrong. In the end, it is likely that the simulations will confirm much of the analytic work, but will give them a richer and more complete physical basis and provide a more detailed comparison to the observations than was possible with the simplified analysis.

12.4.1.2 Current Limitations of MRI Simulations

Such a grand set of simulated results, however, is still many years away. Despite recently moving to full three-dimensional MHD simulations in a more realistic general relativistic (Kerr) gravitational field, the MRI simulations are still in their infancy. There are two reasons for why this is true: limitations of present-day computers and limitations of the physics currently being simulated.

Despite the relentless advance in the power of supercomputers, present-day machines are still inadequate to simulate the microscopic details of MHD turbulence. A good simulation would range over length scales from the large accretion disk size (say, $1000 r_S$) down to the scale on which turbulence turns into heat (a few ion Larmor radii or somewhat less than a meter). For a $10 M_\odot$ black hole, this would require $(3 \times 10^9 \text{ cm} / 30 \text{ cm})^3 = 10^{24}$ separate simulation “cells” or grid points – much more than can fit in any current super computer – plus many years to run the simulation. As a result, current “global” simulations do not actually simulate the conversion of MHD turbulence into heat through reconnection and particle acceleration. Instead, they rely on an artificial *numerical* viscosity to dissipate the turbulence, and the energy balance must be handled carefully in order that the energy lost by this artificial viscous process be converged into heat. The best simulations cover accretion flows 10 – $100 r_S$ in extent, with cell sizes in the range $\sim 0.01 r_S$ ($\sim 300 \text{ m}$).

In computational terms, these limitations mean that the global simulations are not yet in the “convergence regime”. That is, as the computational mesh becomes more refined, the details of the simulations do not merely become more accurate,

they may fundamentally change character. Therefore, while simulations of accretion disks are very detailed, beautiful, and impressive, there is some nagging concern that these results may differ from those with much higher resolution – as much so as the α -models now differ from MRI simulations. Good numerical simulation groups, therefore, perform at least one or two simulations of a large set at a given resolution with much higher resolution in order to determine whether this non-convergence is serious or not.

The second limitation of current MRI simulations is that, as of this writing, only a very few groups have added the effects of radiative emission and transport. In most cases, the energy equation used is adiabatic and, except for the previously-mentioned numerical errors (artificial viscosity) that reconnect the magnetic field and convert it into heat, there is no significant computation of thermal energy balance. These “non-radiative accretion flow” (NRAF) simulations are useful for studying certain kinds of accretion flows, in particular the ADAF-type ones, which are almost completely non-radiative. An inspection of the previous sections in this chapter, however, shows that an understanding of most of the observable properties *and behavior* of accreting sources lies in the radiative and other thermal terms of the energy equation.

At the present time, therefore, we should consider the two types of accretion disk studies – α -model and MRI simulations – to be complementary, each with its own advantages. The α -models allow us to predict model spectra and compare these with observations; and they point the way for future simulations. But they undoubtedly oversimplify the angular momentum transport and the microscopic structure of the accretion flow. Current MRI simulations, on the other hand, give a detailed understanding of the physical processes that drive the accretion flow, but they cannot yet be compared to many observations of black hole engines. As time goes on, however, full numerical simulations certainly will win out, becoming the tool of choice as they have in other similar fields.

12.4.2 Local MRI Simulations

We shall discuss three types of MRI simulations in this section: local, semi-local, and global. The local, or “shearing box”, simulations do not suffer as much from not resolving the microscopic properties of the turbulence as the global ones do. Local simulations consider the behavior of only a small volume of differentially-rotating plasma. This volume can be quite small, so the *linear* resolution can be very high, even down to the dissipation length scale if one wishes. At this level one can study how MHD turbulence develops and how it transports angular momentum, but one cannot study the structure of an entire accretion disk.

Besides demonstrating the power of the MRI, the local simulations have produced several important results:

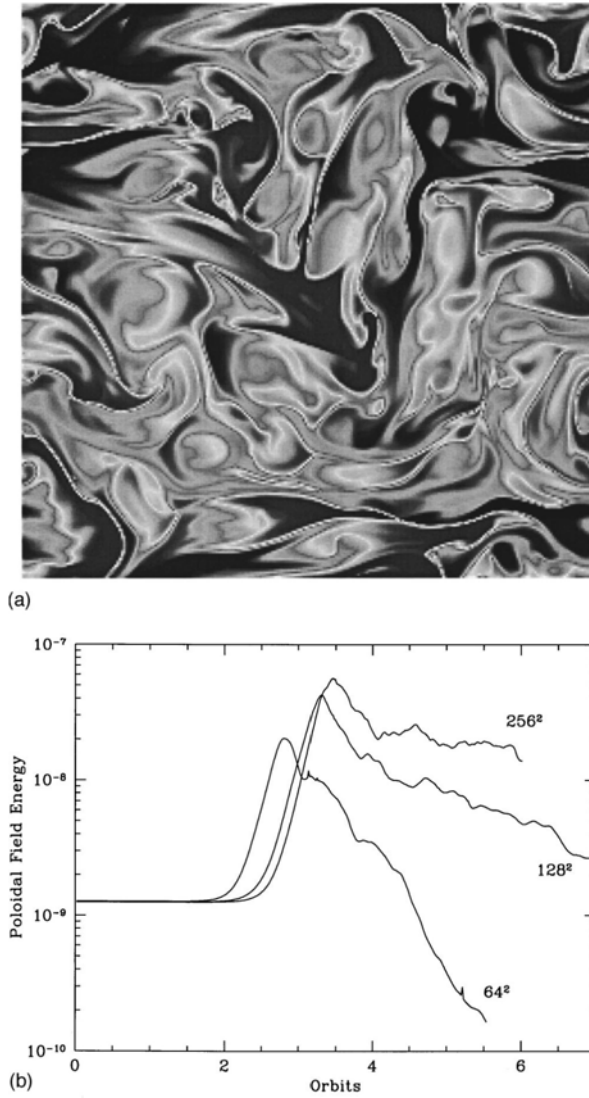


Fig. 12.11: Results of two-dimensional, axisymmetric simulations of MHD turbulence generated by the magneto-rotational instability. *Top:* A “shearing box” gray-scale plot of angular momentum with cylindrical radius R increasing from left-to-right and disk height Z increasing from bottom-to-top. *Bottom:* Curves showing evolution of the poloidal magnetic field energy, labeled with the grid resolution (in number of cells squared). While magnetic energy is lost with time in all these axisymmetric simulations, the more fine the grid, the less the decay rate becomes. Reprinted from Fig. 21 of [367] with permission; copyright (1998) by the APS.

- The MHD turbulence is three-dimensional, and therefore requires three-dimensional computer simulations in order to fully simulate a sustained turbulent dynamo. This is a numerical demonstration of the Cowling “anti-dynamo” theorem.
- The effective viscosity parameter produced by fully-developed MRI-induced turbulence is $\alpha \sim 0.01\text{--}0.1$ – similar to the ranges suggested by Shakura and Sunyaev in their classic 1973 paper.
- The “viscous” stress is actually made up of *both* magnetic (“Maxwell”) and turbulent (“Reynolds”) stress, in a fraction of 3 or 4 to 1, respectively. That is, 75–80% of the torque is produced by magnetic stresses, while 20–25% is produced by turbulent stress.
- The local power spectrum of MRI turbulent fluctuations is similar to that of incompressible turbulence $\propto f^{-11/3}$, the so-called Kolmogorov spectrum.

Fig. 12.11a shows a two-dimensional axisymmetric simulation of the MRI that was presented in Balbus and Hawley’s *Review of Modern Physics* paper in 1998 [367]. The figure presents a gray-scale plot of angular momentum, with cylindrical radius R along the horizontal and Z along the vertical axes. These results have many of the characteristics of the true, three-dimensional MRI. There is a rotational shear between the left and right-hand sides of the box, with the initial fluid velocity going into the paper (in the ϕ direction). The flow quickly develops into turbulence in a few dynamical times, and that turbulence is dominated by eddies that are nearly as large as the region that is being simulated. In a real accretion disk, the largest eddies would be nearly equal in size to the height of the accretion disk.

Fig. 12.11b shows how the magnetic field energy in this two-dimensional simulation evolves with time. While it rises quickly as the turbulence grows, it also decays quickly after about three orbits. The decay is related to the “anti-dynamo” theorem proved by Thomas George Cowling when he was at University College in Swansea, UK [450]. The theorem states that an axisymmetric MHD flow cannot sustain a growing magnetic field. The flow must be inherently *three* dimensional for a true dynamo to exist. While developed during Cowling’s studies of sunspots, the anti-dynamo theorem holds for any MHD flow, including accretion disks around black holes.

When true three-dimensional simulations are performed, however, the magnetic (and turbulent) stresses grow to a level far greater than that in the two-dimensional simulations; and they reach a saturation level that does not decay with time. Fig. 12.12 shows that, after a few orbits in a three-dimensional simulation, the Maxwell and Reynolds stresses saturate to values of 0.01–0.04 (relative to the local pressure) and in a ratio of 3–4 to 1. Their sum is ~ 0.05 , which is close to the canonical value of $\alpha \sim 0.03$ that we used in analytic models in the previous sections of this chapter.

Finally, Fig. 12.13 shows the “power fluctuation spectrum” of this turbulence. The spectrum is very steep – with $P \propto k^{-11/3}$ (a Kolmogorov-like spectrum; or $dP/dk \propto k^{-4.7}$), where $k = 2\pi f/c_s$, f is the frequency of fluctuations, and c_s is the sound speed. The spectrum peaks at the lower frequencies or longer wavelengths, where $\lambda = 2\pi/k \approx 1/4\text{--}1/3$ of the size of the region being simulated (*i.e.*, $1/4$ to $1/3$ the height of the disk at each radius).

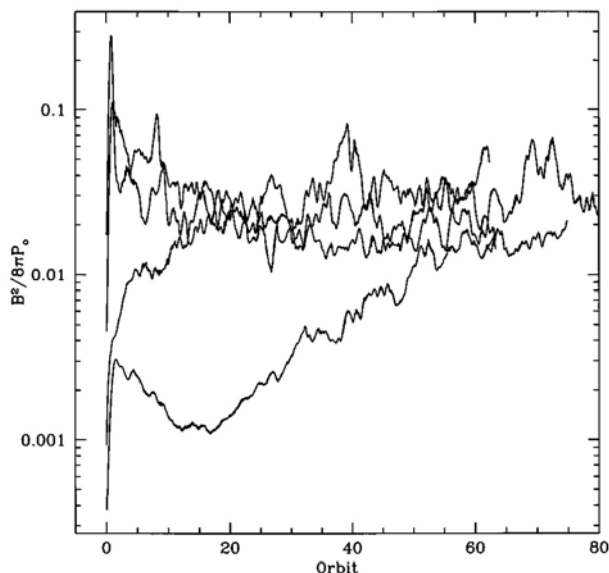


Fig. 12.12: Similar to Fig. 12.11b, but for MHD simulations in three dimensions. Note that the time axis extends for more than a factor of ten longer than the previous two-dimensional simulations and that the ratio of magnetic to gas pressure ($\sim\alpha$) rises quickly to a value of 0.02–0.03 and *does not decay*. The different curves start with differing random initial magnetic field conditions, but all lead to the same level of turbulence. Reprinted from Fig. 23 of [367] with permission; copyright (1998) by the APS.

12.4.3 Semi-Local MRI Simulations

By semi-local MRI simulations, we mean ones that cover one dimension of the disk rather completely (usually this is the vertical height [Z] of the disk) but do not simulate a full disk, even in two-dimensional axisymmetry. The most relevant and interesting simulations in this category are those by Neal Turner (of the University of California, Santa Barbara, and now at the Jet Propulsion Laboratory) in 2004 [453] and, later, by Omer Blaes (of UCSB), Shigenobu Hirose (of the earth Simulator Center in Yokohama, Japan), and Julian Krolik (of Johns Hopkins University) in the period 2009–2011 [452, 454]. These authors studied the very difficult problem of simulating the vertical structure of a radiation-pressure-dominated accretion flow using full radiation-MHD simulations of the MRI. Results of these studies are very important for understanding accretion flow at rates near the Eddington rate $\dot{M}_{\text{Edd}}$. A crucial question here is, do the magnetic turbulent stresses scale with the radiation pressure when it is dominant (as is assumed in the classical Shakura–Sunyaev α -model) or do they scale with the gas pressure or in some other manner? It is the straight $t_{\phi R} = \alpha p = \alpha(p_g + p_r)$ scaling that is responsible for the secular and thermal disk instabilities that are believed to be at work in high-luminosity X-ray

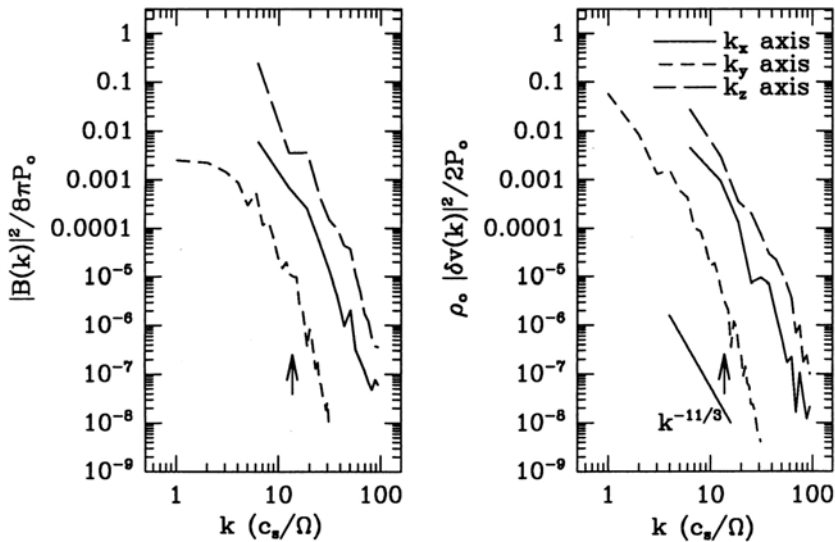


Fig. 12.13: Turbulent power spectra of magnetic energy (left) and kinetic energy (right), in local dynamical units. The wave number k is in units of Ω/c_s , which is approximately the inverse of the disk half-height H (eq. (12.9)). Arrows show the position in k -space of the critical wavenumber $k = 2\pi/\lambda_c$, below which the perturbations are unstable to the MRI. Waves of length $\sim H$, therefore, are the largest unstable modes and contain the most power in the spectrum. Reproduced from Fig. 4b of [451], by permission of the AAS.

sources like GRS 1915+105 and GX 339-4. The results of these studies have developed into a very interesting story that has not been told fully yet.

Turner [453] found that his results were not quite as simple as the α -model predicted. Fig. 12.14 shows the vertical structure of a radiation-pressure-dominated, MRI-turbulent disk; it actually has three layers, not one. First, there is a core at the disk midplane, in which magnetic and turbulent stresses are less than the gas pressure, and the plasma mass density is high. Above and below that there are two upper layers where escaping magnetic field produces dissipation even without much angular momentum transfer. (Note: all three layers in Turner's simulations were optically thick to escaping radiation; the disk photosphere was outside the simulation boundaries.) Nevertheless, because the angular momentum transport stresses are significantly *larger* than the gas pressure, and because they seem to follow the profile of the radiation pressure (at least in the intermediate and outer layers), the assumption that $t_{\phi R} \propto p_r$, and not proportional to p_g (or some other quantity), seems to be a reasonable one.

Subsequently, however, Hirose, Krolik, and Blaes [452] performed simulations that included layers still higher in the atmosphere, ones in the optically *thin* region of the disk. Curiously, they found that, even when the ratio $p_r/p_g \sim 10$, their simulations were *thermally stable*, not unstable, over times at least as long as 40 cooling times. That is, the rising and falling of the disk height found by Turner was

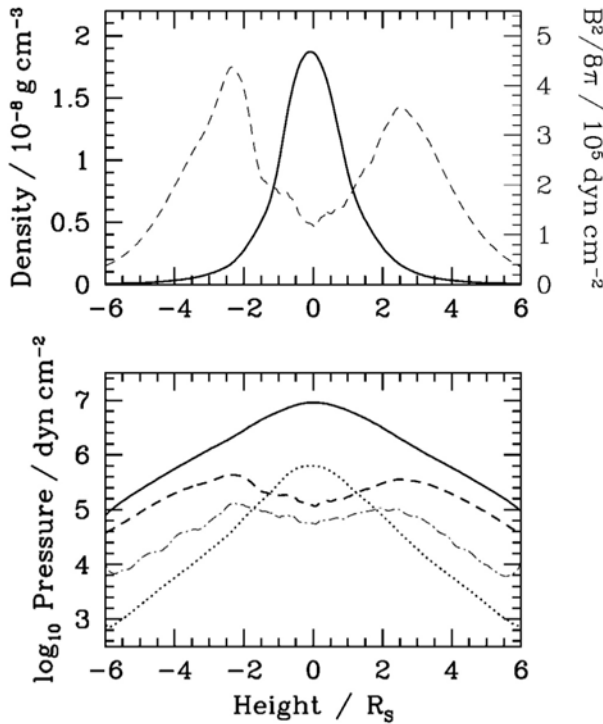


Fig. 12.14: Plots of the variation with vertical height of density and various pressure and stress components in a semi-local simulation of a radiation-pressure-dominated accretion disk. Zero on the horizontal axis denotes the disk midplane. *Top:* Linear plot of density (solid line) and magnetic pressure (dashed line), showing the “three-layer” structure of the disk: a dense, weakly-magnetized core near the midplane where the accretion stress is less than gas pressure, plus two higher-latitude layers (above and below the midplane) in which the magnetic pressure dominates over the gas pressure (but still is smaller than the radiation pressure). In simulations that also include the optically thin structure (which lies at even higher latitudes than shown here), the radiation pressure drops below the magnetic pressure there, creating another two layers above and below the midplane [452]. *Bottom:* Logarithm of different pressure and stress components: radiation, magnetic, and gas pressure (solid, dashed, and dotted lines); plus total accretion stress (dot-dashed line). Reproduced from Fig. 1 of [453], by permission of the AAS.

stopped and stabilized by losses of radiation in the optically thin layers. Yet, in a time-averaged sense (over many cooling times) the stress and total pressure still were related as $t_{\phi R} \propto p_r$!

Because this proportionality remains valid over many cooling times, this raises the possibility that the secular (*i.e.*, the “Lightman–Eardley”) instability, with its potential limit-cycle behavior, still will occur [454]. Indeed, the stress does appear to have an inverse relation with density.

One of the key aspects of the above thermal stability is that the rate of vertical radiation transport by diffusion is capped at the critical rate (above which the plasma *would* go thermally unstable). However, the disk still has more heat to get

rid of, so to accomplish this it brings this heat to the surface as photons trapped in an advective upwelling, *i.e.*, as thermal convection, but in a radiation-pressure-dominated environment. Such convection was actually predicted in the late 1970s and early 1980s by Mitchell Begelman and the author [455, 439] to be a general aspect of radiation-pressure-dominated accretion flows for many classes of viscosity laws, including the Shakura–Sunyaev α -model. (See also Fig. 13.3(right).)

Convection is, in fact, a natural response to a thermal instability (too steep a temperature gradient in a gravitating plasma). And the upwelling of photon dominated eddies, and the subsequent release of those photons in the optically thin atmosphere, will indeed constitute time-dependent variations in the thermal radiation emitted by the disk. What is not clear at present is if those variations will be similar in time scale and behavior to those that would have occurred had the accreting plasma not undergone this convection.

So, it appears from more detailed numerical simulations that the original Shakura–Sunyaev α -model assumption was a remarkably good one, possibly more accurate than those investigators themselves thought at the time. Furthermore, observations (through the display of instabilities in the high accretion rate states) seem to imply a similar conclusion. However, the correspondence between the thermal behavior of actual radiation-pressure-dominated black hole accretion disks and that of disks that have been simulated with complex radiation-MHD is not yet clear. In particular, the fate of the upwardly-convected photons is unclear. Are they dissipated as heat, creating a thermally-driven ADIOS-type wind (Section 13.2), or are they released as a strong radiation field, creating a radiation pressure-driven wind (Section 13.1)? Answers to these questions probably will need full global radiation pressure-dominated-simulations.

12.4.4 Global MRI Simulations

In global MRI simulations, the evolution of the entire region near the black hole is followed, not just a small piece of the disk. If the same sort of grid resolution is used (100–1000 cells in each dimension), then the linear size of a single fluid cell is much larger than the size of the entire box in the above local simulations! Therefore, the linear resolution of global simulations is much, much smaller than that of the local ones. Still, global simulations provide a reasonable approximation to real accretion disks, because the angular momentum transport is dominated by large-scale eddies, ones that are a significant fraction of the disk height.

12.4.4.1 Non-Radiative MRI Simulations

Since most global MRI simulations performed heretofore have been non-radiative, we will discuss those first. It is important to realize that, because the flow in these simulations is adiabatic, there really are only two free parameters to specify: the

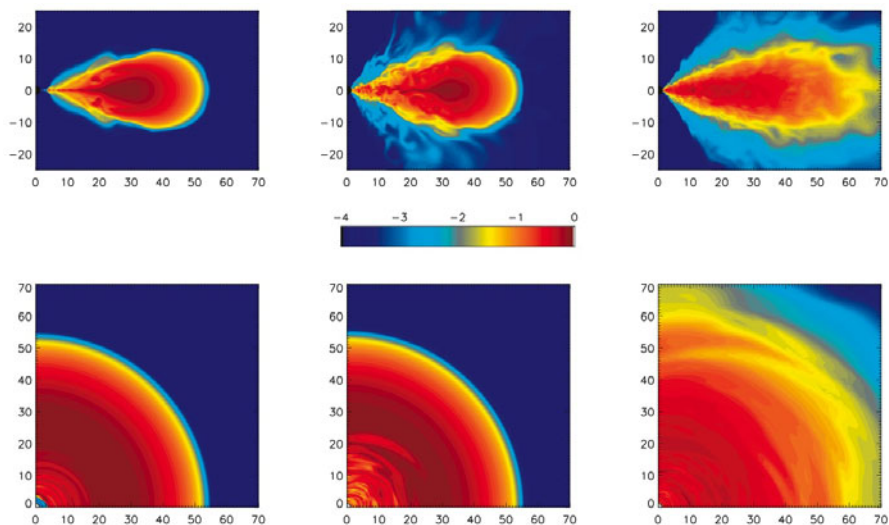


Fig. 12.15: Development of turbulence in a global, three-dimensional MRI simulation of accretion flow around a Kerr black hole with $j = 0.90$. From left to right, the three columns of figures show, respectively, $\log \rho$ at 1, 2, and 10 orbits after start of the simulation, with the axes given in units of r_g . Each top figure is a cut through the $\phi = 0$ meridional plane, while each bottom one is an equatorial cut. Only one quadrant of the entire sphere was simulated, with periodic boundary conditions at $\phi = 0$ and $\phi = \pi/2$. (Note that the thickness of the disk in the top panels is considerably smaller than, for example, that in Figs. 12.17a and 15.14A. This is because the simulations shown here were performed with a numerical code that conserves only internal energy, not total energy. In this case, quite a bit of viscous energy generated is lost from the grid without ever heating the plasma. This fact was pointed out in detail by [355]. Most groups now use total energy conservation codes, in which *all* viscous-generated heat is taken into account; this is absolutely necessary for studies that involve radiative losses, like those in Fig. 12.17.) Reproduced from Fig. 2 of [456], by permission of the AAS.

amount of matter to place in orbit about the black hole initially and the spin parameter of the black hole j . In the analytic theory above, we had to specify the black hole mass M , the viscosity parameter α , the accretion rate $\dot{M}$, and the radius R at which we wanted to evaluate the structure. (The black hole spin j also would be a parameter if we were to consider general relativistic models.) Because there is no radiation to consider in these MRI simulations, however, every black hole of any mass will appear the same, producing a very hot plasma with X-ray and γ -ray temperatures (10^{10-12} K). Furthermore, since these simulations generate viscosity from a nearly infinitesimal initial magnetic field, neither α nor B is a parameter of the flow. Rather, they are *results* of the simulation. Finally, instead of specifying an accretion rate, simulators specify the mass of an initial torus of material in orbit around the black hole; the simulation then determines the accretion rate, which now can be a function of position (R , Z), not just a constant value. (The position R in the disk at which quantities are evaluated is replaced by a full three-dimensional grid of cells in $(R$, Z , ϕ).)

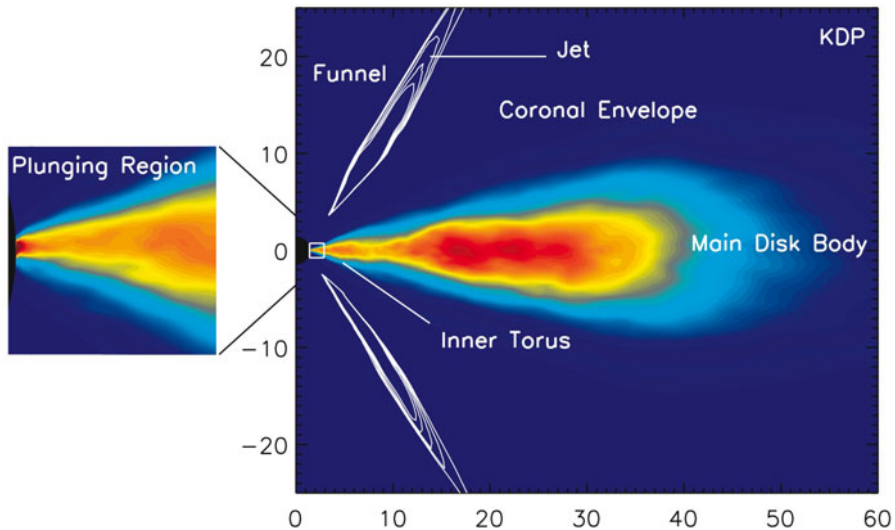


Fig. 12.16: Results of the simulation in Fig. 12.15, averaged over the tenth orbit. (Density ρ is now shown rather than $\log \rho$.) Key features are labeled: main disk, corona, funnel, and initial jet-like outflow along the funnel–corona boundary. Reproduced from Fig. 3 of [456], by permission of the AAS.

Fig. 12.15 and Fig. 12.16 show the results of one very complex simulation by Jean-Pierre De Villiers (of the University of Alberta, Canada) and his colleagues [456]. The simulation uses full general relativistic three-dimensional MHD flow in a Kerr metric with a rather high black hole spin of $j = 0.90$. The horizon radius, therefore, was only $r_H = 1.436 r_g$, with a substantial ergosphere between r_H and $r_E = 2 r_g$. Looking at the disk from the side and from the top in Fig. 12.15, we see that it is quite turbulent. This causes somewhat irregular accretion into the black hole. It also is quite hot, with $H/R \sim 0.2$ and with a corona having $H/R \sim 1$.

Matter in the regions of this corona near the rotation axis, especially near the “funnel–corona” boundary, is in a strange state [439]. In fact, if it is in equilibrium, then it must be spinning faster than the normal Keplerian rate for the cylindrical radius $R = r \sin \theta$. Gravity is balanced by the component of centrifugal force in the *spherical* radial direction by rotating at an angular frequency of

$$\Omega \geq \left(\frac{GM}{r^3} \right)^{1/2} \frac{1}{\sin \theta}$$

However, if there is a temporary rapid squeeze of the funnel toward the rotation axis (by a particularly powerful eddy, for example), then conservation of angular momentum during the squeeze demands that the new angular velocity be

$$\Omega' = \Omega \left(\frac{\sin \theta}{\sin \theta'} \right)^2 \geq \left(\frac{GM}{r^3} \right)^{1/2} \frac{1}{\sin \theta'} \left(\frac{\sin \theta}{\sin \theta'} \right)$$

where θ' is the new, smaller opening angle of the funnel. This new expression for Ω' exceeds the equilibrium value for angle θ' by the factor $\sin \theta / \sin \theta'$. During a squeeze, therefore, material at the funnel–corona boundary will move outward by centrifugal force alone. Material that is outflowing because of such processes is labeled “jet” in Fig. 12.16.

Whether or not this is a true jet, however, is not apparent from this simulation. Unless it is confined by a strong toroidal magnetic field, the very rapidly-rotating material may have a tendency to spray outward in the cylindrical direction, once it escapes the confines of the funnel. There no longer will be a strong inward pressure gradient holding the “jet” together. Furthermore, the outflow from the funnel will not be very fast – only of order 0.3–0.4 c , the escape speed from the distance from the black hole where the material was ejected (10–15 r_g). However, similar simulations by Jon McKinney (see Section 15.1.4 and Figs. 15.14 and 15.15), which lasted for more than a thousand orbital times, showed that this bipolar outflow does, indeed, eventually form a collimated, relativistic MHD jet [457]. The development of such a jet from these conditions is discussed in Chapters 14 and 15.

12.4.4.2 Radiative MRI Simulations

In 2009 P. Christopher Fragile (of the College of Charleston, South Carolina) and D. Meier have begun investigating the behavior of MRI simulations when radiative losses of the hot plasma (Bremsstrahlung, synchrotron, Compton cooling) are taken into account [355]. The purpose of these simulations is to see if the cooling leads to magnetically-dominated accretion flows, as was suggested earlier in this chapter.

Fig. 12.17 shows the results of two two-dimensional MRI simulations, one without and one with cooling. We find that the cooling does indeed cause the ADAF-like flow to collapse and the magnetic field to become more important. However, the structure of the disk interior to $\sim 100 r_g$ in the cooled simulation is not like that of the classical MDAF solution discussed above (equations (12.85) and (12.87), and Fig. 12.9). In fact, it appears to follow the low and high angular momentum accretion rate solutions of the *super*-Alfvénic inflow case. In the region of the flow from $R \approx 110$ down to $R \approx 70$, both the azimuthal velocity and magnetic field seem to remain roughly constant – just like the low angular momentum accretion rate, super-Alfvénic flow. Interior to $R = 70$, however, the azimuthal field appears to wind up to large values, and the azimuthal velocity is nearly vortical ($\propto R^{-0.7}$). $R = 70$ may be the point where the MDAF inflow changes character.

Both the low and high angular momentum accretion rate solutions have the following scalings

$$\begin{array}{ll}
 \rho \propto R^{-1} & V_R = \text{constant} \\
 T \propto R^{1/2} & \Sigma = \text{constant} \\
 p \propto R^{-1/2} & B_R \propto R^{-1} \\
 H = \text{constant} & B_Z \propto R^{-1} \\
 \tau_{\text{es}} \propto R^{-1} & V_{\text{AR}} \propto R^{-1/2}
 \end{array}$$

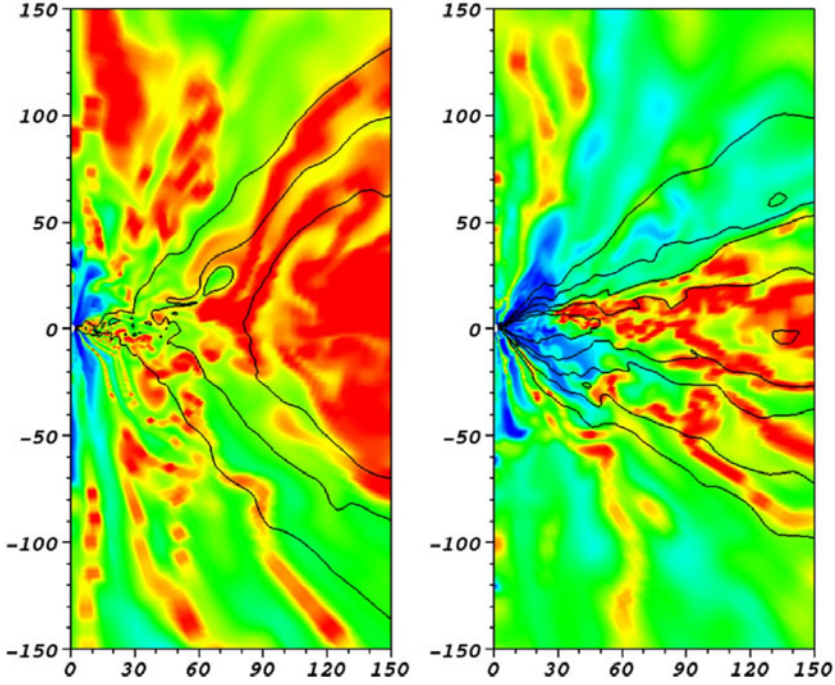


Fig. 12.17: The effects of adding radiative losses to a global MRI simulation like that in Fig. 12.15. Each panel shows a color, meridional plot of the plasma beta ($\log \beta_p$) for a different two-dimensional MRI simulation of accretion flow onto a Schwarzschild black hole. Axes are labeled in units of r_g . *Left:* No radiative cooling, like that in Fig. 12.15 and 15.14A. Note that the disk is dominated by low magnetic field plasma ($10 < \beta_p < 100$; yellow-red). (Note also that the “jet” is quite broken up and not well simulated; this was due to a numerical issue that occurred for very high magnetic field strengths and has been corrected in later versions of this simulation code.) *Right:* With the addition of radiative losses due to synchrotron, Compton, and thermal Bremsstrahlung processes. Note that cooling produces a thinner disk inflow and a much more inhomogeneous distribution of magnetic dominance; now there are regions of moderate ($\beta_p = 1$; green) and even high ($\beta_p < 0.1$; blue) magnetic field. Reproduced from Fig. 7 of [355], by permission of the AAS.

but they have different solutions for the azimuthal velocity and magnetic field:

$$V_\phi = \text{constant}$$

$$B_\phi = \text{constant}$$

and

$$V_\phi \propto R^{-1}$$

$$B_\phi \propto R^{-2}$$

respectively. The low angular momentum accretion rate case predicts that B_Z^2 magnetic pressure supports the flow against gravity in the R direction. The high angular

momentum accretion rate case, on the other hand, predicts that there is force balance between B_ϕ^2 magnetic pressure and $B_R B_\phi$ magnetic tension.

Why did the numerical simulations produce a different MDAF result from the one we were expecting? There are many possible reasons for this. First of all, since we have not yet developed a method of showing secular stability of the MDAF solutions, we must consider the possibility that some of them are unstable; if some are, then they would never form, in a numerical simulation, instead converting into some other form of magnetized inflow. Another possibility is that the axisymmetry of our computation has suppressed non-axisymmetric modes (*e.g.*, non-axisymmetric Parker modes) which would have caused the inflow to form a magnetic field more dominated by radial B_R . Nevertheless, the midplane inflow in Fig. 12.17 is not dominated by azimuthal magnetic field (except in the innermost regions), so it does not represent an inflow whose magnetic field might buckle if we had added a third (ϕ) dimension in the simulation. Another point to note is that the simulation does not cover an extremely large amount of time – only seven orbits. It is possible that the evolution might go through several of the MDAF solutions before settling into one. Indeed, sources like GRS 1915+105 are seen to enter a dozen or more accretion states, each with differing jet and QPO properties. It is possible that some of these states represent different MDAF solutions, plus other non-magnetically-dominated accretion flows. As we have seen from the above chapter, nature has many different choices of accretion inflow, even for the same black hole mass and accretion rate. As long as these choices are reasonably stable, any one of them might be chosen at some time in the lifetime of a source.

12.5 Chapter Summary

12.5.1 What We Know

While spherical accretion is of some interest in the astrophysics of black hole engines, *disk* accretion is by far the dominant mode used by nature to power them. Even a small amount of angular momentum in an infalling plasma will form a centrifugally-supported disk around an object as compact as a black hole.

Material in this disk cannot accrete further inward toward the black hole unless it loses some angular momentum. While angular momentum can be lost through wind or jet outflows, the primary means that accretion flows use appear to be magnetic. If the field is weak, the Keplerian shear drives turbulence through a magneto-rotational instability (MRI). The turbulent eddies transport angular momentum outward via turbulent and magnetic stresses in the accretion flow, allowing much of the matter to flow inward and only a small amount (that carrying the excess angular momentum) to move outward. The process of the turbulent magnetic field reconnecting and dissipating heats the plasma and provides the main conversion mechanism to change gravitational energy lost by the inflowing material into heat and light. Often

this complex magnetic turbulence is approximated as a viscous process, in order to study the basic observational properties of accretion flows. When the magnetic field is strong, turbulence cannot form. However, angular momentum still can be transported over strong magnetic field lines that connect the inner accretion flow to the outer part.

When a steady-state accretion flow solution is developed, it is important to consider its secular and thermal stability. Determining that a solution is stable probably means that the accretion flow is a viable candidate model for explaining observed black hole engines. However, if an instability is found, more work needs to be done. An unstable solution may mean that it is never realized in nature. On the other hand, the flow may enter an oscillatory or hysteretic type of time-dependent behavior, cycling through different structures and never coming to an equilibrium. The answer to these questions can be determined only through time-dependent numerical studies of the accretion flow.

For example, accretion solutions like the SLE model are thermally unstable and never seem to be realized in nature. Such a hot disk, however, can be stabilized by a strongly advective inflow, as in the ADAF model. In the case of accretion near the Eddington limit, the radiation-pressure-dominated “inner” portion of the accretion disk displays a cyclic behavior that may explain unstable behavior seen in bright X-ray sources and quasars.

In this chapter we also have spent much time discussing accretion at rates well below Eddington ($\dot{m} < 0.1$). Most models generally agree that a geometrically thick, advection-dominated inflow develops (possibly through disk evaporation) at a few thousand Schwarzschild radii. The internal temperature of this inflow is generally the virial temperature, and increases inward as R^{-1} . The differences between models all occur in the inner one or two hundred Schwarzschild radii, where the virial temperature significantly exceeds 511 keV or 5.9×10^9 K. The two-temperature ADAF model assumes that the flow inside this point maintains the ions (protons and other atomic nuclei) at this hot virial temperature (up to 10^{12} K), but the electrons stay much cooler at $\sim 1\text{--}2 \times 10^9$ K. The MDAF model, on the other hand, maintains that both ion and electron temperatures drop well below the virial temperature (again staying at $\sim 1\text{--}2 \times 10^9$ K), but the accretion inflow is supported by a strong, well-ordered magnetic field that develops from the cooling, magnetically turbulent material. Both models, therefore, predict similar hard X-ray emission for the low-accretion-rate, hard X-ray state. On the one hand, the MDAF concept is much newer, and not so thoroughly studied, as the thin disk and ADAF concepts. So it is rather early to judge its robustness, or staying power as a model, until much more work is done. Nevertheless, it has some nice properties that will be useful in explaining certain observed characteristics of black hole engines. A strongly-magnetized, rotating magnetosphere could be responsible for some observed QPOs, as well as for driving some of the powerful jets that are observed.

12.5.2 What We Do Not Know

There is a great deal we still do not know about black hole engines. We will touch on only a few items here. Perhaps the greatest mystery is the nature of the quasi-periodic oscillations that are observed in microquasars. There are so many of them, and they occur in so many different accretion states, that it is not clear if all are produced by the same physical mechanism, or if there are several mechanisms at work. Ideas that have been put forward are disk or coronal oscillations or possibly rotational phenomena (hot spots in the disk or entire semi-rigid magnetospheres). So far, no QPO has been seen in a “first principles” MRI simulations that can be identified readily with an observed quasi-periodic phenomenon. What is needed is not simply the identification of oscillatory behavior in a simulation, but, additionally, the association of that behavior with other phenomena in the simulation (*e.g.*, jets) in the same way that QPOs are observed to associate with those phenomena in nature. This would provide a clear indication of what was causing the QPO and why certain QPOs correlate with other accretion phenomena. A related question, that involves both observation and theory, is: do supermassive black hole engines produce QPOs and we simply do not have the signal-to-noise to detect them yet, or are they truly absent? If the latter, why would only stellar-mass black holes produce this phenomenon? Our understanding of accretion, therefore, will be quite incomplete until we can determine how and why these quasi-periodic oscillations are generated.

The next most pressing issue is the behavior of accretion flows when radiation pressure dominates (*i.e.*, near the Eddington accretion limit). Semi-local simulations seem to imply that the $t_{\phi R} = \alpha p$ assumption is a reasonable one, but there have not been any *global* radiation dominated simulations to study more fully how a black hole engine operates in a high-accretion-rate state. Do secular and thermal instabilities actually occur in the simulations? Do these scale with black hole mass as observed (*i.e.*, $\tau_{\text{cyc}} \propto m^{2/3}$)? *Are the cyclic phenomena observed in stellar and supermassive black holes (i.e., GRS 1915+105 and 3C 120) even the same physical process seen in different mass objects, or should we really not compare the two?*

A third issue involves the MRI itself. While there is strong theoretical evidence that MRI-induced turbulence is how nature transports angular momentum in actual accretion flows, there has yet to be a good observational method in which this can be independently verified. Contrast this with the processes of nuclear burning in stars, which can be verified by comparing observed cosmic isotopic abundance ratios with predicted ones, theoretical and observed globular cluster HR diagrams, and so on. Furthermore, no MRI simulation has yet been made of a single observed black hole system, let alone used to show that the MRI is definitely operating in that system. At the present time, the simulations are used to study the *physics* of the MRI process, not the *astrophysics* of black hole engines. We still do not have a good physical explanation of why accreting black holes enter into different accretion states, nor even good simulations of some simple radiative states like the original Shakura–Sunyaev disks. We, therefore, have a long way to go before computer simulations will help us fully understand the black hole engines that we observe. Nevertheless, ultimately it will be MRI simulations like these that will do just that.

Chapter 13

The Thermal Exhaust System: Radiation- and Thermally-Driven Winds and Jets

Exhaust systems vent spent fuel away from the combustion chamber in order to allow fresh fuel to enter. Black hole engines actually have five different exhaust systems, each with a different purpose. In the past two chapters, we already have dealt with three of those systems:

1. First, the emission of photons by the viscous-heated fuel carries away *heat and binding energy* from the fuel and keeps it at a low temperature, so it does not explode out of the accretion disk combustion chamber.
2. Second, the transport of angular momentum by the viscous processes vents *angular momentum* to the outer portions of the accretion disk, allowing the accreting material to sink further toward the black hole, where it can generate even more energy.
3. Third, the fuel entering the horizon, and becoming part of the accreting object, is another type of exhaust system that is unique to black hole systems. Once inside the hole, the heat of that material can no more affect the engine than if it had been ejected to infinity. The mass of that heat (ε/c^2), as well as the matter itself, becomes part of the black hole. So, not only are their engines very energy-efficient, black holes even use spent fuel to build their engine blocks.

In some accreting systems, these three comprise the major forms of exhaust that allow the engine to chug along in a steady state, for millions of years perhaps, being controlled mainly by the supply of fuel from outside.

However, there are two important instances in which a more active, and visible, exhaust system is needed. When the rate of accretion becomes so large that the pressure of radiation becomes important, then radiation-pressure-driven winds help to self-limit the amount of material that is allowed to fall toward the black hole or neutron star. Like #3 above, this fourth is a “mass” and energy exhaust system, but it involves expulsion because the engine is being fueled too fast. If this can be done in a steady, rather than explosive, manner, then the engine can again continue to operate steadily, but at a reduced, safer fueling rate. The fifth exhaust system is another “angular momentum” one that involves the complicating action of magnetic fields *outside* of the disk. It will be discussed in depth in Chapters 14 and 15.

13.1 Radiation-Driven Winds

The outward radiation pressure force can exceed the force of gravity in several ways. For example, when the photon luminosity exceeds the Eddington luminosity (equation (1.1)), the outward force on the free electrons can drive matter out of the system. The outward acceleration caused by the radiation pressure in this case is given by

$$a_r^{e-} = \frac{\kappa_{\text{es}} L_{\text{rad}}}{c 4\pi r^2} = \ell_{\text{rad}} \frac{GM}{r^2}$$

Here L_{rad} is the radiative luminosity of the source, κ_{es} is the electron scattering opacity, M is the mass of the central object, and ℓ_{rad} is the Eddington ratio (equation (2.17)). When $\ell_{\text{rad}} > 1$, the outward radiation pressure force is greater than that of gravity, and an outflow will occur.

The photon luminosity, however, does not have to be above the Eddington limit for a strong radiation-driven wind to develop. We can have $\ell_{\text{rad}} < 1$ but still have other kinds of opacity – particularly bound–bound absorption by atomic lines – cause a_r to approach or exceed the acceleration of gravity

$$a_r^{\text{lines}} = \frac{\kappa_{\text{lines}} L_{\text{rad}}}{c 4\pi r^2} \equiv (1 + \mathcal{M}) \ell_{\text{rad}} \frac{GM}{r^2} \geq \frac{GM}{r^2}$$

where the line radiation “force multiplier” ($\mathcal{M} \gg 1$) can cause the radiation pressure to greatly exceed the force of gravity. The same process occurs in massive star atmospheres, and line opacity-driven winds are believed to be the cause of strong outflows emanating from O and B stars. They likely will play a role in black hole engines as well.

13.1.1 Line-Driven, Sub-Eddington Winds

Several models have been proposed for line-driven winds from black hole accretion disks. One model that potentially can explain both the broad emission *and* absorption lines seen in quasars and Seyfert nuclei was developed in 1995 by Norman Murray, James Chiang, and their colleagues at the Canadian Institute for Theoretical Astrophysics (CITA) at the University of Toronto [125]. The general picture of the wind was shown earlier in Fig. 2.16(bottom) in Chapter 2. When the accretion rate in the disk is close to the Eddington rate ($\dot{m} > 0.2$), the “inner” disk is rather thick ($H \sim \text{constant} \propto \dot{m}$; see Section 12.2.2). In addition, if the “inner” region is thermally unstable, it will spend some of its Lightman–Eardley limit cycle time in a very thick ($H \sim R$) ADAF-like state as a hard X-ray source ($T \sim 10^8$ K). Beyond the “inner” region, the disk is flared ($H \propto R$), allowing its atmosphere to see the luminous disk center.

At a radius $R_{\mathcal{M}}$ where the surface temperature of the disk is cool enough to have only partially-ionized atoms ($T_* \simeq 10^{4-5}$ K), the atomic line absorption opacity at

wavelengths 300–3000 Å will be high enough to have a large force multiplier $\mathcal{M}$. From equation (12.11) this gives

$$R_{\mathcal{M}} = (5000\text{--}10^5) r_g m^{-1/4} \dot{m}^{1/4}$$

For a $1.4 M_{\odot}$ neutron star the middle of this range corresponds to about 5×10^9 cm, while for a quasar with a $10^9 M_{\odot}$ black hole this corresponds to about 2×10^{16} cm, or ~ 0.01 pc.

13.1.1.1 The Basic Disk Wind Equation

Section 12.1.1 introduced the analysis of spherical accretion and winds. The wind/accretion equation for this (non-relativistic) radiation-driven case is slightly different from that in equation (12.2)

$$\frac{d \ln V}{d \ln r} = - \left\{ \frac{\left[1 - (1 + \mathcal{M}) \tilde{\ell}_{\text{rad}} \right] (f GM/r) - 2a^2 + da^2/d \ln r}{V^2 - a^2} \right\} \quad (13.1)$$

r is the *spherical* radius, of course, and $a = (p_g/\rho)^{1/2}$ is the *isothermal* gas pressure sound speed (instead of the adiabatic sound speed used in equation (12.2)). It is the use of a instead of c_s that adds the radial gradient of a^2 on the right-hand side.

This equation also contains a few complications which we will note, but will not discuss in detail. The Eddington ratio ℓ_{rad} has been replaced by one that is attenuated by the electron scattering optical depth $\tilde{\ell}_{\text{rad}} \equiv \ell_{\text{rad}} \exp(-\tau_{\text{es}})$. This is necessary because of the disk geometry: electrons scatter radiation away from the disk surface; unlike spherical geometry, that radiation will never be scattered back again to act as an outward force. Also, there is a factor $f \equiv 1 - R_{\mathcal{M}}/r$ that causes the wind to start at zero velocity at the disk radius $R_{\mathcal{M}}$ where it is launched.

13.1.1.2 Wind Terminal Velocity

Note that the singular point of this equation is *not* where $V = a$. This is because of the two extra derivatives on the right-hand side that, if brought over to the left, would modify the denominator from $V^2 - a^2$ to something else. One of those derivatives is, of course, $da^2/d \ln r$. The other is hidden inside the force multiplier $\mathcal{M}$, which is a function of the optical depth in each absorption line τ_L . In a normal *static* situation, τ_L would be simply proportional to the entire extent of the atmosphere. However, when the atmosphere is accelerating, a single frequency of radiation from the central source ν is absorbed only over a narrow range of velocity $V - \Delta V/2$ to $V + \Delta V/2$, where $\Delta V \simeq V_{\text{th}}$ is the thermal width of the absorption line in velocity units. So, the velocity at which most of the absorption takes place depends on the central line frequency ν_0

$$V = \left(\frac{\nu_0}{\nu} - 1 \right) c$$

with $\nu < \nu_0$. During the wind acceleration, the range in velocities $V \pm \Delta V/2$ occurs over only a finite radius in the atmosphere: $\Delta r \approx \Delta V/|dV/dr| = V_{\text{th}}/|dV/dr|$. So, the force multiplier is a function of the local velocity gradient

$$\mathcal{M}(\tau_L) = \mathcal{M} \left(\kappa_L \rho \frac{V_{\text{th}}}{|dV/dr|} \right)$$

The actual terminal velocity of the wind V_∞ depends on the exact functional form for $\mathcal{M}$, but a rough approximation is given by

$$V_\infty \approx \left(\frac{4GM}{R_{\mathcal{M}}} \right)^{1/2}$$

This gives $\sim 3000 \text{ km s}^{-1}$ for the $1.4 M_\odot$ neutron star X-ray binary case and $\sim 25,000 \text{ km s}^{-1}$ for the $10^9 M_\odot$ quasar case. These results compare very well with observations of the neutron star X-ray binary Circinus X-1, which shows outflow at about 2000 km s^{-1} . And they also compare well with observations of broad absorption line (BAL) QSOs, which show outflows at $5000\text{--}50,000 \text{ km s}^{-1}$.

The model also predicts broad emission lines (BELs) interior to the BALs with Keplerian widths similar to those at $R_{\mathcal{M}}$. That is, the BEL gas is predicted to be the atmospheric gas of the disks, gas before it is significantly accelerated into a wind. These work out to about $V_\infty/2$ or $\sim 1500 \text{ km s}^{-1}$ in Cir X-1 and $12,000 \text{ km s}^{-1}$ in the BAL QSO case. These also are typical of the emission lines in these objects.

13.1.1.3 The X-Ray Heating Problem

The fact that radiatively-driven wind models work well when the accelerated gas is only 10^{4-5} K is rather surprising. Accreting black holes are strong X-ray sources, with radiation temperatures of $T_r \sim 10^8 \text{ K}$. This X-ray radiation is believed to originate near the center of the accretion disk from a corona above the cooler accretion disk. These hard photons should raise the temperature of the disk surface by Compton scattering the cooler electrons to an equilibrium “Compton” temperature of

$$T_C \approx \frac{1}{4} T_r \approx 2.5 \times 10^7 \text{ K}$$

However, the observed winds are nowhere nearly that hot. Cir X-1 shows X-ray absorption lines more typical of $\sim 10^6 \text{ K}$ and BAL QSOs show UV lines of $\sim 10^{4-5} \text{ K}$.

Why are these winds so cool? There are a number of possibilities:

1. Murray and Chiang suggested that cool, dense gas is drawn up from the surface of the disk, just inside the launching point, by the wind itself. This is thought to be accomplished by pressure gradients induced by the wind loss. This “hitchhiking

- gas” then shields the accelerating outflow from the X-rays but not from the UV photons that are doing the acceleration. (See Fig. 2.16(bottom).)
2. Another possibility is that Compton heating of the disk surface interior to R_M produces the shielding gas – in the form of a cool corona. That corona would attenuate X-rays from the central source, allowing a wind to be launched from a radius R_C (beyond which Compton heating of the surface would be unimportant). R_C would differ from R_M , of course, and it is not known at this time whether this would give as successful a model as the standard Murray and Chiang model.
 3. The presence of strong Compton-heating X-ray emission comes from the assumption that sources that show winds are simply normal objects, viewed along the surface of the disk. A third possibility, however, is that Cir X-1 and BAL QSOs are not like other X-ray sources and quasars. Indeed, they are thought to be radiating very close to the Eddington limit ($\ell_{\text{rad}} > 0.5$), whereas quasars are thought to be somewhat less luminous ($\ell_{\text{rad}} \sim 0.1$). When $\dot{m} \rightarrow 1$, the accretion flow may enter a cooler, optically thick, advectively-cooled “slim” disk state. That is, the inner hot corona that normally produces the X-rays may be quenched by the geometrically thick inflow. This would leave only the cooler disk accretion itself, implying central temperatures from equation (12.52) of $\sim 10^7$ K for Cir X-1 and $\sim 10^5$ K for BAL QSOs.
 4. In Section 16.4.1 we suggest yet another explanation: the outflow is not line-driven at all but, rather, an advection-dominated inflow–outflow solution (ADIOS) wind, driven by thermal pressure of the hot component in the two-phase medium that may develop in the unstable Shakura–Sunyaev “inner” disk region (Section 12.2.2). (ADIOS winds will be discussed later in this chapter in Section 13.2.) Again, it is the cool component that hitchhikes with the hot component, and it is lifted off the surface of the inflow by pressure gradients, but the latter develop simply because the ADIOS outflow is naturally unbounded to begin with.

13.1.2 Continuum-Driven, Super-Eddington Winds

When the disk accretion rate exceeds the Eddington limit ($\dot{m} > 1$), then the wind takes on quite a different character. Instead of being driven from the surface of the disk “middle” or “outer” region, the radiation force on the electrons *in the “inner” disk alone* potentially can exceed gravity ($\ell_{\text{rad}} > 1$). The “inner” region of the disk puffs up into a thick, radiation-pressure-supported, toroidal (almost spherical) structure. (See equations (12.35), where $H/r_{\text{isco}} \approx \dot{m}$.) If the luminosity generated exceeds L_{Edd} , an optically thick, essentially spherical wind will be driven from the disk “inner” region. While the optical depth of this wind is certainly a function of the mass-loss rate, generally the wind will be so dense that, even for moderately high super-Eddington accretion rates ($\dot{m} = 10\text{--}100$), the surface of last scattering or “scattersphere” will be at a radius of $10^{3-5} r_g$. This wind, therefore, could some-

times engulf a large fraction of the entire accretion disk, making the accreting black hole look more like a luminous nova than an X-ray source.¹

13.1.2.1 Physical Structure of Super-Eddington Winds

Introduction

The Injection Radius. The basic structure of a super-Eddington disk wind is shown in Fig. 13.1. Near the spherical radius²

$$r_i = 8.2 \times 10^5 \text{ cm } m \dot{m}$$

the disk is bloated by radiation pressure. In our model we will assume that all the material exiting the system in the wind is injected at r_i , although in reality some of it comes from the region $r < r_i$ and perhaps a little beyond as well. Furthermore, because the wind is super-Eddington, the material flowing in it will have the same equation of state properties as that in the disk “inner” region: $\kappa_{\text{es}} \gg \kappa_{\text{ff}}$ and $p_{\text{r}} \gg p_{\text{g}}$.

The parameters that describe the wind structure are similar to those used for the accretion disk flow itself: black hole mass m , accretion rate $\dot{m}$, and viscosity parameter α . However, instead of studying the inflow properties in cylindrical radius R near the disk equator, we will be interested in those of the wind along a spherical radius r , also centered on the black hole. As with the disk itself, α mainly tells how rapidly the gravitational energy released in the accretion flow at r_i is turned into heat. This ultimately determines the density and wind lift-off velocity at this radius, and it sets the stage for the ensuing wind structure.

The Singular/Critical Radius. Outside the injection radius, the wind density and pressure drop with spherical radius, and the outflow velocity increases. In this part of the wind the flow is *subsonic*, a slowly-expanding nearly hydrostatic envelope. At some radius r_c , however, the flow passes through a critical (“sonic”) point³ and becomes kinetic energy dominated. We define $\zeta > 1$ as the ratio of the critical radius to the injection radius

$$r_c \equiv \zeta r_i \tag{13.2}$$

¹ This subsection is based on D. Meier’s work in the late 1970s and early 1980s [455, 458, 459, 460].

² Again, the numerical coefficients of our results here differ by factors of order unity from those in [458, 459, 460], mainly because of slightly different assumptions about opacity, *etc.* and in order to match standard, slim, and super-Eddington models.

³ This is a critical point in the classical causal sense, as discussed in Section 9.6.2, where flow characteristics appear when accelerating from subsonic to supersonic flow. It also is a singular point of the wind equation.

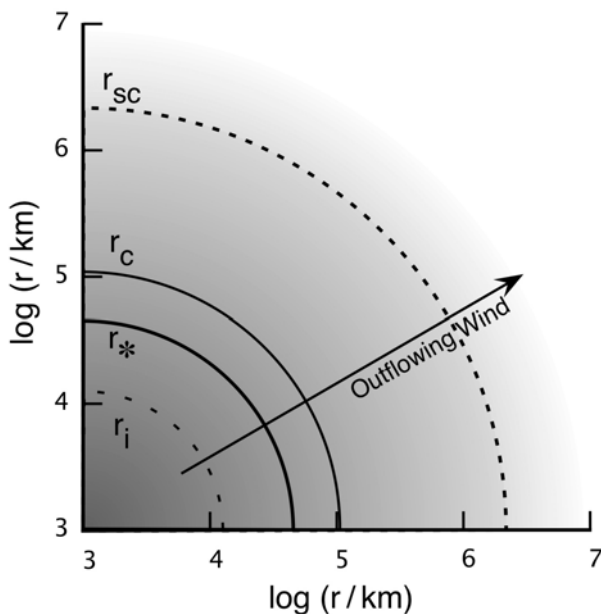


Fig. 13.1: Schematic example of a super-Eddington wind, drawn in $\log r$ - θ space. Matter and thermal energy are injected by some unknown process at r_i . (In most discussions in this book that process will be accretion flow along the equatorial plane.) The hydrostatic envelope expands subsonically, accelerating from r_i to r_c , the radius of the critical point in the spherical wind. Near r_c the acceleration of the wind (due to optically thick radiation pressure gradients) drops off rapidly with radius when the speed of the outflow exceeds the radiation pressure sound speed. At r_* the outflow becomes optically thin to free-free and bound-free photon absorption, but it still is quite optically thick to electron scattering. It is not until $r = r_{sc}$ that τ_{es} becomes less than unity. The figure is a graphical representation of the warm model for super-soft X-ray source P098 shown in the right-hand column of Table 16.4 and Fig. 16.13, with estimated outflow rate of $\sim 3.5 \times 10^{-5} M_{\odot} \text{ yr}^{-1}$. Models with the photosphere forming in the supersonic part of the flow ($r_* > r_c$) also are possible for even higher rates of mass outflow.

Beyond r_c the flow becomes increasingly difficult to accelerate, because its speed is so fast that it leaves the system before dynamical forces have much time to act upon it. The wind achieves a terminal velocity not much greater than the critical velocity, so $V_{\infty} \approx V_c$.

The Scattersphere Radius. At some point the wind will become optically thin to electron scattering, at the scattersphere radius r_{sc} . Most models that we will describe here will do so in the supersonic portion of the flow; that is, $r_{sc} > r_c$. (This is *not* to be confused with the photosphere radius r_* , which is the surface of last *absorption*. Because electron scattering dominates, the last free-free absorption will occur well interior to the scattersphere, so $r_* < r_{sc}$.)

The Wind Structure Equations

Even though there may be some initial departure from spherical symmetry near the injection region, generally the wind will be spherically symmetric throughout most of its structure. Furthermore, because $r_i \gg r_S$ for large $\dot{m}$, and because the wind velocities are not too close to the speed of light, we will be able to use Newtonian gravity and radiation–hydrodynamic equations (5.63) to (5.65) instead of the relativistic versions. In spherical coordinates and symmetry, they are

$$\frac{1}{r^2} \frac{d(\rho V r^2)}{dr} = 0 \quad (13.3)$$

$$V \frac{dV}{dr} + \frac{GM}{r^2} + \frac{1}{\rho} \frac{dp_g}{dr} = \frac{\kappa_{es} H_r}{c} \quad (13.4)$$

$$V \frac{d\varepsilon}{dr} + \frac{(p + \varepsilon)}{r^2} \frac{d(r^2 V)}{dr} = -\frac{1}{r^2} \frac{d(r^2 H_r)}{dr} + \rho \dot{q} \quad (13.5)$$

where V is the outward radial velocity, $\varepsilon = \varepsilon_r + \varepsilon_g \approx \varepsilon_r$ and $p = p_r + p_g \approx p_r$ are the total internal energy and pressure, H_r is the outward flux of radiation, and the energy generated by viscous dissipation occurs only in the injection region

$$\begin{aligned} \rho \dot{q} &= \alpha \frac{(p + \varepsilon)}{\tau_{\text{dyn}}} & (r \leq r_i) \\ &= 0 & (r > r_i) \end{aligned}$$

Here τ_{dyn} is the local dynamical time, which is essentially the free-fall time.

Subsonic Structure of the Wind

The Injection Region. The structure of the injection region is similar to that of the Shakura–Sunyaev “inner” disk region or the slim disk, but now it is advectively cooled by an *outward* wind flow. Defining L_T as the total power generated in the injection region

$$L_T \equiv \int_i 4\pi r^2 \rho \dot{q} dr \approx 4\pi r_i^3 \alpha \frac{(p + \varepsilon)}{\tau_{\text{dyn}}}$$

Assuming $V \ll a \ll c_s = (\Gamma p / \rho)^{1/2}$, the structure equations (13.3) to (13.5) become

$$\Delta \dot{M} = 4\pi r_i^2 \rho_i V_i \quad (13.6)$$

$$L_{\text{rad}} = \left(1 - \frac{p_g}{p}\right) L_{\text{Edd}} \quad (13.7)$$

$$\Delta \dot{M} \left[\frac{(p + \varepsilon)}{\rho} \right] = L_T - L_{\text{rad}} \quad (13.8)$$

and

$$L_{\text{rad}} = 4\pi r^2 H_r$$

is the total radiative luminosity. We can ignore the gas partial pressure $p_g/p \ll 1$ in this radiation-pressure-dominated case, so equations (13.6) to (13.8) tell us that, in equilibrium, the photon luminosity of the wind is very close to the Eddington luminosity ($L_{\text{rad}} \simeq L_{\text{Edd}}$), and the wind mass-loss rate is just enough to carry off the remaining enthalpy generated above the Eddington luminosity

$$\Delta \dot{M} = \frac{(L_T - L_{\text{rad}})}{(p + \varepsilon)/\rho}$$

Because the injection region is optically thick, the radiative flux can be expressed in the diffusion approximation as

$$H_r = -\frac{c}{\kappa_{\text{es}} \rho} \frac{dp_r}{dr} \quad (13.9)$$

This leads to an alternate, but familiar, expression for the radial force balance

$$\frac{1}{\rho} \frac{dp}{dr} = -\frac{GM}{r^2}$$

where we now have combined p_g and p_r into the total pressure p . This gives, approximately, $p/\rho \approx GM/4r$, which is exact when $p \propto r^{-4}$. Equation (13.8) then becomes

$$\Delta \dot{M} = \frac{4\pi c r_i}{\kappa_{\text{es}}} \left(\frac{L_T}{L_{\text{Edd}}} - 1 \right) = \dot{M} \left(\frac{L_T}{L_{\text{Edd}}} - 1 \right) \quad (13.10)$$

All we need now is an expression for the total power generated L_T in terms of $\dot{M}$, and we will be able to solve for the wind mass-loss rate and all of the other variables in the injection region (ρ_i , V_i , *etc.*).

To determine L_T , we assume a very simple accretion model in which the rate $\dot{M}$ is composed of two components:

1. the material that makes it only down to the injection radius r_i and then is ejected ($\Delta \dot{M}$), and
2. the material that makes it all the way to the black hole ($\delta \dot{M}$) and releases $\epsilon_{\text{acc}} \dot{M} c^2$ of energy.

The total, then, is

$$\dot{M} = \Delta \dot{M} + \delta \dot{M} \quad (13.11)$$

The total amount of energy generated in this simple model is

$$L_T = \epsilon_{\text{acc}} \delta \dot{M} c^2 + \frac{GM}{r_i} \Delta \dot{M} \quad (13.12)$$

Equations (13.10) to (13.12) now can be solved, and they have two solutions. The first is the no-wind solution

$$\begin{aligned} L_T &= \epsilon_{\text{acc}} \dot{M} c^2 \\ \Delta \dot{M} &= 0 \\ \delta \dot{M} &= \dot{M} \end{aligned}$$

which includes the slim disk solutions, when $\epsilon_{\text{acc}} < 1$. The second is the super-Eddington wind solution⁴

$$\begin{aligned} L_T &= L_{\text{Edd}} \left[1 + \frac{r_g}{\epsilon_{\text{acc}} r_i} (\dot{m} - 1) \right] = \left(2 - \frac{1}{\dot{m}} \right) L_{\text{Edd}} \\ &\approx 2 L_{\text{Edd}} \end{aligned} \quad (13.13)$$

$$\Delta \dot{M} = \dot{M} - \dot{M}_{\text{Edd}} \quad (13.14)$$

$$\delta \dot{M} = \dot{M}_{\text{Edd}} \quad (13.15)$$

When a super-Eddington wind is produced, only one $\dot{M}_{\text{Edd}}$ can finally accrete onto the black hole. The rest of the original accretion rate emerges in a strong wind, and does so from the rather large radius $r_i \approx \dot{m} r_g / \epsilon_{\text{acc}}$. For $\dot{m} \gg 1$ only *two* Eddington luminosities of power are generated, one to produce the photon radiation and one that is advected outward as enthalpy in the wind. Equations (13.13) to (13.15) now allow us to solve for the conditions in the injection region. The solutions will be given below with the rest of the subsonic structure.

Structure from r_i to the Critical Point. For $r > r_i$ we can ignore the energy generation term in equation (13.5). So, for the remainder of the wind envelope, we have the simpler energy equation

$$V \frac{d\varepsilon}{dr} + \frac{(p + \varepsilon)}{r^2} \frac{d(r^2 V)}{dr} = -\frac{1}{r^2} \frac{d(r^2 H_r)}{dr} \quad (13.16)$$

Still assuming $V \ll c_s$, the solution to equations (13.3), (13.4), and (13.16), the radial subsonic structure can be expressed in terms of the injection region conditions

$$\rho = \rho_i (r/r_i)^{-3} \quad (13.17)$$

$$V = V_i (r/r_i) \quad (13.18)$$

$$p = p_i (r/r_i)^{-4} \quad (13.19)$$

with the photon luminosity remaining constant with r , essentially supporting the envelope with its radiation pressure. If we now define a dimensionless *spherical*

⁴ It is helpful here to remember that the definition of $\dot{M}_{\text{Edd}}$ includes the efficiency ϵ_{acc} (equation (11.4)).

radial coordinate (*cf.* the cylindrical $x \equiv R/r_{\text{isco}}$ in Chapter 12)

$$z \equiv \frac{r}{r_i}$$

we can express the entire subsonic structure of the wind in the same manner as we did for different regions of the accretion flow. For $\dot{m} \gg 1$

r_i	$=$	$8.3 \times 10^5 \text{ cm}$	m	$\dot{m}$		(13.20)
ρ	$=$	$2.1 \times 10^{-5} \text{ g cm}^{-3}$	α^{-1}	m^{-1}	$\dot{m}^{-1/2} z^{-3}$	
p	$=$	$2.3 \times 10^{15} \text{ dyn}$	α^{-1}	m^{-1}	$\dot{m}^{-3/2} z^{-4}$	
V_Z	$= +$	$9.2 \times 10^9 \text{ cm s}^{-1}$	α		$\dot{m}^{-1/2} z$	
τ_{es}	$=$	5.5	α^{-1}		$\dot{m}^{1/2} z^{-2}$	
$\Delta \dot{M}$	$=$	$1.4 \times 10^{18} \text{ g s}^{-1}$		m	$\dot{m}$	
$L_{\text{acc, wind}}$	$=$	$2.1 \times 10^{38} \text{ erg s}^{-1}$		m		

The structure in the injection region is found by setting $z = 1$, of course, and the structure at the critical point is found by setting z equal to the (still unknown) parameter ζ (equation (13.2)).

A quick comparison of equations (13.20) with the slim disk equations (12.52) or the “inner” disk region equations (12.35) shows that all these structures are essentially identical, if we recognize the injection region as occurring at a cylindrical radius $x_i = \dot{m}$. That is, the distance from the black hole, interior to which the super-Eddington wind ejection occurs, grows linearly with accretion rate; and the conditions there are uniquely determined by $L_{\text{acc}} = L_{\text{Edd}}$, $p = p_r$, $\kappa = \kappa_{\text{es}}$, and the accretion flow being geometrically thick from that point inward.

Location of the Critical Point and the Supersonic Structure of the Wind

In principle, the structure beyond the critical point is even easier. As we saw in the line-driven wind case, above the critical point the inertia terms dominate the conservation of momentum

$$V \frac{dV}{dr} \approx 0$$

yielding a simple solution for equations (13.3), (13.4), and (13.16)

$$\begin{aligned} \rho &= \rho_c (r/r_c)^{-2} = \rho_i \zeta^{-3} (r/r_c)^{-2} \\ V &= V_c = V_i \zeta \\ p &= p_c (r/r_c)^{-3} = p_i \zeta^{-4} (r/r_c)^{-3} \end{aligned}$$

In addition, the scattersphere occurs when $\tau_{\text{es}} \approx \kappa_{\text{es}} \rho_{sc} r_{sc} = 1$ or at the radius

$$r_{sc} = \tau_{\text{es}, i} \zeta^{-1} r_i$$

It is clear, therefore, that the key to solving the supersonic structure of super-Eddington winds is determining the position of the critical point ζ . Doing so is not easy, and will require learning some new concepts in radiation-driven wind theory.

Two Possible Energy Equations, Two Possible Sonic Points. The electron scattering, radiation-driven wind equation presents an interesting puzzle. On the one hand, if the energy equation is radiative (the diffusion of photons dominates over their dragging), then equation (13.16) becomes

$$\frac{1}{r^2} \frac{d(r^2 H_r)}{dr} = 0$$

implying that the radiative flux varies as $H_r \propto 1/r^2$, causing the radiation pressure force to act like an anti-gravity. Combined with the conservation of momentum, the wind equation is similar to equation (13.1) but with a few simplifications (the force multiplier $\mathcal{M} = 0$, no attenuation in spherical geometry, and ignoring the factor f)

$$\frac{d \ln V}{d \ln r} = - \left\{ \frac{(1 - \ell_{\text{rad}}) (GM/r) - 2a^2 + da^2/d \ln r}{V^2 - a^2} \right\} \quad (13.21)$$

If the wind is largely isothermal ($da/dr = 0$, as might be the case when Compton heating is important) or if the temperature drops as a power law in radius ($d \ln a^2/d \ln r = -n$), then the gradient in the numerator of the right-hand side would simply change the coefficient of a^2 , and the critical point still would lie at the gas pressure sonic point $r = r_{s,g}$, where $V = a$.

On the other hand, if the energy equation were adiabatic, not radiative, then equation (13.16) would be

$$V \frac{d\varepsilon}{dr} + \frac{(p + \varepsilon)}{r^2} \frac{d(r^2 V)}{dr} = 0$$

which, as we have seen many times, has the solution $p = A \rho^\Gamma$ (see Sections 5.2.2 and 9.3.1). In this case, we can use equation (13.9) to rewrite the radiation force term in equation (13.4) as a simple gradient of the radiation pressure, and then add the two terms in p_r and p_g . This leads to a quite different wind equation,

$$\frac{d \ln V}{d \ln r} = - \left\{ \frac{(GM/r) - 2c_s^2}{V^2 - c_s^2} \right\} \quad (13.22)$$

with a different critical point – at $r = r_s$, where $V = c_s$. Since the total sound speed c_s (which includes radiation pressure) is much, much larger than the gas pressure sound speed a , both the terminal velocity of the wind ($V_\infty \sim c_s$) and the radius of the critical point will be quite different from what they were in the radiative case. *The energy equation is key to determining where the critical point lies.*

More Important Radii: The Wind Trapping Radius and the Adiabatic Radius. Where does the wind change from adiabatic flow to radiative? One might think that

this occurs at the wind trapping r_{tr} , where the photon flux H_r equals the advective $F_{\text{adv}} \approx V(p + \varepsilon)$, or

$$r_{\text{tr}} = \frac{c}{\kappa_{\text{es}} \rho V} \quad (13.23)$$

This is essentially the same trapping radius discussed for spherical accretion, but now it describes the *outflow* of trapped photons, not inflow.

However, it is not specifically the trapping radius where the flow becomes radiative. The proper terms to compare are not the fluxes, but the *divergence* of the fluxes in the energy equation (13.16). Using the subsonic solution in equations (13.20), we find that the magnitude of each advective term on the left-hand side of equation (13.16) is approximately

$$\left| V \frac{d\varepsilon}{dr} \right| \approx \frac{12 V p}{r}$$

The magnitude of the radiative term on the right-hand side can be found by substituting H_r from the momentum equation (13.4)

$$\begin{aligned} \left| \frac{1}{r^2} \frac{d(r^2 H_r)}{dr} \right| &= \left| \frac{c}{\kappa_{\text{es}} r^2} \left[\frac{d}{dr} \left(r^2 V \frac{dV}{dr} \right) + \frac{d}{dr} \left(\frac{r^2}{\rho} \frac{dp_g}{dr} \right) \right] \right| \\ &\approx \frac{3c}{\kappa_{\text{es}} r^2} a^2 & V^2 < a^2 \\ &\approx \frac{3c}{\kappa_{\text{es}} r^2} V^2 & a^2 < V^2 \end{aligned}$$

Therefore, the radius r_{ad} of transition from adiabatic to radiative flow depends on its location relative to the two sonic points (*i.e.*, how fast the wind is expanding locally):

$$\begin{aligned} r_{\text{ad}} &= \frac{c_s^2}{a^2} r_{\text{tr}} & r_{\text{ad}} < r_{s,g} \\ &= \frac{c_s^2}{V^2} r_{\text{tr}} & r_{s,g} < r_{\text{ad}} < r_s \\ &= r_{\text{tr}} & r_s < r_{\text{ad}} \end{aligned} \quad (13.24)$$

which is good to within factors of order unity.

We now can identify four different wind cases, depending on the location of the adiabatic radius. The optically thin case

- Case A ($r_{\text{ad}} < r_{s,c}$): The flow is radiative and optically thin, essentially the same as the line-driven wind in Section 13.1.1 but without the force multiplier. The critical point lies at $r_{s,g}$ and the terminal velocity is roughly the sound speed there $a(r_{s,g})$.

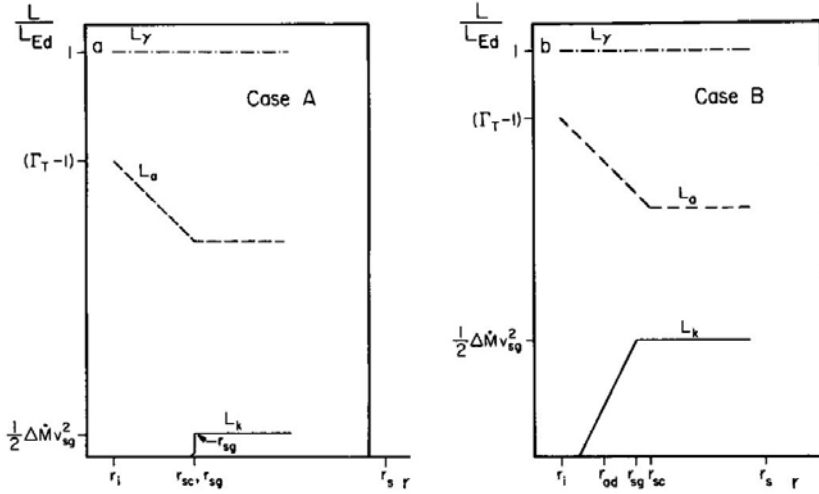


Fig. 13.2: Physical structure of the super-Eddington wind cases discussed in the text [458]. $L_\gamma = L_{\text{rad}}$ is the photon luminosity, which for the most part is the Eddington limit. L_a and L_k are the advective and kinetic luminosity, respectively. In these simple models the wind is assumed to accelerate until it reaches a critical point, whereupon the outward speed saturates at a characteristic value, and a uniform velocity spherical wind ensues. In most cases L_γ dominates throughout the outflow, with the kinetic luminosity the lowest of the three. (*Continued opposite.*)

and three optically thick cases

- Case B ($r_{\text{ad}} < r_{s,g} < r_s$): Radiative, but optically thick winds with the critical point also at $r_{s,g}$ and the $V_\infty \approx a(r_{s,g})$.
- Case C ($r_{s,g} < r_{\text{ad}} < r_s$): Gas sonic point is in the adiabatic region, but the total sonic point is in the radiative portion of the wind. The critical point actually lies at r_{ad} (see below) and the terminal velocity is $V_\infty \approx V(r_{\text{ad}})$.
- Case D ($r_{s,g} < r_s < r_{\text{tr}}$): First introduced by Martin Rees [461], these winds are fully trapped, and have all their acceleration take place in the adiabatic portion of the outflow. The critical point is at r_s and the terminal velocity is approximately $V_\infty \approx c_s(r_s)$. These winds also have enormous mass-loss rates, very high L_T/L_{Edd} , and even have super-Eddington *photon* luminosities ($L_{\text{rad}} > L_{\text{Edd}}$).

Each of these cases is illustrated schematically in Fig. 13.2.

Location of the Critical Point in Case C. For super-Eddington accretion flows onto black holes that produce winds, the strange and interesting Case C is, by far, the most common wind structure. In this case, however, the critical point of the wind equation actually is at the adiabatic radius r_{ad} . The arguments are as follows. The critical point cannot be at the gas pressure sonic point $r_{s,g}$, because the flow there is still advective and adiabatic ($r_{s,g} < r_{\text{ad}}$). The wind flow there is determined by equation (13.22). The wind near $r_{s,g} \ll r_s$, therefore, still behaves adiabatically

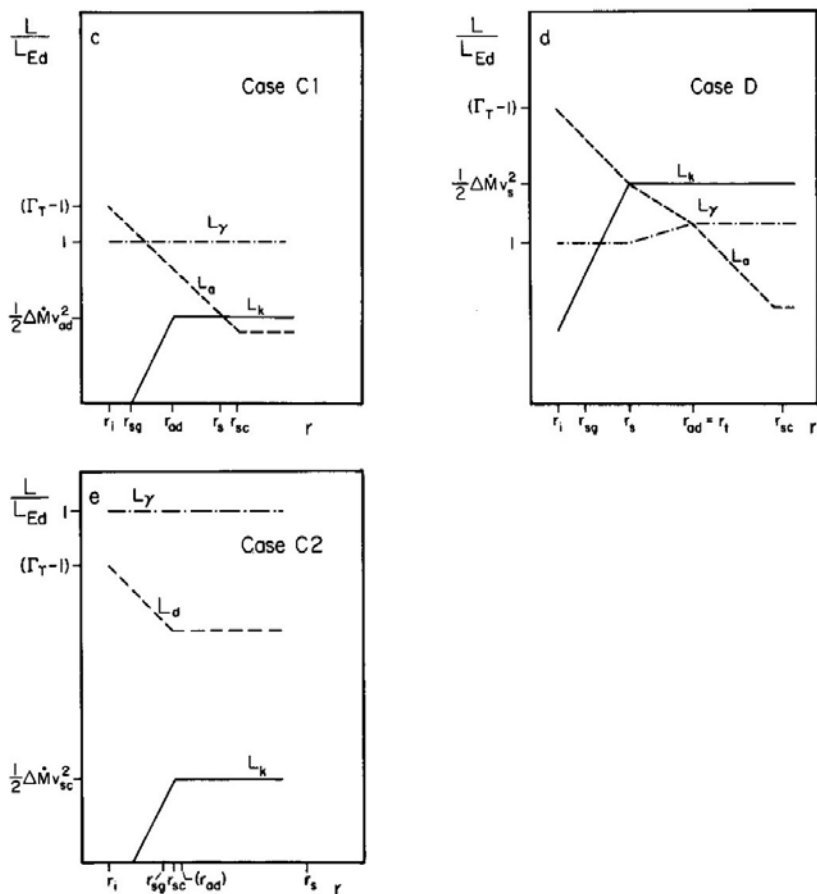


Fig. 13.2: (Continued from opposite.) However, Case C1 and D (see text) begin with the advective luminosity dominating; most of that is lost to adiabatic expansion, with some converted to kinetic luminosity. In Case D the photons are trapped in the outflow until $L_\gamma \geq L_a$, so advective luminosity can be converted into photon luminosity in the supersonic region, allowing the latter to exceed the Eddington limit. Reproduced by permission of the AAS.

and continues to accelerate through $V = a$, without losing any steam (*i.e.*, without the numerator of the equation (13.22) changing sign).

On the other hand, the critical point cannot be at the total pressure sonic point r_s , either. In Case C, the flow at the total pressure sonic point is radiative ($r_{ad} < r_s$), so the wind equation there is given by equation (13.21). But, at r_s the flow there must be very gas pressure supersonic ($V = c_s \gg a$). So, from equation (13.21), the velocity must be constant with radius, and must have had this uniform flow throughout the radiative portion of the wind. So, if the wind must have $V \propto r$ in the adiabatic portion of the wind, and constant velocity in the adiabatic portion, then the only place where the acceleration can become unimportant is in the vicinity of

the adiabatic radius r_{ad} . *The critical point in Case C must be at $r_c = r_{\text{ad}}$, and the terminal velocity must be approximately the velocity of the wind at that point $V_\infty \approx V(r_{\text{ad}})$.* This behavior of the critical point in Case C-like winds was confirmed with numerical simulations [462].

Wind Structure from r_c Outward. Equations (13.17) to (13.19) now can be solved for r_c and, hence, the position of the critical point

$$\zeta = \frac{r_{\text{ad}}}{r_i} = 1.41 \alpha^{-1/2} (1 - \dot{m}^{-1})^{-1/4} \approx 1.41 \alpha^{-1/2}$$

For $10^{-2} < \alpha < 1$ the critical point lies within an order of magnitude of the thickened portion of the super-Eddington accretion disk.

This value of ζ now can be used to solve for the supersonic physical structure of the wind envelope for Case C winds coming from a super-Eddington accretion disk around a black hole:

$r_c =$	$1.17 \times 10^6 \text{ cm}$	$\alpha^{-1/2} m$	$\dot{m}$			(13.25)
$\rho =$	$1.48 \times 10^{-5} \text{ g cm}^{-3}$	$\alpha^{-1/2} m^{-1}$	$\dot{m}^{-1/2}$	z^{-2}		
$p =$	$1.63 \times 10^{15} \text{ dyn}$	$\alpha^{-1/2} m^{-1}$	$\dot{m}^{-3/2}$	z^{-3}		
$V_Z = +$	$1.30 \times 10^{10} \text{ cm s}^{-1}$	$\alpha^{1/2}$	$\dot{m}^{-1/2}$			
$\tau_{\text{es}} =$	3.9	$\alpha^{-1/2}$	$\dot{m}^{1/2}$	z^{-1}		
$r_{\text{sc}} =$	$3.2 \times 10^6 \text{ cm}$	$\alpha^{-1/2} m$	$\dot{m}^{3/2}$			

where z is, once again, r/r_i (*not* r/r_c), and r_{sc} is the surface of last scattering, which lies well beyond the photosphere r_* . Equations (13.20) and (13.25) give the same results at the critical point $z = \zeta$.

Cases A, B, and D are not as interesting here, but are discussed in [458, 459, 460] if the reader is interested.

13.1.2.2 Thermal Structure of Super-Eddington Winds

Even with only Case C to consider, the determination of the thermal structure still can be complex. The injection region could be optically thin or thick to free–free emission. And, if the latter, the surface of last absorption (the true photosphere) could lie in the subsonic or supersonic portion of the wind. Also, Comptonization effects could be quite important in the cooling and spectrum formation. All of these will affect the appearance of black holes accreting well above the Eddington limit.

For simplicity, we will consider only one of these cases here: an injection region optically *thick* to free–free absorption, with the photosphere forming in the *subsonic* portion of the wind and with Compton cooling being unimportant. That is,

$$r_i < r_* < r_c < r_{\text{sc}}$$

This case is the most important one for finding super-critically accreting black holes.

Thermal Structure of the Injection Region

Under these conditions the thermal structure of the injection region, like the physical structure, is essentially identical to that in the “inner” disk region or in the slim disk, if we set $x_i = \dot{m}$

$$\begin{array}{llll} T_i & = & 3.1 \times 10^7 \text{ K} & \alpha^{-1/4} m^{-1/4} \dot{m}^{-3/8} \\ \tau_{i,*} & = & 5.7 \times 10^{-3} & \alpha^{-17/16} m^{-1/16} \dot{m}^{29/32} \\ L_{\text{acc,wind}} & = & 2.1 \times 10^{38} \text{ erg s}^{-1} & m \end{array} \quad (13.26)$$

Photosphere Conditions

With T_i now determined, we can calculate the observed color temperature of this wind by assuming thermal equilibrium, using the radiation pressure law in the sub-sonic portion

$$\frac{T_*}{T_i} = \left(\frac{p_*}{p_i} \right)^{1/4} = z^{-1}$$

and setting the absorption optical depth (equation (12.32)) to unity. The result is

$$\begin{array}{llll} r_* & = & 4.5 \times 10^4 \text{ cm} & \alpha^{-17/28} m^{27/28} \dot{m}^{85/56} \\ \rho_* & = & 0.133 \text{ g cm}^{-3} & \alpha^{23/28} m^{-25/28} \dot{m}^{-115/56} \\ T_* & = & 5.7 \times 10^8 \text{ K} & \alpha^{5/14} m^{-3/14} \dot{m}^{-25/28} \\ \tau_{\text{es}*} & = & 1.89 \times 10^3 & \alpha^{3/14} m^{1/4} \dot{m}^{-15/28} \end{array} \quad (13.27)$$

Fitting to Soft X-Ray Observations

Suppose that we do find a candidate source for a super-Eddington wind (*i.e.*, black hole or neutron star accreting with $\dot{m} \gg 1$). Then it is possible to develop and test the super-Eddington wind model for such a source with only an estimate of the source’s bolometric photon luminosity L_* and its color temperature T_* . The central object mass can be determined by assuming that it is radiating at the Eddington limit

$$m = \frac{L_*}{1.25 \times 10^{38} \text{ erg s}^{-1}} \quad (13.28)$$

and a very useful parameter⁵

$$\xi_{\text{sub}*} \equiv \alpha \dot{m}^{-5/2} = 1.92 \times 10^{-8} m^{3/5} T_*^{14/5}$$

⁵ The subscript “sub*” refers to a *sub*sonic photosphere; $\xi_{\text{sup}*} \equiv \alpha \dot{m}^{-3}$ would be the counterpart in a wind with a *super*sonic photosphere.

which is a function only of m and $T_{*6} \equiv T_*/10^6$ K. Like T_{*6} , several other properties of the wind at the photosphere (r_* , ρ_* , and the optical depths $\tau_{\text{es}*}$ and $\tau_{\text{ff/bf}*}$) are functions only of m and $\xi_{\text{sub}*}$. This means that we can derive many properties of the wind without assuming anything about the α parameter.

However, if we wish to develop a full wind model, we do need to choose a reasonable value for the viscosity/energy-generation parameter α . For example, from equations (13.27) the accretion rate can be cast in the following manner

$$\dot{m} = 1.22 \times 10^3 \alpha^{2/5} m^{-6/25} T_{*6}^{-28/25} \quad (13.29)$$

but this expression is not a strong function of α . In fact, this estimate for $\dot{m}$ varies by only a factor of 6 over the entire range $0.01 < \alpha < 1.0$. With a reasonable value for α , we can derive a large amount of information about the candidate super-critically accreting black hole candidate – densities, temperatures, pressures, optical depths, and radii for the injection region, photosphere, wind critical point, and last scattering surface – from equations (13.20), (13.25), (13.26), and (13.27). Some of the most useful expressions are the following

$$\begin{array}{llll} r_i &= 1.01 \times 10^9 \text{ cm} & \alpha^{2/5} & m^{19/25} T_{*6}^{-28/25} \\ r_* &= 2.2 \times 10^9 \text{ cm} & & m^{3/5} T_{*6}^{-17/10} \\ r_c &= 1.43 \times 10^9 \text{ cm} & \alpha^{-1/10} & m^{19/25} T_{*6}^{-28/25} \\ r_{\text{sc}} &= 1.36 \times 10^{11} \text{ cm} & \alpha^{1/10} & m^{16/25} T_{*6}^{-42/25} \\ V_\infty &= 3.7 \times 10^8 \text{ cm s}^{-1} & \alpha^{3/10} & m^{3/25} T_{*6}^{14/25} \end{array} \quad (13.30)$$

It may seem from the dimensional coefficients above that $r_* > r_c$. However, for any black hole above a couple of solar masses we find that, $r_c > r_*$, so the photosphere actually does form in the subsonic part of the wind. Therefore, the subsonic assumption that leads to equations (13.27) is valid for most astrophysical black holes but perhaps not for all neutron stars accreting supercritically.

13.1.2.3 Slim Disks vs. Super-Eddington Winds

There now appear to be two different, yet seemingly plausible, models for how a black hole appears when it accretes above the Eddington rate. In Chapter 12 we presented the “slim disk” model as the $\dot{m} > 1$ version of the standard Shakura–Sunyaev disk solutions. In this latter model, *most or all* the accreting material enters the black hole and, in the standard solution, *no* wind is emitted. In the super-Eddington wind model, on the other hand, most of the accreted material is recycled outward in a powerful wind, except for the approximately one $\dot{M}_{\text{Edd}}$ that is needed to generate the energy. In both cases, the *photon* luminosity remains at $L_{\text{rad}} \approx L_{\text{Edd}}$. (This is required for an equilibrium solution.) However, the appearance of the two solutions is quite different for the same accretion rate. For example, in the slim disk case, for $\dot{m} = 10$ –100 the object appears as an X-ray source; from equation (12.11)

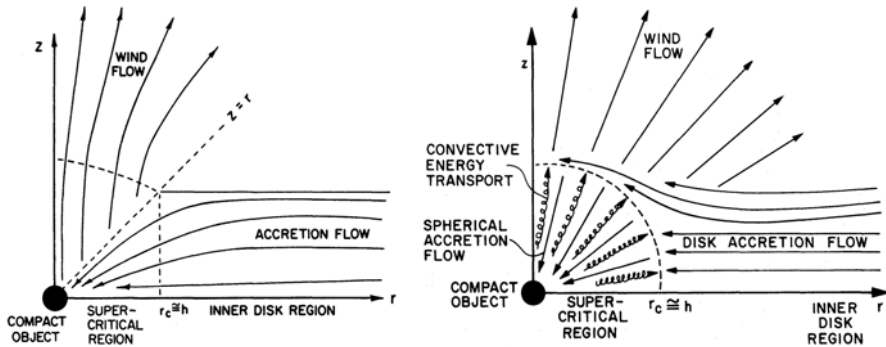


Fig. 13.3: Two possible wind scenarios for the super-Eddington accretion case. *Left:* Slim disk driving a loosely bipolar outflow; unless viewed from a polar direction, the central slim disk still would be visible to observers. *Right:* Essentially spherical outflow, with the optically thick wind covering, and obscuring, most or all of the accretion inflow in the central regions. Adapted from Fig. 6 of [455]. Reproduced by permission of the AAS.

$$T_{*, \text{slim}} \approx 3.0 \text{ keV } m^{-1/4} (\dot{m}/30)^{1/4}$$

On the other hand in the wind case (equation (13.27)), the object appears more like a cool, super-soft X-ray or even far-UV source

$$T_* \approx 0.67 \text{ keV } \left(\frac{\alpha}{0.03} \right)^{5/14} m^{-3/14} \left(\frac{\dot{m}}{30} \right)^{-25/28}$$

To make matters worse, *as the accretion rate increases to even higher super-Eddington values, the slim disk gets hotter while the wind gets cooler.*

Here we suggest a solution to this dilemma that is physically plausible and works well for a number of sources discussed in Chapter 16. First, we take the same view for the slim disk super-Eddington accretion flows (which are *radiation* pressure and advection-dominated) that Blandford and Begelman took for ADAFs (which are *gas* pressure and advection-dominated) – namely, that *the plasma in any super-Eddington inflow is unbounded, and a wind outflow usually will occur.* Fig. 13.3 then shows two possible accretion/wind scenarios for the super-Eddington accretion case that were suggested by the author in 1977 and 1979 [463, 455]. Fig. 13.3(left) shows a super-Eddington wind that is driven above the slim disk in a moderately bipolar fashion. The slim disk inflow can be seen near the equatorial regions, stretching all the way to the black hole. Unless the system is viewed from the north or south polar directions, an observer will be able to see down into the central regions of the engine and view the slim disk itself, plus any X-ray radiation that it may produce. Fig. 13.3(right), on the other hand, shows a source in which the wind ejection is largely spherical. Since the wind covering factor in this case is nearly 1.0, the slim disk, and any X-ray emission it produces, would be obscured from an observer.

In Section 16.2.4 we tentatively identify both steep power-law X-ray binaries and super-soft sources as being black holes accreting above the Eddington limit

($\dot{m} > 1$). The SPL sources still show some hard emission, but the super-soft ones show none. While the SPL sources are believed to be accreting near the Eddington limit, models of the super-soft ones suggest accretion rates of $\dot{m} = 60\text{--}130$ (Table 16.4). Therefore, the transition from bipolar to spherical wind (where the slim disk becomes fully covered by the wind) should be somewhere between $\dot{m} \approx 1$ and $\dot{m} \approx 60$. For demonstration purposes we will choose this boundary to occur at $\dot{m}_{\text{SEW}} \sim 10$. Clearly, a lot of detailed investigation will be needed in order to test this picture and better understand super-Eddington winds and their covering properties.

13.2 Thermally-Driven, ADIOS Winds: A Glimpse into the Disk–Wind Interaction of All Advection-Dominated Accretion Flows

In Chapter 12 we showed that there are two closely-related advection-dominated accretion flows with the same physical structure equations, but with very different thermal structure:

- Slim disks, with *super*-Eddington accretion rates. These are so optically *thick* to radiation that in most of the flow photons are trapped and advected toward the black hole.
- ADAFs, with very *sub*-Eddington accretion rates. These are very optically *thin*, but the thermal (heat) energy is trapped in the flow and advected toward the black hole.

The only real difference between the two is in their effective optical depth τ_* . They both are radiatively inefficient and advective.

In the previous section we discussed the possibility that slim disks, in fact, drive strong super-Eddington winds that are *so* optically thick that they can enshroud the advective, radiation-pressure-dominated accretion flow near the black hole and, instead, appear more like a variable O or soft X-ray star. In the steady state, these powerful winds serve to cool the accretion flow of excess heat.

Now, because slim disks and ADAFs are such close analogues of each other, it is natural to ask the question, “Are cooling winds possible in the ADAF case as well as the slim disk case?” The answer is, “yes”, and such winds were studied by Roger Blandford (then at Caltech) and Mitchell Begelman of University of Colorado, in 1999 [464]. These authors gave these winds the rather clever name *ADIOS* (advection-dominated inflow–outflow solution) *winds*. In these models, ADAFs say “*adios*” to most of their inflowing matter as it is expelled to large radii by a strong thermally-driven outflow.

13.2.1 ADIOS Basics and Motivation

The likelihood of outflow from advective accretion flows was noted in Narayan and Yi's original paper in 1995. The energy per unit gram, sometimes called the “Bernoulli” constant

$$Be = \frac{1}{2} (V_\phi^2 + V_R^2) - \frac{GM}{R} + \frac{\Gamma}{\Gamma - 1} \frac{p}{\rho} \quad (13.31)$$

was found to be *positive* in all advection-dominated accretion flows. We can see this immediately by plugging our own numbers from equations (12.8), (12.9), (12.19), and (12.51) into the above Bernoulli constant expression. This gives $Be = [3\Gamma/4(\Gamma - 1) - 1/2 + 9\alpha^2/32] V_{phi}^2$. For any reasonable adiabatic index $Be > 0$. For example, $\Gamma = 5/3$ gives $Be \approx 1.4 V_\phi^2$, and $\Gamma = 4/3$ gives $Be \approx 2.5 V_\phi^2$.

We might wonder whether this positive Bernoulli constant is due to our simple assumptions of full Keplerian support in the R direction and/or to pressure support in the Z direction. Indeed, we really did not solve the full advective physical structure in our simple treatment in Chapter 12. In part, this is true, and our simple models *do* overestimate the Bernoulli parameter. However, a more careful analysis (see [441, 464]) that solves the radial structure self-consistently shows that the positive energy of the inflow still persists. The more correct answer is, for any adiabatic index Γ ,

$$Be = V_\phi^2$$

which still is always greater than zero. Therefore, even if the inflowing material were ejected to infinity, it nevertheless would have positive energy. So *advection-dominated inflows are unbound*, and, because $V_\phi^2 \propto 1/R$, the material's “unbinding” energy *increases* as it accretes closer to the black hole.

The reason for this excess energy is not due to the assumption of self-similar radial scaling of the advection-dominated accretion solution. Rather, it is due to a very fundamental property of those flows. At a given radius R , more energy is transported *outward* via viscous torque *from the flow interior* – and dissipated into heat at R – than is advected from the outer regions beyond R . This deposition of excess energy from the flow interior occurs at *all* radii until the inflow approaches the black hole horizon. There the torques begin to vanish, and no more viscous energy is transported outward. So, for most of the inflow, the energy content grows inward, remaining unbound by an amount equal to V_ϕ^2 .

13.2.2 ADIOS Equations

The only real difference between Blandford and Begelman's ADIOS models and the classical ADAF models is the *assumption* in the former that mass loss occurs *everywhere* in the advective inflow. And this assumption is implemented in such a

way that the Bernoulli constant is kept *negative* and the inflowing matter bound to the black hole throughout the inflow. In order to obtain radial scaling laws, these authors assume that the accretion rate decreases with decreasing radius as⁶

$$\dot{M}(R) = \dot{M} \left(\frac{R}{R_0} \right)^n \quad R < R_0 \quad (13.32)$$

where $\dot{M}$ is the original accretion rate, and the exponent of R lies in the range $0 \leq n < 1$. This, unfortunately, adds another free parameter (n) to the problem, but without a clear indication of how thermal winds lift off the inflow, there is not much more that can be done.

With $\Gamma = 5/3$ (gas pressure domination), the structure equations (12.8) and (12.9) now become

$$V_\phi^2 - \frac{GM}{R} + \left(\frac{5}{2} - n \right) \frac{p}{\rho} = 0$$

with

$$\frac{p}{\rho} = \frac{GM}{R} \left(\frac{H^2}{R^2} \right)$$

(Note that $V_\phi \neq (GM/R)^{1/2}$ now.) We also have a new conservation of angular momentum equation that replaces equation (12.10)⁷

$$\dot{M}(R) R V_\phi - 4\pi R^2 \alpha p H = \lambda \dot{M}(R) (GMR)^{1/2}$$

The right-hand side of the above equation represents the angular momentum that is *not* lost in the wind but, instead, is passed on to smaller radii by the accretion flow. Its scaling with R was chosen to preserve the scaling $V_\phi \propto R^{-1/2}$, and $\lambda > 0$ is assumed, since the left-hand side represents the net angular momentum transported inward by competing accretion and viscous transfer processes.

Finally, in order to determine the disk temperature we need to balance viscous heating with advective cooling by the wind in the energy equation

$$4\pi R \alpha p H V_\phi - \dot{M}(R) \left(\frac{1}{2} V_\phi^2 - \frac{GM}{R} + \frac{5}{2} \frac{p}{\rho} \right) = \epsilon_{\text{adios}} \dot{M}(R) \frac{GM}{R} \quad (13.33)$$

which replaces equation (12.50). The right-hand side of this equation represents the energy transported from R *outward*, with $\epsilon_{\text{adios}} \geq 0$ being the fraction of energy lost in the ADIOS wind. Note that this equation also preserves the $V_\phi \propto R^{-1/2}$ scaling. With these equations the Bernoulli constant becomes

⁶ Note that this is similar to the Meyer *et al.* evaporation model in Section 12.2.2 that builds up the ADAF from the cool disk, but in the opposite direction ($n > 0$ here rather than $n < 0$ in equation (12.59)). Now the ADAF itself is evaporating to infinity via a strong wind.

⁷ We again have dropped the factor $f = [1 - (R/r_{\text{isco}})^{-1/2}]$.

$$Be = n \frac{p}{\rho} - \frac{1}{2} V_\phi^2 \quad (13.34)$$

which can be negative or positive, depending on the choice of the free parameters.

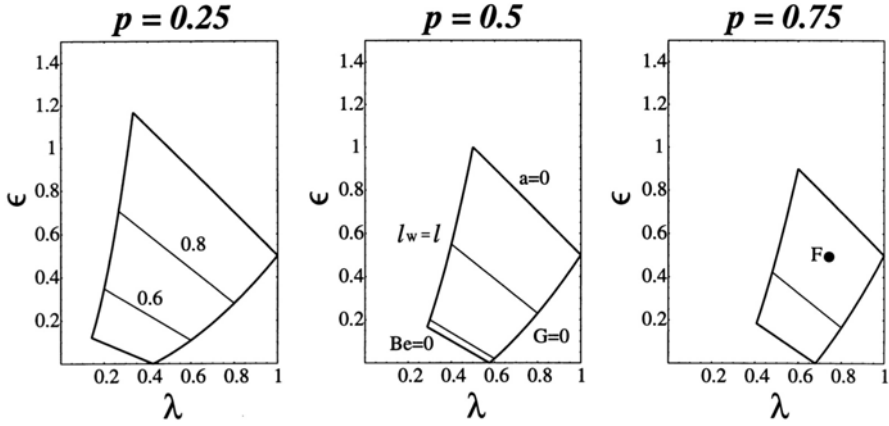


Fig. 13.4: The three-dimensional box-shaped region in n (called p in these panels), λ , and ϵ_{adlos} -space that contains allowable solutions to the ADIOS wind problem. Various types of solutions and their properties are discussed in the text. Figure reprinted from Fig. 1 of [464], by permission of J. Wiley and Sons, Ltd.

13.2.3 ADIOS Models

The ADIOS models have a total of three new free parameters (n , λ , and ϵ_{adlos}) in addition to m , $\dot{m}$, and α . Not all choices of the new parameters give physical solutions. Blandford and Begelman chose to limit solutions with the following conditions enforced at all radii R in the inflow:

1. $H > 0$ (finite disk thickness).
2. $W_{\phi R} > 0$ (outward transport of angular momentum).
3. $\ell_{\text{wind}} > \ell_{\text{inflow}}$ (angular momentum per unit mass carried by the wind reduces that in the disk; *i.e.*, some angular momentum is transported outward by the wind in addition to that transported outward by accretion torques).
4. $Be < 0$.

This creates a box-shaped region that is allowed in $(n, \lambda, \epsilon_{\text{adlos}})$ -space, which is illustrated in Fig. 13.4.

Like the MDAF models, the type of solutions available are many, and they depend on the rate of angular momentum inflow (as well as energy and matter flow). Some possibilities noted by Blandford and Begelman include:

- $n = \lambda = \epsilon_{\text{adios}} = 0$. Non-rotating, Bondi accretion with no wind.
- $n = \lambda = 0$, $\epsilon_{\text{adios}} \neq 0$. No wind, but with radiative losses; energy, but not angular momentum, is carried away.
- $n = 0$, $\lambda = 1$, $\epsilon_{\text{adios}} = 1/2$. Energy and angular momentum are carried away, but there is virtually no mass loss. This could occur in a magnetically-dominated wind. (We will discuss such winds in more detail in the next chapter, but with a full MHD, not ADIOS, formalism.)
- Some special cases on the boundaries of the allowed region in parameter space:
 - $\ell_{\text{wind}} = \ell_{\text{inflow}}$. Wind carries away only its own angular momentum.
 - $Be = 0$. Marginally bound inflow with or without a wind.

As a concrete example, Blandford and Begelman chose $n = \lambda = 3/4$ and $\epsilon_{\text{adios}} = 1/2$, an intermediate solution that carries off both angular momentum and energy. Some results of this choice are:

$$\begin{aligned}
 Be &= -0.35 \frac{GM}{R} \\
 V_{\text{wind } \infty} &= 0.41 \left(\frac{2GM}{R} \right)^{1/2} \\
 V_R &= 0.56 \alpha \left(\frac{GM}{R} \right)^{1/2} \\
 V_\phi &= 0.9 \left(\frac{GM}{R} \right)^{1/2} \\
 \frac{H}{R} &\approx 0.3
 \end{aligned}$$

Almost all the original $\dot{M}$ is expelled in this wind. Only a small fraction $\dot{M}(R_{\text{in}}) = (R_{\text{in}}/R_0)^{3/4} \dot{M}$ actually accretes into the black hole.

The best way to assess which ADIOS solutions are viable is to perform simulations of accretion inflows using full MHD simulations of the MRI. This will necessarily introduce a magnetic field into the wind (as well as into the accretion inflow; see Chapter 12). It is therefore likely that winds from real advection-dominated inflows have an hydromagnetic character to them. The next chapter, therefore, will discuss outflows in which a magnetic field plays an important role.

13.3 Thermally-Driven Jets

Up until now, this chapter has dealt exclusively with thermal winds – outflows that are either largely spherical or, at least, not axial in nature. We now turn to axial outflows (jets) that are driven by gas pressure alone. The primary feature that distinguishes thermally-driven jets from the magnetically-driven ones in Chapter 15 is that, in the former, the magnetic field is never dominant. That is, the plasma

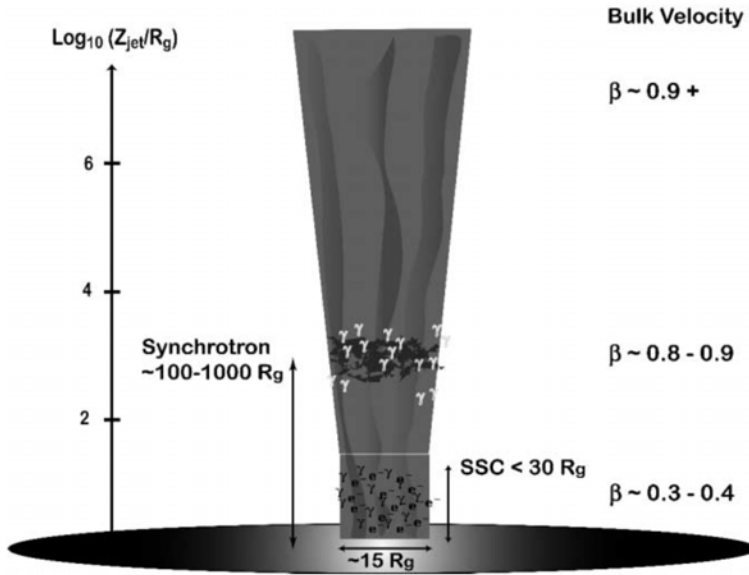


Fig. 13.5: Schematic diagram of the Markoff thermal jet model, along with typical results from models of many actual sources with jets. An accretion disk (bottom) produces a “nozzle”, about 15–20 r_g in diameter (possibly the inner disk corona) and twice that in height. The nozzle accelerates a jet upward initially at sonic speeds. This nozzle region is responsible for synchro-self-Compton radiation seen in the X-ray region of the spectrum. After reaching supersonic speeds, the jet flow continues upward in a cone shape (expanding laterally at the sound speed) with half-opening angle equal to the Mach angle. The resulting vertical pressure gradient further accelerates the flow. A few tens to hundreds of r_g above the disk a strong shock forms, accelerating the internal jet particles and creating a nonthermal radio-emitting plasma downstream from the shock. The emission from this jet plasma produces the broad-band synchrotron spectrum seen from infrared to X-ray energies. Reproduced from Fig. 1 of [468], by permission of the AAS.

$\beta_p = 8\pi p_g/B^2$ is of order unity (see Section 12.3.2). There are potentially several types of models of thermally-driven jets. Here we shall discuss only the model developed by Heino Falcke and Sera Markoff in the year 2000 [465], when both were at the Max Planck Institute for Radio Astronomy in Bonn, along with Markoff’s subsequent enhancements to it [466, 467, 468, 469]. This model has been used to analyze several black hole jet engines, from stellar-mass to supermassive ones.

13.3.1 Physical Structure

The general configuration of this thermal jet model is shown in Fig. 13.5. The jet is assumed to be accelerated in a cylindrical region of radius R_{base} , by some unspecified means (probably thermal pressure), up to some height Z_{base} , whereupon the

jet speed reaches the sound speed and then becomes a supersonic flow.⁸ Beyond the sonic point, the jet plasma (assumed to have an adiabatic index of $\Gamma = 4/3$) expands laterally at the sound speed, but continues to accelerate forward at greater than the sound speed, because of longitudinal pressure gradients created by the free expansion. The conservation of momentum equation in the Z direction for this acceleration is given by equation (9.141), with the conserved quantities given in Section 9.5.6. This can be re-written as

$$\gamma \rho c^2 \beta_Z \frac{\partial[\gamma \beta_Z (1 + \mathfrak{h})]}{\partial Z} = - \frac{\partial p}{\partial Z}$$

where ρ is the rest mass density, $\beta_Z = V_Z/c$ is the axial velocity, γ is the axial Lorentz factor $(1 - \beta_Z^2)^{-1/2}$, and $\mathfrak{h}$ is the specific enthalpy (see equation (9.150)). With the assumptions of adiabatic expansion (no cooling, so $p \propto (\gamma \beta_Z)^{-\Gamma} Z^{-2}$) and lateral confinement, this equation becomes

$$\frac{\partial(\gamma \beta_Z)}{\partial Z} \left\{ \frac{[(\Gamma + \xi)/(\Gamma - 1)] (\gamma \beta_Z)^2 - \Gamma}{\gamma \beta_Z} \right\} = \frac{2}{Z} \quad (13.35)$$

where the expression $\xi = \{\gamma \beta_Z / [\Gamma(\Gamma - 1)/(\Gamma + 1)]\}^{1-\Gamma}$ is a function of the solution $\gamma \beta_Z$. Equation (13.35) can be integrated numerically to obtain a solution, and then the latter can be inverted to determine $\beta_Z(Z)$ or $\gamma(Z)$ independently. This model works well for very low-luminosity sources like Sgr A* in our Galaxy and the central black hole source in M81. It is able to fit the radio and X-ray spectra of these sources well using several model parameters, most of which are tightly constrained by the data, with only a few being completely free[465, 470].

However, in other cases like black hole X-ray binaries and more luminous AGN with jets, the model had to be modified to also include a particle acceleration region far above the black hole in the jet – presumably a strong, stationary shock in the flow – and Compton reflection components to better fit the X-ray portion of the spectrum [466, 467, 468, 469]. The shock is needed to accelerate a nonthermal particle distribution at the shock distance (Z_{sh} or Z_{acc}), which in turn is needed to fit a nonthermal, jet synchrotron spectrum that stretches from the millimeter region all the way out to the X-rays in these latter sources.

⁸ In reference [470] this cylindrical structure was called the “nozzle”, and indeed, for purely hydrodynamic flows, this is an accurate description. We shall refrain from using that term, however, and simply call it the subsonic acceleration region, or subsonic region for short. We shall see in Chapter 15 that in MHD-driven jets, which we believe to be operating in virtually all black hole engines, the actual “nozzle” can extend to great distances (hundreds of r_g perhaps) because magnetic acceleration dominates considerably over gas pressure alone and continues to be effective over great distances from the black hole.

13.3.2 Plasma Conditions, Synchrotron Spectrum, and Comparison with Data

The second part of the thermal jet model is to compute the density and Lorentz factor distribution of relativistic particles, as well as the magnetic field strength, in the outflowing plasma jet so one can calculate the synchrotron spectrum of the jet itself. This then can be compared with observations so that many of the free parameters can be constrained by the data. The initial aspects of this plasma and emission model were outlined by Heino Falcke and Peter L. Biermann [471]. Sera Markoff and her colleagues recognized that the strong shock (Fig. 13.5) also was needed to fit the synchrotron portion of the spectrum and that, for X-ray binaries, a self-Compton reflection component was needed to fit the hard X-ray portion as well [466, 468]. Here we will describe the spectral models only qualitatively; the reader may consult the above reference for more details.

The main step in this process is to calculate the field strength and particle density at the sonic point. One does this by specifying the jet luminosity (as a fraction of the accretion disk luminosity), the mass-loss rate in the jet (as a fraction of the accretion rate), and the relativistic electron particle distribution at the sonic point. The assumed distribution in this case is not very elaborate; often one assumes only a mono-energetic particle distribution and then obtains a detailed distribution and spectrum from the variation in plasma conditions as the jet accelerates. With a knowledge of B_0 and ρ_0 at the sonic point, and a solution to equation (13.35), one is now in a position to calculate the spectrum of the entire jet, sub- and supersonic portions, using standard synchrotron emission formulae [470].

Once a spectrum is obtained, it is compared with broad-band flux data, from radio to X-ray. If the computed spectrum does not fit, the process is iterated, changing parameters like the height of the subsonic flow region, the radius of the jet at its base, the jet luminosity fraction, the jet outflow fraction, and the initial energy of the relativistic particles. The main conclusions of modeling many sources in this manner are discussed in Section 16.2.2.

Chapter 14

The Non-Thermal Exhaust System I. Rotating Magnetospheres that Drive the Turbo Exhaust

While accretion disks transport away *most* of the angular momentum in the accreting gas, they still leave material entering the black hole orbiting at nearly-Keplerian speeds. Black hole mergers also can spin *up* the final product (Section 8.2), leaving a more rapidly spinning black hole as well. Therefore, it is possible that, at some point, the central object will have accreted so much angular momentum that it begins to approach its maximum rotation rate. For black holes, this means that $j \rightarrow 1.0$. In order to continue to accrete steadily, it must shed angular momentum efficiently (from itself, the incoming matter, or both) and in a manner not possible using accretion disk viscous transport alone. In this case magnetic fields are important in carrying away excess angular momentum, and this leads to magnetized winds and jets. While a small amount of matter may be expelled in this process, the main excess quantity being carried away by these winds is angular momentum.

The next two chapters add the important effects of a magnetic field to the wind exhaust discussion of the previous chapter. These are not simple concepts or processes, but they are the essence of jet production in the universe, not only in black hole systems but also in stellar systems of many types (including the early sun). The present chapter discusses the rotating magnetosphere turbine, which drives the initial outflow from the vicinity of the black hole. This is analogous to the injection region for the thermal/radiative winds of the previous chapter.

A rotating stellar magnetosphere is remarkably like a turbine exhaust fan, with a few crucial differences:

- The fan “blades” are composed of stiff magnetic field lines, rather than steel or some other solid substance.
- the exhaust material is highly ionized, and magnetized, plasma, rather than a neutral gas.
- From the point of view of the engine, the most important component of the exhaust is not the matter in the exhaust nor its thermal energy, but rather the rotational kinetic energy that it carries away.

Earlier in this book we treated neutron stars as near-black holes – relativistic, gravitating objects whose properties are similar to those of true black holes in many

respects. We will do the same here also, using the rotating magnetosphere of a neutron star pulsar to introduce many of the concepts of black hole magnetospheres. The main difference between neutron stars and black holes in this regard is that, like the earth, sun, and other stars, a pulsar's magnetic field is anchored in the *stellar* material, and rotates with the star. A black hole's magnetic field, however, is anchored in the *accreting* material itself and rotates with the orbiting plasma, whose own rotation can be affected by the rotating black hole itself.

14.1 Isolated Pulsars: Rotating Neutron Star Magnetospheres with No Accretion

The behavior of a magnetosphere depends greatly on whether the object is a neutron star or black hole and whether that star is accreting matter or not. We shall discuss these four cases in turn, beginning with the simplest first.

14.1.1 Pulsar Basics: Dipole Radiation, Spindown Power, and Light Cylinder Radius

Immediately after the discovery of pulsars in 1968 there began a flurry of theoretical activity to understand these new objects. The basic rotating magnetic dipole model ([472, 473]) quickly became adopted. It predicted pulsar spindown, which was routinely observed, and led to the famous $P \dot{P}$ formula for the magnetic field (equation (3.3)). The derivation of this formula is very straightforward. If the pulsar is a rotating spherical star, then its rotational energy is given by

$$\mathcal{E}_{\text{rot}} = \frac{1}{2} I \Omega^2$$

where Ω is the star's rotation rate in radians per second, and I is the moment of inertia. The energy loss, therefore, is $\dot{\mathcal{E}}_{\text{rot}} = I \Omega \dot{\Omega}$ or, in terms of the period

$$\dot{\mathcal{E}}_{\text{rot}} = -\frac{4\pi^2}{P^3} I \dot{P} \quad (14.1)$$

Now, in Section 6.6.1 we derived the radiation energy loss rate for an *electric* dipole (equation (6.107)). A similar formula also can be derived for the total power emitted by a *magnetic* dipole

$$\mathcal{P}^{\text{EM}} = \frac{2\mu^2 \Omega_f^4}{3c^3} \sin^2 \Theta \quad (14.2)$$

where μ is the magnetic dipole moment, Ω_f is the angular velocity of the magnetic field, and Θ is the angle between the dipole vector μ and rotation axis Ω_f . Assuming that $\Omega_f = \Omega$, $\Theta = 90^\circ$, and taking the relation between magnetic moment and magnetic field B_{NS} at the neutron star surface ($R = R_{\text{NS}}$) as $\mu = B_{\text{NS}} R_{\text{NS}}^3$, the magnetic dipole radiation power of a neutron star pulsar becomes

$$\mathcal{P}^{\text{EM}} = \frac{2 B_{\text{NS}}^2 R_{\text{NS}}^6 \Omega_f^4}{3 c^3} \quad (14.3)$$

Equating $\dot{\mathcal{E}}_{\text{rot}} = -\mathcal{P}^{\text{EM}}$, we can solve for the neutron star magnetic field in terms of two observables

$$B_{\text{NS}} = \left[\frac{3 c^3}{8 \pi^2} \frac{I}{R_{\text{NS}}^6} \right]^{1/2} (P \dot{P})^{1/2}$$

or

$$B_{\text{NS}} = 3.2 \times 10^{19} \text{ G} (P \dot{P})^{1/2} \quad (14.4)$$

(which is equation (3.3)) where P is measured in seconds and $\dot{P}$ is unitless (seconds per second). Equation (14.4) gives the result found using detailed neutron star models, but a uniform density sphere of radius 10^6 cm and $1.4 M_\odot$ gives a close approximation to the coefficient ($3.4 \times 10^{19} \text{ G}$).

The luminosity in equation (14.3) can be written in two other ways that will be very useful in the sections below:

$$\mathcal{P}^{\text{EM}} \approx \frac{B_L^2 R_L^4 \Omega_f^2}{c} \quad (14.5)$$

or, alternatively

$$\mathcal{P}^{\text{EM}} \approx B_L^2 R_L^3 \Omega_f \quad (14.6)$$

Here R_L is the radius of the light cylinder (where the tangential velocity $V_f = R_L \Omega_f = c$)

$$R_L \equiv \frac{c}{\Omega_f} \quad (14.7)$$

and B_L is the average magnetic field strength at the light cylinder, which we find from the dipole formula for the variation of magnetic field strength with radius ($B_{\text{NS}} R_{\text{NS}}^3 = B_L R_L^3$, at least out to the light cylinder). We also have dropped the factor of $2/3$, which is of order unity.

The light cylinder radius is important in the study of rotating electrodynamic and MHD winds. Note that, even though $R \Omega_f > c$ for $R > R_L$, there is no violation

of causality. It is the magnetic field that rotates at this superluminal rate and not the plasma. In fact, the field structure adjusts itself so that any plasma trapped in the field lines still rotates at *less* than the speed of light (*i.e.*, $V_\phi < c < R\Omega_f$), with the superluminal magnetic field slipping azimuthally through the subluminal plasma.

Both formulae in equations (14.5) and (14.6) occur repeatedly in MHD wind theory in various forms, with R_L replaced by the characteristic size of the rotator. Therefore, *the light cylinder radius should be considered to be the typical size of a relativistic rotating magnetized system.*

14.1.2 The Basic Force-Free Pulsar Magnetosphere

While representing a pulsar as a rotating magnetic dipole helps explain its basic dynamical properties, the model does not explain why pulsars radiate nor why they emit that radiation in one or more narrow rotating beams. This requires understanding what happens *inside* the neutron star's rotating magnetic field, how particles produce the radiation, and where those particles come from in the first place. Most radio pulsars generally show no evidence of a companion nor of any material accreting onto the neutron star. Yet there *must* be plasma intermixed with the magnetic field, as we soon shall see.

14.1.2.1 The Goldreich–Julian Charge Density

At the same time people were developing a theory of pulsar magnetic dipole radiation, Peter Goldreich and William H. Julian, of Caltech, were establishing the basic principles of pulsar magnetospheres [474]. The first principle they showed was that a *stationary, rotating* magnetosphere *must* be filled with charge. This follows simply from the ideal Ohm's law (equation (9.105) with α and β zero for a neutron star)

$$\mathbf{E} = -\frac{\Omega_f R}{c} \mathbf{e}_\phi \times \mathbf{B} \quad (14.8)$$

(which gives us the electric field $\mathbf{E}$) and from Gauss's law (equation (9.117), again with α and β zero). Together these tell us that the charge particle density must be

$$\rho_q = \frac{\nabla \cdot \mathbf{E}}{4\pi} = -\frac{\Omega_f \cdot \mathbf{B}}{2\pi c} \equiv \rho_{\text{GJ}} \quad (14.9)$$

which is a charge *number* density of

$$n_q = 0.07 \text{ cm}^{-3} \frac{B}{P}$$

For the Crab Pulsar, for example, this means that the density of charges is nearly 10^{13} cm^{-3} . This is far from a vacuum.

14.1.2.2 The Goldreich–Julian Magnetosphere Cartoon

While Goldreich and Julian were not able to solve the structure of the magnetosphere quantitatively, they were able to *qualitatively* deduce its properties and sketch a cartoon of how the solution should appear. Fig. 14.1 shows this cartoon from their 1969 paper. Several important features are notable:

- Closed magnetic field lines lie *only inside* the light cylinder. Outside that radius, field lines must remain open, with an outgoing wind of plasma streaming outward along those open lines.
- At the equator these open field lines reverse polarity as one crosses from the northern hemisphere to the south. Because $\mathbf{J} = c\nabla \times \mathbf{B}/4\pi$, there must be a sheet of current flowing along the equator.
- The field line that closes just inside the light cylinder (but also continues outward along the equator) is called the “separatrix”, because it separates open from closed field lines. The place on the equator where the separatrices from the north and south meet and then form the current sheet is called the Y-point.
- Finally, field lines that are tangent to a $Z = \text{constant}$ (horizontal) surface have $\boldsymbol{\Omega} \cdot \mathbf{B} = 0$ and, therefore by equation (14.9), must have zero charge density there. Above the null surface the charge density is all of one sign and below the null surface it is of the opposite sign. Because charges cannot cross-field lines, this creates a “null” (zero charge) surface that extends all the way to the light cylinder, flanked by positive and negative charge outflows. Beyond the light cylinder, charges of one sign flow outward along the equator while those of the opposite sign flow outward at higher latitudes. This creates a *closed current system* in each hemisphere, with outward and return current.

Only a few percent of the energy output of a pulsar is in the emitted radiation (radio, optical, hard γ -ray). The great majority is in the MHD wind/jet outflow, which consists of both outflowing charged particles and electromagnetic Poynting flux. Eventually that energy is radiated also – by the surrounding pulsar wind nebula or “plerion”, rather than by the pulsar itself.

14.1.2.3 The Outgoing Wind

To fully understand the outgoing wind we must use relativistic MHD techniques, not just FFDE. We will present the details on MHD winds later in Section 15.1. However, it is instructive to discuss at this point a few of the aspects of the outgoing wind that pertain to force-free, or nearly force-free, outflow.

As the wind particles travel outward along a magnetic field line, there are really two main surfaces of interest: the light cylinder and the causality limit. Beyond the light cylinder the tangential speed of a rotating magnetic field line formally exceeds the speed of light. At the causality limit, which lies well outside the light cylinder, the outflow becomes causally disconnected from the magnetosphere. Beyond that point the structure of the outflow no longer can back-react on the engine that created

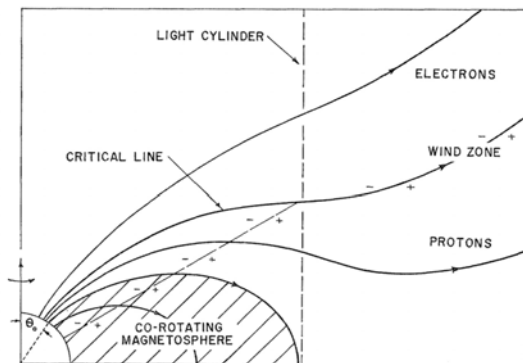


Fig. 14.1: Concept cartoon drawing of a pulsar magnetosphere. Pulsar is at lower left and magnetic field lines are given by the curved arrows. Rotation inflates the magnetosphere in the equatorial direction. Beyond the light cylinder $R_L = c/\Omega_f$ field lines no longer close across the equator, instead stretching out to infinity. When the pulsar rotates, charged particles must flow outward along the field lines, requiring that they be replenished continuously near the magnetic poles by a particle creation process described in the text. However, at points where the magnetic field is normal to the rotation vector Ω , the charge must be zero along this surface, generating regions of positive and negative charge. Reproduced from Fig. 1 of [474], by permission of the AAS.

it. We will learn in Section 15.1 that, for a relativistic MHD wind, the light cylinder occurs very near where the outflowing wind reaches the Alfvén speed. And the causality limit occurs, not at the point where the outflow reaches the magnetosound speed, but rather where the collimation speed of the jet *toward the rotating axis* equals c_{ms} . This “modified fast point” occurs rather far from the central engine.

These relativistic wind concepts raise an interesting question. How can a magnetic field line rotate faster than the speed of light, in fact orders of magnitude faster near the causality limit? The answer is that a magnetic field line is not a true physical entity; it is only a contour of magnetic flux – a mathematical entity. What *is* subject to the cosmic speed limit is the plasma itself. Even though the plasma is frozen to the magnetic field, its speed is kept below c because the field *bends backward substantially near and beyond the light cylinder*. As the magnetic field rotates, it slips through the plasma, slicing through it at an angle small enough to keep the plasma’s azimuthal and poloidal velocities under c . This coiling backward of the field produces a hoop stress (B_ϕ^2/R) that pinches and collimates some of the outflow into a jet along the rotation axis.

14.1.2.4 Numerical Models of Pulsar Magnetospheres: Thirty Years of Hard Work and, Finally, Success!

The story of developing accurate models of pulsar magnetospheres is similar to the one for numerical simulations of black hole mergers that we related in Section 8.2.

Although quite a lot of work was done between 1969 and 1999, as with numerical relativity, we will discuss it only briefly and then jump right to the success stories.

The first three decades of effort in pulsar magnetosphere research concentrated on the easier, two-dimensional “aligned rotator” problem – the one depicted in Fig. 14.1 – where Ω and μ are aligned ($\Theta = 0$). The goal was to confirm, or refute, Goldreich and Julian’s picture and, eventually, explain pulsar radiation itself. The situation lends itself to axisymmetry, as well as stationarity, and can be described well by a strong field with essentially no material forces. The pulsar equation (9.137), therefore, can be used to obtain the structure of this “force-free” magnetosphere.

As we pointed out in Chapter 9, however, the assumptions (especially stationarity) come at a price: there is no obvious way to specify the current (I) or its derivative (I') with respect to the potential Ψ . One must choose reasonable functions for the product $I I'$ and see how well they do.

Fig. 14.2(left) shows the resulting pulsar magnetosphere when the current distribution is a monopole one

$$I_m = -\frac{1}{R_L} \Psi \left(2 - \frac{\Psi}{\Psi_0} \right)$$

Here $\Psi_0 = 1.742 \Psi_{\text{sep}}$, where Ψ_{sep} is the separatrix field line. Note that, while Ψ is continuous at the light cylinder, its gradient is not. Such strange solutions near the light cylinder caused some investigators to suggest that perhaps shock waves or other discontinuities formed in the wind outflow – something not expected in a force-free electromagnetic field.

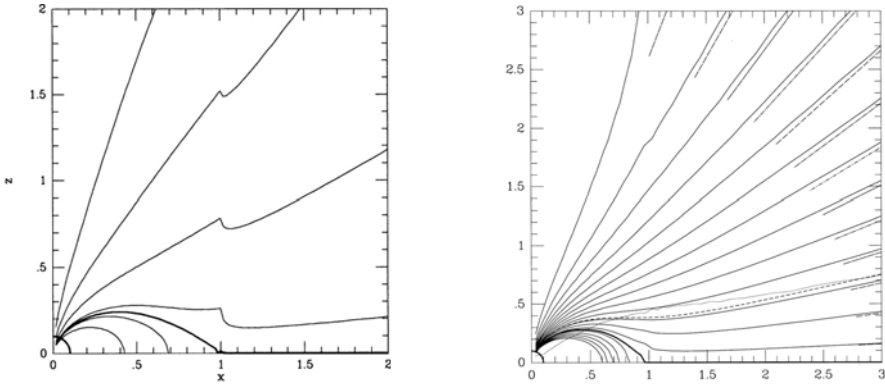


Fig. 14.2: Comparison between specifying (left) and solving for (right) the current distribution in a pulsar magnetosphere. While a reasonable guess produces a good approximation to the pulsar magnetosphere, the field lines are only C^0 continuous at the light cylinder. Requiring C^1 continuity and solving for the current distribution produces a physically acceptable solution. Reproduced from Figs. 1c and 3 of [475], by permission of the AAS.

The first solution of an aligned pulsar rotator *with* smooth flux lines was obtained by Ioannis (a.k.a. John) Contopoulos (of the University of Crete), Demosthenes Kazanas (of NASA’s Goddard Space Flight Center), and Christian Fendt (of Potsdam Astrophysical Institute in Germany) [475]. In retrospect, the solution method is quite straightforward; but breakthroughs often take years to find the right technique. Contopoulos and his colleagues realized that the current distribution $I(\Psi)$ *must* be solved for, but what were the constraints on this function? They pointed out that both Ψ and I vary *vertically* along the light cylinder surface. So, if one *constrained* the product $I I'$ to be continuous across that surface, then the field Ψ and its derivative across the light cylinder also will be continuous. Fig. 14.2(right) shows these authors’ final result. The aligned pulsar rotator *indeed does* look like Goldreich and Julian’s original cartoon!

The Contopoulos *et al.* solution showed that the luminosity of an aligned rotator is *not* zero, unlike the prediction of the vacuum formula (14.2). Even when $\Theta = 0$, the power is

$$\mathcal{P}^{\text{EM}} = \frac{\mu^2 \Omega_f^4}{c^3}$$

(see also [457].) The reason for this discrepancy is that the pulsar numerical model is not a vacuum solution. It includes the Goldreich–Julian charge and currents necessary to create the magnetosphere. These rotating charges radiate even when the magnetosphere is aligned.

A few questions arose about this new result. First, the Y-point does not seem to meet precisely at the intersection of the light cylinder and the equator; it occurs a little bit inside $x = 1$. And the slope of the separatrix there is finite; *i.e.*, it forms a cusp at $Z = 0$. This differed from Goldreich and Julian’s predictions. Which was correct; the numerical results or Goldreich and Julian’s cartoon?

These questions were investigated by Dmitri Uzdensky (of the Kavli Institute for Theoretical Physics [KITP] at UCSB) in 2003 using perturbation methods (a Taylor expansion of the equations) near the Y-point [476]. He concluded that the Y-point does indeed occur slightly inside the light cylinder, contrary to the cartoon, but the separatrix does indeed make a 90° angle with the equator, contrary to the numerical model. The latter discrepancy was due to using a finite-sized grid in the early computations. So, the Y-point is really a “T-point”.

14.1.2.5 Full FFDE Simulations of Aligned and Oblique Pulsar Magnetospheres

Additional questions that arise from Fig. 14.2(right) are, why do we still see kinks in some field lines at the light cylinder? And, what does the situation look like when the magnetic axis and rotation axis are *not* aligned? Furthermore, now that numerical models include charge and current embedded in the magnetosphere, how does that

change the spindown luminosity from the vacuum formula in equation (14.2), and how does this relate to the Contopoulos *et al.* result?

One very useful way of answering these questions is to drop the stationary assumption ($\partial/\partial t = 0$) in order to see how the magnetosphere *evolves* toward a steady state. Ultimately, we will want to drop the axisymmetric assumption ($\partial/\partial\phi = 0$) as well, in order to study full three-dimensional simulations of misaligned pulsar magnetospheres. These enhancements cannot be handled by the pulsar equation. We *must* use the full time-dependent, three-dimensional force-free electrodynamics equations (9.122) to (9.124) in Section 9.5.2, with $\alpha = 1$ and $\beta = 0$ (flat space). These equations are then used to evolve $\mathbf{B}$ and $\mathbf{E}$ to a steady state, subject to the force-free condition on the current density $\mathbf{J}$.

Techniques for handling the time-dependent FFDE equations were developed by Anatoly Spitkovsky (then at Stanford University and now at Princeton) [477]. As we mentioned earlier in Section 9.5.2, the FFDE equations do not *explicitly* enforce the ideal Ohm's law, allowing $|\mathbf{E}| > |\mathbf{B}|$ in some regions of the computational grid during the simulation. When this occurred, Spitkovsky simply reset $|\mathbf{E}|$ to $|\mathbf{B}|$ and allowed the simulation to continue. It is important to realize that in structures like these, where one runs the simulation for long periods of time, the transient effects are less important than the final steady state. While setting $|\mathbf{E}| = |\mathbf{B}|$ might introduce strange transient effects, *not* doing so means that the simulation will crash, reaching no steady solution at all! Spitkovsky's fix is, by far, the lesser of the two evils.

Fig. 14.3 shows a time sequence of Spitkovsky's two-dimensional simulation that leads to a steady-state magnetosphere structure that looks very similar to Goldreich and Julian's original cartoon. The field lines are all quite smooth at the light cylinder, as expected. And the behavior of the structure near the Y-point is very interesting. In these simulations it has no steady-state solution. Sometimes lines beyond the light cylinder close in a cusp-like shape, but then are thrown open by the ejection of a "plasmoid" (a cloud of plasma). This leaves a closed separatrix behind the light cylinder that *does* make a 90° angle with the equator. Spitkovsky's spindown luminosity agrees with that of Contopoulos *et al.* ($\approx \mu^2 \Omega_f^4 / c^3$).

Fig. 14.4a shows Spitkovsky's result for a pulsar with an *oblique* angle of 60° between the rotation and magnetic axes. As expected, the results are rather complex, but nevertheless similar to what was expected. Near the pulsar the field is a rotating dipole, but far from it waves propagate out in a spiral pattern. Fig. 14.4b shows that the total spindown luminosity of an oblique rotator is very similar to the sum of that for the aligned rotator plus a component that varies with the obliqueness angle just like the vacuum solution does. That is,

$$\mathcal{P}^{\text{EM}} \simeq \frac{\mu^2 \Omega_f^4}{c^3} (1 + \sin^2 \Theta) \quad (14.10)$$

This gives a slightly modified $P \dot{P}$ formula of

$$B = 2.6 \times 10^{19} \text{ G} \left[\frac{P \dot{P}}{1 + \sin^2 \Theta} \right]^{1/2} \quad (14.11)$$

This means that pulsars can produce the power they do with magnetic fields 20–40% weaker than we thought – not a large amount, but significant nonetheless.

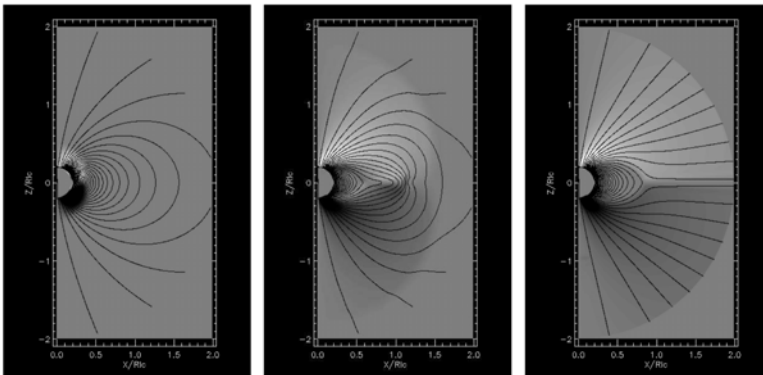


Fig. 14.3: Development of a steady state, aligned pulsar magnetosphere during a time-dependent FFDE simulation. *Left:* Initial dipole configuration with no rotation. *Middle:* Expansion of equatorial structure after rotation is turned on. *Right:* Establishment of open field lines, current sheet along the equator, and Y-point at the light cylinder/equator intersection. Reprinted with permission from Fig. 1 of reference [478]. Copyright (2005), American Institute of Physics.

14.1.2.6 Particle Creation, Acceleration in “Vacuum Gaps”, and the Origin of the Pulsating Radiation

In the above discussion we have assumed that there is sufficient charge embedded in the magnetosphere to create the Goldreich–Julian charge density (equation (14.9)). It is a valid question to ask, where does this charge come from? Of course, one answer might be that the charged particles come simply from plasma that existed when the pulsar formed, or when the pulsar passes through some interstellar cloud. There is a problem with these answers, however. The rotating pulsar drives a strong wind, which expels the magnetospheric plasma. In fact, as we shall learn in the next chapter, the velocity of the wind at the light cylinder (V_L) is of order the Alfvén speed, which is several tenths the speed of light. So, the time scale to eject most of the plasma inside the light cylinder is

$$\tau_L \equiv \frac{R_L}{V_L} = \frac{c}{V_L \Omega_f} < \frac{2\pi}{\Omega_f} = P$$

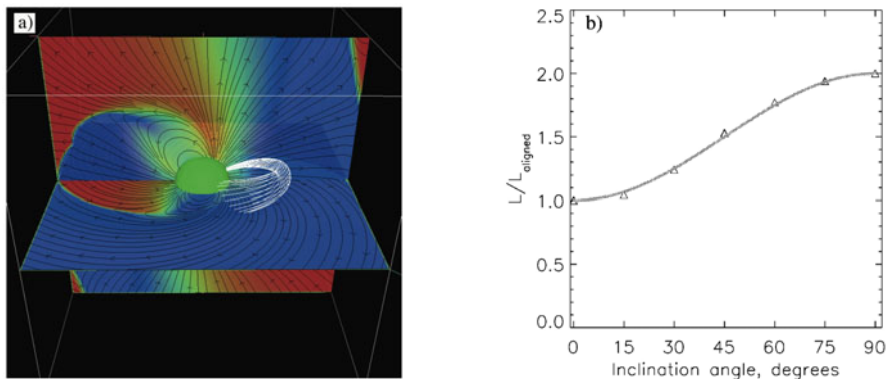


Fig. 14.4: Structure and angular dependence of oblique rotating pulsars. (a) Cuts through the rotational axis (vertical, meridional plane) and equator (horizontal plane) of a 60° -inclined rotator, showing the magnetic field lines and their direction. The white lines show a typical flux tube. (b) Plot of pulsar output electromagnetic power vs. inclination angle; even the aligned rotator has significant output power. Reproduced from Fig. 3 of [477], by permission of the AAS.

– less than one rotation period! *Plasma must be coming from somewhere at a prodigious rate just to maintain the Goldreich–Julian charge density.* How is that accomplished?

Particle Creation

One source, of course, might be the very heavy material on the pulsar surface, but in most cases the pulsar is not hot enough to boil those ions off high into the magnetosphere. A different source of plasma was identified by Marvin A. Ruderman and Peter G. Sutherland of Columbia University in 1975, an exotic process called “vacuum breakdown” [479]. It is actually the loss of charge by the wind itself that *begins* this process. As the charge density drops, $\nabla \cdot \mathbf{E} = 4\pi\rho_q$ still is maintained, but $\mathbf{E}$ no longer can remain at the Ohm’s law value (equation (14.8)). Put another way, there is no longer sufficient charge and current to short out a general electric field (one that is independent of the rotating magnetic field’s emf). $|\mathbf{E}| > |\mathbf{B}|$ is then physically possible, and some of that new electric field is parallel to $\mathbf{B}$; that is, $\mathbf{E} \cdot \mathbf{B} \neq 0$. This is especially important at the magnetic poles, where an electric potential “gap” develops, with a height of a few percent of the neutron star radius

$$h \approx R \left(\frac{R}{R_L} \right)^{1/2} \approx 1.4 \times 10^4 \text{ cm } P^{-1/2}$$

and an electric potential difference equal to the emf ($\nabla^2\Phi = 4\pi\rho_{GJ}$) or

$$\begin{aligned}
\Phi &\approx \frac{\Omega_f B h^2}{2\pi c} \\
&\approx 3.3 \times 10^9 \text{ statvolts} \frac{B_{12}}{P^2} \\
&\approx 2 \times 10^{12} \text{ volts} \frac{B_{12}}{P^2}
\end{aligned}$$

This large potential gap breaks down in a manner similar to sparking in earth's atmosphere (lightning), but in a more exotic manner: it *creates* charge from the vacuum electric field.

The development of this strong electric field occurs before all charges are gone, and the field then accelerates the remaining ones to highly relativistic energies (10^{12} eV) across the gap, giving electrons Lorentz factors of $\gamma_e = 10^{12} \text{ eV}/m_e c^2 = 2 \times 10^6$ or greater. Because these fast electrons still must stream out along the magnetic field, they are accelerated normal to their paths by the curved field lines. This leads to curvature radiation of photons with γ -ray energies of

$$E_\gamma = \frac{3}{2} \frac{\gamma_e^3 \hbar c}{R_{\text{curv}}} \approx 240 \text{ MeV}$$

where $\hbar = 1.05457 \times 10^{-27} \text{ erg s}$ is Planck's constant divided by 2π and $R_{\text{curv}} \approx 10^6 \text{ cm}$ is the radius of curvature of the magnetic field lines. The energy of these γ -ray photons is much larger than the $2 \times 0.511 = 1.022 \text{ MeV}$ threshold needed to create $e^\pm$ pairs as they scatter off the 10^{12} G pulsar magnetic field. So, the acceleration of charged particles in the magnetosphere leads to many more electrons and positrons.

Vacuum breakdown is a self-limiting process, however. As the $e^\pm$ charges are created, the magnetosphere near the poles once again begins to fill with charge. The strong electric field strength shorts out, and the ideal Ohm's law situation is restored when ρ_q achieves the Goldreich–Julian charge density. In practice, of course, this will never happen, because the wind is continually ejecting charge. So an approximate steady state may be reached, in which the gap is just large and strong enough to replace the charge lost in the wind.

Pulsar Emission

In addition to the curvature radiation at the poles, there also will be synchrotron radiation as well as inverse Compton radiation of those photons on the 10^{12} eV electrons and positrons, spanning wavelengths from the optical to the hard γ -ray. Furthermore, there will be low-frequency (radio) waves from disturbances generated in the magnetic field lines at the poles. All of this radiation will stream out primarily along the magnetic poles, creating two oppositely-directed rotating beams of radiation that sweeps around the sky. The essence of a pulsar, therefore, is its rapidly rotating magnetic field. This generates an emf that drives a wind of charged particles. And the efforts of the pulsar to regenerate these charges, by the vacuum

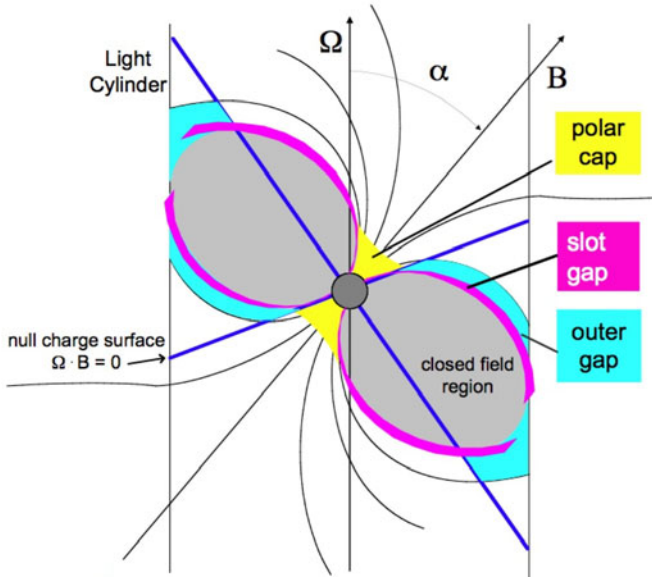


Fig. 14.5: Primary (yellow) and secondary (magenta and cyan) gaps in the pulsar magnetosphere. Depending on the angle Θ (here α) between the pulsar rotation axis and the dipole magnetic field, particles created in the polar cap regions may not replenish all charges ejected by the wind. Charges are not likely to cross the null (zero) charge surfaces (*i.e.*, where the local magnetic field is normal to the rotation vector Ω), shown by thick blue lines. (See also Fig. 14.1.) Therefore, secondary spark gaps, particularly the outer gaps, may develop between these null charge surfaces, producing additional regions where charge can be created. Figure kindly supplied by A. Harding; reproduced with permission.

breakdown process, generates the radiation that we see in the pulses. When the pulsar eventually spins down, the wind, the particle production, and the radiation all will turn off, leaving a dead, but still magnetized, neutron star.

Secondary Gaps and Pulse Structure

Particle production at the pulsar magnetic poles may replenish some, but not all, of the charge lost in the wind. This is because charge flow may be blocked across the zero-charge “null surfaces”. As a result, there may be additional potential gaps in key regions of the magnetosphere where more particle creation and radiation could occur (see Fig. 14.5). The position, shape, and strength of these gaps will depend on the detailed magnetosphere structure, which is only now beginning to be understood. Nevertheless, observations of pulse properties already are being used to construct the properties of these gaps by groups such as Alice Harding’s at NASA’s Goddard Space Flight Center.

14.2 Pulsars in Close Binary Systems: Rotating Neutron Star Magnetospheres with Accretion

When the neutron star pulsar is in a close binary system with a normal star, it can accrete material from its companion in the manner detailed in Chapter 11. This flood of plasma into the magnetosphere makes some of the analysis more complex and some of it simpler. For example, the force-free assumption no longer is valid in this case: we *must* use the full equations of MHD, although we probably can get away with using the non-relativistic versions (equations (9.152) to (9.155)), because the flow velocities and gravitational potential are nearly Newtonian.

Fig. 14.6 shows the general picture for an aligned rotator. The ram pressure of the accretion flow (ρV^2) compresses the magnetic field along the equator until it balances the magnetic pressure at a cylindrical radius R_m . For a dipole magnetic field and simple free-fall accretion, this occurs at the magnetic radius

$$R_m = 1.07 \times 10^8 \text{ cm } \dot{m}^{-2/7} B_{12}^{4/7} \quad (14.12)$$

where $\dot{m} \equiv \dot{M}/\dot{M}_{\text{Edd}}$ as before, and B_{12} is the pulsar *surface* magnetic field in units of 10^{12} G. Matter collects at R_m until its vertical scale height becomes comparable to that of the field. Then material flows along the closed field lines to the central star. In this state the pulsar is said to be in the “accretion” regime.

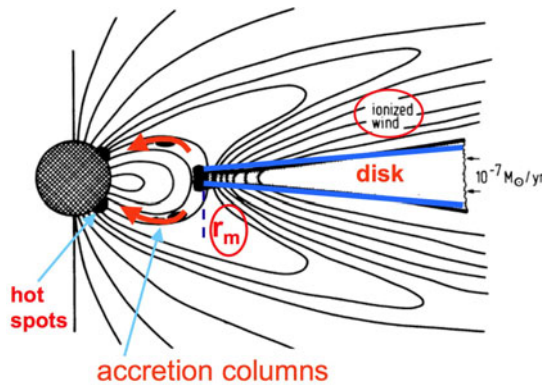


Fig. 14.6: Effect of an intruding accretion disk on an aligned rotating pulsar and the subsequent effects of the strong field on the accretion flow. The accretion “ram pressure” is balanced by the magnetic pressure at R_m , inside which the inflow occurs in accretion columns along field lines, impacting the pulsar surface in hot spots that lie near the magnetic poles. Note that an outgoing ionized wind also is driven by the pulsar rotation. Original figure by M. Camenzind (Fig. 12 of [480]) and adapted by M. Romanova. Reproduced with permission by Springer Publishing Co.

Above the disk at R_m , the rotating field lines inflate outward. If the rotation is fast enough, these field lines open up, creating an outflowing plasma wind that is much denser than that of an isolated pulsar. This latter type of system is said to be

in the “propeller” or “wind turbine” regime. The dividing line between accretion inflow and turbine-driven outflow depends on the ratio of the magnetic radius to the “co-rotation radius”, *i.e.*, where $\Omega_f = \Omega_{\text{disk}} = \Omega_K$ or

$$R_{\text{CR}} = 1.50 \times 10^8 \text{ cm } P^{2/3}$$

If $R_m < R_{\text{CR}}$, we have slow rotation and accretion flow; if $R_m > R_{\text{CR}}$, we have rapid rotation and a turbine-driven exhaust. The critical period for this change in behavior is found by equating R_m and R_{CR} , giving

$$P_{\text{crit}} = 0.60 \text{ s } \dot{m}^{-3/7} B_{12}^{6/7}$$

For X-ray pulsars with longer periods than this, accretion dominates; when their period is shorter, the propeller wind dominates.

These issues have been studied in detail by the Cornell–Moscow group (Marina Romanova and Richard Lovelace at Cornell University, Galina Ustyugova and Aleksander Koldoba at the Russian Academy of Sciences in Moscow, and their colleagues [481, 482, 483].)

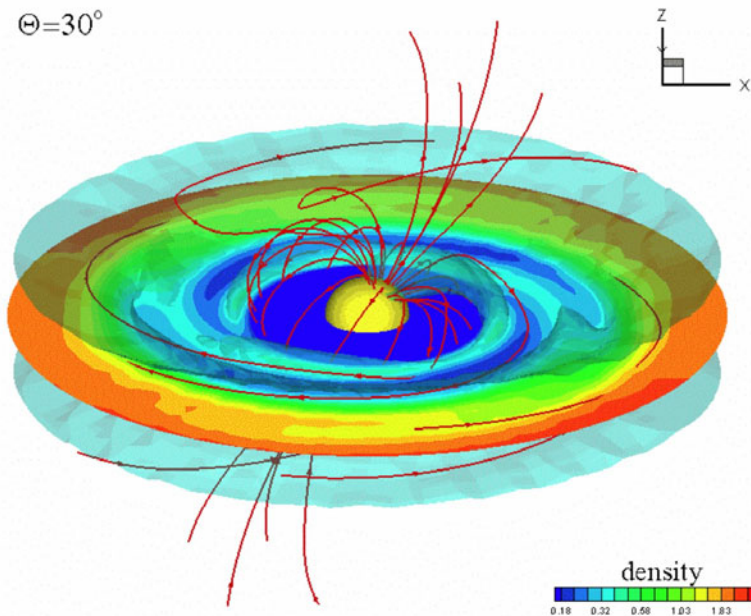


Fig. 14.7: An accreting pulsar in the accretion regime. The pulsar and disk rotation axes are aligned here, with the star’s *magnetic* axis oblique to the rotation by 30° . The disk is shown with translucent isodensity surfaces for better visibility into its interior. Similar to Fig. 5 of reference [481]. Color version kindly provided by M. Romanova; reproduced with permission.

14.2.1 Plasma Flow onto Pulsars in the Accretion Regime

We will look at the accretion regime first ($R_m/R_{CR} < 1$) by Romanova *et al.* [481, 482]. Their simulations are fully three-dimensional, so they can handle oblique pulsar rotators *and* non-axisymmetric structure in the accretion flow.

14.2.1.1 Moderate Accretion Flow

Fig. 14.7 shows a typical system with a $\Theta = 30^\circ$ misalignment between the disk and the pulsar magnetic field. (The disk and pulsar rotation axes are assumed to coincide.) The density surfaces are translucent in the disk midplane and in the disk atmosphere so we can see the layers that lie below. Fig. 14.8 displays two density surfaces of this model with opaque color, which hide the magnetic field lines but clearly show how the accretion flow rises up onto the closed lines in the northern and southern hemispheres and then falls onto the magnetic polar regions.

The two spots where the accretion impacts the pulsar surface become very bright in X-rays and then rotate with the pulsar period. At least one of these spots is visible to an observer, creating a binary X-ray pulsar. (See the light curve in Fig. 14.9A.)

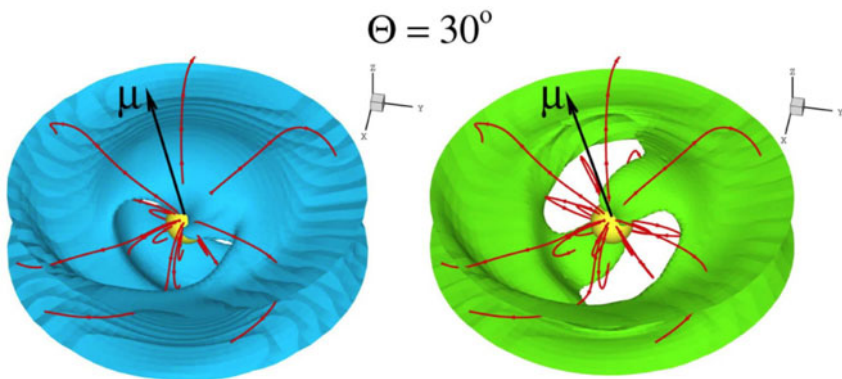


Fig. 14.8: Two different (now opaque) density surfaces for the system in Fig. 14.7, each showing the accretion flow along the field lines and the impacts near the magnetic poles. Similar to Fig. 4 of reference [481]. Color version kindly provided by M. Romanova; reproduced with permission.

14.2.1.2 High-Rate Accretion Flow

At some point the rate of accretion flow will begin to overwhelm the magnetic field. Of course, too much matter could push R_m all the way down to the pulsar surface. This critical accretion rate occurs when the ram pressure of accretion equals the

magnetic field pressure at the surface ($R_m = R_*$), or

$$\dot{m} \geq 1.27 \times 10^7 B_{12}^2 \equiv \dot{m}_{\text{cr}}$$

Except during star mergers it is unlikely that this enormous accretion rate ($>10^7 \dot{M}_{\text{Edd}}$) will be approached, let alone exceeded.

Romanova *et al.*, however, found an intermediate regime where the accretion “sneaks” in between the magnetic field lines at many more places than just at the poles [482]. Fig. 14.9 compares the high accretion rate regime with the moderate one discussed above. Note that the flow is unstable to forming many tongues of accretion, although the magnetic field still is partially successful in keeping the main accretion flow at bay. Note also the, now erratic, light curve that results, instead of the periodic X-ray pulsar curve in the moderate case.

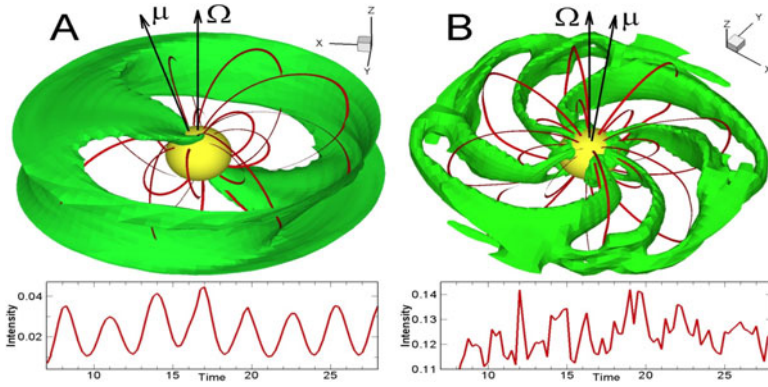


Fig. 14.9: Comparison of structure and generated light curves for accretion below and above the threshold for unstable inflow. (A) Accretion below the threshold rate produces nicely periodic light curves at twice the pulsar rotation frequency, as each hot spot presents itself to the observer. (This is identical to the situation in Figs. 14.7 and 14.8.) (B) Accretion above the threshold rate produces a chaotic light curve, generated by the numerous hot spots in which the inflow impacts the pulsar surface. Reproduced from Fig. 1 of [482], by permission of the AAS.

However, the threshold for unstable accretion flow is still very high for a pulsar: about 1% of $\dot{m}_{\text{cr}}$

$$\dot{m} \gtrsim 13 B_{10}^2 \quad (14.13)$$

for a weaker magnetic field of 10^{10} G. This somewhat super-Eddington rate of accretion is easily achievable ($\sim 4 \times 10^{-7} M_{\odot} \text{ yr}^{-1}$) in binary systems (see Chapter 11) and may be important for the Z sources.

So some would-be X-ray pulsars could be in a state where accretion through a weaker magnetic field is unstable, leading to flickering, but not pulsating, X-rays.

14.2.2 Plasma Flow onto Pulsars in the “Propeller” or Wind Turbine Regime

When the magnetic radius lies outside the co-rotation radius ($P < P_{\text{crit}}$), the rapidly-rotating pulsar begins to drive the accreting material away in a turbine-driven exhaust [483]. Fig. 14.10 shows the general picture of this inflow–outflow situation. Most of the matter flows out along cones with a wide opening angle. However, there is a less dense, but fast, component that collimates along the rotation axis into a jet. Ustyugova and her colleagues find that most of the energy and angular momentum carried away by this wind actually flows out in this jet.

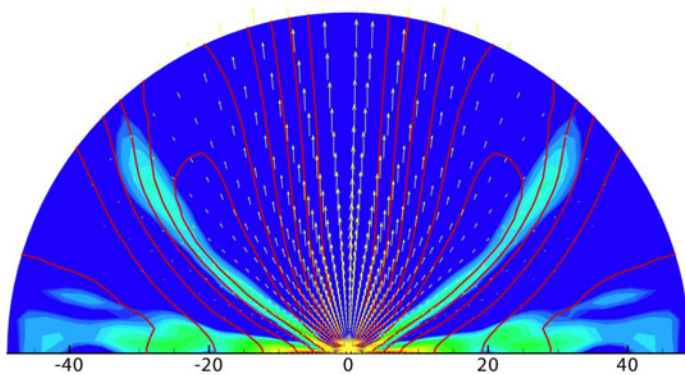


Fig. 14.10: An accreting pulsar in the propeller regime [483]. Pulsar rotation drives two types of outflow when the rotation is fast enough: a broad wind with a large opening angle (wide blue-green wings), which carries most of the outflowing mass, and a collimated jet along the polar axis (arrows), created by the magnetic tower effect (see Section 14.5.3), which carries most of the energy and angular momentum lost from the system. Color figure kindly provided by M. Romanova; reproduced with permission.

What is the total wind power and the spindown time of the turbine? There are two possible answers to the power question. There is substantial electromagnetic power in the rotating pulsar, which is given by a formula that is similar to equation (14.5)

$$\mathcal{P}^{\text{EM}} = B_m^2 R_m^3 \Omega_f = 5.1 \times 10^{36} \text{ erg s}^{-1} \dot{m}^{6/7} B_{12}^{2/7} P^{-1} \quad (14.14)$$

However, the turbo-exhaust also propels most of the accreted matter from $R = R_m$ to $R = \infty$. This needs a mechanical power comparable to the accretion luminosity to do this heavy lifting

$$\mathcal{P}^{\text{MECH}} \approx \frac{GM\dot{M}}{R_m} = 3.6 \times 10^{36} \text{ erg s}^{-1} \dot{m}^{9/7} B_{12}^{-4/7} \quad (14.15)$$

Which power dominates the pulsar spindown in the propeller regime, $\mathcal{P}^{\text{EM}}$ or $\mathcal{P}^{\text{MECH}}$? The answer is $\mathcal{P}^{\text{EM}}$. In fact, the two are equal when the pulsar period is ~ 2.4 times the critical period P_{crit} , and then for shorter periods $\mathcal{P}^{\text{EM}}$ increases faster. So, in the propeller regime ($P < P_{\text{crit}}$) we also find that $\mathcal{P}^{\text{EM}} > \mathcal{P}^{\text{MECH}}$. The spindown time of the wind/jet-producing X-ray pulsar then is

$$\frac{\mathcal{E}_{\text{rot}}}{\mathcal{P}^{\text{EM}}} = \frac{\frac{1}{2} I \Omega_f^2}{B_m^2 R_m^3 \Omega_f} = 136 \text{ yr } \dot{m}^{-6/7} B_{12}^{-2/7} P^{-1}$$

This is only a few hundred to a few thousand years for binary X-ray pulsars that have spun up to periods below ~ 1 s. So, if a binary X-ray pulsar is spun up to the wind turbine stage, it will remain in that state only for a small fraction of its total life. It will spin down in perhaps a thousand years or so, turn off its outflow, and, once again, enter the accretion/X-ray pulsar phase.

14.3 “Isolated” Black Holes with Magnetospheres

Black hole magnetospheres have much in common with those of pulsars. Indeed, the study of the latter has been crucial to the understanding of the former. Like pulsars, black hole magnetospheres rotate rapidly, produce a light cylinder (or surface), generate a magnetized wind that can potentially collimate into a jet, and can couple that outflow to the spinning black hole, removing angular momentum and rotational energy. Furthermore, there can be both “isolated” black hole magnetospheres, where particle creation replenishes the outflowing wind, as well as ones in which accretion supplies the wind material. This subsection deals with the “isolated” case.

As discussed at the beginning of this chapter, however, there are some key differences between black hole magnetospheres and those of other stars and planets. We already mentioned that, under most astrophysical circumstances, black holes are not expected to have significant *intrinsic* magnetic fields. Any field near a black hole horizon must be supported by currents in the ambient plasma surrounding, and accreting onto, the hole, not by currents on or inside the horizon itself. This means that all appreciable black hole magnetic fields are anchored in plasma outside the horizon, and those fields are temporarily *pressed* onto, or pulled into, the black hole by the accreting material. (An appreciably magnetized black hole, therefore, is never truly isolated, hence the quotation marks in the above title.) Therefore, the rotation of this field, which generates the wind, can be caused as much by the *orbital motion* of the plasma as by the *frame dragging* of that plasma by the rotating black hole metric. Some of the angular momentum carried away by the jet and wind may be due to the hole’s rotation, but some also may be due to the orbiting accreting material itself. Disentangling these two processes is still an unsolved problem, both theoretically as well as observationally.

The second major difference between magnetospheres of black holes and other stars is that, in addition to the outflowing wind that is ejected to (positive) infinity, there also is a wind that flows *inward* to the negative infinity of the black hole horizon. This concept of an ingoing wind is much more than a formality. It has its own light surface, as well as surfaces where the inflow speed reaches and exceeds the slow, Alfvén, and fast magnetosonic speeds, plus a surface where the inflow eventually becomes causally disconnected from the rest of the magnetosphere. Because a black hole magnetosphere must have both outflow and inflow, there must be a stagnation point/surface where the two flows balance and where much of the wind's plasma is supplied by particle creation or by accretion flow near the equator. Because the ingoing wind is causally connected both to the black hole ergosphere and to the outgoing wind, the former plays a crucial role in coupling the magnetosphere to the black hole rotation and extracting rotational energy and angular momentum.

The first authors to study black hole magnetospheres in detail were Roger Blandford and Roman Znajek [484], then at the Institute of Astronomy in Cambridge, UK. We will begin with a discussion of their pioneering analytic work and then discuss more papers that developed numerical models.

14.3.1 No Strong Magnetic Fields around “Isolated” Black Holes

We begin by discussing in a bit more detail the issue of why black holes cannot support strong magnetic fields. In principle, if a rotating black hole is charged, it can support quite a large magnetic field. Chapter 7 showed that black holes can have the properties of a giant elementary particle, including spin, charge, and a magnetic moment of strength

$$\mu_H = \frac{Q G M}{c^2} j$$

where Q is the total charge on the hole. Since $\mu_H = B_H r_H^3$, this implies a magnetic field on the horizon of

$$B_H = 2.4 \times 10^{19} \text{ G} \frac{q j}{\left(1 + \sqrt{1 - j^2}\right)^3} m^{-1} \quad (14.16)$$

where the dimensionless charge q is the total charge divided by the maximum possible on the hole

$$q \equiv \frac{Q}{G^{1/2} M} = \frac{Q}{5.14 \times 10^{29} \text{ statcoulombs}} m^{-1} \quad (14.17)$$

The problem with needing a large intrinsic black hole magnetic field is that it requires a large amount of electric charge to remain on the horizon, which is very difficult to maintain in the presence of particle creation or accretion of plasma.

Blandford and Znajek estimated that, if the magnetic field exceeded roughly

$$B_{\text{crit}} \approx 2.5 \times 10^5 \text{ G } j^{1/2} m^{-1/2} \quad (14.18)$$

or if the reduced charge exceeded the minuscule value of

$$q_{\text{crit}} \approx 10^{-14} j^{-1/2} m^{1/2} \quad (14.19)$$

then sufficient $e^\pm$ pairs would be created (in the same manner discussed for pulsars in Section 14.1) that any charge above this value would be discharged from the hole rapidly. At most, then, we expect magnetic fields no higher than B_{crit} on relatively isolated black holes. These field strengths are much weaker than the 10^{8-9} G ones being brought in by accreting matter (*e.g.*, equations (12.31), (12.35), (12.47), *etc.*). So any intrinsic magnetic field on a black hole is likely to be dynamically unimportant. For most purposes, therefore, black holes are assumed to be intrinsically unmagnetized, except for field anchored in the surrounding plasma.

14.3.2 Singular Surfaces of the Grad–Schlüter–Shafranov Equation

The development of black hole magnetosphere theory also parallels that of pulsars. Blandford and Znajek derived a form of the axisymmetric and time-independent GSS equation (the parent of the pulsar equation) in order to treat the black hole problem. This equation again implicitly makes the force-free assumption that only electromagnetic forces are important and that they are in balance. This assumption is valid for black holes undergoing very little accretion, such as the Galactic center supermassive black hole or single, stellar-mass black holes in the ISM.

Our form for the GSS equation is in equation (9.129)

$$\begin{aligned} \nabla \cdot \left\{ \frac{\alpha}{R^2} \left[1 - \frac{R^2}{\alpha^2 c^2} (\Omega_f - \omega)^2 \right] \nabla \Psi \right\} + \\ \frac{(\Omega_f - \omega)}{\alpha c^2} \frac{d\Omega_f}{d\Psi} (\nabla \Psi)^2 + \frac{1}{2\alpha R^2 c^2} \frac{dI^2}{d\Psi} = 0 \end{aligned} \quad (14.20)$$

The solution for Ψ serves both as the flux function (A^ϕ) and scalar potential (A^t) in the vector four-potential, with zero poloidal components (the two differing only by an additive constant). Ω_f again is the field angular velocity, while ω is the local angular velocity of the metric frame dragging (equation (7.47)), and α is the lapse function. $R = g_{\phi\phi}^{1/2} = \Sigma \sin \theta / \rho$ is the closest thing to a cylindrical radius in the Kerr metric.

The GSS equation has two sets of singular surfaces, and as in hydrodynamic wind theory, conditions on these surfaces can be used to constrain the solutions of the equation. As before, these conditions are internal to the equation itself, neither

boundary nor initial conditions but rather “regularity” conditions. A regularity condition simply allows the solution of an equation to pass smoothly through a surface where the denominator becomes zero.

14.3.2.1 The Light Surfaces

Unlike the pulsar form of the GSS equation, the full form above possesses more than one light surface. These surfaces occur when the coefficient of $\nabla\Psi$ in the first term in equation (14.20) is zero

$$\alpha c = R |\Omega_f - \omega|$$

(Note, however, that the entire first term in equation (14.20) does *not* vanish on the light surfaces. That occurs at the causality limits; see below.)

Let us consider the non-rotating Schwarzschild black hole case first ($\omega = 0$). There is, of course, a pulsar-like *outer* light surface for large R ($\gg r_S$), where $\alpha \approx 1$. Its shape is approximately cylindrical ($R_{L,out} \approx c/\Omega_f$).

However, there also is a light surface at small R

$$R_{L,in} = \alpha_{L,in} \frac{c}{\Omega_f}$$

where $\alpha_{L,in} = (1 - r_S/r_{L,in})^{1/2} \ll 1$, or in spherical coordinates

$$r_{L,in} = r_S + r_{L,in}^3 \sin^2 \theta \frac{\Omega_f^2}{c^2} \approx r_S \left(1 + \frac{r_S^2}{r_{L,out}^2} \sin^2 \theta \right) \quad (14.21)$$

As on the outer surface, this is where the field azimuthal velocity is formally equal to the local speed of light, but now it occurs deep in the black hole potential well. For uniform field rotation (Ω_f is the same everywhere), $r_{L,in} = r_S$ at the pole and is only slightly larger than r_S at the equator. The field lines again bend backwards, in the opposite direction to Ω_f (Fig. 14.11(top)) in the case when the black hole is non-rotating (Schwarzschild) or rotating in the same sense as the accretion flow (prograde Kerr).

The retrograde rotating Kerr black hole case ($\omega < 0$) adds the possibility that $\Omega_f - \omega$ will change sign at some point, with the black hole rotating faster than the magnetic field. When this occurs, then the bent field lines at the inner light surface will lead, rather than lag, the field rotation, because the field line must drift backwards against the rotating metric (see Fig. 14.11(bottom)).

In the non-force-free case (full relativistic MHD) the light surfaces lie near Alfvén surfaces, where the poloidal flow speed equals the poloidal Alfvén speed.

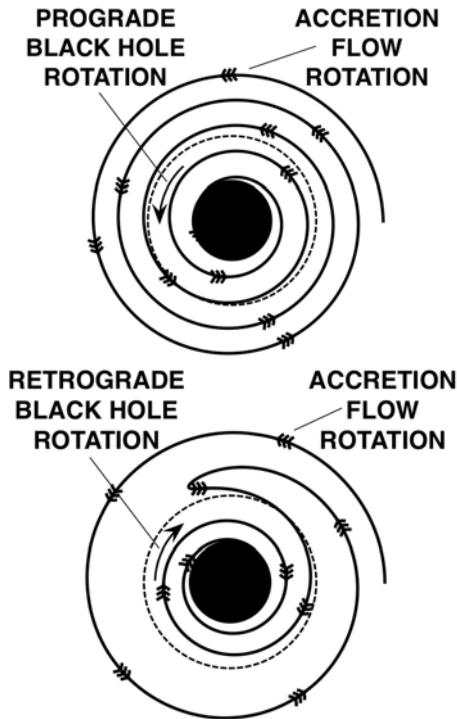


Fig. 14.11: MHD accretion onto rotating black holes, as viewed from the north polar axis of the accretion flow. Horizon is shown as a black spot and the hole’s equatorial ergospheric radius as a dotted line. The inner light cylinder typically lies well inside the ergosphere near the horizon. *Top:* A maximal *prograde* black hole that spins in the same direction as the accretion flow. Inspiral occurs along the magnetic field/stream lines. *Bottom:* A maximal *retrograde* black hole that spins in the opposite direction to the accretion flow. Inspiral still follows the field/stream lines, but it reverses its rotation as it nears the ergosphere and eventually joins the rotation of the metric.

14.3.2.2 The Causality Limit Surfaces

In the force-free case the causality limits occur at the horizon and infinity. That is, formally, any disturbance in the electromagnetic structure anywhere outside the black hole horizon can affect the overall structure of the magnetosphere. In the general case the horizon is specified by

$$\alpha = 0$$

Plugging this into equation (14.20), and being careful to take proper account of the geometric factors in the ∇ gradients, we find that the divergence term now *does* vanish. This allows one to integrate over $d\Psi$ to obtain the following condition on the poloidal current at the horizon

$$I [\Psi_H(\theta)] = (\Omega_f - \omega) \frac{R}{\rho} \frac{d\Psi_H}{d\theta} \quad (14.22)$$

[485, 486]. This is the so-called “Znajek horizon boundary condition”, but as discussed above, it actually is a regularity condition. Specifying the current and scalar potential Ψ to be related in this manner on the horizon simply allows the solution for Ψ to pass smoothly through the horizon without singularities.

In the non-force-free case the inner causality surface lies just outside the horizon and the outer one moves in from infinity to a large, but finite, distance from the black hole (see Section 15.1). The outer, so-called “fast magnetosonic separatrix surface” plays an important role in jet collimation.

14.3.3 The Basic Force-Free Black Hole Magnetosphere

Blandford and Znajek used perturbation techniques¹ to obtain an analytic solution to equation (14.20) for slowly rotating black holes ($j \ll 1$) with magnetic field lines *all* radiating outward in the same direction – the magnetic monopole solution. The field strength is given by

$$B^{\hat{r}} = B_0 \frac{r_g^2}{\Sigma} \quad (14.23)$$

$$H^{\hat{\phi}} = \frac{r_g^2}{Rc} (\Omega_f - \Omega_H) B_0 \sin^2 \theta$$

where Ω_f and $\Omega_H (= cj/2r_H)$ are both uniform and constant angular velocities, and B_0 is the characteristic strength of the magnetic field near the horizon. The radial field has a $1/r^2$ dependence, just like that we used in equations (12.87) for the MDAF “end game”. Blandford and Znajek matched $H^{\hat{\phi}}$ onto Michel’s [487, 488] pulsar wind solution for a monopole in flat spacetime:

$$H^{\hat{\phi}} = -\frac{\Omega r_g^2}{Rc} B_0 \sin^2 \theta$$

and deduced that the uniform angular velocity of the magnetic field must be

$$\Omega_f = \frac{1}{2} \Omega_H = \frac{c}{4r_H} j \approx \frac{c}{8r_g} j$$

where the approximation holds for $j \ll 1$. The final result for the azimuthal field, then, is

¹ We actually follow Komissarov’s discussion of Blandford and Znajek’s paper here [322].

$$H^{\hat{\phi}} = -\frac{r_g}{8R} j B_0 \sin^2 \theta \quad (14.24)$$

Note that we can deduce (from equation (9.134)) that the integrated poloidal current must be

$$I = \frac{1}{8} c r_g j B_0 \sin^2 \theta$$

Other quantities (D , E , ρ_q , *etc.*) also can be computed.

Blandford and Znajek, of course, realized that a magnetic monopole solution was unphysical and artificial. So they modified it to a “split monopole” one (a very small dipole located inside the black hole) with opposite magnetic poles in the two hemispheres

$$B^{\hat{r}}(\pi - \theta) = -B^{\hat{r}}(\theta)$$

(see Fig. 14.12). This solution is artificial as well, unless there is disk of plasma on the equator surrounding the black hole that supports the current sheet between the two hemispheres. ($B^{\hat{r}}$ changes sign on the equator, so $\mathbf{J} = \nabla \times \mathbf{B}/c$ must be large there.)

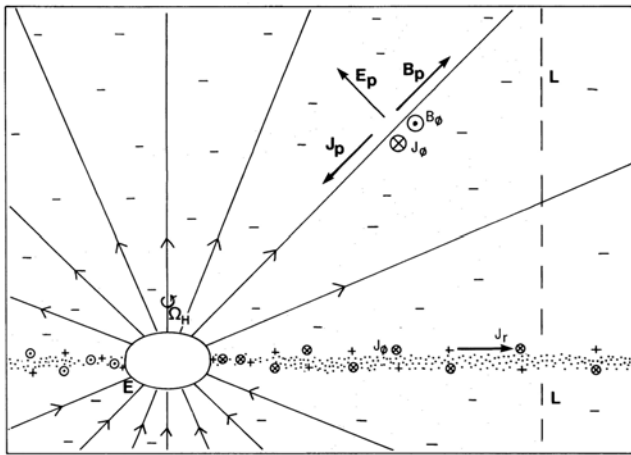


Fig. 14.12: The “split monopole” black hole magnetosphere model, with spin Ω_H , radial field lines, current sheet on the equator, and light cylinder (L). (Compare with the pulsar magnetosphere in Figs. 14.2 and 14.3, which also has an equatorial current sheet and light cylinder.) The current sheet can be more easily maintained if there is a disk of plasma on the equator. Reproduced from Fig. 2a of [484], by permission of J. Wiley and Sons, Ltd.

Just as in the pulsar case, a rotating magnetic field will generate an outflowing electromagnetic power (Poynting flux). Blandford and Znajek calculated this power and found a familiar expression (*cf.* equation (14.5)):

$$\begin{aligned}
\mathcal{P}^{\text{EM}} &= \frac{B_{\text{E}}^2 r_{\text{E}}^4 \Omega_f^2}{c} = \frac{B_{\text{H}}^2 r_{\text{H}}^4 \Omega_f^2}{c} \\
&\approx \frac{1}{16} B_{\text{H}}^2 r_g^2 c j^2 \\
&= 4.1 \times 10^{37} \text{ erg s}^{-1} \left(\frac{B_{\text{H}}}{10^9 \text{ G}} \right)^2 \left(\frac{M}{M_{\odot}} \right)^2 j^2 \\
&= 4.1 \times 10^{47} \text{ erg s}^{-1} \left(\frac{B_{\text{H}}}{10^5 \text{ G}} \right)^2 \left(\frac{M}{10^9 M_{\odot}} \right)^2 j^2
\end{aligned} \tag{14.25}$$

where r_{E} and r_{H} are the black hole ergosphere and horizon radius. We have used the $j \rightarrow 1$ limit in equation (14.25) (*i.e.*, $r_{\text{H}} \rightarrow r_g$), which reduces the power by a factor of 4, bringing it closer to more modern estimates. Therefore, like a pulsar, a rotating black hole with a magnetosphere generates electromagnetic power (a magnetized particle wind and jet), removing angular momentum as well as rotational energy and spinning down the black hole in the process. This effect, first pointed out by these authors, is called the “Blandford–Znajek process” or BZ effect.

As an example, we shall assume that the magnetic field strength on the horizon is of the same order as that at the ISCO. Then, if we plug in the magnetic field from various accretion solutions in Chapter 12 (which are valid in various ranges of accretion rate), we obtain²

$$\begin{aligned}
\mathcal{P}^{\text{EM,BZ}} &\approx L_{\text{Edd}} \dot{m} j^2 & \dot{m} < \dot{m}_A \\
&= 0.03 L_{\text{Edd}} \alpha^{-1/10} m^{-1/10} \dot{m}^{6/5} j^2 & \dot{m}_A < \dot{m} < \dot{m}_{\text{I}} \\
&= 0.5 L_{\text{Edd}} \dot{m}^2 j^2 & \dot{m}_{\text{I}} < \dot{m} < 1 \\
&\approx L_{\text{Edd}} \dot{m} j^2 & 1 < \dot{m}
\end{aligned} \tag{14.26}$$

where the ADAF-thin disk accretion rate boundary $\dot{m}_A$ is given by equation (12.54), and the “middle”–“inner” region boundary $\dot{m}_{\text{I}}$ is given (roughly) by equation (12.60). Below $\dot{m}_A$ an ADAF-like corona (with a possible magnetically-dominated inner MDAF) will dominate the inflow structure. In the supermassive black hole case (no SS middle region) the spindown time for the BZ process is then of order

$$\tau_{\text{spindown}}^{\text{EM,BZ}} \equiv M_{\text{red}} c^2 / \mathcal{P}^{\text{EM,BZ}} \approx 5 \times 10^7 \text{ yr } \dot{m}^{-1} \tag{14.27}$$

for $j \ll 1$. We are fairly confident of the first two above disk models, as they involve stable and well-determined accretion structures (ADAFs and thin, cool disks, respectively). However, the expression for $\dot{m}_{\text{I}} < \dot{m} < 1$ uses the unstable solution for the Shakura–Sunyaev “inner” region. And, when the accretion rate is super-Eddington ($\dot{m} > 1$), there are both theoretical (Section 15.1.1) and observational (Section 16.2.4) questions as to whether a *radio* jet is produced. However, this does not necessarily eliminate the possibility of a powerful, non-radio-emitting, black hole rotation-driven outflow.

² These and equations (15.17) and (15.20) are purposefully tooled to be similar. So they are valid to only one significant figure, if that.

14.3.4 Numerical Models of Force-Free Black Hole Magnetospheres

As with many areas of black hole research, this field has benefited by the recent development of numerical modeling and simulation techniques. Two people in particular, Serguei Komissarov of Leeds University, UK, in 2001 [489, 490] and, a few years later, Jon McKinney then at Harvard University in 2006 [491], developed two-dimensional force-free electrodynamics simulation codes based on the equations in Section 9.5.3. These differ significantly from those of standard FFDE (Section 9.5.2) in that alternate FFDE *explicitly* includes Ohm’s law and, therefore, can ensure always that $\mathbf{B} \cdot \mathbf{D} = 0$. Komissarov’s and McKinney’s codes carry this point one step further: they allow for a finite resistivity η_q (equation (9.41)), and, unlike our discussion in Section 9.2.2, they allow for the resistivity to be different along magnetic field lines (generally much smaller) than across them.

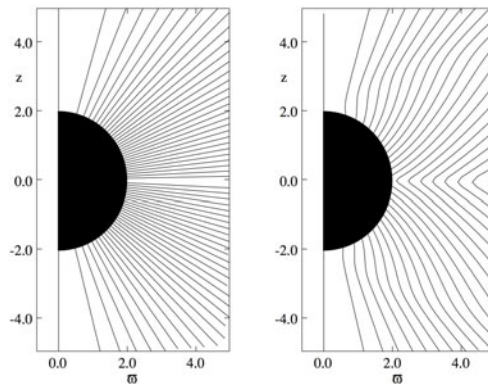


Fig. 14.13: Breakdown of the current sheet when there is no equatorial plasma disk (for a Schwarzschild black hole). Both panels show contours of the magnetic vector potential A_ϕ (*i.e.*, magnetic field lines). *Left:* Initial conditions (Blandford–Znajek split monopole solution in Fig. 14.12). *Right:* Magnetosphere after only five light-crossing times ($t = 5 \tau_g$). The magnetosphere is expelled from the black hole a short time later. Reproduced from Fig. 3 of [322], by permission of J. Wiley and Sons, Ltd.

Komissarov had studied the stability of the Blandford–Znajek split monopole solution in his major 2004 paper [322] (see Fig. 14.13, left panel). While not shown, the direction of the magnetic field lines just above the equator in the northern hemisphere are opposite in direction to those just below it in the southern hemisphere. This creates a current sheet on the equator (since $\mathbf{J} \propto \nabla \times \mathbf{B}$); however, in FFDE codes, there is no disk of plasma to support these large sheets of current. In Komissarov’s code, oppositely-directed magnetic field lines that touch are allowed to reconnect at a speed approaching c . Therefore, the current sheet dissipates (see the right panel in Fig. 14.13), and the magnetic field is expelled from the black hole after $t \sim 5 \tau_g$.

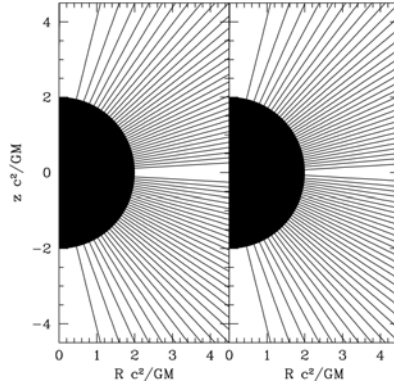


Fig. 14.14: Similar to Fig. 14.13, but with essentially zero charge drift enforced across the equator (and for a slowly-rotating Kerr hole with $j = 0.1$). *Left:* Initial Blandford–Znajek split monopole solution. *Right:* Magnetosphere after $t = 5 \tau_g$. In this case there is no breakdown of the current sheet and no expulsion of the magnetosphere. Reproduced from Fig. 4 of [491], by permission of J. Wiley and Sons, Ltd.

A paper in 2006 by McKinney took another look at the split monopole solution, this time enforcing zero (or nearly zero) charge drift across the equatorial current sheet [491]. (This condition mimics the presence of a large ionized accretion disk on the equator.) In this case McKinney finds quite different results (Fig. 14.14). Even after $t = 5 \tau_g$ the magnetic field structure remains intact. As BZ predicted, the split monopole solution is, indeed, stable as long as the current sheet is maintained and does not reconnect. This means that the split monopole solution is at least one physically reasonable picture of an “isolated” black hole magnetosphere.

14.3.4.1 Simulations of a Black Hole Immersed in an External Magnetic Field

The viability of the split monopole solution still raises some physical questions. Specifically, can such a configuration ever be achieved? The magnetic field lines penetrate the horizon, and this is specifically at odds with a famous magnetosphere solution found in 1974 by Robert M. Wald of the University of Maryland [492]. If a black hole is immersed in a uniform magnetic field, then, in Boyer–Lindquist coordinates, the electromagnetic field will have the following structure:

$$\begin{aligned}
 B^{\hat{r}} &= -\frac{B_0}{2 g_{\theta\theta}^{1/2} g_{\phi\phi}^{1/2}} \frac{\partial}{\partial \theta} (g_{\phi\phi} + 2 j r_g g_{t\phi}) & E^{\hat{r}} &= -\frac{B_0}{g_{rr}^{1/2}} \frac{\partial}{\partial r} (g_{t\phi} + 2 j r_g g_{tt}) \\
 B^{\hat{\theta}} &= \frac{B_0}{2 g_{rr}^{1/2} g_{\phi\phi}^{1/2}} \frac{\partial}{\partial r} (g_{\phi\phi} + 2 j r_g g_{t\phi}) & E^{\hat{\theta}} &= -\frac{B_0}{g_{\theta\theta}^{1/2}} \frac{\partial}{\partial \theta} (g_{t\phi} + 2 j r_g g_{tt}) \\
 B^{\hat{\phi}} &= 0 & E^{\hat{\phi}} &= 0
 \end{aligned} \tag{14.28}$$

In particular, this solution (see Fig. 14.15) shows that, for rotating black holes, $\mathbf{B}$ never penetrates the horizon; it is rejected by the black hole, just as a superconductor rejects a magnetic field. This is sometimes called the “Meissner effect” after Walther Meissner (of the Technical University of Munich) who studied this effect in 1933 in superconducting terrestrial materials [493]. Because we have expressions for both $\mathbf{B}$ and $\mathbf{E}$, we are able to *solve* for the equilibrium angular velocity Ω_f of the magnetic field from equations (9.131) and (9.132), which give

$$\Omega_f = -\frac{c E^{\hat{\theta}}}{R B^{\hat{r}}}$$

$$\text{or } R/R_L = -E^{\hat{\theta}}/B^{\hat{r}}.$$

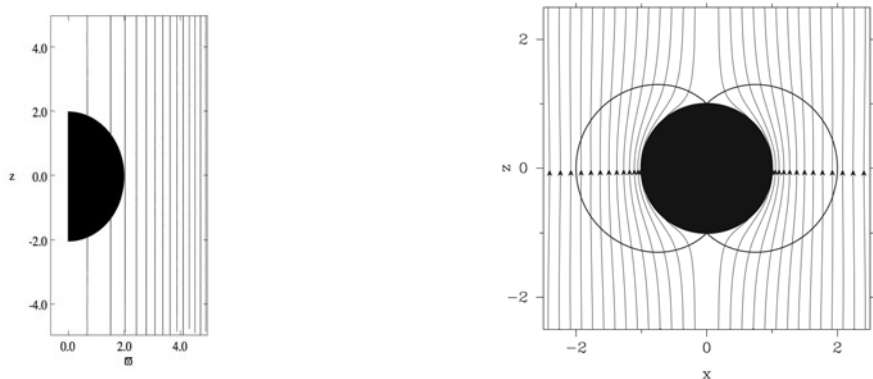


Fig. 14.15: The Wald *vacuum* solution for a black hole immersed in a uniform magnetic field. *Left:* For a Schwarzschild hole with $j = 0$; the magnetic field lines penetrate the horizon (or come extremely close to doing so). Reprinted from Fig. 1 of [322]. *Right:* For a maximal Kerr hole with $j = 1$; the magnetic flux is totally expelled from the horizon, while for values in the range $0 < j < 1$ some flux does penetrate and some is rejected. Thick solid line shows the ergosphere boundary. Reprinted from Fig. 1a of [494]. Reproduced by permission of J. Wiley and Sons, Ltd.

How can we reconcile the horizon penetrating BZ split monopole solution with the Wald flux-rejecting uniform field one? The answer was pointed out by Komissarov and McKinney in a joint 2007 paper: the BZ solution is plasma-filled, like the Goldreich–Julian pulsar solution, while the Wald’s is a vacuum solution [494]. In the latter, not only does the current vanish ($I = 0$ because $B^{\hat{\phi}} = 0$; see equation (9.132)), but the charge density ρ_q also is zero.

On the other hand, with *non-zero* charge and current densities, *and* with some black hole rotation, two things happen to change the electromagnetic field configuration. First, while the strong poloidal electric field (equation (14.28)) may be partially shielded by charge separation, the remaining electric field drives a strong poloidal current $I \neq 0$. Second, this current generates a significant azimuthal field $B^{\hat{\phi}}$ whose hoop stress pinches the poloidal magnetic field, drawing it into the horizon.

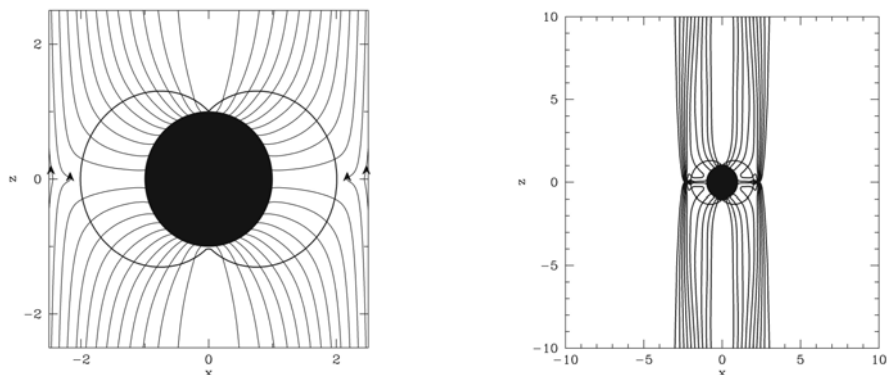


Fig. 14.16: Charge- and current-filled solution for a maximal Kerr black hole in a uniform magnetic field. *Left:* Same ergospheric region as in vacuum case in Fig. 14.15(right). Magnetic flux is now drawn *toward* the black hole, where it is available for extracting rotational energy and angular momentum. *Right:* Larger scale view of the same system, but showing contours of field angular velocity Ω_f . Note Poynting jet produced along the black hole rotation axis. Reproduced from Fig. 2 of [494], by permission of J. Wiley and Sons, Ltd.

Fig. 14.16(left) shows the final state of the rapidly-rotating Wald solution when plasma charge and current densities are allowed to rise to their Goldreich–Julian values. The field is no longer expelled from the black hole; it is “accepted with open arms”. *Bringing magnetized plasma near a rotating black hole, therefore, creates precisely the field configuration necessary to drive the BZ process and extract angular momentum and rotational energy from the black hole.* Fig. 14.16(right) shows contours of the angular velocity $\Omega_f = -cE^{\hat{\theta}}/RB^{\hat{r}}$ of this solution on a larger-scale. Note the highly-collimated, cylindrical Poynting-flux jet. The BZ effect is alive and well, and produces a (presumably) highly relativistic, collimated jet within $5\text{--}10\ r_g$ of the black hole.

14.3.4.2 The Ingoing Wind

Komissarov and McKinney made another, even more important, point. The fact that the field lines penetrate the horizon does not really matter; it is not the rotation of the horizon that actually drives the jet. It is the rotation of the *ergosphere* that does this, coupled with an *ingoing* wind of plasma that is causally connected to the outgoing Poynting wind/jet. This was first pointed out in 1990 by Brian Punsly and Ferdinand Coroniti of UCLA [495] and later that same year by Masaaki Takahashi and colleagues at Nagoya University [496]. The latter group also pointed out that energy extraction could be accomplished only if the Alfvén ingoing surface (or, in the force-free case, the ingoing light surface) occurs inside the ergosphere. These points were emphasized and demonstrated later by Koide in his 2002 papers on the numerical simulation of MHD ergospheric winds. We will discuss MHD simulations

more in the next subsection (14.4); here we will discuss the force-free aspects of the ingoing wind, which were elucidated in Komissarov’s important 2004 paper.

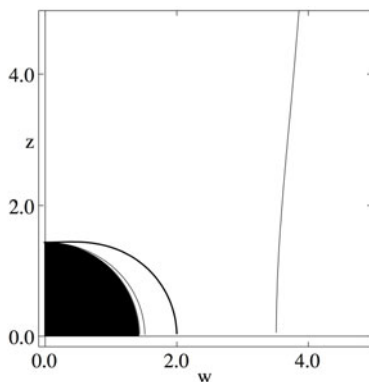


Fig. 14.17: The two light cylinders for a rotating Kerr black hole ($j = 0.9$). The one for the outgoing wind is shown by the thin vertical line near $r = 4 r_g$, while that for the ingoing wind is the thin curve visible near the horizon and inside the ergosphere boundary (thick line with equatorial radius $r_E = 2 r_g$). Reproduced from Fig. A1 of [322], by permission of J. Wiley and Sons, Ltd.

The ingoing wind is, qualitatively, a mirror image of the outgoing wind, but mirrored in a spheroidal (not flat or cylindrical) surface near the stagnation point (where the outgoing and ingoing winds both originate). Fig. 14.17 shows the light surfaces of both winds for a magnetosphere that is rigidly rotating at $\Omega_f = 0.7 \Omega_H$ around a black hole with $j = 0.9$. Komissarov noted several important points about these winds. Neither light surface is truly cylindrical; the outgoing one lies outside the ergosphere (thick line), while the ingoing one lies inside. The sense of curvature of the two light surfaces is generic and due to the structure of spacetime itself. Particle creation occurs generally everywhere between the two light surfaces, but most strongly near the stagnation point (not shown in the figure). The position of the latter depends on the details of the solution, but generally will be near or inside the ergosphere where $B^2 < D^2$ and the electric field cannot be fully shielded.

The relativistic ingoing wind also has a causality limit, which is nearly coincident with the horizon itself.

Finally, for a rigidly-rotating magnetosphere, the magnetic field lines in the ingoing wind bend *forward*, rather than backward, because of frame dragging of the plasma by the black hole. Inside the inner light surface the field rotates more *slowly* than spacetime, and that motion relative to the spacetime is *superluminal* there (just as the field rotates superluminally outside the outer light surface). Nevertheless, as before, the velocity of any plasma in the rotating spacetime must be *subluminal*, with the field lines slipping backward relative to the plasma. So the net effect is to produce a coiled magnetic field with a helicity opposite in sense to the black hole’s rotation direction. Each field line, therefore, forms a continuous helix with the same sense, from horizon, through each light surface, to infinity (see Fig. 14.11).

14.4 Black Hole Magnetospheres in the Presence of Accretion Flow

We now turn to the fourth, and most complex magnetosphere case: when the central object is a black hole (not a neutron star) and the material in the outflow is supplied by an outside source. Like the isolated black hole case in the previous subsection, the ingoing flow will be as important as the outgoing one. But, like the accreting neutron star case, the magnetic field likely will not be force-free. The dynamical forces of the plasma flow may affect, or sometimes even dominate, the structure of the black hole magnetosphere.

To compound matters, the subject of accreting black holes with outflowing magnetospheres and jets actually encompasses the entire subject of this book: the black hole engine. Therefore, in this subsection we will be very careful to limit the discussion only to the accretion inflow very near the black hole and its coupling to the launching of an exhaust outflow. The supply of that matter by accretion, and the complex dynamical and thermal structure of the inflow down to only a few r_S was discussed already in detail in Chapters 11 and 12. And the acceleration and collimation of the outflowing exhaust into a high-speed MHD wind/jet will be discussed in the following section of this chapter.

14.4.1 Origin and Generic Structure of the Black Hole Magnetic Field

There are three possible sources of the black hole magnetic field in the presence of accretion:

1. Produced by an intrinsic black hole magnetic moment.
2. Created from magnetic flux dragged in by the accreting matter.
3. Generated *in situ* by local dynamos in the accretion inflow.

We shall look at each of these in turn.

14.4.1.1 The Very Weak Intrinsic Black Hole Magnetic Field

In Section 14.3 we showed that an isolated black hole can support only a weak charge q_H on its horizon and, therefore, only a weak magnetic moment and field (less than $\sim 10^5 \text{ G m}^{-1/2}$). In the presence of accretion the critical magnetic field strength is even lower ($< (0.011\text{--}21) \text{ G m}^{-1}$); if the accreting matter is ionized, most of the charge on the black hole can be carried away by an electric current in the inflow, discharging the black hole in a time of order τ_g .

Consider the case when the black hole is *positively* charged, and the accreting material is an ionized electron–proton plasma. The electrons will be attracted both

gravitational and electrically, and so will certainly accrete into the black hole. However, the protons could be attracted or repelled, depending on the competition between the gravitational and electrical forces on each proton. When

$$\frac{q_H e}{r^2} \geq \frac{G M m_p}{r^2}$$

the protons will be repelled electrically away from the black hole with a greater force than they are attracted gravitationally. This outflowing wind of positive charges represents a current flowing out of the black hole: more negative electrons will be accreted than protons. The hole, therefore, will lose its positive charge until q_H drops below a critical value. At that point, the black hole will have a dipole magnetic field no stronger than $B_{H,\text{crit}} \equiv q_H c^4 j / G^2 M^2 = 21 \text{ G } j \text{ m}^{-1}$.

The situation is even worse when the expelled charge-carrying particles are lighter electrons or positrons, *i.e.*, when the hole is *negatively* charged, or when the accreting material is an electron–positron plasma. m_p is replaced in the above expression by m_e , yielding a discharged black hole with a critical magnetic field a factor of m_e/m_p smaller ($B_{H,\text{crit}} = 0.011 \text{ G } j \text{ m}^{-1}$). These magnetic field strengths are so small that they are completely unimportant astrophysically. Accreting black holes, therefore, can be considered uncharged and intrinsically unmagnetized.

14.4.1.2 Accretion of Extrinsic Magnetic Flux

That the source of the magnetic field in some accreting sources may be external is a real possibility, and it corresponds to the magnetically-dominated accretion flow discussed in Section 12.3. When magnetic flux is conserved and accreted, at least one component of the magnetic field will increase inward as $\propto R^{-2}$ (*e.g.*, $B_R \propto R^{-1} H^{-1} \propto R^{-2}$). This is a substantial enhancement of the field inward and corresponds to the magnetic pressure increasing as $p_m \propto B^2 \propto R^{-4}$. However, the gravitational forces generally increase in a similar manner ($GM\rho/R \propto R^{-3} R^{-1}$). So, magnetic forces may dominate dynamical (pressure) forces, but they will not dominate gravity in this situation.

Can magnetic fields dragged inward become dominant in an accretion flow? The answer depends partially on whether the plasma inflow is gas-pressure-dominated and adiabatic or not. If it is, then the dynamical forces increase inward as $p_g \propto \rho^{5/3} \propto R^{-5}$, which is *steeper* than the inward increase of magnetic pressure and steeper than that of gravity; *i.e.*, plasma pressure can halt an inward accretion flow, *and* become dominant over an initially much stronger magnetic pressure. On the other hand, if the pressure is radiation or relativistic particle dominated ($p_r \propto \rho^{4/3} \propto R^{-4}$), then a initially dominant magnetic field can remain so. Finally, if the plasma radiates so dramatically that it remains, say, isothermal ($p_g \propto \rho \propto R^{-3}$), then an initially *weak* magnetic field can become dominant inward if flux is conserved.

The answer to magnetic dominance also depends on the amount of angular momentum in the inflow and if angular momentum is conserved in the flow or carried

away by some means. If conserved, then the rotational forces ($\rho V_\phi^2 \propto R^{-3} R^{-2} \propto R^{-5}$) increase inward as rapidly as an adiabatic ideal gas. Even in isothermal inflow, rotation can halt inward accretion and shape the inflow much more than magnetic forces.

14.4.1.3 Magnetic Flux Generation *in situ* by Shear-Driven Dynamos in the Accretion Inflow

This case will occur whenever turbulent or rotational shear develops in the inflow. It corresponds to the type of accretion discussed extensively in Section 12.2. As implied above, *in situ* flux generation, via shear dynamos, will occur near black holes whenever rotational forces dominate somewhere in the inflow. And this will occur in most accretion inflows, unless there is an efficient and very rapid mechanism for transferring angular momentum outward. Except possibly in a hard accretion state, where the magnetic field may indeed dominate rotational forces, *in situ* generation of flux is expected to be the most likely source of magnetic field in accretion flows. Even our magnetically-dominated accretion flows in the previous chapter assume that the magnetic flux they drag inward was generated in a turbulent accretion inflow at a somewhat greater distance from a black hole, not dragged in from another star or from the interstellar medium.

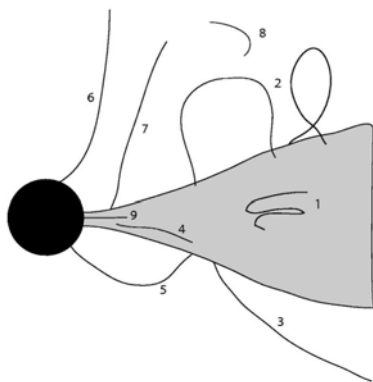


Fig. 14.18: Extension of the Blandford magnetic field line classification, adding #8 (coronal loops) and #9 (direct connection of plasma inflow to black hole horizon). Reproduced from Fig. 5 of [497], by permission of the AAS.

14.4.1.4 Classification of Magnetic Field Lines

In a 2002 paper entitled “To the Lighthouse” (given at a conference on *Lighthouses of the Universe*), Roger Blandford suggested a classification scheme for seven dif-

ferent types of magnetic field lines in accretion flows near black holes [362]. In a subsequent paper in 2004, Shigenobu Hirose and his colleagues at Johns Hopkins (Julian Krolik) and University of Virginia (Jean-Pierre De Villiers and John Hawley) [497], after having performed extensive simulations of such flows, added two more types of field line to the classification, which we adopt below. The nine types of field line are (see Fig. 14.18):

1. Field lines amplified in the turbulent accretion flow by the MRI.
2. Closed loops of field that connect one accretion disk radius to another, through the corona, and energize that corona when they reconnect.
3. Open field lines that connect the accretion disk to an outflow and drive an outflowing MHD wind.
4. Field lines that connect the orbiting accretion flow outside the ISCO to the plunging inflow inside r_{isco} .
5. Coronal field lines that connect the disk to the horizon itself.
6. Open field lines that connect the horizon to a strong relativistic Poynting-flux jet.
7. Open field lines that connect the flow inside the ergosphere to an outflowing wind or jet.
8. Loops of field that connect different parts of the corona and which can close and reconnect.
9. Field lines that connect the plunging flow to the horizon itself.

Chapter 12 dealt mainly with Types 1, 4, and 9 and Section 14.3 with Type 6. This section will discuss Types 2, 3, 5, and 7 and will say more on Type 9. The details of the outflow along lines 3, 6, and 7 are discussed in detail in Section (15.1). Type 8 will not be discussed much in this book.

14.4.2 Structure of a Magnetosphere with an Accretion Disk that Does Not Reach the Black Hole

We begin with the simplest modification to the force-free black hole magnetosphere discussed in Section 14.3: the addition of a conducting disk on the equator. As a further simplification, we will assume for the moment that the disk does not extend all the way to the black hole horizon or even into the ergosphere. Such a configuration is most appropriate for low accretion rate sources ($\dot{M} \ll \dot{M}_{\text{Edd}}$) and, indeed, bears a striking resemblance to the MDAF models discussed in Section 12.3. Initially, we also will assume that the plasma in the disk is not even accreting, although it may be rotating.

While this configuration is far simpler than reality, it still is sufficiently complex that we will introduce it in steps. We begin with the easiest case where *nothing* rotates.

14.4.2.1 Vacuum Magnetosphere of a Non-Rotating Black Hole Where Neither Disk Nor Magnetic Field Rotate

The steady-state structure of the magnetosphere in this case is given by the GSS equation (9.129) with $\Omega_f = \omega = 0$ (which implies $\rho_q = 0$) and $I = 0$. In Schwarzschild–Hilbert coordinates this becomes (with curvilinear Schwarzschild geometric factors taken into account) simply

$$r^2 \frac{\partial}{\partial r} \left(\alpha^2 \frac{\partial \Psi}{\partial r} \right) + \sin \theta \frac{\partial}{\partial \theta} \left(\frac{1}{\sin \theta} \frac{\partial \Psi}{\partial \theta} \right) = 0$$

where $\alpha^2 = (1 - 2GM/c^2)$ is the square of the Schwarzschild lapse function. This equation was solved in 2001 by Akira Tomimatsu and Masaaki Takahashi (then at Nagoya University and the Aichi University of Education, respectively), assuming that the only current in the system is confined to the disk itself [498]. The disk has an inner edge (defined to be at $r = r_0$), and the disk surface current is assumed to drop off as a power law ($dI/dA \propto r^{-s-2}$). Therefore, if $s > 0$, most of the current generating the magnetosphere lies in a loop near the inner disk edge at r_0 , which produces a toroidal structure of closed poloidal field lines near that radius.

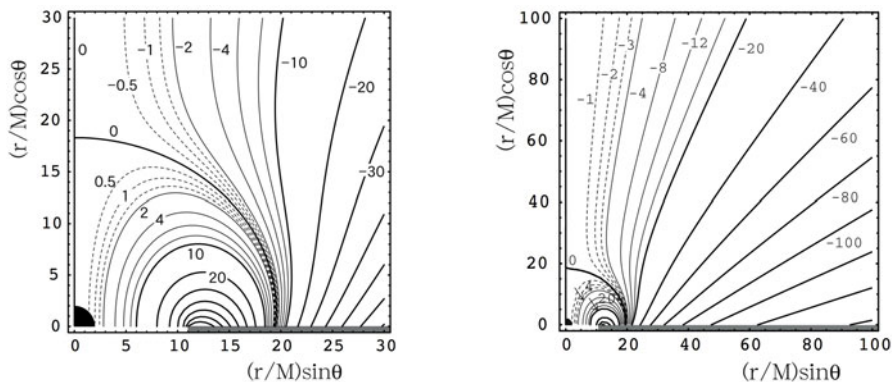


Fig. 14.19: *Vacuum* solution for a magnetosphere around a Schwarzschild black hole with a thin, non-rotating accretion disk on the equator. Note that field lines avoid the horizon, as in other vacuum solutions, even in this non-rotating case. *Left:* Inner $\sim 30 r_g$, showing both closed and open field regions, separated by the separatrix (marked 0). Dashed lines show the separatrix flux tube, which begins near the disk and expands into the closed and open regions. *Right:* Same solution for $r \lesssim 100 r_g$. Reproduced from Figs. 1 and 2 of [498], by permission of the AAS.

Fig. 14.19 shows the inner and outer structure of Tomimatsu and Takahashi’s magnetosphere. While we have simplified the problem quite a bit, the model nevertheless shows several general features expected in more complex global magnetospheres with accretion flows:

- There are open field lines extending from the disk to infinity. An outflowing MHD wind/jet is expected along these (Type 3) field lines.
- There also are closed field lines extending toward the black hole. But, because the magnetosphere is a vacuum solution, the field cannot intersect the horizon.
- Finally, there is a *separatrix* field line that marks the boundary between the closed and open systems, just as in the pulsar case.

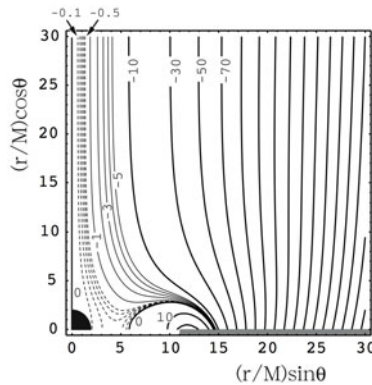


Fig. 14.20: Similar to Fig. 14.19, but with an externally applied uniform magnetic field (as in Fig. 14.15). The external field dominates most of the structure, although there still exists a closed field region near the disk inner edge. Reproduced from Fig. 4 of [498], by permission of the AAS.

Tomimatsu and Takahashi also studied the case when the black hole is immersed in a strong, uniform magnetic field – the equivalent of the Wald problem (Section 14.3) but with a conducting disk, supporting a current sheet, on the equator. Fig. 14.20 shows how this external uniform field modifies the magnetosphere with additional magnetic field lines near the polar axis.

While no one expects that an astrophysical black hole will be immersed in a strong external field from a powerful cosmic source, the reverse is possible: the structure of the magnetosphere may be such that, at large distances, there is a net, quasi-uniform magnetic field. Fig. 14.20 is, therefore, a model for that situation.

14.4.2.2 Non-Vacuum Solution for a Non-Rotating Black Hole with Rotating Disk and Field

Analytic Models. The next level of complexity adds in possible rotation of the disk and magnetic field, and, in doing so, introduces plasma charge density and poloidal currents that may be required in the magnetosphere itself to support B^ϕ ($\omega = 0$, $\Omega_f \neq 0$, $I \neq 0$). The GSS equation regains most of its complexity in this case:

$$\nabla \cdot \left\{ \frac{\alpha}{R^2} \left[1 - \frac{R^2 \Omega_f^2}{\alpha^2 c^2} \right] \nabla \Psi \right\} + \frac{1}{2\alpha c^2} \left[\frac{d\Omega_f^2}{d\Psi} (\nabla \Psi)^2 + \frac{1}{R^2} \frac{dI^2}{d\Psi} \right] = 0 \quad (14.29)$$

Taking into account rotation of the field means that we must deal with the two light surfaces, where

$$\alpha^2 c^2 = R^2 \Omega_f^2(\Psi)$$

Note that we assume that Ω_f is constant along a given field line but not necessarily over all space; *i.e.*, $d\Omega_f/d\Psi \neq 0$. The function $\Omega_f(\Psi)$ is determined simply by setting Ω_f equal to the (relativistic) Keplerian angular velocity at the footpoint of the field line. This determines the poloidal current function $I(\Psi)$ uniquely (Section 14.1). The outer light surface is quasi-cylindrical and can be handled in the same way as Contopoulos did in the pulsar case – constraining Ψ and its gradient to be continuous across this light surface. The structure of the outflowing wind in the black hole case, therefore, should be qualitatively similar to that in the pulsar case (Figs. 14.1 and 14.2(right)). In 2004 Dmitri Uzdensky (while at UC Santa Barbara) solved equation (14.29) for the *inner* portion of the black hole magnetosphere, treating the inner light surface in a similar manner [499].

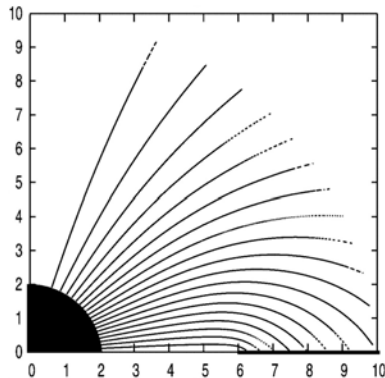


Fig. 14.21: Charge- and current-filled solution for the Schwarzschild black hole magnetosphere with a Keplerian disk. As in Fig. 14.19 the field lines create a closed structure near the disk inner edge, but now they penetrate the horizon, as in other charge-filled solutions. (Note: disks rotating at different rates, including zero, produce structures virtually identical to this one.) Reproduced from Fig. 3 of [499], by permission of the AAS.

Uzdensky's results are shown in Fig. 14.21. Unlike the structure of Tomimatsu and Takahashi, here all closed field lines penetrate the black hole horizon. The difference, of course, is that Uzdensky allowed $I \neq 0$ in the magnetosphere, allowing field lines to penetrate the horizon just as they did in Komissarov and McKinney's joint force-free paper (Section 14.3). Surprisingly, this approximately uniform ra-

dial magnetic field structure is independent of the field rotation rate. Even in the non-rotating field case ($\Omega_f \rightarrow 0$) and even though $I \rightarrow 0$ in the magnetosphere in that case, there still exists a *current sheet* inside the disk inner edge ($r_S < r < r_{\text{isco}}$), which keeps the field lines from closing onto themselves. The apparent robustness of the radial field configuration in a black hole magnetosphere is the main reason this structure was chosen as the “end game” of magnetically-dominated accretion flows in Section 12.3. Indeed, black hole magnetospheres with ingoing winds *are* MDAFs.

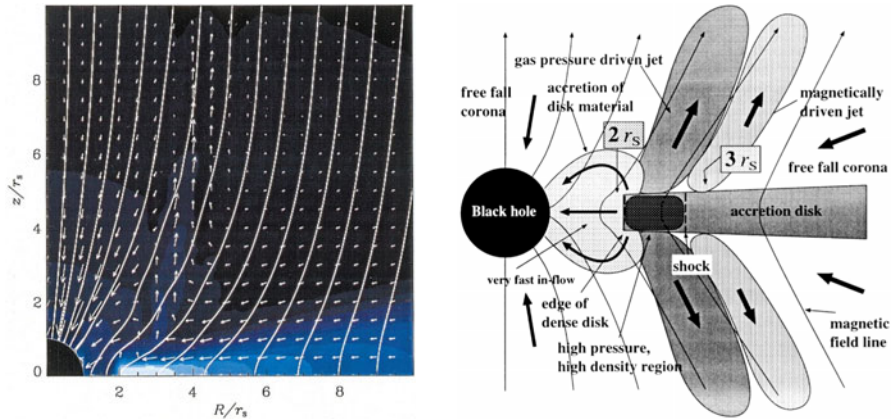


Fig. 14.22: General relativistic MHD simulation of a Schwarzschild black hole with a magnetized Keplerian accretion disk. *Left:* Black hole is at lower left, and accretion disk (blue-white) along the equator is threaded with (white) field lines from an initial Wald solution. Coordinates are labeled in units of $r_S = 2r_g$. Accretion of the plasma drags those field lines onto the black hole, where they penetrate the horizon. An outgoing wind/jet rises from the rotating accretion disk, but not from the non-rotating hole. *Right:* A schematic diagram showing the various parts of the inflow and outflow. Note the ingoing and outgoing winds (heavy black arrows). Reproduced from Figs. 6d and 15a of [500], by permission of the AAS.

Numerical Simulations. The above analytic models are steady state, of course, but the non-steady problem actually was investigated numerically some six years earlier. In 1998–99 Shinji Koide (then at Toyama University), Kazunari Shibata (Kyoto University), and Takahashi Kudoh (National Astronomical Observatory of Japan [NAOJ]) performed some of the first MHD simulations of Schwarzschild black holes accreting magnetized disk and corona material [500]. And these were, in fact, the first GRMHD simulations to show that a jet (of any kind) can be driven from an accreting black hole. In perhaps their most realistic simulation, Koide *et al.* began with a Keplerian disk, with an inner edge at $R \approx 6r_g$, and a freely falling hot corona filling the space above that disk. (This is essentially a Shakura–Sunyaev disk with an ADAF-like corona above.) The entire initial model was threaded with a uniform Wald field and then allowed to evolve.

These authors' final model, at $t = 104 \tau_g$, is shown in Fig. 14.22(left). This is enough time for the inner disk at $6 r_g$ ($3 r_S$) to rotate only once and for much of its material to sink toward the black hole to $4 r_g$ due to the transfer of angular momentum outward by the strong field. While not yet in a steady state, we nevertheless can identify all of the main features expected in an accreting Schwarzschild hole from our above discussion, plus some additional ones discussed later in this chapter:

1. A stagnation point has developed above the disk at $R \approx 5 r_g$ ($2.5 r_S$). At that point the accretion feeds both ingoing and outgoing winds.
2. The ingoing wind, not being in a steady state, drags field lines toward the black hole, rather than simply flowing along them.
3. The outgoing wind travels, more or less, along Type 3 field lines and actually divides into two parts (see Fig. 14.22(right)): a “magnetically-driven wind/jet” from the outer thin part of the disk and a “gas pressure-driven wind/jet” from the inner, gas pressure bloated portion of the disk.
4. We now can identify these two winds as being, essentially, “cold” and “warm” MHD winds, respectively (Section 15.1).
5. Finally, at high latitudes we see that there exists only an *ingoing* wind (accretion) along Type 6 field lines that thread the black hole horizon. The reason for this, of course, is that this black hole, and therefore these field lines, do not rotate, so no outgoing wind can occur there. This is expected to change for rotating holes.

One odd property of these ejected jets is that their speeds (out to $Z \approx 20 r_g$) are only moderately relativistic ($V = 0.88c$ or $\gamma \approx 2$). Such speeds are among the slowest observed for black hole jets. This possible discrepancy between theory and observation worried some investigators at the time. However, we now know this result actually to be very consistent with both jet theory and observations. Quite a bit of jet acceleration occurs between $20 r_g$ and $1000 r_g$ (see Section 15.1), so the velocities of $\gamma \sim 3\text{--}30$ observed by VLBI, at distances of >0.1 pc ($>2000 r_g$) from a central $10^9 M_\odot$ black hole likely would be much slower ($\gamma \sim 2$) when first launched.

14.4.2.3 Rotating Black Hole with Rotating Disk and Field: Closed Field Solution

The case where everything rotates (disk, field, and black hole) also was investigated by Uzdensky the following year (at Princeton in 2005), again only for the inner portion of the magnetosphere [486]. The equation to solve in this case is now the *full* GSS one (9.129). He again imposed regularity conditions at the ingoing light surface and causality limit (horizon) and allowed each field line to have the (relativistic) Keplerian angular velocity where it was anchored in the rotating disk ($\Omega_f = \Omega_K$). His magnetosphere solutions near a rotating black hole (Fig. 14.23) look similar to those in Fig. 14.21 for a non-rotating black hole, but with a very important addition: we now see where the separatrix between the closed and open magnetic field regions lies and find that the former field lines cannot extend from the black hole out to

arbitrarily large disk radii. Instead, the last field line to intersect both black hole and disk now does so *at a finite radius* R_{sep} ; field lines beyond the separatrix radius are open.

Furthermore, R_{sep} is a *decreasing* function of the spin parameter j . That is, the transfer of angular momentum from the black hole to the disk becomes highly concentrated near the disk inner edge as the black hole spin increases. Thus, the magnetospheres have a poloidal magnetic field structure qualitatively similar to those of Tomimatsu and Takahashi, with inner closed field lines, a separatrix, and outer open field lines (*cf.* Fig. 14.19).

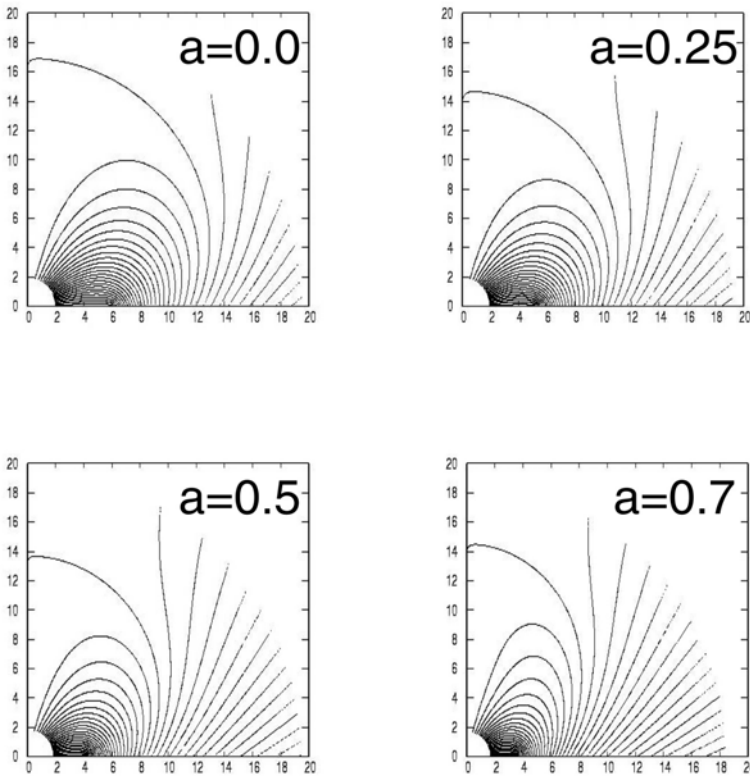


Fig. 14.23: Charge- and current-filled solutions for Kerr black hole magnetospheres as a function of the spin parameter j (here given as a). The upper left panel is an expanded version of Fig. 14.21 and now shows both closed and open field structure. The separatrix radius and closed field region shrink as j increases. (No solutions were found for $j > 0.7$.) Reproduced from Fig. 5 of [486], by permission of the AAS.

Why is this the case? Why *must* there be open field lines anchored in the disk for a rotating black hole? Qualitatively the reason is as follows. In addition to the poloidal field structure in Fig. 14.23, the field lines also have a toroidal component

and, therefore, are helical in shape. When the black hole rotates much more rapidly than much of the disk, the angular velocity difference between where the separatrix penetrates the horizon and it is anchored in the disk $(\Omega_f - \omega) = (\Omega_f - \Omega_H)$ can be large. While the field can slip to some extent along the horizon, nevertheless it must be highly twisted there (see Fig. 14.11). In a stable configuration the hoop stress of this toroidal field near the horizon is balanced by the tension of the rest of the field line as it stretches out to the disk. However, for any black hole with non-zero rotation, there must be a disk radius (R_{sep}) beyond which the hoop stress of the coiled part is too much for the modest tension in the outer portion. No stationary solution can be found for closed field lines beyond the radius that connects the separatrix with the black hole. (In a time-dependent simulation, for field lines beyond the separatrix and possibly for all field lines for $j > 0.7$, the coiled part and outer part may disconnect from each other, with the inner coil propagating upward as a Poynting jet and the outer part becoming open above the disk. We will discuss this issue more in the subsection below on magnetic towers (14.5.3).)

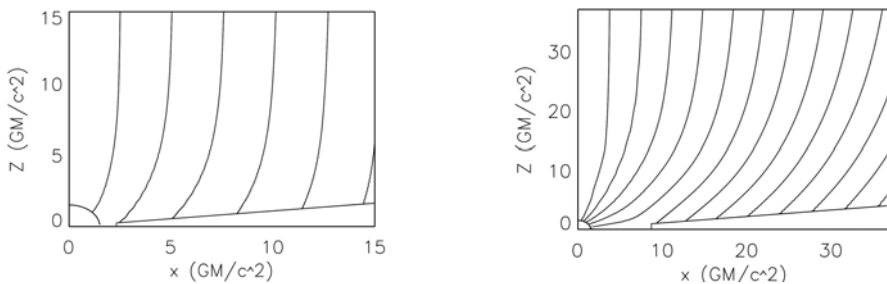


Fig. 14.24: Open field black hole magnetospheres with both a rotating hole and a Keplerian disk. The hole is the (unblackened) quarter-circle at lower left, and the disk is the thin wedge near the equator. Length units are in r_g , but note that the panels have different linear scales. *Left:* A rapid *prograde* Kerr black hole with $j = 0.9$. The ISCO of this disk lies at $2.32 r_g$, leaving only a small gap between it and the horizon radius at $r_H = 1.56 r_g$. *Right:* A rapid *retrograde* Kerr black hole with $j = -0.9$. The ISCO in this case lies at $8.72 r_g$, leaving a substantial gap between it and the horizon (also at $1.56 r_g$). Adapted from Figs. 5 and 6 of [501]. Reproduced by permission of the AAS.

14.4.2.4 Rotating Black Hole with Rotating Disk and Field: Open Field Solution

The same problem also was solved by David Garofalo (of the University of Maryland and, more recently, the Jet Propulsion Laboratory) [501]. Instead of looking for closed magnetic field solutions of the Grad–Schlüter–Shafranov equation, he sought *open* field solutions with both the disk and black hole magnetic field lines stretching to infinity, similar to the field structure in Fig. 14.22. See Fig. 14.24. This produced a significant flux ‘gap’ between the horizon and ISCO, resulting in much more open

magnetic flux being deposited on the black hole horizon. He also was able to obtain solutions for a wider range of black hole spins, $-0.9 < j < 0.9$.

Why would this approach be more successful in producing model magnetospheres than the closed field approach? The answer might lie with the “magnetic tower” effect, discussed below in Section 14.5.3. When there is significant shear between the black hole horizon and the ISCO of the accretion disk, the closed field loops in Fig. 14.23 will be sheared and expanded in the same manner as disk loops would be (see Fig. 14.35). Eventually they will break and become open. Initially, the field lines accreted onto the hole will be of the opposite sense to those in the disk. But, eventually, the field in the disk also will accrete onto the black hole and overwhelm and annihilate any anti-parallel field lines. This will result in the more stable, open-field magnetospheres in Fig. 14.24.

Garofalo applied these new magnetosphere results to calculate a more accurate, and interesting, spin-dependence of the Blandford–Znajek effect. We can approximate his results here and display them graphically. First, his results for the black hole magnetospheres can be approximated by assuming that the accretion flow from the ISCO to the horizon conserves magnetic flux. This yields

$$B_H = B_{Z,\text{isco}} \frac{r_{\text{isco}}(j)^2}{r_H(j)^2}$$

Expressions for $r_H(j)$ and $r_{\text{isco}}(j)$ are given in equations (7.48) and (7.60), respectively. To obtain an expression for $B_{Z,\text{isco}}$ as a function of j , we use either equation (12.52) for an ADAF or equation (12.31) for a thin cool disk and use $x = r_{\text{isco}}/6r_g$ in the B_Z expression. Finally, we recognize that, if the shear between disk and hole is small enough, then the closed-field Uzdensky magnetospheres may, indeed, be appropriate, leading to *no* significant Blandford–Znajek jet. To approximate this, we multiply by a magnetic tower efficiency factor that is proportional to $[(\Omega_H - \Omega_{\text{isco}})/\Omega_H]^p$, where we take $p = 1$ at this time, lacking any other information. This modifies equation (14.25) to be similar to the pulsar power

$$\mathcal{P}_{\text{BZ}}^{\text{EM}} = \frac{B_H^2 r_H^4 \Omega_H (\Omega_H - \Omega_{\text{isco}})}{c} \quad (14.30)$$

Fig. 14.25 shows the results of this calculation for the two more reliable disk models (hot ADAF (solid line) and cool Shakura–Sunyaev “middle” region (dashed line)) at the accretion rate boundary $\dot{m} = \dot{m}_A$. Compare the figure with the first two of equations (14.26). The two curves are separated by almost two orders of magnitude in jet power, but the spin-dependence of each is not at all symmetric about $j = 0$, except for the region where $|j| \ll 1$. Because of the large gap between ISCO and horizon for large negative spin, the square of the magnetic field produces a much stronger jet in equation (14.30) than for large positive spin. Indeed, as $j \rightarrow 1$, the shear between hole and disk may be so weak that a strong closed-loop (Uzdensky-type) magnetosphere may develop, leading to no jet whatsoever for high prograde black hole spins. There also are two additional places where the BZ jet power may

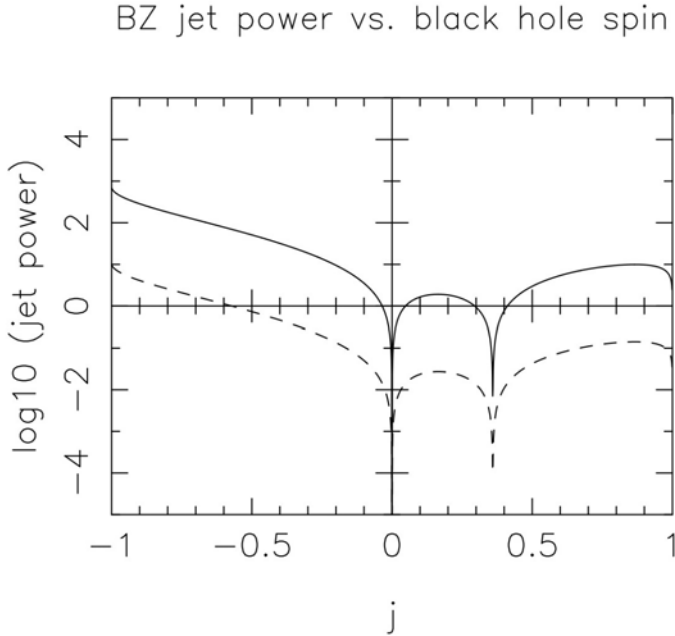


Fig. 14.25: Plot of the electromagnetic jet-production power of the Blandford–Znajek process vs. black hole spin j (eq. (14.30)) for two different types of accretion flow. *Solid line:* ADAF (eq. (12.52)). *Dashed line:* Thin cool disk (eq. (12.31)). The most powerful jets occur when the accretion flow is geometrically thick ($H \sim R$) and the black hole rotation is retrograde. Significant decreases in the BZ jet power may occur when $\Omega_H = 0$ or $\Omega_H = \Omega_{\text{isco}}$.

go to zero: the Schwarzschild case (black hole is not rotating) and $j \approx 0.36$ (where the hole and ISCO rotate at the same rate).

Not counting the sharp excursions to zero, the jet power can vary by about four orders of magnitude from the peak of the solid curve at $j = -1$ to the top of the hump in the dashed curve at $j \approx 0.15$, due to both spin and jet suppression effects.

14.4.2.5 The Nature of the Ingoing and Outgoing Winds

Up until this point we have not considered the nature of the actual accretion inflow or wind outflow along the field lines in Fig. 14.23. We will discuss the wind outflow in some detail in Section 15.1. The ingoing wind was considered by Kouichi Hirotani and his colleagues at Nagoya University in 1992 – more than a decade earlier than Uzdensky’s work [502]. They anticipated (but did not actually solve) the magnetosphere structure in Fig. 14.23 and then considered how accretion inflow would take place along the field lines and into the black hole. This was essentially the first paper to consider MDAF-like accretion inflows.

One of Hirotani *et al.*'s most important conclusions was that the energy and angular momentum of the ingoing plasma depended much more on the shape of the magnetic field near the black hole than on its initial conditions near the inner edge of the disk. In particular, the plasma's "energy at infinity" actually can become negative before it enters the black hole. That is, the plasma can give up *more* than its ρc^2 rest mass energy to an outgoing Poynting MHD wind! When this plasma accretes into the hole, *it will spin the hole down*. This is called the magnetic Penrose mechanism and is similar to the classical particle Penrose mechanism discussed in Section 7.5.1.3.

There are a few final points to note. First, we shall discuss the case when there is an asymptotic uniform magnetic field in the fully rotating case. Just as Tomimatsu and Takahashi obtained (for the non-rotating case) open field lines that extended along the polar axis (Fig. 14.20), we therefore also expect field lines *penetrating* the black hole polar cap to occur for the more complex rotating case (e.g., Fig. 14.23). Such (Type 6) field lines will have both ingoing and outgoing light surfaces. And, because this portion of the magnetosphere is isolated from the disk, the source of plasma in the winds *must* be either particle creation or plasma drift across the field lines (due to a small but finite resistivity in the plasma).

Secondly, note that, while the open field lines threading the hole of course *directly* extract energy from the black hole and drive a jet (through the BZ process), even the closed field lines that connect the black hole and the disk also can *indirectly* extract black spin and drive a wind or jet. The angular momentum deposited in the disk by the black hole (using Type 5 field lines) must go somewhere – either transferred further *outward* in the disk by the MRI turbulence, or transferred *upward and outward* by the open (Type 3) field lines. Of course, these open lines also remove angular momentum accreted by the disk itself, so the angular momentum conservation process is complex. However, as the black hole spin increases ($j \rightarrow 1$) the inner edge of the disk will be dominated by angular momentum deposited there by the black hole, not by accretion. So jets driven by Type 3 field lines near the disk inner edge will definitely be black-hole powered, not accretion-powered.

Thirdly, and finally, it is important to note the expected direction of matter flow along the Type 3 and Type 5 field lines. Matter will initially accrete inward inside the disk due to the MRI process, of course, but some of it eventually will be expelled upward along the magnetosphere field lines. The open (Type 3) field lines will drive an outgoing MHD wind, and along the closed (Type 5) field lines will flow an ingoing MHD wind into the black hole.

14.4.3 Structure of a Rotating Black Hole Magnetosphere with an Accretion Disk that Reaches into the Ergosphere

We now are ready to tackle the more difficult problem of the accretion disk reaching into the ergosphere of a rotating black hole and accreting toward the horizon. That is, the black hole torque is transferred by Type 7 and 9 field lines rather than by Types

3 and 5; and Types 7 and 9 may be physically connected into a single field line. This case is more appropriate for the innermost structure of sources with a high accretion rate (Shakura–Sunyaev disks and some ADAFs). Most of the studies in this area use MHD, taking into account how the plasma flow shapes the magnetic field as well as how the magnetic field affects the plasma flow into the black hole. Strangely, this more complex case was tackled by pioneering studies one or two decades *earlier* than the simpler case where the disk did not enter the ergosphere and not that long after the first force-free papers published by Blandford and Znajek.

14.4.3.1 Early Models

The first investigator to do so was Sterl Phinney in his 1983 Ph.D. thesis at the University of Cambridge. This work established the basic physics and flow structure of the system, including the importance of all three MHD critical points (slow, Alfvén, and fast) in the ingoing wind. The problem was further developed in Punsly and Coroniti’s 1990 paper [495], introduced earlier, which worked out the structure and evolution of the magnetic field and current system in the inflow and which demonstrated the importance of the ergospheric disk ($r_H < r < r_S$) in driving an outflowing MHD jet (with a Type 7 field line). The models in these early works are rather specific and elaborate. They assume that the plasma in the ingoing and outgoing winds is an $e^+ e^-$ pair one that is created by high-energy γ -ray photons from the hot accretion disk corona. The ingoing pair wind can intersect the disk or hole itself, while the outgoing pair wind escapes to infinity. This differs from more recent MHD accretion simulations (Section 15.1.4), which assume that the outflow contains $p^+ e^-$ matter from the accretion inflow. At the present time, however, both scenarios are considered plausible.

Most of the aspects of Punsly and Coroniti’s model are shown in Fig. 14.26, taken from their paper. Note the parabolic shape of the poloidal field and how it intersects the disk midplane. (There also is a strong toroidal field component that is not shown.) As in the pulsar and Blandford–Znajek force-free models, when pair production occurs in the magnetosphere the same field line can carry both an ingoing and outgoing wind (a Type 7/9 field line). The particle creation zone is a stagnation region from which the inflow and outflow emanate. This again differs from more recent simulations, in which the accretion inflow of baryonic matter into the black hole occurs along Type 9 field lines, while the wind outflow occurs along Type 7 or even Type 3. An interesting aspect of Punsly and Coroniti’s model, shown in the lower left of Fig. 14.26, is the prediction that magnetic islands can develop from the disk inflow very near the black hole horizon. The poloidal field is expected to pinch off near the equator and reconnect at an “X-point”. This leaves an isolated current ring (“O-point”), which shrinks in size, further reconnects, and accretes into the black hole.

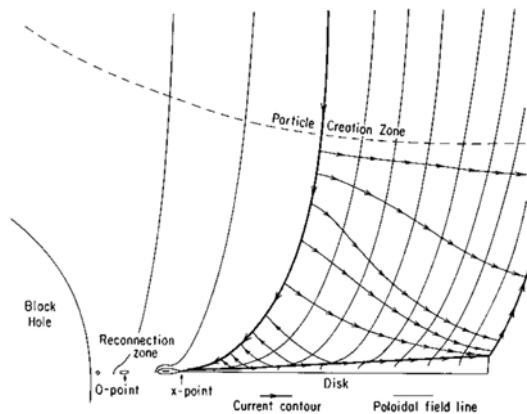


Fig. 14.26: Picture of current flow and magnetic structure of an ergospheric wind around a maximal Kerr black hole (at lower left) with a thin accretion disk (along the equator). Thick arrows show the current flow, while thin lines show the magnetic field. Formally, for $j = 1$, the ISCO should be at the horizon (in Boyer–Lindquist coordinates), but the inner edge of this disk breaks up into X-points, which reconnect into O-points, that separately accrete into the hole. Note the stagnation point where particles are created for both the ingoing and outgoing wind. Reproduced from Fig. 8 of [495], by permission of the AAS.

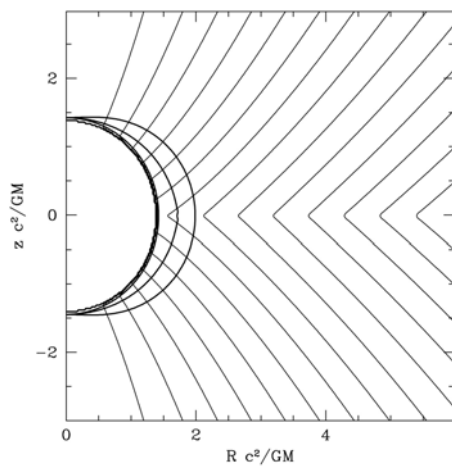


Fig. 14.27: Steady-state GRMHD simulation of a Kerr ($j = 0.9$) black hole with a Keplerian disk on the equator. Also shown are the horizon (thick semi-circle), inner light surface just outside the horizon (see also Fig. 14.17), and the ergosphere boundary. Note the parabolic lines like those in Fig. 14.26 and the monopole structure of the magnetosphere near the horizon (as in other examples in this chapter). Reproduced from Fig. 8 of [503], by permission of J. Wiley and Sons, Ltd.

14.4.3.2 Simulations

Numerical simulations of this process have confirmed most of the aspects of this magnetosphere model. The stationary magnetic structure was studied in 2007 by Jon McKinney and Ramesh Narayan at Harvard (Fig. 14.27) using force-free simulations [503]. Note the parabolic structure of the magnetic field that intersects the midplane disk, and note the split monopole-like structure for the magnetic field that intersects the horizon at higher latitudes. This simulation has features of both the Punsly–Coroniti and Blandford–Znajek/Uzdensky/Garofalo magnetic field solutions.

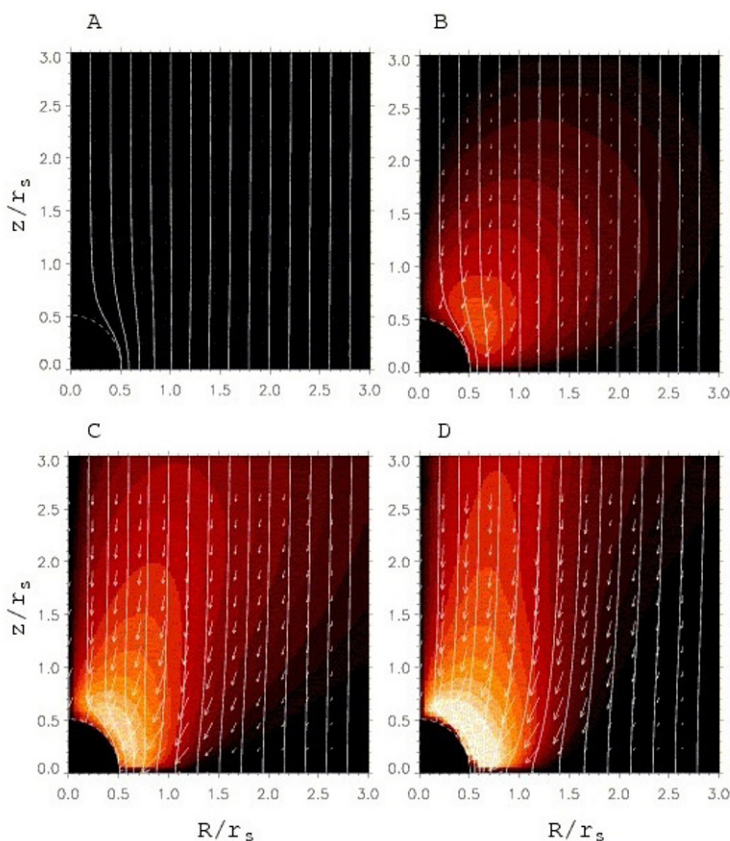


Fig. 14.28: GRMHD simulation of the development of an ergospheric wind, and BZ jet, from an extreme ($j = 0.99995$) Kerr hole with a tenuous plasma filling the computational region. (No accretion disk is involved, making this simulation more like one with an ADAF inflow.) Coordinates are labeled in $r_s = 2r_g$, and the color shows the logarithm of the toroidal magnetic field strength. Arrows show the plasma flow velocity vectors. The simulation times are: (A) $t = 0$, (B) $t = 2\tau_g$, (C) $t = 8\tau_g$, and (D) $t = 14\tau_g$. The details of the structure at the final time step are shown in Fig. 14.29(left). From Fig. 1 of reference [504]. Reprinted with permission from AAAS.

The ability of the ergosphere to drive a strong wind outflow was confirmed by Koide, Shibata, Kudoh, and the author in 2002 [504]. They followed the evolution of plasma trapped in an initial Wald-type magnetic field around a very rapidly rotating black hole ($j = 0.99995$; Fig. 14.28) for a total simulation time of $t = 14 \tau_g$. The plasma accreted toward the hole, developed a magnetic field structure similar to Punsky and Coroniti's, and ejected a Poynting jet similar to that in previously discussed simulations (here and Sections 14.1 and 14.3).³ Of particular interest is an equatorial ring of material inside the ergosphere (Fig. 14.29(left)), which developed *negative* energy at infinity. (See Section 7.5.1.3.) As predicted by Hirotani and his colleagues, plasma in the ergosphere can transfer so much energy to an outgoing MHD wind (arrows) that it can give up *more* than its rest mass! Fig. 14.29(right) shows a 3-D view that illustrates the toroidal, as well as the poloidal, component of the magnetic field.

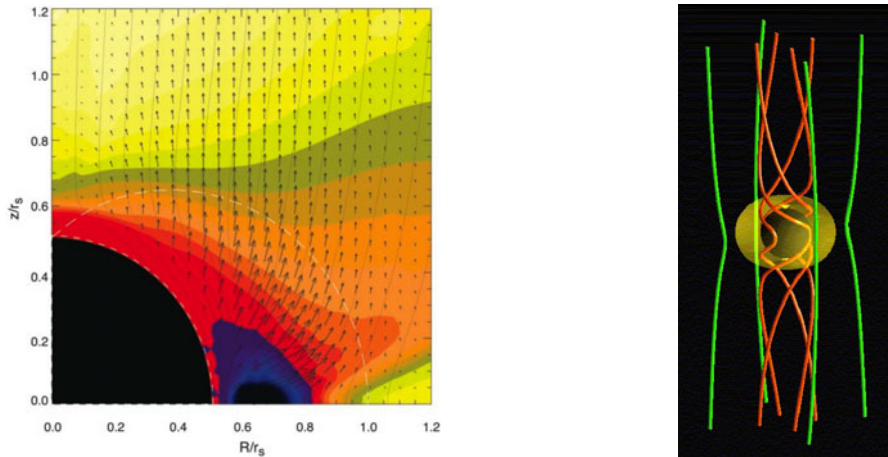


Fig. 14.29: Two views of the final state in Fig. 14.28D. *Left:* Color plot of the energy-density-at-infinity for the plasma near the black hole; green/yellow is positive (unbound), red/purple are approximately zero (rest mass bound to the black hole), and the blue/black torus has negative energy (will decrease the black hole's reducible mass/spin when accreted). Here arrows show Poynting flux density vectors. *Right:* Magnetic flux tube plot of the same information in Fig. 14.28D showing the Poynting jet produced inside the ergosphere. From Figs. 3 and 2 of reference [504]. Reprinted with permission from AAAS.

Finally, in 2005 Serguei Komissarov performed general relativistic *MHD* simulations of ergospheric winds using horizon-penetrating Kerr–Schild coordinates [505]. This allows him to study the flow up to, and through, the black hole horizon. Again, like Koide *et al.*, Komissarov started with a Wald-like magnetic field with plasma trapped in the field lines and, again at early times ($t = 6 \tau_g$) a negative en-

³ Note that this simulation was done in Boyer–Lindquist coordinates, rather than the Kerr–Schild ones of McKinney and others, so magnetic field lines or plasma will never be dragged across the horizon in a finite amount of simulation time.

ergy ring developed deep inside the ergosphere (red color). However, at much later times ($t = 60 \tau_g$) that ring disappeared down the black hole and did not re-form. His final steady state of the simulation is shown in Fig. 14.30(left). The magnetic Penrose process, therefore, may be a transient effect and may not occur in the steady state. Another point to note are the development of magnetic islands as predicted by Punsly and Coroniti. The O-point current rings repeatedly form and disappear down the black hole in Komissarov's simulation.⁴ Finally, note the dotted line, which is the Alfvén (or “light”) surface of the *ingoing* MHD wind. Compare this with the ingoing light surface in the force-free (matter-free) case in Fig. 14.17.

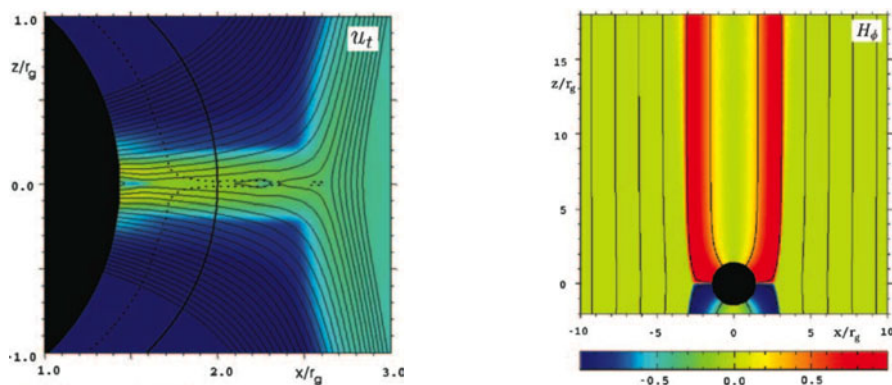


Fig. 14.30: Disappearance of the negative energy torus and ergospheric wind in the steady state for a $j = 0.9$ Kerr hole. Coordinates are labeled in units of r_g . *Left:* Color plot of specific energy-at-infinity at late times ($t = 60 \tau_g$). An ergospheric wind and negative energy torus do, indeed, develop at early times ($t \sim 6 \tau_g$), but then the torus is accreted and is not formed again. While the accreting matter continues to form X- and O-point accretion packets (as in Fig. 14.26), the magnetic flux of the steady state outgoing wind appears to penetrate the horizon only, forming a Garofalo-like magnetosphere (Fig. 14.24). *Right:* Color plot of the toroidal magnetic field strength on a somewhat larger scale, at $t = 60 \tau_g$; a steady BZ jet has formed, as in Fig. 14.16. Reproduced from Figs. 2b and 4c of [505], by permission of J. Wiley and Sons, Ltd.

Fig. 14.30(right) shows the steady state of Komissarov's outgoing wind on an intermediate scale; this can be compared with the outgoing Poynting wind in his force-free simulation in Fig. 14.16. The surprising result is that the outflow is still primarily Poynting flux, even when MHD simulations are used. That is, *the wind/jet ejected from the ergospheric region near the black hole horizon is likely to be a highly relativistic outflow.*

⁴ Note that this simulation, as most others, assumed *ideal* MHD. Therefore, *no* reconnection should occur and no closed current rings should form. The fact that they do is solely because of *artificial* numerical resistivity. Artificial resistivity also is responsible for making the MRI work in numerical simulations and, as in that case, these results should be viewed with some caution. The true role of actual resistivity in black hole plasmas is largely unexplored at the time of this writing.

14.4.4 Relativistic Plasma Outflow from a Pure Split Monopole Magnetosphere

In one of those rare occurrences in science, in 2004 within only a few weeks of each other two investigators on different sides of the globe independently performed the same numerical experiment, got the same exciting result, and submitted those results for publication. Both Serguei Komissarov in the UK [360] and Shinji Koide in Japan [506] studied the case of a rapidly rotating black hole with *radial* magnetic field lines (Type 6, the original Blandford–Znajek model), but with the magnetosphere filled with a tenuous plasma.

Koide’s simulation followed the evolution of this system to $t = 21 \tau_g$, Komissarov’s to $t = 170 \tau_g$. But the results were the same. Unlike the ergosphere disk cases, a very strong relativistic ($\gamma \geq 2$) outflow developed. These were the first demonstrations that a black hole can drive a relativistic wind with a speed similar to that observed in macro- and microquasars. The outflow initially is radial (Fig. 14.31), but it is driven by the same coiled field lines that are seen to collimate such outflow into jets far from the hole in many models and simulations.

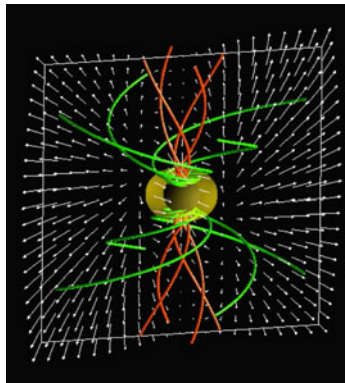


Fig. 14.31: Magnetic flux tube plot at $t = 21.4 \tau_g$ of an MHD simulation of a $j = 0.99995$ extreme Kerr black hole with initially radial field lines filled with tenuous plasma. The Lorentz factor of the outflow is $\gamma \sim 2$ at this early time but is expected to be higher as it is further accelerated by the rotating magnetic field. Reproduced from Fig. 2 of [506], by permission of the AAS.

The results in this section indicate that black holes can drive jets in many ways: transfer angular momentum to a disk and drive a disk wind; drive a transient wind from the ergospheric plasma (perhaps intermittently); drive a relativistic outflow if the magnetosphere is radial; or drive a strong Poynting (electromagnetic) outflow if parabolic field lines thread the horizon. Which mechanism is used will depend on the configuration of the black hole magnetosphere and, consequently, on the type of accretion flow at the time. The same black hole may drive several different types of jet in its lifetime – perhaps even more than one at the same time.

14.5 Jet Launching from Accretion Disks Only

Accretion disks not only can be a source of magnetic field with which to build a black hole magnetosphere, they also can launch plasma outflows on their own (see Fig. 14.22). This section discusses launching mechanisms that involve only the accretion disk itself.

Launching a jet from an accretion disk actually is a difficult proposition, particularly if the disk is thin. Before a jet can be accelerated efficiently from a disk to high speed by MHD forces, the material first must be lifted out of the deep (tidal) gravitational potential well near the disk midplane. The (Newtonian) gravitational potential is $\psi = -GM/(R^2 + Z^2)^{1/2}$, so the tidal force (per unit mass) that holds the accretion disk material near the equatorial plane is

$$-|\nabla\psi|_Z = -\frac{GMZ}{(R^2 + Z^2)^{3/2}} \approx -\frac{GMZ}{R^3}$$

For $Z \ll R$, therefore, it is very difficult to lift the disk material out of the tidal potential, because the restoring force at a given radius R actually *increases* with distance Z above the disk midplane. Only when we have $Z \gtrsim R$ do we get a fall-off in vertical gravitational force with height ($\propto 1/Z^2$). The most important aspect of launching a jet from an accretion disk, therefore, is simply lifting material in the disk upward to a height $Z \sim R$ where it then can be accelerated efficiently by electromagnetic/MHD processes.

Gordon Ogilvie and Mario Livio were some of the first authors to point out this problem [519], in 1998. To solve it, these authors suggested that it was primarily plasma pressure (such as occurs in a hot corona) that lifts the jet material initially out of the tidal potential well. One can think of this as a solution that uses dynamical forces (sound or slow MHD waves) to do the job. However, there are at least two other types of solutions to the launching problem: a magnetocentrifugal one, which uses a large torsional Alfvén wave to do the initial lifting, and a magnetic pressure one, which is essentially a fast MHD wave solution. That is, each possible launching mechanism corresponds to a different characteristic wave speed in the magnetized fluid plasma.

14.5.1 Magnetocentrifugal (Alfvén) Launching

The first disk jet-launching process to be suggested was the magnetocentrifugal one, by Roger Blandford and David Payne (both at Caltech at the time) [507]. This was in conjunction with their theoretical work on the acceleration and collimation of jets, which is discussed in Chapter 15. The basic magnetocentrifugal process involves strong rotating magnetic field lines that are anchored in the accretion disk. Under certain conditions, this scenario can unbind the plasma from the disk and “fling” it outward centrifugally, away and upward from the accretion disk.

14.5.1.1 The Effective Gravitational Potential for Plasma in a Strong, Rotating Magnetic Field

Imagine that we have an infinitely thin, axisymmetric accretion disk (the internal structure of which is of no consequence here), and that a strong magnetic field protrudes from the surface of that disk. At the disk surface ($R, \phi, Z = 0$) the field lines are assumed to be poloidal only but can make an angle θ_f with respect to the vertical Z axis. That is, $\mathbf{B}_d = (B_R, 0, B_Z)$. Because the field is anchored in the rotating, Keplerian disk, these field lines rotate. And, because the field strength is high, any plasma trapped in the field lines does not bend the field much at all: the plasma acts like a lightweight jewelry bead on a stiff wire of magnetic field.

Now consider a single magnetic field line that is anchored at the point $(R_0, 0, 0)$. (The azimuthal coordinate ϕ is set to zero here as it is representative of any other angle ϕ in our axisymmetric disk.) The rotational velocity of this field line is

$$\Omega_f = \Omega_K(R_0) = \left(\frac{GM}{R_0^3} \right)^{1/2}$$

and above and beyond that point that same field line will have that same rotational velocity everywhere, not just at R_0 and $Z = 0$. The effective potential of this gravitating, rotational system is the Roche one (equation (11.55)) with $M_{*1} = M$ and $M_{*2} = 0$ (for zero-mass plasma test particles)

$$\varphi(R, Z) = -\frac{1}{2} R^2 \Omega_f^2 - \frac{GM}{(R^2 + Z^2)^{1/2}} \quad (14.31)$$

Note that a rotational potential energy term, with a constant Ω_f , is included in the effective potential. This is because the plasma is constrained to live on the rigidly-rotating magnetic field line, just as it is constrained to live in the gravitational potential of the black hole. And the rotational potential is negative because it takes energy to move the plasma “bead” on the rotating magnetic field “wire” from large R toward $R = 0$, where the rotational part of the potential vanishes.

Fig. 14.32 from Blandford and Payne’s paper shows some equipotential surfaces ($\varphi = \text{constant}$) in the (R, Z) -plane. the thick solid line is the surface of marginal stability ($\varphi = -\frac{3}{2} GM/R_0$), where the effective potential is equal to that at the anchor point. For field lines in the “stable” region, plasma will fall along the field line back onto the disk. On the other hand, for field lines in the “unstable” region, plasma will either accrete inward toward the Z axis at $R = 0$ or move outward toward $R \rightarrow \infty$ in a centrifugally-driven wind. For the latter outflow to take place, Blandford and Payne found that such *acceleration occurred only if the magnetic field lines lay at an angle greater than 30° from the rotation axis.*

A model using BP’s magnetocentrifugal launching, plus their acceleration and collimation theory, is shown in Fig. 15.2 in the next chapter.

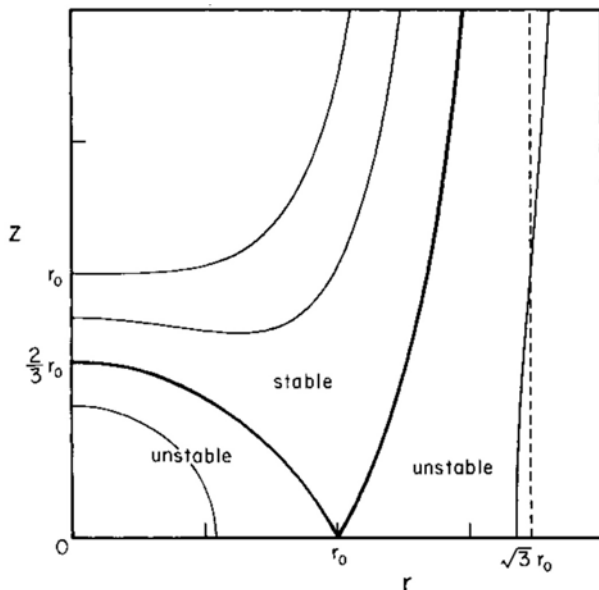


Fig. 14.32: Contours of Blandford and Payne's effective potential (eq. (14.31)) of a plasma with a constant rotation rate $\Omega_f = \Omega_K(r_0)$, including both gravity and centrifugal effects. This diagram shows the Roche potential for a massless ($q = 0$) particle at $r = r_0$, but in the *poloidal* plane (R – Z), rather than in the equatorial plane of Fig. 11.5. The region at left (corresponding to that inside L1) is unstable to infall, while the region at right (corresponding to that outside L2) is unstable to outflow. The portion for $Z > 0$ lies above the orbital plane and usually is not shown in Roche potential plots. Reproduced from Fig. 1 of [507], by permission of J. Wiley and Sons, Ltd.

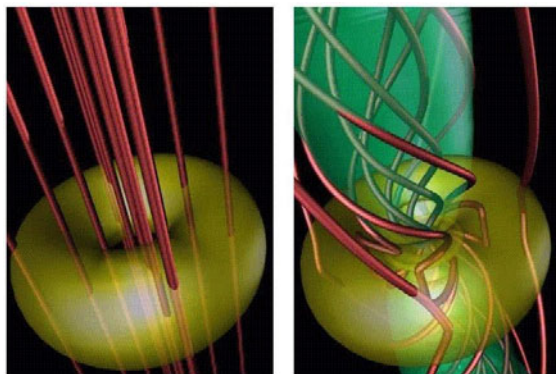


Fig. 14.33: Initial (left) and final (right) conditions for a recent Uchida and Shibata-like simulation [509] of the sweeping pinch mechanism. The geometrically thick Keplerian torus orbits a compact object lying at its center; red lines show magnetic flux tubes. The jet (blue-green plasma) is accelerated by the upward magnetic pressure gradient and collimated by the pinch force. Based on Fig. 1a of [508], kindly provided by the late Professor Uchida. From Fig. 5 of reference [510]. Reprinted with permission from AAAS.

14.5.2 Magnetic Pressure (Fast-Mode) and Gas Pressure (Slow-Mode) Launching

Since the BP paper was published, several groups have simulated MHD winds and, in particular, studied full two-dimensional MHD simulation of winds without the self-similar assumption. The first investigators to look at the MHD of winds from disks were the late Yutaka Uchida (then at Tokyo Astronomical Observatory) and Kazunari Shibata (then at Aichi University of Education in Japan) in 1985 [509], only a few years after Blandford and Payne’s initial paper. They performed time-dependent simulations of jet material being launched from thick accretion disks (tori) that showed the basic features of a “sweeping pinch” mechanism of jet formation. A more modern version of these simulation is shown in Fig. 14.33. While related to the BP concept, the initial acceleration from the disk is *not* magneto-centrifugal but rather caused by the lifting and pinching action of a large toroidal magnetic field that builds up near the torus inner edge and above. The uncoiling of the toroidal field ($-dB_\phi^2/dZ$) upward produces a vertical acceleration, while the $B_\phi^2/4\pi R$ hoop stress squeezes the plasma toward the axis. That is, the mechanism is more “spring” action than “fling”. Note that the lifting action also is aided initially by the large pressure in the thick accretion torus, which creates the distended, nearly spherical, plasma distribution. This is a warm, not cold, MHD outflow.

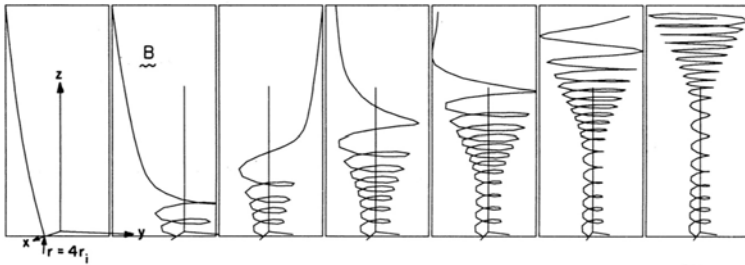


Fig. 14.34: Three-dimensional view of a magnetic field line in the Ustyugova *et al.* simulation of jet outflow from a thin accretion disk. The initial launching of the jet is actually performed by high gas pressure, which lifts the material upward and out of the disk potential well (second and third panels). The magnetic pressure gradient is not effective in accelerating the jet upward until the fourth or fifth panel. The effect of collimation by the pinch effect is seen in the sixth and seventh panels. Reproduced from Fig. 3 of [511], by permission of the AAS.

The first investigators to perform simulations using an (essentially infinitely) thin accretion disk boundary on the equator were a Russian group, led by Galina Ustyugova of the Keldysh Institute, in 1995 [511]. They, indeed, did find jet outflows from their accretion disks, but they also found that they had to assume a fairly hot plasma in the disk $(c_s/V_K)^2 = 0.90$! That is, the coronal material itself that was launched into the jet was essentially a geometrically thick disk. While it was the magnetic field that ultimately accelerated and collimated the outflow, it was the gas pressure

that initially lifted the material out of the deep black hole tidal potential and allowed the MHD mechanisms to work. (See Fig. 14.34.) In fact, for any simulation where the disk plasma has a scale height comparable to its cylindrical radius, it is gas pressure that actually accomplishes the initial jet launching.

14.5.3 The Magnetic Tower Mechanism: Fast-Mode Launching with Closed Field Lines

Most jet-launching mechanisms that we have just discussed assume that the rotating poloidal magnetic field lines that help launch the jet outflow are open, *i.e.*, Types 3, 6, or 7. By analogy, the sun also generates open field lines and does drive an outflow along them. However, the sun's open field lines usually are created from closed loops in coronal mass ejections. Is there an analogous process that can launch jets in which open Type 3, 6, and 7 field lines are created from closed Types 2 and 5?

Yes, the process is called the *magnetic tower*, and it is depicted schematically in Fig. 14.35. In this figure we see a Type 2 field line open up due to differential rotation, rise vertically and be squeezed toward the axis by external pressure, and form a tall, slender tower. Close to the black hole, the field structure looks similar to the open field lines in Figs. 14.21 to 14.23, anchored in the disk and rising to great heights. However, instead of being anchored at “infinity”, the poloidal field lines loop back and are anchored at the other end in the same disk but at a much larger radius, where the rotation is much slower.

14.5.3.1 Basic Physics of Magnetic Towers

The magnetic tower process was first proposed by Donald Lynden-Bell [512] of the University of Cambridge, UK. We will use both his analysis and that of Dmitri Uzdensky and Andrew MacFadyen (of the Institute for Advanced Study in New Jersey) in our discussion below [513]. As with our previous jet launching discussions in this chapter, the magnetic tower requires only a poloidal magnetic field and rotation – specifically, *differential* rotation between the inner and outer anchor points of a closed field line (with the inner anchor point rotating faster).

Let us do a simple analysis of Fig. 14.35d. Poloidal magnetic field lines are generally labeled with their flux Ψ (see equation (9.130)). So, for a magnetic tower with radius R_0 and poloidal magnetic field B_0 , we have

$$B_0 = \frac{\Psi_0}{R_0^2}$$

Here Ψ_0 is an input parameter of the magnetic tower problem. Now, the tower in Fig. 14.35d is confined along its axis by the external pressure p_{ext} , which is a second parameter of the problem, so in pressure equilibrium we must have

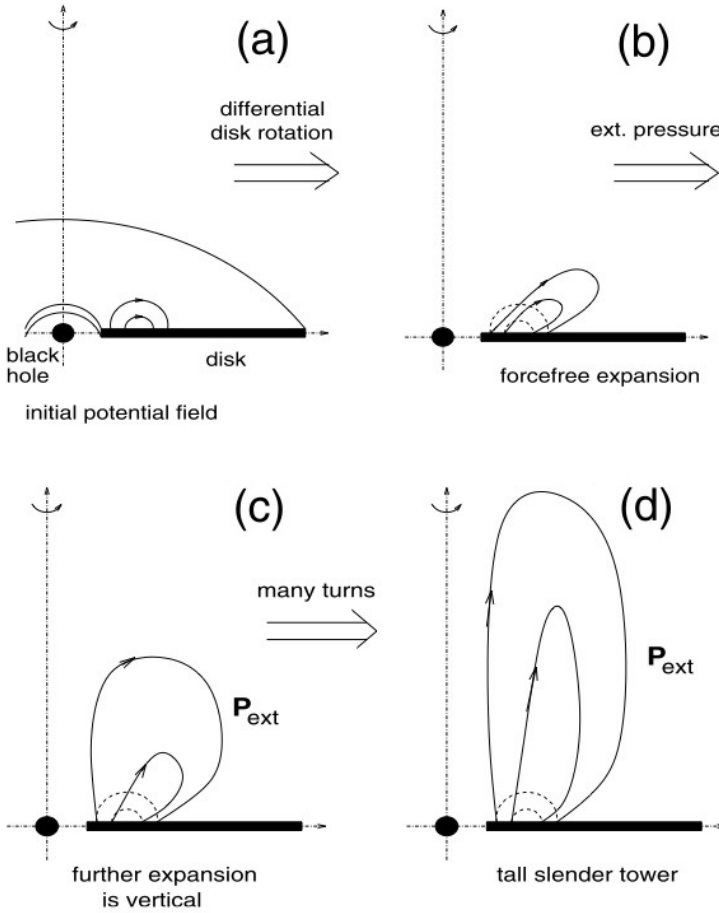


Fig. 14.35: Development of a magnetic tower jet. (a) Initial conditions in the disk (tilted here for easy viewing); a magnetic flux tube connects two different radii in the disk. (b) Early inflation of the flux loop, showing before (dotted) and after (solid) magnetic states. (c) Eventual confinement of the tower in R by external gas pressure; (d) After many disk rotations the initial loops have evolved into a rising magnetic tower. Adapted from Fig. 1 of [513], by permission of the AAS.

$$B_0 \approx (8\pi p_{\text{ext}})^{1/2}$$

which tells us that, in terms of Ψ_0 and p_{ext} , the radius of the tower must be

$$R_0 \approx \left(\frac{\Psi_0^2}{8\pi p_{\text{ext}}} \right)^{1/4} \quad (14.32)$$

Next we shall estimate the vertical growth rate of the tower. It is important to realize that, even though not shown in Fig. 14.35d, the vertical field lines near the

disk inner edge are being continually twisted relative to the outer (more slowly rotating) field lines. So the entire field structure is helical, not simply poloidal. If we let the third parameter of the problem $\Delta\Omega$ be the difference in angular velocity between inner and outer field anchor points, then after a time t the total toroidal flux inside the rotating magnetic tower $\Psi_\phi(t) \equiv B_\phi(t) Z_{\text{top}}(t) R_0$ will be

$$\Psi_\phi(t) = \Psi_0 \Delta\Omega t \quad (14.33)$$

where $Z_{\text{top}}(t)$ is the height of the tower at time t . But, for the tower to remain in force balance, we expect the toroidal and poloidal magnetic fields to be comparable

$$B_\phi(t) = B_0$$

that is, B_ϕ is largely independent of t . From equation (14.33), then, we find that the top of the tower rises linearly with time:

$$Z_{\text{top}}(t) \approx R_0 \Delta\Omega t \quad (14.34)$$

or $V_{\text{top}} \approx R_0 \Delta\Omega$. For a Keplerian disk, this means that $V_j \approx V_K$, similar to what is observed for actual jets. This process is also called by the alternate name the “spring” jet mechanism, because it creates collimated outflow using a continually rotating coil of magnetic field that lifts material by uncoiling upward.

Lynden-Bell performed a more rigorous analysis and obtained

$$Z_{\text{top}}(t) = 0.306 R_0 \Delta\Omega t$$

He also determined that the external pressure p_{ext} is absolutely essential to forming a slender tower. If $p_{\text{ext}} = 0$, the uncoiling field would form a 60° angle with the rotation axis (see Fig. 14.35b) and continue to splay outward forever. However, if p_{ext} were very small but finite, R_0 would continue to increase outward until $B_0^2/8\pi \approx p_{\text{ext}}$, again confining the tower and producing a jet.

14.5.3.2 Simulations of the Magnetic Tower

The first true magnetic tower simulations were performed around the time Lynden-Bell was writing his paper by D. Meier, Samantha Edgington of Caltech, and Patrick Godon of JPL, using an MHD code developed by Kevin Lind, David Payne, and Meier [514]. Like Uchida and Shibata’s work a decade earlier, these simulations also were an attempt to simulate the Blandford–Payne centrifugal MHD wind mechanism (see Section 15.1.1). However, unlike Fig. 14.33, the initial poloidal magnetic field was placed at a substantial angle ($\sim 54^\circ$) from the rotation axis, not parallel to it (see Fig. 14.36(upper left)). The expectation was that the wind would flow out at the $\sim 54^\circ$ angle and collimate in a manner similar to Uchida and Shibata’s.

The surprise, however, was that the jet rose from the *center* of the accretion disk, carrying its magnetic field with it (see Fig. 14.36(lower left)). While the field lines

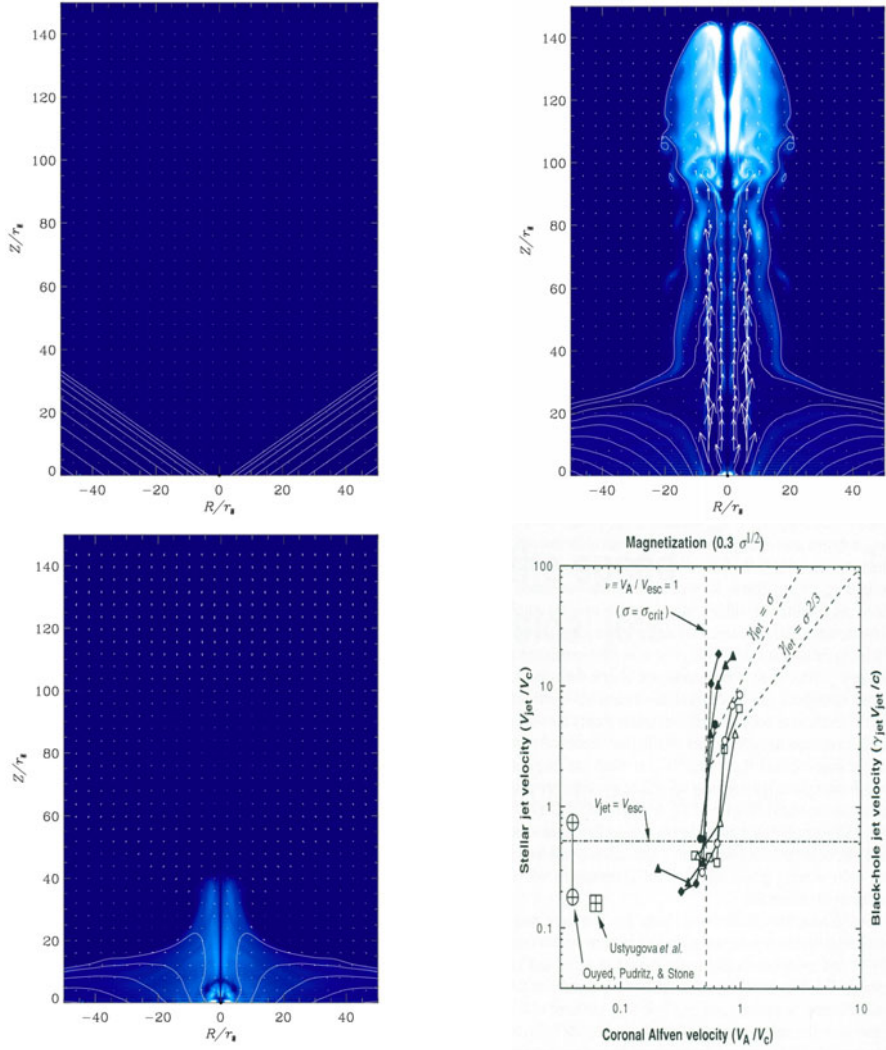


Fig. 14.36: First simulations of the magnetic tower process and demonstration of a “magnetic switch” effect. *Upper left:* Initial magnetic state for both simulations shown, with an infinitely thin equatorial Keplerian disk and tenuous plasma in hydrostatic equilibrium above it. Initial magnetic field lines lie at a 54° angle from the rotation axis and are poloidal only (coordinates in units of r_g). *Lower left:* MHD simulation of a disk corona with an Alfvén velocity *below* the escape speed ($V_A/V_{\text{esc}} = 0.38$), shown at a time $t = 650 \tau_g$. Color shows the strength of the toroidal component of the magnetic field. Note how poloidal field lines rise in a tower fashion, carrying plasma and increased toroidal field. *Upper right:* Similar simulation, at the same time, but with a coronal Alfvén velocity $\sqrt{10}$ higher ($V_A/V_{\text{esc}} = 1.21$) and, therefore, *above* the escape speed. Arrows show the local fluid flow speed. Note the well-formed jet head with much faster speed. *Lower right:* Plot of internal jet speed for many simulations with different V_A/V_{esc} (horizontal axis) and different initial magnetic polar angle (open triangles: 8° ; open squares: 24° ; open circles: 38° ; filled triangles: 54° ; filled diamonds: 68° ; and filled circles: 83°). Simulations shown are first filled triangle on left (lower left panel) and third filled triangle from right (upper right panel), with a ratio of jet speeds of ~ 30 . Adapted by permission from Macmillan Publishers Ltd: Figs. 2 and 3 of reference [514], copyright (1997), and from the author’s own results.

were formally open, they completely enclosed the jet, with both upward twisted field close to the rotation axis and returning field lines outside the tower. We did not recognize at the time that this simulation was an example of Lynden-Bell’s magnetic tower.⁵ Moreover, we noticed a dramatic increase in jet speed when V_A became greater than the escape speed V_{esc} from the center of the disk (Fig. 14.36(upper and lower right)). We called this sudden change in speed the “magnetic switch”.

The concept of the magnetic switch was revisited in 1999 by D. Meier [46]. It was found to exist not only when magnetic field strength alone was increased but also when the magnetic field angular velocity Ω_f alone was increased. Field angular speeds above Keplerian could occur, for example, if a rapidly rotating black hole were coupled to an accretion disk with a magnetic “bridge”, as in Figs. 14.21 and 14.23. A more general magnetic switch criterion then emerged: a magnetic tower jet speed should dramatically increase when the MHD power input into the jet

$$L_{\text{MHD}} = B_0^2 R_0^3 \Delta\Omega$$

exceeded a critical luminosity, given by the escape energy divided by the free-fall time

$$L_{\text{crit}} = \frac{E_{\text{esc0}}}{\tau_{\text{ff0}}} = 4\pi \rho_0 R_0^2 \left(\frac{GM}{R_0} \right)^{3/2}$$

or

$$\Delta\Omega V_{A0}^2 > \Omega_K V_{\text{esc0}}^2 \quad (14.35)$$

Near the inner edge of a Keplerian disk (with $\Delta\Omega \approx \Omega_K$) this criterion reduces to the original one discovered by us in 1997.

In 1998 Marina Romanova and her colleagues in Russia and at Cornell studied the physics of the magnetic tower in great detail [515]. Particularly telling is Fig. 14.37 from their paper, which shows the opening of a single closed coronal field loop in one of their simulations. The dominant force causing the field line to inflate is, of course, the $\mathbf{J} \times \mathbf{B}$ magnetic force. In fact, it acts in a largely force-free manner here, unlike in our earlier magnetic switch simulations. As predicted by Lynden-Bell, magnetic towers can have most of their output energy in the form of magnetic Poynting flux – at least initially and close to the black hole. In the year 2000 this was demonstrated numerically by the same group, this time led by Galina Ustyugova and now joined by colleagues at Los Alamos National Laboratory, Hui Li and Stirling Colgate [516].

The magnetic tower has been studied by several Japanese groups as well. A particularly interesting study in 2002 by Takahiro Kudoh (NAOJ), Ryoji Matsumoto, and Kazunari Shibata showed that a jet can arise quite locally in an accretion disk, especially if it were geometrically thick (see Fig. 14.38) [517]. Unlike the simulations shown in Fig. 14.36, where the tower rose from near the disk midplane, in this

⁵ The author is grateful to M. Nakamura and P. C. Fragile for pointing out the magnetic tower nature of these simulations.

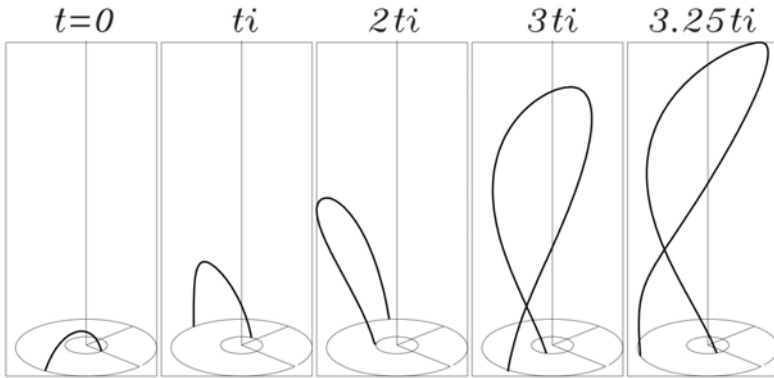


Fig. 14.37: Three-dimensional view of the opening of a field loop in the magnetic tower process. As the inner footpoint rotates somewhat faster than the outer one, the loop is stretched, inflated, and eventually wrapped around the rotation axis, producing a strong toroidal field component. Reproduced from Fig. 6 of [515], by permission of the AAS.

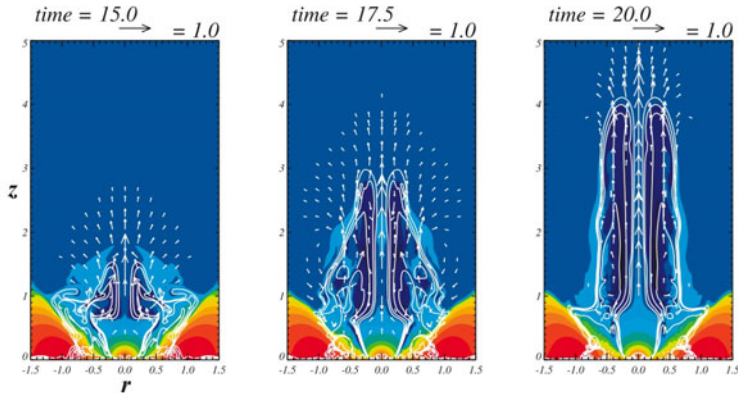


Fig. 14.38: Demonstration of the robustness of the magnetic tower, which can be generated at virtually any part of a magnetized accretion inflow where there is differential rotation, including at high latitudes in a thick disk. Reprinted from Fig. 2a of [517], with permission of the Publications of the ASJ.

case the tower rises from coronal loops that form high on the inner surface of the thick torus. The tidal gravitational forces of the black hole are considerably weaker there than they are near $Z = 0$, making it easier to lift material from the high corona than from the disk midplane.

Finally, Koide, Kudoh, and Shibata demonstrated in 2006 that a magnetic tower can be formed very near a rapidly rotating ($j = 0.99995$) black hole in a magnetic bridge between the hole and its accretion disk [518]. Fig. 14.39 shows the development of a nice magnetic tower at late times. As with other ergospheric winds, this probably will relax into a BZ-type Poynting outflow (Fig. 14.30) after many τ_g .

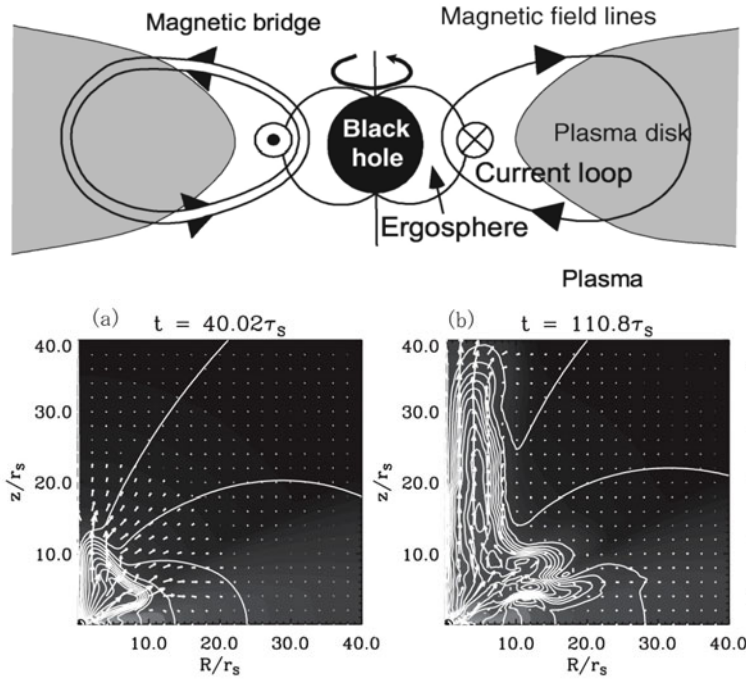


Fig. 14.39: Formation of a magnetic tower by an extreme Kerr ($j = 0.99995$) black hole and disk, where both are connected by a Type 2/4 field line (Fig. 14.18) created by a current loop in the inflow. *Top:* Diagram of the initial conditions, showing field lines connecting disk to ergospheric plasma. *Bottom:* Early (left) and late (right) evolution of the magnetic tower from the ergosphere. Reprinted from Figs. 1 and 11 of [518]; copyright (2006) by the APS.

14.5.3.3 The Magnetic Switch Effect Explained

By recognizing the Meier *et al.* simulations as magnetic tower ones, we now can understand the magnetic switch's basic physical properties and why its behavior is so different from that expected from Blandford and Payne's centrifugally-driven jet launching mechanism (see Section 15.1.1). The standard analysis of the magnetic tower ignores the tidal gravitational forces, but during the initial launching gravity actually is very important.

The behavior of the magnetic tower, and the role the “magnetic switch” plays in it, can be understood as follows:

- As long as the magnetic energy pumped into the rising tower is not enough to unbind the plasma from the tidal potential (*i.e.*, $L_{\text{MHD}} < L_{\text{crit}}$), the magnetic tower will struggle to lift the plasma up from the disk midplane. In the normal thin accretion disk situation ($\Delta\Omega \approx \Omega_K$; $V_A \ll V_{\text{esc}}$) the magnetic tower mechanism will have difficulty producing any kind of vertical, collimated out-flow (Fig. 14.36(lower left)), because $\Delta\Omega V_A^2 \ll \Omega_K V_{\text{esc}}^2$.

- In order for a powerful jet to be produced, largely free of the grip of the black hole's tidal gravitational field, at least one of the following conditions must be met:
 - The disk may become geometrically thick, with the closed field loops originating near $Z \approx R$ where the tidal force is much weaker (Fig. 14.38). This is an example of plasma pressure assisting the magnetic tower.
 - Alternatively, if the disk were thin, the rotation rate of the magnetic field still could exceed a critical value

$$\Delta\Omega > \Omega_{crit} = \Omega_K V_{esc}^2 / V_A^2$$

so that Poynting energy input into the plasma could exceed any gravitational effects in a free-fall time. Fig. 14.39, with a very rapidly rotating black hole and magnetic bridge, is an example of this.

- In order to eject a jet from a compact corona of a thin *Keplerian* disk, the magnetic field needs to be strong enough so that $V_A > V_{esc}$, and the corona must rotate at the disk's Keplerian speed. We then have $\Delta\Omega V_A^2 > \Omega_K V_{esc}^2$, which again imparts more Poynting energy in a free fall time than the escape energy of the plasma. This is the classic magnetic switch (Fig. 14.36).

The magnetic switch of our 1997 and 1999 papers is, therefore, simply a Poynting flux solution to the problem of lifting matter out of the tidal potential that was posed by Ogilvie and Livio in 1998 [519]. In order for powerful jets to be launched from accretion flows near black holes, either magnetic or plasma pressure forces (or both) must be strong enough to lift the jet plasma quickly out of the tidal potential and into the $Z \geq R$ region where the classical magnetic tower mechanism can do its work.

14.5.3.4 Magnetic Towers in Global MRI Simulations

As a final example of the magnetic tower, and as a prelude to the discussion of MHD winds in the next chapter, we shall briefly discuss another paper in the original series series of simulations by Universities of Alberta and of Virginia and Johns Hopkins authors (see Section 12.4.4), this time by Shigenobu Hirose and colleagues [497]. Recall that De Villiers' original simulations (discussed in that section) were fully relativistic, three-dimensional, MHD ones that began with a simple thick plasma torus, with closed poloidal-only loops of magnetic field, orbiting a black hole. The differential rotation then sheared the magnetic field, generated turbulence from the MRI, and caused an accretion flow to ensue. In addition, the simulations also launched a substantial jet outflow, whose strength increased significantly as the black hole was allowed to spin faster [520].

What is not usually emphasized, however, is that these jets (and indeed all jets generated by simulations that begin with closed poloidal field lines), *must be generated by Lynden-Bell's magnetic tower mechanism*. And, in this case, the process that overcomes Ogilvie and Livio's dilemma is not the magnetic switch but rather

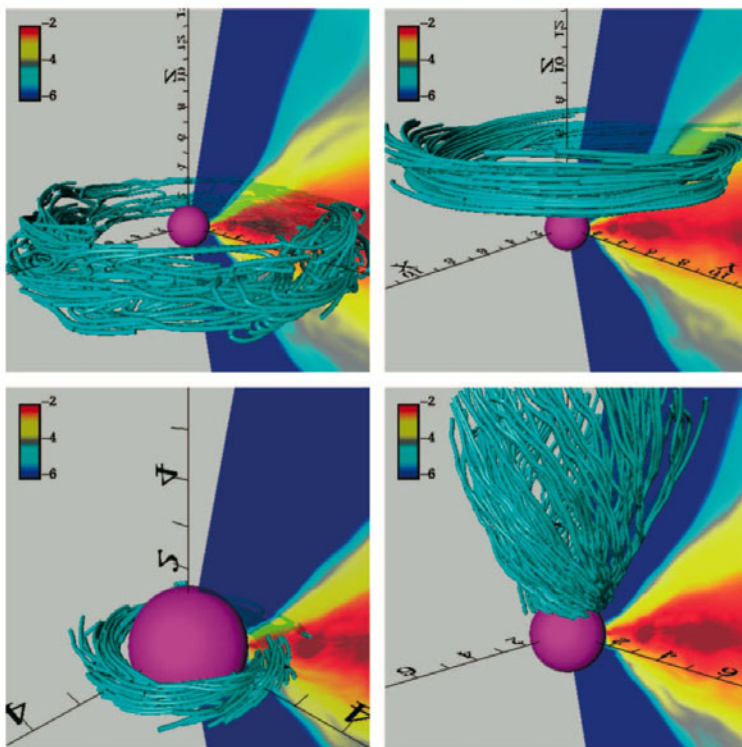


Fig. 14.40: Field lines from four different regions in an accreting $j = 0.9$ Kerr black hole simulation with a radiatively inefficient (ADAF-like) disk. Logarithm of density is shown by the color, with the scale given by the vertical bars. *Upper left:* Near disk midplane, where the MRI creates a tangled field with a strong toroidal component. *Upper right:* In the disk corona, where increased shear creates a mostly-toroidal field. *Lower left:* Near the ergosphere boundary on the equator, where metric shear also enhances the toroidal field. *Lower right:* In the funnel region, where super-Keplerian shear creates a magnetic tower and jet outflow. Reproduced from Fig. 6 of [497], by permission of the AAS.

the strong internal plasma pressure of the thick accretion flow. Fig. 14.40 shows four different sets of magnetic field lines from the Hirose *et al.* paper. Near the disk mid-plane the field is tangled and turbulent (upper left), becoming increasingly toroidal as one approaches the black hole (lower left) or moves to higher disk latitudes (upper right). Note, in particular, that the accreting plasma is in a very geometrically thick inflow, as shown by the color shading. At very high latitudes (lower right) one sees that a rising vertical structure of strong, open helical magnetic field lines has developed from the originally closed ones. Moreover, the process of accretion has attached the tower to the rapidly rotating ($j = 0.9$) black hole. The output of this tower is mainly Poynting flux, in a slender jet that is collimated along the rotation axis. This confirms the identification of this structure as a magnetic tower, driven in this case largely by black hole rotation.

Chapter 15

The Non-Thermal Exhaust System II. Magnetic Winds and Jets

Launching of the MHD wind is only the beginning of the nonthermal exhaust outflow. Just as the thermal exhaust is accelerated away from the injection region by pressure gradients, so the MHD wind also is significantly accelerated, and collimated into a jet, beyond the launching region. Furthermore, just as the thermal wind becomes causally disconnected from the engine beyond a critical (sonic) surface, and subsequently expands away freely and supersonically, so the nonthermal exhaust is accelerated through a series of MHD critical and separatrix surfaces, eventually becoming a freely propagating jet flow beyond the final separatrix surface.

We therefore recognize two distinct regions in the jet:

- The jet nozzle, which accelerates and collimates the jet flow into a narrow, relativistic beam (analogous to the *subsonic* thermal wind structure).
- The ballistic jet itself, which has become causally detached from the engine and no longer influences, or is accelerated by, the latter (analogous to wind *supersonic* structure).

We will discuss each of these in detail in the sections below.

15.1 Magnetized Wind and Jet Theory: The Formation of Jets

While it is obvious now (in light of MRI simulations, for example), that any wind from an accreting disk will carry a non-negligible magnetic field (plasma $\beta_p \leq \alpha^{-1} \sim 100$), it was not so obvious in the late 1970s and early 1980s. And, while a great deal of work had been done on magnetized pulsar winds from the late 1960s through the 1970s, the connection had not yet been made between accretion disk winds and jets in extragalactic radio sources (the only type of jet known at that time).

It was in this environment, then, that two pioneers in the field – Richard Lovelace at Cornell University and Roger Blandford (then at the University of Cambridge) –

suggested (less than one month apart) that jets were launched, and collimated from accretion disks around black holes by magnetohydrodynamic processes.¹

Both of these groups elaborated on this idea over the next several years, each with major papers on their results. A 1982 paper by Blandford and David Payne (both at Caltech at the time) emphasized the ejection and collimation of a magnetized accretion disk wind into a jet flow [507]. Blandford and Payne (or BP for short) is the paper most often cited as the explanation for the production of jets in black hole systems, as well as in protostars and cataclysmic variable systems. Lovelace's group also released two major papers, in 1986–87, on MHD winds from disks [521, 522]. These papers emphasized self-consistent models of the accretion flow and wind launching using a magnetosphere approach that was relativistic and similar to that used for pulsars (see Section 14.1). Extensions of these two groups (Blandford's through Königl at Chicago and Begelman in Boulder; Lovelace's through Ustyugova in Moscow and Contopoulos in Athens) have largely shaped the basic theory of jet acceleration and collimation. In this section we will begin with the Blandford and Payne approach to describe MHD winds and jets and describe the work, particularly by Contopoulos and Bogovalov, that transformed the basic understanding of the underlying physical processes that drive MHD jets. We will end with the Vlahakis and Königl theory of warm, relativistic jets, which can be considered a joint collaboration between descendants of the original two jet theory groups.

15.1.1 *Non-Relativistic, Cold MHD Wind and Jet Theory*

In order to render the first discussions of MHD wind theory more tractable BP made several simplifying assumptions, in particular that the plasma was “cold” (had no appreciable temperature or thermal pressure) and that the flow was non-relativistic. We will relax these assumptions in the following subsections, leading to a quite sophisticated and fairly realistic theory of warm, relativistic MHD winds and jets.

There are a number of additional key points to be made about the BP work that we need to mention:

- Their basic jet-launching mechanism was the magnetocentrifugal one (Section 14.5.1), based on the force balance between gravity and centrifugal effects on the plasma along field lines. This type of mechanism was necessitated because of the cold plasma assumption. Subsequent work by other authors showed that

¹ Lovelace and Blandford actually used the term “electrodynamic” in their papers, rather than “magnetohydrodynamic”. Some investigators like to draw a clear distinction between the two, but the difference really is only one of degree. The equations, in fact, are nearly identical, especially if one compares relativistic MHD and ED. Both sets consider the magnetic *and* electric fields, and both assume the frozen-in condition ($\mathbf{E} = -\mathbf{V} \times \mathbf{B}/c$). The only real difference is the role of plasma forces in the flow. MHD can have significant hydrodynamic forces, while ED has none. As used in pulsar and jet theory, ED is, in fact, simply a special case of relativistic, ideal MHD: the plasma density and pressure vanish. We therefore will *not* distinguish between relativistic MHD and ED here and often will use the two interchangeably.

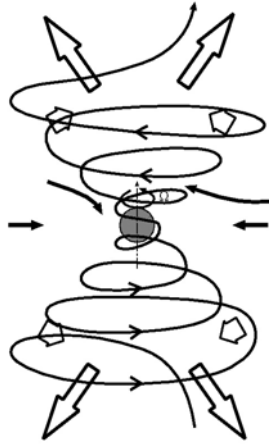


Fig. 15.1: Basic parts of the Blandford–Payne jet acceleration and collimation process: magneto-centrifugal near the compact object (“fling”); magnetic pressure gradient acceleration by decreasing B_ϕ^2 (“spring”; long, white arrows); and magnetic “pinch” to collimate the outflow (short white arrows). From Fig. 3 of reference [510]. Reprinted with permission from AAAS.

this condition could be relaxed if the plasma were hot and created a geometrically thick disk or torus (see Figs. 14.33 and 14.34).

- Their basic jet formation mechanism, however, is very general and can be described as being due to three types of magnetic forces (see Fig. 15.1):
 - The initial outward centrifugal acceleration (“bead on a rotating wire”) described above, sometimes called the “fling” process.
 - Magnetic pressure upward in the Z direction. As more matter is flung outward, it accumulates in the rotating field lines and retards their rotational motion. This creates a *toroidal* magnetic field component, which expands upward in the Z direction. This is sometimes called the “spring” process.
 - Constriction (collimation) of the flow by magnetic coils. The “hoop stress” of the helical field also squeezes the outflowing plasma toward the rotation axis. This is a manifestation of the plasma “pinch” process.
- Blandford and Payne also discussed two new flow singular points, in addition to the gas sonic point discussed in Chapter 13. These are the Alfvén point (AP) and the modified fast point (MFP). BP emphasized the Alfvén one, but downplayed the role of the MFP. However, the MFP may be an even more important feature of jet flows than the Alfvén point and could be related to several observational features of black hole exhaust systems.

The bottom line is that rotating magnetic fields threading an accreting plasma can turn a flat, rotating inflow into a collimated, supersonic outflow. Such are the power and geometry-altering capabilities of magnetic fields.

15.1.1.1 Cold, Axisymmetric, Stationary, Self-Similar MHD Flow

Full MHD is four-dimensional (three spatial dimensions plus time). BP simplified this complex problem by looking for solutions that were axisymmetric ($\partial/\partial\phi = 0$) and time-independent ($\partial/\partial t = 0$). They also made the additional assumptions that the plasma was cold ($p_g = p_r = 0$) and that the solution is self-similar in the spherical-polar radial coordinate $r = (R^2 + Z^2)^{1/2}$.

The cold plasma assumption is simply the strong magnetic field assumption stated quantitatively, $p_g \ll B^2/8\pi$, so that we can ignore the plasma pressure forces entirely in the disk corona and wind. Close to, and certainly inside, the disk itself this assumption breaks down and pressure dominates (see Chapter 12), and the wind velocity will be subsonic ($V < c_s = (\Gamma p/\rho)^{1/2}$). Here we simply ignore the dense part of the flow entirely and assume $c_s \ll V \ll V_A$, where

$$V_A = \frac{B_p}{(4\pi\rho)^{1/2}} = \left(\frac{B_R^2 + B_Z^2}{4\pi\rho} \right)^{1/2}$$

Note that, because of axisymmetry, Alfvén waves cannot travel in the ϕ direction, so the Alfvén speed V_A does not include any B_ϕ .

This is the first place in this book where the self-similar assumption is discussed. It is closely related to our search for radial scaling laws for accretion disks presented in Chapter 12. Such assumptions take the complex two-dimensional MHD problem and turn it into a simpler one-dimensional one. But, in which dimension do we make the symmetry assumption when we are discussing jets? Some choices are:

- Cylindrical Z and then solve only for the R dependence. However, presupposing that the structure in the Z direction was self-similar would be of no help. We are trying to *learn* about the vertical structure of accelerating jet flows, so this approach simply would be “begging the question” on the Z dependence.
- Cylindrical R and then solve only for the Z dependence. This is better and would be similar to accretion disk R -scaled solutions; we could impose Keplerian rotation in the outflow, but we would learn little about its collimation.
- Spherical θ and solve only for the spherical r dependence. This is similar to the spherical wind and accretion flow studies of Sections 12.1 and Chapter 13; it assumes diverging outflow and no collimation.
- Spherical-polar r and solve only for the θ dependence. This is the symmetry direction chosen by Blandford and Payne. Quantities were *not* assumed to be constant with r (i.e., $\partial/\partial r \neq 0$), but they were assumed to be separable, e.g.,

$$V_Z(r, \theta) = \mathcal{V}_Z(r) f(\theta)$$

where $\mathcal{V}_Z(r)$ and $f(\theta)$ can be determined separately from differential equations that are functions of, respectively, r and θ only. This allows collimation to occur if it is warranted, but also allows the winds to diverge into conical or radial sprays, if that is required instead by the equations.

It must be remembered that any symmetry assumption implies that waves can propagate *only in one dimension* – here in the θ direction. There can be flow in r (*i.e.*, $\mathcal{V}_Z(r)$), but only the acceleration/deceleration of the flow in the θ direction is actually solved by the self-similar equations. We return to this important point later.

15.1.1.2 The Blandford–Payne MHD Wind Equation

The BP wind equation is fairly complex; yet it is the simplest of all four MHD flows that are discussed in this section, and therefore the best example. We therefore will spend a little time fully deriving it (and only it). In the subsections below we shall simply write down the other, more complex wind equations and refer the reader to the actual papers that derived and solved them.

The multiple symmetry assumptions reduce the $\mathbf{r}$ and $\mathbf{V}$ vectors to

$$\mathbf{r} = (R, \phi, Z) = (R_0 x, \phi, R_0 y) \quad (15.1)$$

$$\mathbf{V} = (V_R, V_\phi, V_Z) = (x', g, f) \left(\frac{GM}{R_0} \right)^{1/2} \quad (15.2)$$

where x , f , and g are functions of the independent variable

$$y = \frac{Z}{R_0}$$

R_0 is the value of R on the equator ($x = 1, y = 0$), and

$$x' = \frac{dx}{dy}$$

is the slope of the field/stream line. Here x is the scaled radius of a given field line, and g is the θ -dependence of the azimuthal velocity V_ϕ .

The equations for relativistic, stationary, axisymmetric MHD flow are derived in Appendix F. Their non-relativistic form can be extracted by setting $\gamma \approx 1 + \frac{1}{2}V^2/c^2$ and ignoring terms in the main equations of order V^2/c^2 or $\hbar/c^2$ or higher. The angular momentum and energy equations, (F.13) and (F.16), can be combined with the ϕ -component of the velocity equation (F.10) to obtain a quartic equation

$$T - f^2 \mathcal{U}^2 = \left[\frac{(\lambda - x^2) \mathcal{M}_{\text{NR}}^2}{x(1 - \mathcal{M}_{\text{NR}}^2)} \right]^2 \quad (15.3)$$

where T is related to the effective gravitational potential

$$T \equiv -2 [\varphi(R, Z) - \varphi(R_0, 0)] = x^2 + 2S - 3$$

S is related to the true gravitational potential

$$S \equiv \frac{1}{(x^2 + y^2)^{1/2}}$$

$\mathcal{U}$ is related to the poloidal velocity

$$\mathcal{U} \equiv (1 + x'^2)^{1/2}$$

λ is the dimensionless angular momentum constant

$$\lambda \equiv \frac{\ell}{(GM R_0)^{1/2}}$$

and $\mathcal{M}_{\text{NR}}$ is the Alfvén Mach number in this non-relativistic case

$$\mathcal{M}_{\text{NR}} \equiv \frac{V_p}{V_A} = \left[\frac{V_R^2 + V_Z^2}{(B_R^2 + B_Z^2)/4\pi\rho} \right]^{1/2} \quad (15.4)$$

which can be re-written as

$$\mathcal{M}_{\text{NR}}^2 = \kappa x f J$$

where κ is the dimensionless flux constant

$$\kappa \equiv k \mathcal{U}_0 \left(\frac{GM}{R_0 B_0^2} \right)^{1/2}$$

$k = 4\pi\rho V_p/B_p$ is the “ratio of mass flux to magnetic flux” (Appendix F), and J is related to the spreading of the magnetic field lines

$$J \equiv -y^2 \frac{\partial(x/y)}{\partial y} = x - y x'$$

which controls how the Mach number behaves as the wind diverges or converges. Normally, in a cold plasma (with $c_s = 0$) the Alfvén and magnetosonic speeds V_A and c_{ms} would be equal. However, because of our assumption of axisymmetry here, Alfvén waves lose their ability to travel in the azimuthal direction. Therefore, in this case the two speeds are *not* equal

$$V_A^2 = \frac{B_R^2 + B_Z^2}{4\pi\rho}$$

$$c_{ms}^2 = \frac{B_R^2 + B_Z^2 + B_\phi^2}{4\pi\rho}$$

This fact is important in understanding the structure of this MHD wind.

The trick used by Blandford and Payne to derive a final MHD wind equation (and by several others since then) was to differentiate the dimensionless energy (Bernoulli) equation (15.3) and combine it with a component of the momentum

equation (these authors used the Z -component, equation (F.18)). The result yielded an equation for the cold, non-relativistic case that is similar to previous wind equations we have already seen (12.2), (12.5), (13.1), and (13.21)

$$\boxed{\frac{d\mathcal{M}_{\text{NR}}}{dy} = \frac{\mathcal{N}_{\text{NR}}}{\mathcal{D}_{\text{NR}}}} \quad (15.5)$$

where the numerator is rather complex²

$$\boxed{\mathcal{N}_{\text{NR}} = \frac{\mathcal{M}_{\text{NR}} S^2}{2 x T} \left\{ 2 \mathcal{M}_{\text{NR}}^4 y (x^2 - \lambda) J \right.} \\ \left. - (\mathcal{M}_{\text{NR}}^2 - 1) \left[\frac{5}{4} T + x^2 - S \right] x (y + x x') \right.} \\ \left. - (\mathcal{M}_{\text{NR}}^2 - 1)^2 \left[y (x^2 + T) - f^2 (y + x x') \right] J \right\} \quad (15.6)$$

but the denominator is rather simple

$$\boxed{\mathcal{D}_{\text{NR}} = (\mathcal{M}_{\text{NR}}^2 - 1) (t_{\text{NR}}^2 - 1)} \quad (15.7)$$

with the *magnetosonic* Mach number (in the θ direction only) given by

$$t_{\text{NR}} = \frac{V_{\theta}}{c_{ms}} = \frac{V_{\theta}}{B/(4\pi\rho)^{1/2}} \quad (15.8)$$

15.1.1.3 The Non-Relativistic MHD Wind Singular Surfaces

A magnetized plasma has three characteristic wave speeds: the fast and slow magneto-acoustic speeds, V_F and V_S , and the Alfvén speed V_A (see Section 9.6.2). Therefore, based on the simpler (spherical) Parker wind theory (Section 12.1.1), which has one wave speed and one singular surface, we would expect to find three singular surfaces, each where the flow velocity reaches each of the magneto-acoustic and Alfvén wave speeds. However, because we have made the cold plasma assumption here ($c_s = 0$, so $V_S = 0$ and $V_F = c_{ms}$), equation (15.5) has only two singular surfaces: the “Alfvén surface”, where the Alfvén speed is reached ($\mathcal{M}_{\text{NR}}^2 = 1$), and the “modified fast surface” (MFS), where the only allowed fast magnetosonic

² We shall discuss three more MHD wind equations in this section, each adding more physics and more complexity. But we shall show the numerator only in this rather simplified cold, non-relativistic case. It should be clear to the reader that, even with the self-similar assumption, realistic MHD wind equations for collimating jets can be very complex indeed.

velocity is reached ($t_{\text{NR}}^2 = 1$). The term “modified” is used here because the two-dimensional, axisymmetric MHD equations in Appendix F have been modified by the assumption of self-similarity.

Investigators in this field sometimes refer to singular points instead of surfaces. This occurs because one often considers a single streamline. (In the case of MHD winds, the streamline and magnetic field line are the same.) A singular point, therefore, is where a specific stream or field line crosses a singular surface, with the entire surface being the locus of all singular points of that type. The sonic and the modified fast surfaces, for example, are often referred to as the “sonic point” and “modified fast point” (or MFP), especially when some symmetry assumption (sphericity or self-similarity) is being made so that one point can represent all others on the surface. We often use that common terminology below.

The Alfvén surface is well known and non-controversial. When $\mathcal{M}_{\text{NR}} = 1$, in order for equations (15.3) and (15.5) to remain finite and well behaved (“regular”), we need to have $x = \lambda^{1/2} \equiv x_{\text{Alf}}$, which is the (dimensionless) Alfvén radius

$$x_{\text{Alf}} \equiv \frac{R_{\text{Alf}}}{R_0}$$

R_{Alf} is the cylindrical distance from the rotation axis where matter, if it *could* rotate at the same angular velocity as the magnetic field, would have a specific angular momentum equal to that ℓ , which is conserved along that field line

$$R_{\text{Alf}} = (\ell/\Omega_f)^{1/2} = (\ell/\Omega_K)^{1/2} \quad (15.9)$$

The classical fast surface (where $V_p = c_{ms}$) also is not controversial; however, it is *not* a singular surface of the BP equations. Generally (but not in all solutions), once the poloidal velocity exceeds V_A , it will not be long before the accelerating flow exceeds c_{ms} . However, the surface where $V_p = c_{ms}$ corresponds to neither $\mathcal{M}_{\text{NR}} = 1$ nor $t_{\text{NR}} = 1$. *There is no classical fast singular surface*; the solution simply accelerates smoothly through the fast magnetosonic speed.

However, the modified fast surface *has* been the subject of controversy. In the BP self-similar equations the MFP occurs where the plasma velocity *toward the axis* (V_θ) equals the magnetosonic speed. The MFP seems to be a singular point, rather than the classical fast point, because of the self-similar assumption. And, while the Alfvén and classical fast surfaces have been identified in time-dependent numerical MHD simulations, the MFP has not. The reason for this, however, is not necessarily that the MFP does not exist. Instead, the first two surfaces occur relatively near the accretion disk/black hole region, while the modified fast surface will occur very far from the disk – at $Z \gg R_0$, possibly outside the computational domain of most simulations. The MFP also implies that the jet will be significantly modified beyond the MFP. Flow toward the rotation axis at greater than the magnetosound speed may either bounce (perhaps multiple times), or possibly create a collimation shock and restructure the flow entirely.

Blandford and Payne discussed this dilemma at some length and concluded that there should be only one singular surface: the Alfvén one. The MFP, they argued,

must lie at infinity. We return to this point later in this section. But, to foreshadow that discussion, we note here that the MFP (and the Alfvén point) is not simply a mathematical concept created by a symmetry assumption. It is related to the general MHD concept of separatrix surfaces (see Sections 9.6.2 and 15.1.3). The MFP is not only very important in jet production; it may, in fact, be the key to creating independent and stable jet flows. Whether or not a jet becomes kinetic-energy-flux-dominated (KFD), or remains Poynting-flux-dominated (PFD), depends on whether an MFP (*i.e.*, fast magnetosonic separatrix surface) forms in the flow or not.

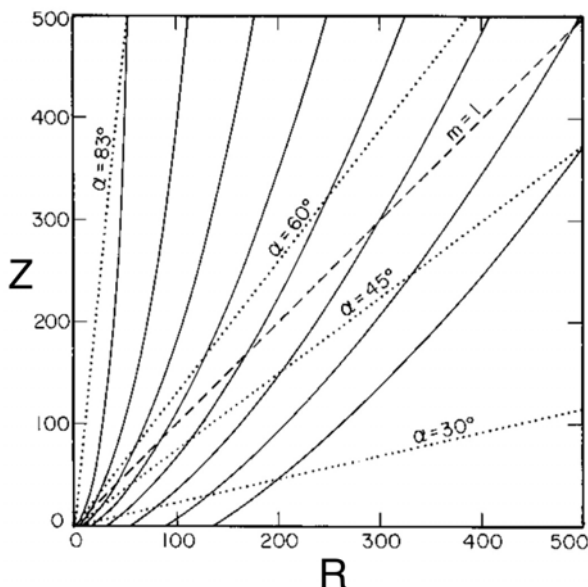


Fig. 15.2: A semi-analytic solution of the BP wind eq. (15.5) showing field/stream lines (solid lines) of the flow in cylindrical coordinates, with the infinitely thin accretion disk on the equator. Remaining lines are all rays that make a constant angle with the Z axis: dashed line is a special case where the Alfvén Mach number is unity (each crossing with a solid line is a different Alfvén point on a different streamline), and dotted lines are labeled with magnetic field pitch angle values. Reproduced from Fig. 3 of [507], by permission of J. Wiley and Sons, Ltd.

15.1.1.4 Solutions and Simulations of Non-Relativistic Blandford–Payne MHD Winds

Fig. 15.2, taken from Blandford and Payne’s paper, shows flow streamlines (and magnetic field lines) for a solution with $\kappa = 0.03$ and $\lambda = 30$, projected into the (R, Z) -plane. Note that each radial line ($\theta = \text{constant}$) intersects each streamline at the same angle. This is a consequence of the r -self-similar assumption. The values of $\chi = \tan^{-1} |B_\phi/B_p|$ (which are labeled as α in the figure) are not the polar angle

but rather values of the magnetic “pitch angle”, which also are constant along radial r lines. The final opening angle of the jet produced in this model is $\theta_j \approx 6^\circ$.

The group whose numerical investigations most closely paralleled the semi-analytic, cold MHD wind work of BP is one composed of Ruben Krasnopolsky (at Caltech when they published their 1999 paper), Zhi-Yun Li (University of Virginia), and Roger Blandford (also then at Caltech) [523]. In order to mimic the BP results, they paid particular attention to boundary conditions, both at the infinitely thin accretion disk at $Z = 0$ and at the outflow boundaries ($R = R_{max}$, $Z = Z_{max}$), making sure that the wind structure was determined only by conditions in the Keplerian accretion disk. They then ran the simulation for a very long time, until it reached the point where its structure stopped changing – a quasi-steady state.

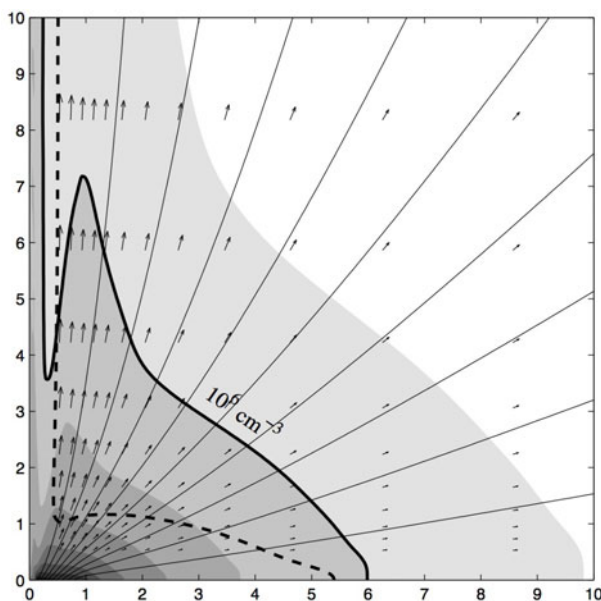


Fig. 15.3: Magnetic field/stream lines for a steady, non-self-similar solution to the BP MHD wind. Axes are in astronomical units. Thick solid line and shading indicates density contours ($\sqrt{10}$ change for every shade), and the thick dashed line shows the classical fast magnetosonic surface. The solution produces a highly collimated jet out to at least $Z = 100$ AU (not shown). Reproduced from Fig. 4a of [523], by permission of the AAS.

Fig. 15.3 shows a typical Krasnopolsky *et al.* model for a non-self-similar, fully two-dimensional MHD wind. Compare this figure with Fig. 15.2. The first thing to note is that collimation of outflows by magnetocentrifugal winds works; and it does so independent of the self-similar assumption. The flow near the axis is quite collimated, can be considered a jet, and carries a substantial fraction of the ejected matter (20%) in a dense flow. Secondly, the fast magnetosonic surface (dashed line) occurs close to the disk, except near the rotation axis (where there is a hollow region

with no jet flow). Thirdly, the simulation goes out to 100 disk radii (100 AU, not shown in the figure), and there still is no sign of a modified fast surface. That is, there is no place where the flow is toward the Z axis, let alone at a speed greater than the magnetosonic speed. Therefore, the MFP in this model is either at infinity or, at least, at a point much farther than 100 AU away. These simulations, therefore, do not address the question of whether the jet may become kinetic energy dominated or not and whether or not it may be severely pinched or crushed into the axis.

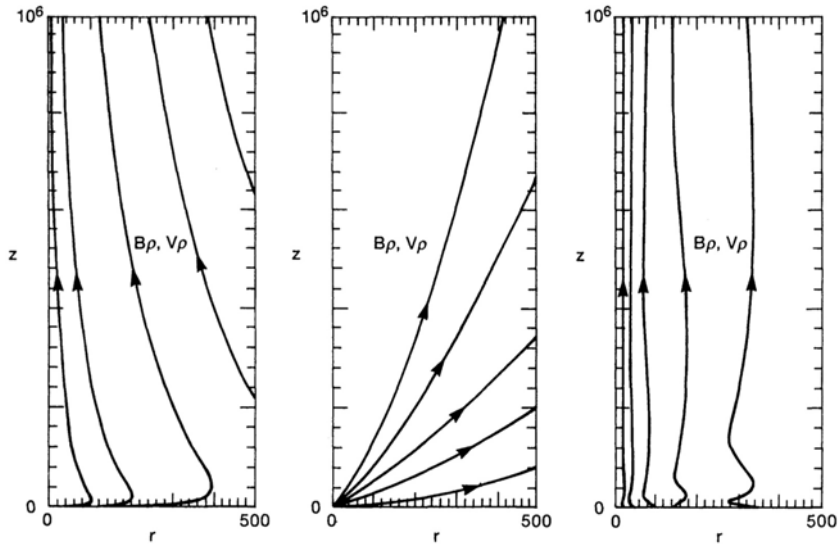


Fig. 15.4: Magnetic field/stream line plots for three Contopoulos and Lovelace models. Each has the same κ , λ , and rotational velocity profiles (BP notation), differing only in their magnetic field/current distribution parameter: left, $F = 0.8$ (similar to BP); center, $F = 0.9$; right, $F = 1.02$. Only the latter is current-carrying. Reproduced from Fig. 2 of [524], by permission of the AAS.

In 1994, however, John Contopoulos and Richard Lovelace at Cornell were able to obtain a much more general set of solutions to the MHD wind equations than the BP ones [524]. They showed that simply by varying the amount of current traveling down the jet, *i.e.* the exponent F of the self-similar scaling law,

$$B_\phi \propto R^{F-2}$$

they were able to obtain not only solutions like Blandford and Payne's ($F = 3/4$), but also ones that recollimate toward the axis or even oscillate. (See Fig. 15.4.) Such a great sensitivity to a small parameter meant that a much greater variety of MHD wind and jet solutions were possible from MHD theory, and foreshadowed recollimating solutions that *do* pass through an MFP. The total current in the Z direction is proportional to $R^2 |\nabla \times \mathbf{B}| \propto R^{F-1}$, so jets with $F > 1$ are called “current-carrying” jets.

15.1.1.5 Energy, Mass, and Angular Momentum Loss Rates

Blandford and Payne derived a set of three algebraic equations that give the energy loss (total luminosity L_T), mass-loss rate ($\Delta\dot{M}$), and angular momentum loss (torque $\mathcal{G}$) as a function of the parameters λ , κ , x' , and the disk inner and outer radii ($R_{0,min}$, $R_{0,max}$). Since $B_{Z0} = B_0/(1 + x_0'^2)$, we can write the BP's expressions for L_T , $\Delta\dot{M}$, and $\mathcal{G}$ in a somewhat simpler manner that can be compared with results of other winds discussed in the next subsections

$$\begin{aligned} L_T &\equiv \frac{1}{2} (\lambda - 3/2) \left(\frac{R_{0,max}}{R_{0,min}} \right) \frac{\Psi_{out}^2 \Omega_{out}}{R_{0,eff}} \\ \Delta\dot{M} &= \frac{L_T}{\langle V_\infty \rangle^2} \\ \mathcal{G} &= \frac{\lambda \Psi_{out}^2}{R_{0,eff}} \end{aligned} \quad (15.10)$$

where the outer magnetic flux, angular velocity, and effective disk radius for the torque are

$$\begin{aligned} \Psi_{out} &\equiv \int_0^{R_{0,max}} B_{Z0}(R) R dR = \frac{3}{4} B_{Z0}(R_{0,max}) R_{0,max}^2 \\ \Omega_{out} &= \left(\frac{GM}{R_{0,max}^3} \right)^{1/2} \\ R_{0,eff} &= \frac{8}{9\kappa} R_{0,max} \end{aligned}$$

and the average wind/jet speed at infinity is (assuming that half the total power emerges as kinetic energy)

$$\langle V_\infty \rangle = \left[\frac{\lambda - 3/2}{\ln(R_{0,max}/R_{0,min})} \right]^{1/2} \left(\frac{GM}{R_{0,min}} \right)^{1/2} \quad (15.11)$$

For Blandford and Payne's standard model ($\kappa = 0.03$, $\lambda = 30$) and taking $R_{0,max}/R_{0,min} \approx 10$, we find

$$\begin{aligned} V_\infty &\approx 3.5 V_{\phi 0,min} = 3.5 \left(\frac{GM}{R_{0,min}} \right)^{1/2} \\ R_{0,eff} &\approx 30 R_{0,max} \approx 300 R_{0,min} \end{aligned}$$

and, therefore,

$$L_T \approx 4.8 \frac{\Psi_{out}^2 \Omega_{out}}{R_{0,max}} = 4.7 \mathcal{G} \Omega_{out}$$

$$\Delta \dot{M} \approx 0.04 \frac{\Psi_{out}^2}{R_{0,max}^3 \Omega_{out}}$$

$$\mathcal{G} \approx 1.0 \frac{\Psi_{out}^2}{R_{0,max}}$$

These formulae can be used to compute approximate energy, mass, and angular momentum loss rates for typical non-relativistic MHD wind/jet outflows from accretion disks.

15.1.1.6 Generalized Energy Loss Rates: Jet Ejection from Different Accretion Flow Models

In Sections 12.2 and 12.3 we found that the dependence of B_Z and even V_ϕ on R can vary between inflow models or even within a single model, especially in the MDAF case. Therefore, in order to compute the jet luminosity from a model accretion flow, we will need a more general form of equation (15.10). To obtain this we will add the parameters

$$F \equiv \frac{\partial \ln B_{Z0}}{\partial \ln R_0} + 2 \quad (15.12)$$

and

$$W \equiv -\frac{\partial \ln V_{\phi 0}}{\partial \ln R_0} \quad (15.13)$$

For example, when the scaling of B_{Z0} is the same as Shakura and Sunyaev's "outer" disk solution ($B_{Z0} \propto R^{-5/4}$) we have $F = 3/4$. And when the flow is Keplerian ($V_{\phi 0} \propto R^{-1/2}$) we have $W = 1/2$.

General Expressions for Jet Power

The generalized expression for the BP MHD wind power is

$$L_T = \kappa (\lambda - 3/2) \int_{R_{0,min}}^{R_{0,max}} B_{Z0}^2(R) V_{\phi 0}(R) R dR$$

We now assume that the leading parameter factor $\kappa (\lambda - 3/2)$ is of order unity, and we will drop it from the equations. When $2(F - 1) - W < 0$ (power dominated by outflow near the inner radius), then equation (15.10) has the more general form

$$L_T = L_{T,in} \equiv \frac{B_{Z0}^2(R_{0,min}) R_{0,min}^2 V_{\phi 0}(R_{0,min})}{W - 2(F - 1)} \quad (15.14)$$

On the other hand, when $2(F - 1) - W > 0$ (power dominated by outflow near the outer radius), this expression becomes

$$L_T = L_{T,out} \equiv \frac{B_{Z0}^2(R_{0,max}) R_{0,max}^2 V_{\phi 0}(R_{0,max})}{2(F - 1) - W} \quad (15.15)$$

And, when $2(F - 1) - W = 0$, each logarithmic radial section contributes equally

$$L_T = L_{T,log} \equiv B_{Z0}^2(R_{0,min}) R_{0,min}^2 V_{\phi 0}(R_{0,min}) \ln \left(\frac{R_{0,max}}{R_{0,min}} \right) \quad (15.16)$$

These expressions are remarkably similar to the form for the pulsar power expressed in equation (14.6), but with the light cylinder radius replaced by either the inner or outer accretion flow radius.

Table 15.1: Types of possible jet outflows from accretion inflow models in Chapter 12

Accretion Inflow	F	W	$2(F - 1) - W$
SS disk “outer” region (eqs. (12.26))	13/16	1/2	−7/8
SS disk “middle” region (eqs. (12.31))	31/40	1/2	−19/20
SS disk “inner” region (eqs. (12.35))	1/4	1/2	−2
Slim disks & ADAFs (eqs. (12.52))	3/4	1/2	−1
MDAF transitional flow (eqs. (12.66))	2	1/2	3/2

Jet Power in Accretion Flows as a Function of Accretion Rate

Table 15.1 shows most accretion flow models discussed in Chapter 12. (The true MDAF regions have been ignored, as their magnetic structure is not expected to produce open field lines except in the transitional flow and ADAF portions, or when anchored in the black hole horizon itself (Blandford–Znajek process).) The standard turbulent disks (Shakura–Sunyaev and ADAF models) have $2(F - 1) - W < 0$, so any jet outflow from those disks would be dominated by power generated in the inner portion of the accretion flow. The transitional flow region is interesting. It has $2(F - 1) - W > 0$, indicating that the jet outflow from this region is produced primarily by this accretion flow at its outer edge. So, if the ADAF is indeed truncated by an MDAF at some radius in the ADAF interior, we would expect that *most of the jet power in this type of flow would be generated at the MDAF/ADAF boundary*.

Equations (15.14) and (15.15) can be used with Table 15.1 to produce estimates of the jet power from the specific accretion models discussed in Sections 12.2 and 12.3. To obtain definite and reasonable numbers, we will use $\alpha = 0.3$ for the slim,

ADAF, and MDAF models, but we will leave the viscosity parameter for the thin SS disk models a free parameter. Both the ADAF and transitional flow regions scale with the ϑ parameter in the same linear manner, where ϑ is, again, the T_i/T_e saturation ratio that characterizes the MDAF interior to the hot ADAF. If we require that the jet luminosity not exceed the accretion power, then the results of this exercise are as follows:

$$\begin{aligned}
 L_J &= L_{\text{Edd}} \dot{m} (\vartheta/820) & \dot{m} < \dot{m}_A \\
 &= 0.03 L_{\text{Edd}} \alpha^{-1/10} m^{-1/10} \dot{m}^{6/5} & \dot{m}_A < \dot{m} < \dot{m}_I \\
 &= 0.5 L_{\text{Edd}} \dot{m}^2 & \dot{m}_I < \dot{m} < 1
 \end{aligned} \tag{15.17}$$

As with equations (14.26) $\dot{m}_A$ is given by equation (12.54), and $\dot{m}_I$ is given (roughly) by equation (12.60). For $\vartheta = 1$ (MDAF starting at $r \approx 5000 r_g$) the jet power is less than that produced by an SS thin disk. But, if $\vartheta \geq 820$, most of the MDAF structure lies inside the ISCO, leaving only the ADAF to produce the jet. In this latter case the jet power is

$$L_J = L_{T,\text{adf}} \approx L_{\text{Edd}} \dot{m} \tag{15.18}$$

which is still significantly below the Eddington limit because $\dot{m} < \dot{m}_A \ll 1$.

For accretion rates in the range ($\dot{m}_A < \dot{m} < \dot{m}_I$) the stable SS “middle” region is the appropriate disk model. Note that $\dot{m}_I > \dot{m}_A$ occurs only for black holes with masses less than $\sim 10^5 M_\odot$ (equation (12.60)). As the accretion rate increases, the observable signature of black holes with masses larger than this should transition directly from a low/hard state (ADAF) to a thermally and secularly unstable state that is characteristic of a SS “inner” disk accretion flow. Note that this does not mean that a SS “middle” region disk cannot exist for accretion rates in the range $\dot{m}_I < \dot{m} < \dot{m}_A$. It simply means that in that accretion rate range the observable properties of the source will be dominated by those of an ADAF-like corona; some thermal component of the cool disk, however, may still be visible in the 1–10 keV X-ray spectrum.

For high, but sub-Eddington, accretion rates ($\dot{m}_I < \dot{m} < 1$), an SS “inner” accretion disk region (equations (12.35)) appears. However, this region is thermally and secularly unstable. A jet from the “inner” region may manifest itself intermittently (Section 12.2.2) on a time scale equal to its secular/viscous time (equation (12.40))

$$\tau_{\text{acc}} = \frac{R_{M-I}}{V_R} = 7.0 \text{ s } \alpha^{-2/3} m^{4/3} \dot{m}^{2/3} \tag{15.19}$$

A weak jet also will continue to be ejected from the stable “middle” region, whose inner radius grows with $\dot{m}$. This leaves an increasingly large gap in the central disk where the “inner” region should be and keeps the steady jet quite weak in between the intermittent powerful jet ejections from the unstable “inner” region.

For super-Eddington accretion flows, the slim disk model (equations (12.52)) is appropriate. In this case the jet total power is of order the *super*-Eddington accretion power

$$L_J = L_{T,\text{slm}} \approx L_{\text{Edd}} \dot{m} \quad 1 < \dot{m} \quad (15.20)$$

(identical to equation (15.18)) and could considerably exceed the radiation losses, which are only equal to the Eddington luminosity. However, the time-dependent behavior of the entire source will be slaved to the instabilities occurring in the “inner” region (where $H < R$), which now lies outside the super-Eddington slim disk region. Super-Eddington accretion flows, therefore, still may produce jets, but they also will be intermittent on the “inner” region secular time scale in equation (15.19).

Jet Suppression at Low Accretion Luminosities

In some X-ray binary sources the jet is seen to suddenly decrease in luminosity. Equations (15.17) show at least two different methods for suppressing the (primarily ADAF) jet power. The first, suggested by D. Meier [432], occurs when the disk becomes geometrically thin and optically thick ($\dot{m} > \dot{m}_A$), resulting in a greatly reduced vertical magnetic field to power the jet. In this case the suppression factor is of order $60/\dot{m}_A^{1/5} \sim 100$. The second method of suppressing the jet power is to create an MDAF in the interior of the ADAF. This truncates the ADAF jet launching radius to $\sim R_0$ (equation (12.64)), resulting in the jet being launched much farther from the black hole than a standard ADAF (in which the jet is launched deep in the potential well near the ISCO). In this case the suppression factor is $R_0/r_{\text{isco}} \sim 820/\vartheta$.

Jet Suppression at Super-Eddington Accretion Luminosities

Even when microquasars like GRO J1655-40 were first discovered [232], it was apparent that a radio jet was not ejected until the source’s luminosity dropped below L_{Edd} . Since then there have been many suggestions for why super-Eddington objects might not produce an observable jet. In 1996 this author suggested [525] that jet launching would be disrupted either by Papaloizou–Pringle instabilities [526, 527] or by the strong disk radiation pressure or wind. PP instabilities, however, require a hard inner reflection point in the disk (centrifugal barrier or stellar surface) to be effective, and this is not likely when a soft ISCO exists. Given the strong similar physical structure in advective disks dominated by radiation pressure (slim disks) and in those dominated by gas pressure (ADAFs; see equations (12.52)), radiation or wind pressure also seems unlikely explanations. ADAFs are apparently quite efficient at producing radio (Section 16.2.4); why not slim disks? Both have similar internal ratios of dynamical to magnetic forces.

Another unlikely possibility is that slim disks may be more geometrically thin than their ADAF cousins. This would reduce the vertical magnetic field in slim disks by a factor of H/R and suppress the ejected jet power by that factor squared (similar to our model for the SS “middle” region in the second expression in equations (15.17)).

One property possessed by super-Eddington disks, and not by ADAFs, is a strong radiation field. In his discussion of BAL QSOs, Robert Antonucci [528] suggested that these quasars do have jet-like bipolar outflows, but they never develop non-thermal particle distributions because of Compton cooling by the strong radiation field. That is, jets from super-Eddington sources *are* produced, but they remain thermal like the jet in SS433 (another suspected super-Eddington source) and therefore undetected in the radio. There has been much discussion since on this question [529, 530, 532, 531, 533], but most of it appears to be applied to GRB jets that already are highly relativistic, not to Antonucci’s original question of either jet formation or *particle* acceleration in a high-intensity radiation environment. Also, interest in this subject seems to have waned after the turn of the millennium. Therefore, at present we do not seem to have a good theory for computing the radio powers of jets from super-Eddington sources.

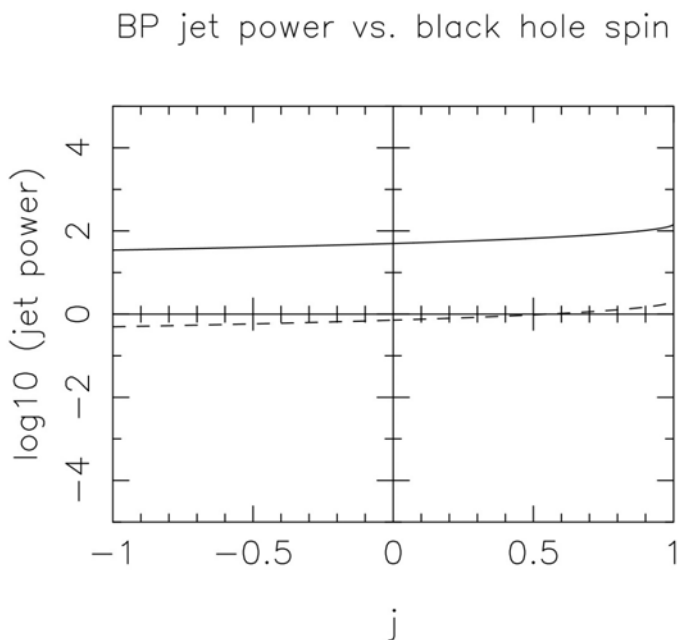


Fig. 15.5: Similar to Fig. 14.25 but for the Blandford–Payne (disk-launched) jet (eq. (15.21)). As before, the thick solid line is for an ADAF-like inflow while the dashed line is for the Shakura–Sunyaev “middle” region disk solution. Generally, the disk jet power is a much stronger function of accretion flow type than of the spin of the black hole itself.

The Spin Dependence of the Blandford–Payne Effect

Just as we presented the spin-dependence of the Blandford–Znajek effect in Section 14.4.2, we also can apply the same principles to accretion disk-driven jets and plot their relative power as a function of black hole spin (including retrograde, as well as prograde, accretion).

$$\mathcal{P}_{\text{BP}}^{\text{EM}} = B_{Z,\text{isco}}(j)^2 r_{\text{isco}}(j)^3 \Omega_{\text{isco}}(j) \quad (15.21)$$

Again, as in David Garofalo’s computation of the spin-dependence of the Blandford–Znajek effect [501], $r_{\text{isco}}(j)$ is computed from equation (7.60), and for $B_{Z,\text{isco}}$ we use either equation (12.52) for an ADAF or equation (12.31) for a thin cool disk and use $x = r_{\text{isco}}/6r_g$ in the B_Z expression. $\Omega_{\text{isco}}(j)$ is the relativistic Keplerian angular velocity in the Boyer–Lindquist metric at r_{isco} .

Fig. 15.5 shows the results of this calculation, again for the two most reliable disk models (hot ADAF and cool Shakura–Sunyaev “middle” region) at the accretion rate boundary $\dot{m} = \dot{m}_A$. Note that the accretion disk jet power rises gradually, and monotonically, as the black hole spin increases toward +1. This occurs because r_{isco} decreases monotonically with j , causing the disk magnetic field and rotation rate to monotonically increase in equation (15.21), with a net dependence of about $\mathcal{P}_{\text{BP}}^{\text{EM}} \propto r_{\text{isco}}^{-1}$.

15.1.2 Relativistic, Cold MHD Wind and Jet Theory

The classical BP MHD wind theory is ideally suited for applications to jets from protostars, symbiotic stars, and other non-relativistic systems. However, black hole jets are often relativistic, reaching speeds of $\gamma = 3\text{--}300$ ($V = 0.95\text{--}0.999995c$). In order to see if MHD processes can accelerate jets to such high speeds, it is important to develop a relativistic theory of MHD winds before it can be applied to the jets observed (or inferred) in microquasars, AGN, and γ -ray bursts.

The study of relativistic MHD winds had been the subject of pulsar investigations since the late 1960s (see Chapter 14). However, Max Camenzind (now at the University of Heidelberg) was one of the first pioneers in the application of relativistic MHD winds to AGN winds and jets [534, 535, 536]. In the mid-to-late 1980s he showed that

- The magnetic field can effectively collimate the flow into a jet-like form.
- Acceleration can continue far past the classical magnetosonic point.
- Final jet Lorentz factors of at least $\gamma \sim 5$ can be obtained.

This was a considerable breakthrough and pointed to MHD processes as being a likely candidate for explaining AGN jets. (Previous suggestions that hydrodynamic or radiation–hydrodynamic processes by other authors were able to achieve only $\gamma < 2$, even when the flow was confined and collimated by the narrow polar funnel of a geometrically thick accretion disk.)

The study that was closest to a relativistic generalization of the Blandford and Payne classic paper, and which was built on Camenzind's early relativistic wind work, was one in 1992 by Zhi-Yun Li, Tzihong Chiueh, and Mitchell Begelman (Li and Begelman at University of Colorado and Chiueh at the National Central University in Taiwan), which we will sometimes refer to as LCB [537]. They started with the stationary, axisymmetric *relativistic* MHD equations, which we derived in Appendix F. These equations include the effects of a high Lorentz factor and the possible inertia of the kinetic, internal, and pressure energies associated with the plasma (which were ignored by Blandford and Payne's non-relativistic treatment). These equations also take into account an *electric* force perpendicular to the magnetic field lines, which makes the collimation of the flow more difficult in the relativistic case than the non-relativistic one.

In the end, however, like Blandford and Payne, LCB made the same cold plasma assumption ($h = 0$). And they also assumed r -self-similarity, which, for the relativistic case, means that they must ignore gravity altogether (not simply assume a non-relativistic gravitational field). Together, these assumptions mean that both the slow magneto-acoustic and escape speeds are ignored (*i.e.*, assumed to be so small as to not be important). However, the loss of gravity also frees up the magnetic field scaling in r , allowing a greater variety of poloidal current distributions along the flow (*i.e.*, the parameter F). As Contopoulos and Lovelace found, this can lead to a greater variety of converging, diverging, or even oscillating jet solutions.

15.1.2.1 The Cold, Relativistic Wind Equation and its Singular Surfaces

LCB's final relativistic MHD wind equation can be found using techniques similar to BP's: take the derivative of the Bernoulli equation (F.16) with respect to the independent variable (here the radius R) and combine it with the component of the momentum equation that is *perpendicular* to the magnetic field (equation (F.18), the so-called "cross-field" or "trans-field" equation). The resulting relativistic MHD wind equation again has the form³

$$\frac{d\mathcal{M}_R}{dx} = \frac{\mathcal{N}_R}{\mathcal{D}_R} \quad (15.22)$$

where $\mathcal{M}_R$ is the relativistic Alfvén Mach number

$$\mathcal{M}_R \equiv \frac{\gamma V_p}{U_A} = \gamma \left[\frac{(V_R^2 + V_Z^2)}{(B_R^2 + B_Z^2)(1 - x^2)/4\pi\rho} \right]^{1/2} \quad (15.23)$$

The relativistic Alfvén speed is given by

³ Note that equations (15.5) and (15.22) differ in that one follows the acceleration in y (Z), while the other follows it in x (R). Ideally, of course, one would like to perform the derivatives along the field line itself, but either Z or R is a reasonable substitute.

$$U_A = \left[\frac{(B_R^2 + B_Z^2)(1 - x^2)}{4\pi \rho} \right]^{1/2}$$

and x is now the radius scaled in terms of the light cylinder radius

$$x \equiv \frac{R}{R_L} = \frac{R\Omega_f}{c}$$

rather than the arbitrary fiducial radius R_0 used in the non-relativistic case.

The numerator $\mathcal{N}_R$ of equation (15.22) is too complex to write down here (and of little importance to our discussion), but the denominator is again given by

$$\mathcal{D}_R = (\mathcal{M}_R^2 - 1)(t_R^2 - 1) \quad (15.24)$$

where t_R is the relativistic modified fast magnetosonic Mach number

$$t_R \equiv \frac{\gamma V_\theta}{U_{ms}} \equiv \frac{\gamma V_\theta}{\left\{ \left[(B_R^2 + B_Z^2)(1 - x^2) + B_\phi^2 \right] / 4\pi \rho \right\}^{1/2}} \quad (15.25)$$

Here, once again, V_θ is the component of the velocity in the cylindrical θ direction and is negative (or positive) for a jet collimating toward the positive (or negative) Z axis.

The relativistic corrections in equations (15.23) and (15.25) (compared to the non-relativistic equations (15.4) and (15.8)) are rather simple: the characteristic speeds now apply to the proper poloidal velocity

$$U_p = \gamma V_p$$

and the effect of the poloidal magnetic field drops to zero at the light cylinder ($x = 1$). In the non-relativistic BP case we are always well inside the light cylinder, because $R^2 \Omega_f^2 \ll c^2$; so the $(1 - x^2)$ factor does not appear in the earlier equations.

Requiring $\mathcal{N}_{NR} = 0$ at the Alfvén point leads to an equation for the Alfvén radius similar to equation (15.9)

$$R_{\text{Alf}} = \left[\frac{\ell}{\mu \Omega_f} \right]^{1/2} \quad (15.26)$$

where μ is the total energy of the plasma divided by its rest energy

$$\mu \equiv \epsilon \left(1 + \frac{\psi}{c^2} \right) = 1 + \frac{Be}{c^2} \quad (15.27)$$

(See equation (F.16).) Again, the Alfvén radius occurs where the angular momentum of any plasma that *could* rotate with the field line angular velocity Ω_f would equal the conserved value ℓ (which is determined at the base of the flow).

Li, Chiueh, and Begelman also discuss the fast and modified fast points at length. They also point out that the equations are not singular at the classical fast point.

And, as in the BP equations, it is not obvious from the momentum or energy equation alone that the modified fast point is a singular point. *Only when the momentum and energy equations are combined is it clear where the singular points are.* This behavior of wind equations goes all the way back to Section 12.1.1, where we derived our first simple wind equation and found that the location of the sonic point depended on whether the energy equation was adiabatic or isothermal. And it was especially important in locating the critical point in Case C in the more complex radiation-driven winds of Section 13.1.2.

15.1.2.2 Solutions for Cold, Relativistic MHD Winds

Most of the relativistic MHD wind solutions found by LCB collimated to a narrow cylindrical jet flow. Once the specific value of the total energy μ that moved the MFP to infinity was found, the solutions became functions of only two parameters: the magnetization parameter (which also is important in pulsar winds)⁴

$$\sigma_M \equiv \frac{\Omega_f^2 \Psi}{k c^3} = \frac{\Omega_f^2 \Psi B_p}{4\pi \gamma \rho V_p c^3} \quad (15.28)$$

(where Ψ is, again, the magnetic flux that is conserved along a field line) and the position of the Alfvén point

$$x_{\text{Alf}} \equiv \frac{R_{\text{Alf}}}{R_L}$$

which determines exactly how PFD the jet is (as $x_{\text{Alf}} \rightarrow 1$, the jet becomes fully electrodynamic, essentially free of material forces). In order to examine a typical LCB relativistic jet, we do not need to be concerned with exactly how PFD the jet is, so we will study how things change only with σ_M – a combination of rotation rate, magnetic field strength, and relative mass flux in the outflow.

Fig. 15.6 shows solutions to LCB's equations, as computed by Contopoulos [538] for the case when $\sigma_M = 10$. Despite the repulsive electric force between magnetic field lines in the relativistic case, the $\mathbf{J} \times \mathbf{B}$ magnetic pinch force still is able to collimate the wind into a jet. Fig. 15.6(left) (a $\log R / \log Z$ -plot) and Fig. 15.6(right) (a proper velocity *vs.* $\log Z$ -plot) illustrate the need to choose a specific μ for the flow in order to push the MFP to infinity: curve (3) shows $\mu = 21.8034$ and is a breeze solution that decelerates and halts at a very large radius; curve (2) is only 10^{-3} less (21.8024) but reaches an infinite Lorentz factor at a finite height; the cylindrical solution (dashed line) is between these two solutions. Neither of the former two are physically acceptable jet models. For the critical solution, LCB found that the Lorentz factor γ reaches $\gamma_F = 2.87$ at the classical fast point and eventually achieves $\gamma_\infty = 11.30$.

⁴ Note that the magnetization parameter is defined with quantities conserved along a magnetic field line, so σ_M also remains constant along a field line.

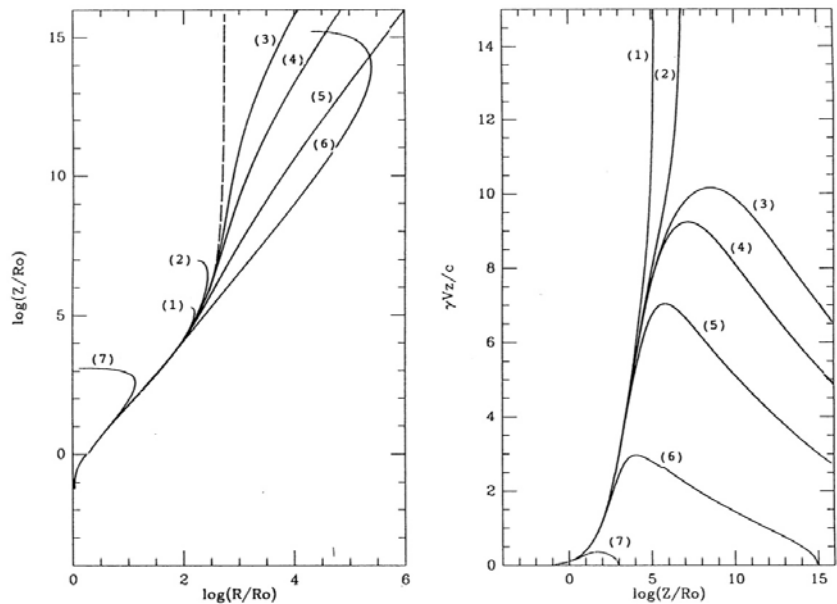


Fig. 15.6: Solutions to the equations of Li, Chiueh, and Begelman [537] as calculated by Contopoulos. *Left:* Field lines on a log–log plot for seven different models, plus a limiting cylindrical one (dashed line). *Right:* Proper velocity ($U = \gamma V$) for each of the models as a function of height Z . Solutions (1) and (2) have infinite Lorentz factor at a finite height; solutions (3) to (7) are breeze solutions. Only the cylindrical solution is a plausible one. Reproduced from Fig. 5 of [538], by permission of the AAS.

Table 15.2: Lorentz factors for different jet-launching magnetizations

σ_M	γ_∞
1	2.42
5	6.47
10	11.30
20	20.85
50	49.37

Table 15.2 shows the final jet Lorentz factors for similar jet models with different magnetizations. We see that, to a rough approximation $\gamma_\infty \approx \sigma_M$ for these models. This result is similar to what was found for pulsar winds years earlier.

Energy-Loss and Mass-Loss Rates

LCB also computed an energy loss rate L_T for their cold relativistic MHD wind. We also can estimate a mass-loss rate if we assume that half of the energy loss goes into kinetic energy

$$\frac{1}{2} L_T \approx \gamma_\infty \Delta \dot{M} c^2$$

(which from Table 15.2 is approximately $\sigma_M \Delta \dot{M} c^2$). This gives us estimates that can be compared with equations (15.10)

$$\boxed{\begin{aligned} L_T &= \frac{F}{4\pi(F-1)} \frac{\Psi_{out}^2 \Omega_{out}}{R_{L,out}} \\ \Delta \dot{M} &\approx \frac{L_T}{2 \sigma_M c^2} \end{aligned}} \quad (15.29)$$

where Ψ_{out} and Ω_{out} have the same definitions as in the non-relativistic BP wind discussion, and the outer light cylinder radius is evaluated with the outer angular velocity

$$\boxed{R_{L,out} \equiv \frac{c}{\Omega_{out}}} \quad (15.30)$$

Comparing with the non-relativistic formulae, we see that the relativistic results are similar, if we take the effective launching radius to be the outer light cylinder radius $R_{L,out}$, not simply the outer disk radius $R_{0,max}$. And in the relativistic case the velocity at infinity can be very near the speed of light, unlike the non-relativistic case, where it could only reach of order the escape speed near the disk *inner* radius $R_{0,min}$. The higher the magnetization σ_M , the more relativistic the outflowing jet can be, and the less mass it will carry.

For the LCB standard model ($F = 1.0526$, $\sigma_M = 10$), the energy and mass-loss rates are

$$\begin{aligned} L_T &\approx 1.6 \frac{\Psi_{out}^2 \Omega_{out}}{R_{L,out}} \\ \Delta \dot{M} &\approx 0.080 \frac{\Psi_{out}^2}{R_{L,out}^3 \Omega_{out}} \end{aligned}$$

which again are similar to the non-relativistic case but with $R_{L,out}$ replacing $R_{0,max}$.

15.1.2.3 Relativistic Winds with Substantial Toroidal Field at the Base

The cold, centrifugally-driven wind theory assumes that there is very little toroidal field in the launching region ($R_{0,min} < R_0 < R_{0,max}$) and that B_ϕ only develops as the field lines are bent backward by the inertia of the outflowing, rotating wind. However, there is some reason to believe that disks and their coronae may have a

substantial, perhaps dominant, toroidal magnetic field.⁵ Independent of LCB, and only shortly thereafter, Contopoulos also developed a relativistic theory of MHD winds and jets, in which he emphasized a finite toroidal field near the wind base [539]. He concluded that current-carrying jets (ones with strong B_ϕ near the base, or $F > 1$) attain good collimation at large Z . However, only a very limited range of these solutions reached high Lorentz factors ($\gamma_\infty \gg 1$). These were ones with $F = 1 + \delta F$, where $\delta F \ll 1$, and again with a very specific value of μ . Also, unlike the non-relativistic solutions of Contopoulos and Lovelace, their relativistic counterparts did not experience radial oscillations. The oscillations were suppressed by two relativistic effects: the repulsive electric force and the greater inertia of the relativistically-moving plasma. Taken at face value, the results of LCB, Contopoulos and others seem to imply that *highly collimated, relativistic jets may require a rapidly-rotating poloidal magnetic field, but with very little initial toroidal component near the central engine*. If confirmed, this has strong implications for the structure of the launching region in accretion flows around black holes.

15.1.3 Non-Relativistic, Warm MHD Wind and Jet Theory

Cold MHD wind theory was useful for an initial understanding of jet production and collimation. However, it has many problems, both observational and theoretical. These problems have prompted theorists to improve the models significantly by taking plasma pressure forces into account – so-called warm MHD wind and jet theory.⁶

For example, the need to allow only certain jet energies μ (as in the relativistic cold jets), without any physical justification, seemed very contrived. Furthermore, in the mid-late 1990s investigators began to realize that the cold, axisymmetric, stationary MHD wind/jet models were *causally connected* throughout the smooth flow – from the black hole accretion disk to infinity – which was decidedly at odds with observations of actual jets. This revolution in thinking arose because of new insight into the nature and behavior of critical and singular surfaces and into the propagation of waves in the flow by multiple physical processes (plasma pressure, magnetic field, *etc.*). These new theoretical insights are not simple, but they are very important. *It can be argued that they are the very essence of jet production*, so we will spend some time discussing them in this subsection.

⁵ This is especially relevant in view of recent MRI disk simulations (Section 12.4).

⁶ By “warm” we mean winds where the plasma pressure is important, but not dominant. We reserve the term “hot” for the reverse situation (where the magnetic field is important but not dominant), as is the case in turbulent accretion flows, for example.

15.1.3.1 Observational Insight into Jet Acceleration Processes

Observations of actual relativistic jets also did not fit into the cold MHD wind picture of smooth, causally-connected flow:

- Very far from the black hole (on the VLA scale of $10^{4-5} \text{ kpc} \approx 10^{8-9} r_S$) AGN radio jets (particularly FR II sources) appear to be dominated by kinetic energy and plasma pressure, not Poynting flux and magnetic forces. This conclusion occurred early in the study of two-dimensional hydrodynamic and MHD jet flows [542, 543, 544]; see Section 15.2.
- Secondly, closer to the black hole (on the VLBI scale of $1-10 \text{ pc}$ or $10^{4-5} r_S$) observations of shocks moving through AGN jets had indicated that the magnetic field in the jet at that distance may be tangled, not the well-ordered structure that was predicted by cold MHD jet theory [545]. More recent polarization observations by Denise Gabuzda and her colleagues [546] have indicated that in FSRQs (whose parent objects are FR II sources) the magnetic field is parallel to the jet axis, while in BL Lac blazars (whose parent objects appear to be FR I sources) the magnetic field is normal to the jet axis. While possibly consistent with a lack of shocks in FSRQs, and a strong occurrence of them in BL Lacs, recently these results also have been interpreted as evidence for open, or tight, helical magnetic field structures, respectively [547]. Whatever the case, it is clear that, after traveling a distance of only a few tens of thousand Schwarzschild radii, the jet already knows whether it will be an FR II or an FR I.
- Thirdly, models of broad-band emission (radio through γ -ray) of AGN and microquasar jets [468] indicate that the objects consist of an inner jet flow that produces infrared through soft X-ray emission, a strong shock no more than $10^3 r_g$ from the black hole that accelerates particles to highly relativistic velocities and produces γ -rays, and a radio-emitting jet flow downstream of the shock. In the outer jet the radio emission clearly is produced not only by relativistic particles accelerated at the shock, but also ones accelerated *in situ* (at the site of the radio emission itself).

These observations were more consistent with a jet that becomes considerably modified at a finite (albeit rather large) distance ($\sim 10^{3-4} r_S$) from the black hole, transitioning to a flow dominated by kinetic energy and plasma pressure (and characterized by a possibly tangled and disrupted magnetic field), rather than remaining dominated by a well-ordered electromagnetic field throughout.

In order for the MHD theoretical ideas on jet acceleration and collimation *and* observations and analysis of actual jets both to be correct, the smooth, magnetically-dominated acceleration must be at least partially disrupted somewhere near $Z \geq 10^3 r_g$ so it can transition to a hot, tangled, turbulent, kinetic dominated flow. And, most importantly, this kinetic dominated flow must not affect or disturb the inner smooth MHD acceleration in the region near the black hole ($Z < 10^3 r_S$). That is, *the region where most of the acceleration takes place must be causally disconnected from the outer jet flow.*

15.1.3.2 New Theoretical Insights into Critical Surfaces, Singular Surfaces, and Causality in Warm MHD Flows

New theoretical insights into MHD wind theory were introduced to the field of astrophysics during the 1990s in papers by Sergei Bogovalov (at the RIKEN institute in Japan and the Moscow Institute of Engineering Physics), John Contopoulos, Christophe Sauty (of the Observatoire de Paris), Edoardo Trussoni (of the Osservatorio Astronomico di Torino), and Kanaris Tsinganos (of the University of Crete). These concepts also have been applied extensively in the analysis of solar and terrestrial plasma flows. (See, *e.g.*, [548].) They boil down to three important points:

- First, unless the flow remains normal to surfaces of symmetry (as in spherical inflow or outflow), singular surfaces do not coincide with critical surfaces (where flow characteristics appear or disappear). In fact, they *avoid* each other. For example, in purely hydrodynamic accelerating flow there are generally *two* important surfaces (one critical and one singular). And in a general accelerating MHD flow there are a total of *six* important surfaces (three critical and three singular). These numbers are potentially *doubled* when there are both outgoing and ingoing winds emanating from a single stagnation surface (see, *e.g.*, Fig. 15.15).
- Second, each (mathematically) singular surface also has a *physical* description and definition, independent of the precise structure of the mathematical wind equation. This means that *all* important surfaces can be discussed *independent of any specific symmetry assumptions being made* (*e.g.*, independent of the assumption of self-similarity).
- Third, the solution to the mathematical equation for this outflow *must* consider all three singular surfaces that correspond to the three physical separatrix surfaces. That is, one cannot assume that the slow magnetosonic speed is unimportant in the early part of the flow. And, *one cannot assume that the modified fast magnetosonic surface lies at infinity any more than one can assume that the sonic point in the solar wind lies at an infinite distance.*

The importance of these concepts, and how they relate to warm MHD winds, is best illustrated by presenting Bogovalov's concrete example of a general, warm MHD wind outflow from a magnetized star that is undergoing acceleration skew to the flow direction, *e.g.*, collimation (Fig. 15.7). The poloidal component of the outflow in this diagram occurs along the (dashed) poloidal magnetic field lines from the star surface (at left) toward interstellar space (at right). When the plasma is warm, the slow magnetosound speed V_S , and even the cusp speed V_c (Section 9.6.2), become important, and the fast magnetosound speed V_F (which also includes the gas sound speed) also is important. However, unless the plasma is very hot, and dominated by gas pressure, we will have the ordering $V_c < V_S < V_A < V_F$. Therefore, the flow begins at left in Fig. 15.7 below the cusp speed ($V < V_c$), passes through the intermediate ranges $V_c < V < V_S$, $V_S < V < V_A$ and $V_A < V < V_F$, and finally achieves supermagnetosonic speeds $V_F < V$.

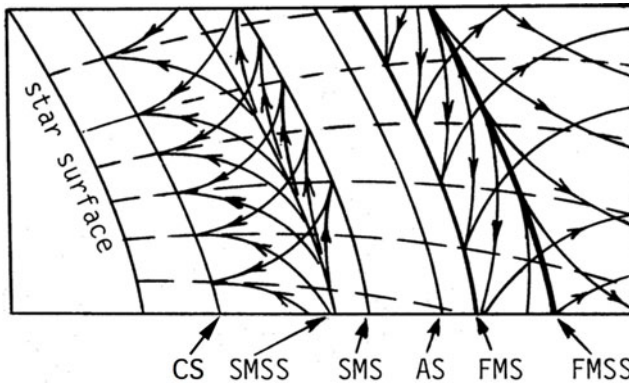


Fig. 15.7: Causal structure of an accelerating, collimating MHD flow from a stellar surface (left). Only the magneto-acoustic characteristics are shown (heavy solid arrows), along with the magnetic field/stream lines (dashed). Wind flow is toward the right. Also shown, and labeled, are the critical surfaces where characteristics appear or disappear (cusp surface (CS), and slow and fast magnetosonic surfaces (SMS and FMS)) and the separatrix surfaces where characteristics change sign along the flow (slow and fast magnetosonic separatrix surfaces (SMSS, FMSS), and the Alfvén surface (AS)). The Alfvén characteristics are not shown, but they point along the field/stream lines away from, and changing sign at, the AS. Adapted from Fig. 1c of [366]. Reproduced by permission of J. Wiley and Sons, Ltd.

General Features of Warm MHD Flow: Critical and Separatrix Surfaces

The general flow diagram in Fig. 15.7 is similar to the linearly accelerating one that was presented in Section 9.6.2, but with some crucial differences that are caused by the curvature in the streamlines.

Critical Surfaces. There still are three *critical* surfaces (where sets of characteristics appear or disappear) that occur when the *phase* velocity of a wave equals the flow speed:

- The cusp surface (CS, where $V = V_c$).
- The slow magnetosound surface (SMS, where $V = V_S$).
- The fast magnetosound surface (FMS, where $V = V_F$).

Between the stellar surface and the cusp surface ($V < V_c$) the elliptic flow is essentially in magnetohydrostatic equilibrium. Between the CS and SMS ($V_c < V < V_S$) the flow is slow-mode hyperbolic: slow-mode information can travel only in the direction of the characteristic arrows (sometimes only *upstream*!). Fast-mode MHD waves, of course, can travel in any direction in this region. Between the SMS and FMS ($V_S < V < V_F$) the flow remains fast-magnetosonically elliptic; fast-mode information continues to travel in any direction in this region. Finally, downstream of the FMS ($V_F < V$) the flow is now fast-mode hyperbolic. Yet, between the FMS and FMSS, fast-mode information can travel both up and downstream!

Separatrix Surfaces and the Sonic Horizon. However, whereas there was only one separatrix surface (where characteristic directions change sign) in the linearly accelerating case before, there now are three:

- The slow magnetosonic separatrix surface (SMSS), occurring in the slow hyperbolic region between the CS and SMS.
- The Alfvén surface (AS) as before, occurring in the trans-Alfvénic elliptic region between the SMS and FMS.
- The fast magnetosonic separatrix surface (FMSS), occurring in the fast hyperbolic region – well *beyond* the FMS.

As with the AS, the magneto-acoustic separatrix surfaces occur when the direction of information propagation along streamlines changes sign. The SMSS occurs when the slow-mode characteristics change from pointing only upstream to pointing both upstream and downstream. Similarly, the FMSS occurs when the fast-mode characteristics change from pointing both upstream and downstream to only downstream.

All separatrix surfaces have the property that conditions on those surfaces can affect properties of the flow not only downstream but also upstream. The FMSS is particularly interesting. Note that, while one edge of each Mach cone on that surface crosses the FMSS, the other edge is *tangent* to that surface. That is, the FMSS, not the FMS, is the *magnetosonic horizon*. Beyond the FMSS all Mach cones point downstream, whereas inside the FMSS one edge of each Mach cone points partially *upstream*. This behavior is identical to that of a black hole horizon, with the light speed replacing the fast magnetosound speed.

Singular Surfaces. The concept of a separatrix surface is a *physical* one, involving the physical properties of information flow and causality. However, when an actual equation for the MHD flow is derived, the separatrix surfaces become *singular* surfaces of that equation, where the denominator of the acceleration equation vanishes. And the process of specifying regularity conditions on those three singular surfaces (setting the numerator of that equation to zero) is simply specifying information at each separatrix surface to ensure that the flow is smooth throughout its run.

Characteristics and Causality in General Accelerating MHD Flow

What is the origin of the complex network of characteristics in Fig. 15.7? Fig. 15.8 shows a schematic of the Friedrichs diagram for this kind of curved-streamline MHD flow. Compare this figure with Figs. 9.10 and 9.11 in Section 9.6.2; the acceleration normal to the flow distorts the group velocity curves from their symmetry in linearly-accelerating flow. This distortion is directly responsible for the unusual variety of characteristics in the flow diagram, for the appearance of two additional separatrix surfaces, and for the critical and separatrix surfaces *not* coinciding.

Points A and B in the distorted and undistorted diagrams have identical characteristic behavior. Point A in both is a sub-cusp speed and therefore lies in the elliptic region of the flow diagram. Point B is super-cusp in both Figs. 9.11 and 15.8 and, in

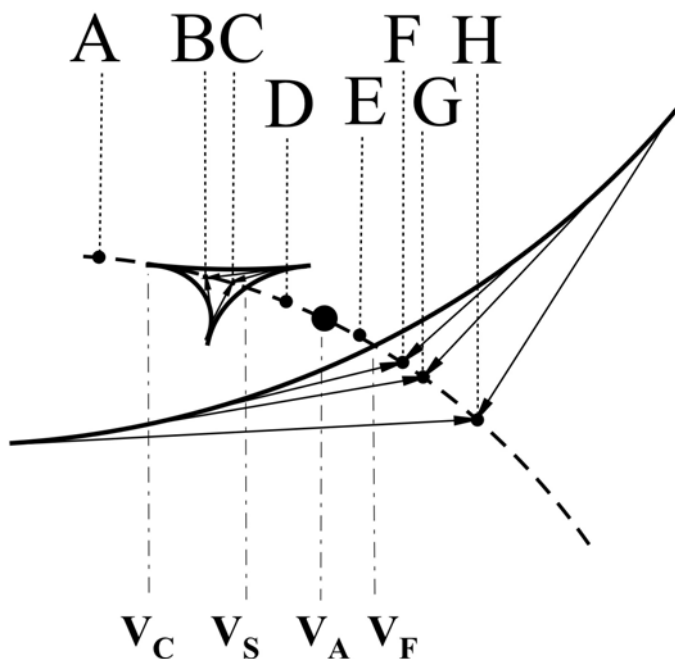


Fig. 15.8: Schematic MHD group polar diagram distorted by acceleration normal to the field/stream (dashed) line, *e.g.* by collimation. Similar to Fig. 9.10, it shows slow, fast, and Alfvén mode branches (thick solid lines and large spot) as well as places where different characteristic phase speeds (V_c , *etc.*) are achieved in the flow. Also shown are the (thin) characteristic arrows, for points A–H, that generate local Mach cones. In three cases (C, F, and G) the component of a characteristic along the streamline changes sign, indicating the formation of a separatrix surface: C and F have Mach cones pointing partially upstream and partially downstream; G has one edge normal to the streamline. B’s Mach cone (in the sub-slow region) points fully upstream and H’s (in the super-fast region) points fully downstream.

both cases, forms characteristics with the upper and lower trailing slow-mode group velocity branches.

The behavior of Point C, however, differs in the distorted diagram. It still is in the super-cusp region, but, because the lower part of the slow-mode group curve is compressed in Fig. 15.8, the lower characteristic jumps (discontinuously) from the trailing to the leading branch. It is this change in sign of information flow that forms the separatrix surface (SMSS) in the slow magnetosonic hyperbolic region in Fig. 15.7. The upper characteristic remains on the (stretched) upper trailing slow-mode branch – right up to the SMS critical surface.

There is no change in the behavior of Points D and E in the trans-Alfvénic region. The Alfvén characteristics still change sign at the Alfvén (separatrix) surface.

However, Points F and G in the fast magnetosonic hyperbolic region of Fig. 15.7 behave differently from their linearly accelerating counterparts. Because of the non-linear distortion of the group polar in Fig. 15.8, the upper fast characteristic can

have a component that transmits information *upstream*, even if that flow is at a speed greater than the fast magnetosound speed. This is decidedly different from Fig. 9.11, where the characteristic direction was normal to the velocity at the FMS and downstream beyond that. As the flow accelerates further, the orientation of this upper characteristic rotates to be increasingly more downstream. When the acceleration reaches Point G, the direction of information flow actually changes sign, again creating a separatrix surface, this time in the fast magnetosonic hyperbolic region.

The behavior of Point H in the distorted Fig. 15.8 is the same as in Fig. 9.11. In the flow region beyond the FMSS the flow forms characteristics that only transfer information downstream, causally disconnecting that flow with the earlier part.

Application to Warm, Self-Similar, Stationary, Axisymmetric MHD Winds and Jets

What do these new concepts imply for the stationary, axisymmetric MHD wind and jet models, especially for those that assume r -self-similarity? In the Parker and other θ -symmetric wind models the surfaces of symmetry were spherical shells, so both waves and flow were restricted to travel in the $e_{\hat{r}}$ direction. The r -self-similar models, on the other hand, are less restrictive and quite different. The surfaces of symmetry are cones centered on the Z axis, with their apexes at the origin ($r = 0$). In this case waves can travel *only* in the spherical-polar $e_{\hat{\theta}}$ direction. However, the plasma flow is still along field lines, which have components in both the $e_{\hat{r}}$ and $e_{\hat{\theta}}$ directions. So in the MHD wind/jet models the critical and the separatrix/singular surfaces still will be ones of symmetry (cones). However, because the flow may be accelerating normal to streamlines, and hopefully collimating, the two different sets of surfaces will not coincide. The flow should look something like Fig. 15.7, not Fig. 9.12.

In the case of warm r -self-similar winds the singular surfaces will be the familiar “modified fast surface” (the FMSS), often referred to as the “modified fast point” or MFP and the Alfvén surface, plus a new one in the sub-slow magnetosonic region, created by allowing a finite temperature plasma – the “modified slow surface” (the MSS or SMSS) or MSP. None of these surfaces can be ignored, because they are the places in the models where information must be specified in order to keep the solution and flow regular. The MFP is particularly important. It is where the jet flow becomes causally disconnected from the inner jet acceleration. That is, *the MFP (not the classical fast point) is the exit point of the jet engine nozzle.*

It is the set of regularity conditions on each singular surface that determines the character of the MHD flow, not requirements that the MFP lie at infinity or the MSP lie at zero distance. For example, Li, Chiueh, and Begelman found out that forcing the MSP to be at $Z = 0$ and the MFP to be at $Z \rightarrow \infty$ implied a unique value of the total energy per unit mass of the flow μ . However, μ (along with total angular momentum, mass flux, *etc.*) is set by conditions at the base of the outflow, not by where we place the MFP: μ could be quite different for different physical systems.

It is the state of the outflow near the black hole that will determine where the MFP will occur, not the other way around.

So, for physically acceptable (causal) solutions to the warm MHD flow equations, the MSP, Alfvén point, and the MFP all must be taken into account in the same manner as the sonic point is used in the Parker wind. The solution must pass through all three singular points; and the regularity conditions there (formed by setting the numerator to zero at those points) must be applied to each wind solution. In addition to respecting the physics of high-speed MHD flow, this also allows theoretical jets to become causally disconnected from the jet engine at large distances (say, $10^{3-7} r_S$). This feature is important, because the observations of actual jets are telling us that the outer jet flow is, indeed, causally disconnected from the acceleration process.

Enforcing one regularity condition in a singular flow equation is not trivial, however. Enforcing three is even more difficult. However, in the years 1998–2000 Nektarios Vlahakis (presently at the University of Athens) and his colleagues finally succeeded in constructing *non-relativistic, warm*, MHD wind/jet models that passed smoothly through all three singular points of the r -self-similar MHD wind equation. And, in the years 2001–2004, Vlahakis and Arieh Königl (of the University of Chicago) succeeded in constructing *relativistic, warm* MHD jet models with the same desirable properties. We now will discuss them below.

15.1.3.3 The Non-Relativistic, Warm, Self-Similar MHD Wind Equation

In warm MHD wind theory the counterpart to the Blandford and Payne 1982 work is a paper by Vlahakis, Tsinganos, Sauty, and Trussoni in the year 2000 (which we will refer to as VTST) [549]. This paper marks the first time that jet models with finite temperature were constructed that passed through all three singular surfaces. While their notation and terminology is slightly different from BP's, the concepts are sufficiently close that we will continue with the latter's in our discussions here, rather than create a new set of coordinates and parameters.

Only one new dimensionless parameter needs to be added to our discussion, the plasma parameter β_p . Since this parameter is not conserved along a streamline, we must choose a reference value along each magnetic field line. Therefore, VTST chose the value at the Alfvén point

$$\beta_{pA} \equiv \frac{p_A}{B_A^2/8\pi} \quad (15.31)$$

In cold MHD theory $\beta_{pA} = 0$, so this parameter measures the “warmth” of the plasma wind.

The equation that VTST derived for warm MHD winds has a structure similar to equation (15.5)

$$\frac{d\mathcal{M}_{W,NR}}{d\theta} = \frac{\mathcal{N}_{W,NR}}{\mathcal{D}_{W,NR}} \quad (15.32)$$

except now the polar angle θ is the independent coordinate instead of Z or R . Both the numerator and denominator of this equation are too complex for the current discussion, but the denominator does vanish at the Alfvén point

$$V_p^2 = V_A^2 \quad (15.33)$$

and also when the following condition is met

$$V_\theta^4 - V_\theta^2 c_{ms}^2 + c_s^2 V_{A,\theta}^2 = 0 \quad (15.34)$$

Comparing with equations (9.187) and (9.188), we find that this equation is satisfied when $V_\theta^2 = V_S^2$ or $V_\theta^2 = V_F^2$, *i.e.*, at the non-relativistic *modified* slow/fast points.

If the modeling were to proceed in the same manner as Blandford and Payne did, we would begin by first settling on a type of gas being simulated (*e.g.*, $\Gamma = 5/3$) and specifying that the wind is being driven by a Keplerian, Shakura–Sunyaev disk ($F = 3/4$ and $\Omega \propto x^{-3/2}$). Then, we would attempt to solve for a wind/jet model for a given rate of mass loading κ , dimensionless angular momentum λ , and plasma β_{pA} at the Alfvén point – three free parameters of the flow. The MHD equations then would be solved along a single “reference” magnetic field/stream line, such that the solution passed through all three singular points. In the process, the satisfying of these three regular conditions would give us three features of the flow: the position of the Alfvén surface θ_A , the value of the acceleration there ($a_A = d\mathcal{M}_{W,NR}/d\theta|_A$), and the angle $\psi_A = \tan^{-1} x'_A$ that the magnetic field makes in the (R, Z) -plane with the Z axis. In this manner the regularity conditions would determine the geometry of the flow solution but not constrain the fundamental (conserved) parameters of the problem.

Vlahakis *et al.*, however, took a slightly different approach. They chose to specify the parameters κ , θ_A , and ψ_A , and then to *derive* the values of λ and β_{pA} (and, of course, a_A) that were implied by their choice of mass loading and wind warmth. Beginning at $\theta = \theta_A$, where the solution was forced to pass through the Alfvén point, VTST then had two quantities they could adjust in order to obtain a solution that passed through the modified slow and fast points: λ itself and the acceleration a_A (which ultimately determined β_{pA}). This procedure still allows the entire $(\kappa, \lambda, \beta_{pA})$ parameter space to be explored (by repeatedly choosing different values of θ_A and ψ_A). It does *not* restrict the regular solutions to, for example, a few special energies μ , as was the case in previous treatments of the cold MHD winds.

15.1.3.4 Solutions and Simulations of Warm MHD Winds and Jets

Self-Similar Solutions

The Warm MHD Wind Flow Diagram. Fig. 15.9 shows a typical solution for a warm MHD wind/jet that is comparable to Fig. 15.2 from Blandford and Payne’s version for the cold, non-relativistic MHD wind. The heavy solid black lines that curve upward are the same as before: magnetic field/stream lines projected into the

(R, Z) -plane. Unlike BP's solution, however, these streamlines re-focus toward the Z axis at high latitudes, particularly beyond the modified fast surface ($\mathcal{M}_{mf} = 1$).

The elliptic and hyperbolic regions of the flow are denoted by gray-shaded and white areas, respectively. Compare this actual jet model with Bogovalov's schematic diagram for general accelerated MHD flow in Fig. 15.7. Topologically, the two diagrams are the same, with the same set and ordering of critical and separatrix surfaces. The critical surfaces that mark the boundaries between the two types of flow are the cusp ($\mathcal{M}_c = 1$), classical slow magnetosonic ($\mathcal{M}_s = 1$), and classical fast ($\mathcal{M}_f = 1$) surfaces. Two of the separatrix/singular surfaces, the modified slow ($\mathcal{M}_{ms} = 1$) and modified fast ones, are shown, but the Alfvén surface is not. It occurs inside (and roughly bisects) the elliptic flow region between the classical slow and fast surfaces. Note in particular how the streamlines smoothly cross all three singular surfaces, including the modified fast one near the polar axis.

The light gray lines show the characteristics that remain continuous throughout the hyperbolic regions, such as the outward flowing ones that begin at the classical fast surface ($\mathcal{M}_f = 1$) and extend to infinity. (The arrows have been added by the present author.) These eventually cross the modified fast surface. The darker gray lines show the characteristics that form the separatrix surfaces, such as those that are tangent at the modified fast surface at left and terminate at the classical fast surface. It is this information flow that keeps the region between the classical and modified fast surfaces causally connected to the acceleration. Information from the modified fast surface can affect all parts of the classical fast surface, and that surface can affect the entire (disk) region where the jet is initially accelerated.

In the narrow conical region on the other side of the modified fast (separatrix) surface, the dark gray characteristics are outgoing and appear to be radial lines. They are not truly radial, however; they curve slightly toward the polar axis, in an effort to become parallel to the streamlines. In that conical region the jet is causally *disconnected* from the acceleration region at lower latitudes. This is important, because the refocusing of the streamlines in this region may create effects (shocks, turbulence, *etc.*) that otherwise might disrupt the structure of the jet engine outside that cone.

Causality and Information Flow in Warm MHD Winds from Black Holes. How does the FMSS, instead of the FMS, tell the rest of the accelerating flow that it is to become the magnetosonic horizon? The answer is in figures like Fig. 15.7 and Fig. 15.9. The flow beyond the FMS still is causally connected to it via the characteristics that arch back and intersect the FMS. So, changes in the flow that occur on a part of the FMSS can be communicated (via magnetosonic waves) back to the outer fast magnetosonic surface. From there information can be communicated back to the section of the engine that created that part of the FMSS via fast waves in the elliptic flow region below the FMS. This communication of information will take some time – on the order of 10–20 magnetosound crossing times – but the flow eventually *will* set up a sonic horizon that lies much closer to the polar axis than the classical fast surface. This entire structure up to the FMSS should be thought of as the jet engine and the narrow cone is its nozzle.

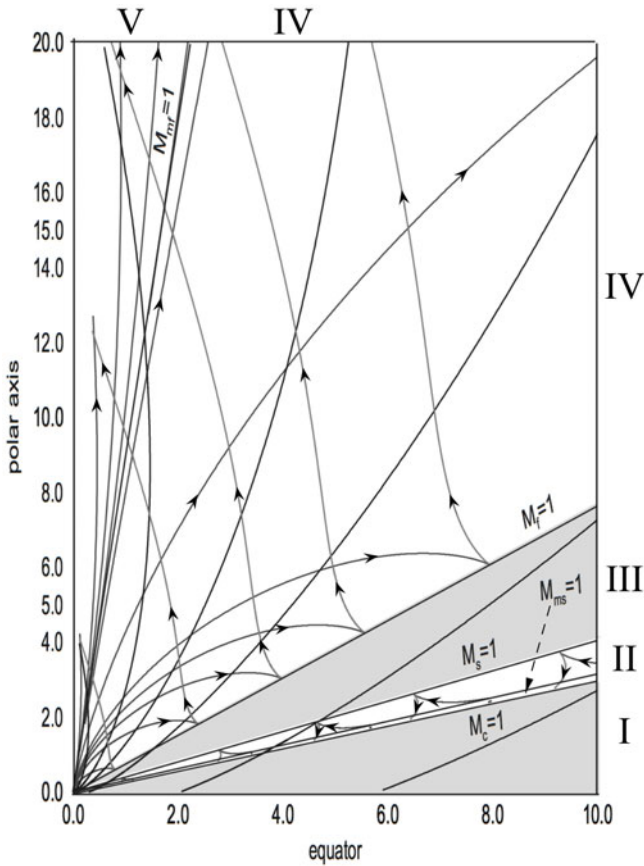


Fig. 15.9: Causality diagram for a radially self-similar jet model. Compare with the schematic in Fig. 15.7. Dark black upwardly curved lines (without arrows) are the magnetic field/stream lines; gray lines (many with arrows) show the characteristics. Beginning at the equator we have the following regions. (I) Hydrostatic sub-cusp region (shaded, bounded by the equator and the CS [$\mathcal{M}_c = 1$]), which is fully causally connected with slow-mode and other waves. (II) Super-cusp/sub-slow region (unshaded wedge, bounded by the CS and SMS [$\mathcal{M}_s = 1$]); the SMSS ($\mathcal{M}_{ms} = 1$, dashed arrow) lies in this region and is barely visible near the CS; slow-mode information travels up and downstream between the SMSS and SMS, but only upstream from the SMSS; however, the region still is fully connected with fast-mode waves. (III) Super-slow/sub-fast region (shaded, bounded by the SMS and FMS [$\mathcal{M}_f = 1$]); the AS ($\mathcal{M}_{W, NR} = 1$, not shown) lies in this region. (IV) Causally-connected, super-fast region (unshaded, bounded by the FMS and FMSS [$\mathcal{M}_{mf} = 1$]); although the fluid speed exceeds the fast-mode speed V_F here, the FMSS nevertheless can affect the FMS via the rightward arching characteristics that connect both; because everything below the FMS also is causally connected, all parts of the FMSS can affect all the structure of the jet engine! (V) Causally-disconnected, super-fast region (unshaded, bounded by the FMSS and polar axis); no characteristics from this region reach the rest of the structure; shocks and other disruptions can occur in this conically-shaped region and not affect the acceleration and collimation of the jet. Adapted from Fig. 4 of [549]. Reproduced by permission of J. Wiley and Sons, Ltd.

How much time is needed for the engine structure to settle down? In a black hole engine the magnetosound speed will be close to the speed of light, say $0.5c$. If the principal FMSS/MFP occurs at, say, $1000 r_g$, then 20 magnetosound crossing times would give a time of

$$\tau_{equ} \approx 20 \frac{1000 r_g}{0.5 c} = 2 s \frac{M}{10 M_\odot}$$

For GRBs and microquasars only a few seconds are required to set up the jet nozzle – plenty of time except, perhaps, for the short-duration GRBs. In the case of a $10^7 M_\odot$ Seyfert black hole, the engine would need less than a month to form, while a large $10^9 M_\odot$ quasar may need half a decade. However, these conservative estimates are sufficiently short, compared to the source lifetimes, that we can expect a full FMSS/MFP to form in most black hole jet engines. And if the principal MFP occurred closer to the black hole, say at only $100 r_g$, then the required time would be 10 times shorter.

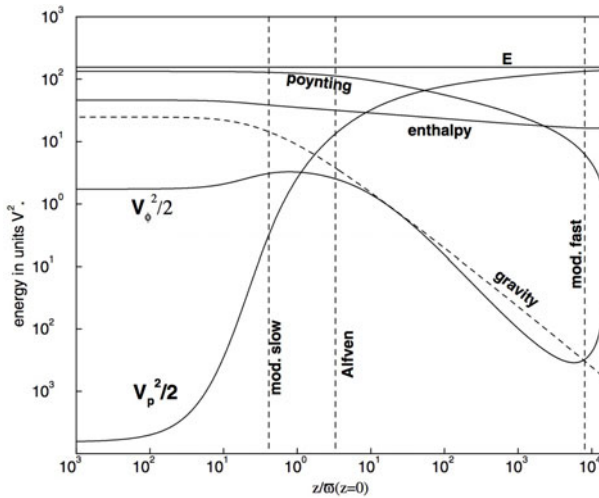


Fig. 15.10: Plot of components of the conserved total energy E per unit mass (*i.e.*, $\mu - 1$) along a field line for a model similar to that in Fig. 15.9 vs. height Z (in units of the field line radius at the equator [$Z = 0$]). Note how Poynting energy is converted into poloidal kinetic energy. Rotational energy is never very important, and gravity becomes unimportant beyond the modified slow point. The enthalpy remains moderately warm, but never exceeds either the Poynting or kinetic energy. Reproduced from Fig. 7 of [549], by permission of J. Wiley and Sons, Ltd.

Energy Conversion in the Warm MHD Wind. Fig. 15.10 shows how the different types of energy change along a stream line for a jet model (that is similar to, but not exactly the same as, the model in Fig. 15.9) vs. the cylindrical axial coordinate Z in units of the radius of the field anchor point R_0 in the accretion disk. The

vertical dashed lines in the figure show the locations of the singular points where streamlines cross the three singular surfaces. As in Blandford and Payne's models, gravitational effects *are* taken into account. The horizontal line at the top is the Bernoulli constant that is conserved in the flow. As expected in an MHD jet, the energy flow is dominated initially by Poynting (electromagnetic energy) flux. The thermal component (enthalpy) is much less, although still substantial and important in the early stages of the outflow. Acceleration up to the modified slow point actually is by gas pressure, just like the Parker wind. Between the MSP and Alfvén point acceleration is primarily by the centrifugal effects of the rotating magnetic field (an Alfvén wave). From the Alfvén point to the MFP (at $Z/R_0 \sim 20$) the acceleration is caused initially by the gradient of toroidal magnetic pressure (dB_ϕ^2/dZ) and then by the increasingly constricting hoop stress (B_ϕ^2/R), both fast magnetosonic effects. Beyond $Z/R_0 \sim 100$ the flow switches from being Poynting-flux-dominated (PFD) to kinetic-flux-dominated (KFD). And, beyond the MFP the rapid pinching of the jet causes the rotational velocity of the jet to rapidly increase, due to angular momentum conservation.

Non-Self-Similar Simulations

Numerical simulations that illustrate many of VTST's points about warm winds actually were published not only before *their* work but also 7 months before Krasnopolsky *et al.*'s similar work on *cold* MHD winds! The study was performed by Galina Ustyugova and her colleagues at the Russian Academy of Sciences in Moscow and with Marina Romanova and Richard Lovelace at Cornell, the same group to first perform MHD wind simulations in 1995–1999 [511, 550]. Unlike other studies, the Russian group's outflowing winds began with substantial gas pressure ($c_s \approx V_A/3$), and these authors paid particular attention to the sonic causality (Mach cones and characteristics) at the boundaries of the simulation region.

Fig. 15.11 shows an R – Z -plane with poloidal magnetic field lines and velocity vectors. Three surfaces also are shown: two critical surfaces (the FMS (upper dashed line) and the SMS (barely visible lower dashed line)) and the Alfvén separatrix surface (middle dashed line). (Note that the cusp surface, SMSS, and FMSS are not shown, however.) Also shown are the Mach cones on the outflow boundaries. If the flow were hyperbolic everywhere on these boundaries, and if the Mach cones all were pointed outward, then this simulation would completely describe the final steady state MHD outflow. However, we will see that this is not the case.

Qualitatively this diagram compares well with the self-similar one in Fig. 15.9 and with Fig. 15.2 (although with the addition of a slow surface). However, some features in this simulation are different from the self-similar model. The critical and separatrix surfaces are not cones centered on the Z axis. Instead the classical fast and Alfvén surfaces are roughly horizontal. This means that the numerical simula-

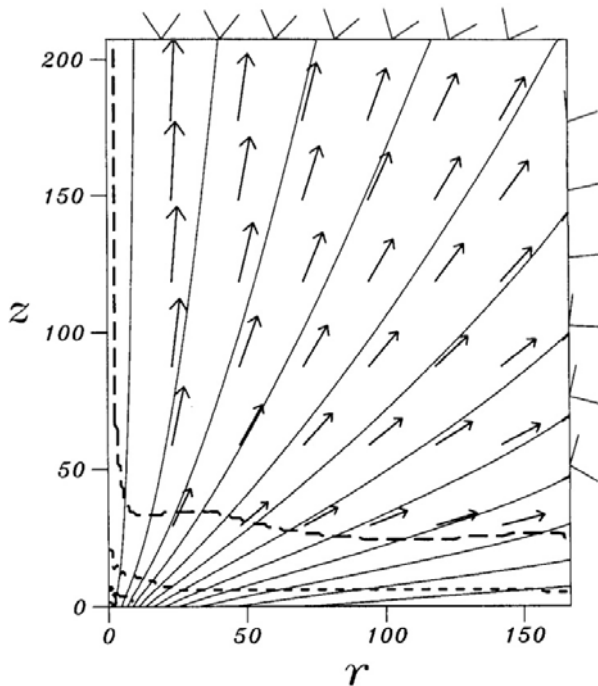


Fig. 15.11: Magnetic field/stream diagram from a non-self-similar simulation of MHD outflow from a geometrically thin disk (on the equator). Surfaces shown by the dashed lines are: SMS (barely visible at lower left), AS (short dashes), and FMS (long dashes); Mach cones are shown on the outer boundaries of the flow. Note the Mach cone at lower right, which presumably intersects the sub-fast-magnetosonic region below the long dashed line. Information, therefore, can travel from the outflowing wind beyond the FMS back into the sub-fast region. So, the jet engine is still causally connected to at least part of the outflow. Adapted from Fig. 7 of [550]. Reproduced by permission of the AAS.

tion does not produce r -self-similar flow.⁷ Nevertheless, the characteristic structure is qualitatively very similar to that in Fig. 15.9. The Mach cones *widen* as one approaches lower latitudes (see lower right in Fig. 15.11). And the right-hand characteristic of the Mach cone eventually intersects the classical fast surface, just as it does in Fig. 15.9. Furthermore, recall that the flow inside the classical fast surface (below roughly $Z \sim 25$) is *fully causally connected* by fast magnetosonic waves. Therefore, just as in Fig. 15.9, because the super-magnetosonic flow still can influence the classical fast surface, and because the sub-fast region is fully causally connected, many (if not all) parts of the super-fast region still can causally affect the jet acceleration region. Furthermore, this influence can propagate from *beyond the*

⁷ Actually, non-self-similarity was to be expected, since the magnetic field boundary conditions in the thin accretion disk on the equator are not self-similar: *i.e.*, B_z/B_R is not a constant.

right-hand boundary back into the region where Mach cones do not exist ($Z \lesssim 25$).⁸

Unfortunately, therefore, these simulations cannot be used to judge whether an analogy to the modified fast surface (*i.e.*, the FMSS) can form or not. Despite the outward-pointing Mach cones, outflow beyond the boundary still can influence the acceleration inside the computational region. The simulation has to have a much larger computational region for all characteristics intersecting the FMS to be handled properly, *and* it has to be evolved for a time long enough that a true FMSS would be able to set itself up.

15.1.4 Relativistic, Warm MHD Wind and Jet Theory

15.1.4.1 The Relativistic, Warm MHD Wind Equation and its Singular Surfaces

The warm MHD counterpart to the LCB and Contopoulos papers on relativistic jet models is contained in a series of papers by Vlahakis and Königl in the early 2000s, particularly a paper published in late 2003 [368]. The derivation of the wind equation begins with the relativistic MHD equations (Appendix F). This time the effects of the specific enthalpy $\mathfrak{h} = (\varepsilon + p)/\rho$ are taken into account. Vlahakis and Königl (VK), of course, also assumed r -self-similarity, and, like LCB, they ignored gravity. This again is done because, in the relativistic case, one cannot examine models with a Keplerian disk, since relativistic flows require $\Omega \propto R^{-1}$ on the equator.⁹ However, the run of magnetic field with R along $Z = 0$ (*i.e.*, the parameter F) still can be freely specified, just as in LCB's cold MHD treatment.

The warm relativistic MHD wind equation has the same form as equations (15.22) and (15.32), but now the proper Alfvén speed contains relativistic enthalpy terms as well

$$U_A = \frac{B_p^2 (1 - x^2)}{4\pi \rho (1 + \mathfrak{h}/c^2)} \quad (15.35)$$

where, again, $x = R/R_L$ is the cylindrical distance of a point on the magnetic field line in units of the light cylinder radius and $B_p^2 = B_R^2 + B_Z^2 = B_r^2 + B_\theta^2$. The numerator and denominator of the warm, relativistic MHD wind equation ($\mathcal{N}_{W,R}$ and $\mathcal{D}_{W,R}$) were too complex for even VK to publish in their paper. However, we still

⁸ This result also is true for the spherical-polar simulations performed by Ustyugova *et al.* in the same paper.

⁹ Actually, VK developed a clever scheme to place the origin of the MHD self-similarity away from the black hole, but still on the Z axis. This allowed one to choose an arbitrary disk angular velocity Ω (see their Fig. 2). However, because this adds another free parameter (the height of the disk equator), and because VK ignored gravity in the end anyway, this rather clever coordinate system is seldom used in jet studies.

can describe here the important points of the equation that they solved. In particular, the denominator again vanishes

- At the relativistic (modified) Alfvén surface, where

$$(\gamma V_\theta)^2 = U_{A,\theta}^2 \equiv \frac{B_\theta^2 (1 - x^2)}{4\pi \rho (1 + \mathfrak{h}/c^2)} \quad (15.36)$$

- And at the relativistic modified slow and fast surfaces

$$(\gamma V_\theta)^4 - (\gamma V_\theta)^2 U_{ms}^2 + U_s^2 U_{A,\theta}^2 = 0 \quad (15.37)$$

where

$$U_s^2 = \frac{c_{s,R}^2}{(1 - c_{s,R}^2/c^2)}$$

is the proper gas sound speed (*not* the slow magnetosonic speed),

$$U_{ms}^2 = U_s^2 + \frac{B_p^2 (1 - x^2) + B_\phi^2}{4\pi \rho (1 + \mathfrak{h}/c^2)}$$

is the proper magnetosound speed, and

$$c_{s,R}^2 = \frac{\Gamma p}{\rho (1 + \mathfrak{h}/c^2)}$$

is the relativistic adiabatic sound speed (equation (9.195)), which also includes relativistic enthalpy terms. The roots of equation (15.37) are, of course, the modified slow (−) and fast (+) speeds in the rest frame of the black hole

$$(\gamma V_\theta)^2 = \frac{1}{2} \left\{ U_{ms}^2 \pm \left[U_{ms}^4 - 4 U_s^2 U_{A,\theta}^2 \right]^{1/2} \right\}$$

The reader may notice that the proper velocities U_A , U_s , and U_{ms} are *not* the same quantities that we derived in the discussion of basic relativistic MHD waves (equations (9.192), (9.195), and (9.196) in Section 9.6.2). There are two reasons for this. First of all, those quantities were derived in the *rest frame* of the moving fluid. Here they have been transformed to the “lab” frame of the black hole, *e.g.*, Boyer–Lindquist coordinates, and expressed as proper velocities (U) rather than V . This accounts for the magnetic field being absent in the denominator of the Alfvén speed here (equation (15.35)) and for the U ’s not being limited to the speed of light. It also accounts for the $(1 - x^2)$ factor, which zeroes out the contribution of the poloidal magnetic field at the light cylinder.

Secondly, axisymmetry has been assumed here, so the expression for the Alfvén speed again does not include the magnetic azimuthal component B_ϕ . Indeed, when radial self-similarity also is assumed, the Alfvén speed is reduced to the single θ component (equation (15.36)), because transverse magnetic field perturbations are

not allowed in the spherical r direction either. However, because magneto-*acoustic* waves are compressive, the magnetosound speed still does include B_ϕ .

Equations (15.36) and (15.37) give us the singular surfaces for all types of self-similar MHD jet outflows: warm, cold, relativistic, or non-relativistic. For example, if U_s were to vanish (cold limit), the lower root of equation (15.37) would be $V_\theta^2 = 0$, placing the modified slow surface at the beginning of the outflow. And the modified fast surface would occur at the same location as LCB's (*cf.*, equation (15.25)). Furthermore, these equations reduce to the warm (and cold) non-relativistic results as well: since $\mathbf{V}_p$ and $\mathbf{B}_p$ are parallel, equation (15.36) implies that $(\gamma V)^2 = U_A^2$, which reduces to equation (15.33) when $\gamma \rightarrow 1$ and $x \rightarrow 0$. Likewise, equation (15.37) reduces to equation (15.34) in the non-relativistic limit.

15.1.4.2 Solutions

Conceptually, the modeling begins as in the previous cases, by choosing a type of gas (usually $\Gamma = 4/3$ for a relativistic plasma) and the magnetic flux gradient (usually $F \approx 1$, plus or minus a small value in the range 0.001–0.01). However, the main model parameters for a given magnetic field line are a little different from the VTST non-relativistic case and more similar to those of LCB: the field line angular velocity Ω_f ; the magnetization parameter σ_M (instead of κ ; see equation (15.28)), here again *without* the relativistic entropy factor $\mathfrak{h}$; the position of the Alfvén point in light cylinder units x_{Alf} (instead of λ in the non-relativistic models); and a dimensionless parameter indicating the warmth of the flow

$$\mathcal{Q} \equiv \frac{k^2}{4\pi} \left[\frac{\Gamma}{(\Gamma - 1)} \frac{K_\Gamma}{c^2} \right]^{1/(\Gamma - 1)}$$

(instead of the plasma β_{pA} at the Alfvén point), where $K_\Gamma = p/\rho^\Gamma$ is the coefficient of the adiabatic equation of state. The cold approximation is $K_\Gamma = 0$, or $\mathcal{Q} = 0$.

As before, the solution should determine geometric properties of the solution, such as the angle of the Alfvén surface θ_A and the angle of the field line there ψ_A . In practice, however, θ_A and ψ_A may be specified, along with σ_M and x_{Alf} , and then parameters like $\mathcal{Q}$ and Ω_f (or even μ) may be adjusted until all three singular points are crossed smoothly.

One important point deserves mentioning, however. Despite the extra effort made to treat the modified fast point properly in the equations and despite the considerable amount of discussion on causality in MHD winds/jets, in the end Vlahakis and Königl decided to accept only solutions whose flow was asymptotically cylindrical. That is, all their 'trans-Alfvénic' solutions placed the final critical surface (FMSS) at infinity, just as BP and LCB had, and *not* at a finite distance as Vlahakis and his colleagues had done earlier with non-relativistic, warm winds. Placing the MFP at infinity is, in fact, a solution of $\mathcal{D}_{W,R} = 0$, as well as other finite distances, and it corresponds to a unique value of μ , just as it did in the LCB cold wind theory. Vlahakis and Königl's argument for doing this was that recollimating jet flows (and de-

celerating “breeze” solutions) were physically unacceptable solutions. This may be the case for certain types of sources (*e.g.*, FR I jets), but recollimating flows may be the solution of choice for FR II sources (see Section 15.2) and microquasars, whose spectra have been shown by Sera Markoff and her colleagues to contain evidence for a strong, particle-accelerating shock far from the black hole. Also, decelerating, failed jet solutions may be relevant in very low-luminosity AGN like Sgr A*.

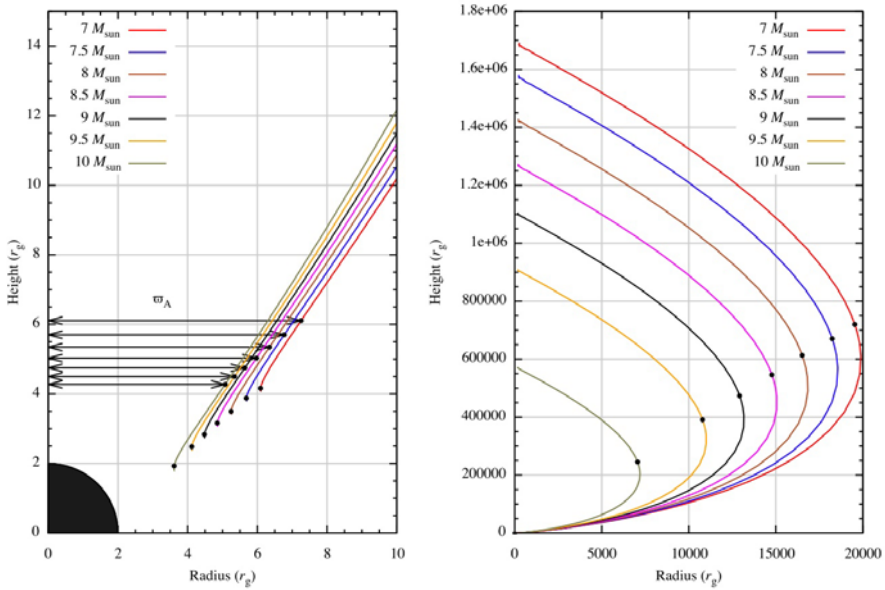


Fig. 15.12: Magnetic field/stream lines on two very different linear scales for several models of warm MHD winds/jets from black holes. *Left:* Region very near the black hole for models with the same *linear* Alfvén point radius, but different black hole masses, showing positions of both the Alfvén point (rightmost arrowhead) and MSP (near innermost solution point) in each model. *R* and *Z* axes have the same linear scale, and a Paczyński–Wiita potential was used in these models to handle gravity near the event horizon. *Right:* Same field/stream lines plotted on a scale more than 10^5 times larger (in height *Z*), with colors corresponding to those in the left-hand plot and MFP points marked. The jets begin to recollimate before the magnetosonic horizon and then converge onto the polar axis, where a strong shock may form. Note that the vertical scale is about a factor of 50 larger than the radial scale: *i.e.*, these jets are very narrow and highly collimated. Figure kindly supplied by P. Polko; reproduced with permission.

Recently Peter Polko and Sera Markoff (both of the University of Amsterdam) and D. Meier have been solving the Vlahakis and Königl equations with both the MFP and MSP singular points in mind [551, 552]. In order to obtain physically plausible solutions for the modified slow point, these authors also added a gravity

term to the equations.¹⁰ They found relativistic solutions with the MSP very close to the jet origin and the MFP at a very large, but finite, distance from the black hole.

Fig. 15.12 shows a representative field line for each of several *different* jet solutions (not simply self-similar shiftings of the same field line). (Note the linear distance scales on the two plots in the figure. The axes in the left plot have the same scale and show structure interior to $20 r_g$. The scale of the horizontal axis of the right-hand plot is over 50 times smaller than that of its vertical axis. In addition, the scale of the vertical axis in that plot is 10^5 times larger than that of the left-hand one.) The red curve, with a black hole mass of $7 M_\odot$, is typical of the jet physics (even though this particular solution may not fit data from any actual black hole source). The slow magnetosonic separatrix surface (MSP in these models) lies near the black hole and in the rotating plasma inflow that drives the jet. The Alfvén surface lies not that much farther away from the black hole and is consistent with MHD simulations of jet outflows from disks (see Fig. 15.3 and Fig. 15.11). However, the fast magnetosonic separatrix surface (MFP here) lies at great distances from the black hole, and occurs at a place where the jet begins to recollimate. This recollimation eventually leads to the flow colliding with itself on the axis, where (depending on the strength of the collision) either a strong bounce or shock will occur. This shock is reminiscent of the shock that Markoff needed in X-ray binaries in order to fit the radio spectrum (Section 13.3.1). Note that this dramatic change in the character of the flow will not affect the “jet engine” that launches the flow: the FMSS is the magnetosonic horizon, so beyond that surface the flow is causally disconnected from, and therefore cannot affect, the jet base.

These jet solutions have at least six free parameters:

the warmth of the flow (Q), the magnetization (σ_M ; eq. (15.28)), the dimensionless Alfvén radius in units of the light cylinder radius ($x_{\text{Alf}} = \varpi_A/R_L$), the light cylinder radius itself in units of r_g , the scaling of physical parameters like magnetic field with cylindrical radius (*i.e.*, an accretion disk model), and the mass of the black hole $M_\bullet$. Fig. 15.13 shows how the heights of the three singular points vary with two of these, Q and σ_M . Other things being equal, as the flow becomes hotter, the modified slow point (related to the gas pressure sonic point) rises toward the height of the Alfvén point. Similarly, as the magnetization increases, the Alfvén point moves higher than the MSP, and the MFP increases dramatically – by an order of magnitude for only a modest increase in magnetization. By adjusting the values of the input parameters, and by fitting a predicted radio/X-ray spectrum to a given source, a physical model for that source can be developed, just as has been done using thermally-driven jets (Section 13.3).

¹⁰ The addition of a gravity term is not strictly a self-similar step. However, the equations are self-similar in the non-relativistic regime, as they become the equations of Vlahakis *et al.* [549] discussed above. Well above the disk gravity becomes negligible, and the equations then become the self-similar relativistic equations of Vlahakis and Königl [368]. One can think of this process as matching a relativistic solution without gravity, which extends to large distances from the black hole, onto a non-relativistic solution *with* gravity, which extends down into the disk and launching point near the black hole. In this case, however, the entire solution is continuous and has continuous derivatives.

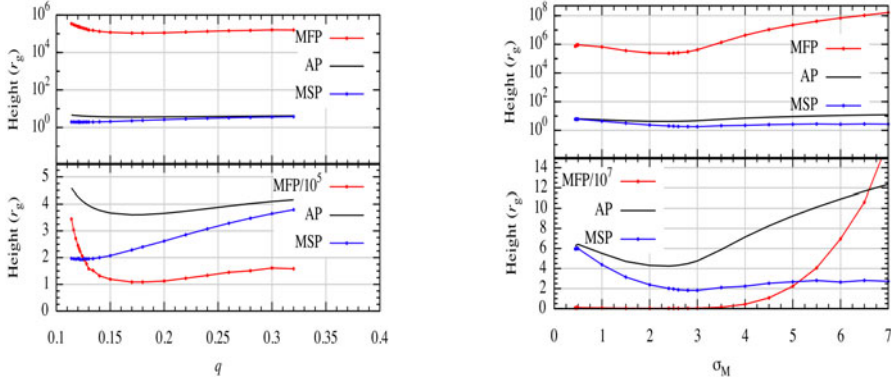


Fig. 15.13: Comparison of positions of the MSP (blue), AP (gray), and MFP (red) for different models of warm, self-similar MHD winds/jets from black holes, using both logarithmic (top) and linear (bottom) scales. *Left:* Different models varying only $\mathcal{Q}$ ($= q$ here) of the jet at its base. For cold flows, the MSP approaches zero height, while the MFP increases in height. For very warm flows, the MSP approaches the Alfvén point and the MFP values level off. *Right:* Different models varying only the magnetization parameter σ_M . Low magnetization flows appear similar to very warm flows at left, while high magnetization ones appear to resemble cold flows. Figure kindly supplied by P. Polko; reproduced with permission.

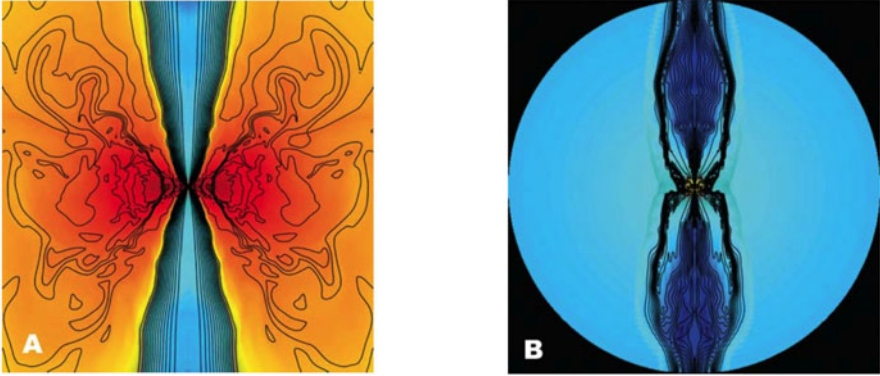


Fig. 15.14: Plots of rest mass density $\log \rho/\rho_{\text{disk}}$ and magnetic field lines for a large, two-dimensional, axisymmetric GRMHD simulation of an MRI turbulent disk that launches a large-scale jet from the vicinity of a $j = 0.9375$ black hole. (A) The inner $100 r_g$ at the final time ($t = 1.4 \times 10^4 \tau_g$), showing both high density (red is $\sim 2 \rho_{\text{disk}}$) and tangled magnetic field. Note the very geometrically thick disk and corona, the continued turbulence, and the well-formed axial jet. (B) Same as (A), but for the full simulation in the spherical volume $r \leq 10^4 r_g$. Note the recollimation of the jet that occurs near the boundaries. However, a full MFP has not yet formed; this is not expected to happen until several fast magnetosound crossing times or $\sim 4\text{--}8 \times 10^4 \tau_g$. Adapted from Figs. 2b and 1b of [457]. Reproduced by permission of J. Wiley and Sons, Ltd.

15.1.4.3 Simulations of Warm, Relativistic MHD Winds and Jets

So far, there does not appear to be *any* simulation of relativistic MHD outflows that unequivocally forms a MFP/FMSS nozzle and a KFD super-magnetosonic jet outflow. One simulation that might be relevant, however, was performed by Jon McKinney (then at the Center for Astrophysics at Harvard) [457]. The simulation actually was that of an ADAF-like, MRI-driven accretion flow, with the potential for MHD wind outflow. As mentioned in Chapter 12, these MHD simulations form PFD jet flows similar to those in the inner parts of the self-similar jet models. McKinney's simulation was fully relativistic in both the MHD flow and the gravitational potential of the accreting black hole. McKinney, however, did something other investigators have not done: he let the simulation, and the jet production, continue for thousands of black hole light crossing times ($\tau_g = GM/c^3$). Fig. 15.14A shows the inner density and magnetic field structure of the accretion flow and the highly-collimated PFD jet already well formed by $Z < 100 r_g$. Fig. 15.14B shows the same plot for the outer region (up to $Z \sim 10^4 r_g$). The jet expands laterally for much of its length, presumably because the external confining pressure decreases as $p \propto r^{-5/2}$ in this model. However, the jet is re-focused beyond $\sim 3000 r_g$ by the magnetic hoop stress. Note the oscillatory nature of the poloidal magnetic field lines there.

Fig. 15.15 shows the northern hemisphere of Fig. 15.14 on a log–log plot. Illustrated here, in one diagram, is much of the physics that we have discussed in this chapter. The red lines are the poloidal magnetic field lines B_p , which can be compared with (linear) plots like Fig. 15.9. In a steady state the poloidal velocity V_p would be parallel to these lines, but that is not necessarily the case in this time-dependent simulation, especially at the larger distances. Let us look at the other classes of surfaces:

- **The Computational Boundaries:** blue arcs at $r \approx 0.9 r_g$ in lower left-hand corner and at $r \approx 10^4 r_g$ at upper right, plus the ordinate and abscissa.
- **Black Hole Surfaces:**
 - *Horizon:* Dark blue arc at $r_H \approx 1.35 r_g$.
 - *Ergosphere:* Green arc inside the slow magnetosonic surface.
- **The Stagnation Surface (between ingoing and outgoing winds):** The very wavy black curve beginning at $Z \sim 6 r_g$. On this surface $V_p = 0$ and particle creation or diffusion into this region must take place in order to maintain the Goldreich–Julian charge density.
- **The Ingoing Wind:**
 - *Ingoing SMS:* Wavy magenta line inside the stagnation surface and outside the ergosphere.
 - *Ingoing Light Surface:* Not shown, but lies near the ingoing cyan Alfvén surface.
 - *Ingoing AS:* Cyan arc just outside the horizon.
 - *Ingoing FMS:* Essentially coincident with the dark blue horizon.

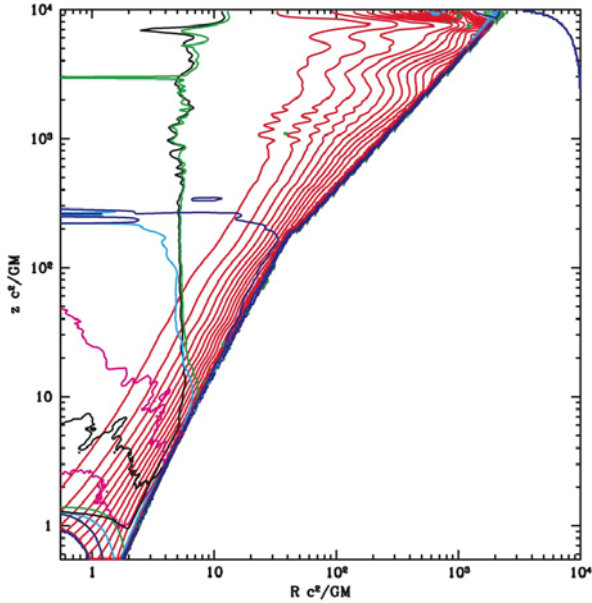


Fig. 15.15: Log–log plot of important surfaces for simulation in Fig. 15.14. Curves are as follows: outer and inner thick black (computational boundaries); red (field/stream lines); green (outer light cylinder radius and ergospheric boundary; inner light surface not shown); blue (outer and inner FMS); cyan (outer and inner AS, which lie near light surfaces); magenta (outer and inner SMS). The three thinner black lines each have a different character: inner (horizon), middle wavy (stagnation surface, from which both inner and outer winds emanate), and long curve beginning near horizon and following the light cylinder (where $B_\phi = B_r$). Reproduced from Fig. 11 of [457], by permission of J. Wiley and Sons, Ltd.

• The Outgoing Wind and Jet:

- *Outgoing SMS:* Wavy magenta line outside the stagnation surface.
- *Outgoing AS:* Cyan line beginning at $Z \sim 250 r_g$ and continuing along the outgoing light cylinder.
- *Outgoing Light Cylinder:* Green line with roughly constant $R \sim 5\text{--}7 r_g$.
- *Outgoing FMS:* Dark blue line beginning at $Z \sim 300 r_g$ and extending well to the right of the light cylinder.

The long black line that begins near $Z \approx r_H$, arcs to the jet boundary, and then parallels the outer light cylinder is where the pitch angle of the magnetic field is roughly 45° ($B_\phi = B_r$). The accretion inflow and extended atmosphere lie in the blank white space at right in the diagram. In this very impressive simulation the jet traveled out to ten thousand r_g , where it began to undergo oscillation and shocking (see Fig. 15.15). Based on Vlahakis and Königl’s equations, this is in a region where we might expect an FMSS to form in McKinney’s simulation, including the initial oscillations and subsequent shock at distances of 1000–2000 r_g .

Despite this tantalizing possibility, however, there are a few of reasons why an FMSS may not have formed. One, of course, is that even this simulation may have not been run long enough to complete the process. Twenty crossing times at $0.5c$ for an FMSS occurring at $1000\text{--}2000 r_g$ implies a necessary evolution time of $4\text{--}8 \times 10^4 \tau_g$, whereas the simulation was run for about 1/6 to 1/3 of this. Fig. 15.15 also shows significant numerical boundary effects, which may have affected the flow for $Z \gtrsim 3000 r_g$. A third issue relates to a problem pointed out in a 2007 paper by Komissarov and his colleagues [553]. One can use Fig. 7 of [457] to determine if total energy is conserved along a magnetic field line or not by computing

$$\mu = \gamma_\infty = \gamma_\infty^{\text{MA}} + \gamma_\infty^{\text{EM}}$$

where “MA” stands for matter (kinetic, thermal) and “EM” for electromagnetic energy. One finds, however, that this quantity does *not* remain constant along the field line for which the figure is plotted. Instead, μ drops by almost two orders of magnitude, and most of the loss occurs in the region $Z < 100 r_g$. Where this energy has gone is not clear, but the implications are that, by $Z \sim 1000 r_g$ the jet may have weakened sufficiently that any forming FMSS and subsequent collimation shock would be only moderate at best.

The bottom line is that many more simulations like this need to be performed in order to study numerically the formation of the magnetosonic horizon (FMSS). The elliptic flow region inside the FMS must not be disturbed by boundary or other issues. The simulation needs to run long enough for that elliptic region to communicate well with a potential FMSS region (at least 20 magnetosound crossing times). And special attention needs to be paid to conservation of energy in the numerical method, not only globally but also locally.

15.2 Beyond the Magnetosonic Horizon: Propagation of Ballistic Jets

Perhaps the least understood aspect of jet production is how the flow transitions from an accelerating and collimating one to a ballistically propagating jet entity, free of any further influence on the central engine. To be sure, once the flow passes through the final separatrix surface (*i.e.*, the FMSS), the jet becomes causally disconnected from the central engine. But, how does this transition take place? Beyond the FMSS the jet flow becomes kinetically dominated and is rapidly converging. Does this convergence create a strong shock, disrupt and reconnect the magnetic field, and create a turbulent, weakly-magnetized, hot plasma? Or, does the converging flow bounce more gently, retaining much of its toroidally magnetized nature. In fact, do all jets even form an FMSS, or do some “fizzle” and never become formally causally disconnected from the engine? And can we identify different astrophysical sources with any of these possible outcomes?

A complete set of answers to all of these questions does not exist at this time and will require many simulations like that of McKinney in order to explore all the possibilities. Instead of following the flow through the transition to causal disconnection, therefore, in this final section we shall *begin* by assuming that the jet has, indeed, become causally disconnected and then study the behavior of a variety of ballistic jets with different properties. We then can compare such simulations with observations of actual astrophysical jets and *infer* what kind of transitions the initial accelerating flow in those real sources must have made near the FMSS.

In effect, we now shall begin at the jet head, where it impacts the interstellar or intergalactic medium and work our way back toward the FMSS.

15.2.1 Hydrodynamic Jets: A Model for FR II Sources

In 1973 several serious problems of understanding how radio sources are powered by an active galactic nucleus were outlined by Malcolm Longair, Martin Ryle, and Peter Scheuer (of the Mullard Radio Astronomy Observatory in Cambridge, UK). They also suggested an important solution: the energy powering the lobes did indeed come from the radio galaxy nucleus, and it was supplied continuously in the form of a relativistic beam or jet [540, 541]. This really was the first serious suggestion that jets were at work in these sources, more than half a decade before such jets were actually imaged by the Very Large Array radio telescope.

The basic properties of a jet flow striking a denser external ambient medium then were worked out by Roger Blandford and Martin Rees in 1974 [28]. While their model for how the jet flow is accelerated (twin hydrodynamic exhaust nozzles) is no longer favored for astrophysical jets, their model for the jet hot spots and radio lobes is still the best explanation for the extended structure of FR II sources. On the basis of Hargrave and Ryle’s map of Cygnus A (Fig. 2.3), these authors concluded that the jets that fed those lobes from the galactic nucleus were flowing supersonically. The jets then formed a strong reverse shock (the “working surface” or “Mach disk”) at the location of the hot spots.

Fig. 15.16 shows Blandford and Rees’s schematic drawing of what happens at an FR II hot spot. Because material from the ambient medium must flow into the lobe region (across the bow shock **B**), the lobe and hot spot at the jet head *must* move away from the central black hole at a speed V_{head} , which can be found by equating (in the frame of the contact discontinuity **D**) the ram pressure of the jet with the ram pressure of the inflowing external medium

$$\frac{L_{jet}}{\pi r_{jet}^2 V_j} = \rho_{ext} V_{head}^2$$

For a relativistic jet its power and speed are $L_{jet} = \gamma_{jet} \rho_{jet} c^2 (\pi r_{jet}^2 c)$ and $V_j = c$, where ρ_{jet} includes the enthalpy of the jet particles as well as their rest mass. (For a jet composed of a relativistic plasma, with random particle energies $\gamma_{particle} \gg 1$,

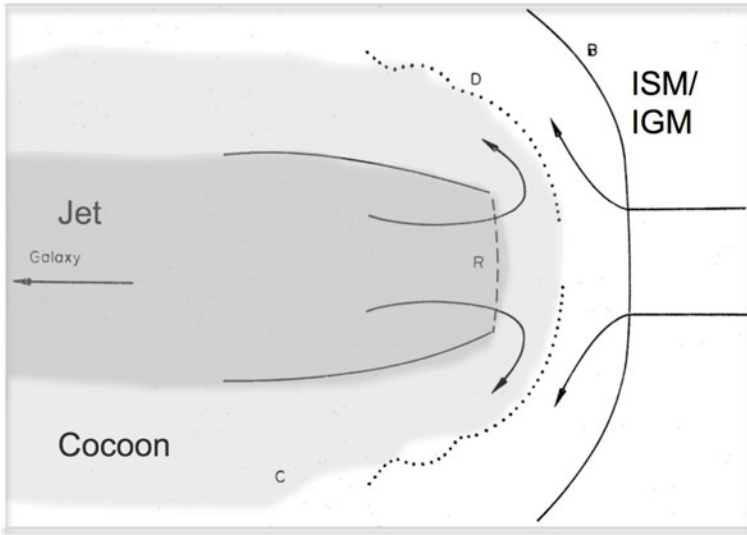


Fig. 15.16: Schematic diagram of a hot spot's structure at the end of a supersonic jet, drawn in the rest frame of the “Mach disk” or “reverse shock” (**R**). Jet material (medium gray) travels toward the right, passes through the Mach disk, where it is heated and then forms the hot cocoon (**C**). External material from the interstellar or intergalactic medium flows toward the left in this frame, is heated at the bow shock (**B**), and forms a hot confining medium around the cocoon. The contact discontinuity (**D**) separates the two different shocked media. Adapted from Fig. 2 of [28]. Reproduced by permission of J. Wiley and Sons, Ltd.

the enthalpy will overwhelmingly dominate the jet inertia rather than the plasma rest mass.) This gives

$$\frac{V_{head}}{c} = \left(\frac{\gamma_{jet} \rho_{jet}}{\rho_{ext}} \right)^{1/2}$$

So, the advance of the jet head will remain non-relativistic ($V_{head} \ll c$) as long as the jet is very light (much less dense than the very tenuous intergalactic medium, *i.e.*, $\rho_{jet} \ll \rho_{ext}/\gamma_{jet}$). For a typical intergalactic density of $10^{-27} \text{ g cm}^{-3}$, a typical jet Lorentz factor of $\gamma_{jet} = 10$, and jet particle Lorentz factor of $\gamma_{particle} \approx 10^5$, this means that the jet must be much less dense than the equivalent of one relativistic electron particle per cubic meter!

15.2.1.1 One-Dimensional HD “Jet” Simulations: The Shock Tube Jet

A true *shock tube* is an experimental device for studying supersonic shock propagation in the laboratory. In the classical version of the experiment a tube of gas is divided lengthwise into two parts by a barrier that can be quickly opened fully, one part with high-pressure gas and one with low-pressure gas. Opening the barrier

suddenly creates a large pressure discontinuity that drives a shock wave into the low-pressure gas. Properties of shock waves can be studied by varying the densities and pressures of the two gases.

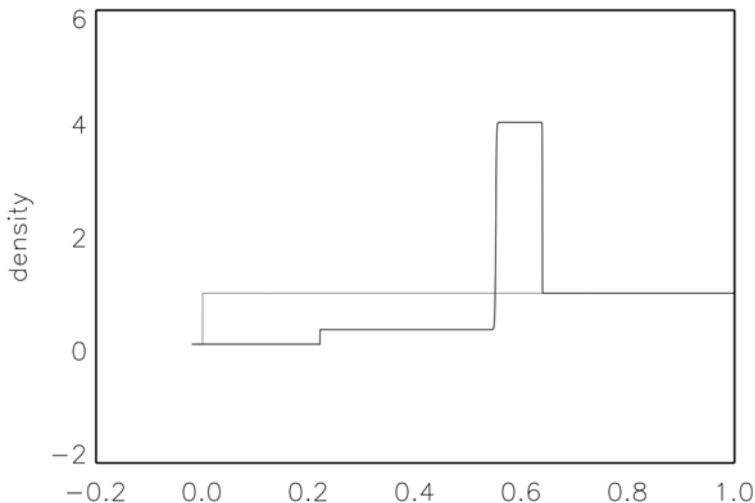


Fig. 15.17: Plot of density vs. distance for a simulation of a simple one-dimensional supersonic flow from the left at a speed of $0.98c$ ($\gamma = 5$). Thin gray line shows initial conditions with the “jet” density equal to 0.1 times the external density ($\rho_{\text{ext}} = 1.0$). Black line shows conditions at $t = 0.66$. Compare with two-dimensional plots in Figs. 15.16 and 15.18: reverse shock lies at $x \approx 0.2$, contact discontinuity at $x \approx 0.55$, and “bow” shock at $x \approx 0.65$. Simulation performed by D. Garofalo using M. Nakamura’s SRMHD code in hydro-only mode.

Shock tubes can be simulated relatively easily with a one-dimensional hydrodynamic (HD) code. A particularly relevant problem is shown in Fig. 15.17, where we have induced a shock, not with a pressure discontinuity but with a supersonic inflow from the left, at a speed of $0.98c$ ($\gamma = 5$). This simple simulation has most of the features outlined by Blandford and Rees:

- A “reverse” shock at $x \approx 0.2$, corresponding to the Mach disk, where the high-speed inflow is initially decelerated (to $\sim 0.8c$).
- A constant pressure region, extending from $0.2 \leq x \leq 0.65$ that contains both “jet” material injected from the left and “ambient medium” material that have been shocked.
- A contact discontinuity at $x \approx 0.55$ separating the heavier ambient medium from the lighter jet.
- Finally, a strong “forward” shock, corresponding to the jet bow shock where the ambient medium pressure and density are compressed by the inflowing jet.

The only feature missing from the basic jet picture in Fig. 15.16 is the over-pressured nature of the post-shock jet/ambient material region and their lateral expansion (in

a second dimension) into the cocoon. So, while the one-dimensional shock tube cannot provide a quantitative picture of jets, it can be used to qualitatively illustrate most of a jet's important features. This will be important below when we discuss MHD jets.

15.2.1.2 Two-Dimensional HD Jet Simulations

Non-Relativistic 2-D Studies

Eight years after Blandford and Rees's analysis, their prediction of light supersonic jets was confirmed by the detailed two-dimensional hydrodynamical simulations performed by Michael Norman, Larry Smarr, Karl-Heinz Winkler, and Michael Smith [542]. They all visited the Max Planck Institute for Physics and Astrophysics in Garching, Germany, and used the powerful supercomputer there at the Max Planck Institute (see Fig. 15.18).¹¹ All the parts of Blandford and Rees's schematic diagram can be seen near the jet head in this figure: bow (forward) shock, shocked ambient medium, contact discontinuity between the latter and the cocoon (shocked jet material), Mach disk/working surface (reverse shock), and supersonic jet. Note the density (and pressure) gradients in the shocked ambient medium and cocoon that drive the material radially away from the jet axis, just as was drawn in Fig. 15.16. Because the Mach disk is such a strong shock, one can expect a lot of particle acceleration and synchrotron emission there if the plasma were even weakly magnetized. The structure of these two-dimensional hydrodynamic (HD, not MHD) simulations, therefore, is quite similar to that of FR II radio sources, *e.g.*, Cygnus A (Figs. 2.3 and 2.5).

At the time (1974 to the early 1980s), this result was not nearly as surprising and significant as it is now. Drawing on the results of the previous sections, we now know that the flow near the central engine must be *strongly* magnetized ($\beta_p \ll 1$) in order to reach the observed relativistic speeds. Yet, here we have concluded that FR II sources behave as if their plasma were largely *unmagnetized* ($\beta_p \gg 1$). In the FR II case, at least, the jet *must* have transitioned, at some point beyond (but perhaps near) the FMSS, from a warm, magnetically and Poynting-flux-dominated, accelerating flow to one that is hot, gas pressure and kinetically dominated. *One possible explanation for this is that the rapid pinching of the jet beyond the final separatrix surface in FR II sources may indeed be strong enough to destroy the well-ordered magnetic field and convert its energy into heat (i.e., accelerate the jet plasma particles to relativistic energies).* Such a dramatic change in an FR II jet's

¹¹ While powerful supercomputers for full two-dimensional simulations had been available at US government laboratories for some time before this work (*e.g.*, LeBlanc and Wilson's computation of the collapse of a rotating magnetic star in 1970 [373]), these non-governmental authors had to travel to Germany to produce such nice results. This embarrassment to the US scientific community, plus the subsequent efforts by several of the authors of this famous paper, eventually led to the US Congress funding the National Science Foundation supercomputer centers nationwide. These centers continue to support high-performance scientific computation (in all fields) to this day.

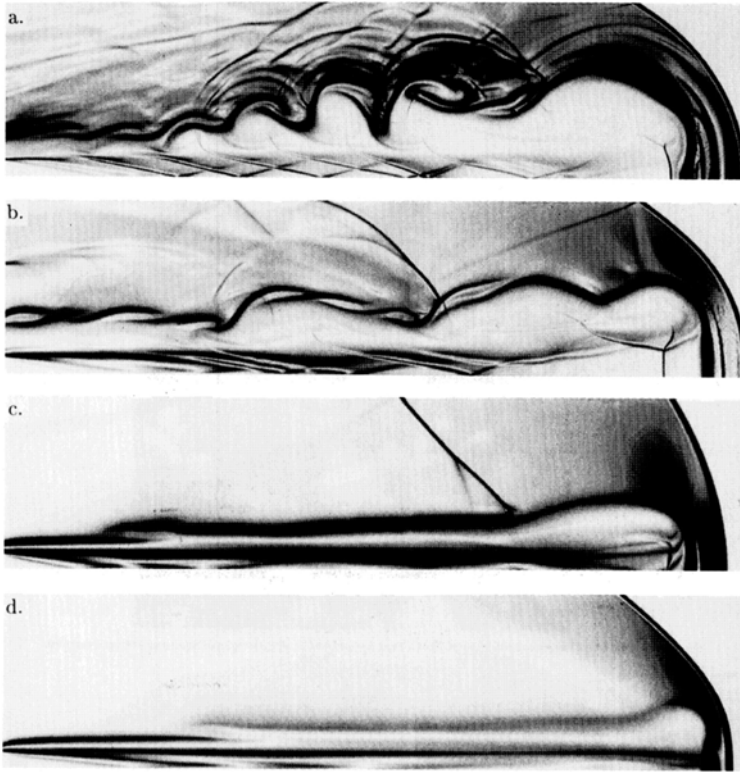


Fig. 15.19: Four different simulations of two-dimensional, axisymmetric hydrodynamic jets, showing effects of making the flow more relativistic, both in speed (top three) and in adiabatic index (bottom). (a): Speed: $v = 0.3c$ ($\gamma = 1.05$), adiabatic index $\Gamma = 5/3$; (b): $v = 0.98c$ ($\gamma = 5$), $\Gamma = 5/3$; (c) $\gamma = 10$, $\Gamma = 5/3$; (d) $\gamma = 10$, $\Gamma = 4/3$. Non-relativistic jet at top appears to have a turbulent shocked ambient medium and cocoon, like Fig. 15.18. But, highly relativistic flow (c) stabilizes this turbulence, and having a more relativistic plasma ($\Gamma = 4/3$) increases that stability. Reproduced from Fig. 1 of [554], by permission of the AAS.

15.2.1.3 Three-Dimensional HD Jet Simulations

Non-Relativistic 3-D Studies

Of course, the assumption of axisymmetry in the two-dimensional studies can potentially suppress many instability modes. Indeed, Philip Hardee, of the University of Alabama, one of the leading experts on jet stability, and his colleagues showed that it is the non-axisymmetric modes (particularly the helical, elliptical, and triangular Kelvin–Helmholtz surface modes) that tend to go unstable the most quickly and then disrupt a jet [556]. Fig. 15.20 shows how the jet cross-section changes with time. We can see triangular and elliptical modes coupling, along with some helicity (displacement of the densest point from the main jet axis). Fig. 15.21 shows the effects



Fig. 15.20: Gray-scale temperature plots of cross-sections through a three-dimensional hydrodynamic jet that becomes increasingly unstable along its length. (Lighter gray indicates higher temperature.) Cuts are taken along the jet length at (7.5, 15.0, ..., 37.5, and 45.0) jet radii. While jet is roughly laminar early on, it becomes more turbulent downstream. Reproduced from Fig. 10 of [556], by permission of the AAS.

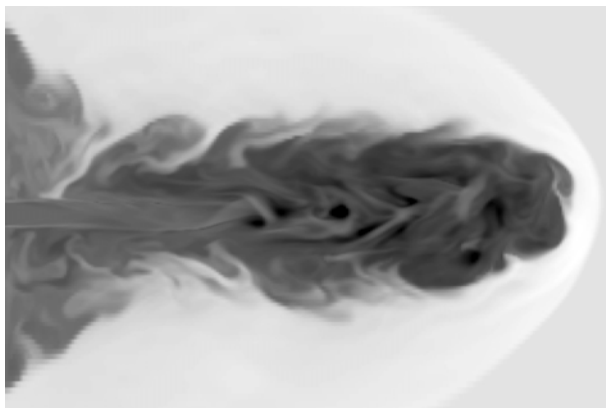


Fig. 15.21: Full three-dimensional realization of the growth of turbulence in a light (density ratio $\eta_\rho = 0.02$), supersonic ($\mathcal{M} = 6$) hydrodynamic jet. Reproduced from Fig. 1 of reference [557], by permission of the ASP.

of those modes on a three-dimensional simulation of Mach 6 jet by David Clarke of St. Mary's College in Halifax, Nova Scotia in 1996 [557]. Even a jet this supersonic is disrupted by non-axisymmetric Kelvin–Helmholtz instabilities, broadening and slowing the jet significantly. Does this disruptive effect of non-axisymmetry on supersonic jets pose a problem for our understanding of black hole jets (which appear to remain stable for long distances)? In order to answer this question, we need to examine full three-dimensional simulations of relativistic jets.

Relativistic 3-D Studies

The first three-dimensional (magneto-) hydrodynamic simulations of relativistic jets were performed by Ken-Ichi Nishikawa of LSU and his colleagues in 1998 [558]. Their simulations of a $\gamma = 4.56$ ($V_j = 0.976c$) jet showed that, while the jet could be deflected by a strong magnetic field, it remained rather stable. A year later Miguel-Angel Aloy of the University of Valencia, Spain, and his colleagues performed much higher resolution three-dimensional HD simulations of a $\gamma = 7$ ($V_j = 0.990c$) jet [559]. These authors did not investigate the effects of deflection or Lorentz factor on the stability of their relativistic jet. However, they did study the effects of increasing the spatial resolution of the simulation and found that the jet cocoon became more prominent and more structured as resolution is increased. Nevertheless, the body of their relativistic jet itself appeared quite stable, unlike that of Clarke's non-relativistic jet in Fig. 15.21.

The first authors to look specifically at how jet stability varies with relativistic speed were Hughes, Mark Miller (then at Washington University in St. Louis), and Duncan, in 2002 [560]. Jets with $\gamma = 2.5$ ($V_j = 0.916c$), 5, and 10 were deflected by a strong, oblique shock (rather than a magnetic field) and found to become more stable the higher the Lorentz factor. The most interesting experiment, however, was the introduction of a precession of the jet axis at its base (Fig. 15.22). For the $\gamma = 5$ jet, a wiggle of only $\pm 11^\circ.25$ showed that this jet initially continued to propagate for ~ 50 jet radii, but then disrupted like Clarke's non-relativistic jet. However, recent work in 2009 by Hughes (private communication) shows that a $\gamma = 10$ jet, with a similar precession angle, does *not* disrupt (Fig. 15.23), even when this direction change is quite rapid. Instead, the sudden change in jet direction create two separate jets that propagate independently in different directions.

These simulations clearly show that, while Kelvin–Helmholtz instabilities are seriously disruptive to non-relativistic jet flows, and even affect the cocoons of relativistic jets, they do not appear to cause many problems for jets with $\gamma \geq 10$. Even highly relativistic precessing jets seem to be largely stable, creating multiple “bullet” jets rather than disrupting them.

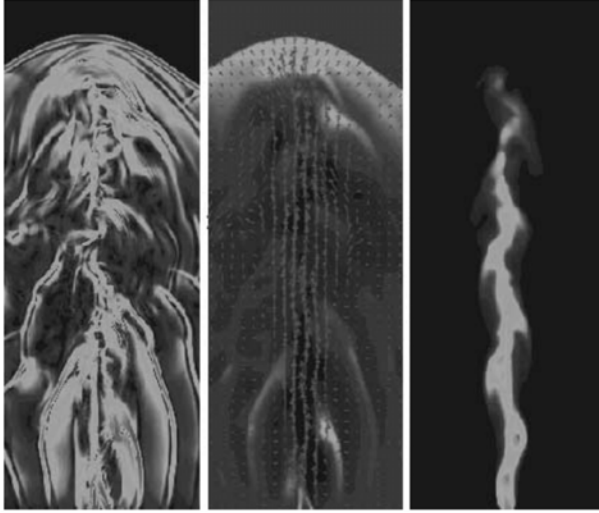


Fig. 15.22: Three different views of a three-dimensional $\gamma = 5$ hydrodynamic jet that has an $11^\circ.25$ precession at its base. *Left:* Pressure. *Center:* Velocity vectors with linear pressure in the background. *Right:* Lorentz factor. The inner core has $\gamma \approx 5$, while the dark outer material has $\gamma \approx 1$. Disturbed jets at this modest Lorentz factor still appear to eventually go turbulent. Reproduced from Fig. 13 of [560], by permission of the AAS.

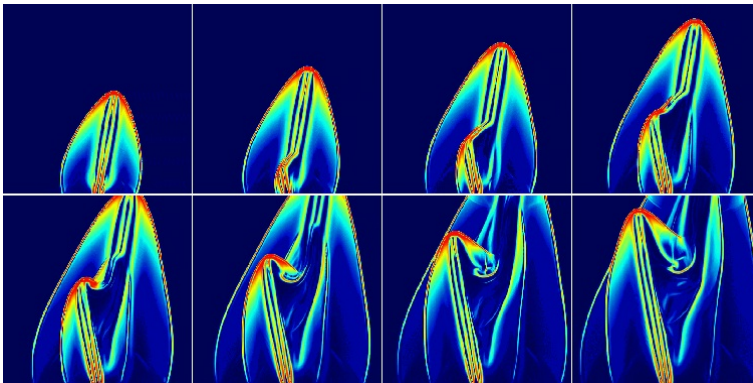


Fig. 15.23: Eight time frames from a recent three-dimensional simulation of a $\gamma = 10$ hydrodynamic jet with a precession angle similar to that in Fig. 15.22. While the $\gamma = 5$ jet was disrupted under similar circumstances, this more relativistic jet is not, even with rather rapid directional changes. Highly relativistic hydrodynamic flow, therefore, appears quite stable to significant disturbances. Figure kindly supplied by P. Hughes; reproduced with permission.

15.2.1.4 FR I Sources as Disrupted FR II Flows

By the beginning of the 1990s, once the basic structure and dynamics of Fanaroff and Riley Class II sources were understood, investigators turned toward understanding the lower-luminosity, and more abundant, FR I sources (see Section 2.2.1), using the two-dimensional and three-dimensional hydrodynamic physics and simulation codes that were available at that time.

Spontaneous Disruption at Low Jet Mach Numbers

One of the first investigations to consider seriously this issue was done in 1985 by Geoffrey Bicknell of Mount Stromlo and Siding Spring Observatories and Australian National University [561]. He based his results on laboratory experiments with subsonic and supersonic hydrodynamic jets, which show that low Mach number jets are more susceptible to turbulence than highly supersonic flows. He concluded that FR I flows were low Mach number ($\mathcal{M}_{\text{jet}} \equiv V_j/c_{s,\text{jet}} \leq 2$) FR II-like jets that had become turbulent and then slowed to subsonic speeds. The critical Mach number could be even much higher than 2 if the jets had very low densities compared with the interstellar and intergalactic medium. Clarke's 1996 and Hughes' *et al.* 2002 simulations of disrupting low Mach number and low-density jets are good examples of Bicknell's spontaneous disruption mechanism.

In 1995 Bicknell applied this idea to the Owen *et al.* [42, 43, 44] result that the dividing line between FR I and II sources in the radio-optical plane is a strong function of galaxy luminosity (*i.e.*, galaxy mass; see Section 2.2.1) [45]. He was able to predict not only the slope of this dividing line but also its intercept on the radio axis. The Bicknell model is still the leading explanation for many of the statistical properties of FR I sources.

Induced Disruption by a Strong Shock

A second method for disrupting supersonic hydrodynamic jets is for them to interact with a strong external shock. Such shocks can be produced by extragalactic accretion of gas onto a radio galaxy [562, 563].

In 1988 Mike Norman and his colleagues performed two-dimensional slab (x, y) HD simulations of moderately supersonic ($\mathcal{M}_{\text{jet}} = 2\text{--}5$) jets that interact with a strong shock in the external medium [562]. They found that, if the Mach number of the shock were greater than that of the jet, the jet would disrupt and expand into a large, subsonic plume (Fig. 15.24). This result is consistent with the later simulations of shock/jet interactions performed by Hughes and his colleagues in 2002 and 2009 (discussed above).

Finally, another method of inducing jet disruption that is related to the shock-discontinuity model was studied by the author, Alberto Sadun of Agnes Scott College, and Kevin Lind of NRAO in 1991 [563] and by Nakamura and his colleagues

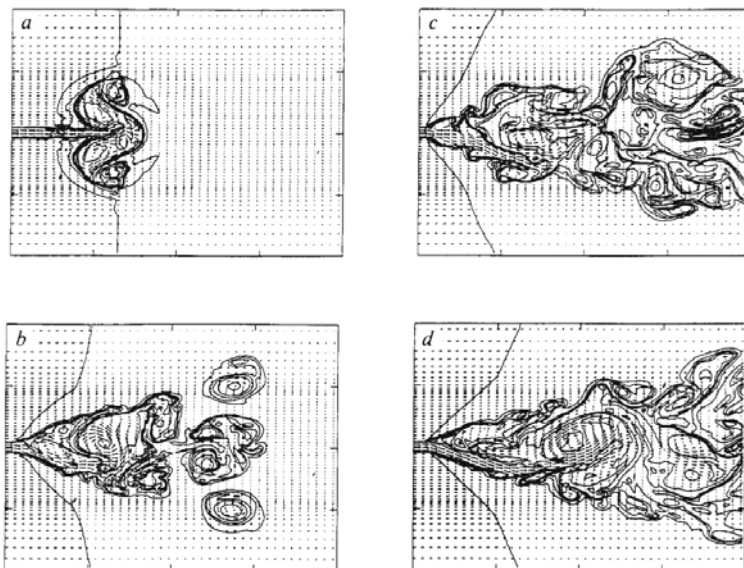


Fig. 15.24: Density contours and velocity vectors for four time sequences from a two-dimensional (“slab”) simulation of a low Mach number jet that emerges from a higher Mach number wind (both traveling from left to right). The jet is disrupted as long as $\mathcal{M}_{\text{wind}} > \mathcal{M}_{\text{jet}}$. Reprinted by permission from Macmillan Publishers Ltd: Fig. 3 of reference [562], copyright (1988).

in the context of the magnetic tower (see Section 14.5.3) [565, 566, 567]. Infalling gas onto a galaxy can produce an accretion shock with a strong pressure discontinuity (p being high inside the shock radius (nearer the galaxy) and very low outside it). When a jet flowing outward from the galaxy passes through this discontinuity into the low-pressure region, it no longer is pressure confined by the ambient medium. It can expand freely for quite a while, until the internal jet pressure drops to the new external ambient pressure. The broad, poorly-collimated flow then quickly decelerates in a manner similar to the shock-induced deceleration above.

This is only a brief introduction to the ways in which an FR I radio source might form. More methods are discussed in the next subsection.

15.2.2 Magnetohydrodynamic Jets: A Model for Perhaps Many FR I Sources

The introduction of a strong magnetic field into a supersonic jet can dramatically change the character of the flow, as we have seen previously in the first section of this chapter. A magnetic field can alter, or destroy, the hot spot/cocoon structure of the jet head and introduce kinks and wiggles in the jet body. In addition, because

there are two shock systems in MHD (slow and fast; see Section 9.6.2) instead of HD shocks only, a strong magnetic field can create a variety of interesting, and possibly useful, shock features in the jet outflow. Some of these aspects are consistent with features of at least some FR I sources. Unlike FR II sources, therefore, FR I radio galaxies may be a very heterogeneous lot, with some being disrupted hydrodynamic flows and others being more complex magnetohydrodynamic ones.

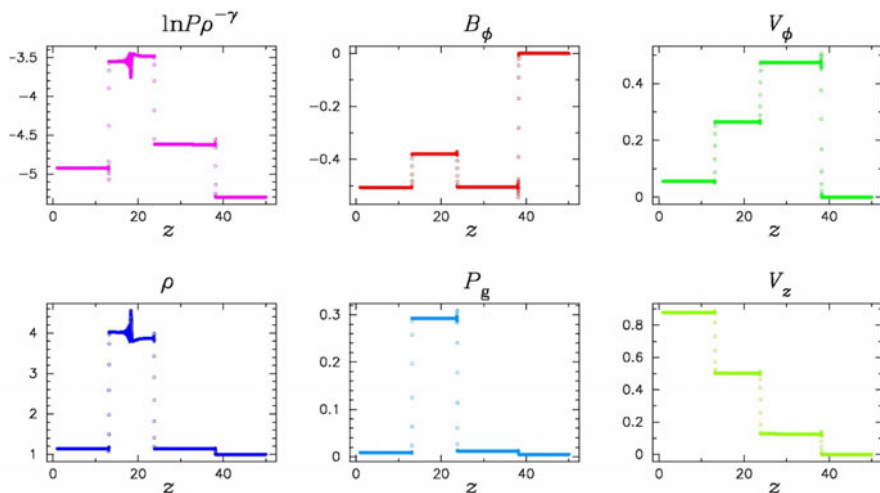


Fig. 15.25: Plots of various physical quantities with distance for a simulation of a simple one-dimensional super-slow magneto-acoustic “jet” from the left. The flow, however, is not super-fast, so it produces only three shocks. (See the p_g and V_z plots at lower right.) From left-to-right they are: reverse slow, forward slow (slow-mode bow shock), and forward fast (fast-mode bow shock). The contact discontinuity can be seen between the two slow-mode shocks in the entropy (upper left) and density (lower left) plots, but not in the others. (The sharp spike in ρ is a numerical artifact.) The most interesting aspect of MHD jets is the manner in which B_ϕ and V_ϕ seem to show rotation and helical field reversals. Actually, V_ϕ (also V_z) is antisymmetric about the region with the contact discontinuity, while B_ϕ is symmetric about the discontinuity. Figure from Fig. 4.3 of [568]; reproduced with permission of M. Nakamura.

15.2.2.1 One-Dimensional MHD “Jet” Simulations

As with the HD jets in the previous subsection, one-dimensional shock tube simulations can illustrate many of the features of an MHD jet. Fig. 15.25 shows a one-dimensional (often referred to as 1.5-dimensional because the azimuthal quantities V_ϕ and B_ϕ also are evolved) computation that is similar to Fig. 15.17. However, now the simulation includes a strong magnetic field (B_z) that threads the jet inflow and external medium as well as a torsional Alfvén wave (B_ϕ) that propagates with the inflowing jet ($V_z = 0.89$; $\rho = 1.2$). In this figure the inflow is *faster* than its

own slow magneto-acoustic speed but *slower* than its fast magneto-acoustic speed. That is, $V_{S,jet} < V_Z < V_{F,jet}$.

In the density, V_Z , and gas pressure p_g evolution diagrams of Fig. 15.25 we see very familiar features:

- A contact discontinuity between the shocked jet and shocked ambient medium ($Z \approx 18$; where ρ changes sharply but p_g does not).
- A reverse (slow-mode) shock, signifying a slow-mode Mach disk ($Z \approx 13$).
- A forward (also slow-mode) shock, signifying a slow-mode bow shock ($Z \approx 23$).
- A staircase in V_Z , where the flow velocity steps down at each succeeding shock (forward or reverse).

In addition to the above, there now also is a *fast-mode* forward (bow) shock ($Z \approx 38$) that travels ahead of the slow shock pair. This fast shock occurs because the inflow speed is faster than the *ambient medium's* fast magneto-acoustic speed.

The behaviors of V_ϕ and B_ϕ , however, are completely new. V_ϕ produces a staircase that *increases* in level toward the bow shock – the reverse of the V_Z behavior. The seemingly complex behavior of B_ϕ also appears strange but actually has a very simple interpretation:

1. The head of the propagating torsional Alfvén wave occurs at the forward fast-mode shock, and B_ϕ is strongly modified there. Behind that shock, however, the wave train maintains its input twist of $B_\phi \approx -0.5$.
2. However, between the two slow-mode shocks something strange happens: the forward slow shock untwists the field somewhat, and then the reverse shock twists it back again.

The nice symmetry of this interpretation seems to be broken by the V_ϕ and V_Z velocity plots, that is, until one transforms to the translating *and* rotating frame of the flow in between the two slow-mode shocks (see Fig. 15.25). In the frame with V_ϕ and V_Z zero in that region, the jet and ambient material flow in from left and right, respectively. And the equal and opposite azimuthal velocities of these two inflows are both slowed to zero at the two shocks. As with the HD jet, the system has a very simple symmetry around the contact discontinuity.

This unusual behavior of the toroidally magnetized flow between two slow-mode shocks in *three-dimensional* jets is potentially important for understanding certain types of FR I radio galaxies, as we shall see below.

15.2.2.2 Two-Dimensional MHD Jet Simulations: A Possible Glimpse into Jet Behavior just beyond the Magnetic Horizon

Non-Relativistic 2-D Studies

The first two-dimensional magnetized jet simulations were performed in 1986, only four years after the Norman *et al.* work on HD jets, by David Clarke, Mike Norman,

and Jack Burns (all at the University of New Mexico at the time) [543]. They computed the evolution of a jet with a strong *torsional* field (plasma $\beta_p = 0.2$), showing that highly magnetized jets have a significantly different structure from purely HD ones. Shortly after this study, in 1989, Kevin Lind (NRAO), David Payne and D. Meier (both at JPL), and Roger Blandford performed a comprehensive study of how decreasing the plasma β_p of the jet below unity triggers a dramatic onset of the Clarke *et al.* MHD jet flow [544].

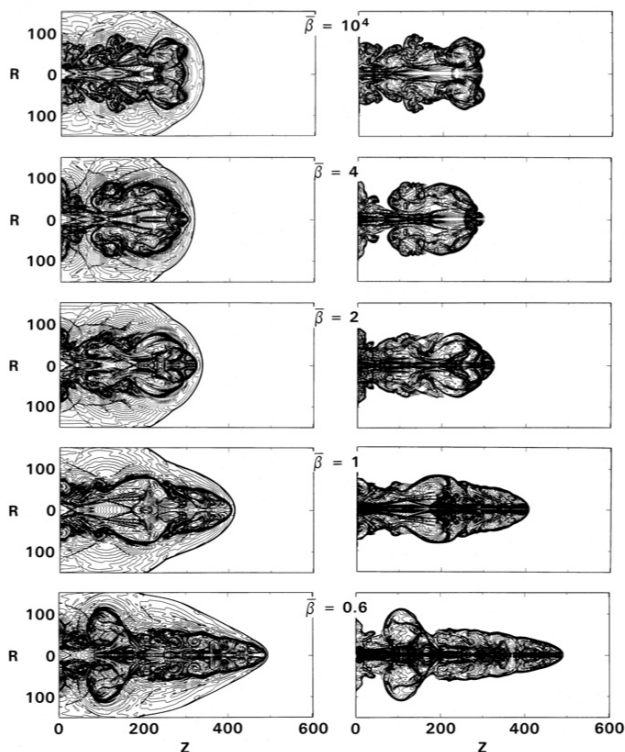


Fig. 15.26: Five different simulations (one per row) of two-dimensional, axisymmetric non-relativistic MHD jets with strong toroidal magnetic field, all shown at the same model time ($t = 75$). *Left column:* Contours of constant density. *Right column:* Contours of constant B_ϕ . *Top two panels:* A hydrodynamic jet (plasma $\beta_p = 10^4$) with a Mach number of $\mathcal{M}_{\text{jet}} = 6.1$, adiabatic index $\Gamma = 5/3$, and density ratio $\eta_\rho = 0.1$. The remaining simulations ($\beta_p = 4, 2, 1$, and 0.6) have the same initial parameters except for the relative strength of B_ϕ . As the magnetic field increases in strength, the cocoon diminishes in importance and the “nose cone” and speed of the bow shock both grow. Reproduced from Fig. 12 of [544], by permission of the AAS.

Fig. 15.26 shows jets with $\beta_p = 10^4, 4, 2, 1$, and 0.6 – all at the same model time and all with a Mach number of $\mathcal{M}_{\text{jet}} = 3.5$ and density ratio $\eta_\rho = 0.1$. The high β_p jet appears just like an HD jet, with a bow shock, contact discontinuity, Mach disk, and cocoon surrounding the inflowing jet. However, as the magnetic field becomes

more important, more of the cocoon material is forced forward by the strong toroidal pinch into a fast-moving *nose cone*¹² that flows ahead of the rest of the jet. In Clarke *et al.*'s $\beta_p = 0.2$ case, almost all of the shocked jet material is forced into the nose cone with little left to be sent into the cocoon.

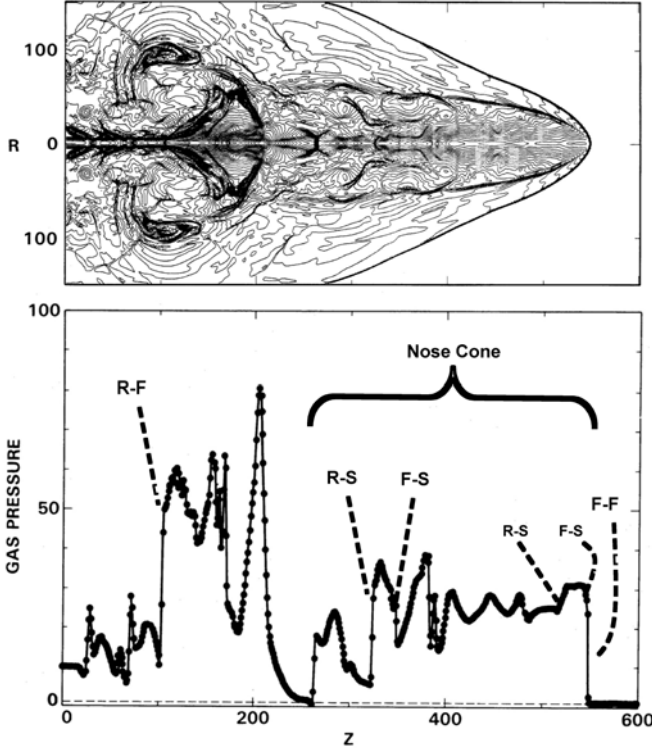


Fig. 15.27: Pressure and shock conditions for the final time step ($t = 85$) for the highly magnetized ($\beta_p = 0.6$) model in Fig. 15.26. *Top:* Pressure contours. *Bottom:* Linear plot of pressure along the rotation axis ($R = 0$). The injected jet speed here is super-fast magneto-acoustic, which produces a reverse-fast shock ($Z \sim 100$) in addition to those in Fig. 15.25. Ambient pressure is unity, so the nose cone is over-pressured by a factor of ~ 30 . Note the *pairs* of reverse-slow and forward-slow shocks, with one forward-fast bow shock at right. Examination of successive views like this one show repeated ejections of jet components from the magnetic “chamber” near $Z \lesssim 200$. Adapted from Figs. 6b and 9 of [544]. Reproduced by permission of the AAS.

The best way to identify most of an MHD jet’s features is by looking at a two-dimensional contour plot of the pressure of the $\beta_p = 0.6$ simulation (Fig. 15.27, top panel) and comparing with a plot of the gas pressure measured along the Z axis at the same time (Fig. 15.27, bottom panel). We note the following:

¹² Dubbed so by Clarke *et al.*

- Starting at the left in Fig. 15.27, the first two X-type shocks are not very significant. They appear to be related to transient effects created when the jet exits the inlet at $Z = 0$ or to the impact of cocoon eddies on the jet cylindrical surface.
- The strong reverse shock (“Mach disk”) at $Z \approx 110$ is a *fast-mode* shock that decelerates the super-fast magneto-acoustic flow to sub-fast, but still super-slow, velocities.
- The jet flow beyond the fast-mode Mach disk remains supersonic with respect to the denser, *unmagnetized* external medium. The strong bow shock ($Z \approx 550$) is, therefore, an HD one, neither fast nor slow MHD shock, because it occurs in the external medium. And the contact discontinuity travels very near the head of this bow shock, in the uniform pressure region just behind it ($Z \approx 540$).
- Most of the remaining features in the region $120 \leq Z \leq 530$ are either multiple slow-mode shock pairs or waves. While each of these shock pairs is similar to the single pair in Fig. 15.17, the fact that many are generated by the flow indicates that much more happens in two-dimensional MHD flow than in 1-D.
- Finally, we note the strange (and largely stationary) spike shock pair at $Z \approx 200$. Its significance is discussed below.

The top panel of Fig. 15.27 shows the two-dimensional pressure contours of this simulation. Dense collections of parallel contour lines indicate axisymmetric shock features. Our analysis in Lind *et al.* showed that the nose cone lies in the region $Z \geq 200$, and the higher pressure conical feature bounded by strong shocks (just prior to $Z \approx 200$) is a magnetic chamber that periodically belches out fast pulses of material (slow-mode shock pairs) and feeds the trans-fast ($V_S \ll V_{nosecone} \leq V_F$) nose cone flow. The nature of the narrow shock pair at $Z \approx 200$ (see bottom panel of Fig. 15.27) is very unusual. This feature is almost stationary, so it is definitely a slow-mode shock pair. However, it is the disappearance of this feature that initiates a burp of magnetized material into the nose cone, and its re-appearance ends the forward ejection event. Once the quasi-stationary shock pair is re-established, it actually forces magnetized material *backward* along the axis (from $Z \approx 200$) and then outward into the cocoon at $Z \approx 150$. If an observed radio source displayed a quasi-periodic ejection of slow-mode shock pairs from a quasi-stationary shock region, it could be an important diagnostic of a strongly, toroidally-magnetized jet flow. Note that even the nose cone is strongly magnetized. Its toroidal field confines a gas pressure maximum on the axis that is up to 20–30 times that of the ambient medium.

So, an axisymmetric strongly-magnetized, super-magnetosonic jet develops a complex quasi-stationary magnetic structure that creates a new jet-like flow (the nose cone) that is internally super-slow-magneto-acoustic and trans-fast-magneto-acoustic, but still quite supersonic with respect to the surrounding ambient medium.

Relativistic 2-D Studies

In 1999 Serguei Komissarov performed relativistic axisymmetric simulations of the same phenomenon, with Lorentz factors up to $\gamma = 10$ [564]. His results showed that

nose cones develop in relativistic flows as well. However, he discovered a new and important point: it is not the value of the initial plasma β_p of the jet that determines the size of the nose cone, but rather the final value (β_{p2}) of the *shocked jet plasma* (beyond the Mach disk)

$$\beta_{p2} = \frac{1}{12\Pi} \left[(1 - 4\Pi) + (1 + 16\Pi + 16\Pi^2)^{1/2} \right]$$

where

$$\Pi \equiv \frac{B^2/4\pi}{(\gamma_{\text{jet}} - 1)\rho c^2} \approx \frac{B^2/4\pi}{\frac{1}{2}\rho V_j^2}$$

is the initial ratio of Poynting to kinetic energy flux in the jet. Requiring that a nose cone begins to form ($\beta_{p2} < 1$) means that the initial jet must have at least a moderate Poynting ratio ($\Pi > 0.2$).

15.2.2.3 Three-Dimensional MHD Jet Simulations

Non-Relativistic 3-D Studies

Of course, in order to fully understand the propagation of simulated magnetized jets, and their possible application to observed jets, we will need large studies of three-dimensional PFD flows – both sub- and super-magneto-acoustic – at highly relativistic speeds and with high spatial resolution. MHD simulation technology is not quite there yet, but should be within the next decade. High-resolution, three-dimensional relativistic MHD codes, capable of evolving for long periods of time, are now coming into existence.

Two studies that have all of these elements, except relativistic flow, were performed by Masanori Nakamura, with his colleagues at the Science University of Tokyo in 2001 [568], and with D. Meier at the Jet Propulsion Laboratory in 2004 [569]. The results of these studies have given great insight into the behavior of strongly-magnetized jets and have suggested possible explanations for at least some observed FR I radio sources.

The three-dimensional setup and evolution of one of Nakamura's simulations is shown in Fig. 15.28. A poloidal magnetic field threads all space, and a torsional Alfvén wave with plasma embedded (the jet) is injected at left. Initially, field and azimuthal velocity (color) are tightly wound. The presence of a small slow-mode shock pair can be seen at $t = 4.0$, nearly half-way down the jet, where the azimuthal velocity V_ϕ is lower (whiter) than elsewhere. By $t = 9.0$ this shock pair has lengthened considerably, and signs of kinking of the magnetic field are beginning to form near the reverse shock. Finally, by $t = 11.0$ the kinks in the field between the two slow mode shocks are fully developed.¹³ In the nonlinear phase ($t \geq 12.0$)

¹³ The helical kink instability is the $m = 1$ current-driven (CD) mode, not a Kelvin–Helmholtz (KH) one.

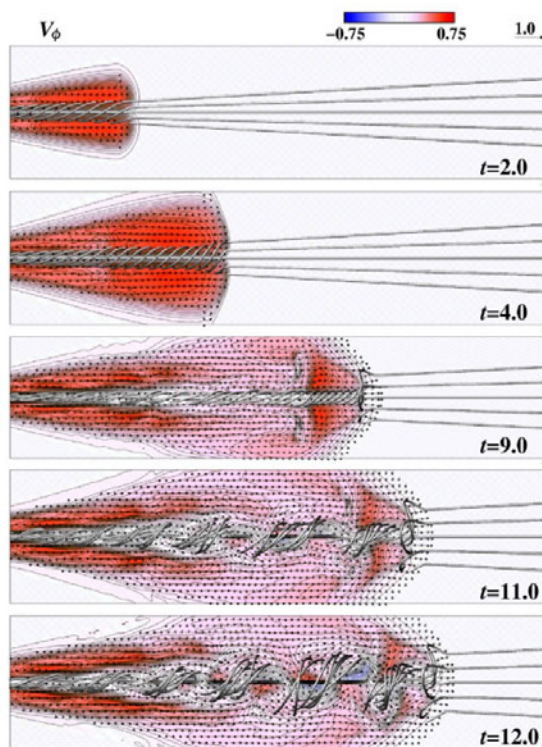


Fig. 15.28: Simulation of the three-dimensional development of the helical kink instability in an MHD jet. Magnetic field lines are shown as silver tubes, poloidal velocity vectors as arrows, and toroidal velocity V_ϕ of the plasma as a color scale. A sudden drop in the rotational speed (see also Fig. 15.25) causes the plasma to be unstable to the helical kink instability, which forms in the region between the two slow-mode shocks. Figure from Fig. 5.7 of [568]; reproduced with permission of M. Nakamura.

the kink propagates beyond the two slow shocks ($V_A > V_S$, after all), forward and backward.

In our 2004 paper we looked at this helical kink instability in more detail, studying how it varies with the amount of Poynting-flux domination and with the steepness of the density and pressure gradient in the surrounding *external* medium. Our results are presented in Figs. 15.29 and 15.30 and can be summarized as follows:

- Three-dimensional super-slow-mode jets develop on their axis the classical three-shock structure seen in the one-dimensional simulations (compare each V_Z plot in Fig. 15.29 with that in 15.25).
- The development of a nose-cone-like feature (a super-slow flow well ahead of the forward slow shock) is facilitated, as expected from Komissarov's work, by increasing the Poynting-flux domination of the flow (compare in Fig. 15.30 Model A-2 with A-1); the low ($\Pi = 1.5$) PFD jet (A-2) looks similar to a classical HD

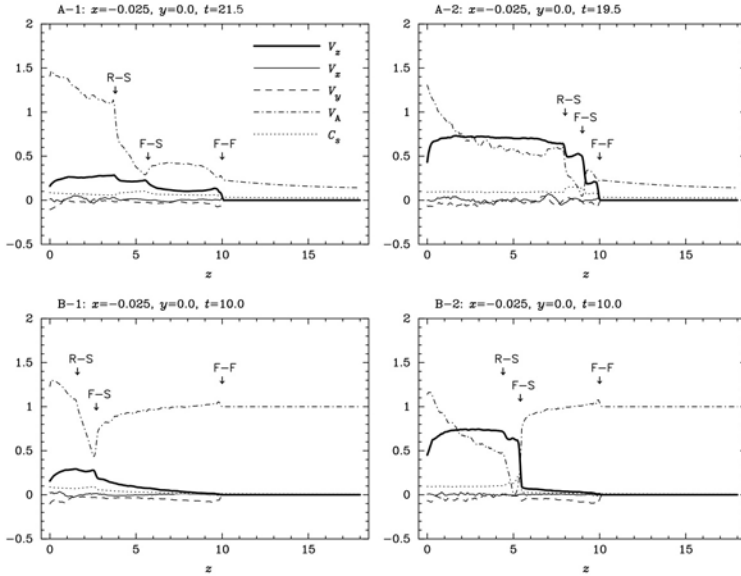


Fig. 15.29: Plots of various velocity quantities for four different three-dimensional super-slow magneto-acoustic jet simulations. Only the thick solid line (V_Z) is of interest here. *Left column:* Highly Poynting-flux-dominated ($II \approx 0.9$; Case 1) simulations. *Right column:* Mildly Poynting-flux-dominated ($II \approx 0.6$; Case 2) simulations. *Top row:* Simulations with a shallow gradient in the external medium density ($\rho \propto Z^{-2}$; Type A). *Bottom row:* Simulations with a steep gradient in external density ($\rho \propto Z^{-3}$; Type B). Note the position of the reverse-slow, forward-slow, and forward-fast shock or wave fronts. Reproduced from Fig. 8 of [569], by permission of the AAS.

jet, while the high ($II = 9$) PFD (A-1) looks like a two-dimensional nose-cone jet.

- What was *not* expected was that development of a fast nose cone *also* is facilitated by a steep density and pressure gradient (compare in that same figure A-2 with B-2 or even A-1 with B-1). Furthermore, the steep gradients in the external medium tend to produce some additional stability to helical kinking.
- Each time a helical kink develops in the flow, it starts to do so between the two slow-mode shocks.
- In the highly PFD jets (A-1 and B-1), the kink is more pronounced, as expected, but it does not appear to ever fully disrupt the jet.

Analysis by Nakamura and Meier showed that it is the rotational inertia of the plasma that stabilizes the MHD jet against the Kruskal–Shafranov criterion for forming a three-dimensional helical kink (V_ϕ must exceed $V_{A\phi}$). And it is the increase in toroidal magnetic field strength in the rotating frame between the two slow shocks that triggers this instability. This poses an interesting question for future *relativistic* MHD simulations: can relativistic inertia of a rotating, magnetized plasma render even highly PFD jets stable to the helical kink? Or, are all PFD jets subject to the helical kink on some level?

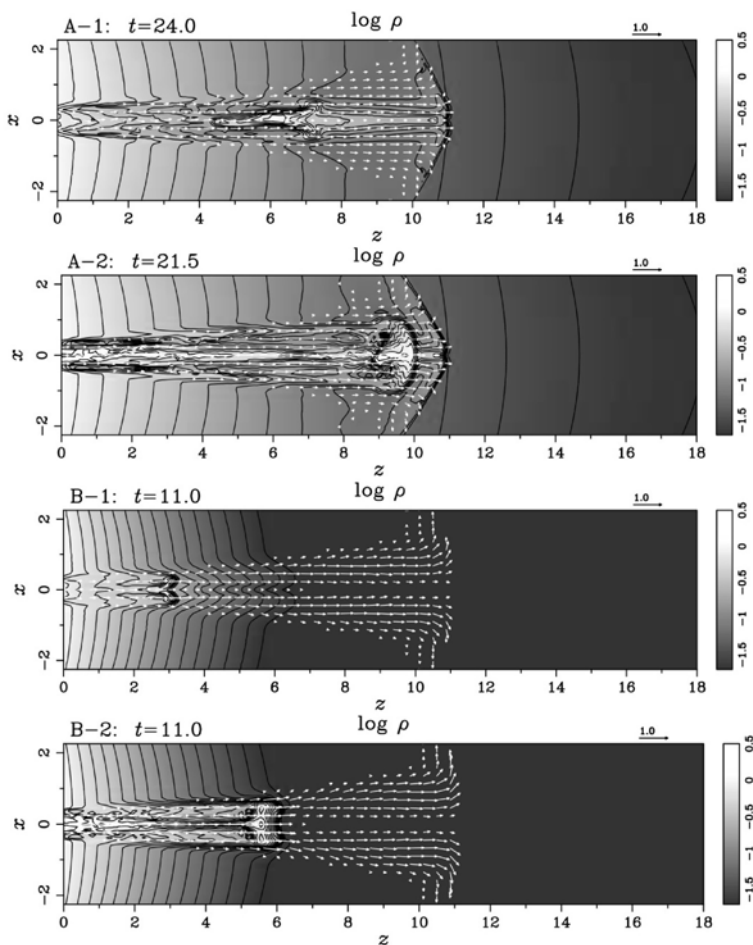


Fig. 15.30: Plots of all four simulations in Fig. 15.29 when each had reached roughly the same distance from the injection point ($Z \sim 10$). Contours and shading show the density distribution, and poloidal velocity vectors are shown as arrows. The highly Poynting-dominated simulations (A-1 and B-1) show a large fast MHD wave preceding the main jet. Increasing the steepness of the external density gradient (going from Case A-1 to Case B-1) increases the prominence of this wave. Decreasing the Poynting-flux domination (going from case A-1 to A-2, or going from case B-1 to B-2) decreases the prominence of this wave. In simulation A-2 the fast wave barely leads the forward-slow shock. Adapted from Figs. 4b, 5b, 6b, and 7b of [569]. Reproduced by permission of the AAS.

Yosuke Mizuno of the University of Alabama in Huntsville and his colleagues Philip Hardee (of the University of Alabama, Tuscaloosa) and Ken-Ichi Nishikawa (also of UAH) revisited this issue in 2011 with (essentially non-relativistic, $\gamma = 1.02\text{--}1.05$) MHD jet simulations [570]. They found that the CD helical kink instability also can be mitigated by surrounding the jet with a sheath of more slowly-moving wind. As the relative speed between jet and wind ($V_j - V_w$) increases, the shear surface between the jet and wind causes the kink to grow more slowly, propagate downstream with a greater speed, and transition to nonlinear structures (*e.g.*, disruption) at later times. The growth of the kink instability is further reduced if the helical twist of the field decreases with lateral distance from the jet body. So, even without relativistic effects, there appear to be several ways in which helical kink growth can be slowed significantly.

Relativistic 3-D Studies

One of the goals of performing three-dimensional *relativistic* simulations of jet propagation is to determine if, as in the non-magnetized hydrodynamic case, relativistic flow can help stabilize an MHD jet (see Section 15.2.1.3). Mizuno and his colleagues, therefore, undertook a study of $\gamma = 2.5$ jet flow and the effects of adding a magnetic field to the jet body and an external wind sheath around the jet [571]. The group's results for their mildly relativistic *hydrodynamic* jet were similar to those found by Hughes, Miller, and Duncan (Section 15.2.1.3): low Lorentz factor hydrodynamic jets develop KH instabilities rather quickly and fully disrupt.

However, when a magnetic field parallel to the jet velocity was added, the flow became much more stable to KH instabilities (see Fig. 15.31a–c).¹⁴ Furthermore, when a more slowly moving wind sheath was added around the jet, the KH instabilities were almost fully suppressed (Fig. 15.31d–f). Mizuno and his colleagues derived a stability criterion for such jets

$$(V_j - V_w)^2 < \frac{V_{As}^2}{\gamma_j^2 \gamma_w^2}$$

where the shear surface Alfvén speed is

$$V_{As}^2 = [\gamma_j^2 \rho_j \xi_j + \gamma_w^2 \rho_w \xi_w] \frac{(B_j^2 + B_w^2)}{4\pi \rho_j \xi_j \rho_w \xi_w}$$

and the subscripted quantities V , γ , ρ , $\xi \equiv (1 + \mathfrak{h}/c^2)$, $\mathfrak{h}$, and B are the jet or wind velocity, Lorentz factor, rest mass density, inertia per unit mass, enthalpy per unit mass, and magnetic field, respectively.

While these studies made significant advances in the understanding of KH instabilities, there still is the open question of the CD helical kink instability in rela-

¹⁴ Note that, because the magnetic field was not twisted, and the Poynting flux therefore vanished, these jets were *stable* to the helical kink.

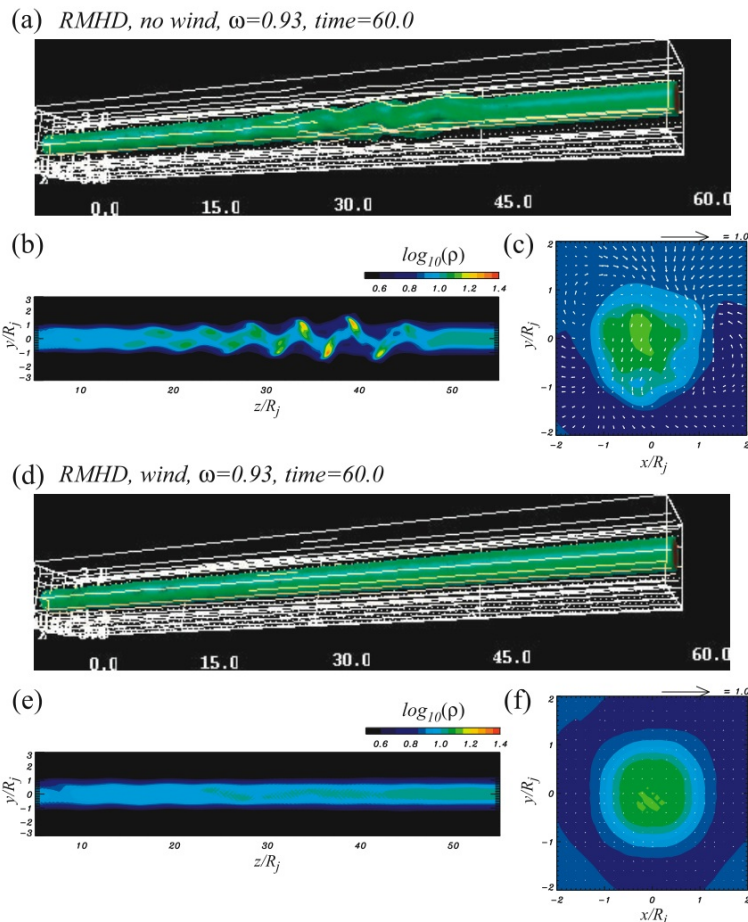


Fig. 15.31: Comparison of relativistic MHD jet propagation without and with a sheath of MHD wind surrounding the jet. *Top (a–c):* Relativistic ($\gamma = 2.5$) jet without a wind sheath. (a) three-dimensional volume rendering, (b) two-dimensional cut through jet body, (c) two-dimensional cross-sectional cut of jet at $Z = 30 R_j$. Colors show logarithm of density, white traces magnetic field lines, and arrows indicate plasma velocity. *Bottom (d–f):* Same as a–c, but with a $V_w = 0.5c$ ($\gamma = 1.15$) wind encasing the faster jet. The wind clearly enhances the jet’s stability. Reproduced from Fig. 2 of [571], by permission of the AAS.

tivistic flow. An initial investigation into this area was performed by Jon McKinney and Roger Blandford in 2009 [572]. They performed MRI/jet simulations like the McKinney one discussed in Section 15.1.4.3, but this time in full three-dimensional, relativistic GRMHD. They also studied the effects of different seed disk magnetic field topologies, particularly quadrupolar vs. dipolar.

The simulations that began with a seed *quadrupolar* magnetic field did *not* produce a steady relativistic ($\gamma \gtrsim 3$) jet. This was caused by plasma being regularly

leaked into the polar regions near the disk by unstable magnetic fields, making the jet heavy and difficult to accelerate.

Simulations that began with a seed *dipolar* magnetic field *did* produce a steady relativistic jet, up to $\gamma \approx 10$ and an half-opening angle of $\sim 5^\circ$ at a distance of $Z \approx 10^3 r_g$. This occurred in spite of strong, non-axisymmetric turbulence in the accretion flow. Fig. 15.32 is the three-dimensional analog to Fig. 15.14, for a black hole with a very similar spin ($j = 0.92$). The non-axisymmetric distortions, most visible in Fig. 15.32(right) appear to originate in the accretion flow, but do not seem to affect the stability of the axial relativistic jet (orange material in that panel). This high-speed flow has many of the properties discussed above that tend to stabilize jets to KH and/or CD instabilities: relativistic flow speeds ($\gamma \sim 10$), relativistic rotation of the jet plasma and magnetic field, steep external density/pressure gradients, strong axial magnetic field, and a surrounding wind sheath (yellow and blue material in Fig. 15.32(right)). This rather stable jet is on its way toward a possible magnetosonic horizon, ultimately to form an FR I or FR II radio source.

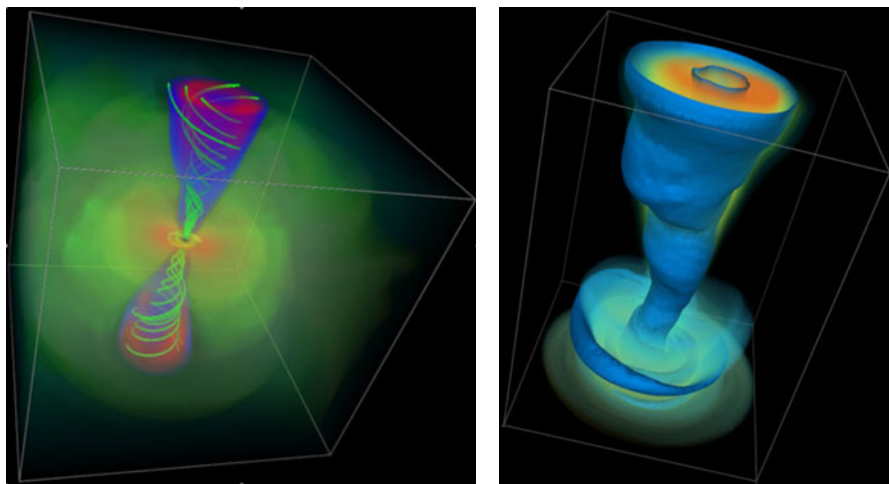


Fig. 15.32: Volume rendering for a fully three-dimensional GRMHD simulation of an MRI turbulent disk that launches a large-scale jet from the vicinity of a $j = 0.92$ black hole. (Compare with two-dimensional simulation in Fig. 15.14.) *Left:* The inner $100 r_g$ (like Fig. 15.14A) at $t = 4 \times 10^3 \tau_g$, showing disk (yellow/orange), wind sheath (green), and relativistic MHD jet (red/blue color and green field lines). *Right:* Northern hemisphere for full simulation at same time, out to $10^3 r_g$. (The linear length of this jet is about 10% of that in Fig. 15.14B.) Shown are the outer disk and wind (cyan near bottom), outer and inner boundaries of the outer perturbed jet (blue), and the axial $\gamma = 10$ relativistic jet (orange). Reproduced from Figs. 1 and 2 of [572], by permission of J. Wiley and Sons, Ltd.

15.2.3 Possible Scenarios for Behavior of Jets beyond the Magnetosonic Horizon

While the definitive three-dimensional relativistic MHD simulations have not yet been performed, we can synthesize a possible scenario for jet behavior as they exit the central engine's magnetic nozzle from the results of the previous two subsections. Much more sophisticated future simulations will determine whether the scenario below is valid or must be modified.

15.2.3.1 Passage through the Magnetosonic Horizon

As the jet flow approaches the final separatrix surface (FMSS), it already is highly super-fast-magnetosonic, as indicated by the self-similar models. After passing through the FMSS even the component of jet speed *toward* the axis ($-V_\theta$) exceeds the fast magneto-acoustic speed. The flow, then, will quickly converge near the axis and likely form a strong recollimation shock, slowing the flow from a highly super-fast speed to a trans-fast-magneto-acoustic one. Because this shock will be associated with the largely stationary FMSS, it will appear as a standing feature in the flow, at least on time scales of order the travel time from black hole to FMSS. Beyond the recollimation shock the flow will be decidedly *non*-self-similar and possibly even time-dependent and non-axisymmetric. Stationary, axisymmetric self-similar models, therefore, will not be useful beyond this shock.

15.2.3.2 Passage through the Strong Recollimation Shock

The Ideal MHD Case: A Possible Model for Magnetized FR I Jets

When the plasma enters the strong, stationary recollimation shock, its energy flow will be about equally divided between Poynting and kinetic energy flux and its internal energy will be quite cold. If the flow emerges from the shock with little magnetic field dissipation, then the plasma will remain highly kinetic and magnetized, but now be very warm, as some of the kinetic energy has been converted into internal energy. To first order, magnetic and internal particle energy will be in rough equipartition. Therefore, its behavior will be much like the MHD simulations in the previous subsection. The jet will retain a large amount of Poynting flux and develop a fast, long “nose cone” jet ahead of the recollimation shock, as we have seen in both the two-dimensional and three-dimensional simulations. Well downstream of this shock, however, the internal energy generated there will have been radiated, most likely by synchrotron and Compton processes, leaving the jet partially kinetic and magnetized but again relatively cool.

Because the pre-recollimation-shock flow is super-fast-magneto-acoustic, the post-recollimation-shock flow may periodically eject slow-mode shock pairs into

the nose cone flow from an oscillating magnetic chamber (see Fig. 15.27).¹⁵ Each of the slow shock pairs will travel downstream at the rapid nose cone speed; however, they each will lengthen at a rate that is a few times the slow speed as they do so. The jet, therefore, will display multiple and detached helical kink structures along its length, each growing longer and broader laterally as they travel away from the galaxy center and down an ambient pressure gradient. Since little lobe/cocoon structure occurs before the standing recollimation shock (*i.e.*, very near the galaxy; see Fig. 15.27), and only the nose cone beyond it, the jet should appear as a broadening flow beyond this shock with no apparent cocoon – that is, as an FR I source.

The Non-Ideal MHD Case: FR II Jet Formation?

The behavior of the jet when the flow through the recollimation shock is magnetically dissipative is much more speculative at this point. Indeed, even the dissipation mechanisms are still unclear (anomalous resistivity, Weibel instability, two-stream instability, *etc.*). This is an important quest for GRBs, for example, but we will not speculate on this issue here. However, the net effect of dissipation of the magnetic field will be to produce rapid conversion of magnetic energy to internal particle energy in flaring, reconnection events. This will considerably reduce the jet internal magnetic field energy but not reduce the kinetic energy any more than in the non-dissipative case. The flow will be quite hot initially, due to massive particle acceleration in both the recollimation shock and the flares. Again, downstream of the recollimation shock heating and particle acceleration will subside, and the plasma will be cooled by radiative processes. The difference between this and the ideal MHD case is that the jet will be dominated by kinetic flux only, with both its magnetic and internal energies being rather weak.

Eventually, this kinetic flow will be decelerated at a termination shock (Mach disk) far from the galactic nucleus, causing the source to appear similar to an FR II one. As pointed out in Section 15.2.1, if kinetic FR II-type flows are disrupted by strong external shocks, they will form FR I-type lobes even without the aid of strong magnetic fields. So, there may be multiple mechanisms for forming sources with an FR I morphology.

¹⁵ Note, however, that the existence of such a device is not certain, as we have seen such a chamber in two-dimensional, axisymmetric, super-magneto-acoustic simulations but not in three-dimensional, sub-magneto-acoustic ones.

Chapter 16

Putting it All Together: Black Hole Engines of All Sizes

This chapter attempts to summarize what we know, and do not know, about how black hole engines of all known sizes operate. It is meant as a companion chapter to the first four, but written with the benefit of the remaining eleven. The goal is to unite black hole theory and observations to present a broad, if incomplete, picture of black hole engines in the universe. The chapter includes some basic and well-known information about these objects, some new and recently published ideas, and even some more speculative concepts that look good as explanations for what is going on, but have not yet been tried by fire in discussions in the astrophysical literature and the greater community. Because the newest of these ideas may evolve during the referee process, and even after publication, we shall be careful to note which material is fairly well accepted by astrophysicists and which is newer and not established. It is this chapter, along with Chapter 4, that perhaps will require the most revision in future editions of this book. Such is the nature of astrophysical science.

16.1 Neutron Stars: $1 M_{\odot}$ Near-Black Holes

When discussing each class of object, I find it useful to compare object sizes to something nearby and familiar. In this case a $1.4 M_{\odot}$ neutron star has a radius of ~ 10 km and, therefore, is similar in size to a good-sized town, but it has the mass of the sun.

16.1.1 Neutron Star Formation

It is generally agreed that neutron stars are formed in iron-core-collapse supernova explosions of stars whose mass is in the approximate range $9\text{--}21 M_{\odot}$ (see Sections 3.3 and 10.1). This conclusion is deduced from measuring the abundances of heavy

elements in the universe and comparing with accurate simulations of what elements are made in the interiors of stars in this mass range.

Unfortunately, while there is strong evidence that stars in this mass range explode, supernova theorists believe that they understand how the core-collapse supernova process works only for stars at the low end of this range. During the core collapse the “accretion” of infalling iron occurs on a dynamical time scale (equation (11.66)), with the rate approaching $10 M_{\odot} s^{-1}$ ($10^{34} g s^{-1}$; equation (11.68)). In this extremely dense state the hot proto-neutron star (PNS) is optically thick to neutrinos, and neutrino radiation transfer can play an important role, especially if it can heat the infalling material near the accretion shock. This can reverse the infall and produce an outflowing wind that drives away the stellar mantle and envelope. This outflow will appear as an explosion to observers.

However, this process works only in the narrow mass range (9–11 $M_{\odot}$ – essentially only B2 stars). Unfortunately, these stars account for less than one-third of all known core-collapse supernovae. In detailed numerical simulations of stars in the mass range 11–21 $M_{\odot}$ infall of the massive iron core overwhelms the outward pressure of the neutrino-heated material, crushing the erstwhile neutron star into a black hole. Nature clearly does not leave black holes for most stars that begin with masses of 9–21 $M_{\odot}$, so there must be another method by which most neutron stars are created. The method that this author favors is the explosive action of an MHD bipolar outflow from just outside the rapidly-rotating, highly magnetized PNS itself.

16.1.1.1 Phenomenological Arguments

Phenomenological arguments alone, that MHD jets play an important role in supernovae, can be quite compelling. Fig. 16.1 shows a schematic diagram of the evolution of stars of all types and the branches that jet formation in those stars can take. When stars are first formed, they almost universally produce protostellar jets. It is likely that such jets are the process by which protostars shed their excess angular momentum (acquired from the infalling molecular cloud material) and become the relatively slowly-rotating objects they are on the main sequence. (See, *e.g.*, equations (15.12).)

Jets also are produced by stars at the *ends* of their lives. Bipolar outflows are quite common in dying stars: planetary nebulae and γ -ray bursts both are believed to be jetted outflows (Section 3.6). Furthermore, even dead stellar remnants can be re-kindled and can produce jets when material accretes onto them from the interstellar medium (ISM) or from a binary companion. Finally, even young isolated pulsars produce jets simply by creating particles in their rotating magnetospheres. In short, any time there is stellar shrinkage, collapse, or accretion (situations where a large amount of angular momentum will likely be added to a magnetized stellar object or remnant) a jet will be generated.

The “odd one out” in Fig. 16.1 is the core-collapse supernova event (dashed ellipse region). It also is an infall event, the proto-neutron star should have a significant amount of angular momentum, and the remnant (a pulsar or magnetar) definitely

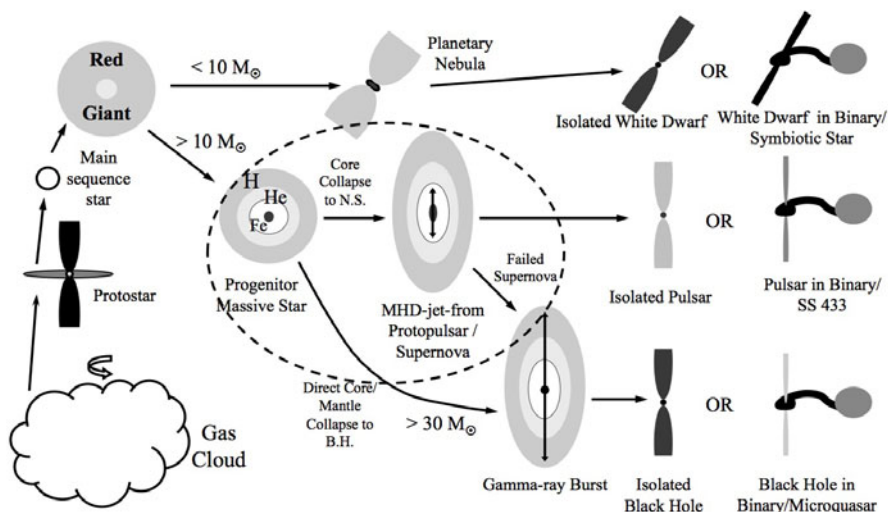


Fig. 16.1: The formation, evolution, and death of stars, with an eye toward neutron star, black hole, and jet formation. *Left:* Collapse of an interstellar cloud forms a protostar (with a jet) that lives most of its life on the main sequence, becoming a red giant (or supergiant) in the late stages of its life. *Top:* Solar-type stars end their lives in the planetary nebula and white dwarf phase, which often produces a bipolar outflow, during and after death. *Middle and bottom:* Massive stars produce neutron star or black hole remnants. When black holes form, powerful jets also are produced, in a γ -ray burst event. A simple phenomenological unification model (dotted ellipse) states that when neutron stars are formed, powerful jets are produced as well, in supernova events. Reprinted from Fig. 25.2 of reference [573], with the permission of CUP. Copyright ©2006 Cambridge University Press.

will be magnetized (see Section 10.1.2). Therefore, there is strong circumstantial evidence that these events also will involve strong bipolar MHD flows, just like their lower-mass and higher-mass siblings.

16.1.1.2 Observational Evidence

The best observational evidence that collapsing supernova cores *also* produce bipolar outflows is the recent work by Lifan Wang, Craig Wheeler, and their colleagues (see Section 3.3). Asymmetric supernova explosions, therefore, may be the missing link between dying low-mass stars that produce bipolar planetary nebula jets and massive stars that produce GRB jets. And the additional energy from the supernova core jets provides an attractive explanation for why stars in the range $11\text{--}21 M_{\odot}$ (and not just $9\text{--}11 M_{\odot}$) leave neutron star remnants, and spread their nucleosynthetic products around the Galaxy (instead of locking all of this material up into black holes).

16.1.1.3 Theoretical Arguments

The energetics for MHD jet-driven supernova explosions were discussed in Section 10.1.2 where we found that a rotating proto-neutron star with a period of ~ 120 ms (similar to that computed by Fryer and Heger for evolved post-collapse cores in rotating stars) would generate about 10^{51} erg of additional explosion energy in a bipolar outflow (similar to what Woosley needed to match the heavy element abundances in his supernova simulations). More details of MHD jets themselves were discussed in Section 15.1.1, where the MHD power generated in such a non-relativistic outflow was given in equations (15.10). This luminosity can be cast in a form similar to the pulsar power equation (14.6)

$$L_T \approx \frac{B_{\text{PNS}}^2 R_{\text{PNS}}^3 \Omega_{\text{PNS}}}{2} \\ = 4 \times 10^{50} \text{ erg s}^{-1} \left(\frac{B_{\text{PNS}}}{10^{15} \text{ G}} \right)^2 \left(\frac{R_{\text{PNS}}}{50 \text{ km}} \right)^3 \left(\frac{P_{\text{PNS}}}{1 \text{ s}} \right)^{-1} \quad (16.1)$$

with light cylinder magnetic field and radius replaced by the proto-neutron star magnetic field and radius. For a PNS cooling time of ~ 1 s, we find that, to generate $\sim 10^{51}$ erg in this amount of time, the PNS magnetic field has to be fairly high ($\sim 10^{15}$ G), strong compared to pulsar fields (10^{12-13} G) but weak compared to the virial limit ($\sim 10^{17}$ G). In 2003 Shizuka Akiyama and her colleagues at The University of Texas found that $> 10^{15}$ G fields can be generated in a fraction of this cooling time by the magneto-rotational instability (Section 9.6.3) above a newly-formed PNS that is differentially rotating with respect to the infalling iron mantle [574].

One thing is clear about MHD jet-driven explosions, if they are to be the solution to the supernova problem: any such bipolar outflow cannot originate inside the proto-neutron star. In 1976 this author, and his colleagues at The University of Texas, showed that such a jet would eject far too much highly-neutronized material into the ISM (Section 10.1.2 and [575]). When summed over all core-collapse supernovae, it would overproduce the heavy “*r*-process” elements by a factor of 1000. Taking the outflow launch site to be outside the PNS – in the infalling iron region where the binding energy is less – also makes it easier to form an MHD explosion.

Finally, it is worth noting that an MHD bipolar jet of ionized iron also could be the source of the pulsar rocket needed to propel the Crab and Vela Pulsars, and probably many others, to velocities of $100\text{--}200 \text{ km s}^{-1}$ (Section 3.1.1). For a bipolar outflow of 10^{51-52} erg, a 150 km s^{-1} neutron star kick, *along the pulsar rotation axis*, could be produced with a jet asymmetry of only 3–30%.

So, while detailed supernova simulations in two or more dimensions, with magnetohydrodynamics included, still need to be done, all aspects of this possible solution to the core-collapse supernova problem look good: energetics, time scales, MHD, actual observations of bipolar supernova explosions, and even similar phenomenology to the behavior of lower and higher-mass stellar deaths. The model

represents quite a paradigm shift from relying on nuclear physics and hydrodynamics alone to blow up stars. But it may be a shift worth making in the long run.

16.1.1.4 Other Methods of Neutron Star Formation

Besides the core-collapse supernova mechanism discussed above, Table 10.1 shows that there really are very few other processes that end up with a neutron star where there was not one existing already. Surprisingly, accretion from a normal star onto a C/O white dwarf in a cataclysmic variable (“singly-degenerate” star) definitely does *not* work (Section 10.3). This process slowly increases the white dwarf mass toward the Chandrasekhar limit of $1.4 M_{\odot}$, but then violently ignites a carbon detonation (Type Ia) supernova before the critical mass is reached.

Only one other possibility is known: coalescence of doubly-degenerate (WD+WD) binaries. However, as discussed in Section 10.3.6, 96% of all such binaries are either not old enough to have merged yet or not massive enough to form a $>1.4 M_{\odot}$ end product. And, of the 4% remaining, most of those mergers will produce a C/O WD that explodes as a Type Ia supernova. Only occasionally will an ONeMg white dwarf (with little carbon) combine with another WD, avoid the carbon detonation stage, and collapse to a neutron star – a process several orders of magnitude more rare than standard core-collapse supernovae occurring inside stars $>8 M_{\odot}$.

16.1.2 Isolated Neutron Stars

Moving to the right of the neutron star-formation event in Fig. 16.1, we reach the isolated neutron star stage, after the stellar envelope has been fully dispersed into a supernova remnant. An isolated pulsar can become visible in at least three ways: (1) accrete material from the ISM, (2) generate its own pulsar emission if it possesses a strong enough magnetic field and rapid enough rotation, and (3) if the magnetic field is strong enough, generate X-ray and γ -ray emission when that field decays or reconnects in a *magnetar* outburst.

16.1.2.1 Neutron Stars Accreting from the Interstellar Medium: Dead Pulsars

Accretion onto an isolated neutron star, especially one with a “runaway” speed of $\sim 150 \text{ km s}^{-1}$ is, by far, the least important of these three. Such objects are more furnaces than engines, producing some heat and light but little mechanical power.

From equation (11.9) the accretion rate of a $1.5 M_{\odot}$ object traveling at this speed through the ISM ($n = 1\text{--}100 \text{ cm}^{-3}$) is only 2×10^8 to $2 \times 10^{10} \text{ g s}^{-1}$, yielding $10^{28.3\text{--}30.3} \text{ erg s}^{-1}$ in luminosity. In Eddington units, this accretion rate is only $\dot{m} = 2 \times 10^{-10}$ to 2×10^{-8} , which would be optically thin and emit in the hard X-ray ($\sim 100 \text{ keV}$).

As discussed in Section 3.1, we expect a couple hundred dead pulsars within Geminga's distance (~ 157 pc), with the nearest being ~ 40 pc away. In 2012 the Nuclear Spectroscopic Telescope Array (NuSTAR) focusing hard X-ray telescope will launch. At 50 keV X-ray energies NuSTAR will have a 3σ detection limit of $10^{28-29} \text{ erg s}^{-1}$ for sources at distances of 40–160 pc. This would allow the spacecraft to see isolated dead neutron stars as hard X-ray sources even if they were streaming through nearby interstellar clouds only $1\text{--}10 \text{ cm}^{-3}$ in density. So, even weak, isolated neutron stars might be detectable in the very nearby solar neighborhood.

16.1.2.2 Rotating Magnetized Neutron Stars: Live Radio Pulsars

The energy that powers radio-emitting pulsars comes, not from accretion, but from the rotational energy of the neutron stars themselves, with periods ranging from a few milliseconds to a few seconds. And the means by which each pulsar taps that rotational energy is the star's (electro-) magnetic field. They are giant electric generators that drive currents (and particles) into the ISM. More importantly, however, if the concept of MHD jet production during neutron star formation that was presented above is correct, then pulsars are the remaining vestiges of the powerful central engine that exploded the star long ago.

The electromagnetic power is given by the now familiar formula (*e.g.*, equations (14.6), (14.5), and (14.3))

$$\begin{aligned}
 \mathcal{P}^{\text{EM}} &= \frac{2 B_L^2 R_L^3 \Omega_L}{3} \\
 &= \frac{2 B_L^2 R_L^4 \Omega_L^2}{3 c} \\
 &= \frac{2 B_{\text{NS}}^2 R_{\text{NS}}^6 \Omega_{\text{NS}}^4}{3 c^3} \\
 &= 4 \times 10^{35} \text{ erg s}^{-1} \left(\frac{B_{\text{NS}}}{10^{12} \text{ G}} \right)^2 \left(\frac{R_{\text{NS}}}{10 \text{ km}} \right)^6 \left(\frac{P_{\text{NS}}}{100 \text{ ms}} \right)^{-4}
 \end{aligned} \tag{16.2}$$

where $\Omega_{\text{NS}} = 2\pi/P_{\text{NS}}$ is the angular speed of the pulsar rotation (and of its strong magnetic field), and R_{NS} and B_{NS} are the neutron star radius and surface magnetic field strength. The “light cylinder” radius $R_L = c/\Omega_L$ is the distance outward along the equator to where the tangential rotation speed of the magnetic field equals the speed of light,¹ and the magnetic field strength there $B_L = B_{\text{NS}} (R_{\text{NS}}/R_L)^3$ is the dipole magnetic field's strength at that radius.

Most of this electromagnetic power is used to generate an outflowing wind of particles and very low-frequency ($\nu = \Omega_{\text{NS}}/2\pi$) electromagnetic waves, creating a

¹ Even though the magnetic field can rotate at speeds faster than that of light, the plasma trapped in this field does not exceed, or even reach, light speed. This is because the rotating magnetic field bends backward, allowing the plasma to follow the laws of special relativity, while the field slips through it at super-light speeds.

giant pulsar wind nebula or *plerion* a few parsecs in size. This nebula emits throughout the electromagnetic spectrum, particularly in the radio, optical, X-ray, and γ -ray. Twin jets are generated in the direction of the pulsar rotation axis, in the manner described in Chapter 15. As pointed out in Section 3.1.1, asymmetry in these observed jets could *not* be responsible for the cosmic rockets that propelled the Crab and Vela Pulsars to $>100 \text{ km s}^{-1}$ speeds along their rotation axes. Even if all the energy in equation (16.3) over the Crab Pulsar’s 1000-year lifetime went into the jet ($E = 10^{49} \text{ erg}$), the *total* momentum in the jet ($p = E/c = 5 \times 10^{38} \text{ g cm s}^{-1}$) would be only 1% of that needed to produce the Crab Pulsar’s 180 km s^{-1} motion. As pointed out above in Section 16.1.1, the pulsar kick easily could have been provided by a powerful MHD jet along the rotation axis that was produced in the original core collapse. The current pulsar jets in Fig. 3.3, therefore, are simply a faint (13 orders of magnitude weaker) remnant of the explosive MHD event that took place when the pulsar was formed.

The radio pulsar emission, which is used routinely to find these objects, is another *ten* orders of magnitude fainter ($10^{25-28} \text{ erg s}^{-1}$) than the plerion itself. This is both a testament to the extraordinary sensitivity of radio telescopes and to the fact that this emission is highly beamed along the magnetic dipole (not rotation) axis.

The origin of this radiation is one of the most interesting aspects of pulsar theory. It can be shown (Section 14.1.2) that a rotating magnetic field *must* have electric charges trapped within that field, at the level of the “Goldreich–Julian charge density” (equation (14.9)). But the pulsar particle wind is so powerful that all of this charge would be ejected in less than one rotation period – a fraction of a second. How, then, are these charges replenished, in order to maintain the electromagnetic structure of the pulsar? The non-equilibrium that is generated by a lack of charges creates enormously strong electric fields (10^{12} V). These electric fields, in turn, create and accelerate $e^{\pm}$ pairs with Lorentz factors of $\gamma_{e\pm} \sim 10^6$ or more. The process reaches a quasi-steady state, where most of the Goldreich–Julian charge is maintained, but enough of an electric field also persists to keep the particle creation replenishing those charges carried out by the wind.

The primary places where these “spark gaps” are created are in the pulsar *magnetic* polar caps (see Fig. 14.5). Synchrotron and self-Compton radiation is generated in the optical–X-ray range, along with radio radiation generated from small disturbances in the polar cap magnetic field.

16.1.2.3 Flares in Strongly-Magnetized Neutron Stars: Magnetar Emission and Outbursts

When a rotating neutron star’s magnetic field is in the magnetar range (10^{14-15} G), its lifetime in the pulsar phase can be quite short – so short (months to a few years) that we have yet to observe a *pulsar* of this type. In Section 14.1.1, as a result of Spitkovsky’s simulations, we found out that the spindown rate of a pulsar is even faster when the magnetic field poles lie at an angle $\Theta > 0$ to the rotation axis. This leads to a $P - \dot{P}$ formula (introduced in Section 3.1.1 and originally derived in equa-

tion (14.4)), that yields magnetic field 20–40% weaker than we thought (equation (14.11))

$$B_{\text{NS}} = 2.6 \times 10^{19} \text{ G} \left[\frac{P_{\text{NS}} \dot{P}_{\text{NS}}}{1 + \sin^2 \Theta} \right]^{1/2}$$

For $\Theta = 90^\circ$, this gives a spindown time of

$$\begin{aligned} \tau_{\text{spindown}} &= \frac{P_{\text{NS}}}{2 \dot{P}_{\text{NS}}} \\ &= 7 \times 10^4 \text{ yr} \left(\frac{M_{\text{NS}}}{1.5 M_\odot} \right) \left(\frac{B_{\text{NS}}}{10^{12} \text{ G}} \right)^{-2} \left(\frac{R_{\text{NS}}}{10 \text{ km}} \right)^{-4} \left(\frac{P_{\text{NS}}}{100 \text{ ms}} \right)^2 \end{aligned} \quad (16.3)$$

Normal pulsars (10^{12-13} G), therefore, have spindown ages of 10^3-5 yr . Magnetars, however, have spindown ages of only a few weeks to 7 yr.

Now, pulsar death is believed to occur when the rotating, magnetized neutron star no longer can produce electron–positron pairs in its magnetic polar caps. Empirically, in Fig. 3.1, we see that this occurs when the rotational period becomes greater than about 5 s [576]. (The magnetar candidates generally have such long periods and have no detectable radio emission.) Magnetars, therefore, can be thought of as “dead” pulsars that have spun down long ago and now are producing their anomalous X-ray (pulsating) radiation and the repeating soft γ -ray outbursts by means other than the classical pulsar processes.

Although we have not discussed much magnetar physics in this book, in the interest of completeness, we shall include a brief discussion of the processes in these stars, drawing mainly from a recent review [577]. The main process that produces the X-ray emission in AXPs is thermal emission from a warm neutron star crust, heated by decay of the strong magnetic field. This decay occurs because of the Hall effect (*e.g.*, the h_q term in equations (9.33) and (9.40)) and not electrical resistivity (the η_q term in those equations). Initially, the very strong magnetic and neutron star crustal forces are in static equilibrium; however, as the field decays, stresses build up on the crust, causing it to crack and shift which, in turn, heats the crust to X-ray temperatures.

In the repeating outbursts of SGRs, a more energetic magnetic process occurs. While a magnetar is spinning slowly enough that radio pulsar emission is no longer produced, the magnetic dipole radiation is still applying a torque to spin down the rotating neutron star. This torque is applied through the magnetic field, into the neutron star crust, and eventually is felt through viscous forces throughout the star. As in accretion disks, this is believed to initially enhance the azimuthal magnetic field in and below the crust through shear in that crust. And, as in accretion disks, the enhanced azimuthal field repeatedly decays in flaring reconnection events. These flares are similar to solar flares, but they produce so many $e^\pm$ pairs that the plasma is optically thick and produces thermal, rather than nonthermal, γ -ray emission. The

energy released in these flares ranges from 10^{41} to 10^{45} erg s $^{-1}$, the latter being the very rare outbursts (only one per star per 50–100 yr).

16.1.3 Accreting Neutron Stars in Binary Systems

The case of isolated neutron stars was complicated enough, with parameters like magnetic field strength and rotation speed. The addition of a binary companion and the possibility of mass transfer to the compact object increases the complexity even more. We now must contend with the rate of accretion onto the neutron star, the type of companion star (high-mass or low-mass main sequence star, white dwarf, *etc.*), and even the type of orbit in the binary system. We will not be able to treat all of the different possibilities in detail here. But we will try to outline the most important ones and make a theoretical connection with the observations discussed in Section 3.2.

In that section we learned that there are two main types of accreting neutron stars in binary systems: the very young (<20 Myr) high-mass X-ray binaries (HMXBs) with a massive O/B companion star of $10\text{--}20 M_{\odot}$, and the rather old (>1 Gyr) low-mass X-ray binaries (LMXBs) with a low-mass companion of $\sim 1\text{--}2 M_{\odot}$ or less. The neutron star in an HMXB tends to be a pulsar in the X-rays. Its magnetic field strength is similar to that of field radio pulsars ($10^{12\pm 1}$ G), but the periods tend to be much longer, 0.05–700 seconds [578]. A likely reason for this spindown is the propeller mechanism in the presence of accretion (see below and Section 14.2). Accretion in the HMXBs tends to occur by the wind mechanism. From equation (11.65) wind accretion rates depend on the mass of the compact object, the orbit, and the strength of the wind and tend to lie in the range $0.001 < \dot{m} < 1$. At the high accretion rates the plasma will be in a geometrically thin, Keplerian disk with an emission radiation temperature of ~ 1 keV (softer X-ray region; equation (12.11)), while at the low accretion rates the inflow will be in a geometrically thick, optically thin, nearly spherical ADAF with an emission temperature that is nearly 100 times higher (hard X-ray region; equation (12.58)).

The LMXB systems further subdivide into three types: X-ray pulsars, Z sources, and atoll objects. Here we will adopt the general classification [189, 190] that atoll sources have a weak magnetic field ($10^{8\pm 1}$ G), Z sources a moderate field ($10^{10\pm 1}$ G), and X-ray pulsars a strong field ($10^{12\pm 1}$ G). The pulsar periods are often very short – a few milliseconds. Accretion in the LMXBs tends to occur by the Roche lobe overflow mechanism, either on a nuclear time scale ($\dot{m} < 0.005$; equation (11.64)) or on the shorter thermal time scale ($\dot{m} \leq 10$; equation (11.63)), leading to similar low/hard or high/soft states in the LMXB case as well. This accretion again is the likely reason for both the wide variation in magnetic field among the different sources and the unusually *short* pulsar periods: after billions of years of being assaulted by accreting plasma, the magnetic fields decay from their initial high strengths; furthermore, the Roche lobe overflow mechanism tends to transfer a lot of orbital angular momentum (with similar vector orientation) to the neutron star

over those billions of years, spinning up the neutron star to nearly breakup speeds (equation (11.70)).

Below we shall concentrate on the very diverse LMXB source class. Using theoretical arguments from Chapters 12 and 14, we then will show how the different magnetic field strengths can lead to these seemingly different types of sources.

16.1.3.1 Binary X-Ray Pulsars

To make matters easier we shall start by looking at the binary X-ray pulsars. Ironically, these strongly-magnetized objects are the simplest. Figs. 3.5b and 14.9A give a good idea of how these objects might appear. The magnetic field is so strong that, in many cases, the accretion flow reaches the surface of the neutron star *only* by traveling along the strong field lines and accreting onto the magnetic poles.

While we do not have a detailed theory of accretion in this situation, we still can develop a simple idea of how such an object might appear spectrally. It should not take a very rapid accretion rate to cause the accreting spot at each magnetic pole to be optically thick. However, since each of these radiating spots has a small area ($\delta A \ll 4\pi R_{\text{NS}}^2$) compared to the entire neutron star surface, the amount of energy radiated by both spots essentially comes from only that area

$$\epsilon_{\text{acc}} f \dot{M} c^2 = 2 \delta A \sigma T_*^4$$

or

$$T_* = 1.38 \text{ keV} \left(\frac{M_{\text{NS}}}{1.4 M_{\odot}} \right)^{-1/4} \dot{m}^{1/4} \left(\frac{\delta \Omega}{0.1} \right)^{-1/4} \left(\frac{R_{\text{NS}}}{10 \text{ km}} \right)^{-1/2} \left(\frac{f}{0.5} \right)^{1/4} \quad (16.4)$$

where $\delta \Omega \equiv \delta A / 4\pi R_{\text{NS}}^2$, and f is the fraction of $\dot{M}$ that falls onto the optically thick polar spots.

This is only part of the story, however. Surrounding the accretion column above each polar spot will be an optically thin, hot coronal plasma, cooled by inverse Compton scattering, at temperatures similar to those in hot, optically thin accretion flows. For lack of a better model, we will make the same assumption made in the ADAF and MDAF models in Chapter 12, namely that $T_e \approx 10^9 \text{ K} = 86 \text{ keV}$.

In the next subsection we shall argue that the horizontal branch (HB) of Z sources is essentially identical to the accretion flow in a binary X-ray pulsar, but with Z sources having a weaker magnetic field. There are two reasons for this identification:

1. First, some binary X-ray pulsars (particularly ones in Be systems), in addition to showing an HB at low accretion rate, also show a partial normal branch (NB) at higher luminosity, with timing properties similar to those of Z sources [579].
2. Second, even when a partial NB does not exist in a binary X-ray pulsar, such as Her X-1, the source still exhibits a horizontal branch [580] in which the hard

color (here 9.3–14 keV/4.7–9.3 keV, or “band3 / band2”) decreases only slightly as the counts in the 2–4.7 keV band increase.

16.1.3.2 Z Sources

Turning the First Corner

We are now in a position to propose the first of two important properties of our model for accreting, magnetized neutron stars. *We suggest here that, in all of these sources, the transition from the HB to the NB is caused by the effect discovered by Marina Romanova and her colleagues and depicted in Fig. 14.9:* when the accretion exceeds a critical rate (which depends on the magnetic field strength), the flow onto the magnetized neutron star changes from two simple optically thick polar spots to multiple spots peppered over the neutron star surface. As depicted in Fig. 14.9B, although occurring in spiral streamers, much of this flow becomes increasingly equatorial, and therefore disk-like.

Based on the Romanova *et al.* simulations the critical accretion rate for this transition is given by equation (14.13). However, this result still is rather uncertain, both because of the uncertain nature of MHD simulations at the present time and because of the uncertain magnetic field strengths in observed sources. Adopting $\dot{m} = 0.1$ as the transition accretion rate in Z sources and van der Klis’s estimate of 10^{10} G as a typical Z source magnetic field [189, 190], we arrive at a modified version of equation (14.13) for the transition from polar accretion to broken spiral equatorial streamers

$$\dot{m} \gtrsim 0.1 \left(\frac{b}{0.1} \frac{B}{10^{10} \text{ G}} \right)^2 \equiv \dot{m}_B \quad (16.5)$$

where b is a measure of our uncertainty in the magnetic field strength (here, about a factor of 10).

Qualitatively, after the transition, the spectrum of the Z source should behave as follows. As more spots appear near the neutron star equator, $\delta\Omega$ in equation (16.4) will increase much faster than $\dot{m}$, leading to a softening of the soft color of the Z source. This additional cool emission will dramatically increase the Compton cooling of the hot coronal material, causing the latter to be subsumed into the increasingly disk-like equatorial accretion flow. This loss of hot coronal material will cause a dramatic decrease in the hard color. The softening of both the soft and hard colors are indications that the source has entered the NB.

Turning the Second Corner

The transition from the NB to the flaring branch (FB) in Z sources, then, likely marks the completion of disk formation. Here we take this event to occur at $\dot{m} \sim 0.3$. On

the FB a true Keplerian disk now exists nearly all the way down to the neutron star surface, and the spectrum should be roughly like a multi-colored thin disk one from equation (12.11)

$$T_* = 0.61 \text{ keV} \left(\frac{M_{\text{NS}}}{1.4 M_\odot} \right)^{-1/4} \left(\frac{\dot{m}}{0.3} \right)^{1/4} \left(\frac{R}{2 R_{\text{NS}}} \right)^{-3/4} \quad (16.6)$$

However, the threshold for a radiation-pressure-dominated “inner” disk region to form is somewhere between $\dot{m}_{\text{I}} \sim 0.1$ (equation (12.34)) and ~ 0.4 (equation (12.60)); for discussion purposes we will take this boundary to be $\dot{m}_{\text{I}} \approx 0.2$. So for $\dot{m} \geq 0.3$ a substantial “inner” region should exist as the Z source turns the second corner. For $\alpha = 0.01$ and $m = 1.4$, the outer radius of this region should be

$$R_{\text{M-I}} \approx 134 \text{ km} \left(\frac{\dot{m}}{0.3} \right)^{16/21}$$

or $\sim 13 R_{\text{NS}}$. Since most of the accretion luminosity will be coming from this region, its properties will substantially affect the observed behavior of the Z source on the FB.

In the classical model of a disk “inner” region (equations (12.35)) the optical depth to electron scattering is large

$$\tau_{\text{es}} \approx 3000 \left(\frac{\dot{m}}{0.3} \right)^{-1} \left(\frac{R}{2 R_{\text{NS}}} \right)^{3/2}$$

but the optical depth to free–free absorption, even with multiple electron scatterings taken into account, is much smaller

$$\tau_* = 10 \left(\frac{\dot{m}}{0.3} \right)^{-2} \left(\frac{R}{2 R_{\text{NS}}} \right)^{93/32}$$

and can be less than unity near the neutron star surface. The entire $13 R_{\text{NS}}$ “inner” disk region is potentially thermally unstable (Section 12.2.2), which would lead to flickering and flaring. The fact that this is, indeed, the “flaring branch” lends credence to identifying the FB as a fully-developed, near-Eddington accretion flow.

Because of the low absorption optical depth, especially as $\dot{m} \rightarrow 1$, we can obtain a rough estimate of the radiation temperature when $\tau_* < 1$ by simply setting it equal to the central temperature of the disk in that region (equations (12.35))

$$T_{\text{c}} = 5.5 \text{ keV} \left(\frac{R}{2 R_{\text{NS}}} \right)^{-3/8}$$

(Note that this is independent of accretion rate.) This is nearly ten times hotter than if the disk were optically thick (T_* above), so it is not surprising that the flaring branch has a hot component that becomes more prominent as $\dot{m}$ increases and τ_* becomes optically thin.

16.1.3.3 Atoll Sources

Atoll sources have the weakest magnetic field strengths in the van der Klis classification ($\sim 10^8$ G) [189, 190]. From equation (16.5) we see that $\dot{m}_{cf} \approx 10^{-5}$ in such sources, so even very low, ADAF-like accretion rates are expected to disrupt the neutron star magnetic field and form an equatorial torus/disk in the magnetosphere. This should not be too surprising: at $\dot{m} \sim 10^{-3}$ disk magnetic fields (10^{7-8} G, equations (12.47)) are comparable to this weak neutron star surface field, and the former fall off with radius much more slowly ($\propto R^{-5/4}$) than a stellar dipole ($\propto R^{-3}$).

The weak-field atoll source is, therefore, often considered to be the neutron star equivalent to the black hole. The island state (IS), which begins rather hard in color, but softens as $\dot{m}$ increases is the analogy of the black hole low/hard state. For this state we again use the ADAF/MDAF approximation $T_e \approx 10^9$ K = 86 keV. One thing needs to be kept in mind, however: in this low accretion rate state ($\dot{m} \ll 1$), a black hole will have a very low efficiency (equations (12.57) and (12.88)), while an atoll neutron star source's efficiency will remain near 20% (Section 11.1.1). *The X-ray luminosity of an atoll source, therefore, will be proportional to $\dot{m}$, while a low-luminosity black hole's will be proportional to $\dot{m}^2$.*

The banana state, which behaves similarly to the flaring branch of the Z sources (in QPO, as well as color-color, properties) may be the analogy of the black hole high/soft state [189, 190]. For this state we use the Shakura–Sunyaev disk surface temperature again in equation (16.6). However, the similarity between the shapes of the Z and atoll color-color diagrams is qualitative, at best. That is, *the fact that an atoll color-color plot is L-shaped, and looks like the bottom of a Z diagram (see Fig. 3.3), does not mean that they necessarily represent the same accretion process.* The atoll banana state behavior occurs at far lower $\dot{m}$ than the Z source FB behavior; the former's weak field is easily disrupted by the accretion flow even at low $\dot{m}$. Furthermore, the IS in the atoll source is not exactly like the Z source NB state. And the transition from IS to lower banana (LB), while probably related to formation of an equatorial disk, may be more related more to a transition from ADAF to thin disk than a transition from equatorial accretion spots to a full disk.

Finally, the upturn at the high-luminosity end of the upper banana (UB) state is interesting. It is similar to that in the Z source FB, which we have suggested above as being due to the increasing importance of an “inner” radiation-pressure-dominated disk region. The same could be true in the atoll banana state as well, as van der Klis's estimates of $\dot{m}$ there also significantly exceed 0.1.

16.1.3.4 Jets from Neutron Stars in Binaries

Jet Production at High Accretion Rates: Jet Suppression

Chapters 14 and 15 showed that rotating stars with magnetic fields make jets, and stars with magnetized accretion disks can do the same. Accreting stars in binary systems (particularly Z and atoll sources) produce jets, and they have both intrinsic

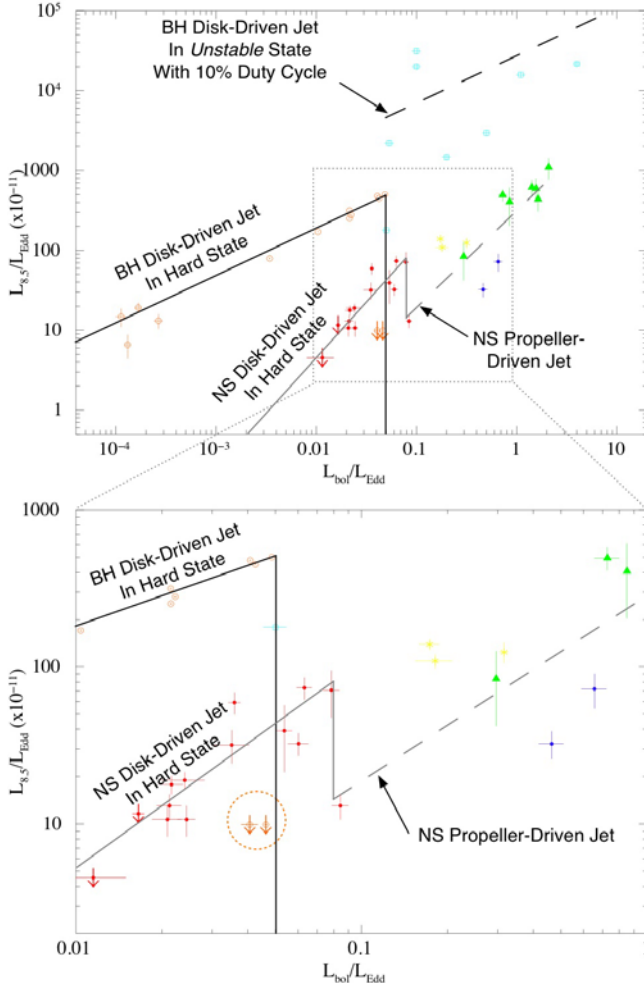


Fig. 16.2: Radio vs. bolometric luminosity (latter computed from the X-ray flux) in Eddington units for both binary black holes (open points) and neutron stars (filled points). *Top:* Entire accretion range, spanning six orders of magnitude in $\ell_{\text{rad}} \equiv L_{\text{bol}}/L_{\text{Edd}}$. For both species, transition from low/hard state at left to higher soft states at right is delineated by a vertical line (at $\ell_{\text{rad}} \approx 0.05$ for BHs and 0.08 for NSs) that shows the sudden suppression of the jet radio power. Open orange circles (including two upper limits at the transition) show data for GX 339-4 in the hard state; through these points is drawn a solid black line found by plotting eq. (16.24) for L_R against eq. (16.10) for L_{bol} . Open cyan boxes show data for several transient and outbursting binary BHs; through these is drawn the same curve (dashed line), assuming that all jet energy is generated in 10% duty-cycle outbursts. Filled red points show data for several atoll sources in the hard state (4U 1728-34, Aql X-1, 4U 1608-52 and 4U 0614-09); through these is drawn a solid gray line from eq. (16.7). Remaining points show atoll sources in a soft (blue) or outbursting (yellow) state and high-luminosity Z sources (green); through these is fit (gray dashed line) a propeller-driven jet from eq. (14.14) ($\propto L_J^{1.4} \propto \dot{m}^{1.2}$). *Bottom:* A zoom into the jet suppression region. Black hole jets are heavily suppressed in the higher accretion states, while neutron star jets are not. Our model explains this difference as suppression of the disk jet only; the pulsar propeller jet remains strong up to the Eddington accretion rate and beyond. Reproduced from Fig. 4 of [581], by permission of J. Wiley and Sons, Ltd.

magnetic fields and accretion disks. So, which mechanism is responsible for those jets – the pulsar propeller mechanism (Section 14.2) or the MHD disk wind process (Sections 14.5.3 and 15.1)? This turns out not to be an easy question to answer.²

At first the answer would seem fairly straightforward. Simone Migliari (now at the European Space Astronomy Centre (ESAC) in Madrid) and Rob Fender performed a detailed study of binary neutron stars with jets and found that the jet radio power is approximately proportional to the accretion rate $\dot{m}$, both for atoll and Z sources (Fig. 16.2) [581]. This is a classic signature of MHD disk winds. To see this, we can estimate the jet luminosity from the disk central magnetic field using the Blandford–Payne mechanism. Equations (15.17) show that most disk winds should have a dependence of $\mathcal{P}^{\text{EM,BP}} \propto \dot{m}^p$, with the exponent lying in the range $1.0 < p < 1.2$ (except for the unstable, near-Eddington radiation-pressure-dominated region, which will have $p = 2.0$).

The pulsar propeller mechanism, however, has about as strong an $\dot{m}$ dependence of $\mathcal{P}^{\text{EM,prop}} \propto \dot{m}^{0.86}$ (equation (14.14)). This is caused, not by a disk magnetic field strength increasing as $\dot{m}$ increases, but rather by the magnetic radius R_m decreasing with increasing $\dot{m}$ (equation (14.12)). This allows an increasingly strong part of the pulsar dipole field to be exposed to the accretion flow, thereby increasing the propeller wind/jet luminosity.

Furthermore, as we will discuss in more detail in the next section, jets from black holes are suppressed (by at least a factor of 50) as $\dot{m}$ approaches and exceeds 0.05 (see open circles in Fig. 16.2). Very little radio jet emission is seen above this accretion rate. This is widely believed to be an accretion-disk-related phenomenon, and we discuss a possible theoretical explanation below. In the neutron star case, jet suppression does occur (at $\dot{m} \approx 0.08$; again, see filled circles in Fig. 16.2), but it is much less dramatic (of order a factor of 5 or less in atoll sources [581]), and then *neutron star jets regain their growth in power with $\dot{m}$, up to the Eddington accretion rate and beyond* (Fig. 16.2).

Why might black hole jets be suppressed substantially but not neutron star jets? The answer may lie in the primary differences between neutron stars and black holes: unlike the latter, the former have a hard surface (which may or may not affect jet production) and an intrinsic dipole magnetic field (which almost certainly will affect the way jets are produced). From this premise we can propose the following tentative resolution of the problem:

- We suggest that neutron star jets in binaries are produced by *both* mechanisms, magnetized accretion disk and rotating magnetized pulsar. The pulsar jet would not be affected by disk magnetic field suppression, *if* the neutron star magnetic field near the stellar surface is strong enough to not be disrupted completely by the accretion flow. This allows the propeller jet to work even up to $\dot{m} = 1$ and beyond. The disk-driven jet, however, will be suppressed significantly as $\dot{m} \rightarrow 0.1$, just as it is in black holes. A suppression of atoll jets by only a factor of 5 then would imply that, in the hard state, the disk-driven jet contributes 80%

² This question is equivalent to asking whether the Blandford–Znajek or Blandford–Payne mechanism is responsible for black hole jets.

or less of the radio emission, while the rotating magnetized neutron star itself contributes at least 20%. After suppression ($\dot{m} > 0.1$) the pulsar propeller jet then produces at least 90% of the jet luminosity.

- Black hole jets also may be a combination of black hole spin-driven (Blandford–Znajek) and disk-driven (Blandford–Payne) jets. However, the source of the magnetic field in both cases is the accretion disk flow. Therefore, when the disk magnetic field is suppressed, *both* BP and BZ jet also will be suppressed. That is, the lack of an intrinsic magnetic field in the black hole means that black hole jets will remain suppressed.

Jets from Z sources in the $0.1 < \dot{m} < 1$ range are only a factor of 3–4 more powerful than atoll sources. This also is consistent with the mechanism being the pulsar propeller in that range. The dependence of the propeller jet luminosity on neutron star surface magnetic field strength is rather weak in equation (14.12) – $\mathcal{P}^{\text{EM,prop}} \propto B^{2/7}$ – implying that Z and atoll differ in magnetic field strength by a factor of ~ 50 –130, very similar to the factor of 100 suggested originally by van der Klis [189, 190].

The fact that for $\dot{m} > 0.1$ atoll sources’ jet luminosities continue to increase as $\dot{m} \rightarrow 1$ (two blue crosses near the lower right-hand part of Fig. 16.2) indicates that, while the atoll pulsar magnetic field is easily disrupted, the surface field must be strong enough to drive a jet in the presence of this high accretion rate.

Jet Radio Power at Low Accretion Rates

One quantity that we have not been able to compute theoretically in this book is the actual *radio* emission produced by a jet. While we can compute the *total* jet luminosity (kinetic plus magnetic), the conversion of that luminosity into radio emission is a rather complex process. However, several authors’ discussion of the conversion of jet luminosity to radio synchrotron power [582, 583, 584, 585] leads one to adopt the scaling relation $L_{R,\text{acc}}^{\text{NS}} \propto L_{\text{Edd}} L_J^{1.4}$. The greater than unity power law variation with L_J , and the proportionality constant, embody the details of the efficiency of conversion to radio luminosity. Therefore, following Migliari and Fender’s arguments and fitting the proportionality constant to their data at the point just before jet suppression begins, we obtain

$$L_{R,\text{acc}}^{\text{NS}} = 7 \times 10^{-10} L_{\text{Edd}} \left(\frac{L_J}{0.08 L_{\text{Edd}}} \right)^{1.4} = 1.22 \times 10^{29} \text{ erg s}^{-1} \left(\frac{\dot{m}}{0.08} \right)^{1.4} \quad (16.7)$$

For example, for the problem discussed at the beginning of this section (a bare neutron star traveling through the interstellar medium), with $\dot{m} \approx 10^{-10}$ – 10^{-8} , such an object would have to pass within 0.007–0.037 pc in order to be detected as a 1 mJy source at earth. This is well within the distance of the Oort cloud – far too close for any random neutron star to approach. Therefore, detection of accreting dead neutron stars in the ISM by radio means is out of the question.

The above equation (16.7) is valid only for jets produced by neutron stars; black hole binaries produce much stronger jets in their low/hard state. As Migliari and Fender state, neutron stars remain “stubbornly” less radio-loud than black holes for a given X-ray luminosity (by about an order of magnitude). Jet luminosity and suppression will be discussed more in the stellar-mass black hole section below.

16.2 Stellar-mass ($10\text{--}100\ M_{\odot}$) Black Holes: Rosetta Stones for Deciphering the Ultimate Engines

For the remainder of this chapter we will deal exclusively with true black holes, which we divide into mass classes that differentiate not only their size but also when, where, and how these objects are formed and how they obtain the accreting plasma that powers them as engines. We begin with black holes with masses typical of those of stars (in the range of $10\text{--}100\ M_{\odot}$). The physical size of a $10\ M_{\odot}$ black hole (30 km in radius) is similar to that of a large city, while a $100\ M_{\odot}$ black hole is the size of a typical country (or province/state in a large country).

16.2.1 *Stellar-mass Black Hole Formation*

Massive stars that die in a core-collapse supernova explosion leave only one of two types of compact remnant: a neutron star or a stellar-mass black hole (see Fig. 16.1). The formation of neutron stars and of stellar-mass black holes, therefore, are closely linked, both physically and astrophysically. The process of stellar-mass black hole formation is actually rather simple: for some reason the collapse of a neutronized iron core cannot be halted, even by neutron degeneracy pressure (Section 9.3.1) and the central region of the PNS is crushed to a density so high that an horizon forms and grows outward (Section 8.3.3). The center of Fig. 16.1 shows two possible evolutionary pathways to a black hole:

- A failed supernova, in which a PNS forms *and* an MHD jet-assisted supernova ensues, but the mantle is too massive for the supernova power to fully expel it.
- A direct collapse of a very large (several $M_{\odot}$) iron core.

The main questions that have yet to be answered in stellar-mass black hole formation are

1. In what mass range are black holes, instead of neutron stars, made by core collapse, and how does this change with the initial stars’ metallicity? In general, we expect black holes to come from stars more massive than those that produce neutron stars.
2. What are the physical reasons that cause some stars to crush their central PNS cores to black holes and some to not do so?

For this subject we will not speculate on, nor present, any new possible mechanisms. Instead, we simply shall review the issues presented in the previous chapters.

16.2.1.1 Theoretical Arguments

In Section 10.1.2 we discussed the extensive set of supernova explosions computed by Woosley, Heger, and Weaver, in which they artificially inject an extra 10^{51} erg of spherical explosion energy somewhat outside the surface of the PNS. The type of remnant left by such an explosion was strongly dependent primarily upon the initial mass of the exploding star. For stars that start with approximately solar abundances, neutron stars begin to be made inside stars that begin above $\sim 9 M_{\odot}$ and black holes are made in stars that above $\sim 21 M_{\odot}$. Woosley and his team also ran models with zero metal abundance. They saw little difference between those two extremes in this moderate range; so $9\text{--}21 M_{\odot}$ for neutron star formation, and $>21 M_{\odot}$ for black hole formation, may be a fairly robust (theoretical) result. However, we note here that models were not run for abundances that were twice solar or greater. The latter is more typical of the stars being formed at present, whereas solar abundances reflect the Galactic conditions about 5 billion years ago, when our Galaxy was only half its current age. We will discuss this in more detail below.

For stars with initial masses above $\sim 50 M_{\odot}$, however, Woosley *et al.*'s results were strongly dependent on the stars' initial metallicity. Because all massive stars radiate near the Eddington luminosity limit, their atmospheres already are hovering near the point where they will produce radiation-pressure-driven winds. (This physics is similar to that discussed in Section 13.1 for winds coming from *accreting* stars.) So, the addition of even a small amount of metals in those stars' atmospheres will increase the radiative opacity by many times the simple electron scattering value, resulting in a large "force multiplier". High-metallicity massive stars, therefore, will have a very strong radiation-pressure-driven wind that could reduce a star's mass significantly before it has a chance to explode, or even reach its final stages of stellar evolution. For example, for solar abundances, a $60 M_{\odot}$ star may be reduced to only an $8 M_{\odot}$ or less Wolf-Rayet star, losing all of its hydrogen, and perhaps even its helium and carbon mantles. Woosley *et al.* discuss the possibility of a "neutron star" window, in the initial mass range $50\text{--}80 M_{\odot}$ where neutron stars are formed, while black holes may be formed in the ranges $21\text{--}50 M_{\odot}$ and $>80 M_{\odot}$.

16.2.1.2 Observational Evidence

Nucleosynthesis arguments (Section 10.1) and supernova rates (Section 3.3) are good methods of estimating the neutron star-formation rate. However, they are not much help in estimating the black hole formation rate. This is because predicted metal abundances and supernova rates are strong functions of the *lower* initial stellar mass limit for neutron production but weak functions of that for black hole production.

As discussed in Section 3.5.1, a reasonable way to estimate the stellar mass above which black holes form (instead of that for neutron stars) is to measure the relative numbers of neutron stars and black holes and then work backward to the NS/BH stellar mass boundary. This works as follows. Using the Scalo initial mass function (Section 10.4.1), the number of massive stars produced in a given initial mass range $m_{\star\min} < m < m_{\star\max}$ is proportional to

$$N(m_{\star\min} < m < m_{\star\max}) \propto (m_{\star\min}^{-1.3} - m_{\star\max}^{-1.3})$$

So the ratio of black holes to neutron stars is then

$$R_{\text{NS:BH}} = \left(\frac{m_{\star\min,\text{NS}}^{-1.3} - m_{\star\max,\text{NS}}^{-1.3}}{m_{\star\min,\text{BH}}^{-1.3} - m_{\star\max,\text{BH}}^{-1.3}} \right)$$

Taking $m_{\star\min,\text{NS}} = 9 \text{ M}_{\odot}$, $m_{\star\max,\text{NS}} = m_{\star\min,\text{BH}}$, and $m_{\star\max,\text{BH}} \rightarrow \infty$, we can solve for the minimum initial stellar mass that forms black holes

$$m_{\star\min,\text{NS}} \approx 9 \text{ M}_{\odot} (1 + R_{\text{NS:BH}})^{1/1.3}$$

So, what is the observed ratio of neutron stars to black holes currently being produced in our Galaxy? Since HMXBs are young systems currently being produced, the HMXB numbers in Table 3.2 gives the best estimate of current ratios

$$R_{\text{NS:BH}} = 15\text{--}20$$

with the lower number including persistent sources only. This gives a value of

$$m_{\star\min,\text{BH}} \approx 75\text{--}95 \text{ M}_{\odot} \quad (16.8)$$

for the initial stellar mass above which black holes are currently forming in their interiors.

The LMXB numbers in Table 3.2, however, give a long-term (and less certain) average over the history of the Galaxy of $R_{\text{NS:BH}} = 3$. That is, black hole formation was much more common in the distant past. This gives $m_{\star\min,\text{BH}} \approx 25 \text{ M}_{\odot}$, which is more consistent with the $21 \text{ M}_{\odot}$ value we have been quoting.

Why might the two estimates be so different? One reason could be bias in the neutron star and black hole HMXB death rates. The estimate of *current* compact object formation rates assumes that these HMXB death rates are equal. However, if neutron star HMXB systems live longer than black hole ones, then the ratio of neutron stars to black holes would be biased high. However, this probably is not the case. The lifetime of an HMXB is equal to the lifetime of the O/B companion star, which is a strong function of the mass of that O/B star. It turns out that the spectral, and therefore mass, range of companion stars in HMXB systems is remarkably small, O7 to B0 or $20\text{--}25 \text{ M}_{\odot}$. So, all HMXB systems have lifetimes of $\sim 10\text{--}15 \text{ Myr}$, the ratio of observed numbers *does* reflect their current formation rates, and the conclusion that $m_{\star\min,\text{BH}} \approx 75\text{--}95 \text{ M}_{\odot}$ remains credible.

If the values for $m_{\star\min,\text{BH}}$ derived from HMXB and from LMXB numbers reflect real differences between the current and average black hole formation rates, then a possible astrophysical explanation would be that black hole formation was much more common in the distant past than currently (by a factor of 5–7). There now may be a significantly wider “neutron star gap” than that estimated by Woosley’s team, because the current Galactic metallicity is at least a factor of two greater now than solar (which reflects conditions when the sun was formed at about half the Galaxy’s present age). O/B star mass-loss rates may be so high now that stars with initial masses $\lesssim 75 M_{\odot}$ or so will form only neutron stars. So, it is possible that, while supernovae will continue to explode at a relatively steady rate over the life of the Galaxy, the fraction of those stellar deaths that form black holes will continue to decrease as the Galaxy further ages. Having dropped from one out of three in the past to one out of twenty today, black hole formation in our Galaxy may disappear altogether in the distant future.

16.2.2 Black Hole Astrophysics in a Nutshell

The physics that governs how accreting black holes behave is somewhat different from that of neutron stars. Therefore, we need to spend a few moments outlining the most important equations that relate accretion theory to black hole observations. As we do so, it might be helpful to remember the main differences between these two kinds of objects:

1. Matter, heat, light, and magnetic energy all can be lost into the black hole horizon, while neutron stars have a hard surface.
2. Black holes cannot maintain a strong intrinsic magnetic field; any magnetosphere or jet generated near a black hole comes only from the matter accreting toward the hole *outside* the horizon.

16.2.2.1 Accretion Power: Black Hole Luminosity in the Ultraviolet to X-Ray Band

Low Accretion Rates ($\dot{m} < \dot{m}_A$)

For both the ADAF model, and the MDAF augmentation to it, the efficiency of accretion can be very low (equations (12.57) and (12.88))

$$\epsilon_{\text{acc,adf/mdf}} = C \left(\frac{\dot{m}}{\alpha^2} \right) \epsilon_{\text{acc}}$$

where α is the Shakura–Sunyaev viscosity parameter (Section 12.2.1), C is a constant of order unity, and ϵ_{acc} is usually taken to be about 0.1. (Recall that $0.053 < \epsilon_{\text{acc}} < 0.42$ for black holes of all rates of spin $-1 < j < 1$; Section

7.5.) This leads to a bolometric luminosity of (equations (12.58) and (12.87))

$$L_{\text{acc,adf/mdf}} \approx C L_{\text{Edd}} \left(\frac{\dot{m}}{\alpha^2} \right) \dot{m} \quad (16.9)$$

Recalling that $\dot{m}$ includes ϵ_{acc} (Section 11.1), we can calibrate the above equation with Migliari and Fender’s observations in Fig. 16.2 to give $L_{\text{acc,adf/mdf}} \approx C L_{\text{Edd}} \dot{m}^2 / \dot{m}_A$ for $\dot{m} < \dot{m}_A \approx 0.05$ (which corresponds to $C \approx 1.8$ for $\alpha = 0.3$), or³

$$L_{\text{acc,adf/mdf}} \approx 2.5 \times 10^{39} \text{ erg s}^{-1} m \dot{m}^2 \quad (\dot{m} < \dot{m}_A) \quad (16.10)$$

with most of the luminosity coming out in the hard X-ray band

$$T_{*,\text{adf/mdf}} \sim 100 \text{ keV} \quad (\dot{m} < \dot{m}_A) \quad (16.11)$$

– independent of black hole mass. So, ADAF/MDAF theory is consistent with the hard state depicted in the Fig. 3.12 of Cygnus X-1.

Intermediate-to-High Sub-Eddington Accretion Rates ($\dot{m}_A < \dot{m} < 1$)

For sub-Eddington accretion rates above the critical value of $\dot{m}_A \approx 0.05$ the inflow switches from a hot advection-dominated one to a cooler, thin accretion disk with a standard efficiency ϵ_{acc} and a bolometric luminosity of $L_{\text{acc,thin}} \approx C L_{\text{Edd}} \dot{m}$ (equations (12.30) and (12.35)) or

$$L_{\text{acc,thin}} \approx 1.25 \times 10^{38} \text{ erg s}^{-1} m \dot{m} \quad (\dot{m}_A < \dot{m} < 1) \quad (16.12)$$

Equation (12.13) shows that the spectrum from such a “multi-colored black body” rises with frequency as $\nu^{1/3}$ and has a spectral temperature of about (equation (12.11))

$$T_{*,\text{thin}} \approx 1.3 \text{ keV } m^{-1/4} \dot{m}^{1/4} \quad (\dot{m}_A < \dot{m} < 1) \quad (16.13)$$

At high accretion rates, therefore, stellar-mass black holes appear as X-ray (0.5–1.0 keV) sources, while supermassive ones should appear as ultraviolet sources ($\sim 10^5$ K). Standard thin disk accretion theory, therefore, is compatible with the soft state of Cygnus X-1 in Fig. 3.12, and with the “big blue bump” in quasars (Section 2.2) being due to cool thermal accretion disk emission.

³ The viscosity parameter α (a measure of how rapidly heat is generated in the accretion flow) can be different in different types of inflow. Typically we take $\alpha \approx 0.01$ – 0.03 in disk-like flows and $\alpha \approx 0.3$ in ADAF-like coronal flows.

Special Consideration for High Sub-Eddington Accretion Rates ($\dot{m}_I < \dot{m} < 1$)

When a Shakura–Sunyaev “inner” region exists ($\dot{m} > \dot{m}_I \approx 0.2 m^{-1/8}$), the accretion flow is secularly (possibly cyclically [238]) unstable on a secular (Lightman–Eardley) time scale (equation (12.40))

$$\tau_{\text{acc}} = \frac{R_{M-I}}{V_R} = 12 \text{ min} \left(\frac{M_\bullet}{10 M_\odot} \right)^{4/3} \left(\frac{\dot{m}}{0.3} \right)^{2/3} \quad (16.14)$$

(for $\alpha = 0.03$). Table 16.1 shows example secular time scale values for black holes of different masses and different accretion rates (the unstable state currently being discussed and the two super-Eddington states discussed below). (Note that the ADAF, MDAF, and stable soft states do not have an unstable Shakura–Sunyaev “inner” region, so they are not included in the table.) The secular time grows faster than linearly with black hole mass $M_\bullet$. And it grows as $\dot{m}^{2/3}$ with normalized accretion rate because the radius of the unstable “inner” region itself grows with accretion rate, as well as black hole mass. Note that it is quite possible that the instantaneous peak accretion rate in these cycles may reach, or even exceed, the Eddington value. Therefore, the instantaneous jet luminosity also may equal, or exceed, the Eddington luminosity for brief periods of time.

In addition, because the “inner” region is thermally unstable as well, we also quote the range of thermal time scales, from R_{M-I} (equation (12.41))

$$\tau_{\text{th}}^{\text{max}} = 1.0 \text{ s} \left(\frac{M_\bullet}{10 M_\odot} \right)^{8/7} \left(\frac{\dot{m}}{0.3} \right)^{8/7} \quad (16.15)$$

down to R_{isco} (equation (12.42))

$$\tau_{\text{th}}^{\text{min}} = 0.049 \text{ s} \left(\frac{M_\bullet}{10 M_\odot} \right) \quad (0.2 m^{-1/8} < \dot{m} < 1) \quad (16.16)$$

Table 16.2 shows example thermal time scale ranges for the same range of black hole mass and accretion rates as in Table 16.1. The upper limit on the range is given by equation (16.15) and the lower range by equation (16.16) or (16.20), depending on the accretion rate. Tables like these will be very useful for estimating variability time scales for black holes of different masses and accretion rates. Note that, in this case, the free-fall time is about 1/60th of the minimum thermal time: $\tau_{\text{ff}} = 0.55 \alpha \tau_{\text{th}} \approx 0.017 \tau_{\text{th}}$.

Super-Eddington Accretion Rates ($1 < \dot{m}$)

Whether the “slim” disk model or the super-Eddington wind model is the correct description of an accretion flow’s observable properties when $\dot{m} \gg 1$, they both

Table 16.1: Secular time scales for high accretion rate states vs. black hole mass

Accretion state	$\dot{m}$	Black hole mass (M _⊙)									
		1	10	100	10 ³	10 ⁴	10 ⁵	10 ⁶	10 ⁷	10 ⁸	10 ¹⁰
Unstable	0.03							23 yr	500 yr	11 kyr	230 kyr
Unstable	0.1				2 d	1.3 mo	2.3 yr	50 yr	1.1 kyr	23 kyr	500 kyr
Unstable	0.3	30 s	12 m	4 hr	4 d	2.7 mo	5 yr	110 yr	2.3 kyr	50 kyr	1.1 Myr
Slim disk	3.0	2.6 m	0.9 hr	20 hr	18 d	1.1 yr	23 yr	500 yr	11 kyr	230 kyr	5 Myr
Super-Edd wind	30	12 m	4 hr	4 d	2.7 mo	5 yr	110 yr	2.3 kyr	50 kyr	1.1 Myr	23 Myr

Table 16.2: Thermal time scale ranges for high accretion rate states vs. black hole mass

Accretion state	$\dot{m}$	Black hole mass (M _⊙)									
		1	10	100	10 ³	10 ⁴	10 ⁵	10 ⁶	10 ⁷	10 ⁸	10 ¹⁰
Unstable	0.03							0.06–0.4 d	0.6–6 d	0.2–2.7 mo	1.5–40 yr
Unstable	0.1				0.08–1 m	0.8–13 m	0.14–3 hr	0.06–1.5 d	0.02–0.8 mo	0.015–0.8 yr	1.5–150 yr
Unstable	0.3	5–70 ms	0.05–1 s	0.5–14 s	0.08–3 m	0.8–45 m	0.14–10 hr	0.06–6 d	0.02–3 mo	0.015–3 yr	1.5–600 yr
Slim disk	3.0	0.025–1 s	0.25–14 s	0.4–3 m	0.4–45 m	0.07–10 hr	0.03–6 d	0.01–3 mo	0.008–3 yr	0.08–40 yr	0.8–600 yr
Super-Edd wind	30	0.8–14 s	0.14–3 m	1.4–45 m	0.2–10 hr	0.09–6 d	0.03–3 mo	0.026–3 yr	0.26–40 yr	2.6–600 yr	0.026–8 kyr

Table 16.3: Typical longitudinal magnetosound crossing times for a relativistic jet vs. black hole mass

Crossing time	Black hole mass (M _⊙)									
	1	10	100	10 ³	10 ⁴	10 ⁵	10 ⁶	10 ⁷	10 ⁸	10 ¹⁰
10 ³ τ _g	5 ms	0.05 s	0.5 s	0.08 m	0.8 m	0.14 hr	0.06 d	0.02 mo	0.016 yr	1.6 yr
10 ⁴ τ _g	50 ms	0.5 s	5 s	0.8 m	8 m	1.4 hr	0.6 d	0.2 mo	0.16 yr	16 yr
10 ⁵ τ _g	500 ms	5 s	50 s	8 m	80 m	14 hr	6 d	2 mo	1.6 yr	160 yr

predict the same approximate radiative luminosity $L_{\text{acc,super-Edd}} \approx L_{\text{Edd}}$ (equations (12.53) and (13.20)) or

$$L_{\text{acc,super-Edd}} \approx 1.25 \times 10^{38} \text{ erg s}^{-1} m \quad (1 < \dot{m}) \quad (16.17)$$

However, the “slim” disk and super-Eddington wind models predict very different observed spectral temperatures (equations (12.53) and (13.27)):

$$T_{*,\text{slm}} \approx 3.0 \text{ keV } m^{-1/4} \left(\frac{\dot{m}}{30} \right)^{1/4} \quad (16.18)$$

$$T_{*,\text{wnd}} \approx 0.67 \text{ keV } m^{-0.21} \left(\frac{\dot{m}}{30} \right)^{-0.89} \quad (16.19)$$

That is, the “slim” disk should appear as a slightly hot source that gets *hotter* as $\dot{m}$ increases. However, the super-Eddington wind should appear as a soft X-ray or UV source that gets *cooler* as $\dot{m}$ increases. It is possible that both types of super-Eddington accretion flows can be observed in nature: a visible “slim” disk for moderately super-Eddington accretion rates, and an invisible “slim” disk buried inside an optically thick wind for highly super-Eddington rates. We suggest candidate sources for both below. However, at this time, we have no good way of determining when this transition takes place. So, for illustrative purposes we will be assuming that the source displays a visible “slim” disk for $1 < \dot{m} \lesssim 10$ and above that it shows only a powerful wind.

Also, for $\dot{m} > 1$, the unstable Shakura–Sunyaev “inner” region does not extend all the way down to the ISCO – only to the radius where the accretion rate becomes super-Eddington locally ($r_{\text{tr}} \approx r_i \approx \dot{m}$). So in this case

$$\tau_{\text{th}}^{\text{min}} = 0.049 \text{ s } \left(\frac{M_{\bullet}}{10 M_{\odot}} \right) \dot{m}^{3/2} \quad (1 < \dot{m}) \quad (16.20)$$

16.2.2.2 Jet Power and Jet Suppression: Black Hole Luminosity in the Radio Band

In addition to producing thermal emission in the accretion flow, black holes also produce nonthermal emission in the magnetohydrodynamic jet outflows produced either by the rotating accretion flow or by the rotating black hole itself. We will first discuss the total jet luminosity produced and then later the implication for observations in the radio band.

Taking a cue from the above discussion, we will adopt $\dot{m}_A = 0.05$ as the critical black hole accretion rate where the inflow changes from an ADAF/MDAF optically thin, geometrically thick one to an optically thick, geometrically thin disk inflow.

Low Accretion Rates ($\dot{m} < 0.05$): ADAFs vs. MDAFs

Total Black Hole Jet Luminosity. In both ADAF and MDAF theory, the electron temperature saturates at $T_{e,\text{sat}} \sim 100$ keV (equation (16.11) above). In MDAF theory only, the ion temperature also saturates – at a value $T_{i,\text{sat}}$, which is related to $T_{e,\text{sat}}$ by the unknown ϑ -parameter

$$\vartheta \equiv \frac{T_{i,\text{sat}}}{T_{e,\text{sat}}} \quad (16.21)$$

For MDAFs, this parameter lies in the range $1 < \vartheta < 820$ and governs both the size of the MDAF region inside the ADAF (equation (12.64))

$$R_0 = 8.9 \times 10^5 \text{ cm } m \left(\frac{\vartheta}{820} \right)^{-1} = R_A \vartheta^{-1} \quad (16.22)$$

and the luminosity of the jet, which is produced primarily in the ADAF portion only (first of equations (15.17))

$$L_{J,\text{adf/mdf}} = 6.2 \times 10^{37} \text{ erg s}^{-1} \left(\frac{M_\bullet}{10 M_\odot} \right) \left(\frac{\dot{m}}{0.05} \right) \left(\frac{\vartheta}{820} \right) \quad (\dot{m} < 0.05) \quad (16.23)$$

Here R_A is the radius of the entire ADAF/MDAF region and R_0 is the radius of the MDAF region only. If $\vartheta \geq 820$, the ion temperature essentially does not saturate at all, and the entire inner portion of the accretion flow is an ADAF only (*i.e.*, no MDAF). On the other hand, if the energy transfer from ions to electron is so fast that $T_{i,\text{sat}} \approx T_{e,\text{sat}}$, then $\vartheta = 1$, and the entire inner portion of the accretion flow is an MDAF only (*i.e.*, no ADAF). Hybrid ADAF/MDAF accretion structures lie in the broad range between $\vartheta = 1$ and $\vartheta \sim 1000$, and we see that as the ADAF grows in dominance, the luminosity of the jet produced increases accordingly.

Black Hole Jet Power in the Radio Band. An empirical formula for the observed black hole jet radio power also can be obtained from Migliari and Fender’s paper [581]. Instead of that formula, however, we shall use a more general expression, derived from the “black hole fundamental plane”, discovered by Andrea Merloni (now at the Technical University of Munich), Sebastian Heinz (University of Wisconsin), and Tiziana Di Matteo (Carnegie Mellon University) when they all were at the Max Planck Institute for Astrophysics in Munich [582]. The fundamental plane formula relates *observed* radio and X-ray powers of many black hole sources to their derived black hole masses, from $7 M_\odot$ (GRO J1655-40) to $3 \times 10^9 M_\odot$ (M87):

$$\log L_R = 0.6 \log L_X + 0.78 \log m + 7.33$$

If we apply the expression $L_X = L_{\text{Edd}} \dot{m}^2 / \alpha^2$ from equation (16.9), and adjust the coefficient by a factor of 1.4 to fit Migliari and Fender’s data, we arrive at an

expression for the *observed* black hole radio power that should be valid over nearly nine orders of magnitude in black hole mass, for any source lying in the fundamental plane⁴

$$L_{R,\text{adf}} = 6.2 \times 10^{30} \text{ erg s}^{-1} \left(\frac{M_{\bullet}}{10 M_{\odot}} \right)^{1.38} \left(\frac{\dot{m}}{0.05} \right)^{1.2} (\dot{m} < 0.05) \quad (16.24)$$

It is important to remember that this empirical expression is valid only in the hard X-ray accretion state ($\dot{m} \lesssim 0.05$) and, as we shall see below, only valid when this hard state is dominated by an ADAF that extends essentially down to the black hole ISCO. It also is valid only when L_R and L_X are not relativistically beamed.

A Possible General Expression for Jet Radio Luminosity

It is clear from equation (16.24) that the actual radio power varies with mass and accretion rate more steeply than the total jet luminosity does. This is well known from works of several groups on the fundamental plane. One can, therefore, develop an empirical formula for converting *any* jet luminosity L_J into a radio power:

$$L_R(M, L_J) = 2.74 \times 10^{-15} \text{ erg s}^{-1} \left(\frac{M_{\bullet}}{10 M_{\odot}} \right)^{0.18} \left(\frac{L_J}{\text{erg s}^{-1}} \right)^{1.2} \quad (16.25)$$

This expression works for hard-state sources in the fundamental plane over nine orders of magnitude, and we will use it to obtain an approximate jet GHz radio luminosity from the total jet luminosity in the other cases below. This is only a rough estimate of the radio power in all accretion cases, not very accurate (and possibly not even correct) for the higher accretion rate sources. It is noteworthy, however, that even with the high power of the mass $M_{\bullet}$, equation (16.25) produces a radio luminosity that still is well below the actual jet total luminosity; *e.g.*, for $M_{\bullet} = 10^{10} M_{\odot}$ and $L_J = 10 L_{\text{Edd}}$, the ratio $L_R/L_{\text{Edd}} < 0.008$.

Intermediate Accretion Rates ($0.05 < \dot{m} \lesssim 0.2 m^{-1/8}$)

By *intermediate* accretion rates, we mean those that are high enough to produce a geometrically thin, optically thick Shakura–Sunyaev type accretion disk ($\dot{m} > \dot{m}_A$), but low enough *not* to produce a secularly and, possibly thermally, unstable radiation-pressure-dominated “inner” region ($\dot{m} < \dot{m}_I$). For hot, black hole sources this means that the Shakura–Sunyaev “*middle*” region is dominant and that it, and the jet launching process, extend all the way down to the black hole ISCO.

⁴ As mentioned earlier, for a neutron star ($m = 1.4$, $\dot{m} = 0.08$) this expression predicts about a factor of 6 higher radio luminosity than observed (equation (16.7)). Migliari and Fender pointed out the fact that, compared to black holes, neutron stars seem to be underluminous in the radio, even when their lower mass is taken into account.

As we did in the discussion above (near equation (16.6)) we will adopt $\dot{m}_I = 0.2 m^{-1/8}$ as the accretion rate above which such an “inner” radiation-pressure-dominated region forms. For a $10 M_\odot$ black hole, the total jet luminosity for this accretion state is given by the second of equations (15.17)

$$L_{J,\text{interm}} = 1.08 \times 10^{36} \text{ erg s}^{-1} \left(\frac{M_\bullet}{10 M_\odot} \right)^{0.9} \left(\frac{\dot{m}}{0.05} \right)^{1.2} \quad (0.05 < \dot{m} < 0.2 m^{-1/8}) \quad (16.26)$$

which, using equation (16.25), converts to a GHz jet radio luminosity of

$$L_{R,\text{interm}} = 4.8 \times 10^{28} \text{ erg s}^{-1} \left(\frac{M_\bullet}{10 M_\odot} \right)^{1.26} \left(\frac{\dot{m}}{0.05} \right)^{1.44} \quad (0.05 < \dot{m} < 0.2 m^{-1/8}) \quad (16.27)$$

Note that, when the black hole mass exceeds $\sim 10^5 M_\odot$, we have $\dot{m}_I < \dot{m}_A$. So, for massive and supermassive black holes the intermediate accretion state may no longer exist, and the source may transition directly from a hot ADAF/MDAF state to the unstable radiation-pressure-dominated state as $\dot{m}$ increases above $\dot{m}_A$.

Suppression of the Strong, Steady Hard State Jet

It is now clear that there are at least two decidedly different ways in which the steady jet luminosity of an ADAF in the low, hard state can be significantly suppressed: the accretion state can transition to an MDAF ($\vartheta \ll 820$ in equation (16.23)), which causes the jet to be weaker because it is launched further from the black hole than in a standard ADAF; or the source can transition to an optically thick, geometrically thin and cool disk (equation (16.26)), which has a much lower jet power because the vertical magnetic field is weaker.

Jet Suppression by an ADAF $\rightarrow$ MDAF Transition. If the low accretion state acquires a significant MDAF structure inside its ADAF interior ($\vartheta = 820 \rightarrow \vartheta \ll 820$; i.e., $T_i \approx 10^{12} \text{ K} \rightarrow \ll 10^{12} \text{ K}$), the jet then is launched from the radius R_0 , which is a factor of $820/\vartheta$ larger than the last stable orbit. Because the escape energy there is weaker, the jet luminosity is correspondingly weaker by a factor of $\vartheta/820$ (equation (16.23)). For example, even for a modest $\vartheta \approx 82$ (ion temperature saturating at $\sim 10^{11} \text{ K}$ and the jet launching at a radius of $60 GM/c^2$), the jet total luminosity would drop by a factor of 10, and the jet radio power (proportional to $L_J^{1.2-1.4}$) would drop by a factor of 15–25.

Jet Suppression by an ADAF $\rightarrow$ Thin Disk Transition. If, instead, the low accretion state is dominated by an ADAF structure ($\vartheta \rightarrow 820$), there will be a more than

two orders of magnitude (130–300) drop in the jet *radio* power when the source transitions from that low state (equation (16.23)) to the thin disk state (equation (16.26)). This large difference is due mainly to the significant change in disk thickness between the hot and cool accretion states, leading to a significant change in the strength of the vertical disk magnetic field B_{Z0} and a resulting dramatic difference in the inferred jet luminosity (equations (15.14) to (15.16)), which is proportional to B_{Z0}^2 .

In the sub-subsections below we discuss possible observational examples of both types of jet suppression. It is important to realize that possibly *neither* of these mechanisms of jet suppression, nor the sources where we believe these mechanisms may be playing a role, may have much to do with the reason that most quasars appear to be “radio-quiet”.

High Sub-Eddington Accretion Rates ($0.2 m^{-1/8} < \dot{m} < 1$)

For high, sub-Eddington accretion rates, an unstable “inner” region exists. Therefore, we might be tempted to take the *average* jet luminosity in this case to be given by the third of equations (15.17)

$$L_{J,\text{high}} = 6.5 \times 10^{37} \text{ erg s}^{-1} \left(\frac{M_{\bullet}}{10 M_{\odot}} \right) \left(\frac{\dot{m}}{0.3} \right)^2 \quad (0.2 m^{-1/8} < \dot{m} < 1) \quad (16.28)$$

which approaches the Eddington limit (in jet luminosity) as $\dot{m} \rightarrow 1$. However, *this expression is highly suspect*, as it is uncertain how the long-term behavior of the secular and thermal instabilities might affect the average of the jet power.

A more reasonable approach is to assume that the thermal instabilities lead to a two-phase medium, with the dense, cool material settling mostly along the equator and the tenuous, hot material settling mostly in a corona above the disk. (Of course, such a structure still will not be in a steady state, evolving instead on the Lightman–Eardley secular time.) In this case, we might be more correct in treating the system as one involving *two* separate and *not* coincident accretion flows: one of which is similar to a Shakura–Sunyaev “middle” region and the other similar to an ADAF, depending where in the secular cycle the source currently happens to be. In that case, then, it is, in fact, more appropriate to use either equation (16.27) (for the cool phase equatorial material) or equation (16.24) (for the hot phase corona), rather than try to base a radio jet power on equation (16.28).

Super-Eddington Accretion Rates ($1 < \dot{m}$)

Jet Suppression by Super-Eddington Accretion. In situations where the accretion rate is super-Eddington, we shall use the “slim” disk model for the central accretion engine. In that case a jet should be ejected from the inner portion of the “slim” disk

near the black hole. The luminosity of that jet would be relatively steady (as the “slim” disk model is advective and stable), and it would, indeed, be super-Eddington

$$L_{J,\text{super-Edd}} \approx 1.25 \times 10^{39} \text{ erg s}^{-1} \left(\frac{M_{\bullet}}{10 M_{\odot}} \right) \dot{m} \quad (1 < \dot{m}) \quad (16.29)$$

Normally, we would use equation (16.25) to convert this expression to a radio power. However, as discussed in Section 15.1.1 and in the sections below, sources in which we strongly suspect super-Eddington accretion are rather radio-quiet. At present we do not have a good model for jet suppression in a super-Eddington state. However, as a tentative working hypothesis for jet suppression we shall adopt the compelling scenario in which the jet plasma particles remain thermal and never are accelerated to nonthermal energies, because of Compton cooling by the strong radiation field (Section 15.1.1 and [528]). This concept preserves the close relationship between ADAFs and “slim” disks (geometrically thick inflows that have equivalent magnetic and dynamical pressures plus strong jet-producing properties). Yet it explains why the super-Eddington inflows might produce bipolar outflows that do not emit strongly in the radio. It also places strange sources like SS433 within the general model discussed below, and it sets up a nice picture for BAL AGN as having bipolar outflows that form cool clouds. Indeed, work on the cool SS433 jet clouds [586, 587] may be directly applicable to the formation of BAL clouds in a thermal bipolar outflow.

For lack of a better model, at this point we will take the radio power in this case to be simply

$$L_{R,\text{super-Edd}} \approx 0 \text{ erg s}^{-1} \quad (1 < \dot{m}) \quad (16.30)$$

Clearly this is a gross underestimate (ignoring effects like thermal synchrotron), but it will suffice for now.

16.2.2.3 Are Jets Launched Primarily by Rotating Black Holes or by their Rotating Accretion Disks?

The reader may notice that, in this section, we will be able to discuss virtually all aspects of jet production in stellar black hole engines by using the Blandford–Payne equations only. That is, we consider jets driven only by magnetic fields in the rotating accretion flow. Does black hole rotation play any part in jet production in binary X-ray sources at all?

The lack of need for the Blandford–Znajek effect also was noted by Fender *et al.* [588], who looked at the issue in some detail. They compared properties of X-ray binary radio jets with the measured spins of black holes in those systems and concluded that there is no evidence that black hole spins play any role in the production of jets.

In this author’s opinion, however, such a conclusion is very premature. While it is a formally correct result from this careful study, the lack of evidence may have nothing to do with which mechanism (or mechanisms) contribute to jet production in stellar black hole engines. That is, “Absence of evidence is not evidence of absence” [589]. There are several reasons:

1. First, except for sources like GX 339-4, XTE J1550-564, and GRO J1655-40, which have visibly broad X-ray iron lines and therefore $j \rightarrow +1$, it is not clear that *any* accurate spin measurements have been made in X-ray binary systems. Indeed, measurements of the spin in GRS 1915+105, for example, lie in the range $0 < j < 1$. The scatter in all measurements from different groups and methods is much greater than the quoted error bars on each method.
2. Secondly, in both the BP and BZ processes, the *primary* physical quantity that determines the power of the jet is the vertical component (along the rotation axis) of the magnetic field in the disk ($L_J \propto B_z^2$). The black hole has no magnetic field of its own. Rotation is, of course, important, but which object produces that rotation (disk or hole) is secondary – and potentially difficult to distinguish.
3. Thirdly, it is important to realize that *any observed variations in jet properties must be due to variations in accretion flow properties (like magnetic field strength) – regardless of which object produces the rotation*. Black hole spin in a binary system can be affected only on long time scales ($\tau_{\text{spin}} \sim M_\bullet / \langle \dot{M} \rangle \gg \tau_{\text{salp}} \sim 4 \times 10^7 \text{ yr} (\epsilon_{\text{acc}}/0.1)$). Binary black hole spins, therefore, have not changed much since before the death of the dinosaurs, let alone over the lifetime of a single human observer.
4. Fourthly, one might have predicted from the outset that the Blandford–Znajek process may be weaker than the Blandford–Payne power in X-ray binaries. Whatever spin the black hole has now, it is likely to be partly aligned with the angular momentum of the binary itself and, therefore, with the angular momentum of the accreting material as well. (All binary material probably derives its angular momentum vector from the original molecular cloud that collapsed to form the system.) Therefore, the accretion onto the black hole most likely occurs in a prograde sense ($0 < j < 1$). BZ powers, therefore, will be weak, because they only approach those of Blandford–Payne jets as j approaches (but does not equal) +1. (See Section 14.4.2 and Fig. 14.25.)
5. Finally, we also do not expect a jet launched by a rotating black hole horizon (BZ process) to have the same properties as those of a jet launched by an accretion disk, unless the accretion disk reaches well into the ergosphere and near to the black hole horizon, *i.e.*, $j \rightarrow 1$. Instead, we might expect BZ jets to have a very high Lorentz factor – as high as is inferred by GRBs deep inside their collapsar and the fastest AGN jets ($\gamma \sim 40$ –50 or higher). The jet emission, then, might be highly beamed into an angle of $\theta \sim 1/\gamma \sim 1^\circ$. As a result, one might expect the emission even from the base of such a jet to be of higher energy than regular X-ray binary jets, *i.e.*, at very hard X-ray and γ -ray energies.

So, even before we begin any observations, we expect that any attendant BZ jet may have a power less than or of order the BP disk jet; that it may be difficult

to distinguish from the BP jet using variability properties (as both will be equally affected by magnetic conditions in the accretion flow); and that the BZ jet may have a high Lorentz factor, be beamed, and emit at γ -ray energies. It is possible, therefore, that jets driven by black hole spin do exist, but we have not been lucky enough to have observed a γ -ray black-hole-binary “blazar” yet.

An easier, but still very difficult experiment, would be to measure the spin of the *magnetic field* in the jet itself instead of the spin of the black hole. High time resolution, X-ray polarization observations of jets from a single source in different accretion states would have different signatures for hole- and disk-driven jets. Black hole spin always would produce the same spin frequency (in the 300–1600 Hz range for $0.1 < j < 1$ and $M_{\bullet} = 10 M_{\odot}$). However, a significant variability in a source’s field rotation frequency would be a strong indication that this particular jet was, in fact, powered by the accretion disk (BP process) and not by the black hole (BZ process).

16.2.2.4 Jet Acceleration and Collimation

The previous subsections dealt mainly with the *launching* of jets from black hole engines (from Chapter 14 and Section 15.1). It is very fortunate that we can estimate the total and radio power of a jet simply from the rotating plasma that drives that jet. However, we learn little about the detailed physics of the jet itself. How is the jet accelerated and collimated into a supermagnetosonic, relativistic outflow? Where does it become supermagnetosonic, and where does the jet become causally disconnected from the central engine? What parameters control the jet’s final speed (or Lorentz factor) and opening angle? Can we use observations of the jet electromagnetic spectrum and the central engine as diagnostics of jet structure and physical properties?

As discussed in Chapters 13 and 15, we have used several types of tools to help answer such questions: semi-analytic models of thermal jets and of the production, acceleration, and collimation of magnetohydrodynamic jets. The former are routinely compared with data from jetted sources, and soon the latter also will be used for the same purpose. We also have two- and three-dimensional simulations of magnetic jet production, which afford a detailed look at jet physics, but as yet have not been applied much to modeling actual sources.

Recent Thermal Jet Models of Black Hole Broad-Band Data

The Markoff method of fitting thermal jet models to broad-band photon spectra of black hole sources was discussed in Section 13.3. This has been highly successful in deducing many conditions in the jet flow: (1) where (in physical coordinates) the jet becomes supersonic; (2) what the jet’s final Lorentz factor might be; and (3) the plasma conditions within the jet flow (as a function of distance from the black hole) including the synchrotron emission [470, 465]. The addition of Compton re-

flection processes to the computation of the broad-band spectra has produced tighter constraints on the models and their properties [468].

Of particular interest is the discovery by Markoff and colleagues [466] that a particle acceleration region (probably a stationary shock or other compression feature) is needed in the jet flow in order to explain a power law synchrotron spectrum that stretches from the submillimeter to X-ray regions. This discovery has important implications for MHD jet models (discussed below).

The most important results from the thermal jet models, in this author's opinion, are those on the jet geometry. The base of the jet is generally only a few r_g in radius – $R_{\text{base}} \lesssim 20 r_g$ and perhaps as small as $R_{\text{base}} \sim 3 r_g$ in some cases [467]. The latter suggests that some jet production may occur inside the ISCO of a non-rotating black hole or that rotating black holes may be involved. The jet flow reaches the sound speed very early – at a height Z_{base} of only $\sim 2 R_{\text{base}}$, which is not far above the accretion flow itself. It is clear, however, that the great majority of the acceleration of the jet continues well above Z_{base} . The height of the region where particles are accelerated to very relativistic energies varies greatly from source to source, however. Distances as small as $Z_{\text{acc}} \approx 10 r_g$ to $(1-3) \times 10^4 r_g$ have been found.

Present and Future MHD Jet Models of Black Hole Broad-Band Data

Chapters 14 and 15, respectively, discuss the details of models and simulations of MHD jet launching and of MHD jet acceleration, collimation, and release as a causally disconnected entity. The general properties of an accelerating MHD outflow are quite different from those of a thermal one, but they actually are very similar to the properties of jets *deduced* from the studies discussed above.

The causality structure of an accelerating MHD flow *in a steady state* was discussed in Sections 9.6.2 and 15.1. It is very important that astronomers and astrophysicists begin to grasp this subject, so the reader is strongly encouraged to read those sections. We give only a very brief review here, however.

A steady MHD flow crosses three important surfaces as it accelerates to highly supermagnetosonic speeds. These are called *separatrix* surfaces, because the information flow for a particular type of MHD wave is always *away* from each of one of the separatrix surfaces. That is, each of the three separatrix surfaces acts like an internal “boundary”, and each provides an internal constraint, called a *regularity condition*, on the flow. The three surfaces are the slow magnetosonic separatrix surface (or SMSS), the Alfvén surface (AS), and the fast magnetosonic separatrix surface (FMSS). The three constraints are, respectively, that the *normal* component of the velocity to each surface be equal to the slow magnetosound speed, the Alfvén speed, and the fast magnetosound speed. The FMSS plays the role of an MHD event horizon; information can cross the FMSS in one direction only (downstream), unless that information propagates faster than the fast MHD wave (*e.g.*, at the speed of light). Note: since the regularity condition is only on the velocity *normal*, the actual plasma speed through the surface can be much higher if the velocity makes a small

angle to the separatrix surface. This is especially true in jets, which can be highly supermagnetosonic and relativistic as they cross the FMSS and become causally disconnected.

MHD jets, therefore, accelerate in the following manner:

- At the base of the jet the outflow speed is well below the slow magnetosound speed, which is essentially the gas pressure sound speed. *This is the subsonic or “sub-slow” portion of the flow.*
- Gas pressure gradients accelerate the plasma along the field lines until, shortly thereafter, it crosses the SMSS and becomes super-slow (basically supersonic). (See Fig. 15.7.) *This surface occurs near the accretion disk or black hole.*
- Beyond this point, gas pressure forces are no longer effective, so magnetic forces take over, accelerating the flow with a rotating, torsional Alfvén wave. This is the magnetocentrifugal effect that Blandford and Payne [507] studied.
- The flow continues to be accelerated in this manner until it crosses the Alfvén surface and becomes super-Alfvénic. Almost all models and simulations show that the AS lies less than a factor of ten in height above the SMSS. For relativistic MHD jets that are Poynting-dominated, the AS lies near, but just inside, the light cylinder (see Fig. 15.15). *At the moment it is unclear whether the SMSS or the AS corresponds most closely to the sonic surface discussed by Falcke and Markoff [465].*
- With magnetocentrifugal Alfvén waves no longer effective, the total magnetic pressure gradient

$$\frac{dp_{\text{mag}}}{dZ} = \frac{1}{8\pi} \frac{d(B_R^2 + B_Z^2 + B_\phi^2)}{dZ}$$

becomes the dominant jet acceleration mechanism. *This is similar to the Falcke–Markoff gas pressure acceleration in a conical flow, except that now it adds the more dominant magnetic forces rather than just gas pressure alone.* In addition, there is a hoop stress (pinch) component of the force ($B_\phi^2/8\pi$) toward the rotation axis that not only collimates the flow, but also accelerates it slowly toward the jet axis. These two accelerations continue to great distances above the black hole ($10^{2-5} r_g$) until the flow speed toward the axis equals the fast magnetosound speed. This occurs near the FMSS – the third and final separatrix surface.

- Beyond the FMSS two things occur: (1) the total flow speed is so fast that it becomes magnetosonically *decoupled* from the central engine, and (2) the flow races toward the axis and bounces or shocks there in less than a transverse magnetosound crossing time. *It is this stationary compression feature that we tentatively identify with the particle acceleration region that Markoff needs in her models to produce broad-band synchrotron emission in X-ray binary and other radio jets.*

Peter Polko, Sera Markoff, and this author are working toward replacing the thermal jet models with MHD jet models and then fitting the latter to source broad-band spectral data. The first of these models should be completed by the end of 2012 or sooner. Besides realism, the MHD models will allow a more full diagnosis of

jet conditions, using a single, coherent jet structure. The six free model parameters mentioned in Section 15.1.4 have six corresponding observationally-derived parameters (jet luminosity L_J , black hole mass $M_\bullet$, R_{base} , Z_{base} , Z_{acc} , and the base plasma temperature from an accretion model fit to optical or X-ray data). Initially, the new models will give us better diagnostics of physical quantities like jet acceleration, collimation, magnetic field strength, Alfvén Mach number, temperature, *etc.* Eventually, however, the models should provide a deeper understanding of how and when jet production occurs in black hole engines.

So far, simulations have confirmed the above general picture of MHD acceleration and collimation up to and beyond the Alfvén surface. However, the formation of an FMSS and a stationary compression feature at $10^{3-4} r_g$ or more has not been seen yet in the simulations. In fact, *no* two- or three-dimensional simulation has been capable of doing so. In order to set up a steady-state jet outflow, one needs to run the simulation for at least 10–20 *longitudinal* magnetosound crossing times $(1-2) \times 10^5 \tau_g$. This would take perhaps many months, even for a two-dimensional simulation. However, for a black hole system even $10^5 \tau_g$ still is a rather short time: five seconds for a $10 M_\odot$ black hole, eight minutes for a $1000 M_\odot$ IMBH, six days for a $10^6 M_\odot$ Seyfert, and about 16 years for a $10^9 M_\odot$ quasar. Table 16.3 shows typical longitudinal crossing times for our array of black hole masses. So, once a jet begins to be ejected, it does not take very long for all three separatrix surfaces, including the FMSS, to be set up into a relatively steady-state system.

However, if we compare the magnetosound crossing times in Table 16.3 with typical disk thermal times in the unstable state (Table 16.2), we see that some thermal variations can occur faster than it takes to equilibrate the jet. It is possible, therefore, that jet ejection is somewhat erratic near the central black hole, appearing smooth perhaps only on times longer than the longest of thermal times.

16.2.3 *Isolated Stellar-mass Black Holes Accreting from the Interstellar Medium*

In Sections 3.4 and 3.5 we estimated that the number of stellar-mass black holes in the Galaxy could be 15–100 million (about one-third the number of accumulated neutron star remnants), many of which wander around the Galaxy alone without any binary companion. Therefore, the nearest black hole to the earth, on average, may be only ~ 60 pc (~ 200 light years) away. Up to the present time, isolated black holes have been (possibly) detected only by gravitational microlensing techniques. Below we assess the possibility of detecting the energy generated by accretion of the interstellar medium onto such objects by new hard X-ray telescopes like NuSTAR or by very sensitive radio telescopes like the VLA. Unfortunately, the predictions below are rather bleak; at this time gravitational microlensing remains the best method for finding isolated black holes, although some nearby objects could be detectable in the radio as microjansky-level sources with proper motions below 1 mas yr^{-1} .

16.2.3.1 Hard X-Ray Detection? No

The main issue in determining the detectability of an accreting isolated black hole is producing a good estimate of the black hole's *speed* through the ISM. Since we found above that black holes were produced at a much more rapid rate in the past than now, it is reasonable to consider the entire class to be rather ancient, with many having been produced when the Galactic bulge was formed. Indeed, some ancient LMXB microquasars have orbits that travel well out of the Galactic plane [590, 591]. So, a reasonable speed for isolated black holes would be at least $\sim 150\text{--}300\ \text{km s}^{-1}$.

From equation (11.9) the accretion rate of a $10\ M_{\odot}$ black hole traveling in this manner through the ISM is $\sim 10^{10-12}\ \text{g s}^{-1}$, where we have used a rather generous ISM density of $10\text{--}100\ \text{cm}^{-3}$. Nonetheless, this converts to only $\dot{m} \approx 10^{-9}\text{--}10^{-7}$, indicating very low accretion in the hard state. Estimated hard X-ray luminosities, therefore, will be only $10^{22-26}\ \text{erg s}^{-1}$. With NuSTAR's detection flux limit of $\sim 10^{-13}\ \text{erg s}^{-1}\ \text{cm}^{-2}$, objects of these luminosities can be detected as hard X-ray sources only at distances closer than $0.024\text{--}2.4\ \text{pc}$. This is considerably closer than the expected average distance of $60\ \text{pc}$. So, *it is extremely unlikely that we will be able to find a wandering stellar-mass black hole from its hard X-ray emission in the near future*. The main reason for this is the black hole's extreme *inefficiency* in generating hard X-ray emission when $\dot{m} \ll 0.05$.

16.2.3.2 Radio Emission Detection? Possibly

Isolated black holes are, in fact, the original type of object considered by Blandford and Znajek's original work (see Section 14.3). We see from equations (14.26) or (15.17) that, compared to the ADAF/MDAF X-ray luminosity, an accreting black hole in the hard state is dominated by the jet, not the accretion, luminosity. Even with an accretion rate as low as $10^{-9}\text{--}10^{-7}$, the radio power (from equation (16.24)) should be of order $10^{22-24}\ \text{erg s}^{-1}$ at GHz frequencies. So, if we assume a radio detection limit of $1\ \text{mJy}$ (*i.e.*, $10^{-17}\ \text{erg s}^{-1}\ \text{cm}^{-2}$ at $1\ \text{GHz}$), we find that isolated black holes could be detected at distances of $3\text{--}30\ \text{pc}$. And, if we use a detection limit of $\sim 10\ \mu\text{Jy}$, the detection distance moves out by a factor of ten to $30\text{--}300\ \text{pc}$.

Still, even in the most optimistic and most sensitive case, there would be only a few tens of such $10\ \mu\text{Jy}$ black holes scattered over a sky full of many more orders of magnitude of other objects with such weak fluxes. At $100\ \text{pc}$, however, their $300\ \text{km s}^{-1}$ motion would be detectable as a $\sim 0.6\ \text{mas yr}^{-1}$ proper motion, allowing observers to perhaps distinguish the nearby black holes from much more distant weak radio sources.

16.2.4 Stellar-mass Black Holes Accreting in Binary Systems

Accreting black holes in binary systems should, in principle, be simpler than their neutron star counterparts, because the former lack an intrinsic magnetic field. However, the black holes' lack of a hard stellar surface appears to add an equal, if not greater, amount of complexity, as it exposes all the black hole radii that we derived in Chapter 7 (r_{isco} , r_{mb} , r_{ph} , r_{E} , and r_{H}) – and all the relativistic effects that accompany them. The most important of these is r_{H} , of course, which allows matter and energy to be lost from the system without ever being able to emit radiation that observers on earth can detect. But, clearly, there are many relativistic effects that will affect how the accretion behaves. Collectively, these effects allow the inward accretion flow to assume a large number of different states. By some accounts, for example, the strange microquasar GRS 1915+105 may have a dozen and a half different such states, not just the standard two (hard and soft) varieties. And, unless there is something peculiar about the accreting matter in this system (rather unlikely), which accretion state the black hole attains should be governed largely by only one parameter: the *rate* of accretion onto the $14 M_{\odot}$ black hole in that binary system. In this subsection we shall frequently refer to two very good reviews on the properties of black hole binaries [204, 205] by Jeffrey E. McClintock (of the Harvard-Smithsonian Center for Astrophysics) and Ronald A. Remillard (of the Kavli Institute of Astrophysics and Space Research at MIT).

Binary black holes, like binary neutron stars, are divided into HMXB and LMXB systems. Like the neutron star HMXB systems the black hole ones also accrete from a stellar wind that is emitted by its O/B companion star. However, equation (11.65) shows that $\dot{M} \propto M^2$, so the normalized accretion rate $\dot{m}$ actually increases linearly with m . That is, if the neutron star accretion rate is $0.001 < \dot{m} < 1$ (Section 16.1.3 above), then the accretion rate onto a $10 M_{\odot}$ black hole will have a range $0.007 < \dot{m} < 7$. So, moderately super-Eddington accretion rates are possible, even when the black hole engine is powered by accretion of a wind from a high-mass companion star.

Without a strong surface magnetic field, some aspects of black hole LMXB systems are, indeed, less complicated than the neutron star X-ray pulsar, Z, and atoll source sequence. In fact, many LMXB black hole binary properties appear similar to those of atoll sources alone; this might be explained by both objects having weak or non-existent surface magnetic fields. Because stellar-mass black holes are an order of magnitude heavier than neutron stars, the normalized accretion rates for LMXBs on nuclear and thermal time scales (from equations (11.64) and (11.63)), are about an order of magnitude smaller than those for neutron star LMXBs: $\dot{m} < 5 \times 10^{-4}$ and $\dot{m} \leq 1$, respectively.

What is most interesting, however, is the possibility that an HMXB stellar companion could transfer mass to the black hole via Roche lobe overflow, particularly in the thermal case (equation (11.63)), which could produce an $\dot{m}$ up to 10^4 or higher ($>10^{22} \text{ g s}^{-1}$). As we shall see below, such enormous accretion rates would have very noticeable observable consequences for accretion black hole systems, producing something that looks like *neither* an HMXB nor an LMXB.

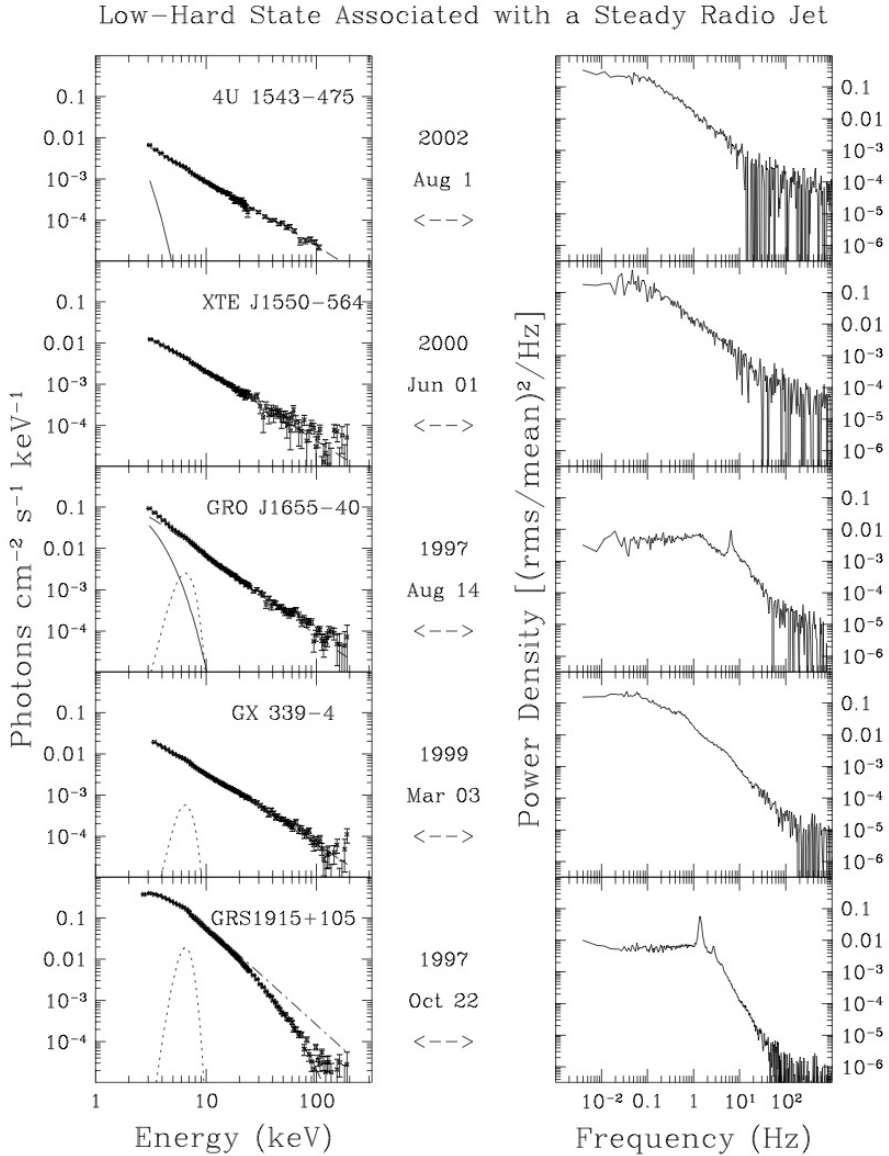


Fig. 16.3: Sample photon (left) and power fluctuation (right) spectra for five black hole binary sources in the hard state. The photon spectra are dominated by a flat power-law spectrum that stretches from 1 keV to at least 100 keV and possibly beyond (see Fig. 3.12). Typically, this is modeled as a hot ADAF/corona in the system. Note the need for several of these to also have a strong iron line (dotted curve), as well as a weak thermal component (solid curve) in some cases – both probably from a cool, geometrically thin disk. Many of the power fluctuation spectra are dominated by a power law component over three or more orders of magnitude in frequency, but a few (GRO J1655-40 and GRS 1915+105) have band-limited noise with a sharp cutoff above a few hertz and a strong and narrow QPO. Reprinted from Fig. 4.11 of reference [204], with the permission of CUP. Copyright ©2006 Cambridge University Press.

16.2.4.1 Binary Black Holes in the “Hard State” with an Associated Steady Jet ($\dot{m} < \dot{m}_A$)

Fig. 3.9 gave a single example of the photon spectrum of a neutron star source in a hard accretion state. Fig. 16.3, on the other hand, shows such spectra for five different *black hole* (microquasar) systems: the sources 4U 1543-475, XTE 1550-564, GRO J1655-40, GX 339-4, and GRS 1915+105 [204].⁵ In most cases the spectra have a power law form ($f_\nu \propto \nu^{-\alpha}$, where the spectral slope $\alpha \approx 0.55$ – 0.60). However, GRO J1655-40 and GRS 1915+105 differ significantly, with $\alpha \sim 0.70$ and 1.34 , respectively. On the right-hand side of Fig. 16.3 we see the power fluctuation spectrum for each of these sources, when the source is in the hard photon state that is depicted in the left-hand side.

Note that the frequencies f of the power fluctuation spectra in Fig. 16.3 range from 1 mHz to 1 kHz and show the energy in fluctuations per unit frequency $\mathcal{P}_f$. To convert these to “equal energy plots”, one must multiply by f to obtain $f\mathcal{P}_f$. Again, 4U 1543-475, XTE 1550-564, and GX 339-4 are similar with power law spectra $\mathcal{P}_f \propto f^{-\beta}$ and $\beta \approx 1.1$ – 1.4 . And, again, GRO J1655-40 and GRS 1915+105 are quite anomalous, with a three-part power spectrum: $\beta_1 \approx 0.066$ (nearly flat in $\mathcal{P}_f$), $\beta_2 \approx 2.5$ (very strongly falling off above a few hertz), and a strong quasi-periodic oscillation (QPO) near the spectral break. This strong increase in power with f , followed by a sharp drop in power for $f > f_{break}$, is sometimes referred to as *band-limited* or *bandwidth-limited* noise.

Hard State Sources Probably Dominated by an ADAF ($\dot{m} \ll \dot{m}_A$)

Since advection-dominated accretion flows (equations (12.50) and (12.58)) are the classical model for objects in a hot accretion state, we first should check to see if ADAFs fit these photon and power spectra. Surprisingly, they do, at least for the three above objects *without* bandwidth-limited noise.

ADAF Photon Spectrum. To estimate the photon spectrum we use a formula similar to equation (12.12) with a “multi-colored” approximation. That is, we assume that each locally-emitted spectrum in the disk is sufficiently peaked at a frequency $\nu \approx kT(R)/h$ that the portion of the *total* spectrum at that frequency is emitted *only* at the radius $R = R(T(\nu))$. In this case, then, the spectrum is given by

$$\begin{aligned} \nu f_\nu &= 2\pi \int_{R_{in}}^{R_{out}} \nu \varepsilon_\nu H_{\text{skin}}(R') R' dR' \\ &\approx 2\pi \varepsilon(R) H_{\text{skin}}(R) R \Delta R \end{aligned}$$

⁵ Note that these spectra are plotted as *number* of photons per second per unit energy – the same kind of plot as in Fig. 3.8. In order to convert them to an f_ν plot, one must multiply by ν (or E). And, in order to convert them to the νf_ν -type plot in Fig. 3.9, one must multiply by ν^2 (or E^2).

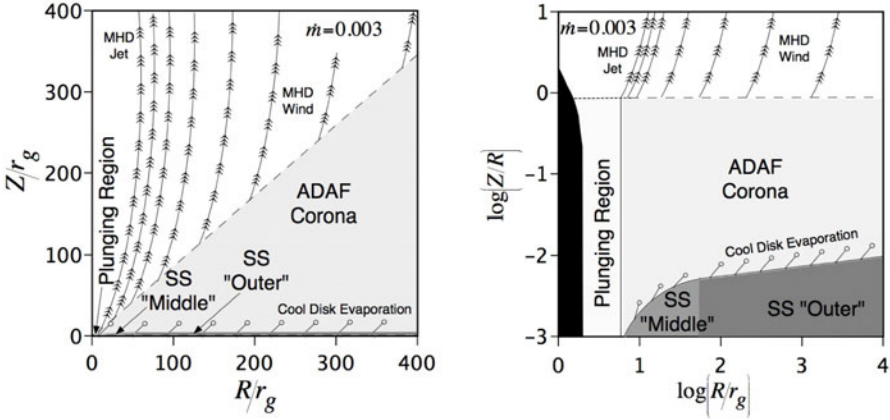


Fig. 16.4: Our model for the hard state, based on the discussion in Section 12.2 and in this chapter: a geometrically thin, evaporating Shakura–Sunyaev disk with a dominant ADAF corona. Disk parameters used are $m = 10$, $\dot{m} = 0.003$, $\alpha_{\text{adf}} = 0.3$, and $\alpha_{\text{ss}} = 0.03$. *Left:* A linear plot of half of a meridional cut through the accretion flow. Note the presence of a strong jet and MHD wind, which were predicted in Section 15.1.1 and simulated numerically in Section 15.1.4. (Compare with a bare ADAF in Fig. 12.6a.) *Right:* Same accretion model but in a skewed-logarithmic plot (introduced in Fig. 12.3d). The thin disk, plunging region, and black hole are now much more visible.

where $\varepsilon(R) = \nu \varepsilon_\nu$ is the total volume emissivity across that peaked spectrum at each radius R , and $H_{\text{skin}}(R)$ is the visible skin depth there for an observer. For optically thin ADAFs $H_{\text{skin}} \approx 2H \approx 2R$, and we take the Bremsstrahlung emissivity $\varepsilon(R) \propto \rho(R)^2 T(R)^{1/2}$. From equations (12.50) and (12.58), using $\nu \propto T \propto R^{-1}$, and taking $\Delta R \propto R$, we obtain $\nu f_\nu \propto R^{-1/2} \propto \nu^{1/2}$ or

$$f_\nu \propto \nu^{-1/2} \quad (16.31)$$

or $\mathcal{N}_E \propto E^{-3/2}$, which is very similar to the sources in Fig. 16.3 that do not have bandwidth-limited noise.

So, the hard power law spectra are consistent with being the sum of many optically thin spectra produced at different radii in a typical ADAF. Note, however, that the observed spectra in Fig. 16.3 sample only the *one-temperature* portion of the ADAF (*i.e.*, $T \leq 10^9 \text{ K} \approx 100 \text{ keV}$). It will be important to also obtain γ -ray spectra for these sources *above* 100 keV in order to determine how hot the electrons get as the ADAF splits into a two-temperature flow.

This discussion of the black hole X-ray spectrum is somewhat different from that usually included in papers and books on the subject. In the latter cases, authors generally assume that the accretion disk corona produces a power law spectrum of the kind we just derived (plus perhaps a thermal accretion disk spectrum added in) and then proceed to calculate the more subtle spectral features (a Compton reflection hump at a few tens of keV plus fluorescence lines, primarily the iron line at 6.4 keV rest wavelength). These spectral features occur when the power law emission from

the corona impinges on a cooler disk which, in this model, may lie along the equator inside a hotter ADAF X-ray corona (Fig. 16.4). See [92, 220] for a more detailed discussion of these reflection/fluorescence features.

ADAF Fluctuation Spectrum. A similar calculation can be performed for the fluctuation spectrum

$$f\mathcal{P}_f \propto \rho V_{turb}^2 H_{skin}(R) R \Delta R \propto R^{1/2}$$

We have assumed here that the energy per unit volume in turbulence is ρV_{turb}^2 and, again, a skin depth proportional to R . Now, at each radius the turbulent frequency $f \approx V_{turb}/H \propto R^{-3/2}$, so $f\mathcal{P}_f \propto f^{-1/3}$, leading to

$$\mathcal{P}_f \propto f^{-4/3} \quad (16.32)$$

which, again, is consistent with the non-bandwidth-limited sources in Fig. 16.3, as well as many more like them. While observations above $f \approx 10\text{--}100\text{ Hz}$ can be affected by measurement noise, it appears that the $f^{-1.3}$ fluctuation spectrum may continue very close in to the black hole in these systems. That is, in these cases the ADAF may also continue all the way to the ISCO, with no MDAF in its interior.

The ADAF Steady Jet. If turbulent ADAF structure exists down to the black hole ISCO, then there is reasonable confidence that the jet is produced in a manner similar to that simulated in Fig. 14.40(lower right) and 15.14. From equation (15.17) this means that

$$L_J \approx L_{Edd} \dot{m}$$

when accreting in pure ADAF mode. This implies an L_R vs. L_X relation, derived from equations (16.24) and (12.58), of

$$L_{R,adf} \propto L_{J,adf}^{1.2} \propto \dot{m}^{1.2} \propto L_X^{0.6}$$

which agrees with the fundamental plane (see Fig. 16.5). So, again, ADAF theory seems to fit nicely most X-ray binary sources in the hard state.

Hard State Sources Possibly Dominated by an MDAF ($\dot{m} < \dot{m}_A$)

Recently, Rob Fender at Southampton (UK) and his colleagues at the University of Amsterdam, Elena Gallo, David Russell, and Paolo Soleri, report the discovery of a new class of hard-state X-ray binary black holes that do not lie in the fundamental plane. These sources have been called “outliers” or “radio-quiet black hole binaries” [588, 592].⁶

⁶ This author cautions against the use of “radio-quiet” to describe these sources. They lie only an order of magnitude or less in radio power below the fundamental plane. On the other hand,

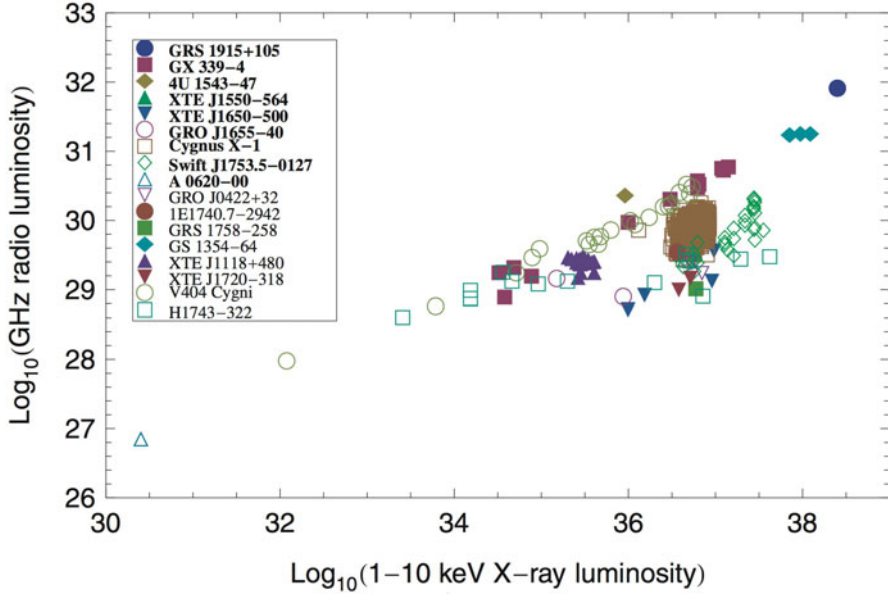


Fig. 16.5: The radio/X-ray luminosity (fundamental plane) plot for hard state black hole X-ray binaries. The upper envelope is the true fundamental plane; the sources that lie beneath are the “outliers”. Reproduced from Fig. 2 of [588], by permission of J. Wiley and Sons, Ltd.

In Fig. 16.5 the outliers are plotted along with the fundamental plane sources. A comparison of these data with the power spectra in Fig. 16.3 (plus similar Figs. 4.12 and 4.15 in [204]) shows a tentative correlation between the outliers and sources with band-limited noise, as well as one between fundamental plane sources and those without band-limited noise (*i.e.*, ADAF-like). Since the magnetically-dominated accretion flows discussed in Chapter 12 were invented (by this author) to explain band-limited noise with a low-frequency QPO, and since MDAF models predict a suppression in jet power, it is intriguing to explore the possibility that (1) the band-limited sources and outliers are, indeed, the same objects and (2) these sources are black hole binaries that are in the MDAF state when observed.

The ϑ -parameter and the MDAF Steady Jet. At this time we do not have a predictive theory for $T_{i,\text{sat}}$, and therefore ϑ . However, we can estimate its value from sources when they are in their band-limited, plateau state by identifying the QPO frequency as the rate of rigid rotation of the MDAF magnetosphere (see equations (12.85) for ν_f):

real radio-quiet quasars (RQQs) can be five or more orders of magnitude weaker in the radio than RLQs. And none of the quasars (RQQs or RLQs) are in an accretion hard state. The “radio-quiet” label, therefore, implies that there is an analogy between radio-quiet QSOs and the black hole binary outliers, when in reality none may exist.

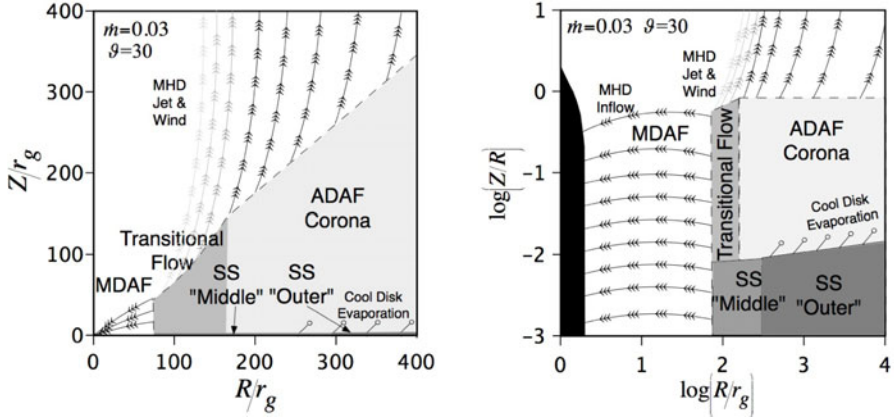


Fig. 16.6: The MDAF model for hard state outlier sources – similar to Fig. 16.4, but with $\dot{m} = 0.03$ and $\vartheta = 30$. Again, linear plot is on the left and skewed-log plot on the right. Thermal properties of MDAFs are very similar to those of ADAFs, but mechanical properties are quite different (see Section 12.3). The accretion rate (and thin disk luminosity!) are up to a factor of ten higher than in Fig. 16.4. In this case the ADAF inside $R_0 = R_A/\vartheta \approx 160 r_g$ cools (possibly because of a significant source of photons from the SS “Middle” region). ADAF pressure support is lost, the plasma β_p becomes unity at $R_1 = R_A \alpha^{2/3}/\vartheta \approx 70 r_g$, and an inward-facing magnetosphere develops. A jet is launched only from the turbulent portion, reducing its luminosity by a factor of $R_1/r_{\text{isco}} \approx 12$ and making the source weak in the radio for its rather strong X-ray luminosity. Silent and relatively cool ($T_i \approx 3 \times 10^{10}$ K, $T_e \approx 10^9$ K) accretion occurs along the magnetosphere field lines and into the black hole. The power fluctuation noise is band-limited above a frequency of $\nu \approx \Omega_K(R_1)/2\pi \approx 5$ Hz. (Compare with the pure MDAF model in Fig. 12.9.)

$$\vartheta \approx 30 \left(\frac{M_\bullet}{10 M_\odot} \right)^{2/3} \left(\frac{\nu_{QPO}}{5 \text{ Hz}} \right)^{2/3} \left(\frac{\alpha}{0.3} \right)^{2/3}$$

For $\alpha = 0.3$, we obtain for GRO J1655-40 ($m = 7$, $\nu_{QPO} \approx 7$ Hz) and for GRS 1915+105 ($m = 14$, $\nu_{QPO} \approx 1.5\text{--}3.0$ Hz) values of ϑ in the range 20–30. The total jet luminosity is still given by equation (16.23),

$$L_J \approx 0.037 L_{Edd} \dot{m} (\vartheta/30)$$

but with $\vartheta \approx 30$, and the equivalent to equation (16.24) for the jet radio power in the MDAF case is

$$L_R = 1.17 \times 10^{29} \text{ erg s}^{-1} \left(\frac{M_\bullet}{10 M_\odot} \right)^{1.38} \left(\frac{\dot{m}}{0.05} \right)^{1.2} \left(\frac{\vartheta}{30} \right)^{1.2} (\dot{m} < 0.05) \quad (16.33)$$

Note that a canonical value of $\vartheta \sim 30$ implies that the ions saturate at a rather cool temperature of $T_{i,\text{sat}} \sim 3 \times 10^{10}$ K, with the MDAF supported primarily by magnetic pressure, rather than ion pressure at the virial temperature.

Jet Radio Power vs. X-Ray Luminosity for Outliers. Fig. 16.5 shows that the L_R vs. L_X relation for outlier sources is different from that for sources in the fundamental plane. The latter show $L_R \propto L_X^{0.6}$ (cf. equation (16.24)), while the outlier sources (not including Cyg X-1) show a steeper function (approximately $L_R \propto L_X^{1.5}$), stretching from XTE J1720-318 (at $\log L_X \approx 36.5$) to GRS 1915+105 (at $\log L_X \approx 38.5$).

How is it possible that the radio jet power has a different dependence on L_X in MDAF theory than in ADAF theory? Clearly it is not due to any difference in L_X dependence between the two accretion states; comparison of $L_{\text{acc,mdf}}$ (equation (12.87)) and $L_{\text{acc,adf}}$ (equation (12.58)) show that MDAF and ADAF accretion luminosities (and spectral temperatures) should be essentially identical. So the difference (if any) must be in the predicted jet radio power in the two theories. For MDAFs we must use equation (16.33)

$$L_{R,\text{mdf}} \propto L_{J,\text{mdf}}^{1.2} \propto \dot{m}^{1.2} \vartheta^{1.2} \propto L_X^{0.6} \vartheta^{1.2}$$

So, for $L_{R,\text{mdf}}$ to be proportional to $L_X^{1.5}$, ϑ also must vary with L_X

$$\vartheta \propto L_X^{0.75}$$

However, if this supposition is true, then the frequency of the low-frequency QPO *also* must vary roughly linearly with L_X . That is, if we are to explain outliers as sources in the MDAF state and explain the steeper slope of outliers in Fig. 16.5, then we must have⁷

$$\nu_{QPO} \propto L_X^{1.1}$$

This is indeed the case for at least one band-limited source: GRS 1915+105; see Fig. 2 in [593].

So, MDAF theory seems to be at least consistent with the jet suppression (“radio quietness”), band-limited noise, steeper L_R vs. L_X slope, and increase in LF QPO frequency with L_X of outlier sources. However, the dependence of ν_f on L_X is an empirically determined relation only; it is not yet a fundamental prediction of MDAF theory.

Finally, since the jet speed then will increase with the ϑ parameter

$$V_j \approx \left(\frac{GM}{R_1} \right)^{1/2} \propto \vartheta^{1/2} \propto L_X^{0.4}$$

V_j also should increase with L_X in outlier sources, starting at non-relativistic speeds and then reaching moderately relativistic speeds as the source becomes more ADAF-like. Unfortunately, jet speed is very difficult to measure in these sources at the present time.

⁷ Note that uncertainties in slopes of L_R vs. L_X in the fundamental plane and for outlier sources put this exponent in the range 0.9–1.1.

Band-Limited Source Photon Spectrum. In addition to differing in their power spectra, band-limited and non-band-limited sources also differ in their photon spectra, although the differences are a bit more subtle. The left side of Fig. 16.3 shows not only the entire photon spectrum (plotted points), but also its breakdown into the standard three components: power law (dot-dashed line), iron line (dotted line), and a thermal accretion disk (solid line). It is clear that the iron line and disk emission are much stronger in the band-limited sources than in the others. It is possible, then, that many of the properties of the band-limited (and possibly outlier) sources (*i.e.*, the possible MDAF structure, strong LF QPO, weaker jet, *etc.*) may be due to the presence of a much more prominent accretion disk inside the band-limited sources. (Compare Figs. 16.4 and 16.6.) It is important to remember that both the ADAF-like corona and the cool interior disk are accretion flows, with angular momentum transport as described in Section 12.2. However, the two flows may be exchanging matter with each other, with $\dot{M}$ in each flow varying with cylindrical radius R .

The other noticeable difference in the photon spectra is that, in the band-limited sources, the hard power law is somewhat steeper than $f_\nu \propto \nu^{-1/2}$. Since this emission is probably produced in the ADAF-like corona, a possible explanation for this steeper power law may be a shallower density gradient in the band-limited sources (*i.e.*, $\rho(R) \propto R^{-k}$, where $k < 3/2$). This would imply that the accretion rate in the corona is *decreasing* with decreasing R as the inflow approaches the black hole; *i.e.*, $\dot{M}_{corona} \propto R^n$, where $n > 0$.

Band-Limited Source Fluctuation Spectrum. The two features of the fluctuation spectrum in band-limited sources that fit well with MDAF theory are the very steep cutoff in power at high frequencies and the low-frequency quasi-periodic oscillation near the power break. These correspond to transitional flow and the rigidly-rotating, strong-field MDAF, respectively. Using equations (12.66), we can estimate the slope of the transitional flow power spectrum as we did for equation (16.32)

$$f\mathcal{P}_f \propto \rho V_{turb}^2 H R \Delta R \propto R^2$$

In this case $f \propto V_{turb}/H \propto R^{-3/2}$, so

$$\mathcal{P}_f \propto f^{-7/3}$$

This is very similar to the $\mathcal{P}_f \propto f^{-7/3}$ shown by GRO J1655-40 and GRS 1915+105 above the spectral break (Fig. 16.3).

Interior to this transition region we expect a rigidly-rotating magnetosphere (MDAF) stretching toward the black hole. If there were no turbulence (fluctuations) in the MDAF (*i.e.*, $V_{turb} = 0$), the QPO would be infinitely narrow and have a frequency equal to the Keplerian rotation rate at the inner edge of the transition flow. In these two sources this inner edge turns out to be roughly $280 r_g$ and $170 r_g$, respectively, for masses of $7 M_\odot$ and $14 M_\odot$.

One serious issue, however, is that neither MDAF theory, nor ADAF theory, predicts a nearly flat ($\mathcal{P}_f \propto f^{0.066}$) power spectrum. A tentative, but plausible, suggestion is that this unusual turbulence comes from the one accretion component that is

not predicted by either MDAF, or ADAF, theory, but that we know likely exists in the band-limited systems: the fairly prominent, cool accretion disk contained inside the corona. In order for the turbulent power ($f\mathcal{P}_f$) to increase inward in the disk, the accretion rate in the disk will have to *increase* inward ($\dot{M}_{disk} \propto R^n$, where $n < 0$) – just the opposite of what we deduced above for the corona. In fact, it is easy to develop an accretion theory when the accretion rate is a power law function of R . For example, the derivation of the Shakura–Sunyaev “middle” region disk structure equations (12.31) does not assume that $\dot{m}$ is a constant; that occurs only when we assume that parameter is a constant in the equations. Therefore, one could simply replace that parameter with $\dot{m}_{isco} x^n$ and allow n to be an unknown to be solved for. Using the data in Fig. 16.3 and assuming that the “middle” region is the proper model, we find that $n \sim -2$ works quite well. (Actually $n = -13/6$ fits the flat power spectra nearly exactly.)

Tentative Model. So, we arrive at a fairly consistent, if very tentative, picture of the band-limited (outlier?) hard state objects: sources in which the ADAF-like corona has an accretion rate that *decreases* toward the black hole because it is condensing into a cool disk inflow along the equator, and that disk’s accretion rate is *increasing* inward as it accepts more condensed material. MDAF theory postulates that, if the energy coupling between ions and electrons is significant enough, the ion temperature in the corona will saturate at a value well below 10^{12} K: for the $\vartheta \sim 30$ found here, this temperature would be of order 3×10^{10} K. Interior to $(820/\vartheta)r_{isco} \sim 150 r_g$, then, the inflow transitions to one supported by magnetic, rather than ion, pressure. Still not fully understood, however, is why *both* the corona and disk are truncated at a similar radius, so that both disk and corona give way to the formation a single, inner, magnetically-dominated (MDAF) inflow.

Why certain sources display this kind of hard state, while many others seem to have a simpler ADAF one, is a mystery. Some sources seem to return to this same state frequently, while others seem to not display the band-limited noise at all. Are some more prone to having a condensing corona? If so, why? Is this a function of accretion rate or perhaps black hole spin? Much more detailed and numerical work is needed in order to verify this picture and answer these and other questions.

16.2.4.2 Binary Black Holes in the “Soft State” ($\dot{m}_A < \dot{m} < \dot{m}_I$)

One of the great advantages of studying black holes of stellar mass is that their accretion properties can change on time scales much shorter than a human lifetime – often in a few months or even hours. Fig. 16.7 shows the same sources as in Fig. 16.3 but now when they are in a higher, soft accretion state. Unlike the hard state, the photon spectrum of the soft state is dominated by the thermal emission of a cool, optically thick accretion disk and the fluctuation spectrum is nearly featureless with a rough $\mathcal{P}_f \propto f^{-1}$ power law. However, as we saw in Figs. 3.8 and 3.12, black holes in a soft state display a hard X-ray tail, which is produced probably by a hot

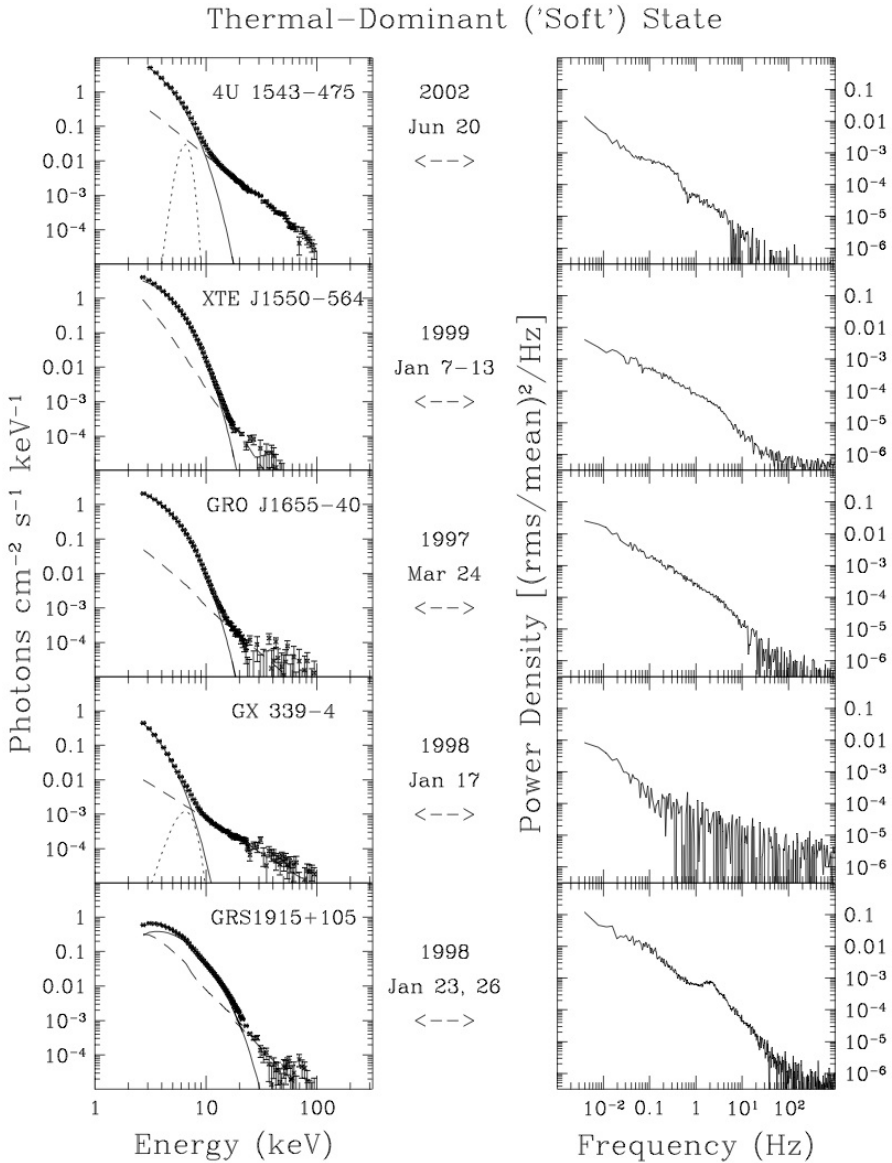


Fig. 16.7: Photon and power spectra for the same five sources in Fig. 16.3 when they enter the soft state. The photon spectra now are dominated by a strong and cool thermal component (thin solid curve) at low energies, but there still is a weak hard component that goes out to several tens of keV. Normally such spectra are modeled as a cool disk with a compact corona (see Fig. 16.8). The power spectra are generally featureless, except for the possibility of a broad QPO. Even sources that were band-limited in Fig. 16.3 now have measurable power out to many tens of hertz. Reprinted from Fig. 4.8 of reference [204], with the permission of CUP. Copyright ©2006 Cambridge University Press.

corona above the disk. This corona is generally believed to be more compact than a full ADAF, but it appears to have a lot of the same properties.

Soft State Source Photon and Power Fluctuation Spectra

We can determine the photon and fluctuation spectra as before, but now we use equations (12.31) for an optically thick, Shakura–Sunyaev “middle” region:

$$\nu f_\nu \propto T_*^4 R \Delta R \propto R^{-1} \propto T_*^{4/3} \propto \nu^{4/3}$$

or

$$\boxed{f_\nu \propto \nu^{1/3}} \quad (16.34)$$

and

$$f \mathcal{P}_f \propto \rho V_{\text{turb}}^2 H_{\text{skin}}(R) R \Delta R \propto R^{7/20} \propto f^{-7/30}$$

or

$$\boxed{\mathcal{P}_f \propto f^{-37/30}} \quad (16.35)$$

The $\nu^{1/3}$ spectrum is the same one found in equation (12.13). In order to obtain the fluctuation spectrum estimate we have used $H_{\text{skin}}(R) \approx 1/(\kappa_* \rho)$ as the $\tau_* = 1$ skin depth (see equation (12.32)) and $\kappa_* = (\kappa_{\text{ff, bf}} \kappa_{\text{es}})^{1/2}$ is the effective absorption opacity in the electron-scattering disk.

While the rough estimate of the fluctuation spectrum is similar to the $\mathcal{P}_f \propto f^{-1}$ observed, the $\nu^{1/3}$ spectrum is nowhere to be seen in Fig. 16.7. Indeed, it has been remarked in the past that the spectra of X-ray binaries look nothing like the Shakura–Sunyaev prediction.

However, this is not the proper way to state the situation. No X-ray spectrum shows the $\nu^{1/3}$ rising slope because that part of the spectrum does not occur in the X-ray band. It actually occurs in the UV to soft X-ray band, much of which is absorbed by hydrogen and other interstellar elements. We see only the warmest part of the thermal spectrum in the few keV range as it falls off exponentially above 10–20 keV. Fig. 16.8 shows plots of the standard model of the soft state.

Jet Suppression in the Soft State

In 1999 Rob Fender reported that, for the same source at approximately the same accretion rate, jets in the soft state were a factor of 50 or more less luminous in the radio than jets in the hard state. (See also Fig. 16.2.) That is, jets in the soft state are *suppressed* relative to the hard state. Migliari and Fender [581] elaborate quite

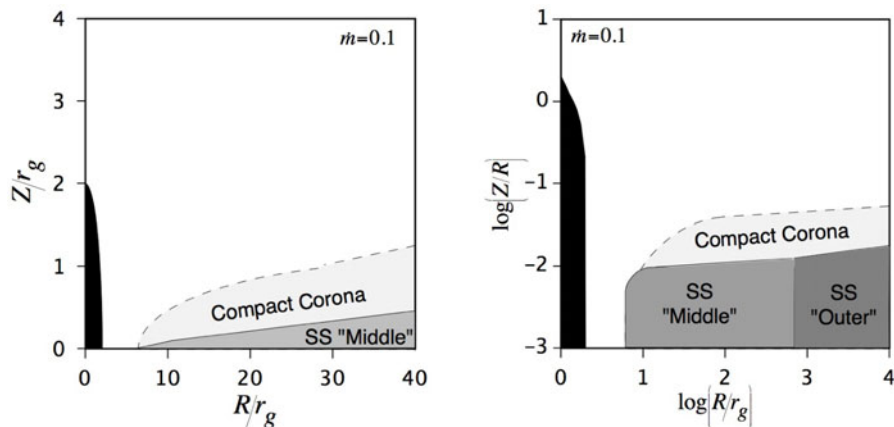


Fig. 16.8: Typical model of the soft state, with a thin Shakura–Sunyaev disk and a compact corona. Plot is similar to Fig. 16.4, but with $\dot{m} = 0.1$. The accretion rate is a factor of three higher than that in Fig. 16.6. Note the stretched vertical scale on the linear plot at left. The compact corona is probably all that is left of the once mighty ADAF in the hard state models, largely quenched by the strong flux of cool photons from the bright SS “Middle” region. Note the absence of a jet from the corona, chiefly (in this model) because the plasma is trapped deep in the thin disk potential well, with no strong pressure forces to lift it out. Power fluctuations of different frequencies come from different radii in the disk and corona, including regions close to the black hole.

a bit on this subject; see also Section 16.1.3. In 2001 this author suggested that the main reason for this type of jet suppression, which occurs as both neutron star and black hole systems transition from a hard state to a soft state, is that the vertical magnetic field strength B_{Z0} is proportional to the ratio of disk height H to disk radius R . (See also [431].) Since the Blandford–Payne jet power is proportional to B_{Z0}^2 (equations (15.14) to (15.16)), the thin, cool disk in the soft state should produce a much weaker jet. This effect has been included in the jet total luminosity equations (16.23), (16.26), (16.28), and (16.29).

Comparison of the ADAF hard state jet power (equation (16.23)) with that in the soft state (equation (16.26)) shows a factor of ~ 57 suppression in *total* jet power. Combined with the observation that jet *radio* power varies more than linearly with total jet power ($L_R \propto L_J^{1.2-1.4}$), this predicts a suppression in radio power of $\sim 130-290$ – in comfortable agreement with the observed > 50 value.

The observed jet suppression in the soft state, like the mild suppression between the outlier and fundamental plane sources, also has been suggested as an explanation of the radio quietness of most QSOs. Again, however, it is very unlikely that the soft state suppression can completely explain the difference between radio-loud and radio-quiet quasars, any more than the outlier jet suppression can. Several of the reasons for this conclusion are similar to those stated for the outlier sources. Real RQQs can be five or more orders of magnitude weaker in the radio than RLQs, and none of the objects identified as quasars (RQQs or RLQs) appear to be in an accretion hard state. (The best hard state candidates are ones *without* strong optical

emission – FR I radio galaxies or LLAGN.) Furthermore, there is every indication that RQQs and RLQs are very similar objects, except that, for some reason, the former produce much weaker jets. They all seem to produce similar thermal emission (*i.e.*, be in a similar accretion state), but produce very different nonthermal (radio jet) emission.

16.2.4.3 An Unrecognized State?: Binary Black Holes in the “Unstable State” ($\dot{m}_1 < \dot{m} < 1$)

The observational definition of an accretion state used by McClintock and Remillard [204, 205] is that the binary black hole source must have the same or similar properties for a significantly long term (many days or weeks) and must not simply be a short-term transitional condition between two other recognized states. This definition suffices for all currently recognized states, plus a new “super-soft” state that we propose below.

However, a theoretician might propose a slightly modified version of this definition: an accretion state occurs when the accretion rate at reasonably large distances from the black hole (say, $10^{3-4} r_g$) remains within a range such that the basic behavior of the source remains similar for long periods of time. At first, this might seem equivalent to the above observational definition. However, it allows for another type of state: an *unstable* one, in which the source’s behavior changes with time, but otherwise displays the same properties over and over again. That is, the instability exhibits cyclic behavior.

There is at least one such unstable state predicted by standard accretion theory, and there appear to be several candidate behaviors occurring in nature that corresponds to such a state. When the accretion rate is large enough to create an unstable Shakura–Sunyaev “inner” region, but still less than or of order the Eddington accretion rate, the accretion flow may exhibit a limit cycle behavior. Fig. 16.9 shows a theoretical accretion flow in just that range; unlike the previous Figs. 16.4, 16.6, and 16.8, an accretion rate this high (~ 0.3 Eddington) has an unstable Shakura–Sunyaev “inner” region (equations (12.35)). In this case the “inner” region attempts to form a stable, thin and cool disk, but that disk may be unstable to secular and thermal instabilities that could heat it to a rapidly-accreting ADAF-like substate; in that substate, then, the material in the “inner” region would bloat into a thick torus and rapidly accrete toward the black hole, producing a strong, moderately relativistic jet, as ADAFs do, but now at perhaps a much higher accretion rate than standard ADAFs. If the instantaneous accretion rate in this ADAF-like substate is of order the Eddington value, then the jet power could easily equal, perhaps even exceed, the Eddington limit during this outburst. The cycle then would repeat, with the material entering the outer radius of the “inner” region once again trying to form a stable, geometrically thin cool disk. See Section 12.2.2 and Fig. 12.5. In effect, the “inner” region cycles through at least two substates, each of which mimics a known state (particularly the soft and hard states). The difference is that the accretion rate in

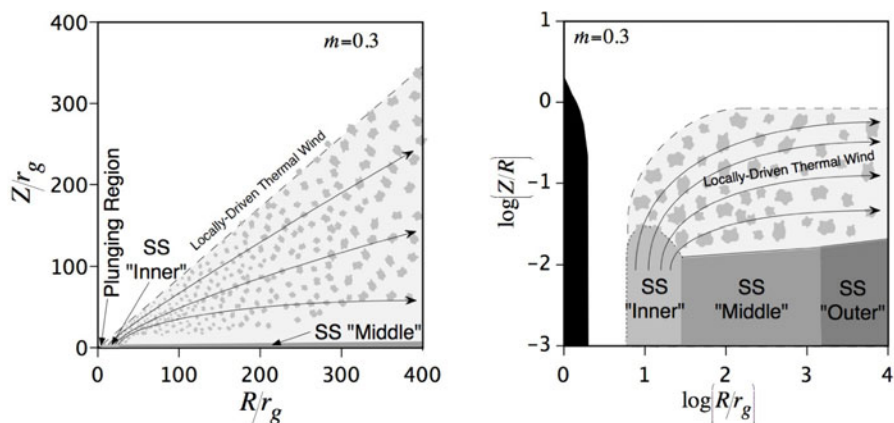


Fig. 16.9: Model of a proposed *new* state, the unstable state, which occurs for accretion rates so high that an unstable SS “Inner” region exists ($0.2 < \dot{m} < 1.0$ for stellar-mass black holes; nearly $0.01 < \dot{m} < 1.0$ for supermassive ones). Plot is similar to Fig. 16.4, but with $\dot{m} = 0.3$. Recent numerical simulations indicate that the inner region may fluctuate on its range of thermal time scales (perhaps producing broad QPOs), but it does not develop thermal runaways [452]. However, it still may be unstable to the Lightman–Eardley instability [454], causing limit-cycle behavior on a secular time scale. A strong thermal wind likely emanates from the “inner” region, driven locally possibly by the near-Eddington radiation pressure, enhanced by a large absorption line force multiplier, or by thermal pressure (like an ADIOS wind), if some thermal heating does occur. This wind may be responsible for QSO broad line and UFO activity. However, a similar candidate wind phenomenon has not yet been identified in the X-ray binary case.

these substates can be larger (sometimes much larger) than the standard states that they mimic.

The 1996–97 observations of GRS 1915+105 by Michael Muno, Edward Morgan, and Ronald Remillard at MIT [593] show several types of cyclic behavior in their Fig. 1. This source is believed to remain near, or sometimes even above, the Eddington accretion rate for very long periods of time. Often the source quickly cycles between a hard substate and a soft substate, sometimes ejecting a jet and sometimes not. The jet ejection, when it occurs, usually happens at the end of a relatively long (several hundred second or longer) hard substate. Usually the QPO frequency increases with the intensity of the hard X-ray radiation, as one might expect from a shrinking MDAF inside an ADAF as the accretion pulse makes its way toward the black hole. Occasionally the hard substates are fairly luminous, not very stable, and separated by short low, *soft* substates, which are quite stable. Several of these different states have been modeled numerically by Tomaso Belloni and colleagues [594] and by Ron Taam and his colleagues [595] as disks that empty and refill or as the temporary formation of a hot corona to explain these various behaviors.

There is a popular misconception that is sometimes voiced about these sources: that the entire disk “inner” region is *ejected* in the explosive jet. Energetically, this *cannot* be the case. The source of the jet energy is gravitational accretion of the “inner” region material toward the black hole. However, it would take all the accretion

energy generated to lift the entire “inner” region to infinity, and even more to accelerate that material to mildly relativistic velocities. Instead, it must be the case that *most* of the “inner” region disappears down the black hole and much of its liberated binding energy is used to eject only a tiny fraction of the “inner” region’s total mass.

It is quite possible that these behaviors may be triggered by the source being in an accretion state in the high but sub-Eddington $\dot{m}$ range where instabilities are expected. However, it also appears that there are multiple kinds of behaviors within this range and, therefore, multiple unstable states, rather than just one. This is, therefore, a very complicated range of accretion rate, and a full understanding of GRS 1915+105 and other sources like it is still perhaps many years off.

16.2.4.4 Binary Black Holes in the “Steep Power Law State” ($1 < \dot{m} \lesssim 10$)

McClintock and Remillard [204, 205] identify another important, and steady, state and argue strongly that it is a true, long-term one, not simply a transitional one between the soft and hard states. They call it the “steep power law” or SPL state, but it often is referred to as the “very high state” (VHS). In their definition this state is characterized by

- “The presence of a power-law component in the X-ray spectrum with a photon index of $\Gamma > 2.4$ ” (*i.e.*, $f_\nu \propto \nu^{-\alpha}$ with spectral slope $\alpha > 1.4$).
- “Either there are X-ray QPOs present (0.1–30 Hz) while the power-law contributes 20% of the total (unabsorbed) flux at 2–20 keV, or the power-law contributes more than 50% of the total flux without detection of QPOs.”
- The SPL/VH state is associated with
 - Episodes of the highest absolute luminosity in many black hole binaries.
 - A distinct γ -ray spectrum.
 - Occurrences of QPOs at both low and high frequency.
- The SPL/VH state is essentially radio-quiet, but the transition to the SPL/VH state may be associated with the ejection of a powerful (“explosive”) transient jet component.

Fig. 16.10 shows the same five sources in Figs. 16.3 and 16.7 when they are in their SPL/VH state. Their photon spectra show a soft thermal component and iron line, in varying degrees of prominence, in addition to the steep power law. In this regard the SPL/VH state seems similar to the much less luminous hard state, but with a much more prominent accretion disk. Also, like the hard state, the SPL/VH state’s fluctuation spectra tend to have $\mathcal{P}_f \propto f^{-\beta}$, with $\beta \approx 1$, but low-frequency QPOs seem to be much more common than in the hard state plots (Fig. 16.3). The presence of a QPO in the SPL/VH state does, indeed, accompany band-limited noise around the QPO frequency. However, with the possible exception of GRS 1915+105, the steepness of the high-frequency noise cutoff, and the flatness of the low-frequency noise lead-in, are not as pronounced as in the band-limited hard state sources. Indeed, in most of the SPL/VHS sources, while a high-frequency cutoff may exist

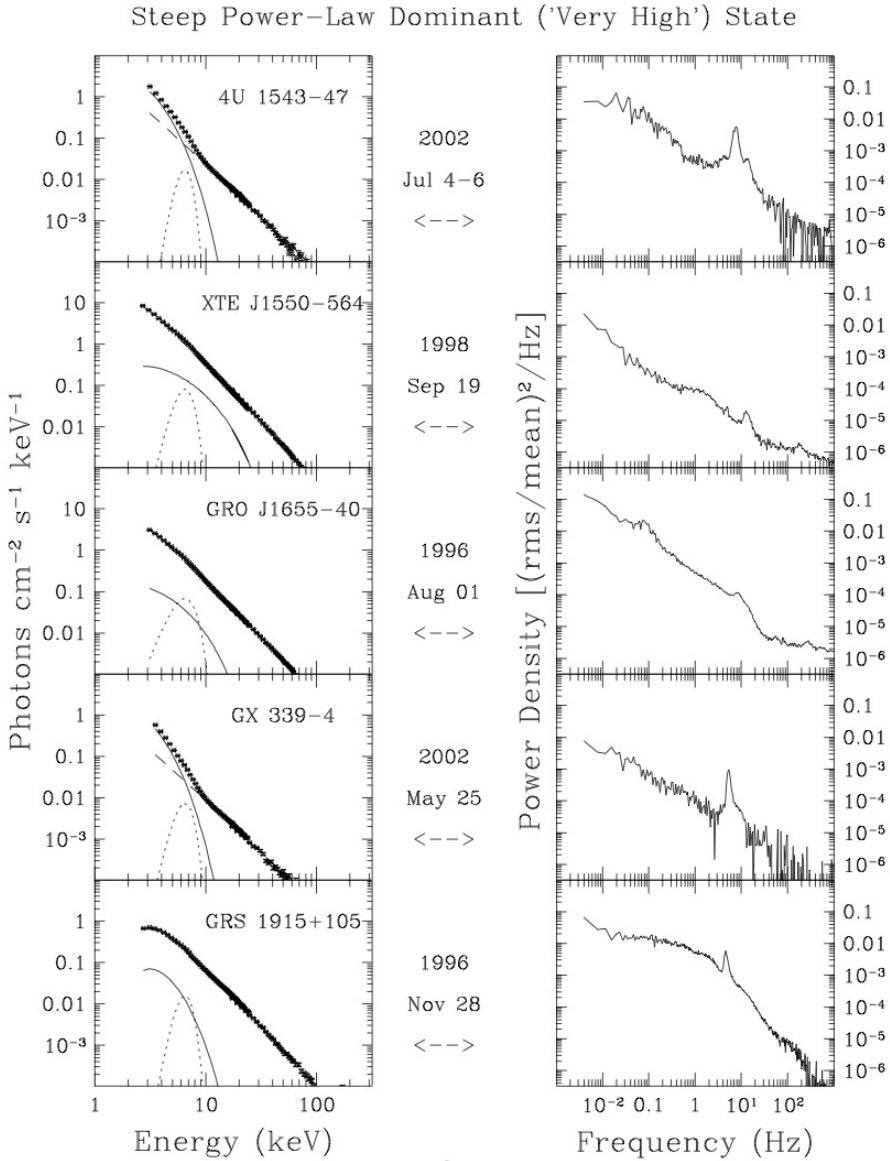


Fig. 16.10: Photon and power spectra for the sources in Figs. 16.3 and 16.7 when they enter the very high/steep power-law state. Most sources have a strong thermal component, an iron line, and a mysterious steep (soft) power-law spectrum. (Compare with the hard power law spectra in Fig. 16.3.) Broad, low-frequency QPOs are now strong and common. These sources are largely radio-quiet, but they do show explosive transient and relativistic jets. Reprinted from Fig. 4.13 of reference [204], with the permission of CUP. Copyright ©2006 Cambridge University Press.

above the QPO frequency, the fluctuation spectrum appears to not only recover, but even flattens out, maintaining measurable fluctuations up to the kHz range!

The physical nature of the SPL/VH state is one of the outstanding problems in stellar-mass black hole research. However, given the discussion of turbulent accretion theory in Section 12.2 and the models of the hard and soft states discussed above, we now can suggest a “straw-man” model that can be used for further, and more detailed, investigation and testing. We suggest that the SPL/VH state may work as follows:

- Because it often is associated with high total luminosity episodes, *we tentatively identify the SPL/VH state as a manifestation of the “slim” disk, radiation-pressure-dominated, accretion flow*, and we attribute the power fluctuations seen in the right column of Fig. 16.10 as being produced in that disk. That is, the SPL/VH state represents moderately super-Eddington accretion. From equations (12.52) and (12.53), this disk should be optically thick to electron scattering but thin to absorption.
- We also tentatively identify the weak thermal component as the cool phase material in the Shakura–Sunyaev “inner” region, whose inner radius is R_{tr} , which moves outward as $\dot{m}$ increases. (This identification is more likely than associating the thermal material with the cooler “middle” region of the disk inflow; the latter would be much cooler than the few keV thermal emission fitted in Fig. 16.10.)
- Like the other power law photon spectra in the hard and soft states, the steep power law emission probably originates in a Comptonizing ADAF-like corona above the inflow. We therefore identify this portion as emanating from the hot material in the upper latitudes of the radiation-pressure-dominated “slim” disk inflow.
- We currently have no strong candidate for the thermal X-ray jet/bipolar outflow that we expect to be ejected from the center of the “slim” disk. Indeed, adiabatic expansion in this outflowing conical wind may cool it to the point where it does not contribute significantly to the X-ray spectrum.

A view of our high $\dot{m}$ SPL/VH state model is shown in Fig. 16.11.⁸ The radiation-pressure-dominated “slim” disk ($\dot{m} > 1$) is a natural choice for the basic physical structure of the SPL/VH state. It explains not only the high luminosities of that state but also the SPL/VH state’s stability. In this regard the “slim” disk is a much better choice than the Shakura–Sunyaev “inner” disk region alone, which is potentially very unstable. (See Section 12.2.) Nevertheless, the “inner” region’s thermal variations (on time scales of ~ 0.05 –4 s for $\dot{m} \sim 1$) might be expected to produce variability in the range of 0.25–20 Hz; broad QPOs in this frequency range indeed are observed in several of these sources. The “slim” disk model also explains why the SPL/VH state has some features similar to those of the hard state. Equations (12.52)

⁸ One slightly strange aspect of the identification of this state as a super-Eddington one is the position of this state in the hardness–intensity (FBG) diagram (Fig. 3.14). Note that the steep power law spectra in Fig. 16.10 are still inherently *harder* than the thermal spectra in Fig. 16.7. So the SPL/VH states lie high and center in the FBG diagram.

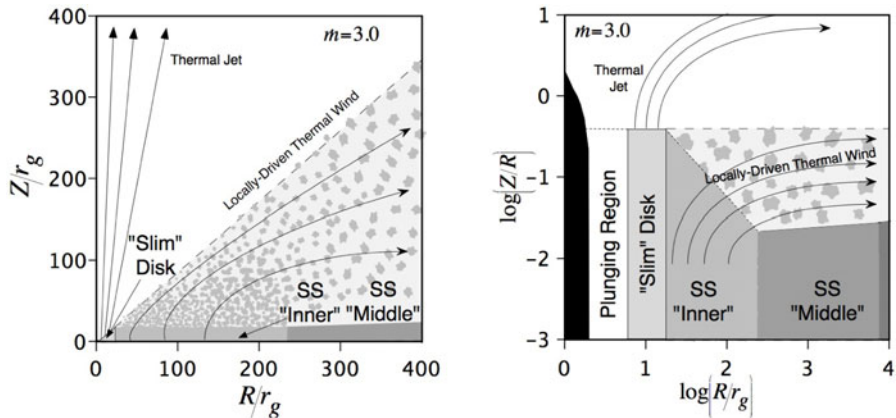


Fig. 16.11: Our model for the SPL state is a moderately super-Eddington one; here the accretion rate is ten times that in Fig. 16.9. Plot is again similar to Fig. 16.4, but with $\dot{m} = 3.0$. The main difference from that previous model is the appearance of a new, *stable*, radiation pressure and electron scattering dominated “Slim” Disk region, whose geometrically thick physical structure is very similar to that of an ADAF but thermal structure is quite different (Section 12.2.2). A large, unstable SS “Inner” region still exists, and may exhibit the Lightman–Eardley instability, as well as continue to emit its locally-driven wind. Like the ADAF, the “Slim” Disk should produce an MHD jet, but these sources are notoriously radio-quiet. One way to reconcile this is to assume that the jet remains thermal, never accelerating particles to highly relativistic speeds. We have drawn in a thermal jet as a suggested solution to this.

show that the basic *physical* structure of “slim” disk flows and ADAFs are *identical*. Only their thermal and luminosity properties are different (equations (12.53) vs. (12.58)). Therefore, their basic fluctuation properties should be rather similar.

Despite appearances, this actually is true for the hard and SPL state power fluctuation spectra. We already have identified the two flat hard-state spectra in Fig. 16.3 as being from an intermediate hard/MDAF state (Section 16.2.4.1). Admittedly, the SPL fluctuation spectra in Fig. 16.10 (with $\beta \sim 1$) are a little shallower than the remaining three true hard state ones and also shallower than the predicted $\beta = 4/3$. But, this discrepancy could be within the errors; or it could mean that each “slim” disk’s accretion rate increases slightly as the flow nears the black hole. As noted above, the broad QPO itself is right in the frequency range where “inner” region thermal variations in modest super-Eddington sources are expected (Table 16.2).

However, the photon spectra in Fig. 16.10 ($\alpha > 1.4$) are nothing like those in the hard state and nothing like what we would predict from having an ADAF-like corona above the “slim” disk ($\alpha = 1/2$, same as in Fig. 16.3). This could mean, of course, that the model is incorrect. However, we also could make the model compatible with the data on SPL sources by assuming that the “slim” disk’s corona is in a mass-losing state, with its accretion rate decreasing with radius as $\dot{M} \propto R^n$. In this case, we would need only a modest $n > 0.45$ in order to obtain the steep, $\alpha > 1.4$ spectra. This mass loss from the corona could occur as the bipolar wind/jet we depict in Fig. 16.11 or, perhaps, as condensation onto the “slim” disk below it.

Finally, given the discussion below of possible super-Eddington wind sources, we might expect that the “slim” disk (at high enough $\dot{m}$) to give way to producing a radiation pressure-driven wind that fully covers the “slim” disk. We have chosen a range of $1 < \dot{m} \lesssim 10$ for the “slim” disk/bipolar wind state, based on the derived accretion rates for sources like P098 below. But, the upper limit of that range is little more than a guess at present.

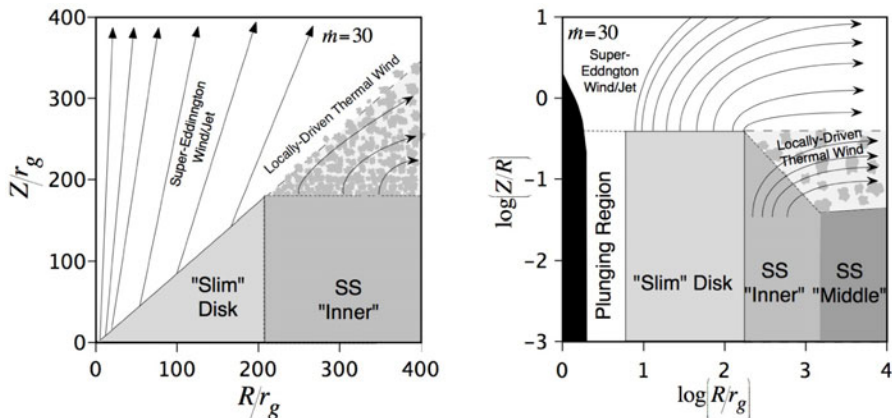


Fig. 16.12: A highly super-Eddington model with ten times the accretion rate in Fig. 16.11. Plot is again similar to Fig. 16.4, but now with $\dot{m} = 30$. In this case the “Slim” Disk part is ten times larger and the super-Eddington outflow is now a broad wind that covers most, if not all, of the central engine; this model is the inflow–outflow realization of the super-Eddington winds discussed in Section 13.1.2 and Figs. 13.1 and 13.3. The SS “Inner” region still can modulate the amount of material feeding the slim disk, which will further modulate the wind outflow rate, as well as its color temperature and opacity. The inner workings of this “sputtering” engine may be hidden inside the large electron-scattering, and photon absorbing, envelope of the wind (see Fig. 16.13).

16.2.4.5 Binary Black Holes in a “Super-Soft State”: Cygnus X-1 on Steroids ($1 \ll \dot{m}$)

The final accretion state that we shall discuss is not recognized as such, in general, because most of the emission lies at or below 0.1 keV. Objects like N1 in M81, P098 in M101, and SS433 show very soft X-ray emission and essentially none above 1 keV [596, 597, 598, 279, 599]. (The thermal jet in SS433, however, *is* visible in X-rays with sensitive enough observations. So, perhaps super-soft sources in other galaxies will reveal X-ray jets as well, if new X-ray telescopes become powerful enough.) These are generally ULX sources with modestly high bolometric luminosity ($\lesssim 10^{39.5} \text{ erg s}^{-1}$), decidedly unlike the hyper-luminous HLX sources ($> 10^{41} \text{ erg s}^{-1}$). They also are distinct from very soft X-ray sources which, nevertheless, also show evidence of a disk and corona: emission up to 10 keV or more,

a 6.4 keV iron line, *etc.* The super-soft state objects have been seen to vary in their color temperatures, while still maintaining essentially the same bolometric luminosity. (See, *e.g.*, Table 2 of [599].)

This latter property is a classic signature of winds in nova explosions: their luminosity remains near the Eddington limit, but their color temperature becomes hotter as the amount of hydrogen being ejected is depleted. We therefore suggest here that the super-soft sources are black holes undergoing highly super-Eddington accretion (*i.e.*, $\dot{m} \gg 1$) and are producing a strong super-Eddington wind outflow – the kind of super-Eddington source first discussed by Shakura and Sunyaev in their famous 1973 paper [351] and studied in detail by D. Meier [463, 455, 458, 459, 460].

Fig. 16.12 shows plots of such an accretion state. The basic underlying accretion flow is the same “slim” disk model as in Fig. 16.11, but now the mass inflow is so great that it produces an essentially spherical outflow from the large trapping radius ($r_i = R_{\text{tr}}$). The outflowing wind is the radiation-pressure-driven counterpart to the gas-pressure-driven ADIOS wind of Blandford and Begelman ([464] and Section 13.2). Mass is lost from the surface of the “slim” disk at most radii, until the inflow begins to approach the black hole. The physics of these super-Eddington winds were discussed in Section 13.1 and a set of relations suitable for analyzing observations of sources in the super-soft state were presented in equations (13.28) to (13.30).

Table 16.4: Models of P098 at two different color temperatures

Parameter	Symbol	Cool spectrum	Warm spectrum
Color temp.	T_*	0.090 keV	0.173 keV
Accretion rate	$\dot{m}$	130	63
Mass-loss rate	$\Delta\dot{M} \approx \dot{M}$	$4.6 \times 10^{21} \text{ g s}^{-1}$	$2.2 \times 10^{21} \text{ g s}^{-1}$
Injection radius	r_i	27, 000 km	13, 000 km
Photosphere	r_*	141, 000 km	46, 000 km
“Sonic” radius	r_c	223, 000 km	107, 000 km
Scattersphere	r_{sc}	$7 \times 10^6 \text{ km}$	$2.2 \times 10^6 \text{ km}$
Wind speed	V_∞	1950 km s^{-1}	2800 km s^{-1}

Table 16.4 shows an analysis of P098 in M101 using these equations at two different epochs that lie about 13 hours apart, one when the source color temperature was 0.090 keV ($T_{*6} = 1.04$) and one when it was 0.173 keV ($T_{*6} = 2.01$). A constant bolometric luminosity of $3.1 \times 10^{39} \text{ erg s}^{-1}$ implies $M_\bullet = 25 M_\odot$. With $\alpha \approx 0.03$, we can derive many essential properties of the massive super-Eddington wind and its relation to the accreting black hole binary.

The derived mass-loss rates, which are within $\sim 1 \dot{M}_{\text{Edd}}$ of the accretion rate, are many tens of Eddington. This implies mass accretion rates of $2\text{--}5 \times 10^{21} \text{ g s}^{-1}$, which can be supplied by an $\sim 8 M_\odot$, Roche-lobe-filling star evolving on a thermal time scale (equation (11.59)). This is a reasonably plausible future scenario for current HMXB black hole binaries (*e.g.*, Cyg X-1) when their companion stars enter the supergiant phase, expand to fill their Roche lobes, and transfer significant amounts of matter to their companion black holes on a thermal time. Fig. 16.13 shows a

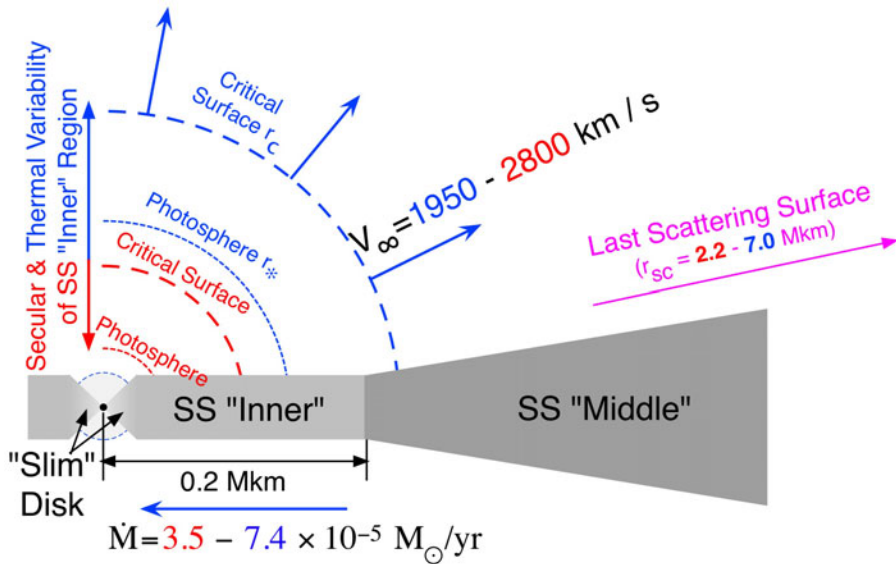


Fig. 16.13: A model for the super-soft X-ray source P098 in M101, using Table 16.4 and eqs. (13.28) to (13.30). Both columns in the table are represented: cool, high $\dot{M}$ state (blue lettering) and warm, lower $\dot{M}$ state (red lettering). Drawing is to linear scale, with the last scattering surface r_{sc} lying 20–30 times the distance of the critical surface r_c . At high $\dot{M}$ even radiation from most of the “inner” region is reprocessed inside the photosphere.

model of the black hole system in P098; the binary companion lies a few tens of million km away from the black hole, well outside the super-Eddington wind.

The injection radius, where the inflow first becomes super-Eddington, is very large ($r_i \approx 730 r_g$). And, as assumed in equations (13.30), the last absorption surface lies inside the wind critical point r_c . The last scattering surface r_{sc} lies at the enormous distance of several million kilometers – several hundred thousand r_g . Nevertheless, this is still probably smaller than a derived binary separation of $r_T \approx 35 \times 10^6$ km, for an $8 M_\odot$ blue supergiant companion with a radius of $R_* \sim 2\text{--}3 \times 10^6$ km. The wind velocities of $2000\text{--}3000 \text{ km s}^{-1}$ are fairly sensitive to the parameter α ($\alpha^{0.3}$). If a means can be devised to measure it, V_∞ could be used to obtain a better estimate of this parameter in very super-Eddington accretion situations.

Such a velocity measurement ($\sim 2000 \text{ km s}^{-1}$) already may have been made for the $\sim 0.2 \text{ keV}$ black hole source in globular cluster RZ 2109 orbiting the galaxy NGC 4472 (see Section 4.2.5 and [278]). Inverting the V_∞ equation (13.30) for α , we obtain

$$\alpha = 0.0065 \left(\frac{V_\infty}{2000 \text{ km s}^{-1}} \right)^{10/3} \left(\frac{M_\bullet}{35 M_\odot} \right)^{-2/5} \left(\frac{T_*}{0.2 \text{ keV}} \right)^{-28/25} \quad (16.36)$$

This is a factor of nearly 5 lower than the $\alpha = 0.03$ we used above. However, $M_{\bullet} = 10\text{--}35 M_{\odot}$ is uncertain by a factor of 3, and T_* is good to only one significant figure. For example, with $M_{\bullet} = 10 M_{\odot}$ and $T_* = 0.15 \text{ keV}$, our estimate would be $\alpha = 0.02$. The main conclusion to draw here is that this estimate of α for a super-Eddington wind is close to what we expect: not extremely small and certainly not greater than unity.

A remaining question is, why the source P098 is so variable? Clearly, the variability is in the wind mass-loss rate and, therefore, the accretion rate at the injection radius (see Table 16.4). A possible source of variability in the accretion rate could be the unstable “inner” region interior to $R_{\text{M-I}}$. The secular time scale there is about 1 day (equation (16.14)), and the range of thermal time scales is $\sim 1\text{--}20 \text{ min}$ (equations (16.15) and (16.20)). This is just the sort of variability seen in P098.

16.3 Intermediate Mass ($10^2\text{--}10^4 M_{\odot}$) Black Holes: What They Might Look Like and How to Find Them

Except for the possibility of the hyper-luminous X-ray sources, no confirmed intermediate mass black hole system has been discovered. Our discussion of how black hole engines in this mass range work, therefore, cannot go much beyond that in Chapter 4. However, we have learned quite a bit about the formation of such objects in Chapter 10. While that discussion is, for the most part, theoretical, it is important for the ongoing search for the elusive IMBHs and also for understanding the genesis of massive and supermassive black holes of considerably larger size.

For reference, the sizes of IMBHs are very similar to those of satellites and planets in the solar system. They are, of course, much more massive – hundreds to many thousands of solar masses in each planet-sized black hole.

16.3.1 Formation of IMBHs

16.3.1.1 Formation of IMBHs in Population I and II Stellar Systems

IMBHs are distinguished from their less massive cousins not only by their masses but also by their exotic means of birth. Fig. 10.1 shows that, for stars with a reasonable amount of heavy elements in their atmospheres (Population I or II), the effects of wind mass loss can be so extreme that even very massive stars ($>100 M_{\odot}$) may not make a black hole greater than a few tens of solar masses in their cores. The only known means of creating an IMBH in Pop I or II systems, therefore, is the Portegies Zwart and McMillan method of direct collapse of a supermassive star in the core of a very compact star cluster (see Section 10.4.2). To briefly summarize this mech-

anism again, if a young star cluster is compact enough (relaxation time less than 25 Myr) and young (age less than 5 Myr), then the stellar collision and coalescence time scale in the cluster core can be so short that a supermassive star will build up in the center before any of the stars merging to form it explode as a supernova. In that case, the supermassive star can collapse directly to a black hole before it can lose an appreciable amount of mass via its own stellar wind. The mass of such a black hole would be of order 0.1% of the cluster mass. Clusters in the range $10^5\text{--}7\text{ M}_\odot$ would produce black holes in the range 100–10,000 $\text{M}_\odot$.

This very attractive mechanism has a number of advantages. It explains why extreme ULX and HLX sources are seen only in galaxies with a high rate of star formation. Young star clusters, and especially the rarer *compact* ones, must form at a higher rate in star-forming galaxies than others, thereby enhancing IMBH production. Furthermore, once formed, the IMBH has plenty of fuel (in the form of gas and stars) in its host cluster, allowing it to shine brightly and even grow. Galaxies like ULIRGs and HyLIRGs must have large numbers of IMBHs and HLX sources in their dust-enshrouded interiors.

The Portegies Zwart and McMillan mechanism also might explain why it has been difficult to find IMBHs in globular clusters. Although the ratio of IMBH-forming young star clusters to the total number of young star clusters is currently very uncertain, it must be rather small, as the former must be very compact to create an IMBH. It is not surprising, therefore, that some globular clusters do not have an IMBH: one would predict that *most* may not, simply from this (admittedly qualitative) discussion. It is important, therefore, to continue the search. Determining this ratio will help to better understand the formation of large star clusters themselves.

It is interesting to mention here the recent discovery of a very massive star (265 $\text{M}_\odot$ named R136a1) in the 30 Doradus star-forming region in the Milky Way's Large Magellanic Cloud [600]. This is clearly a Pop I/II object; it is estimated that its zero-age main sequence (ZAMS) mass was perhaps 320 $\text{M}_\odot$; and its discovery raises many questions in this field. Was R136a1 formed as a single star, or did it coalesce from many O/B stars? Are there other stars in its vicinity with which it might merge into an even more massive object? What will be its final fate? Will it lose most of its current mass and form a stellar-mass black hole? Will it explode as a pair-instability supernova (Section 10.4.1). Or, will it ultimately collapse directly to an IMBH? The answers to these questions will become clear from additional theoretical and observational studies long before the star actually ends its life a million years or so from now.

16.3.1.2 Formation of IMBHs in Population III Stellar Systems

Because the gas that formed the first stars (more than 14 Gyr ago) had essentially *no* heavy elements, the environment for creating IMBHs was considerably richer than it is today. The average star mass was perhaps a hundred times higher than now, making very massive stars of hundreds of solar masses or more fairly common back then (Section 10.3.6). Also, with no heavy elements to enhance the wind mass loss,

such stars could easily proceed directly to the black hole stage, unabated. Black holes formed in two ranges, according to Fig. 10.5: stellar ($3\text{--}50\text{ M}_\odot$) and intermediate (200 to several thousand solar masses). This pristine environment lasted only a relatively short period of time, perhaps only a few hundred million years or so. After that most galaxies' interstellar gas were dirtied with heavy elements, making the production of IMBHs from Pop II, and eventually Pop I, stars much more difficult. However, the Pop III phase must have lasted long enough to produce the seed black holes in the centers of dwarf galaxies that eventually merged into the spheroidal bulges and elliptical galaxies that we observe today.

16.3.2 Accretion-Powered IMBHs in the Local Universe

The story on IMBH engine power is much as it was for stellar-mass black holes, except that the masses of the former are about two orders of magnitude larger. The basic physics was discussed in Section 16.2.2 (equations (16.10) to (16.30)), addressing both X-ray and jet radio luminosities in various accretion states. We therefore can describe most IMBH possibilities without very many new equations.

16.3.2.1 Isolated Intermediate Mass Black Holes Accreting from the Interstellar Medium ($\dot{m} \ll 0.05$)

Stray IMBHs of $\sim 1000\text{ M}_\odot$ may be roaming the interstellar medium in our Galaxy. Their accretion rates still would be very sub-Eddington ($\dot{m} \sim 10^{-7}\text{--}10^{-5}$), with X-ray luminosities of $10^{28\text{--}32}\text{ erg s}^{-1}$ in the hard X-ray band ($\sim 100\text{ keV}$). For a NuSTAR detection limit of $10^{-13}\text{ erg s}^{-1}\text{ cm}^{-2}$, such objects would be detectable only out to $0.24\text{--}2.4\text{ kpc}$. Again, radio detection prospects are more promising: out to $1.5\text{--}37\text{ kpc}$ as a 1 mJy source and $15\text{--}370\text{ kpc}$ as a $10\text{ }\mu\text{Jy}$ source.

The dynamical friction lifetime of $10^{3\text{--}4}\text{ M}_\odot$ objects in our Galaxy's disk is so long (many tens to hundreds of Gyr; equation (11.38)) that if such objects had been formed in the distant past, they still would likely be orbiting at roughly the Galactic radius where they were born. Therefore, while we would expect more IMBHs to lie near the Galactic center ($\sim 7\text{ kpc}$ away), where they would be detectable as 1 mJy sources, nevertheless there may be a few within $1\text{--}2\text{ kpc}$ of the earth that could be detected, and possibly confirmed, both in the radio and hard X-ray.

16.3.2.2 Hard State IMBHs with a Companion ($\dot{m} \lesssim 0.05$)

If an IMBH lies at the center of an old star cluster, it is possible that the former has captured a $1\text{--}2\text{ M}_\odot$ main sequence star as a binary companion. The accretion rate onto the IMBH could rise to $10^{18\text{--}19}\text{ g s}^{-1}$ when that star enters the giant phase, fills its Roche lobe, and transfers matter to the IMBH on a thermal time scale (equation

(11.63)). These rates are still quite sub-Eddington for a $1000 M_{\odot}$ black hole ($\dot{m} \sim 10^{-3}$ – 10^{-2}), but they are now large enough for the $\dot{m}^2$ factor in equation (16.10) to make the source much more visible (in hard X-rays) to NuSTAR than the isolated IMBHs – out to perhaps 0.5–5 Mpc. That is, NuSTAR may detect IMBHs in the hard state in nearby galaxies like M31 (0.6 Mpc) and NGC 5128 (4 Mpc). These objects' radio luminosities ($10^{31-33} \text{ erg s}^{-1}$; equation (16.24)) are now significantly *less* than in the hard X-ray, but still detectable to respectable distances (0.2–1 Mpc as a 1 mJy source or to 2–10 Mpc as a $10 \mu\text{Jy}$ source).

16.3.2.3 IMBHs with a Companion in Higher Accretion States (Soft, Unstable, Super-Eddington)

In order to achieve relative accretion rates in the range $\dot{m} \sim 0.1$ – 1.0 or so, for $10^{3-4} M_{\odot}$ black holes, the actual mass transfer rates need to be in the range $10^{20-22} \text{ g s}^{-1}$. Equation (11.63) shows that the highest of these rates can be supplied only by captured $\sim 10 M_{\odot}$ (early B) stars in thermal mass transfer. The lifetimes of these are only $\sim 30 \text{ Myr}$, which would require a fairly young star cluster of comparable age. Higher accretion rate IMBHs, therefore, should be found in galaxies with fairly high star (and compact star cluster) formation rates. This is precisely the kind of galaxy in which we find ULX sources. Because the range of accretion rate between the hard plateau and soft power law states is only a factor of ~ 20 , *it is quite possible that IMBHs with mass-transferring B star companions may, like GRS 1915+105, also display hard, soft, unstable, and SPL states at different times.*

Soft State IMBHs with a Companion ($0.05 < \dot{m} < 0.08$)

As the black hole mass increases toward $10^5 M_{\odot}$ the range of $\dot{m}$ in which a stable soft state can exist decreases and then disappears (see the discussion after equation (16.27)). If the $\dot{m}_A \sim 0.05$ estimate in Section 16.2.2 is correct, this range can be at most $\dot{m} \sim 0.05$ – 0.1 for IMBHs. That is, $\sim 10^{20-21} \text{ g s}^{-1}$ would be needed to push an IMBH into the soft state.

Soft state IMBHs should have many properties that we often find in ULX (but not HLX) sources:

- Accreting at $\dot{m} \sim 0.1$ Eddington.
- High luminosity ($\sim 10^{40} \text{ erg s}^{-1}$; equation (16.12)).
- Very soft thermal spectra ($\sim 0.1 \text{ keV}$; equation (16.13)).
- But still display the spectral features of an ADAF-like corona with a power law spectrum out to 100 keV or more, an iron line, and Compton reflection; see Fig. 16.7.
- Lack of variability on disk thermal and secular time scales.
- A weak, suppressed radio jet ($L_R \sim 10^{32-33} \text{ erg s}^{-1}$).

So, soft ULX sources with a hard power law tail up to 100 keV or so are good candidates for IMBHs in the stable soft state.

IMBHs in the Unstable State ($0.08 < \dot{m} < 1$)

With the unstable accretion rate range widening (a bit) with increasing black hole mass, it is possible that the set of unstable states available to intermediate mass and large black holes is even more rich than it is for the $14 M_{\odot}$ black hole in GRS 1915+105. One can only speculate right now, of course. However, if an IMBH can be fed at rates approaching Eddington ($\dot{m} \sim 0.1$ – 1.0), the accretion flow indeed may be subject to secular instabilities with time scales a few days (equations (16.14)) and have thermal instabilities in the range of a few seconds to a few minutes (equations (16.15) and (16.16)). This variability would distinguish such sources from the soft state IMBHs, which should be steady and somewhat rare.

The longer (few day) secular variations may make significant excursions in the hardness–intensity diagram (HID; analog of the FBG diagram) and may display radio outbursts on the order of $L_R \sim 10^{34-36} \text{ erg s}^{-1}$. That is, one would likely expect a “jet line” in the HID just as there is one for binary black holes. When comparing with the hard X-ray band, however, *one must use the flux in the 0.1–0.5 keV band to form the hardness ratio* instead of the 2–10 keV flux. The thermal disk will be nearly a factor of ten cooler in IMBHs than in normal binary black hole X-ray sources.

IMBHs in the Steep Power Law State

IMBHs in a moderately super-Eddington state need at least $10^{21-22} \text{ g s}^{-1}$ and could be as bright as $10^{41} \text{ erg s}^{-1}$ in the soft X-ray (0.2–0.3 keV), depending on black hole mass. This would classify the object as an HLX, and we again expect these to occur in star-forming galaxies.

Continuing to adopt our early suggestion that the soft power law spectrum in the VH state is a condensing ADAF-like corona above a cooler “slim” disk, we suggest that the X-ray photon spectrum of the SPL state in IMBHs should appear similar to that of binary black holes. There even should be 6.4 keV and Compton reflection components. However, the thermal spectrum component will be missing – or, more accurately, will be shifted lower in energy by a factor of $(1000/10)^{1/4} \sim 3$ from $\sim 1 \text{ keV}$ to $\sim 0.3 \text{ keV}$ or less (equation (16.18) with $m = 1000$ and $\dot{m} = 3$). The power fluctuation spectra also should look similar, but with any prominent QPO shifted lower in frequency f by a factor of $\sim (1000/10) = 100$ (equation (16.20)). We also note that a thermal bipolar outflow also may be present, but that outflow will have a small covering factor for Eddington or slightly super-Eddington accretion rates.

IMBHs in a Super-Eddington Wind State

The accretion rates required for this kind of state to occur are enormous: $10^{23-24} \text{ g s}^{-1}$. In principle, this could be supplied by a 20–40 $M_\odot$ star in thermal mass transfer to its IMBH companion. More likely, however, such accretion rates would occur during the IMBH growth phase, after coalescence and collapse of the central supermassive star to an IMBH. Additional O/B stars that did not participate in the coalescence still could be captured by the IMBH and feed it at this rate for a few tens of thousand of years.

IMBH sources in this super-soft state would have emission *only* in the soft X-ray ($\lesssim 0.1 \text{ keV}$ or $\lesssim 10^6 \text{ K}$; equation (16.19)), and much of this bolometric luminosity may be absorbed in their host galaxies' ISM. There also would be no X-ray power law emission (hard or steep), because of the Compton thick and/or X-ray-absorbing super-Eddington wind. But there might be a Rayleigh–Jeans ($\propto \nu^2$) spectrum in the optical–UV, modified by electron scattering. However, one should *not* expect to see a shallow $\nu^{1/3}$ multi-color black body spectrum, since the accretion disk is hidden deep inside this optically thick outflow.

As with sources like P098 in M101, the color temperature may rise and fall while maintaining an approximately constant bolometric luminosity of $10^{41-42} \text{ erg s}^{-1}$; but the maximum temperature in this state would be that of a bare “slim” disk at $\dot{m} \sim 1$ (0.2 keV). The secular variations might occur on time scales of a few months to a few years, with thermal disk variations of a fraction of an hour to a few days also affecting the wind outflow rate (equations (16.14), (16.15), and (16.16) with $M \sim 10^{3-4} M_\odot$ and $\dot{m} \sim 30$).

16.4 Massive (10^5 – $10^7 M_\odot$) Black Holes in Spiral Bulges, Pseudobulges, and Dwarf Ellipticals

Traditionally, black holes with masses more than a few tens of thousands of solar masses, all the way up to ten billion solar masses (the largest known), have been lumped into a single category: the supermassive black holes (see Chapter 2) – and with good reason:

1. **Location:** they all seem to exist at the centers of galaxies, and up to the present time, appear to be the only such massive object in that galaxy.
2. **Origin:** most fit on a single $M_\bullet$ – σ relation (Fig. 2.18(right)), implying that their origin and growth processes must be very similar.
3. **Accretion States:** as the accretion rate rises, all black holes in the $M_\bullet > 10^5 M_\odot$ range should develop an unstable Shakura–Sunyaev “inner” region before the ADAF disappears; *i.e.*, $\dot{m}_1 < \dot{m}_A$. This means that there will be no easily detectable SS “middle” region and, therefore, *such black holes should display no stable soft state*. The low-luminosity objects (*e.g.*, LLAGN) will appear ADAF-like hard state sources, while the high-luminosity ones (Seyferts, quasars) will

be variable, unstable state objects in the optical–UV and produce occasional or periodic jet outbursts of varying strength.

However, there also are distinct differences between black holes in the $10^{5-7} M_{\odot}$ and $10^{8-10} M_{\odot}$ ranges:

1. **Morphological:** the million solar mass black holes occur in objects which traditionally have been referred to as “AGN” for many years: Seyfert galaxies (active spirals) or active dwarf ellipticals (LLAGN). The billion solar mass holes, on the other hand, are associated with giant elliptical galaxies: N-galaxies and quasar hosts. Ellipticals provide a very different feeding environment for their central black holes: rapid feeding and growth during galaxy mergers, and then much less activity between these events. Spirals, however, are galaxies that have *not* undergone major mergers in a very long time. But they do have much gas in their disks that can feed a central hole regularly at a relatively smaller rate.
2. **Radio Luminosity:** the relative radio luminosities (L_R/L_{Edd}) of the $10^{6\pm1} M_{\odot}$ objects also are rather low. If AGN and LLAGN ever were strong radio sources, they must have been so in the distant past, not recently. On the other hand, the $10^{9\pm1} M_{\odot}$ holes have a much higher occurrence of radio jets and lobes. This higher rate of jet-production activity may be related to the higher rate of merger activity in their giant elliptical hosts.
3. **Fueling:** $10^8 M_{\odot}$ represents a distinct change in how active objects can be fed and still produce luminous energy. Below this mass the black hole can tidally disrupt main sequence stars (see equation (11.26)). Black holes with $M_{\bullet} \gtrsim 10^8 M_{\odot}$, however, swallow most stars whole. This would contribute to a *growth* in the black hole mass, but not to the fueling of the central engine. In order to create the high luminosities seen in quasars, the fuel must either come from red giants or supergiants or, perhaps more likely, from gas brought into the galaxy center during a merger, *before* it is turned into stars.

We, therefore, will adopt terminology that has recently come into use by the gravitational wave community: black holes in the range $10^{5-7} M_{\odot}$ will be called “massive” and distinguished as a separate class. On the other hand, for those in the range $10^{8-10} M_{\odot}$ we will retain the title “supermassive” here. This chapter section will discuss mainly the massive black hole class, but also will deal issues that are common between the MBH and SMBH classes.

Compared to SBHs and IMBHs, the massive black holes are large astronomical objects in their own right, ranging in radius from a few Jupiter radii to larger than the radius of Mercury’s orbit, *i.e.* spanning the size range of most stars (brown dwarfs to giants). But these black holes contain as much mass as an entire globular cluster of stars or a dwarf galaxy! The nearest MBHs are the one at the center of our Galaxy (8 kpc away, $4.1 \times 10^6 M_{\odot}$ in mass, and in the middle of the MBH mass range) and the one at the center of the Andromeda galaxy (800 kpc distant, $3 \times 10^7 M_{\odot}$ in mass, and at the upper end of the range).

16.4.1 *A Proposed Model for the Central Engines of Seyfert 1s and Quasars, from First Principles*

Before beginning the standard discussion of black hole formation and then accretion, it will be useful to discuss how both the theoretical work discussed in Chapters 12 and 13 and the application of that to binary black holes (Section 16.2) can be applied to massive and supermassive objects. For our example model we will use the black hole mass of $10^{7.5} M_\odot$ – on the border between massive and supermassive objects and about the black hole mass of the famous radio-loud Seyfert 3C 120.

16.4.1.1 Motivation and Constraints

For many years astrophysicists have been struggling with the problem of reconciling standard black hole accretion theory with the basic structure of the nuclei of Seyfert galaxies and QSOs that has been deduced from observations of these objects. To be sure, much progress has been made at distances of, say, 1–100 pc from the central black hole, with very plausible physical pictures for the obscuration in Seyfert 2s (a dusty torus) and the narrow line region (NLR, an ionization cone above the torus); see Section 2.3. And reverberation studies have located the broad line region (BLR) very close in: only $10^{3-4} r_g$ from the black hole (see, *e.g.*, [601]).

However, many unanswered questions remain, such as where does the BLR lie in the accretion flow? Why is it broken up into small clouds? What causes the broad line widths (rotation, turbulence, infall, outflow)? What is the physical nature of the central engine itself, where in the accretion flow does it lie, and why is it always interior to the BLR? In short, *what are the fundamental properties of accretion inflows that favor the ubiquitous generation of BLRs and central ionizing sources?*

Such a model not only would have to answer these questions and be otherwise consistent with the data, it also would have to be compatible with basic accretion theory. In the latter we include ADAFs, MDAFs, Shakura and Sunyaev disks, “slim” disks, and super-Eddington winds – plus any thermal and secular instabilities predicted by this theory (see Section 16.2.2). Finally, it is desirable for this model also to be compatible with, and perhaps even utilize, the knowledge that we have gained from the detailed studies of binary black holes.

Below we suggest such a scenario. It has many advantages, including good phenomenological, structural, and variability properties. It also satisfies the above constraints. However, it currently is offered only as a straw man – to be tested with more detailed studies than we have done here.

16.4.1.2 General Considerations

First, based on the broad optical–UV properties of AGN central engines, we can rule out both low/hard states (ADAF, MDAF) and the super-Eddington wind state

as candidates for our AGN central engine model. Little optical–UV radiation is expected in the hard states. And a strong, cool outflowing and Compton thick wind in the super-soft state is incompatible with our ability to see X-ray emission from most AGN. This leaves only the unstable and “slim” disk states as candidate models. The accretion rates of AGN with BLRs, therefore, must lie in the range $\dot{m}_I < \dot{m} \lesssim 10$, with the upper limit only a guess right now (Section 16.2.4). The value of the Eddington ratio for AGN, therefore, should lie in the range $\dot{m}_I < \ell_{\text{rad}} \lesssim 1$. (Note: even “slim” disks and super-Eddington winds radiate at $\ell_{\text{rad}} \sim 1$.) This conclusion is compatible with Boroson’s principal component plot (Fig. 2.20), and also with a recent analysis of AGN by Moshe Elitzur of the University of Kentucky and Luis Ho of Carnegie Observatories in Pasadena, California [602], which shows that ℓ_{rad} for AGN with BLRs lies in the range ~ 0.01 –1.⁹

Second, while the mechanism that produces a central ionizing source and a BLR may operate in both the unstable state ($\dot{m}_I < \dot{m} < 1$) and in the “slim” disk state ($1 < \dot{m} \lesssim 10$), most AGN appear to radiate in the range $\ell_{\text{rad}} < 1$ or even $\ell_{\text{rad}} < 0.1$. The super-Eddington “slim” disk objects, even if some can be identified, must be a relatively minor component of all AGN and must lie near $\ell_{\text{rad}} \approx 1$. *We therefore conclude that the standard Seyfert/QSO nucleus must be in an unstable state, with $0.01 < \dot{m} < 1$.*

16.4.1.3 The Broad Line Clouds

Taking $\dot{m} = 0.3$ and $M_\bullet = 10^{7.5} M_\odot$ as our canonical unstable state model of a (borderline) Seyfert/QSO, the structure of the accretion flow will appear similar to that in Fig. 16.9. At these accretion rates, there is a substantial secularly and thermally unstable Shakura–Sunyaev “inner” region. The secular time scale in this case is *long* – 11,000 years. So, a human observer would not notice that the source may be exhibiting limit cycle behavior on this time scale. Instead, it would appear to be “permanently” stuck in what really is a temporary substate along the hysteresis loop in the hardness–intensity diagram; this substate could be a soft (*i.e.*, 0.01 keV) state or a hard state or somewhere near the jet line (see Fig. 3.14). Thermal time scales are far shorter (a few days to nearly a year); variability in the optical–UV and X-rays on those time scales should be very noticeable.

If the “inner” region is thermally unstable, it is likely to be a two-phase medium; but the spatial distribution of the two phases may be very different. For example, the equator likely will contain more material in the cool phase in a disk-like structure, with hot bubbles rising out of the disk. The high latitudes should contain more of the hot, optically thin ADAF-like material. The cool phase of the Shakura–Sunyaev

⁹ These authors argue that is the onset of the hard state that quenches the BLR activity. Indeed, if the UV ionizing source disappeared, the BLR probably would also. However, equation (12.34) shows that the unstable “inner” region also disappears for $\dot{m} < \dot{m}_I \approx 0.027$ for $M_\bullet = 10^{7.5} M_\odot$. At the present time, it is difficult to determine whether the BLR activity disappears because $\dot{m}$ goes below $\dot{m}_A$ or below $\dot{m}_I$.

“inner” region, therefore, may be a good source of cool clouds for the BLR and the hot phase a good candidate for the cloud confining medium.

However, *the “inner” region itself cannot be the BLR*; the Keplerian speeds in this region are very high: $0.08\text{--}0.2\ c$ or more. Typical BLR full-width linewidths at half-max are $0.02\ c$ or less. So, if we wish to use the thermally unstable “inner” region as the source of the BLR material, we need to transport that material out to BLR distances (to a few light-days, say $1000\ r_g$) in an outflowing wind. We also need to slow down any cloud rotational and turbulent motions, and do this quickly before the clouds dissipate.

One mechanism for accomplishing this outflow is the radiation-driven wind discussed in Section 2.3.4. A problem with this approach, however, is that it would accelerate the cool clouds with a much larger force multiplier (Section 13.1.1) than the highly ionized ($\sim 10^9\ \text{K}$) coronal gas – which we believe to be present in this model. This would quickly ablate and dissipate the clouds.

Instead, we suggest here that the outflow is a *thermally*-driven wind – essentially an ADIOS one (Section 13.2) – in which the confined cool clouds simply go along for the ride.¹⁰ Our estimates of the final velocity components in the BLR are the following:

- **Rotational:** $V_{\phi\text{M-I}} \sim 0.08\ c$ at $R_{\text{M-I}} = 160\ r_g$. Conservation of angular momentum ($V_{\phi} \propto 1/R$) would reduce this to well below the BLR linewidths of $\sim 0.02\ c$ if $R_{\text{BLR}} > 1.1\ \text{lt-day}$. So, $V_{\phi,\text{BLR}}$ is probably unimportant.
- **Radial:** $V_{\infty,\text{wind}} \approx 0.4\ V_{\phi\text{M-I}} = 0.03\ c$ (Section 13.2). So, outflow velocities should be crucial for determining the broad line widths.
- **Turbulent:** at $R_{\text{M-I}}$ the turbulent velocity is of order $\alpha^{1/2} c_s \approx 0.55\ V_{\phi}$ (for $H/R \sim 1$). Adiabatic expansion ($V_{\text{turb}} \propto R^{-1/2}$) of the turbulent flow would reduce this velocity to $V_{\text{turb},\text{BLR}} \sim 0.02\ c$ (of order the BLR linewidths).

So, if the BLR is at a distance of only $\sim 1000\ r_g$ (2 lt-days), then both radial outflow and turbulent velocities may be important for determining BLR linewidths. However, if $R_{\text{BLR}} \gtrsim 20\ \text{lt-days}$, then even turbulence will be unimportant, leaving only outflow as the prime linewidth candidate. In any case, at outflow rates of $0.03\ c$, the two-phased medium could reach BLR distances as far as 10 lt-days or so in a typical thermal time. In addition, as the outflow expands, the local thermal time scale could increase significantly, as the flow no longer would be subject to the viscous heating/cooling processes associated with shearing accretion disks. The outflowing medium, therefore, could preserve its thermal state even well beyond the BLR distance.

It is worth noting here that this scenario is only partially compatible with the MHD shearing box simulations of radiation-pressure-dominated accretion flow (Section 12.4.3). Those simulations do, indeed, become stratified, with a dense material on the equator and a more tenuous plasma in the upper layers of the disk, where much of the energy may be dissipated. The simulations also confirm Shakura and Sunyaev’s assumption that the stresses transporting angular momentum in this

¹⁰ That is, in this scenario the hot phase material (coronal outflow) makes up the main outflow, while the cool BLR clouds are the hitchhikers in the flow.

region are indeed proportional to the radiation pressure (*i.e.*, $t_{\phi R} \propto p_r$) in a *time-averaged sense*. This makes our use of a secular (Lightman–Eardley) instability quite reasonable [454].

However, our use of the Shakura–Sunyaev–Pringle *thermal* instability is much more controversial. The time scale over which the stresses must be averaged (to obtain the αp_r relation) is considerably longer than a thermal time; this effect conspires to make the simulated radiation dominated disks appear thermally stable [452]. If this result holds up as these semi-local simulations become more sophisticated, fully three-dimensional and global, it would be a serious setback for our understanding of black hole accreting sources. There are many variability phenomena that occur on precisely what we calculate to be the thermal time scale in stellar mass, as well as supermassive, black hole systems: flickering in sources like Cyg X-1, the broad QPOs in SPL sources, the short-term variability in accretion rate of super-soft sources like P098, the rich amount of variability exhibited by QSOs in the optical, and the periodic ejection of jets in compact radio sources like 3C 120 and 3C 273 (not to mention the nice properties of two-phase, thermally-driven, ADIOS-like winds for explaining AGN broad line regions). It may be possible that graduating to full three-dimensional radiation dominated simulations will show some new (thermo-viscous?) effect that occurs on a thermal time scale. Or it may be possible that the microphysics of transferring heat amongst electrons, ions, photons, *etc.* needs to be handled more carefully in MRI simulations. If not, most of the thermal instability explanations discussed in this book will be wrong, and we will have to go back to the drawing board.

16.4.1.4 The Ultra-Fast Outflows (UFOs)

The recent discovery that a large fraction of AGN and QSOs ($\gtrsim 35\%$) have broad absorption in their iron $K\alpha$ line near 6 keV [603, 604] indicates the presence of ultra-fast outflows ($0.04\text{--}0.15\ c$) in many, if not most, AGN. With an estimated covering factor of $0.4\text{--}0.6$, if these winds occur in the equatorial regions, their opening half-angle would be $53^\circ\text{--}66^\circ$ (*i.e.*, $H/R \sim 1.7$) – consistent with an ADAF/ADIOS outflow. We suggest here that these are the same two-phase, thermally-driven outflows that we suggested above were responsible for the BLR. The only difference here is that the ultra-fast material originates from the inside portion of the unstable Shakura–Sunyaev “inner” region ($R = 0.4^2\ GM_\bullet/v^2 \approx 7\text{--}100\ r_g$), while the BLR clouds originate from the “inner” region’s outside portion ($100\text{--}160\ r_g$; approximate model numbers given for $M_\bullet = 10^{7.5}$ and $\dot{m} = 0.3$).

It is interesting to estimate the effects of adiabatic expansion on both types of clouds (UFO *vs.* BLR). If we assume that the cool part of the two-phase medium is similar to that of a Shakura–Sunyaev cool “middle” region disk (equations (12.31)), but at the higher unstable state value of $\dot{m} = 0.3$, then an upper limit on the temperature of the BLR cloud material at the $100\text{--}160\ r_g$ disk source is $\sim (0.4\text{--}0.7) \times 10^6\ \text{K}$. Adiabatic expansion of this non-relativistic baryon gas to $R \gtrsim 1000\ r_g$ would go as $T \propto R^{-2}$, giving $T_{BLR, \text{cloud}} \lesssim (7\text{--}12) \times 10^3\ \text{K}$. Cloud temperatures deduced from

photoionization models are $\sim 10^4$ K, so the thermal state of the BLR clouds will be partially, or completely, dominated by the radiation field of the central ionizing source (as is commonly assumed).

The UFO clouds' thermal state is most certainly dominated by that central ionizing source's radiation field. A similar temperature upper limit begins with material of $\sim (0.7-8) \times 10^6$ K at $(7-100) r_g$ in the accretion disk, but adiabatic expansion out to $R > 0.01$ pc ($6500 r_g$) will reduce this to only $(8-160)$ K. So, for the UFO clouds, their residual thermal content from the original accretion disk material will be negligible. Unless the central source can keep these clouds ionized, however, they may be seen only in absorption, especially in X-ray lines.

16.4.1.5 The Central Ionizing Source

Since the coronal values of $T_e \approx 10^9$ K and $\alpha \approx 0.3$ are probably very similar among all AGN inflows, the only parameter that really determines the corona's Compton depth ($y = 4\tau_{\text{es}}^2 kT/m_e c^2$; equation (12.46)) is the accretion rate $\dot{m}$. For our $\dot{m} = 0.3$ model the accretion rate in the corona alone would be less than 0.3, yielding a Compton parameter of $y < 10$. In this low y case Sunyaev and Titarchuk (see Section 9.4.3 and [356]) showed that cool photons (such as those in the cool disk at ~ 13 eV) will be Comptonized into a power law photon spectrum that stretches from that energy toward the temperature of the Comptonizing medium (here $T_e \sim 100$ keV).

While the slope of this power law UV spectrum will be sensitive to $\dot{m}$, it likely will be similar amongst all AGN with similar Eddington ratios. So, to first order, the scenario we have outlined here produces a central ionizing source that contains much of the optical-UV luminosity and has a power law spectrum. The size of this source will be only somewhat larger than r_{isco} : $< 20 r_g$ or $< 10^{14}$ cm for our $10^{7.5} M_{\odot}$ example. This is much smaller than the $160 r_g$ SS "inner" region, which is, in turn, smaller than the several light-day BLR.

Note that, in this model, the hard X-rays are *not* produced by the Comptonization of the disk UV photons. As in Section 16.2.4, we take the power law X-ray spectra to be produced by Bremsstrahlung and thermal synchrotron in the corona itself [441]. The many emitting coronal radii then produce a multi-colored ADAF spectrum to obtain the power law. Therefore, even if an AGN goes into a low state, and the thin accretion disk disappears, the hard X-ray emission nevertheless will remain.

16.4.1.6 Modifications to Include Narrow Line Seyfert 1s

In order to accommodate the Eddington-radiating NLSy1s in the model, we suggest here that these sources are in a slightly super-Eddington "slim" disk accretion state ($1 < \dot{m} \lesssim 10$). This higher-luminosity structure differs a bit from the unstable state, mainly in having a stable, advective, radiation-pressure-dominated inflow interior to $R_{\text{tr}} = \dot{m} r_{\text{isco}}$ and a possible radiation pressure-driven bipolar outflow above

the “slim” disk. Outside the trapping radius lies the unstable “inner” region, the expanding two-phase corona, *etc.* The other main difference is that the outer radius of the thermally unstable region is a factor of 6 larger ($10^3 r_g$ for $\dot{m} = 3$ and $M_\bullet = 10^{7.5} M_\odot$). So, while the properties of the BLR emission lines certainly will depend on the photoionization process and the radiative details, it is noteworthy that the overall velocity scale will be lower by a factor of $6^{-1/2} = 0.4$. For example, an accreting black hole that produces $\sim 5000 \text{ km s}^{-1}$ BLR linewidths at $\dot{m} = 0.3$, the same system will produce $\sim 2000 \text{ km s}^{-1}$ lines at $\dot{m} = 3$. The “slim” disk scenario for the NLSy1s, therefore, accounts for their typical Seyfert 1 properties, as well as the NLSy1 higher-luminosities and smaller line widths. It also predicts NLSy1s will have longer variability times than regular Seyferts with similar black hole mass.

16.4.1.7 Modifications to Include Broad Absorption Line (BAL) AGN

In this author’s opinion the physical nature of the BAL AGN is one of the great unsolved problems of black hole astrophysics. We will not solve it here, but we do have a suggested scenario that is consistent with standard accretion theory and with the above overall AGN model. The main properties of BAL AGN were outlined in Section 2.3.4. They have high-velocity blue-shifted absorption troughs in *optical–UV lines*, indicating ultra-fast velocities of $0.1 c$ or more. Sometimes the trough is detached from the BLR emission line, indicating material outflowing in a shell or clump with only a small range of approaching velocities. Not surprisingly, these objects are weaker in observed X-rays than other AGN of similar luminosity. Whether that is intrinsic (*i.e.*, seen at all viewing angles), or due simply to the intervening absorbing material, its not clear. The X-ray weakness turns out not to be a critical defining property of the class.

But, surprisingly, these objects also are statistically much weaker in the radio than the general population of QSOs/QSRs. Few, if any, FR II radio quasars, for example, are BAL QSOs. This cannot be accomplished with absorption by BAL clouds in the inner parsec: FR II lobes extend outside the host galaxy to 100 kpc or more. While most known BAL QSOs have a supermassive black hole (see Fig. 2.20), a few lower-mass BAL objects have been found, including some (such as WPVS 007) that have changed from a NLSy1 to a BAL AGN in recent years, and possibly back again [605, 606, 607]. One potentially important result is that the covering factor of BAL clouds seems to be rather small (0.10–0.20; [124]).

BAL objects, therefore, are AGN with strong and fast outflows. In a phenomenological sense, their physical nature can be only one of two classes: either (1) they are otherwise normal AGN, simply viewed in a unique manner, or (2) they are a new kind of AGN state, and their relatively small numbers reflect the probability that all AGN may enter this new state occasionally.

AGN Viewed Favorably?

The fact that accretion disk models predict different kinds of winds¹¹, and that at least 35% of bright AGN show ultra-fast outflows (in the X-ray iron lines) makes very attractive the idea that BAL objects are physically like other AGN. The UFO winds are, indeed, likely to be disk winds, and wind speeds are even similar to BAL outflows. Furthermore, the idea of viewing the object along the disk surface in order to see the outflowing clouds in absorption also fits well with the small covering factor measurements. There is a need, of course, to explain why many AGN show X-ray absorption lines but only a few (the BAL objects) show absorption in the optical–UV.

The real problem with this model, however, is the radio-weak nature of BAL AGN and the serious of FR II BAL QSOs. *BAL AGN simply cannot be just like other AGN.* There must be some property related to their “BAL-ness” that suppresses jet production. Simply viewing an AGN from a favorable direction will not suppress the viewing of isotropically radiating radio hot spots and lobes.

Near-Eddington AGN?

However, we do not need to adopt a radically different model in order to incorporate the rather radio-quiet nature of BAL objects. We simply can use what we have learned about binary black holes in high-luminosity states to suggest that BAL AGN are radio-weaker simply because they are accreting near or at the Eddington value. As discussed above (Sections 12.2.2, 13.1.2, and 16.2.2), while it is not fully understood yet, black holes in a high-luminosity state ($\dot{m} > 1$) do not seem to be able to produce a strong radio jet. Indeed, in the Boroson principal component diagram (Fig. 2.20) BAL QSOs lie near $L = L_{\text{Edd}}$. So a model in which BAL QSOs are simply high $\dot{m}$ AGN is certainly more attractive than one in which they are simply normal sub-Eddington AGN that are viewed in a special way. In this second possible scenario, BAL QSOs are, instead, near-Eddington ($\dot{m} \sim 1$) AGN that are again viewed along the surface of the disk.

Some issues remain. One needs to explain why near-Eddington luminosities generate BAL clouds that produce attached and detached absorption troughs. That is, now that we have put BAL objects into a higher-luminosity class, why do these very high-luminosity AGN produce BAL-type clouds while sub-Eddington ones do not?

Super-Eddington Outflows Viewed Favorably

A more radical approach, and the one recommended here, is to take the accretion rate to be moderately *above* Eddington. That is, we also could put BAL AGN into

¹¹ Above we argued for a thermally-driven, ADIOS-type wind, rather than a radiative line-driven wind. However, at the current level of discussion, this is a technicality and not really important at this point.

the “slim” disk super-Eddington state, just like we suggested above for the NLSy1s. What, then, is the difference between a BAL object and a NLSy1? In this case it *is*, again, the viewing angle (and, to a lesser extent perhaps, the accretion rate). But in this model the angle for viewing the object as a BAL is *axial*, not equatorial. And the BAL outflow is now a distinctly different one from the disk wind in normal AGN: it is the thermal bipolar super-Eddington wind/jet where that wind has not yet spherized (see Figs. 13.3(left) and 16.11).

This solves a number of problems. It identifies even moderately super-Eddington states with an outflow, which we have expected all along. It also respects the low covering factor measured for many BAL objects. That is, with an unspherized, bipolar super-Eddington wind the BLR at $R \sim 1000 r_g$ remains uncovered unless one views it near the bipolar axis through the approaching wind. Since this requires a super-Eddington $\dot{m} > 1$, the Shakura–Sunyaev “inner” region is expected to be larger in radius than in our standard AGN model ($\dot{m} = 0.3$). The emission lines in the BLR, therefore, will be correspondingly narrower *intrinsically*, just like those in NLSy1s. This scenario also explains why some NLSy1s go in and out of a BAL-like state. Precession of the bipolar outflow, or more likely an increase or decrease in $\dot{m}$, would bring the outflow into our line of sight or away from it, turning a NLSy1 into a BAL AGN or *vice versa*.

One also might expect to see some *rare* BAL AGN with a *high* covering factor. These would be objects with a fully spherical wind ($\dot{m} > 10$?) just like the super-soft X-ray sources of stellar mass in Section 16.2.4.

Of course, there are several issues to be solved. But the high accretion rates required ($7\text{--}700 M_\odot \text{ yr}^{-1}$) for the supermassive objects is not one of these. We already observe ULIRGs and HyLIRGs with apparent accretion rates in the middle of this range, and we are fairly confident that supermassive black holes that obey the $M_\bullet\text{--}\sigma$ relation have experienced accretion rates higher than this shortly before both black hole growth and bulge star formation were shut off by a super-Eddington wind. So, nature does achieve these absurdly high accretion rates *and* efficiently turns them into radiative luminosity, even if we do not yet know how. Of course, it may be that by the time the AGN is optically visible the spherized wind may have largely scaled back into a bipolar flow. But eventually high-resolution infrared imaging should become as good as optical imaging and be able to show us this very high outflow state in detail even if it is still obscured in the optical.

One question that remains in this super-Eddington wind scenario for BAL AGN include is how do the BAL clouds form in the bipolar outflow? This is very similar to the question of how SS433 forms cool clouds in its thermal jet, and work done on this latter stellar source may be applicable to BAL AGN. Also, are there any high covering factor BALs visible in the optical? How does the bipolar outflow opening angle vary with $\dot{m}$? At what $\dot{m}$ does the wind become essentially spherical (a problem for all mass ranges)?

16.4.1.8 Radio Properties of Seyferts and Quasars

The expression we use to estimate the observed GHz radio power of a black hole accretion model is given by equation (16.25). The computed radio power, therefore, will depend on the following:

- The black hole mass. Generally, the power increases rapidly as $\propto M_\bullet^{1.32\pm0.06}$.
- Accretion rate $\dot{m}^v$, with $1.2 < v < 2.4$.
- If the source being modeled is in the unstable state, the radio power will depend on where in its hysteretic limit cycle the source is when it is observed:
 - If the source is in a soft *substate*, it will be radio-quiet; equation (16.27), with the model source $\dot{m}$ (of any value), should be used.
 - If the source is in a hard ADAF-like *substate* near the jet line, it will produce a strong radio jet outflow; equation (16.24), again the source’s actual $\dot{m}$ should be used.
 - If the source is in a super-Eddington state (“slim” disk or super-Eddington wind), we adopt the observed result that such binary black hole systems tend to be radio-quiet, not -loud. However, at the present time we do not have a good mathematical expression for computing a model radio power in this case.
- Finally, if the central black hole is rotating rapidly, the output jet power can be dramatically altered (positively or negatively) from our estimates for non-rotating holes in Section 16.2.2. This effect will be discussed in Section 16.5.

For our 3C 120-like model Seyfert/QSO with $M_\bullet = 10^{7.5}\text{ M}_\odot$ and $\dot{m} = 0.3$, the (unbeamed) radio jet power in the unstable state will be $10^{38}\text{ erg s}^{-1}$ ($\log P = 22.0$) when it is in a soft *substate*; and it will be $5 \times 10^{40}\text{ erg s}^{-1}$ ($\log P = 24.7$) when near the jet line in its hardness–intensity diagram. Beamed radio outbursts in 3C 120 ($z = 0.033$) are of order 5 Jy (or $\log P \approx 25$ at 15 GHz; [608]), so this source may indeed be in the latter powerful jet-producing *substate* (*i.e.*, near its jet line).

16.4.1.9 Seyfert Radio Outliers?

As an example of how active this field is, two new papers appeared while this chapter was being drafted, one in the *Astrophysical Journal* [609] and one in the *Monthly Notices of the Royal Astronomical Society* [610]. Both reported on a lack of correlation between radio and X-ray emission in the Seyfert galaxy NGC 4051 (black hole mass of about $1.8 \times 10^6\text{ M}_\odot$). This is distinctly different from the fundamental plane correlation (equation (16.24)), which predicts that radio emission increases strongly as X-ray emission increases, in black hole engines of all masses. Fig. 16.14 shows that NGC 4051 lies well below the standard fundamental plane relation (solid line), and the lack of correlation produces an approximately horizontal line in this diagram. This diagram looks remarkably similar to Fig. 16.5: the fundamental plane relation provides a rough upper bound to the radio/X-ray plot, and there are several

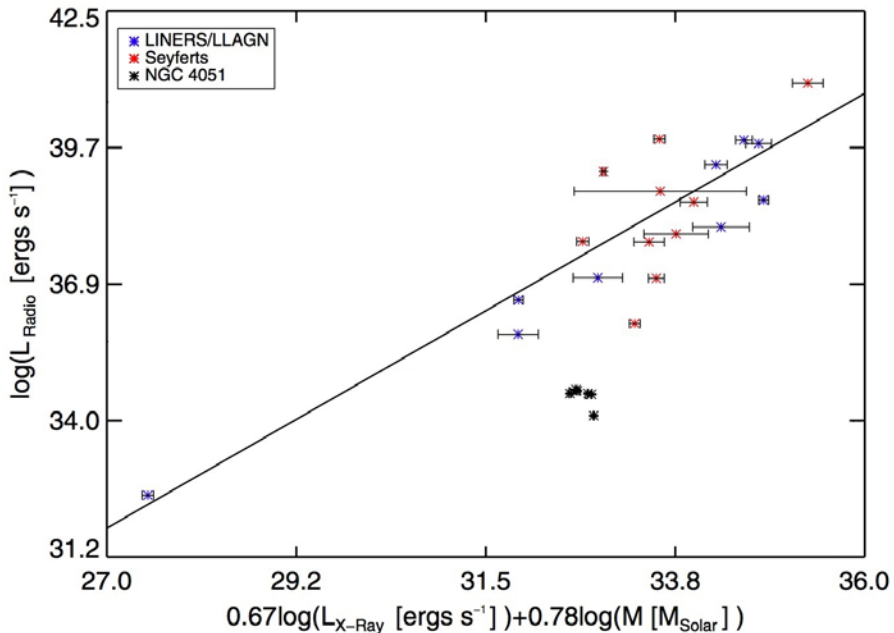


Fig. 16.14: The radio/X-ray luminosity plot for Seyfert galaxies, which bears a striking resemblance to Fig. 16.5. The solid line shows the standard fundamental plane relation for black holes of all mass. Some possible outliers, especially NGC 4051, are seen. Its radio emission is suppressed by a factor of ~ 10 , as with hard state binary black hole outliers, rather than the factor of hundreds or more in soft and steep power-law sources. However, considerable caution must be exercised, as NGC 4051 is a high-luminosity Seyfert, not a traditional low-luminosity hard state object. Adapted from Fig. 8 of [609]. Reproduced by permission of the AAS.

outlier sources that lie below the standard correlation. The horizontal correlation for NGC 4051 also is similar to that of the least radio-loud outliers in Fig. 16.5 (*i.e.*, those at $\log_{10} L_{\text{radio}} \approx 29$). It is tempting, therefore, to identify NGC 4051, and a few other sources in Fig. 16.14, as radio outliers, just like similar stellar black holes in Fig. 16.5. If this is the case, most comments about outliers in Section 16.2.4, and the possible application of MDAFs to the phenomenon, may be applicable to AGN as well.

However, this might be a decidedly wrong interpretation. Fig. 16.14 shows both hard state-type objects (LINERs/LLAGN) and soft state-type objects (Seyferts) on the same diagram. The hard state objects (blue points) do, indeed, all lie close to the fundamental plane, as we might expect. However, the soft state objects (red points plus NGC 4051, an NLSy1 galaxy), if they form a correlation at all, do *not* follow the fundamental plane. It is questionable whether or not the two classes even should be lumped into the same correlation. Indeed, NGC 4051 is a near-Eddington object; labeling it as an “outlier” from the fundamental plane of low-luminosity, hard state objects, as we do in the previous paragraph, seems highly improbable.

However, there yet may be a tenuous connection with MDAF-type accretion flows after all. NGC 4051 is known to have a break in its power fluctuation spectrum [611], indicating that, like certain black hole X-ray binary sources, it also has band-limited noise. Our primary model here for the band-limited nature of the spectrum is the presence of a strongly-magnetized inner disk region (MDAF) that quells the normal disk turbulence. But, if an MDAF is supposed to be a model for hard state objects, how could it have anything to do with an Eddington, or slightly super-Eddington, object such as NGC 4051? The answer may again lie with observations of black hole binaries, as reported in McClintock and Remillard [204]. Their Fig. 4.15 shows that both XTE J1550-564 and GRS 1915+105 can attain an SPL-like state, but with distinctly band-limited noise. The implication is that, just as ADAFs have a high-luminosity counterpart in the radiation-pressure-dominated slim disks, MDAFs also may have a high-luminosity counterpart as well, which manifests itself with band-limited noise in near- or super-Eddington SPL and NLSy1 sources. That is, the high-luminosity, band-limited noise sources like NGC 4051 and XTE J1550-564 and GRS 1915+105 (when in this strange SPL state) may have low radio luminosities compared with their non-band-limited counterparts and, therefore, be high-luminosity radio outliers nonetheless.

16.4.2 *Important Clues to How Massive Black Holes Form and Grow*

There have been some very recent observations that provide important clues to the mystery of how central black holes acquire their properties, and to the meaning of relations like the $M_\bullet\text{--}M_{\text{bulge}}$ and $M_\bullet\text{--}\sigma$ ones (see Fig. 2.18). John Kormendy of The University of Texas, and his colleagues, find that black hole properties correlate *only* with properties of “classical bulges” (central stellar bulges of early-type spiral galaxies or entire elliptical galaxies) [612]. They do *not* correlate with properties of the spiral disks or with “pseudobulges” (central concentration of stars in late-type spirals whose shapes are as flat as the galactic disks themselves). Nor do central black holes correlate with the properties of a galaxy’s dark matter halo [613]. These points argue strongly for the following:

- When a classical bulge exists (which occurs when the bulge has undergone many hierarchical mergers that created rather random stellar motions), the bulge and central black hole grow together and are limited in size by related processes.
- It is highly likely that the central black holes were created, or at least grown, mostly from the same matter that made up the bulge stars: baryons.
- While black holes apparently can be formed and grown in the centers of bulgeless spiral galaxies (ones with only pseudobulges), the process that does so is not (for some reason) as well coupled to the pseudobulge (or gaseous disk) as that which does so in the centers of bona fide bulges.

However, the fact that black hole masses do not correlate well with dark matter, pseudobulge, or gaseous disk properties does *not* mean that the holes could not have been partially (or even completely) fed and grown with dark matter, stars from the pseudobulge, gas from the disk, or even other black holes. It simply means that, whatever process limits the growth of the central black holes in classical bulges, that process also limits the growth of that bulge, *and* it also must be closely related to the bulge's velocity dispersion. Indeed, bulgeless galaxies still do have central black holes, and the most likely sources of material for creating and growing those holes are the gaseous disk, pseudobulge stars, halo dark matter, or other available black holes. The question is, therefore, what process grew holes in bulgeless galaxies and why does that process not correlate with any of these other sources of material?

16.4.3 Formation, Growth, and Fueling of MBHs in Mergers

“Formation” is probably not an appropriate term for black holes above $10^{3-4} M_{\odot}$ in mass. As discussed in Section 10.5, MBHs likely begin in the early universe in mini-halos as IMBHs of $10^{2-3} M_{\odot}$ or so. These mini-halos are believed to merge then in an hierarchical fashion, forming halos roughly twice their original mass, then four times, eight times, and so on. Each of these would be called a “major” merger, because the two merging halos would be about equal in mass. Of course, “minor” mergers also could occur, with one halo being significantly larger in mass than the other. In each merger of either type, the two black holes in each halo are expected to merge eventually into a single central object, thereby contributing to some hierarchical growth in the black hole size also.

In the merger scenario, classical spiral galaxy bulges and dwarf ellipticals are believed to have had similar merger histories; their only difference is that the former eventually acquired massive disks via infall of pristine gas, while the latter have not (or may never) acquire a significant gaseous disk.

The early dwarf galaxies are initially no larger than perhaps a globular cluster in mass. They usually are called mini-“halos” because their mass and gravity are dominated by dark matter. However, the same is true of virtually any galaxy that we see; the Milky Way also could be called a “macro-” or giant halo, which just happens to have a gaseous spiral galaxy inside containing $\sim 10\%$ of the mass. Nevertheless, unless black holes are born and grown by more exotic means than we currently envision,¹² present theories assume that it is this minor, baryonic component that makes up the stars and gas that dissipates into the very compact things that collapse to central black holes.

This point is especially important for forming black holes inside classical bulges. Taking into account the interstellar gas and stars that participate in halo mergers provides another source of material to grow the central holes, in addition to black

¹² *E.g.*, one alternative possibility could be that MBHs and SMBHs are created from dark matter very early in the universe and then serve as nucleation points for galaxies to form. Currently, this idea is not favored by the data.

hole mergers alone. It also ties hole growth to the gas and stars, making it easier to explain why black hole and galaxy bulge properties correlate so well, even though the bulge is well outside the black hole's gravitational sphere of influence.

16.4.3.1 Growth as a Hidden AGN

Initially the gas and population III stars in the mini-halo must have been rather pristine (mainly H and He) with few heavy elements. However, within a few massive star lifetimes (say, 10 Myr; [614]) supernova explosions would have enriched the interstellar material to the level of Pop II stars (metallicity of 10^{-2} solar, or higher). This would have included reasonable amounts of interstellar dust and molecular gas. The new heavy elements have several important consequences:

- Metal cooling drops the average star mass down to $\sim 1\text{ }M_\odot$.
- Stellar mass-loss rates will soar, making it very difficult to make black holes larger than $\sim 20\text{ }M_\odot$.
- Mergers that last more than a few massive star lifetimes will see the production of copious amounts of interstellar dust. That dust will be very efficient in hiding both the formation of additional stars and the feeding of the black hole by accretion of gas and stars.

So, after the hierarchical merger process gets started, virtually all major mergers (and perhaps even minor ones) will grow the black hole behind a thick veil of dust.

Observations in the mid infrared by the Spitzer Space Telescope (Section 2.3.4) show that the ratio of obscured to unobscured quasars is about 4:1 [109, 110].¹³ We expect a similar ratio for mergers of smaller galaxies going nearly back to the mini-halo stage.

Super-Eddington Growth Rates

How fast can a black hole grow in this rapid, hidden accretion phase? There is a common misconception that the black hole can grow no faster than at the Eddington rate $\dot{M}_{\text{Edd}}$ (equation (11.2)). This would set the minimum black hole mass doubling time to the Salpeter time

$$\tau_{\text{salp}} = \frac{M_\bullet}{\dot{M}_{\text{Edd}}} = 4.6 \times 10^7 \text{ yr} \left(\frac{\epsilon_{\text{acc}}}{0.1} \right) \quad (16.37)$$

However, black holes can double their masses on time scales significantly shorter than this while still radiating at (or even below) the Eddington limit. Some possible ways of doing so are: (1) accreting in a super-Eddington phase ($\dot{m} > 1$) in a “slim” disk; and (2) swallowing compact stars whole (other black holes, neutron

¹³ The NuSTAR hard X-ray telescope will be able to find even more obscured AGN than Spitzer did, possibly leading to a refinement in this ratio.

stars, white dwarfs, and even some main sequence stars). That is, the accretion efficiency in equation (16.37) above is a constant of order 5–10% only when the accretion is in a radiative state, preferably a thin disk. When the accretion carries trapped radiation into the hole, the effective efficiency is $\epsilon_{\text{acc}}/\dot{m}$ with $\dot{m} > 1$ (see Section 12.2), for both spherical and “slim” disk accretion. So, in these cases

$$\tau_{\text{doub}} \approx 4.6 \times 10^7 \text{ yr} \left(\frac{\epsilon_{\text{acc}}/\dot{m}}{0.1} \right) = \dot{m}^{-1} \tau_{\text{salp}}$$

For example, if $\dot{m} \gtrsim 10$, a black hole can double its mass in only a few million years.

Compton vs. Absorption Thick X-Rays

There has been a lot of discussion (related to NuSTAR planning) as to whether the interstellar gas and dust surrounding these Type 2, hidden AGN could be Compton thick, thereby significantly altering the X-ray spectrum. However, in the models that we discussed in Section 13.2 there is a potentially more important X-ray altering component that is not discussed much: a super-Eddington wind emanating from the accreting black hole itself. In those models the wind is so optically thick that all the disk emission (including the X-rays) is reprocessed to a thermal temperature of only (equation (16.19))

$$T_{*,\text{wind}} \approx 0.035 \text{ keV} \left(\frac{M_{\bullet}}{10^6 M_{\odot}} \right)^{-0.21} \left(\frac{\dot{m}}{30} \right)^{-0.89}$$

On the other hand, if the ISM were not Compton thick, and if the black hole were in a super-Eddington, “slim” disk phase, then a standard AGN X-ray spectrum would be expected. There are, therefore, three possible configurations that we would find when we observe a known mid-IR, hidden AGN with NuSTAR

1. We might see a (more-or-less) standard AGN X-ray spectrum. We then would conclude that (a) there is no super-Eddington wind and (b) there is no Compton thick medium surrounding the hidden AGN.
2. Or, we might see a Compton modified X-ray spectrum. We then would conclude that (a) there is no super-Eddington wind, but (b) the nuclear material has a Compton depth above unity.
3. Finally, we might observe no X-rays at all above 0.1 keV. Our conclusions then would be ambiguous. Either (a) there is a powerful super-Eddington wind, OR (b) the ISM is very Compton thick. Observations at other wavelengths would be needed to resolve this degeneracy.

16.4.3.2 Accretion and Growth as an Unveiled AGN

In the Andrew King model of black hole and bulge growth (Section 10.5.2), when the hole grows to a large enough mass (equation (10.12))

$$M_{\bullet} \geq M_{\bullet, \max} = 1.2 \times 10^6 M_{\odot} \left(\frac{\sigma_V}{60 \text{ km s}^{-1}} \right)^4$$

the radiation and winds from the black hole (accreting above the Eddington limit) will unbind the bulge interstellar material, turning off both the star formation and the black hole feeding. How quickly this takes place, and through what stages the central engine passes, are some of the current significant puzzles in AGN research. We will discuss all possible phases below – in reverse order from the previous discussions in Sections 16.2 and 16.3.

The Super-Eddington Wind Phase ($\dot{m} \sim 30$)

Initially the unveiled AGN still might be in a very rapid accretion phase (say, $\dot{m} \gtrsim 30$) with an active and dense super-Eddington wind, but an approximately steady bolometric luminosity of $\sim 10^{44} \text{ erg s}^{-1} (M_{\bullet}/10^6 M_{\odot})$. The wind's color temperature will be in the extreme UV ($\sim 0.035 \text{ keV}$ or $T_{*6} \sim 0.41$, for $\dot{m} = 30$ and $M_{\bullet} = 10^6 M_{\odot}$; equation (16.19)). The wind velocity in this case would be $\sim 3800 \text{ km s}^{-1}$ (equation (13.30)). The electron scattering depth would be ~ 4500 (equation (13.27)), the absorption depth of the wind only a little above unity, and the Compton parameter $y \approx 5000$ (equation (12.46) for $\tau_{\text{es}} \gg 1$). With such a large Compton depth, and an absorption depth near unity, we might expect the X-ray spectrum to be significantly suppressed by the outflowing wind. All of these parameters may be variable on thermal time scales in the range 9 d to 3 yr and on secular (~ 2300 yr) time scales.

This is only an example model. Section 16.2.4 and equations (13.30) show how to build a better super-Eddington wind model from a measurement of T_{*} and the bolometric luminosity.

The Super-Eddington “Slim” Disk Phase ($\dot{m} \sim 3$)

If the accretion rate drops into the range $1 < \dot{m} \lesssim 10$, the wind might lose its sphericity, leaving a fairly visible “slim” disk accreting into the black hole, with a bipolar thermal wind or jet. The disk color temperature would be a little hotter ($\sim 0.05 \text{ keV}$ or $T_{*6} \sim 0.62$ for $\dot{m} = 3$ and $M_{\bullet} = 10^6 M_{\odot}$; equation (16.18)). The X-ray spectrum should look a lot like that of a steep power law binary X-ray source (Fig. 16.10) and reflect the now-visible coronal emission. The object should be variable on thermal ($\sim 7 \text{ hr}$ to 3 mo) and secular (~ 500 yr) time scales – somewhat shorter now because of an assumed factor of 10 decrease in accretion rate. We expect this type of object to be radio-quiet, just like the SPL black hole binaries.

Sources in this phase should have (1) *no* strong spherical wind, (2) a strong X-ray spectrum, (3) a central UV ionizing source and a BLR with relatively narrow lines (Section 16.4.1), and (4) rather quiet radio emission. That is, they might appear as a NLSy1s or, if viewed from a polar direction, as a BAL AGN.

The Unstable, High Accretion Rate Phase ($\dot{m} \sim 0.3$)

When the accretion rate drops into the range $\dot{m}_1 \lesssim \dot{m} < 1$, the “inner” portion of the disk itself may be unstable on thermal (~ 1 hr to 6 d) and secular (~ 110 yr) time scales (for $M_\bullet = 10^6 M_\odot$ and $\dot{m} = 0.3$). Depending on where in the limit cycle the source is, the GHz radio power could be rather low when in a soft substate (5×10^{37} erg s $^{-1}$; equation (16.28)) or rather high when in a hard substate near the jet line (4×10^{38} erg s $^{-1}$; equation (16.24) with $\dot{m} = 0.3$), with intermittent outbursts on thermal time scales. Optical–UV luminosities would be $\sim 4 \times 10^{43}$ erg s $^{-1}$ for our $10^6 M_\odot$ example. Such an object would appear like a normal Seyfert nucleus (Section 16.4.1).

The Steady, Low Accretion Rate Phase ($\dot{m} \sim 0.03$)

When the accretion rate drops below $\dot{m}_A \approx 0.05$, a hot ADAF coronal inflow replaces most of the thin cool disk. In the stellar-mass black hole case we tentatively identified such objects as being standard, “fundamental plane” sources in Fig. 16.5. However, near the borderline value of $\dot{m} = 0.03$ there also exist outlier sources with radio power a factor of ten or so below the fundamental plane values. We have tentatively identified the latter objects as ones with an inner MDAF and a strong low-frequency rotational QPO.

Again, without any more information other than binary black holes for comparison, we speculate that moderately low-luminosity AGN might exhibit some sources that lie in the fundamental plane and some that are outliers. In both objects the X-ray luminosity will be essentially the same ($\sim 10^{42}$ erg s $^{-1}$ for $M_\bullet = 10^6 M_\odot$; equation (16.10)). In the fundamental plane, the GHz radio luminosities for this type of object would be $L_R \sim 10^{37}$ erg s $^{-1}$ (equation (16.24)), while for the outliers L_R would be a factor of 10 or so below that. In the outliers we also might expect QPOs in the X-rays with periods of order several hours. But we do *not* expect the thermal or secular instabilities that are believed to accompany thin accretion disks with significant, or dominant, radiation pressure. There still might be a moderate amount of thermal radiation (again in the outlier sources) from an evaporating, thin accretion disk with a color temperature of < 0.02 keV. Most of its emission would be in the EUV, but its shallow $\nu^{1/3}$ spectrum in the far UV might be a secondary indication of this *EUV bump*.

16.4.4 Formation, Growth, and Fueling of MBHs without Mergers

Recently, there has been considerable interest in how black holes might be fueled and grown even when the host galaxy is not experiencing a large merger [612]. Below we discuss fueling by infall of mass lost by dwarf elliptical galaxy stars and by transport of substantial amounts of AGN fuel to the center via inspiraling compact star clusters. While this latter mechanism may be at work in mergers, it also can operate in non-merging galaxies with sufficient gas content.

16.4.4.1 In the Centers of Rather Inactive Dwarf Ellipticals ($\dot{m} \lesssim 0.003$)

With gas content and star-formation rates well below that of spirals, dwarf ellipticals emerging from a recent merger, or in between mergers, generally have only a meager source of fuel for the AGN engine. It is little wonder that they often are the hosts of LLAGN. For a $10^{5-7} M_{\odot}$ MBH, to exceed the maximum ADAF accretion rate of $\dot{m}_A \sim 0.05$, one would need to supply $\sim 10^{22-24} \text{ g s}^{-1}$. As we saw in the previous section, this would require a star captured by the black hole to be in thermal mass transfer and be $10\text{--}40 M_{\odot}$ in size. Clearly this would be rare when the galaxy's star-formation rate is low and few O/B stars are around.

Another, more likely possibility might be infall from a number of old, solar-type stars in the galactic nucleus that are losing mass when they reach the red giant phase. We estimate the Bondi accretion rate for this process to be significantly sub-Eddington (using $V \sim \sigma_V \sim 57 \text{ km s}^{-1} (M_{\bullet}/10^6 M_{\odot})^{1/4}$ and equation (11.9))

$$\dot{m}_B \sim 0.0013 \left(\frac{M_{\bullet}}{10^6 M_{\odot}} \right)^{1/4} n$$

Unless the gas is as dense as a molecular cloud ($n \gg 1$), this accretion rate falls short of $\dot{m}_A$, leaving the central hole a weak, inefficiently accreting object.

We will adopt $\dot{m} = 0.003$ and $M_{\bullet} = 10^6 M_{\odot}$ or our example model. This low accretion rate places the central engine well below the rate where, based on binary black hole cases, we might expect a pure advection-dominated accretion flow with the radio and X-ray lying in the fundamental plane, with luminosities of $\sim 10^{36} \text{ erg s}^{-1}$ (equation (16.24)) and $\sim 10^{40} \text{ erg s}^{-1}$ (equation (16.10)), respectively. Based on the first, second, and fourth panels of Fig. 16.3, one might *not* expect strong QPOs in these very sub-Eddington objects.

16.4.4.2 In the Centers of Bulges in Spiral Galaxies

Almost by definition, an elliptical galaxy is one that has undergone at least one, and probably many, mergers. And, conversely, a spiral galaxy is one that has not had a major merger since its bulge finished forming and slowly accumulating gas has settled into a disk. Yet, spirals seem to periodically become Seyfert galaxies, and

accrete at near-Eddington ratios [615], without undergoing another major merger. Why? Clearly the answer is related to the amount of gas in the disk, but the large amount of angular momentum in that disk provides a puzzle: how does that gas get into the nucleus to fuel the central engine?

In Section 11.1.4 this author expanded on one process, suggested by Ebisuzaki *et al.* [401], by which interstellar gas could be locked up in compact star clusters, which are then trucked rapidly to the center by dynamical friction angular momentum loss, tidally disrupted by the central black hole, and then finally accreted therein, both stars and gas. The stellar disruption works as long as $M_{\bullet} < 10^8 M_{\odot}$. Of course, this mechanism needs much more detailed study and testing before it can be accepted as a major trigger of Seyfert outbursts. However, the scenario is reasonably plausible, it can fuel the central hole to Eddington rates or above, it will increase the bulge stellar mass, and it even can explain star-formation rings at a few hundred parsecs from the galactic center (Section 11.1.4 and [416]).

In the case of a classical spiral bulge, with a relation between hole and bulge mass

$$M_{\bullet} \approx 0.0013 M_{\text{bulge}}$$

the equation for the accretion rate onto the black hole (the last of equations (11.53)) can be written as

$$\dot{M}_{\text{clst}}(r_h) = 0.23 M_{\odot} \text{ yr}^{-1} \eta_g \times \left(\frac{M_{\bullet}}{10^6 M_{\odot}} \right)^{1/4} \left(\frac{M_{\text{clst}0}}{3 \times 10^6 M_{\odot}} \right)^{3/4} \left(\frac{\ln A}{3} \right) \left(\frac{M_{\text{clmp}0}}{10^7 M_{\odot}} \right)^{-1/4} \quad (16.38)$$

where $M_{\text{clmp}0}$ is the mass of the gas clumps that inspiral to the star-formation ring radius, and $M_{\text{clst}0}$ is the mass of the compact clusters formed at that radius and inspiral toward the black hole. In this case the accretion rate is super-Eddington for any central hole smaller than $\sim 10^7 M_{\odot}$, and the accretion rate $\dot{M}$ (in g s^{-1}) depends strongly only on $M_{\text{clst}0}$. So, in such a spiral galaxy, the black hole mass will be limited by the mass of the bulge (which is also growing due to this star cluster infall process), and a $M_{\bullet}-M_{\text{bulge}}$ (and $M_{\bullet}-\sigma$) relation will be maintained.

Eventually, however, the growth rate will go below Eddington when the black hole becomes so massive that the above rate becomes very sub-Eddington. When that happens, the black hole and bulge growth rates will become less coupled. This situation, and the general case of pseudobulges, is discussed below.

16.4.4.3 In the Centers of Pseudobulges in Spiral Galaxies

As discussed in Section 16.4.2, there appear to be at least two methods of growing black holes (and bulges or pseudobulges) in late-type galaxies: (1) the, now classical method, in which the hole grows until its Eddington luminosity drives out the gas that feeds both hole and star formation in the bulge (see also Section 10.5.2) and

(2) one where the growth of hole and stellar content are not so tightly coupled, or limited. We suggest here that the cluster inspiral method is at work in *both* cases, but that this method has an inherent instability when a correlation between bulge and hole mass either does not exist, or no longer can be maintained.

When a good correlation between $M_{\bullet}$ and M_{bulge} does not exist, the accretion rate onto the central hole (see equations (11.53)) has an additional factor compared with equation (16.38) above

$$\dot{M}_{\text{clst}}(r_h) = 0.23 M_{\odot} \text{ yr}^{-1} \eta_g \left(\frac{M_{\bullet}}{0.0013 M_{\text{bulge/pseudobulge}}} \right)^{3/4} \times \left(\frac{M_{\bullet}}{10^6 M_{\odot}} \right)^{1/4} \left(\frac{M_{\text{clst0}}}{3 \times 10^6 M_{\odot}} \right)^{3/4} \left(\frac{\ln \Lambda}{3} \right) \left(\frac{M_{\text{clmp0}}}{10^7 M_{\odot}} \right)^{-1/4} \quad (16.39)$$

and that factor introduces a strong dependence on the ratio of black hole to bulge mass. In the event that the ratio is much larger than the classical $M_{\bullet}-M_{\text{bulge}}$ relation, the hole will be overfed and continue to grow rapidly. However, when that ratio lies below the standard $M_{\bullet}-M_{\text{bulge}}$ relation, the hole will be continue to be undermassive. The latter seems to be the case for many black holes in pseudobulges [612]. A few galaxies (including NGC 2787, with a possible small classical bulge) lie on the classical mass correlation, but the majority have black holes undermassive by factors of $\sim 5-50$. Their central engines may still be fueled by cluster inspiral, but not at a rate sufficient to maintain a proper $M_{\bullet}-M_{\text{bulge}}$ relation. (Note, however, that even a “fully” tidally stripped compact cluster will keep its central IMBH and some stars and gas near it. This still will grow the galaxy’s MBH and produce gravitational waves when the two holes merge, but there will be little fueling of the engine itself.)

What is the physical origin of this “cluster inspiral instability”? The inspiraling clusters are continually being tidally stripped by the dominant tidal component of the galactic gravitational field. If the black hole is undermassive, then its radius of influence (r_h ; equation (11.20)) is small, and more of the inspiraling cluster then will be stripped away, before it sinks to r_h , to become part of the general bulge/pseudobulge stellar content. This leaves less cluster mass to be captured by the hole. On the other hand, when the black hole is overmassive, it has a large r_h , allowing it to capture the inspiraling cluster earlier, stealing more stars and gas from the general bulge stellar population.

In fact, even when bulge and hole are correlated via super-Eddington winds, and produce a good $M_{\bullet}-\sigma$ relation, the black hole feeding rate above still will be affected somewhat by the hole-to-bulge mass ratio. This could cause significant amounts of scatter in the $M_{\bullet}-M_{\text{bulge}}$ relation, even for classical bulges.

16.5 Supermassive (10^8 – $10^{10} M_\odot$) Black Holes: The Ultimate Engines

In terms of physical size, even the MBHs are dwarfed by the supermassive black holes; the latter are probably the largest stars in the universe, some as much as thirty times larger than the biggest red supergiant stars. The horizon of a typical quasar black hole ($10^9 M_\odot$) would encompass the orbit of Uranus (20 AU), and the largest black holes known ($\sim 10^{10} M_\odot$ and 200 AU in radius) would encompass all the trans-Neptunian objects, including Sedna and beyond, with room to spare. The latter objects have the mass of an entire moderate-sized galaxy! Nearby galaxies with SMBHs in their centers are NGC 5128 (Centaurus A; ~ 4 Mpc distant, $(3\text{--}7) \times 10^7 M_\odot$ in mass, at the lower end of the range) and Messier 87 (16 Mpc distant, $(3\text{--}6) \times 10^9 M_\odot$ in mass, at the upper end). It is fair to say that M87 probably was a quasar in the distant past and a radio-loud one as well.

Besides mass, some other observed properties set the SMBHs apart from the MBHs:

- Most importantly, they occur exclusively in a different type of galaxy – the giant elliptical (gE). Current galaxy evolution theory explains ellipticals as having been subject to at least one, and perhaps many, major mergers over the eons, and also subject to many minor mergers (or “harassments”) as well. This idea explains their elliptical shapes, their location in dense regions of galaxy clusters, their large masses, and their large black holes. For example, M87 is an enormous elliptical at the center of the Virgo cluster, and NGC 5128 appears to be an elliptical that has devoured a dusty spiral disk [616, 617].
- Central SMBH engines in elliptical galaxies tend to be much more radio-loud than their MBH counterparts in spirals and dwarf ellipticals (dEs). Much of the dependence on mass ($L_R \propto M_\bullet^2$; equation (16.25)) could be due to the dependence of L_R in the black hole fundamental plane ($\propto M_\bullet^{1.4}$). However, this still leaves an additional unexplained $L_R \propto M_\bullet^{0.6}$ dependence. Increasing merger activity is one obvious candidate mechanism, but how mergers would influence radio luminosity needs to be seriously examined.

Finally, there are a few important theoretical properties of SMBHs that make their engines quite unique:

- The time scales in accreting SMBHs are enormous (Tables 16.1 to 16.3), especially the secular (accretion) times. While MBH secular time scales can be much longer than a human lifetime, those in SMBHs are generally much longer than the entire history of humans (and even primates). Events that may look stable to observers (and even to the black holes’ immediate environs) actually are undergoing unstable transitions. Indeed, at Eddington accretion rates, the secular times (0.1–50 Myr) can be a fair fraction of the merger time itself. Assuming $\dot{m}$ constant may itself be incorrect.
- Also, as mentioned in Section 11.1.3, for central black holes with $M_\bullet \gtrsim 10^8 M_\odot$ (almost all SMBHs) the tidal disruption of normal dwarf (*i.e.*, solar-type) stars

does not fuel the engine. The stars are swallowed whole by the hole. This means that bright supermassive objects near the Eddington limit (quasars, ULIRGs, HyLIRGs) cannot be fueled by the compact cluster inspiral method discussed in Sections 11.1.4 and 16.4.4. They must be fed with interstellar gas *not* locked up in stars.

16.5.1 Formation, Growth, and Fueling of SMBHs

16.5.1.1 Formation and Growth of SMBHs

The formation and growth of SMBHs begins in essentially the same manner as for MBHs (Section 16.4.3). The difference, of course, is that SMBHs continue to grow in mass well beyond $10^{7.5} M_\odot$. In the Eddington luminosity-limited scenario, this can happen only in galaxies whose bulge mass exceeds nearly $10^{12} M_\odot$, *i.e.*, in gE galaxies, which are formed in regions of the universe where the number density and collision rate of galaxies is high (dense groups and clusters).

It is worth pointing out here that almost any kind of collision of galaxies will result in an elliptical galaxy product: major mergers (two gEs, two spirals, or a gE and a spiral) or minor mergers (gE with a dE or dwarf irregular [dIrr]). The only time a spiral product might emerge would be in a minor merger between a giant spiral (gS) and a dwarf of some kind. Even then, the disk could be quite disturbed before the spiral completely recovers, probably with a more massive central bulge and a correspondingly larger central black hole.

16.5.1.2 Growth vs. Fueling in SMBHs

A significant part of the initial growth phase may be driven by compact cluster inspiral. This is an efficient method for storing interstellar gas in stars, rapidly transporting those stars to the central black hole, and then releasing the gas in stellar tidal disruption. As long as $M_\bullet < 10^8 M_\odot$, this process both grows the black hole *and* fuels its central engine.

Above $10^8 M_\odot$, however, cluster inspiral becomes extraordinarily *inefficient* at fueling AGN activity (most stars are swallowed whole); but it still remains a marvelously *efficient* way of silently growing the black hole mass. If the accretion of regular interstellar gas accompanies compact cluster inspiral, the black hole will accept that gas at only the Eddington rate ($\dot{M}_{\text{Edd}}$) or perhaps a few $\dot{M}_{\text{Edd}}$ in a “slim” disk. However, that same hole will accept almost as many stars as the cluster inspiral process will throw at it. The accretion rates possible with cluster inspiral are not exorbitantly high, and they are not a strong function of the bulge mass. The normalized accretion rate is (equation (11.52))

$$\dot{m} \approx 0.33 \eta_g \left(\frac{M_{\text{bulge}}}{7.7 \times 10^{10} M_{\odot}} \right)^{-3/4} \quad (16.40)$$

While this equation is not strictly valid for $\eta_g > 1$, there is a strong indication that, with reasonably high gas-to-star ratios, the SMBH could reach Eddington growth or higher from this silent feeding method. This could lead to the central black hole becoming overmassive over the standard $M_{\bullet}-M_{\text{bulge}}$ relation.

Whether this actually occurs or not, however, will depend on the production rate of compact star clusters anchored by a central IMBH in an AGN galaxy merger. There is, indeed, some evidence that, for $M_{\bullet} > 10^8 M_{\odot}$, the $M_{\bullet}-M_{\text{bulge}}$ relation may be overmassive for redshifts in the range $1 < z < 2$ by a factor of ten [618]. High gas-to-star ratios, along with the compact cluster inspiral method, may provide an alternative explanation to the proposition that SMBHs were formed before their galaxies formed. In fact, if the gas-to-star ratio remains high during black hole growth, when it comes time for the central engine to produce super-Eddington winds and eject most of the interstellar gas, a high gas-to-star ratio galaxy will lose a great deal of its mass, making the black hole appear even more overmassive.

16.5.1.3 Cooking with Gas

Despite the evidence that black holes may grow silently to overmassive sizes early in the process of galaxy formation, there still is much evidence that SMBHs are *fueled* up to the Eddington rate by regular interstellar gas:

- First, the $M_{\bullet}-\sigma$ relation is fairly tight up to masses of $M_{\bullet} \sim 10^{10} M_{\odot}$. Therefore, if Andrew King's Eddington wind feedback mechanism is valid, then radiatively efficient accretion must have occurred some time during the lives of SMBHs in order to limit the bulge and black hole masses simultaneously.
- Secondly, and more importantly, we observe large numbers of AGN whose luminosities are above the Eddington limit for a $10^8 M_{\odot}$ black hole ($10^{47} \text{ erg s}^{-1}$ quasars, BAL QSOs, ULIRGs and HyLIRGs, *etc.*).

Other methods involving cluster inspiral will not work: the clusters themselves will have been stripped of any interstellar gas early on, and any gas brought in by giant and supergiant cluster stars will produce an accretion rate much smaller than that in equation (16.40).

The natural solution is to appeal to the process that most investigators already believe is at work in these objects. Some of the tenuous interstellar gas and dust accreted by the galaxy during a merger must somehow make it all the way down into the final 10 pc near the black hole without being converted into stars. Even if the black hole grows rapidly by inefficient accretion processes, if efficient luminous accretion also occurs simultaneously, the hole's super-Eddington wind will limit the bulge and hole growth when the SMBH reaches the mass given by equation (10.12).

While this is probably the solution to SMBH *fueled*, it still is a great mystery as to *how* the gas loses nearly 100% of its angular momentum to reach inside the

radius r_h . In this author's opinion, numerical simulations of this process are still too crude to even begin to answer this question. Even today's refined grids can create artificial shear viscosities that transfer angular momentum inadvertently. Shocks can transfer angular momentum also, but they can form stars as well. While atomic line cooling and self-gravity may be easy to include in these simulations, processes like star and star cluster formation are not handled well at all yet. The problem seems very daunting to me – essentially simulating the structure and evolution of an entire galaxy, and its nuclear environs as well, with extremely high resolution.

This angular momentum problem is reminiscent of the accretion viscosity problem in the 1970s and 1980s, which was solved by introducing magnetic fields as a key physical component of the simulations. It is very possible that the same is true for this much more complex problem. However, I do not think it is as simple as introducing the MRI into galactic structure and evolution codes. Galactic magnetic fields are likely to be stronger than the relatively weak MRI fields in accretion disks. Stronger fields can transfer angular momentum more efficiently, provide large-scale channels toward the center, and prevent clouds from collapsing and forming stars. One very speculative possibility is that some parts of the ISM have stronger fields than others. In the weak field regions compact star clusters can form and inspiral as discussed earlier. In the more magnetized regions stars do not form; instead, gas loses its angular momentum magnetically and is channeled to the center. Beyond speculation like this, however, the processes of fueling supermassive black holes in giant elliptical galaxies remains a great mystery in my view.

16.5.2 Growth of the Elliptical Stellar Bulge

The formation of stars in a growing galaxy is a complex problem that is well beyond the scope of this book. However, there are a few points related to SMBHs that we need to make.

16.5.2.1 Ratio of Compact to Loose Star Clusters Created in a Forming Galaxy

First, it is assumed by most investigators that bulge growth and black hole growth both occur in mergers (major and minor). So far, we have discussed only the *compact* cluster inspiral method of growing the central black hole. However, it is easy to show that, although this method leaves the compact clusters stripped of most of their stars, this stripping is not nearly sufficient to grow the stellar bulge enough to create a healthy $M_\bullet$ – M_{bulge} relation ($M_{\text{bulge}}/M_\bullet \approx 10^{10} M_\odot / 1.3 \times 10^7 M_\odot = 770$). The ratio of stellar mass stripped to stars entering r_h is

$$\frac{M_{\text{clst}0}}{M_{\text{clst}}(r_h)} = 30 \left(\frac{M_\bullet}{10^8 M_\odot} \right)^{-1/4}$$

(see equation (11.52) with a constant $M_{\text{clst}0} = 3 \times 10^6 M_{\odot}$). This ratio is only 10–30 throughout the mass range $10^8 M_{\odot} < M_{\bullet} < 10^{10}$. The rate at which *compact* clusters release stars to the bulge, therefore, is far too low to grow the bulge by a factor of 25–80! It will, instead, overgrow the central black hole.

Of course, there certainly are other ways to make bulge stars. One method is simply to produce “loose” (non-compact) clusters that are *not* anchored by an IMBH. Indeed, the problem of properly growing both hole and bulge could be solved simply by making compact clusters only 1–4% of the time. The other 96–99% of the time only loose clusters would be created, and these would be fully tidally disrupted before reaching r_h . Since star formation is believed to take place in gravitationally bound molecular clouds anyway, this scenario is not a bad “straw-man” model for stellar bulge and black hole growth in elliptical galaxies.

16.5.2.2 How does a Bulge Catch up with an Overgrown Central Black Hole?

The other issue is related to the possible sudden ejection by a super-Eddington AGN wind of most of the gas in a gas-rich proto- or primeval galaxy ($\eta_g \gg 1$); see discussion after equation (16.40). This not only would leave the black hole appearing overmassive, *it also could unbind the stellar bulge itself* (just as the high-mass lost in a supernova explosion can unbind the binary star system in which it occurs). Are there indeed galaxies strewn about the universe that have been dissipated by AGN feedback?

Probably not. Elliptical galaxies and other bulges sit in dark matter halos that are more massive than the bulge stars, plus gas, by about a factor of ten. Immediately after the loss of all the (dominant) gas, the stellar bulge would initially expand adiabatically and the stellar velocity dispersion would drop precipitously ($\sigma_V \propto T^{1/2} \propto (p/\rho)^{1/2} \propto r^{-1}$). So, the stellar gas would never reach the escape speed of the halo. Instead, in a few dynamical times the bulge would re-collapse, violently relax, “radiate” any excess thermal energy by ejecting a small fraction of stars at high velocity, and then assume a standard elliptical distribution ($\sigma_{V,\text{new}} \propto M_{\text{bulge, new}}^{1/4}$; $R \propto M_{\text{bulge, new}}^{1/2}$) with a greatly reduced mass because of all the gas lost.

In fact, the gas blown out by the earlier super-Eddington AGN may not entirely escape the dark matter halo either. The outward velocity of that gas will be approximately equal to the original velocity dispersion σ_V , whereas the dark matter halo with an order of magnitude more mass will have an escape speed of

$$V_{\text{esc, halo}} = \left(\frac{2 G M_{\text{halo}}}{R_{\text{halo}}} \right)^{1/2} = 3.6 \sigma_V \left(\frac{M_{\text{halo}}}{10 M_{\text{bulge}}} \right)^{1/4}$$

So, $V_{\text{wind}}/V_{\text{esc, halo}} \approx 0.28$. The possibility that much of the gas may not escape from the halo means that a lot of gas may cool and re-collapse into the stellar bulge again at later times. It then will once again attempt to populate the bulge with more new stars – and also grow the central hole.

This sort of speculation raises many questions. Is gas fallback a likely process when the galaxy is gas-, rather than star-, dominated ($\eta_g \gg 1$)? Is a galaxy that is *not* gas-dominated ($\eta_g < 1$) more likely to have its small amount of gas escape the halo entirely in a super-Eddington wind? These questions might be answerable with some single-galaxy simulations with varying proportions of gas, stars, and dark matter.

Furthermore, is it plausible for compact clusters (with IMBHs) to be generated in great abundance in relatively pristine gas-rich galaxies (thereby overgrowing the SMBH relative to the bulge)? And, are there similar reasons why enriched, recycled gas might preferentially form loose clusters, thereby starving the SMBH of much of its growth and helping the stellar bulge to catch up? Such questions are more difficult to answer than the first set, but also premature. It is far from clear at this stage exactly how stars are created at different stages in a galaxy's life and what role, if any, cluster formation and tidal disruption play in building new galaxies.

16.5.3 *Jet Production and Propagation in the Mighty Quasars*

The propensity of supermassive black holes to produce powerful radio jets is well known, and it has the potential to teach us about jet production in other, less massive, black hole systems. We find below that the evidence is very strong that these jets result from the Blandford–Znajek process operating in a rotating SMBH – even as it seems that such evidence in binary *stellar* mass black holes is lacking (Section 16.2.2). So far we have no solution to this apparent paradox.

Because of their great masses, and therefore large linear sizes, SMBH systems give us jets with the greatest *angular* sizes by far – despite their enormous cosmological distances. SMBH jets can be studied with Very Long Baseline Interferometry (VLBI) in the radio, from both ground and space, down to resolutions of only a few Schwarzschild radii – a feat only dreamed about at other wavelengths. Therefore, understanding how SMBH jets are launched, accelerated, and collimated has implications for black hole engines of all masses, especially X-ray binaries and γ -ray bursts. Yet, for some inexplicable reason, the interest of the scientific community in what high-resolution interferometry has to offer is rapidly waning. If this trend is not reversed, the understanding of jets in all black hole systems, and even in proto-star and other stellar systems, could be set back several decades or much longer, as instruments and projects are abandoned and shut down.

16.5.3.1 Quasar Jets are Powered by Rotating SMBHs

The predicted jet luminosity of the Blandford–Znajek process (*i.e.*, black hole-driven jets) was discussed in Sections 14.3.3 and 14.4.2, and its dependence on black hole spin and accretion state was shown in Fig. 14.25. In addition, a similar discussion of the Blandford–Payne process (accretion disk-driven jets) was given

BZ "Radio" power vs. black hole spin

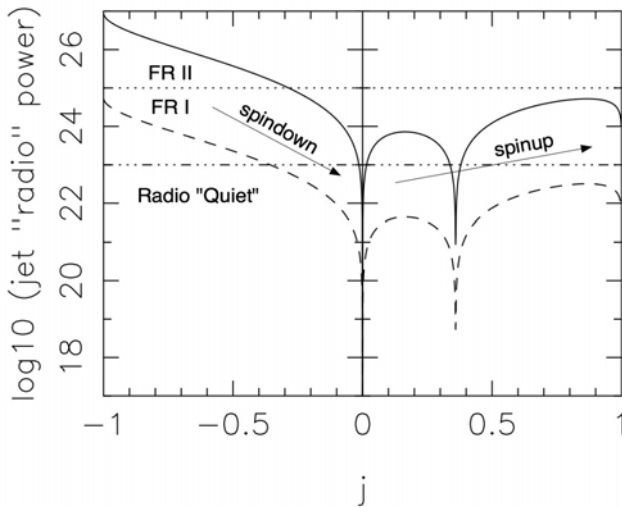


Fig. 16.15: The spin-dependence of the Blandford–Znajek absolute radio power for a $10^9 M_\odot$ black hole. Fig. 14.25 was scaled to radio power with the fundamental plane eq. (16.25), and to a $10^9 M_\odot$ black hole using a maximum radio power of $P \sim 10^{27} \text{ W Hz}^{-1}$. The dashed line shows suppression of the jet due to a thin disk vs. a thick ADAF (solid line). Dotted line shows the typical FR I–II break for galaxies with $10^9 M_\odot$ black holes, and the triple-dot-dashed line shows a rough boundary between radio “loud” and “quiet” sources. The most radio-loud sources are black holes with maximal retrograde spin (relative to the accretion flow rotation) and with geometrically thick inflows. The most radio-quiet ones have cool, geometrically thin disks and positive prograde spin. Dramatic dips occur when the black hole is not rotating or when it and the inflow rotate at the same rate.

in Sections 15.1.1 and 16.2.2, with *its* spin and accretion state dependence shown in Fig. 15.5. The following arguments strongly favor the conclusion that black hole (not accretion disk) spin powers SMBH radio jets:

1. In 2002 Qingjuan Yu and Scott Tremaine, of Princeton University, [619], and Martin Elvis at the Harvard-Smithsonian Center for Astrophysics and colleagues in Italy [620], showed that the average accretion efficiency (ϵ_{acc}) in AGN and QSOs is about 10–15%. This can occur only if most SMBHs are spinning rapidly ($0.7 < j < 1.0$). However, most AGN and QSOs are radio-quiet, implying that rapidly spinning *prograde* ($j > 1$) black holes do not produce strong jets. This result is consistent with the spin-dependence of the Blandford–Znajek effect (Fig. 14.25) but not with the spin-dependence of the Blandford–Payne effect (Fig. 15.5).
2. Recent reliable measurements of black hole spin using the relativistic iron line method (see [621] for a review), show that the most rapidly spinning black holes (sources like MCG 6-30-15 [622, 623] and 1H 0707-495 [624], with $r_{\text{isco}} <$

- $1.6 r_g$ or $j > 0.98$) are radio-quiet (even silent). This is consistent with item #1 above and particularly indicative of the precipitous drop in the Blandford–Znajek power as $j \rightarrow +1$ (Fig. 14.25).
3. Conversely, those sources that have strong radio jets (3C 120 [625] and 3C 390.3 [626], with $r_{\text{isco}} \sim 10$ – $20 r_g$) do *not* show a broad iron line or r_{isco} very near r_H . In fact, the r_{isco} determinations for these sources are much more consistent with rapidly-spinning, retrograde black holes ($j \rightarrow -1$, $r_{\text{isco}} = 9 r_g$). This also is consistent with Fig. 14.25, which predicts much more powerful jets when $j \rightarrow -1$.
 4. Recent studies of the X-ray Compton reflection component in different AGN [627] show that radio-quiet objects have measurable Compton reflection, indicating an optically thick inner accretion disk (*i.e.*, $r_{\text{isco}} \rightarrow r_g$). On the other hand, radio-loud objects have no detectable Compton reflection component, indicating an optically thin accretion flow or gap – again consistent with (although not unique to) $j \rightarrow -1$.
 5. Furthermore, the range in AGN radio luminosity for a given optical luminosity can be four or more orders of magnitude [628, 117]. Fig. 16.15 shows a radio luminosity version of the Blandford–Znajek Fig. 14.25, stretched by a factor of 1.2 in the log (see equation (16.25)) and normalized so the strongest radio power is a maximum of $P \approx 10^{27} \text{ W Hz}^{-1}$ (typical maximum power of a $10^9 M_\odot$ black hole in an extragalactic radio source). The range of predicted radio power for different black hole spins and accretion states is $21.7 < \log P < 26.8$, more than 5 orders of magnitude, even without taking into account the radio silent dips at $j = 0, 0.36$, and 1.0 .

In fact, the Blandford–Znajek process is so good at explaining the basic properties of SMBH jets that almost any addition at all of power from an accretion disk-driven jet actually *destroys* the good agreement between observation and theory. *This is quite a big and unsolved problem in quasar theory.* The Blandford–Payne process works quite well for stellar-mass black hole systems (Section 16.2), and it also works well for protostellar and other stellar systems. It should be at work in AGN systems as well, so why does the Blandford–Payne process seem so absent in AGN and QSRs?

There really can be only a limited number of answers, if our ideas on how jets are formed are at all correct: either (1) the Blandford–Payne process is somehow greatly suppressed (by a factor of 100 or more than suggested by Fig. 15.5) or (2) the Blandford–Znajek process is enhanced somehow by a similar factor. David Garofalo and his colleagues [629] looked at the first possibility and suggested that the black hole and disk magnetosphere interact, producing strong fields and jets for both processes when $j \rightarrow -1$, and weak ones when $j \rightarrow +1$. But no specific quantitative models for either this, or a Blandford–Znajek enhancement, exist at this time.

16.5.3.2 Why are Elliptical Galaxies So Radio-Loud?

Imagine the following scenario in the context of the above model for black hole-driven jets in AGN. A central MBH or SMBH has just endured a major episode of accretion in which the angular momentum of the accreted material overwhelmed the black hole and spun it *up* to high prograde spins. The AGN becomes radio-quiet in the process and eventually turns off as the fuel is exhausted, and the black hole is left spinning rapidly.

A few hundred million years later another accretion event besets the central black hole. If the hole is in a spiral galaxy, and the new material (gas or compact star clusters) comes from the gaseous disk, then the angular momentum of that material will likely be in the same sense as the previous event. The black hole, therefore, will experience a bright accretion event, but it will be *prograde* accretion. That is, according to Fig. 16.15 the AGN should be radio-quiet, or perhaps a weak FR I source for a short time when it nears the jet line in its hardness–intensity diagram.

On the other hand, if the black hole is in an elliptical galaxy, the source of gas will be from a merger (major or minor). Depending on the galaxy’s cluster environment, the merger may have a positive or negative impact parameter with respect to the black hole spin. A truly random impact parameter distribution might produce strongly negative impacts somewhat less than 50% of the time, but still quite often. Detailed simulations could determine a better estimate of this percentage. In this case the object will become a powerful FR II radio source in the above model. *So, a once radio-quiet AGN could suddenly become a radio-loud one, not because the black hole spin suddenly changed, but because the angular momentum vector of the new accreting material is now different.* However, in a Salpeter time or so, two important effects will cause the black hole to spin down (perhaps even with a sudden spin flip in the process [630]). One effect is the BZ process itself that uses the reducible mass to produce the powerful FR II jet (equation (14.27)). The other is the spindown torque of the retrograde accretion onto the black hole, which also will act on a time of order $\tau_{\text{salp}} \dot{m}^{-1}$. (See equation (16.37).) After the spindown, however, the accretion torque will spin the hole up to a radio-weak prograde state again.

Note that this rapid spindown/spinup that turns off the radio jet will take only 10^7 yr or less. Therefore, this is *not* the cosmic evolution of the radio galaxy/quasar population discovered by Maarten Schmidt (Section 2.3.5), which takes place on about a 1 Gyr time scale. Something much more universal, like the cosmic merger rate, must drive such longer-term events.

16.5.3.3 Observations of Acceleration, Collimation and Recollimation Features in M87 and Other Radio-Loud AGN

In Sections 15.1.4 and 16.2.2 we discussed how a jet that is launched by a rotating (electro-) magnetic field continues to be accelerated and collimated by that field and, ultimately, over-collimated enough to bounce or shock on the rotation axis. We also

suggested that the (fairly stationary) particle acceleration zone in Sera Markoff's model fits to jet broadband spectra may indeed be that recollimation feature. Here we additionally suggest that such recollimation features in nearby AGN jets actually may be resolved and imaged with VLBI and, in a few cases, even with the VLA. Detailed studies of the properties of recollimation, and the propagation of the jet beyond such features, could shed further light on the FR I–II dichotomy and perhaps on how other jets (*e.g.*, GRBs) turn Poynting flux into kinetic and particle energy.

Initial Acceleration and Collimation of the M87 Jet in the First $100 r_g$

Several investigators [20, 631] have suggested that the black hole in the VLBA image of M87 (Fig. 2.2, bottom) does not lie in the bright core but rather another $\sim 100 r_g$ or so toward the southeast. That is, the jet “turns on” only at the bright core distance because relativistic particles begin to be accelerated there and then radiate downstream as well in the strong, helical magnetic field.

This is an attractive suggestion for several reasons. It corroborates well with Sera Markoff's models of other black hole sources (in this case mostly X-ray binaries with stellar-mass black holes) which, nevertheless, have similar Eddington ratios ($\dot{m} \sim 10^{-4}$ – 10^{-3}); this provides yet another link amongst black holes of very different masses. The distance of $\sim 100 r_g$ also is similar to that Markoff finds in hard state X-ray binary jets [466]. The suggestion that the M87 jet turns on $\sim 100 r_g$ from the black hole also fits well with the recollimation property of the theoretical jet models discussed in Section 15.1.4 and Fig. 15.12. The implications are that jets, at least in these low-luminosity radio sources, are completely formed and causally disconnected by the time they reach $\sim 100 r_g$ from the central black hole engine.

Continued Acceleration of the M87 Jet over the Next Million r_g

However, VLBI observations of the M87 jet with the European VLBI Network in the range 5×10^4 to $5 \times 10^5 r_g$ by Keiichi Asada of the Astronomical Institute of Astrophysics (ASIAA) and his colleagues show (Fig. 16.16(left); [632]) that a significant increase in speed of the jet (up to $V_{\text{app}} \sim 6c$) continues out to $\sim 5 \times 10^5 r_g$ (or about 1000 mas), where the component HST-1 lies. This is consistent with a conical jet being accelerated, possibly by (magnetic) pressure gradients, as described in Sections 13.3.1 and 15.1.4. This is indeed a strange result. It indicates that the jet is *not* fully formed by $\sim 100 r_g$; instead, it continues to be accelerated over another four orders of magnitude in distance. So, while the jet may be causally disconnected from the central engine by $\sim 100 r_g$, *it is not causally disconnected within itself*; *i.e.*, it is not super-fast magnetosonic. The continued acceleration means that out to nearly $\sim 10^6 r_g$ the jet still remains causally *connected* along its length, probably by fast magnetosound waves.

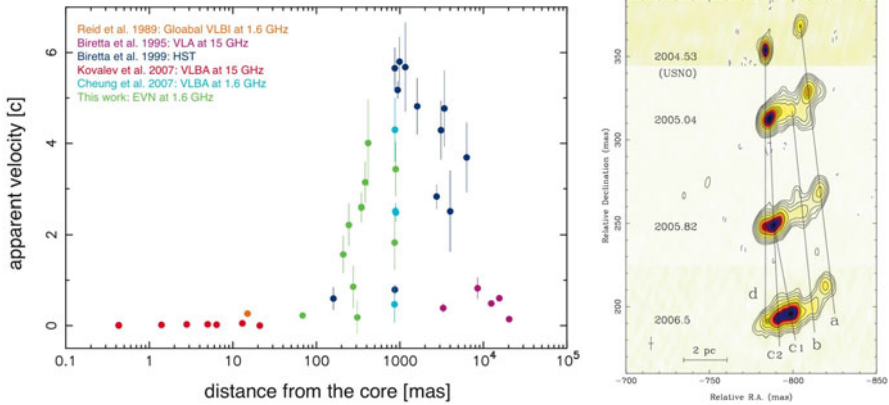


Fig. 16.16: Observational evidence that M87 may have a recollimation shock nearly $10^6 r_g$ from the black hole. *Left:* Apparent velocity data of M87 jet components from six different sources over five decades in distance from the black hole. Most of the acceleration occurs in the range 0.16–0.9 mas ($\sim 1.3\text{--}7 \times 10^5 r_g$), whereupon there appears to be a nearly stationary feature in the HST-1 complex (vertical blue, green, and cyan points). The jet subsequently resumes its nearly $6c$ speed, but eventually falters and becomes sub-relativistic. Reprinted from Fig. 1b of reference [632], with the permission of CUP. Copyright ©2006 Cambridge University Press. *Right:* Birth of high-speed (here, up to $4.3c$) components in HST-1 (not in the black hole core!) over two years’ time, similar to that in an MHD simulation of a recollimation shock (Fig. 15.27). Reprinted from Fig. 3 of reference [633]. Reproduced by permission of the AAS.

A Distant Recollimation Feature in M87 and Other AGN at a Million r_g

A Recollimation Feature at HST-1. Moreover, the activity *within* the HST-1 complex is very striking. Fig. 16.16(right) shows VLBA observations by Teddy Cheung (when at NRAO and Stanford), Dan Harris (at CfA), and Łukasz Stawarz (also at Stanford) of that region over a period of two years [633]. It appears that *superluminal* jet components were ejected from an essentially *stationary* core within HST-1; this occurred ~ 120 pc, nearly one million r_g , from the black hole! Component HST-1d is the stationary “core”, with a speed of $<0.25c$; HST-1c split into two components, HST-1c2 (closer to HST-1d) and HST-1c1 (farther out), that have apparent speeds of $V_{\text{app}} \sim 0.47c$ and $4.3c$, respectively; still farther out, HST-1b has a speed of $\sim 2.5c$; and the most distant knot also had a speed of $2.5c$, but decelerated to only $1.4c$ at the end of 2005. Furthermore, in addition to the radio outburst at HST-1, there also was flaring in both the near ultraviolet and X-ray bands over the period 2004–2006 [634]. Apparently, an ejection event from the central black hole at least 400 years ago recently affected the HST-1 complex this past decade.

In 2010 Masanori Nakamura (then at Space Telescope), along with David Garofalo and this author (both at JPL), proposed a radical new idea [635]: HST-1d is a *second recollimation feature* in the M87 jet, and its behavior in giving birth to new superluminal components mimics that simulated by us (K. Lind and colleagues [544] in 1989 and discussed in Section 15.2.2 and Fig. 15.24). Nakamura and col-

leagues even were able to model the HST-1c double-knot component as *a single MHD jet ejection event* having a Lorentz factor of $\gamma_{\text{jet}} \approx 6.2$, which produces both a forward MHD shock complex (actual speeds of $0.97c$ and $0.94c$ for the forward fast and slow mode shocks) and a reverse shock complex (actual speeds of $0.67c$ and $0.73c$ for the fast and slow reverse modes). Therefore, HST-1d, c2, and c1 all can be explained by a single recollimation MHD shock that ejects a single MHD jet event at $\gamma_{\text{jet}} \approx 6.2$.

Why would M87's jet suddenly collapse and be reborn a million r_g away from the black hole? Nakamura has a promising suggestion. The radius of influence of the $3 \times 10^9 M_\odot$ black hole is $r_h \sim 73 \text{ pc}$ (equation (11.20)). Interior to this radius the star and gas distribution should be cusp-like; indeed, M87 does have such a structure within ~ 1 arcsecond or $\sim 78 \text{ pc}$ (see Fig. 1.2). As the jet travels outward through the cusp region, it sees a decreasing external pressure gradient, expands laterally, and further accelerates (up to $V_{\text{app}} \sim 6c$). Beyond r_h the external star and gas density becomes almost uniform (Fig. 1.2 again), so the jet no longer expands laterally. Instead, magnetic pinch forces now cause the jet to begin collimating again, quickly forming another FMSS and collapsing onto the axis within $\sim 2 r_h = 146 \text{ pc}$ (close to the $\geq 120 \text{ pc}$ distance of HST-1).

There should be no surprise that a jet that extends over nine orders of magnitude in length could have more than one recollimation feature. The one near $\sim 100 r_g$ clearly reflects the launch conditions in the black hole accretion disk or magnetosphere. The HST-1 feature at $\sim 10^6 r_g$, however, tends to reflect the “galactic weather” conditions in the inner cusp and core regions.

The causality properties also should be qualitatively similar. The particle acceleration feature at $\sim 100 r_g$ should be causally disconnected from the central black hole engine. And the complex behavior of HST-1 should be causally disconnected from the conical jet and its acceleration out to the distance of HST-1.

A Recollimation Feature in BL Lac as well? In 2008 Alan Marscher, of Boston University, and his colleagues [608] used simultaneous observations of the blazar BL Lacertae in the radio, optical, X-ray, and γ -ray bands to show that it also has an inner helical jet structure that is terminated by a standing shock (see Fig. 16.17). The picture is surprisingly similar to that deduced by Markoff and Nowak [468]. However, Marscher's deduced shock distance from the black hole is $\sim 10^5 r_g$, similar to that of the HST-1 complex in M87, so these observations say nothing about any possible recollimation feature at $\sim 100 r_g$.¹⁴ In addition, Marscher's team was able to measure a 240° polarization swing in the optical data, indicating a possible shock feature rotating around a helical field. If we take BL Lac's black hole mass to be $2 \times 10^8 M_\odot$, the distance of the recollimation shock to be $\sim 2 r_h$, and the jet line-of-sight angle to be 7° , we obtain a projected distance of $2 \times 10^5 r_g$. The mea-

¹⁴ Several properties of BL Lac make it impossible with current technology to see components as small as an initial $\sim 100 r_g$ recollimation feature. BL Lac is nearly 20 times farther away than M87, its line-of-sight angle toward us is half that of M87 or less, and its black hole is a factor of ten or more smaller than M87's. So, whereas M87's $3000 r_g$ VLBA jet subtends nearly 30 mas , a $3000 r_g$ jet in BL Lac would subtend $< 70 \mu\text{as}$ on the sky.

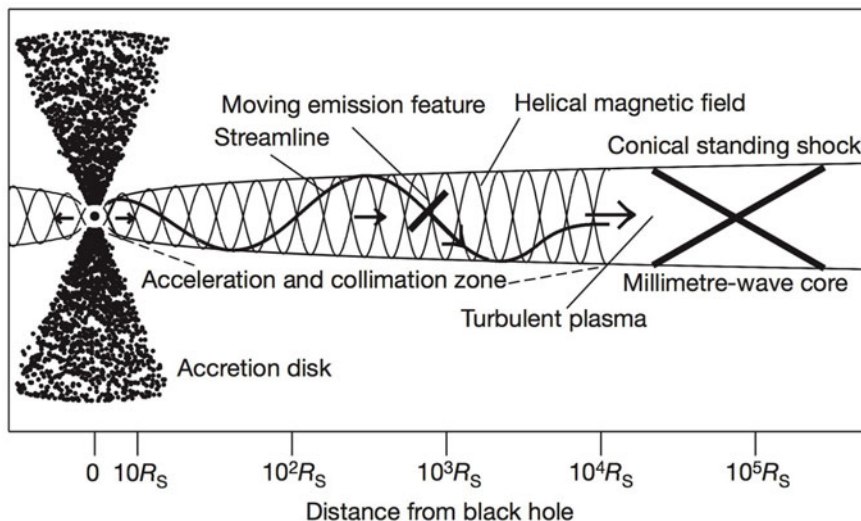


Fig. 16.17: Schematic diagram of the BL Lacertae jet, over several decades in distance from the black hole. The standing shock at $\sim 10^5 r_g$ is the recollimation feature that we discuss in the text. The picture bears a striking resemblance to that in Fig. 13.5, which was deduced by S. Markoff and M. Nowak from X-ray binary jet data. Reprinted by permission from Macmillan Publishers Ltd: Fig. 3 of reference [608], copyright (2008).

sured distance in Marscher's Fig. 1 is $7 \times 10^4 r_g$, or only a factor of ~ 3 lower. This is within the errors, given how uncertain black hole masses are for sources like BL Lac without strong emission lines. Since the Eddington ratio for BL Lac sources is $\dot{m} \sim 10^{-4} - 10^{-2}$ [636], similar to M87, it would be interesting to study additional low-luminosity radio sources to see if more such objects generally possess jet recollimation features just outside their black hole stellar cusps.

16.5.3.4 FR Is vs. FR IIs: Effect of Recollimation Features on Jet Propagation beyond One Million r_g

Because of their violent dynamical nature, as well as important causal properties, recollimation features have the ability to completely change the character of a supermagnetosonic jet and yet not at all affect the structure of the central black hole engine that produces the jet. Therefore, recollimation features may be related to the difference between FR I and FR II sources. But, there are many questions. How might such features affect jet propagation; which features might be more important in this regard (ones near the black hole at $\sim 100 r_g$ or farther out at $10^{5-6} r_g$); and where does this leave the popular Bicknell explanation that FR IIs are supersonic jets that bore through their galaxy's ISM while FR Is are similar, but less powerful, jets that are disrupted by the ISM and decelerated to subsonic speeds (Section 15.2.1)?

FR I Sources

First of all, this author still subscribes to the Bicknell model for the FR I–II break. It has nice properties that explain both the break qualitatively and quantitatively. However, while it is clear that the FR II sources are a fairly homogeneous class, it is quite possible that the FR I sources are not (see Section 15.2.1). The jet in M87, for example, clearly begins to fail and decelerate just beyond HST-1 (Fig. 16.16(left)). While Nakamura’s recollimation scenario is, technically, an effect due to the ISM (as in the Bicknell model), what appears to doom the jet is that it never becomes kinetic-flux-dominated (KFD). Instead, it remains magnetically (Poynting flux) dominated or PFD, undergoing helical kink instabilities like those in Poynting-dominated jet simulations (see Section 15.2.2 and Figs. 15.28 to 15.30) [569]. Perhaps this is indeed what happens in most FR I sources; perhaps it is not. In the Bicknell model the jet *does* become KFD, but without enough thrust to bore through the ISM before decelerating to subsonic speeds. It will take many high-resolution studies of FR I jets, and their magnetic properties, before we can determine whether there are one, two, or many different processes that produce an FR I morphology.

FR II Sources

Even though the FR II sources indeed may be KFD jets that *can* bore through their galaxies’ ISM, serious questions still remain. How did an outflow that was accelerated and collimated by strong rotating magnetic fields become kinetically dominated and magnetically weak? The answer suggested here is that FR II sources undergo a very strong, recollimation event very early in their outward propagation – so strong of a shock, in fact, that the magnetic field is dissipated into relativistic particle creation and acceleration [637]. Conversely, if the recollimation is not sufficiently strong, a “bounce” (instead of a strong shock) may occur that leaves the jet largely PFD out to great distances, as is the case in M87. How this conversion from PFD to matter and kinetic-flux domination occurs is not understood in detail right now. But it is clear that the jet will be left with a high, relativistic speed and with enough matter to be able to bore through the ISM and into the IGM very efficiently.

16.6 Summary: Black Holes in the Universe

The reader should not take the preceding discussion as a definitive explanation of all known black hole engines. Instead, it is meant as an example of the kind of discussion that should occur from now on: integrated consideration of black hole systems in all mass ranges, while comparing and contrasting their accretion and jet behaviors as a function of parameters like normalized mass $m = M_{\bullet}/M_{\odot}$, accretion rate $\dot{m}$, and spin j . The explanations offered here are plausible, but not necessarily widely accepted. Some even have not been submitted for publication yet.

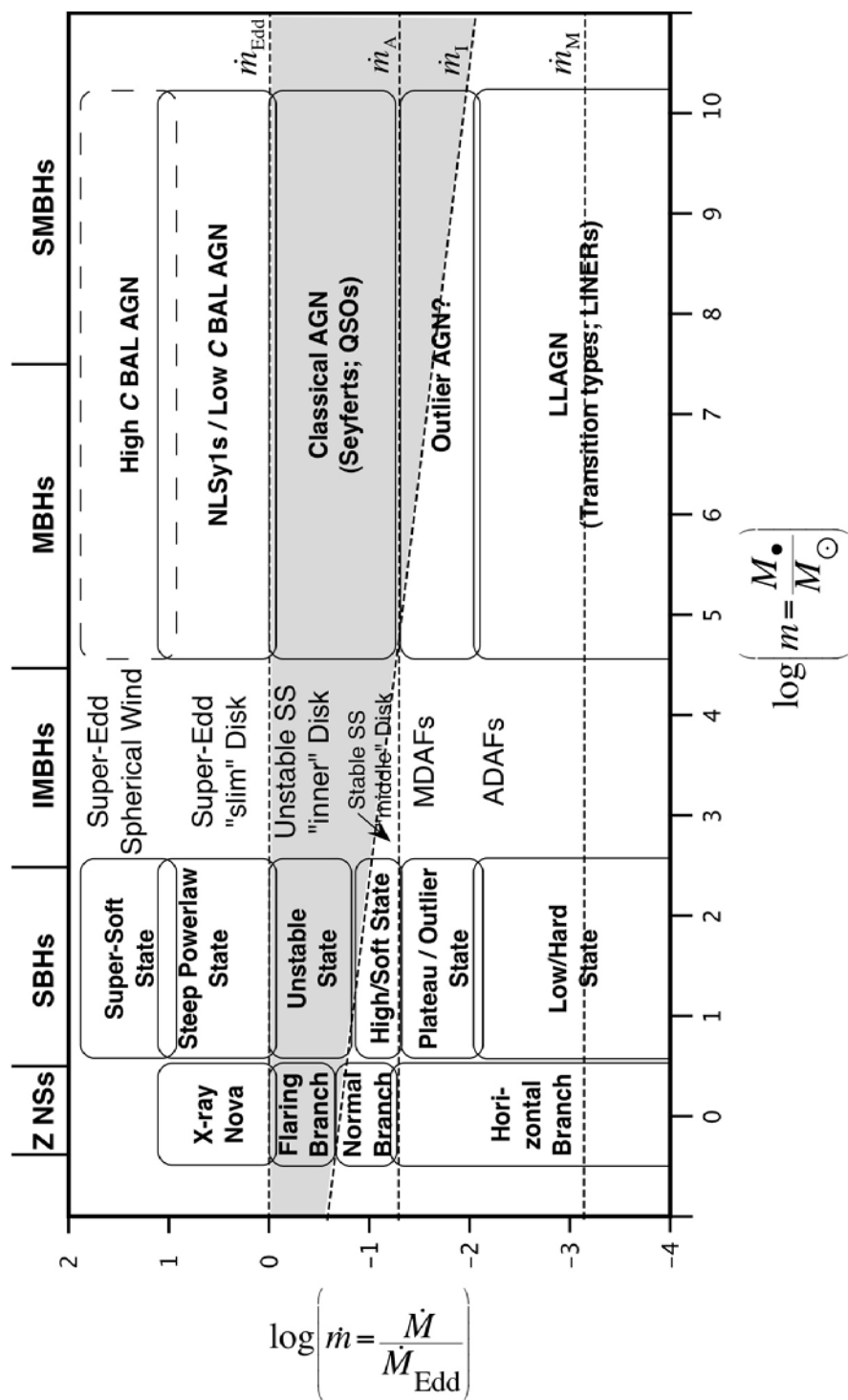


Fig. 16.18: *Previous page:* A theorist’s “principal component” diagram, showing ten decades in compact object mass and six decades in accretion rate. At top, mass ranges indicate object type: neutron star Z sources and stellar, intermediate mass, massive, and supermassive black holes. Boxes with rounded corners and bold lettering indicate proposed $\dot{m}$ location of different observed states of these objects. Lacking any concrete observations, the IMBH column is used to list the proposed theoretical models of accretion states depicted in Figs. 16.4, 16.6, 16.8, 16.9, 16.11, and 16.12 (which use $\alpha = 0.03$ and 0.3 for thin and thick disks, respectively). Dotted lines show accretion rates where different disk states appear: SS “middle” ($\dot{m} > \dot{m}_M$), SS “inner” ($\dot{m} > \dot{m}_I$), ADAF corona ($\dot{m} < \dot{m}_A$), and super-Eddington “slim” ($\dot{m} > \dot{m}_{\text{Edd}}$). Shaded area shows where an unstable SS “inner” region exists; it coincides with objects of high activity (flaring Z sources, black hole binaries with explosive jets, and most classical AGN). The $\dot{m}_A = 0.05$ line is extrapolated from black hole jet suppression points in Fig. 16.2 and is, therefore, highly uncertain. Broad absorption line AGN with high covering factors should exist in galaxies with rapid black hole growth, but they may lie in dust-enshrouded ULIRGs, requiring detailed mid-IR observations with JWST.

In Fig. 16.18 I offer a theorist’s “principal component diagram” with axes m and $\dot{m}$, analogous to Fig. 2.20(right). The mass axis extends from $1 M_\odot$ (the neutron stars) up to the most supermassive black holes known. The normalized accretion rate axis descends to very low rates ($\dot{m} = 10^{-4}$), but not all the way down to very low-luminosity AGN like Sgr A* ($\dot{m} \lesssim 10^{-6}$). That axis also extends well above unity. One always should be aware that, while $L/L_{\text{Edd}} \leq 1$, the amount of matter accreting *toward*, and sometimes into, a black hole can far exceed $\dot{M}_{\text{Edd}}$, depending on the efficiency of accretion and the outflows generated.

Except for the IMBH column, which has very few objects discovered so far, I have highlighted the location of a variety of phenomena in each category, based on theoretical models discussed in this chapter and preceding ones. Also delineated are accretion rate boundaries for the different accretion states listed in the IMBH column. The top horizontal line is $\dot{m}_{\text{Edd}} = 1$; and, while we have depicted the boundary between having and not having an ADAF corona $\dot{m}_A = 0.05$ as horizontal also, this may not be the case. More detailed studies of low-efficiency accretion flows as a function of $M_\bullet$ are needed. The division between having a stable Shakura–Sunyaev (SS) “middle” disk region near the black hole ($\dot{m} < \dot{m}_I$) and having an *unstable* “inner” region there ($\dot{m} > \dot{m}_I$), in fact, does have a weak dependence on black hole mass ($m^{-1/8}$) which, nevertheless, is important over the ten orders of magnitude in mass shown here. Finally, for accretion rates below the boundary $\dot{m}_M$ even the “middle” region disappears, leaving only an “outer” region possible in standard SS accretion theory. (Recall that, while X-ray observations may be dominated by an ADAF corona, this does not preclude an associated thin SS disk, even at low $\dot{m}$. See Fig. 16.4.)

The gray-shaded region shows where an active, but unstable, “inner” region should dominate the accretion flow near the black hole. Note: *many of the suggested explanations for observed black hole engines in this book critically depend on the secular and thermal instability analysis of the SS “inner” region being correct.* If detailed modeling and simulation of accretion flows with $\dot{m}_I < \dot{m} < 1$ show that such accretion is stable, then much of the nice picture discussed in the preceding

sections is incorrect. In its favor, however, the most active and unstable neutron star and black hole systems lie within this gray theoretical zone of instability.

Below are comments on each of the main mass categories, summarizing the problems and challenges that remain therein.

16.6.1 Neutron Stars

Even after a half-century of work, the most critical question in this area is why most massive stars leave neutron star remnants instead of black holes. We have made a case here (Section 16.1.1) that MHD jets from rotating proto-neutron stars assist (if not cause) ejection of the stellar envelope shortly after core collapse in a massive star, saving the nascent neutron star from being crushed. Simulations and detailed studies of this process should be pursued much more vigorously than at present.

Other important issues are:

- Dead isolated neutron stars within 40–160 pc might be detectable by the NuSTAR satellite as hard X-ray sources.
- In Fig. 16.18 we choose to depict only the Z sources – binary neutron stars with substantially larger accretion rates than isolated ones. We explain each of their various manifestations as interactions of different accretion states with the Z star’s moderately strong magnetic field. This model, presented in Section 16.1.3, depends critically on van der Klis’s classification of atoll sources as weak field neutron stars ($10^{8\pm1}$ G) and Z sources as moderately strong ones ($10^{10\pm1}$ G). If this classification turns out to be incorrect, the Z source model described herein will have to be reworked.
- Jets from binary neutron stars may be driven both by the accretion disk (Blandford–Payne mechanism) and by the rotating neutron star (propeller mechanism). The lack of much jet suppression at high $\dot{m}$ could be explained if the disk-driven jet dominates at low accretion rates and still is suppressed for $\dot{m} \gtrsim 0.08$, like black hole accretion disks. However, in the latter case the pulsar still can drive a propeller-driven jet no matter what the accretion rate, so the overall effect of disk jet suppression is minor, as in Fig. 16.2.

16.6.2 Stellar-mass Black Holes

In this author’s opinion, the most important question for black holes of all sizes is why very high-luminosity (super-Eddington) sources are radio-quiet. This seemingly minor issue actually has wide implications. In the stellar-mass black hole case, solution to this problem is key to understanding the steep power law (SPL) state, highly super-Eddington objects like SS433 and P098 in M101, and, in fact, jet production in general. In the massive and supermassive cases it is key to understand-

ing BAL AGN, NLSy1s, and perhaps the entire radio-loud/radio-quiet dichotomy. In the past this author has advocated a model where the disk vertical magnetic field, and therefore jet power, is dramatically smaller for cool, geometrically thin SS “middle” disks than for hot, geometrically thick ADAF flows. However, super-Eddington “slim” disks should be as geometrically thick as ADAFs and, therefore, have as large (or larger) magnetic field strengths. Yet, very high-luminosity sources, in which “slim” disks should be playing a role, all appear to be radio-quiet.

An alternative idea, in which jets still are produced but their internal plasma remains *thermal* (as in SS433), looks more promising than the vertical magnetic field strength explanation. For example, a jet may be launched, collimated, and reach the Alfvén speed, but never pass through its fast magnetosonic separatrix surface (Section 16.2.2) to produce relativistic speeds and a particle-accelerating recollimation feature.

Other important issues are:

- The problem of not forming neutron stars in core-collapse supernova simulations is also a problem for black holes, but in reverse. Discovering why nature creates so many neutron stars also will explain why so few black holes are created in these same events.
- It is *unlikely* that isolated stellar-mass black holes in the galaxy, accreting from the ISM, will be detected in the hard X-ray with NuSTAR; these sources are not great enough in number to be close by and too weak to be detected at their typical distances. However, a few might be detectable in the radio, even with that meager source of fuel.
- We adopt the classical explanation of low/hard state objects as advection-dominated accretion flows (ADAFs) and stable high/soft state sources as SS “middle” region accretion flows.
- We tentatively identify the SPL state as the stable super-Eddington “slim” disk flow (Section 16.2.2).
- We also suggest two new states exist which have not been identified yet:
 - The “unstable” state ($0.1 < \dot{m} < 1$), where the accretion flow should be radiation-pressure-dominated. The SS disk counterpart to this state is the “inner” accretion flow, which is unstable to secular and thermal instabilities. This state accounts for secular loops in the hardness–intensity diagram and thermal flickering or broad QPOs on thermal time scales.
 - The “super-soft” state ($1 \ll \dot{m}$), where the accretion flow drives a strong, optically thick super-Eddington wind that absorbs the hard X-rays emitted by the corona buried within the wind.
- We further suggest that the strange outlier sources may be a magnetically-dominated accretion flow (MDAF) state or sub-state, intermediate in $\dot{m}$ between the ADAF and SS “middle” region states. Compared to low/hard sources, which lie in the radio source fundamental plane, sources in an MDAF state are predicted to produce underluminous radio powers.
- Finally, there appears to be no strong evidence that the jets seen in X-ray binary systems are powered by rotation of the black hole. We do not dispute this con-

clusion here. However, it may be true simply because the evidence is very poor at the present time, or because black hole spin in binary systems are such that the Blandford–Znajek process never will be very strong. This conclusion is in sharp contrast to that for SMBH objects, for which there is ample evidence that black hole rotation plays a significant role (Section 16.5.3).

16.6.3 *Intermediate Mass Black Holes*

Given that very few IMBH candidates have been found, a surprising amount is known theoretically about how they might form, both in the present day and in the early universe. The author is particularly fond of the Portegies Zwart–McMillan process of forming an IMBH through massive star coalescence at the center of a compact star cluster. Compact clusters anchored by an IMBH may have important implications for growing and fueling AGN black holes. The major question in this area, then, is how often are IMBHs formed in this, and other ways, and how often are they accreted by a central MBH or SMBH, along with their attendant stars?

Other important issues are:

- Isolated IMBHs in the Galaxy, accreting from the ISM, are best discovered in the radio (as with isolated stellar-mass black holes).
- IMBHs with a $1\text{--}2\ M_{\odot}$ companion, in the giant phase and undergoing thermal mass transfer, still would be in the low/hard state and visible as hard X-ray sources, not only throughout our Galaxy, but also in nearby galaxies as well.
- IMBHs with a $10\ M_{\odot}$ B supergiant companion, also in thermal mass transfer, could reach near-Eddington accretion rates of $0.1 < \dot{m} < 1$. Even higher rates could be reached, but would require one or more high-mass O star companions, as would be the case when the IMBH is forming at a compact cluster center. All such objects would be most common in star-forming galaxies, but many may be hidden by surrounding dust. Therefore, large infrared telescopes, especially in space, would be the most promising instruments to discover and study such exotic IMBH objects.

16.6.4 *Massive Black Holes*

The most important challenge in the MBH range is to explain, from first principles (e.g., SS disk models), why these massive objects so naturally and routinely form Type 1 engines with central ionizing sources and broad line regions. Looking at Fig. 16.18 we see that Type 1 AGN all lie in the Shakura and Sunyaev instability strip. We, therefore, have outlined a scenario in which the thermal instabilities in the SS “inner” region create a two-phase accretion inflow, with the hot phase escaping in an ADIOS wind that carries, and confines, clouds of the cool phase. The wind

launched from near the “inner”–“middle” boundary is consistent with the optical BLR properties, while that launched from the central portion of the unstable “inner” region can account for the ultra-fast outflows (UFOs) seen in X-rays. Furthermore, the large range of thermal time scales available, even in a single object, is similar to the range of optical variability in typical QSOs. And, for larger black hole masses, the typical variability time scale should increase with mass (see Table 16.2).

The “inner” region is also subject to the Lightman–Eardley instability on much longer secular (accretion) time scales. Unlike in the stellar and intermediate mass black hole cases, in the MBH and SMBH cases, the secular instability operates on time scales much longer than typical observing times, or human lifetimes for that matter (20 yr – 20 Myr; Table 16.1). All observers should be aware that, when viewing QSOs, they may be looking at an unstable accretion flow whose properties may change dramatically thousands of years from now, just as X-ray binaries like GRS 1915+105 cycle through soft and hard substates, producing a jet every 20 minutes.

Another key challenge for MBHs is to explain how they are fed enough material to be fueled at, or above, the Eddington rate. We have presented in this book a scenario in which interstellar gas is stored temporarily in stars in compact clusters anchored by an IMBH and then quickly trucked into the sphere of influence of the MBH by tidal friction. There both clusters and stars are tidally disrupted by the MBH, turning those stars back into interstellar gas again, which finally accretes onto the black hole. This model can keep an MBH in a spiral galaxy supplied at near the Eddington limit and can explain naturally the presence of a star-formation ring a few hundred parsecs from the center of an active galaxy. More detailed work on this mechanism is needed, however, as is further scrutiny by the refereeing process. If it turns out to be viable, compact cluster inspiral may turn out to play an important role in the growth of massive and supermassive black holes, in galaxies both with and without merger activity.

Other important questions are:

- How do NLSy1s and BAL AGN fit into this picture? We argue above that both are near- or super-Eddington objects. In particular, BAL AGN cannot be normal AGN viewed from a favorable angle, but they can be super-Eddington AGN viewed favorably, along their axial, thermal jets.
- Why are MBHs in Seyfert galaxies less radio-loud than their SMBH counterparts in ellipticals? This is addressed a bit in the next subsection.

16.6.5 Supermassive Black Holes

The most important question in this greatest of mass ranges is, how can such an enormously massive black hole be supplied with enough fuel in the galactic nucleus that the hole accretes at, or above, the Eddington limit? The compact cluster inspiral method, which worked well for MBHs, cannot supply matter at nearly that great a rate in this case.

The remaining important questions in the SMBH range all center around jet production and radio loudness. What drives the jets emanating from SMBH systems, black hole spin or accretion disk rotation? Why are SMBHs so much more radio-loud than MBHs? Also high on the list, what causes some jets (*e.g.*, FR IIs) to convert Poynting flux efficiently to kinetic (and matter) flux while others (at least some FR Is) apparently remain Poynting-flux-dominated?

Much of the discussion below is based on Garofalo's new model [501] for producing powerful jets from spinning black holes (see Section 16.5.3). He has shown that the Blandford–Znajek power is drastically different when the accretion disk spin is aligned (prograde) with that of the black hole compared to when it is anti-aligned (retrograde). See also Fig. 16.15. The prograde case produces a relatively weak relativistic jet, while the jet in retrograde case is very strong.

Despite the lack of evidence for the Blandford–Znajek process in stellar black hole jets (Section 16.2.2), we argue above that there is significant evidence that SMBH jets are black hole spin-driven. In fact, there seems to be very little evidence that accretion disk jets are at work in AGN at all – just the opposite of the stellar black holes. This is actually a dilemma: hiding black hole spin-driven jets, while still having accretion disk-driven jets operating is relatively easy; but, hiding an accretion disk-driven jet while still producing a black hole spin-driven jet is very hard. Both types of jet require magnetized accreting matter, which usually will collect in a rotating, accretion inflow that should produce a jet. On the other hand, a black hole can be spun up to rapid prograde spins that produce only a weak jet. Naïvely, then, one would expect a strong, if not relativistic, jet from the disk and *maybe* a Blandford–Znajek jet as well – not systems that seem to have a strong BZ jet only.

A possible answer to the question of why there is a rapid increase in AGN radio loudness with black hole mass is that it is a consequence of the same retrograde/prograde black hole spin jet model. Accretion of spiral galaxy disk gas onto the central MBH will tend to relentlessly spin the hole up to a prograde system, producing a weak jet only; even successive accretion events likely will be prograde. A giant elliptical galaxy, on the other hand, with a much larger black hole, will experience random minor and major mergers, some of which will produce strong retrograde accretion events simply because of the possible difference in alignment between hole and accretion flow. At least qualitatively, then, it should come as no surprise that SMBHs are much more radio-loud statistically than MBHs.

Finally, a possible answer to the problem of converting Poynting-dominated jets to kinetic-dominated ones may lie in recollimation events that occur where the flow pinches itself onto the axis, creating a strong shock in the process. If the collimation shocks are strong enough, they could dissipate the magnetic field and create and/or accelerate particles, which eventually would lose internal energy via adiabatic and radiative processes. As shown in Fig. 15.27, a fast, but relatively cool, weakly-magnetized, and kinetic-dominated jet still would remain. If such a process is important in jets, it may have a strong connection to the FR I–II dichotomy as well. However, despite these early two-dimensional MHD simulations [544, 564] and some analytic models [637], there still is much work left to be done in this area.

Appendices

Appendix A

Mathematical Notation Used in this Book

This appendix gives a brief description of the mathematical notation used in this book, especially in Chapters 5 to 9. In all cases scalar quantities are presented in simple math mode. Examples are mass m or M ; fundamental constants c , G , *etc.*; and Greek letters γ , δ , ϕ , *etc.*. This does *not* include components of vectors and tensors, which are discussed below.

A.1 Vector and Tensor Notation for Two- and Three-Dimensional Spaces

Two and three-dimensional space notation is the same, as neither contains a time coordinate.

A.1.1 Two- and Three-Dimensional Vector and 1-Form Notation

Vectors in two and three dimensions are written as bold capital letters in math mode. Examples are momentum and velocity $\mathbf{P}$ and $\mathbf{V}$; electric and magnetic field $\mathbf{E}$ and $\mathbf{B}$; *etc.* Components of vectors are written in the same font as the geometrical quantity, but not in boldface and with a *upper* index indicating which component it represents. Examples: P^x , V^r , E^y , and B^z .

Roman subscripts are sometimes added to vectors or their components to distinguish which object or type of physics problem is being discussed. Examples are the velocity of object #2 $\mathbf{V}_2$ or orbital velocity in Schwarzschild geometry $V_{\text{orb,SH}}^\phi$.

1-Forms for the above quantities are written in bold lower-case letters, also in math mode. Examples are the momentum and velocity 1-forms $\mathbf{p}$ and $\mathbf{v}$; electric and magnetic 1-forms $\mathbf{e}$ and $\mathbf{b}$; *etc.* Components are of these written in the same

font as the 1-form geometrical quantity, but not in boldface and with a *lower* index. Examples: p_z , v_x , e_r , and b_y .

In flat two- and three-dimensional space, even when the coordinates are curvilinear, there usually is no distinction between vector and 1-form components. This is because they usually are written in a local orthonormal frame rather than a coordinate frame and because the metric in the that frame is simply the identity matrix. For example, in polar coordinates, the velocity V^θ is actually the linear velocity $V^{\hat{\theta}} = r d\theta/dt$, not the speed of the angular coordinate θ ($V^\theta = d\theta/dt$). Therefore, $v_{\hat{\theta}} = V^{\hat{\theta}}$, so there is no distinction between vectors and 1-forms.

A.1.2 Tensor Notation

Contravariant tensors in two and three dimensions are written as bold, *sans serif capital* letters. An example is the three-dimensional stress tensor $\mathbf{T}$. Components are written in non-bold sans serif capital letters, with upper indices: T^{xx} .

Covariant tensors in two and three dimensions also are bold, sans serif letters, but lower-case is used. The covariant version of the stress tensor is $\mathbf{t}$, and its components are written t_{xx} , t_{xy} , *etc.* The metric tensor $\mathbf{g}$ follows this form, but its contravariant version does not, because it is the inverse of the metric $\mathbf{G} = \mathbf{g}^{-1}$.

In order to limit confusion, mixed tensors are avoided in this book.

A.2 Vector and Tensor Notation for Four-Dimensional Spacetime

In order to emphasize the fundamental difference between three-vectors and four-vectors (and between three- and four-dimensional tensors), we use a different notation for spacetime quantities.

A.2.1 Vector and 1-Form Notation in Four-Dimensional Spacetime

As above, four-vector geometric quantities are written as bold capital letters. However, the font is a roman one, not math mode. Examples are four-momentum and four-velocity $\mathbf{P}$ and $\mathbf{U}$; four-current $\mathbf{J}$ and four-potential $\mathbf{A}$. Components also are non-bold characters of the same font and again with upper indices: P^x , U^r , J^y , and A^z .

Similarly, 1-form quantities in four dimensions are written as bold *lower-case* roman letters, with their components in non-bold roman font and having lower indices: p_z , u_x , j_r , and a_y .

A.2.2 Tensor Notation in Four-Dimensional Spacetime

Contravariant tensors in four dimensions are written as bold, *calligraphic* capital letters. Examples are the stress-energy tensor $\mathcal{T}$ and the Faraday and Maxwell tensors $\mathcal{F}$ and $\mathcal{M}$. The components are written as upper indices: $\mathcal{T}^{ww}$, $\mathcal{F}^{xy}$, $\mathcal{M}^{wz}$.

Because calligraphic letters are available only in upper-case, for covariant four-dimensional tensors we re-use the regular math mode letters. (There should be no ambiguity, as this notation is rarely used for three-dimensional 1-forms in this book. See above.) Examples are the metric g with components g_{xx} , *etc* or the Faraday 2-form f . As in three-space, the covariant forms are written with the same letter, but in lower-case. The exception again is the metric, because its contravariant form is also its inverse g^{-1} .

In the rare case when we need to express a tensor with more than two indices (*e.g.*, the Riemann tensor), we use the Fraktur font: $\mathfrak{R}$ or $\mathfrak{R}_{\alpha\beta\gamma\delta}$.

A.3 Miscellaneous Notation

Dipole and quadrupole moments often use similar letters to other physical quantities, so here we distinguish them by writing them in a Fraktur font. The same is true of other rare quantities used herein, including three-force per unit volume $\mathfrak{F}$ in a Fraktur font (as opposed to a simple force F) and four-force per unit volume $\mathfrak{F}$ in a script font.

Occasionally a vector or tensor of the same spatial dimensions and rank will be needed to indicate the *time and spatially-independent coefficients* of a sinusoidal wave function. In that case we use the same letter (lower or upper-case) with a slightly different, but related, font. Examples are the Euler font $\mathcal{A}$ for the coefficient of the three-vector potential A or the blackboard font $\mathbb{A}$ for the four-vector potential A .

Appendix B

Derivatives of Vectors and Tensors: Differential Geometry

Physics, and therefore astrophysics, is described by equations that involve the spatial and time derivatives of vectors and tensors. In order to properly describe physics in a curved spacetime (and even in flat spacetime that is spanned by curvilinear coordinates) we will need to understand how gradients and divergences of these quantities are calculated when the coordinates are *not* Euclidean or Minkowskian. This requires the mathematics of differential geometry, which is a broad subject. Here we provide only a very brief introduction.

Gradients and divergences will depend, of course, on how the metric changes with those coordinates. The derivatives of g are best embodied by a single function, called the “Christoffel symbol” or “connection coefficient”, and it takes into account all possible derivatives¹

$$\Gamma_{\lambda\mu\nu} \equiv \frac{1}{2} \left(\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} \right) \quad (\text{B.1})$$

The indices λ , μ , and ν range over the dimension of the space (*e.g.*, 0–3 for spacetime).

While the connection coefficients appear rather ugly, and the reader may be inclined to ignore them, they are of enormous importance in physics and even our daily lives. They give rise to

- Formulae for the gradient, divergence, and curl in curvilinear coordinates that are routinely used in fluid dynamics, plasma physics, mechanics, *etc.*
- Pseudo-forces like centrifugal and Coriolis forces, which arise because our frame of reference is accelerated.
- The force of gravity, which in Einstein’s view is also a pseudo-force, arising because of the curvature of space and time.

¹ There are a few places in this book where matrix notation is not adequate, so we shall revert to component notation. This is the case here, and in a few other places, where we must deal with quantities with more than two indices.

B.1 Covariant Gradients in Curved Spacetime

The gradient of a geometric quantity (scalar, vector, tensor) increases the number of indices on that quantity by 1: a scalar becomes a vector; a vector becomes a 2-tensor, *etc.* Using connection coefficients, the gradient can be written as follows:

$$\begin{aligned}
 (\nabla\Phi)_\gamma &= \frac{\partial\Phi}{\partial x^\gamma} \\
 (\nabla\mathbf{V}^\beta)_\gamma &= \frac{\partial V^\beta}{\partial x^\gamma} + \sum_{\lambda,\mu} g^{\beta\lambda} \Gamma_{\lambda\mu\gamma} V^\mu \\
 (\nabla\mathcal{T}^{\alpha\beta})_\gamma &= \frac{\partial \mathcal{T}^{\alpha\beta}}{\partial x^\gamma} + \sum_{\lambda,\mu} [g^{\alpha\lambda} \Gamma_{\lambda\mu\gamma} \mathcal{T}^{\mu\beta} + g^{\lambda\beta} \Gamma_{\lambda\mu\gamma} \mathcal{T}^{\alpha\mu}] \quad (\text{B.2})
 \end{aligned}$$

The sums in the above formulae are performed over the number of dimensions in the space or spacetime.

Similar formulae also exist for 1-forms, 2-forms, *etc.*

$$\begin{aligned}
 (\nabla\mathbf{v}_\beta)_\gamma &= \frac{\partial v_\beta}{\partial x^\gamma} - \sum_\mu \Gamma_{\beta\mu\gamma} v^\mu \\
 (\nabla t_{\alpha\beta})_\gamma &= \frac{\partial t_{\alpha\beta}}{\partial x^\gamma} - \sum_{\lambda,\mu} [g^{\mu\lambda} \Gamma_{\alpha\mu\gamma} t_{\lambda\beta} + g^{\mu\lambda} \Gamma_{\beta\mu\gamma} t_{\alpha\lambda}] \quad (\text{B.3})
 \end{aligned}$$

These are called “covariant” gradients, because they can be used in physical laws that are invariant under any coordinate transformation. Furthermore, the formulae are true for *any* metric and so are quite amazing, if rather messy.

We can simplify the look of them, using the matrix notation we used in Chapters 6 and 7 if we define a *simple* derivative ∂_γ to be the operator

$$\partial_\gamma \equiv \frac{\partial}{\partial x^\gamma}$$

∂_γ is similar to ∇_γ , but it takes the derivative of *only the components* of the geometric quantities, not of their unit vectors as well. We then can write equations (B.2) as

$$\begin{aligned}
 \nabla\Phi &= \partial\Phi \\
 \nabla\mathbf{V} &= \partial\mathbf{V} + [\mathbf{g}^{-1} \cdot \boldsymbol{\Gamma}] \cdot \mathbf{V} \\
 \nabla\mathcal{T} &= \partial\mathcal{T} + [\mathbf{g}^{-1} \cdot \boldsymbol{\Gamma}] \cdot \mathcal{T} + \mathcal{T} \cdot [\mathbf{g}^{-1} \cdot \boldsymbol{\Gamma}] \quad (\text{B.4})
 \end{aligned}$$

While these are much more intuitive than equations (B.3), they have much less computational power, because it is not clear over which indices the dot products occur.

B.2 Divergences in Curved Spacetime

The divergence of a geometrical quantity is simply the “contraction” of the gradient. This means that we set one upper index equal to the lower differentiating one and sum. Because it requires an upper index, the divergence applies only to vectors and tensors, not forms. Setting $\gamma = \beta$ in equations (B.2) and summing over β we have

$$\begin{aligned}\nabla \cdot \mathbf{V} &= \sum_{\beta} \left\{ \frac{\partial V^{\beta}}{\partial x^{\beta}} + \sum_{\lambda, \mu} g^{\beta\lambda} \Gamma_{\lambda\mu\beta} V^{\mu} \right\} \\ (\nabla \cdot \mathcal{T})^{\alpha} &= \sum_{\beta} \left\{ \frac{\partial \mathcal{T}^{\alpha\beta}}{\partial x^{\beta}} + \sum_{\lambda, \mu} [g^{\alpha\lambda} \Gamma_{\lambda\mu\beta} \mathcal{T}^{\mu\beta} + g^{\lambda\beta} \Gamma_{\lambda\mu\beta} \mathcal{T}^{\alpha\mu}] \right\} \quad (\text{B.5})\end{aligned}$$

However, these *can* be simplified considerably by introducing the determinant of the metric

$$g \equiv \|\mathbf{g}\| \quad (\text{B.6})$$

Equations (B.5) then become

$$\begin{aligned}\nabla \cdot \mathbf{V} &= \frac{1}{\sqrt{|g|}} \sum_{\beta} \frac{\partial (\sqrt{|g|} V^{\beta})}{\partial x^{\beta}} \\ (\nabla \cdot \mathcal{T})^{\alpha} &= \sum_{\beta} \left\{ \frac{1}{\sqrt{|g|}} \frac{\partial (\sqrt{|g|} \mathcal{T}^{\alpha\beta})}{\partial x^{\beta}} + \sum_{\lambda, \mu} [g^{\alpha\lambda} \Gamma_{\lambda\mu\beta} \mathcal{T}^{\mu\beta}] \right\} \quad (\text{B.7})\end{aligned}$$

Again, we can write these in our matrix notation as

$$\begin{aligned}\nabla \cdot \mathbf{V} &= \frac{1}{\sqrt{|g|}} \partial \cdot (\sqrt{|g|} \mathbf{V}) \\ \nabla \cdot \mathcal{T} &= \frac{1}{\sqrt{|g|}} \partial \cdot (\sqrt{|g|} \mathcal{T}) + [\mathbf{g}^{-1} \cdot \boldsymbol{\Gamma}] : \mathcal{T} \quad (\text{B.8})\end{aligned}$$

In three-dimensional curvilinear coordinates, the metric is often diagonal, with elements $g_{ii} = h_i^2$. The vector divergence in equations (B.7) then gives us the familiar form for the divergence in curvilinear coordinates

$$\nabla \cdot \mathbf{V} = \frac{1}{h_1 h_2 h_3} \sum_{i=1}^3 \frac{\partial (h_1 h_2 h_3 V^i)}{\partial x^i}$$

This equation is useful only for diagonal three-metrics; however, equations (B.7) work for any metric and so are much more powerful.

B.3 The Metric Has No Gradient or Divergence

If we plug the metric tensor g^{-1} into the tensor form of equations (B.2), (B.3), and (B.5), and use the definition of Γ (equation (B.1)), we find that

$$\begin{aligned}\nabla g^{-1} &= 0 \\ \nabla g &= 0\end{aligned}$$

and, therefore, $\nabla \cdot g = 0$. The metric, therefore, has no covariant derivative or divergence, even though its components do change, in general, as we move through spacetime. *The covariant derivative ∇ , therefore, picks up only those changes in a vector or tensor that are independent of the coordinates*, not those that are generated by motion or curvature. This is why ∇ is used to cast the equations of physics in a coordinate-invariant (covariant) form.

In fact, we can now return to equation (B.1) and see how it was derived in the first place. We begin by simply *imposing* the requirement that the metric have no gradient or divergence. The formula for $\Gamma_{\lambda\mu\nu}$, therefore, must involve a linear combination of all possible derivatives of the symmetric metric tensor g

$$\Gamma_{\lambda\mu\nu} \equiv a \frac{\partial g_{\lambda\mu}}{\partial x^\nu} + b \frac{\partial g_{\lambda\nu}}{\partial x^\mu} + c \frac{\partial g_{\mu\nu}}{\partial x^\lambda} \quad (\text{B.9})$$

If we now express ∇g in the general form (B.3), require it to be zero, require $\Gamma_{\lambda\mu\nu}$ to be symmetric in μ and ν , and solve for the coefficients a , b , and c , we recover the formula for Γ in equation (B.1).

Appendix C

Derivation of the Adiabatic Relativistic Stellar Structure Equations

This appendix derives the equations for the structure and evolution of a relativistic star in spherical symmetry. Because of this assumption, there will be no time-dependent quadrupole moment of the star's mass and, therefore, no gravitational radiation emitted. This derivation, while fairly basic in relativistic physics is a little beyond the scope of the main part of the book. Nevertheless, it is an important demonstration of the use of the Einstein field equations (7.21) in solving for the time evolution of a relativistic gravitational field and an excellent example of black hole formation. The equations presented here were first derived by Charles Misner (University of Maryland) and David Sharp (Princeton University) [638], and independently by Michael May and Richard White of the Lawrence Radiation Laboratory (now known as Lawrence Livermore National Laboratory or LLNL) [639]. We will follow May and White's work most closely.

C.1 The Spherical Metric in Mass Coordinates

We begin with the spherically symmetric metric written in a manner similar to the Schwarzschild metric

$$ds^2 = -\bar{a}^2 c^2 dt^2 + \bar{b}^2 dr^2 + r^2 d\Omega^2 \quad (\text{C.1})$$

where $d\Omega^2 = d\theta^2 + \sin^2 \theta d\phi^2$. Indeed, *outside* the star we *do* have Schwarzschild geometry, with $a = b^{-1} = (1 - r_S/r)^{1/2}$. However, inside the star $\bar{a}$ and $\bar{b}$ will be functions of time as well as of the radius r .

For many reasons, detailed in Section 5.2.2, it is more convenient to use a coordinate system in which the independent radial coordinate is the integral of *rest mass* from the center of the star to some fixed specific value

$$m \equiv \int_V \rho dV \quad (\text{C.2})$$

where dV is the *proper* volume (including the curvature of space) and ρ is the rest mass density only (no internal or kinetic energy included). The amount of rest mass m inside this point will remain fixed during the evolution, but the radius of that shell $r(m, t)$ will change with time. It is helpful in this gauge to think of the star as composed of concentric shells, each with a mass $dm = \rho dV$, which can (1) collapse under their own gravitational weight if there are no other forces, (2) press against one another to keep the star in equilibrium if there are restoring pressure forces, or (3) even explode outward if the pressure can overcome gravity. A given value of m sits at the outer edge of a given mass shell and follows that shell's motion as the star evolves.

Switching from r to the mass coordinate m , then, gives us a new metric

$$ds^2 = -a^2 c^2 dt^2 + b^2 dm^2 + r^2 d\Omega^2 \quad (\text{C.3})$$

where a , b , and r are functions of m and the new time variable t only (the spherically symmetric assumption). In particular, r will be the outer radius of the spherical mass shell at position m and radial mass width dm . (To be precise, r will be equal to $1/2\pi$ times the circumference of the m th spherical shell.)

Now, the proper three-dimensional volume element of each spherical shell in this metric is

$$dV = b dm r d\theta r \sin \theta d\phi$$

So equation (C.2) becomes

$$\begin{aligned} m &= \int_0^{2\pi} d\phi \int_0^\pi \sin \theta d\theta \int_0^m \rho r^2 b dm \\ &= \int_0^m 4\pi r^2 \rho b dm \end{aligned} \quad (\text{C.4})$$

Because the differential of both sides of equation (C.4) must be equal (*i.e.*, $dm = 4\pi r^2 \rho b dm$) we immediately can write down the metric coefficient of dm

$$b = \frac{1}{4\pi r^2 \rho} \quad (\text{C.5})$$

As with Newtonian stellar structure, the choice of mass coordinates automatically enforces the conservation of mass.

C.2 The Field Equations and Conservation Laws

The remainder of the metric coefficients (a and r) and the state variables (ρ and internal energy ε) are determined by two Einstein field equations and two conservation laws. We will use the stress-energy-momentum tensor written in the *rest frame*

of each mass shell (*i.e.*, in the moving $[m, t]$ coordinate system) (equation (6.69)). The $\mathcal{G}^{tt} = 8\pi G \mathcal{T}^{tt}/c^4$ Einstein field equation is, then

$$\frac{c^2}{2} \frac{\partial}{\partial m} \left\{ r \left[1 + \frac{1}{a^2 c^2} \left(\frac{\partial r}{\partial t} \right)^2 - \frac{1}{b^2} \left(\frac{\partial r}{\partial m} \right)^2 \right] \right\} = 4\pi G \left(\rho + \frac{\varepsilon}{c^2} \right) r^2 \frac{\partial r}{\partial m} \quad (\text{C.6})$$

and the $\mathcal{G}^{tm}$ component has no source

$$\frac{\partial^2 r}{\partial m \partial t} - \frac{1}{a} \frac{\partial a}{\partial m} \frac{\partial r}{\partial t} - \frac{1}{b} \frac{\partial b}{\partial t} \frac{\partial r}{\partial m} \quad (\text{C.7})$$

Finally, the t and m components of $\nabla \cdot \mathcal{T} = 0$ are

$$\frac{\partial}{\partial t} \left(\rho + \frac{\varepsilon}{c^2} \right) = -\xi \rho \left(\frac{1}{b} \frac{\partial b}{\partial t} + \frac{2}{r} \frac{\partial r}{\partial t} \right) \quad (\text{C.8})$$

$$\frac{\xi \rho c^2}{a} \frac{\partial a}{\partial m} = -\frac{\partial p}{\partial m} \quad (\text{C.9})$$

where we have defined the inertia per unit rest mass to be

$$\xi \equiv 1 + \frac{(\varepsilon + p)}{\rho c^2} \quad (\text{C.10})$$

It is helpful, both mathematically and physically, to define the *gravitational mass*

$$\mathfrak{M} \equiv \int_0^m 4\pi r^2 \left(\rho + \frac{\varepsilon}{c^2} \right) \frac{\partial r}{\partial m} dm \quad (\text{C.11})$$

which *does* include both the internal energy and the kinetic energy, as well as rest mass, inside the shell at coordinate m . It also is useful to to define two quantities

$$u \equiv \frac{1}{a} \frac{\partial r}{\partial t} \quad (\text{C.12})$$

$$W \equiv \frac{1}{b} \frac{\partial r}{\partial m} \quad (\text{C.13})$$

These will turn out to be the m -component of the four-velocity of the mass shell and a geometric factor W telling us how dm and $4\pi r^2 \rho dr$ are related. The coordinate derivatives of the gravitational mass $\mathfrak{M}$ have simple expressions

$$\frac{\partial \mathfrak{M}}{\partial m} = 4\pi r^2 \left(\rho + \frac{\varepsilon}{c^2} \right) \frac{\partial r}{\partial m} = \left[1 + \frac{\varepsilon}{\rho c^2} \right] W \quad (\text{C.14})$$

$$\frac{\partial \mathfrak{M}}{\partial t} = 4\pi r^2 \frac{p}{c^2} u \frac{\partial r}{\partial t} = -4\pi r^2 \frac{a p}{c^2} u \quad (\text{C.15})$$

The first of these follows directly from the mass derivative of equation (C.11). The second can be derived by taking the time derivative of that equation, folding in equations (C.7) to (C.10), and integrating over dm . (It also can be derived by computing the [redundant] $\mathcal{G}^{rr} = 8\pi G T^{rr}/c^4$ Einstein field equation, which is an equivalent amount of work.)

C.3 The Adiabatic, Relativistic Stellar Evolution Equations

We now are in a position to solve for some of the variables and put the equations in simple, familiar forms.

C.3.1 The Mass Shell Geometric Factor

Using equation (C.14) we can replace the right-hand side of equation (C.6) with $G \partial \mathfrak{M} / \partial m$ and integrate over dm . Applying definition (C.13), this gives

$$W^2 = 1 + \frac{u^2}{c^2} - \frac{2 G \mathfrak{M}}{c^2 r} \quad (\text{C.16})$$

If there were no gravity ($G = 0$), the geometric factor W would be simply the Lorentz factor. On the other hand, if we had gravity but no motion ($u = 0$), then W would have the Schwarzschild form $(1 - 2 G \mathfrak{M} / (c^2 r))^{1/2}$.

C.3.2 The Density Equation

Equations (C.5) and (C.12) can be differentiated with respect to t and m , respectively, and plugged into equation (C.7) to yield an equation analogous to the Newtonian density equation (5.58)

$$\frac{1}{\rho} \frac{1}{r^2} \frac{1}{a} \frac{\partial(\rho r^2)}{\partial t} = - \frac{\partial u / \partial m}{\partial r / \partial m} \quad (\text{C.17})$$

This looks exactly like the Newtonian version if we identify

$$\frac{1}{a} \frac{\partial}{\partial t} = \frac{\partial}{\partial \tau} \quad (\text{C.18})$$

as the time derivative in the rest frame of each mass shell (*cf.*, equation (C.12)).

C.3.3 Conservation of Energy Equation

If we substitute the mass metric coefficient (equation (C.5)) and its time derivative into the conservation of energy equation (equation (C.8)), we obtain

$$\frac{\partial \varepsilon}{\partial t} = \frac{(\varepsilon + p)}{\rho} \frac{\partial \rho}{\partial t} \quad (\text{C.19})$$

which is the first law of thermodynamics (equation (5.50)) with no heating or cooling (adiabatic flow). The lack of any heat flow terms ($\mathcal{T}^{tm}$) in the stress-energy tensor is where the adiabatic assumption was made. This ensures that the gas remains isentropic.

C.3.4 Equation of Motion in Mass Coordinates

Because the mass shell coordinate system is a moving (Lagrangian) one, it should come as no surprise that the conservation of momentum equation comes mainly from the Einstein field equation (the rr one) with only some help from equation (C.9), rather than the other way around. To derive it, we differentiate the equation for the geometric factor W with respect to time

$$W \frac{\partial W}{\partial t} = \frac{u}{c^2} \frac{\partial u}{\partial t} - \frac{G}{c^2 r} \frac{\partial \mathfrak{M}}{\partial t} + \frac{G \mathfrak{M}}{c^2 r^2} \frac{\partial r}{\partial t} \quad (\text{C.20})$$

We then substitute into the above equation an expression for $\partial W/\partial t$ derived from equations (C.5), (C.7), (C.9), and (C.13)

$$\frac{\partial W}{\partial t} = -\frac{4\pi r^2}{\xi c^2} \frac{\partial p}{\partial m} \frac{\partial r}{\partial t} \quad (\text{C.21})$$

plus the expressions for $\partial \mathfrak{M}/\partial t$, and $\partial r/\partial t$. The result, after multiplying by c^2/u , is

$$\boxed{\frac{1}{a} \frac{\partial u}{\partial t} = -4\pi r^2 \frac{W}{\xi} \frac{\partial p}{\partial m} - \frac{G (\mathfrak{M} + 4\pi r^3 p/c^2)}{r^2}} \quad (\text{C.22})$$

This is the relativistic version of the Newtonian conservation of momentum in mass coordinates (equation (5.61)). In the non-relativistic Newtonian limit, the mass contributed by pressure and internal energy in equation (C.22) will be negligible ($\xi \rightarrow 1$ and $4\pi r^3 p/\mathfrak{M} c^2 \rightarrow 0$). Also, because of equation (C.9), the lapse function a will become unity throughout the star, so $\partial \tau = \partial t$, and we will have $W \rightarrow 1$ as well. So the equation of motion does indeed reduce to the Newtonian one derived in Chapter 5.

C.3.5 Equation of Motion in Schwarzschild–Hilbert-like Coordinates

We also can write the relativistic equation of motion in terms the radial pressure gradient. It still will be in the Lagrangian frame of reference, but will appear more familiar.

Because r is a monotonic function of m at any time t , we can write

$$\frac{\partial p}{\partial m} = \frac{\partial p}{\partial r} \frac{\partial r}{\partial m} = \frac{\partial p}{\partial r} W b = \frac{W}{4\pi r^2 \rho} \frac{\partial p}{\partial r} \quad (\text{C.23})$$

So we now can convert equation (C.22) into one involving derivatives in proper time τ and the shell radius r

$$\boxed{\frac{\partial u}{\partial \tau} = - \left(1 + \frac{u^2}{c^2} - \frac{r_S}{r} \right) \frac{\partial p}{\partial r} - \left(1 + \frac{4\pi r^3 p}{\mathfrak{M} c^2} \right) \xi \frac{G \mathfrak{M} \rho}{r^2}} \quad (\text{C.24})$$

The relativistic corrections in equation (C.23) have the following interpretations: acceleration term (inertia due to internal energy and pressure must be included); pressure term (one factor of W comes from the $\partial/\partial r$ gradient and one comes from a Lorentz-like boost of the pressure itself); gravitational force (as pressure increases, its gravitational mass must be included, both in the enthalpy inertia and in the inertia of the mechanical force itself).

Appendix D

Derivation of the General Relativistic MHD Equations from Kinetic Theory

The equations of general relativistic magnetohydrodynamics, which play a central role in this book, can be derived from the general relativistic Boltzmann equation in a two-step process. First, we take velocity moments of that equation to generate the *multi-fluid* GRMHD equations. Then we perform weighted sums of those equations, over mass and over charge, to produce conservation laws for mass, charge, four-momentum, and four-current. The derivation presented here follows an article by D. Meier [346] on the generalized Ohm's law (conservation of current).

D.1 The Multi-Fluid Equations of General Relativistic Magnetohydrodynamics

D.1.1 The Zeroth Moment: Conservation of Particle Number

We begin by re-writing the general relativistic Boltzmann equation (9.2) as

$$\mathbb{U}_i \cdot \nabla \mathbb{N}_i + \mathbb{F}_i \cdot \nabla \mathbb{P} \mathbb{N}_i = \dot{\mathbb{N}}_{i, \text{coll}} \quad (\text{D.1})$$

where $\mathbb{U}_i = \mathbb{P}/m_i$ is a function (like $\mathbb{F}_i$), not a coordinate.¹ It can be shown [346] that the momentum integral of the second term on the left and of the collision term on the right vanish. And, because $\mathbb{X}$ and $\mathbb{P}$ are independent phase space coordinates, $\nabla \cdot \mathbb{U}_i \propto \nabla \cdot \mathbb{P} = 0$, so the zeroth velocity moment of the Boltzmann equation becomes simply

$$\nabla \cdot \iiint \mathbb{U}_i \mathbb{N}_i d^4 \mathbb{P} = 0 \quad (\text{D.2})$$

¹ As in Chapter 9, we use the blackboard font to indicate quantities pertaining to a given volume in eight-dimensional phase space ($\mathbb{X}, \mathbb{P}, \mathbb{U}, \mathbb{F}$) and the script font in six-dimensional phase space ($\mathcal{X}, \mathcal{P}, \mathcal{V}, \mathcal{F}$), while regular bold characters are used for average quantities in four- or three-dimensional physical space. See Appendix A.

This can be written in a more familiar form if we decompose $\mathbb{U}_i$ into an average center-of-mass velocity

$$\mathbf{U} \equiv \frac{\sum_i m_i \iiint \mathbb{U}_i \aleph_i d^4\mathbb{P}}{\sum_i m_i \iiint \aleph_i d^4\mathbb{P}} \quad (\text{D.3})$$

and the drift velocity $\mathbb{V}_i$, which is always orthogonal to $\mathbf{U}$, giving us

$$\mathbb{U}_i = \gamma_i (\mathbf{U} + \mathbb{V}_i) \quad (\text{D.4})$$

where the Lorentz factor for each volume of phase space is defined as

$$\gamma_i \equiv -\frac{1}{c^2} (\mathbf{U} \cdot \mathbb{U}_i) = (1 - \mathbb{V}_i \cdot \mathbb{V}_i)^{-1/2} \quad (\text{D.5})$$

The second half of equation (D.5) is true *if* we measure the components of $\mathbb{V}_i$ in the rest frame of the fluid. Note that, while $\mathbb{V}_i$ is formally a four-vector, because $\mathbf{U} \cdot \mathbb{V}_i = 0$, $\mathbb{V}_i$ has only three non-zero (spatial) components in the rest frame of the fluid. The w (time) component of $\mathbb{V}_i$ is zero and can be ignored.

Equation (D.2) now can be written as the conservation of particle species i

$$\boxed{\nabla \cdot \mathbf{n}_i (\mathbf{U} + \mathbf{V}_i) = 0} \quad (\text{D.6})$$

where the particle density of species i is

$$\mathbf{n}_i = \iiint \gamma_i \aleph_i d^4\mathbb{P} = \iiint f_i d^3\mathcal{P} \quad (\text{D.7})$$

and the average drift velocity for that species is

$$\mathbf{V}_i = \frac{1}{\mathbf{n}_i} \iiint \gamma_i \mathbb{V}_i \aleph_i d^4\mathbb{P} = \frac{1}{\mathbf{n}_i} \iiint \mathbb{V}_i f_i d^3\mathcal{P} \quad (\text{D.8})$$

(Here we have used the relation between $\aleph_i$ and f_i (equation (9.4)) and have performed the integral over the mass shell, as discussed in Section 9.1.) With these definitions we see that equation (D.3) implies that the mass weighted drift velocity vanishes.

$$\sum_i \mathbf{n}_i m_i \mathbf{V}_i = 0 \quad (\text{D.9})$$

D.1.2 The First Moment: Conservation of Particle Four-Momentum

The first velocity moment of the general relativistic Boltzmann equation can be obtained by first multiplying equation (D.1) by $\mathbb{U}_i$. This produces a vector Boltzmann

equation

$$\nabla \cdot (\mathfrak{N}_i \mathbf{U}_i \mathbf{U}_i) + \mathbf{U}_i \left[\frac{q_i}{m_i c} (\mathbf{U}_i \cdot \mathbb{F}_i) \cdot \nabla_{\mathbb{P}} \mathfrak{N}_i \right] = \mathbf{U}_i \dot{\mathfrak{N}}_{i, \text{coll}}$$

With $\mathbb{F}_i$ given by equation (9.3), the integral of this equation over momentum four-space yields the conservation of four-momentum for particles of species i

$$\boxed{\begin{aligned} \nabla \cdot [\mathfrak{n}'_i \mathbf{U}_i \mathbf{U}_i + \mathfrak{n}_i \mathbf{U} \mathbf{V}_i + \mathfrak{n}_i \mathbf{V}_i \mathbf{U} + \mathbf{\Pi}_i] \\ = \frac{1}{m_i c} \mathbf{J}_i \cdot \mathcal{F} - \nu \mathfrak{n}_i (\mathbf{U} + \mathbf{V}_i) \end{aligned}} \quad (\text{D.10})$$

where we now see two *new* averaged quantities: the relativistic particle density

$$\mathfrak{n}'_i \equiv \iiint \gamma_i^2 \mathfrak{N}_i d^4 \mathbb{P} = \iiint \gamma_i f_i d^3 \mathcal{P}$$

and the *beamed* drift velocity

$$\mathbf{V}'_i \equiv \frac{1}{\mathfrak{n}_i} \iiint \gamma_i^2 \mathbb{V}_i \mathfrak{N}_i d^4 \mathbb{P} = \frac{1}{\mathfrak{n}_i} \iiint \gamma_i \mathbb{V}_i f_i d^3 \mathcal{P}$$

We also now have a definition of the partial electric current contributed by each particle species

$$\mathbf{J}_i \equiv q_i \mathfrak{n}_i (\mathbf{U} + \mathbf{V}_i)$$

and the partial pressure tensor

$$\mathbf{\Pi}_i \equiv \iiint \gamma_i^2 (\mathbb{V}_i \mathbb{V}_i \mathfrak{N}_i) d^4 \mathbb{P} = \iiint \gamma_i (\mathbb{V}_i \mathbb{V}_i f_i) d^3 \mathcal{P}$$

Equations (D.6) and (D.10) are the general relativistic *multi-fluid* MHD equations for each particle species density $\mathfrak{n}_i$ and velocity $(\mathbf{U} + \mathbf{V}_i)$. These equations do *not* close (*i.e.*, have the same number of equations as unknowns), because $\mathbf{\Pi}_i$ involves the *second* velocity moment, which we have not computed yet. There are only two ways to continue with the calculation and compute $\mathbf{\Pi}_i$: (1) compute the second velocity moment of equation (D.1) (which will only perpetuate the problem by producing an equation that needs the *third* velocity moment) or (2) assume a known equilibrium form for $f_i(\mathcal{P})$, which allows us to explicitly calculate $\mathbf{\Pi}_i$ (and also $\mathfrak{n}'_i$). We choose the second method and further assume that $f_i(\mathcal{P})$ is isotropic over the solid angle in momentum three-space

$$\int \int f_i d^3 \mathcal{P} = f_i 4 \pi \mathcal{P}^2 d\mathcal{P} \equiv \frac{d\mathfrak{n}_i}{d\mathcal{P}} d\mathcal{P}$$

The partial pressure tensor then becomes diagonal

$$\Pi_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & p_i & 0 & 0 \\ 0 & 0 & p_i & 0 \\ 0 & 0 & 0 & p_i \end{pmatrix}$$

where the partial pressure is given by

$$p_i = \frac{1}{3} \int \mathcal{P} \mathcal{V}_i \frac{d\mathbf{n}_i}{d\mathcal{P}} d\mathcal{P} \quad (\text{D.11})$$

since $\mathcal{P} = \gamma_i m_i \mathcal{V}_i$.

D.2 The One-Fluid Equations of General Relativistic Magnetohydrodynamics

The most popular form of the MHD equations eliminates all reference to individual particle species i and considers the system to be composed of a single neutral fluid in which currents are generated by the collective drift of charge.

D.2.1 Conservation of Rest Mass and Four-Momentum

If we multiply equations (D.6) and (D.10) by the particle mass m_i and then sum over all species i , we reduce these equations to the familiar forms

$$\nabla \cdot (\rho \mathbf{U}) = 0 \quad (\text{D.12})$$

$$\nabla \cdot \mathcal{T}_{\text{GAS}} = \frac{1}{c} \mathbf{J} \cdot \mathcal{F} \quad (\text{D.13})$$

The gas stress-energy tensor obtained from this sum is given by

$$\mathcal{T}_{\text{GAS}} = \left(\rho + \frac{\varepsilon}{c^2} \right) \mathbf{U} \mathbf{U} + \frac{1}{c^2} [\mathbf{Q} \mathbf{U} + \mathbf{U} \mathbf{Q}] + p \mathcal{P} \quad (\text{D.14})$$

where the *projection tensor* is given by

$$\mathcal{P} = \frac{1}{c^2} \mathbf{U} \mathbf{U} + g$$

the *rest mass* density is given by

$$\rho \equiv \sum_i \mathbf{n}_i m_i$$

the total internal energy and total pressure are given by the sum over all particle species

$$\varepsilon = \sum_i \varepsilon_i \quad p = \sum_i p_i$$

and the heat flux four-vector is given by

$$\mathbf{Q} = \sum_i \mathbf{n}_i m_i c^2 \mathbf{V}'_i = \sum_i \mathbf{n}_i m_i c^2 (\mathbf{V}'_i - \mathbf{V}_i)$$

(The second equality is valid because of equation (D.9).) In addition to the partial pressure p_i , we now have an integral expression for the internal (kinetic) energy of a relativistic gas

$$\varepsilon_i = (\mathbf{n}'_i - \mathbf{n}_i) m_i c^2 = m_i c^2 \iiint (\gamma_i - 1) f_i d^3\mathcal{P}$$

or

$$\boxed{\varepsilon_i = \int \varepsilon_{iK} \frac{d\mathbf{n}_i}{d\mathcal{P}} d\mathcal{P}} \quad (\text{D.15})$$

where $\varepsilon_{iK} \equiv (\gamma_i - 1) m_i c^2$ is defined as the particle kinetic energy. Since the right-hand side of equation (D.13) can be written as the divergence of a tensor

$$\frac{1}{c} \mathbf{J} \cdot \mathcal{F} = -\nabla \cdot \mathcal{T}_{\text{EM}}$$

where $\mathcal{T}_{EM}$ is given by equation (6.119). Equation (D.13) then becomes, simply,

$$\nabla \cdot \mathcal{T} = 0$$

where $\mathcal{T} = \mathcal{T}_{\text{GAS}} + \mathcal{T}_{\text{EM}}$. This exercise, therefore, shows us specifically how to calculate the stress-energy tensor for a conducting fluid in an electromagnetic field. It is this $\mathcal{T}$ that must be inserted into Einstein's field equations (7.21) in order to generate the evolution equations for the gravitational field.

D.2.2 Conservation of Charge and Four-Current

Another summation over all particle species can be done if we instead multiply equations (D.6) and (D.10) by the particle charge q_i before summing. We then obtain the conservation of charge and of four-current

$$\boxed{\nabla \cdot \mathbf{J} = 0} \quad (\text{D.16})$$

$$\boxed{\nabla \cdot \mathcal{C} = \frac{\omega_p^2}{4\pi} \left[\frac{1}{c} (\mathbf{U} + h\mathfrak{J}) \cdot \mathcal{F} - \eta_q (\rho_q \mathbf{U} + \mathfrak{J}) \right]} \quad (\text{D.17})$$

where the spatial current four-vector is given by

$$\mathfrak{J} \equiv \mathbf{J} - \rho_q \mathbf{U}$$

and is orthogonal to the four-velocity

$$\mathbf{U} \cdot \mathfrak{J} = 0$$

The *total* four-current density and charge density are defined as

$$\begin{aligned} \mathbf{J} &\equiv \sum_i \mathbf{J}_i \\ \rho_q &\equiv \sum_i q_i \mathbf{n}_i \end{aligned}$$

The charge-current tensor looks similar to the stress-energy tensor

$$\mathcal{C} = \left(\rho_q + \frac{\varepsilon_q}{c^2} \right) \mathbf{U} \mathbf{U} + \mathbf{U} \mathfrak{J}' + \mathfrak{J}' \mathbf{U} + p_q \mathcal{P} \quad (\text{D.18})$$

and the *beamed* spatial current is defined as

$$\mathfrak{J}' \equiv \sum_i q_i \mathbf{n}_i \mathbf{V}'_i$$

In addition to ρ_q , some other new charge-weighted thermodynamic quantities appear, like the charge-weighted internal energy

$$\begin{aligned} \varepsilon_q &\equiv \sum_i q_i c^2 \iiint (\gamma_i - 1) f_i d^3\mathcal{P} \\ &= \sum_i q_i c^2 \int (\gamma_i - 1) \frac{d\mathbf{n}_i}{d^3\mathcal{P}} d^3\mathcal{P} \\ &= \sum_i \frac{q_i}{m_i} \varepsilon_i \end{aligned}$$

and the charge-weighted pressure

$$\begin{aligned}
 p_q &\equiv \sum_i \frac{q_i}{3 m_i} \int \mathcal{P} \mathcal{V} \frac{d\mathbf{n}_i}{d\mathcal{P}} d\mathcal{P} \\
 &= \sum_i \frac{q_i}{m_i} p_i
 \end{aligned}$$

Finally, we see new plasma state variables, like the plasma frequency

$$\omega_p = \left[4 \pi \sum_i \frac{q_i^2 \mathbf{n}_i}{m_i} \right]^{1/2}$$

the electrical resistivity

$$\eta_q \equiv 4 \pi \frac{\nu}{\omega_p^2}$$

and the coefficient of the Hall-effect term

$$h \equiv \frac{4 \pi}{\omega_p^2 |\mathfrak{J}|} \sum_i \frac{q_i}{m_i} |\mathfrak{J}|$$

which is related to the classical Hall coefficient R_H as

$$R_H = \frac{h}{\eta_q c} |\mathbf{B}|$$

($|\mathbf{B}|$ being the strength of the magnetic field).

Equation (D.17) is often referred to as the generalized Ohm's law. When the left-hand side is small compared to the terms on the right (*i.e.*, when ω_p^{-1} is much smaller than other time scales in the system), we obtain the classical Ohm's law, with the Hall term

$$\frac{1}{c} (\mathbf{U} + h \mathfrak{J}) \cdot \mathcal{F} = \eta_q \mathbf{J}$$

(which eventually reduces to the well-known $V = IR$ Ohm's law taught in freshman physics and electrical engineering classes). If the time-dependent terms on the left-hand side (in the four-divergence of $\mathcal{C}$) are *not* negligible, the generalized Ohm's law shows how the current evolves toward its equilibrium value given above.

Appendix E

Derivation of the General Relativistic Grad–Schlüter–Shafranov Equation

The Grad–Schlüter–Shafranov equation is a general statement of force-free degenerate electrodynamics (FFDE) under the assumptions of a steady state and axisymmetry. The non-relativistic version is used to study the structure of Tokamak and other terrestrial and solar system fields, while the relativistic version is used to study the electrodynamics of pulsars and black holes. Here I re-derive the GSS equation in the Kerr metric using the general relativistic notation employed in this book, with the electrodynamic definitions of $\mathbf{B}$, $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{H}$ given in Section 9.5.1 (adopted from Komissarov [322]), to arrive at the version published by Uzdensky [486] (but with $c \neq G \neq 1$; see also [640] for the Schwarzschild case). Equation (E.11) below is useful for analyzing the magnetospheres of black holes in Kerr spacetime, as well as those of pulsars in a flat metric.

E.1 The Magnetic Induction Equation

In the Kerr metric in Boyer–Lindquist coordinates, under the assumption of axisymmetry ($\partial/\partial\phi = 0$) Maxwell’s solenoidal condition ($\nabla \cdot \mathbf{B} = 0$) gives the following *poloidal* magnetic induction

$$\mathbf{B}_p = \frac{1}{R} \nabla \Psi \times \mathbf{e}_{\hat{\phi}} \quad (\text{E.1})$$

where $\mathbf{B}_p = \mathbf{B}_p(B_{\hat{r}}, B_{\hat{\theta}}, 0)$ and the cylindrical radius is given by

$$R \equiv \frac{\Sigma \sin \theta}{\rho}$$

and Σ and ρ are the usual Kerr area and radius parameters. The magnetic flux is a function of the poloidal coordinates

$$\Psi(r, \theta) = \frac{1}{2\pi} \int \mathbf{B} \cdot d\mathbf{S} = R A_{\hat{\phi}} \quad (\text{E.2})$$

integrated over a disk surface with radius R that is centered on, and normal to, the rotation axis. $A_{\hat{\phi}}$ is the azimuthal component of the three-vector potential $\mathbf{A}$. The goal of this derivation is to find a single partial differential equation for the magnetic flux function Ψ . Then, using secondary equations we will be able to derive all other electromagnetic quantities ($\mathbf{B}$, $\mathbf{H}$, $\mathbf{D}$, and $\mathbf{E}$) from Ψ .

The *azimuthal* component of the magnetic induction comes from Ampère's law (with $\partial/\partial t = 0$), integrated over the same surface $d\mathbf{S}$

$$B_{\hat{\phi}} = \frac{H_{\hat{\phi}}}{\alpha} = -\frac{I}{\alpha R c} \quad (\text{E.3})$$

where $H_{\hat{\phi}}$ is the ϕ component of the magnetic field, α is the Kerr lapse function, and the current distribution function is

$$I = -\frac{1}{2} \int \mathbf{J} \cdot d\mathbf{S} \quad (\text{E.4})$$

Because I and Ψ are integrals over the same surface, I is a function of Ψ , or

$$I = I(\Psi)$$

Together, equations (E.1) and (E.3) give the complete magnetic induction three-vector for steady-state, axisymmetric FFDE

$$\boxed{\mathbf{B} = \frac{1}{R} \nabla \Psi \times \mathbf{e}_{\hat{\phi}} - \frac{I}{\alpha R c} \mathbf{e}_{\hat{\phi}}} \quad (\text{E.5})$$

E.2 The Electric and Magnetic Field Equations

The degeneracy condition of FFDE ($\mathbf{B} \cdot \mathbf{D} = 0$; see Section 9.5.2) implies that the electric displacement three-vector must be (equation (9.126))

$$\mathbf{D} = -\frac{1}{\alpha c} (\mathbf{V}_f - \alpha \boldsymbol{\beta} c) \times \mathbf{B}$$

where $\boldsymbol{\beta}$ is the Kerr drift vector. (This is also the ideal form of Ohm's law.) Furthermore, in order for $\mathbf{B}$ and $\mathbf{D}$ to be perpendicular, the field velocity $\mathbf{V}_f$ must be in the ϕ direction only. So we now can define a field *angular* velocity to be

$$\Omega_f \equiv \frac{|\mathbf{V}_f|}{R}$$

Plugging in $\mathbf{B}$ from equation (E.5) we obtain $\mathbf{D}$ in terms of Ψ

$$\mathbf{D} = -\frac{(\Omega_f - \omega)}{\alpha c} \nabla \Psi \quad (\text{E.6})$$

where ω is the angular velocity in the Kerr metric. We also get $\mathbf{E}$ in terms of Ψ

$$\mathbf{E} = -\frac{\Omega_f}{c} \nabla \Psi \quad (\text{E.7})$$

Note that, as implied by equations (7.67) $\mathbf{D}$ and $\mathbf{E}$ are the electric field measured in the rotating and fixed frames, respectively. As they both depend on $\nabla \Psi$ only, they are both poloidal functions only.

Finally, now that we know $\mathbf{D}$, we can calculate the magnetic field from the right-hand part of equation (9.124) as

$$\mathbf{H} = \frac{\alpha}{R} \left[1 + \frac{R^2}{\alpha^2 c^2} \omega (\Omega_f - \omega) \right] \nabla \Psi \times \mathbf{e}_{\hat{\phi}} - \frac{I}{R c} \mathbf{e}_{\hat{\phi}} \quad (\text{E.8})$$

E.3 The Charge and Current Densities

From Gauss's law, we can immediately determine the charge density

$$\rho_q = -\frac{1}{4\pi c} \nabla \cdot \left[\frac{(\Omega_f - \omega)}{\alpha} \nabla \Psi \right] \quad (\text{E.9})$$

Calculation of the current density is a little more tricky: it is best done by computing the poloidal ($\mathbf{J}_p$) and toroidal ($J_{\hat{\phi}} \mathbf{e}_{\hat{\phi}}$) components separately and then combining the results.

Because $I = I(\Psi)$, and therefore

$$\nabla I = \frac{dI}{d\Psi} \nabla \Psi$$

then the poloidal current $\mathbf{J}_p$ must be parallel to the poloidal magnetic field $\mathbf{B}_p$. From equations (E.2) and (E.4), this proportionality must be

$$\mathbf{J}_p = -\frac{1}{4\pi} \frac{dI}{d\Psi} \mathbf{B}_p$$

which gives us the poloidal component of the current density.

The toroidal component is found by dotting $\mathbf{e}_{\hat{\phi}}$ into a version of equation (9.122), also with $\partial/\partial t = 0$, to obtain

$$J_{\hat{\phi}} = \frac{c}{B^2} \rho_q (\mathbf{E} \times \mathbf{B})_{\hat{\phi}} + \frac{1}{B^2} (\mathbf{B} \cdot \mathbf{J}) B_{\hat{\phi}}$$

The total current density, then, is the vector sum of both components

$$\mathbf{J} = \rho_q \Omega_f R e_{\hat{\phi}} - \frac{1}{4\pi} \frac{dI}{d\Psi} \mathbf{B} \quad (\text{E.10})$$

That is, the current is the sum of that flowing along the twisted magnetic field (second term) plus those charges that are dragged around as the magnetic field rotates (first term).

E.4 The GSS Equation

The only independent equation that we have not yet incorporated into this discussion is the ϕ component of Ampère's law (right-hand equation (9.123)). (The r and θ components of Ampère's law are redundant with the definition of the current I .) Because the spatial part of the Kerr metric in Boyer–Lindquist coordinates is diagonal, equation (E.8) gives

$$(\nabla \times \mathbf{H})_{\hat{\phi}} = -R \nabla \cdot \left(\frac{c}{R} \nabla \Psi \right)$$

Inserting this and $4\pi J_{\hat{\phi}}/c$ into the ϕ component of Ampère's law, and noting that $\mathbf{D}$ has no ϕ component and $\partial/\partial t = 0$, and combining some terms, gives us the general relativistic version of the GSS equation

$$\begin{aligned} \nabla \cdot \left\{ \frac{\alpha}{R^2} \left[1 - \frac{R^2}{\alpha^2 c^2} (\Omega_f - \omega)^2 \right] \nabla \Psi \right\} + \\ \frac{(\Omega_f - \omega)}{\alpha c^2} \frac{d\Omega_f}{d\Psi} (\nabla \Psi)^2 + \frac{1}{2\alpha R^2 c^2} \frac{dI^2}{d\Psi} = 0 \end{aligned} \quad (\text{E.11})$$

where we have used the relation

$$\nabla \Omega_f(\Psi) = \frac{d\Omega_f}{d\Psi} \nabla \Psi$$

The GSS equation, plus appropriate boundary conditions, gives us an equation for the scalar potential Ψ , under the assumptions of time independence and axisymmetry, in the stationary and axisymmetric Kerr metric.

Appendix F

Derivation of the Equations for Stationary, Axisymmetric Ideal SRMHD in Newtonian Gravity

Ideal, stationary, axisymmetric magnetohydrodynamics is the main method for treating the acceleration and collimation of jets in black hole systems. This appendix begins with the standard ideal MHD vector equations, given in Section 9.5.1, and shows how the assumptions of stationarity and axisymmetry simplify these to the ones used in Section 9.5.6. The discussion generally will follow that in Mestel's 1961 paper on the subject [641], but we will do the derivations using the relativistic equations. While we will retain the possibility of flow near the speed of light, we will assume a Newtonian gravitational field only (*i.e.*, $GM/(Rc^2) \ll 1$), with no appreciable frame dragging (Kerr drift vector $\beta = 0$). The actual use of the resulting equations to study jet acceleration and collimation is in Section 15.1.

The MHD derivations here are the counterpart to the force-free electrodynamic ones given in Appendix E, except there we retained the possibility of a Kerr black hole metric.

F.1 The Axisymmetric, Stationary Equation(s) Parallel to the Magnetic Field

Under the assumptions of stationarity and axisymmetry magnetohydrodynamics generates two main differential equations, one perpendicular to the magnetic field (as in force-free electrodynamics) and a new one parallel to the field. The new equation describes the flow of plasma along the magnetic field lines, and was not needed in FFDE (where we ignored the matter entirely). We will derive the pieces of this equation below from Maxwell's laws of electromagnetism and from conservation laws for fluid flow. Then we will discuss the cross-field equation in the MHD case.

F.1.1 Faraday's and Ohm's Laws and Conservation of Mass: The Frozen-in Magnetic Field

The time-independent form of Faraday's law (9.103) states that

$$\nabla \times \mathbf{E} = 0$$

or $\mathbf{E} = \nabla\Phi + \mathbf{E}_0$, where Φ is the scalar electric potential and $\mathbf{E}_0$ is a vector uniform in space and constant in time. With the ideal Ohm's law (9.105), Faraday's law becomes

$$\nabla \times (\mathbf{V} \times \mathbf{B}) = 0 \quad (\text{F.1})$$

Now, let us decompose $\mathbf{V}$ and $\mathbf{B}$ into poloidal and toroidal components

$$\mathbf{V} = \mathbf{V}_p + \mathbf{V}_t \quad \mathbf{B} = \mathbf{B}_p + \mathbf{B}_t$$

where, for example, in cylindrical coordinates in flat space we have

$$\mathbf{V}_p = V_{\hat{R}} \mathbf{e}_{\hat{R}} + V_{\hat{Z}} \mathbf{e}_{\hat{Z}} \quad \mathbf{V}_t = V_{\hat{\phi}} \mathbf{e}_{\hat{\phi}}$$

Equation (F.1) then can be written as two equations, one in the poloidal plane and one in the toroidal direction

$$0 = [\nabla \times (\mathbf{V} \times \mathbf{B})]_p = \nabla \times (\mathbf{V}_p \times \mathbf{B}_p) = -c \nabla \times (E_{\hat{\phi}} \mathbf{e}_{\hat{\phi}}) \quad (\text{F.2})$$

$$0 = [\nabla \times (\mathbf{V} \times \mathbf{B})]_t = \nabla \times [(\mathbf{V}_p \times \mathbf{B}_t) + (\mathbf{V}_t \times \mathbf{B}_p)] \quad (\text{F.3})$$

Note that equation (F.2) has only one term because $\mathbf{V}_t \times \mathbf{B}_t = V_{\hat{\phi}} B_{\hat{\phi}} (\mathbf{e}_{\hat{\phi}} \times \mathbf{e}_{\hat{\phi}}) = 0$.

F.1.1.1 The Poloidal Velocity – Magnetic Field Relation

First, we will examine the *poloidal* equation (F.2). It has the solution

$$E_{\hat{\phi}} = [\nabla\Phi]_{\hat{\phi}} = \frac{1}{R} \frac{\partial\Phi}{\partial\phi}$$

But the axisymmetric assumption means that $\partial/\partial\phi = 0$ for any function, so $E_{\hat{\phi}} = 0$. That is, $\mathbf{V}_p$ is parallel to $\mathbf{B}_p$

$$\mathbf{V}_p = \mathcal{K} \mathbf{B}_p \quad (\text{F.4})$$

where $\mathcal{K}$ is a scalar function of (R, Z) .

F.1.1.2 The Toroidal Velocity – Magnetic Field Relation and the Field Angular Velocity

Next, we will examine the *toroidal* component (F.3). In cylindrical coordinates this can be written as

$$\frac{\partial}{\partial R} \left[B_{\hat{R}}(V_{\hat{\phi}} - \mathcal{K} B_{\hat{\phi}}) \right] + \frac{\partial}{\partial Z} \left[B_{\hat{Z}}(V_{\hat{\phi}} - \mathcal{K} B_{\hat{\phi}}) \right] \quad (\text{F.5})$$

We now can combine this equation with the axisymmetric version of the solenoidal condition

$$\nabla \cdot \mathbf{B}_p = \frac{1}{R} \left[\frac{\partial(R B_{\hat{R}})}{\partial R} + \frac{\partial(R B_{\hat{Z}})}{\partial Z} \right] = 0 \quad (\text{F.6})$$

to produce simply

$$\mathbf{B}_p \cdot \nabla \left(\frac{V_{\hat{\phi}} - \mathcal{K} B_{\hat{\phi}}}{R} \right) = \mathbf{B} \cdot \nabla \left(\frac{V_{\hat{\phi}} - \mathcal{K} B_{\hat{\phi}}}{R} \right) = 0$$

This means that the gradient of the quantity above in the parentheses is zero along a given magnetic field line. That is, the following is constant along each field line

$$\frac{V_{\hat{\phi}} - \mathcal{K} B_{\hat{\phi}}}{R} = \text{constant} \equiv \Omega_f \quad (\text{F.7})$$

which we identify as the angular velocity of the magnetic field line Ω_f .

Why is $V_{\hat{\phi}}$ not equal to $R \Omega_f$? The reason is that, if the magnetic field has a toroidal component ($B_{\hat{\phi}}$), then no matter what the field rotation rate plasma can flow freely in the $\mathbf{e}_{\hat{\phi}}$ direction at the velocity $\mathcal{K} B_{\hat{\phi}}$, *i.e.*, with the same proportionality as in the poloidal direction. So the total toroidal velocity of the plasma is, therefore,

$$\mathbf{V}_t = V_{\hat{\phi}} \mathbf{e}_{\hat{\phi}} = (\mathcal{K} B_{\hat{\phi}} + R \Omega_f) \mathbf{e}_{\hat{\phi}} \quad (\text{F.8})$$

If there were no field rotation ($\Omega_f = 0$), then plasma would flow along the field line with the same proportionality in all dimensions. On the other hand, if there were no matter (as in FFDE), then $\mathcal{K} = 0$ and $V_{\hat{\phi}}$ simply would be equal to $R \Omega_f$.

Combining equations (F.4) and (F.8), we find that the total three-velocity and total magnetic field are related as

$$\mathbf{V} = \mathcal{K} \mathbf{B} + R \Omega_f \mathbf{e}_{\hat{\phi}}$$

F.1.1.3 Determining the Proportionality Constant

We now can determine the value of $\mathcal{K}$ by considering the conservation of mass equation (9.100)

$$0 = \nabla \cdot (\gamma \rho \mathbf{V}) = \nabla \cdot (\gamma \rho \mathcal{K} \mathbf{B}) + \frac{1}{R} \frac{\partial(R \Omega_f)}{\partial \phi}$$

If we again apply both the axisymmetry and solenoidal conditions, the conservation of mass reduces to the conservation of another scalar along a field line $\mathbf{B} \cdot \nabla(\gamma \mathcal{K} \rho) = 0$. For mathematical purposes we define this scalar to be another constant k divided by 4π

$$\gamma \mathcal{K} \rho = \text{constant} \equiv \frac{k}{4\pi} \quad (\text{F.9})$$

Note that k is *not* unitless; it is the ratio of the constant local poloidal mass flux ($4\pi \gamma \rho \mathbf{V} \cdot d\mathbf{S}_p$) to the constant local poloidal magnetic flux ($\mathbf{B} \cdot d\mathbf{S}_p$), where $d\mathbf{S}_p$ is a small poloidal area vector. The final combination of the laws of Faraday, Ohm, and mass conservation yields the axisymmetric, stationary frozen-in condition

$$\mathbf{V} = \frac{k}{4\pi \gamma \rho} \mathbf{B} + R \Omega_f \mathbf{e}_{\hat{\phi}} \quad (\text{F.10})$$

F.1.2 Conservation of Specific Angular Momentum

We now will use the toroidal component of the momentum equation (9.101), which also makes use of Gauss's (9.117) and Ampère's (9.116) laws, to derive a third scalar constant along a magnetic field line – the angular momentum per unit mass or *specific* angular momentum. The full axisymmetric, stationary vector equation of motion in a Newtonian gravitational potential ψ is

$$\nabla \cdot \mathbf{T} = - \left(\gamma \rho + \frac{\mathfrak{E}}{c^2} \right) \nabla \psi \quad (\text{F.11})$$

The component of this along $\mathbf{e}_{\hat{\phi}}$ has no gravitational force

$$\frac{1}{R} \frac{\partial [R(R \mathbf{T}_{\hat{\phi}\hat{R}})]}{\partial R} + \frac{\partial (R \mathbf{T}_{\hat{\phi}\hat{Z}})}{\partial Z} = 0 \quad (\text{F.12})$$

where the two components of the stress tensor are determined from equation (9.113)

$$R \mathbf{T}_{\hat{\phi}\hat{i}} = \frac{k}{4\pi} \left[\gamma \left(1 + \frac{\mathfrak{h}}{c^2} \right) R V_{\hat{\phi}} - \frac{R B_{\hat{\phi}}}{k} \right] B_i$$

$\mathfrak{h} \equiv h/\rho$ is the enthalpy per unit mass, and $i = (Z, R)$. (Recall that $E_{\hat{\phi}} = 0$ because of axisymmetry.) The momentum equation can be combined with the solenoidal condition again to obtain the conservation of another scalar quantity along a magnetic field line

$$\gamma \left(1 + \frac{\mathfrak{h}}{c^2} \right) R V_{\hat{\phi}} - \frac{R B_{\hat{\phi}}}{k} = \text{constant} \equiv \ell \quad (\text{F.13})$$

which we identify as the angular momentum per unit mass of the plasma ℓ .

F.1.3 Conservation of Specific Entropy

The assumption of an adiabatic equation of state in equation (9.112) leads to a fourth quantity that is conserved along a field line: the entropy per unit mass

$$\frac{\mathcal{S}}{\rho} \propto \frac{p}{\rho^\Gamma} = K_\Gamma = \text{constant} \quad (\text{F.14})$$

While entropy must remain constant along a given field line, different field lines can have different values for K_Γ .

F.1.4 Conservation of Specific Energy

The final equation along each magnetic field line is the conservation of energy per unit mass. The master energy equation (9.102) in Newtonian gravity is

$$\nabla \cdot [c^2 (\mathfrak{P} - \gamma \rho \mathbf{V})] = -\mathfrak{P} \cdot \nabla \psi \quad (\text{F.15})$$

with the axisymmetric, poloidal momentum given by

$$\mathfrak{P} = \frac{k}{4\pi} \left[\gamma \left(1 + \frac{\mathfrak{h}}{c^2} \right) - \frac{R B_{\hat{\phi}}}{k c^2} \Omega_f \right] \mathbf{B}_p$$

Dropping one term in equation (F.15) that is proportional to ψ^2/c^4 and using the solenoidal condition, we obtain a fourth conserved quantity: the Bernoulli constant (specific total energy)

$$(\epsilon - 1) c^2 + \epsilon \psi = \text{constant} \equiv Be \quad (\text{F.16})$$

where ϵc^2 is the total specific internal energy of the plasma, including rest mass

$$\epsilon = \gamma \left(1 + \frac{\mathfrak{h}}{c^2} \right) - \frac{R B_{\hat{\phi}}}{k c^2} \Omega_f \quad (\text{F.17})$$

When (F.16) is combined with (F.10), (F.13), and (F.14), we can eliminate ρ , $V_{\hat{\phi}}$, $B_{\hat{\phi}}$, and p , producing a single equation that relates poloidal velocity to poloidal

magnetic field. This energy equation is essentially the equation of motion along each magnetic field line and is governed by the five free parameters of the problem:

- Ω_f : field line angular velocity;
- k : plasma mass loading of the field line;
- ℓ, K_Γ, Be : specific angular momentum, entropy, and total energy of the plasma.

F.2 The Axisymmetric, Stationary Equation(s) Normal to the Magnetic Field

We have only two remaining equations in the set (9.100) to (9.117) to consider: the R and Z components of the equation of motion (F.11). The projection of these vector components parallel to the poloidal magnetic field is essentially the equation along the field described above. The projection *normal* to the field is the cross-field equation that is analogous to the GSS equation derived in Appendix E for force-free electrodynamics. Formally, the component normal to $\mathbf{B}$ and in the poloidal plane is

$$(\mathbf{e}_{\hat{\phi}} \times \mathbf{b}) \cdot \left[\nabla \cdot \mathbf{T} + \left(\gamma \rho + \frac{\mathfrak{E}}{c^2} \right) \nabla \psi \right] = 0$$

where $\mathbf{b} \equiv \mathbf{B}_p/|\mathbf{B}_p|$. In cylindrical coordinates this equation becomes

$$\boxed{\begin{aligned} & \frac{b_{\hat{Z}}}{R} \frac{\partial(R \mathsf{T}_{\hat{R}\hat{R}})}{\partial R} - \frac{b_{\hat{R}}}{R} \frac{\partial(R \mathsf{T}_{\hat{R}\hat{Z}})}{\partial R} + b_{\hat{Z}} \frac{\partial \mathsf{T}_{\hat{R}\hat{Z}}}{\partial Z} - b_{\hat{R}} \frac{\partial \mathsf{T}_{\hat{Z}\hat{Z}}}{\partial Z} \\ & - b_{\hat{Z}} \frac{\mathsf{T}_{\hat{\phi}\hat{\phi}}}{R} + \left(\gamma \rho + \frac{\mathfrak{E}}{c^2} \right) \left(b_{\hat{Z}} \frac{\partial \psi}{\partial R} - b_{\hat{R}} \frac{\partial \psi}{\partial Z} \right) = 0 \end{aligned}} \quad (\text{F.18})$$

where

$$\begin{aligned} \mathsf{T}_{\hat{R}\hat{R}} &= \left[p + \frac{1}{8\pi} (B^2 + E^2) \right] + \gamma^2 \left(\rho + \frac{h}{c^2} \right) V_{\hat{R}}^2 - \frac{1}{4\pi} (B_{\hat{R}}^2 + E_{\hat{R}}^2) \\ \mathsf{T}_{\hat{R}\hat{Z}} &= \gamma^2 \left(\rho + \frac{h}{c^2} \right) V_{\hat{R}} V_{\hat{Z}} - \frac{1}{4\pi} (B_{\hat{R}} B_{\hat{Z}} + E_{\hat{R}} E_{\hat{Z}}) \\ \mathsf{T}_{\hat{Z}\hat{Z}} &= \left[p + \frac{1}{8\pi} (B^2 + E^2) \right] + \gamma^2 \left(\rho + \frac{h}{c^2} \right) V_{\hat{Z}}^2 - \frac{1}{4\pi} (B_{\hat{Z}}^2 + E_{\hat{Z}}^2) \\ \mathsf{T}_{\hat{\phi}\hat{\phi}} &= \left[p + \frac{1}{8\pi} (B^2 + E^2) \right] + \gamma^2 \left(\rho + \frac{h}{c^2} \right) V_{\hat{\phi}}^2 - \frac{1}{4\pi} B_{\hat{\phi}}^2 \end{aligned}$$

The cross-field equation (F.18) becomes quite complex as we substitute in the expressions for T_{ij} , so we will not do that here. Specific cases of the cross-field equation, and their implications for jet production in black hole engines, are discussed in Chapter 15.

Appendix G

Physical and Astrophysical Constants Used in this Book

Table G.1: Physical constants used in this book

Name	Symbol	Value in cgs/Gaussian units ^a	Value in SI units ^a
Avogadro’s number	N_A	$6.02214 \times 10^{23} \text{ mol}^{-1}$	$6.02214 \times 10^{23} \text{ mol}^{-1}$
Boltzmann constant	k	$1.38065 \times 10^{-16} \text{ erg } K^{-1}$	$1.38065 \times 10^{-23} \text{ J } K^{-1}$
Charge on electron	e	$4.80321 \times 10^{-10} \text{ esu }^b$	$1.60218 \times 10^{-19} \text{ C}$
Electronvolt	e	$1.60218 \times 10^{-12} \text{ erg}$	$1.60218 \times 10^{-19} \text{ J}$
Gas constant	$\mathcal{R}$	$8.31446 \times 10^7 \text{ erg mol}^{-1} K^{-1}$	$8.31446 \text{ J mol}^{-1} K^{-1}$
Gravitation constant	G	$6.6738 \times 10^{-8} \text{ erg cm g}^{-2}$	$6.6738 \times 10^{-11} \text{ J m kg}^{-2}$
Mass of electron	m_e	$9.1094 \times 10^{-28} \text{ g}$	$9.1094 \times 10^{-31} \text{ kg}$
Mass of proton	m_p	$1.6726 \times 10^{-24} \text{ g}$	$1.6726 \times 10^{-27} \text{ kg}$
Planck’s constant	h	$6.62607 \times 10^{-27} \text{ erg s}$	$6.62607 \times 10^{-34} \text{ J s}$
Planck’s constant ($/2\pi$)	$\hbar$	$1.05457 \times 10^{-27} \text{ erg s}$	$1.05457 \times 10^{-34} \text{ J s}$
Radiation constant	a	$7.5658 \times 10^{-15} \text{ erg cm}^{-3} K^{-4}$	$7.5658 \times 10^{-16} \text{ J m}^{-3} K^{-4}$
Speed of light	c	$2.997925 \times 10^{10} \text{ cm s}^{-1}$	$2.997925 \times 10^8 \text{ m s}^{-1}$
Stefan–Boltzmann constant	σ	$5.6704 \times 10^{-5} \text{ erg s}^{-1} \text{ cm}^{-2} K^{-4}$	$5.6704 \times 10^{-8} \text{ W m}^{-2} K^{-4}$
Thomson cross-section	σ_T	$6.65246 \times 10^{-25} \text{ cm}^2$	$6.65246 \times 10^{-29} \text{ m}^2$

^a Source: National Institute of Standards and Technology (<http://physics.nist.gov/cuu/Constants/>).
^b Note: 1 esu = $2997924580 \sqrt{4\pi \epsilon_0} \text{ C}$.

Table G.2: Astrophysical constants used in this book

Name	Symbol	Value in cgs/Gaussian units ^a	Value in SI units ^a
Astronomical unit	AU	$1.496 \times 10^{13} \text{ cm}$	$1.496 \times 10^{11} \text{ m}$
Light year	ly	$9.461 \times 10^{17} \text{ cm}$	$9.461 \times 10^{15} \text{ m}$
Parsec	pc	$3.086 \times 10^{18} \text{ cm}$	$3.086 \times 10^{16} \text{ m}$
Solar mass	$M_\odot$	$1.989 \times 10^{33} \text{ g}$	$1.989 \times 10^{30} \text{ kg}$
Solar luminosity	$L_\odot$	$3.839 \times 10^{33} \text{ erg s}^{-1}$	$3.839 \times 10^{26} \text{ W}$
Solar radius (average)	$R_\odot$	$6.955 \times 10^{10} \text{ cm}$	$6.955 \times 10^8 \text{ m}$
jansky (unit of radiative flux)	Jy	$10^{-19} \text{ erg s}^{-1} \text{ Hz}^{-1}$	$10^{-26} \text{ W Hz}^{-1}$

^a Source: International Astronomical Union
(http://www.iau.org/science/publications/proceedings_rules/units/).

References

1. Pacholczyk, A.G., *Radio Astrophysics* (Freeman, San Francisco, 1970).
2. De Young, D.S., *The Physics of Extragalactic Radio Sources* (University of Chicago Press, Chicago, 2001).
3. Osterbrock, D.E. & Ferland, G.J., *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (University Science Books, Mill Valley, CA, 2005).
4. Rybicki, G.B. & Lightman, A.P., *Radiative Processes in Astrophysics* (John Wiley & Sons, Inc., Chicago, 2001).
5. Krolik, J.H., *Active Galactic Nuclei* (Princeton Series in Astrophysics, Princeton, 1998).
6. Lewin, W. & van der Klis, M., eds., *Compact Stellar X-ray Sources* (Cambridge Univ. Press, Cambridge, 2006).

Chapter 1

7. Eddington, A.S., *Mon. Not. R. Astron. Soc.*, **77**, 596, (1917).
8. Ghez, A., *et al.*, *Astroph. J.*, **689**, 1044–1062, (2008).
9. Sargent, W.L.W., *et al.*, *Astroph. J.*, **221**, 731, (1978).
10. Young, P.J., *et al.*, *Astroph. J.*, **221**, 721, (1978).
11. Macchetto, F., *et al.*, *Astroph. J.*, **489**, 579, (1997).
12. Eddington, A.S., *The Observatory*, **42**, 119, (1919).
13. Mao, S., *et al.*, *Mon. Not. R. Astron. Soc.*, **329**, 349–354, (2002).
14. Bennett, D.P., *et al.*, *Astroph. J.*, **579**, 639, (2002).

Chapter 2

15. Fath, E.A., *Pub. Astron. Soc. Pacific*, **17**, 504–508, (1909).
16. Slipher, V.M., *Lowell Observatory Bulletin*, **80**, 59–62, (1917).
17. Campbell, W.W. & Moore, J.H., *Pub. Lick Obs.*, **13**, part IV, 75–184, (1918).
18. Morgan, W.W., *Astroph. J.*, **153**, 27–30, (1968).
19. Curtis, H.D., *Pub. Lick Obs.*, **13**, part I, 9–42, (1918).
20. Walker, R.C., Ly, C., Junor, W., & Hardee, P.E., *J. Phys.: Conf. Ser.*, **131**, 012053, (2008).
21. Hubble, E., *Astroph. J.*, **62**, 409–435, (1925).
22. Hubble, E., *Astroph. J.*, **64**, 321–372, (1926).
23. Hubble, E., *Proc. Natl. Acad. Sci.*, **15**, 168–173, (1929).
24. Seyfert, C.K., *Astroph. J.*, **97**, 28–41, (1943).
25. Markarian, B.E., *Astrofizika*, **3**, 24–38, (1967).
26. Markarian, B.E., Lipovetskii, V.A., & Stepanyan, Dzh A., *Astroph.*, **17**, 321–332, (1981).
27. Khachikian, E. Ya. & Weedman, D.W. *Astroph.*, **7**, 231–240, (1971).
28. Blandford, R.D. & Rees, M.J., *Mon. Not. R. Astron. Soc.*, **169**, 395–415, (1974).

29. Clarke, B.G., Kellermann, K.I., Cohen, M.H., & Jauncey, D.L., *Science*, **157**, 189–191, (1967).
30. Clarke, B.G., *et al.*, *Astroph. J.*, **153**, L67–L68, (1968).
31. Rogers, A.E.E., *et al.*, *Astroph. J.*, **193**, 293–301, (1974).
32. Hargrave, P.J. & Ryle, M., *Mon. Not. R. Astron. Soc.*, **166**, 305–327, (1974).
33. Waggett, P.C., Warner, P.J., & Baldwin, J.E., *Mon. Not. R. Astron. Soc.*, **181**, 465–474, (1977).
34. Readhead, A.C.S., Cohen, M.H., & Blandford, R.D., *Nature*, **272**, 131–134, (1978).
35. Dreher, J.W., Carilli, C.L., & Perley, R.A. *Astroph. J.*, **316**, 611–625, (1987).
36. Bennett, A.S., *Mon. Not. R. Astron. Soc.*, **125**, 75–86, (1962).
37. Laing, R.A., Riley, J.M., & Longair, M.S., *Mon. Not. R. Astron. Soc.*, **204**, 151–187, (1983).
38. Fanaroff, B.L. & Riley, J.M., *Mon. Not. R. Astron. Soc.*, **167**, 31–35, (1974).
39. Laing, R.A. & Bridle, A.H., *Mon. Not. R. Astron. Soc.*, **228**, 557–571, (1987).
40. Leahy, J.P., *Vistas Astron.*, **40**, 173–177, (1996).
41. <http://www.jb.man.ac.uk/atlas/>
42. Owen, F.N. & White, R.A., *Mon. Not. R. Astron. Soc.*, **249**, 164–171, (1989).
43. Owen, F.N. & White, R.A., *Mon. Not. R. Astron. Soc.*, **249**, 164–171, (1991).
44. Ledlow, M.J. & Owen, F.N., *Astron. J.*, **112**, 9–22, (1996).
45. Bicknell, G.V., *Astroph. J. Supp. Ser.*, **101**, 29–39, (1995).
46. Meier, D.L., *Astroph. J.*, **522**, 753–766, (1999).
47. Gopal-Krishna & Wiita, P. J., *Astron. Astroph.*, **363**, 507–516, (2000).
48. Osterbrock, D.E., Koski, A.T., & Phillips, M.M., *Astroph. J.*, **206**, 898–909, (1976).
49. Costero, R. & Osterbrock, D.E., *Astroph. J.*, **211**, 675–683, (1977).
50. Hazard, C., *Mon. Not. R. Astron. Soc.*, **124**, 343–171, (1962).
51. Schmidt, M., *Nature*, **197**, 1040, (1963).
52. Greenstein, J. L. & Schmidt, M., *Astroph. J.*, **140**, 1–34, (1964).
53. Schmidt, M., *Astroph. J.*, **141**, 1295–1300, (1965).
54. Chiu, H.-Y., *Quasi-Stellar Sources and Gravitational Collapse*, proceedings of the First Texas Symposium on Relativistic Astrophysics, 16–18 Dec 1963, ed. I. Robinson, A. Schild, & E. L. Schucking (Univ. Chicago Press, Chicago, 1965), p. 3.
55. Hutchings, J.B. & Neff, S.G., *Astron. J.*, **96**, 1575, (1988).
56. Hutchings, J.B. & Neff, S.G., *Astron. J.*, **101**, 2001, (1991).
57. Bahcall, J. N., *et al.*, *Astroph. J.*, **452**, L91, (1995).
58. Unwin, S. C., *et al.*, *Astroph. J.* **289**, 109, (1985).
59. Blandford, R. D. & A. Königl, *Astroph. J.*, **232**, 34, (1979).
60. Rees, M.J., *Nature*, **211**, 468–470, (1966).
61. Cohen, M. H., *et al.*, *Astroph. J.*, **170**, 207–217, (1971).
62. Cohen, M. H., *et al.*, *Nature*, **268**, 405–409, (1977).
63. Pearson, T. J., *et al.*, *Nature*, **290**, 365, (1981).
64. Murphy, D. W., *et al.*, *Mon. Not. R. Astron. Soc.*, **264**, 298, (1993).
65. Readhead, A.C.S., Cohen, M.H., Pearson, T.J., & Wilkinson, P.N., *Nature*, **276**, 768–771, (1978).
66. Pearson, T.J. & Readhead, A.C.S., *Astroph. J.*, **328**, 114–142, (1988).
67. Wehrle, A. E., *et al.*, *Astroph. J.*, **391**, 589, (1992).
68. O’dea, C., *et al.*, *Astron. J.*, **96**, 435, (1988).
69. Conway, J. & Murphy, D. W., *Astroph. J.*, **411**, 89, (1993).
70. <http://www.iau.org/public/naming/#stars>
71. Hoffmeister, C., *Astron. Nachr.*, **236**, 233–246, (1929).
72. Andrew, B.H. & MacLeod, J.M., *Astroph. Lett.*, **1**, 243–246, (1968).
73. Schmitt, J., *Nature*, **218**, 663, (1968).
74. Vermeulen, *et al.*, *Astroph. J.*, **452**, L5–L8, (1995).
75. Fossati, G., *et al.*, *Mon. Not. R. Astron. Soc.*, **299**, 433, (1998).
76. Ghisellini, G., *et al.*, *Mon. Not. R. Astron. Soc.*, **301**, 451, (1998).
77. URL: http://heasarc.nasa.gov/docs/cgro/images/epo/gallery/agns/agn_spectra.gif

78. *GLAST: Exploring Nature's Highest Energy Processes with the Gamma-ray Large Area Space Telescope*, NP-2000-9-107-GSFC, http://fermi.gsfc.nasa.gov/public/resources/pubs/gsd/GSD_print.pdf, February (2001).
79. Readhead, A.C.S., *Astroph. J.*, **426**, 51-59 (1994).
80. Kafka, P., *Nature*, **213**, 346, (1967).
81. Schmidt, M., *Astroph. J.*, **151**, 393, (1968).
82. Schmidt, M., *Astroph. J.*, **176**, 303, (1972).
83. Masson, C.R. & Wall, J.V., *Mon. Not. R. Astron. Soc.*, **180**, 193–206, (1977).
84. Scheuer, P.A. & Readhead, A.C.S., *Nature*, **277**, 182–185, (1980).
85. Jackson, C.A. & Wall, J.V., *Mon. Not. R. Astron. Soc.*, **304**, 160–174, (1999).
86. Antonucci, R. & Miller, J., *Astroph. J.*, **297**, 621, (1985).
87. Ebstein, S.M., Carleton, N.P., & Papaliolios, C., *Astroph. J.*, **336**, 103–111, (1989).
88. Cecil, G., Bland, J., & Tully, R.B., *Astroph. J.*, **355**, 70–87, (1990).
89. Evans, I.N., *et al.*, *Astroph. J.*, **369**, L27–L30, (1991).
90. Wilson, A.S., *et al.*, *Astroph. J.*, **419**, L61, (1993).
91. Jaffe, W., *et al.*, *Nature*, **364**, 213–215, (1993).
92. Zdziarski, A.A., *et al.*, *Mon. Not. R. Astron. Soc.*, **269**, L55–L60, (1994).
93. Madejski, G.M., *et al.*, *Astroph. J.*, **438**, 672–679, (1995).
94. Tanaka, Y., *et al.*, *Nature*, **375**, 659–661, (1995).
95. Halpern, J.P., *Astroph. J.*, **281**, 90–94, (1984).
96. Osterbrock, D.E. & Pogge, R.W., *Astroph. J.*, **297**, 166–176, (1985).
97. Osmer, P. & Smith, M., *Astroph. J.*, **210**, 267–276, (1976).
98. Pavlovsky, C., *et al.*, *HST ACS Data Handbook* (STScI, Baltimore, 2006).
99. Stern, D., *et al.*, *Astroph. J.*, **568**, 77–81, (2002).
100. Sanders, D.B. & Mirabel, I.F., *Ann. Rev. Astron. Astroph.*, **34**, 749–792, (1996).
101. Sunyaev, R. A., Tinsley, B. M. & Meier, D. L., *Comments on Astrophysics*, **7**, 183–195, (1978).
102. Meier, D. L. & Sunyaev, R. A., *Scientific American*, **241**, 130, (1979).
103. Smail, I., Ivison, R.J., & Blain, A.W., *Astroph. J.*, **490**, L5–L8, (1997).
104. Chapman, S.C., Blain, A.W., Smail, I., & Ivison, R.J., *Astroph. J.*, **622**, 772–796, (2005).
105. Kaufman, M., *Astroph. Sp. Sci.*, **40**, 369–384, (1976).
106. Meier, D.L., *Astroph. J.*, **207**, 343–350, (1976).
107. Treister, E., Urry, C.M., & Lira, P., *Rev. Mex. Astron. Astrof.*, **26**, 147–148, (2006).
108. Stern, D., *et al.*, *Astroph. J.*, **631**, 163–168, (2005).
109. Donley, J.L., *et al.*, *Astroph. J.*, **660**, 167–190, (2007).
110. Hickox, R.C., *et al.*, *Astroph. J.*, **671**, 1365–1387, (2007).
111. Wright, E.L., *et al.*, *Astron. J.*, **140**, 1868–1881, (2010).
112. Eisenhardt, P.M.E., *et al.*, in preparation.
113. Wu, J., *et al.*, in preparation.
114. Hickox, R.C., *et al.*, *Astroph. J.*, **696**, 891–919, (2009).
115. Leipski, C., *et al.*, *Astron. Astroph.*, **440**, L5–L8, (2005).
116. Lacy, M., in *Astrophysics Update 2*, ed. J. W. Mason (Springer Praxis Books, Berlin, 2006), pp. 195–211.
117. Cirasuolo, M., Celotti, A., Magliocchetti, M., & Danese, L., *Mon. Not. R. Astron. Soc.*, **346**, 447–455, (2003).
118. Ivezic, Z., *et al.*, *Physics with the Sloan Digital Sky Survey*, ASP Conf. Ser., **311**, 347–350, (2004).
119. Hutchings, J.B., *et al.*, *Astroph. J.*, **429**, L1, (1994).
120. Bahcall, J.N., Kirhakos, S., & Schneider, D.F., *Astroph. J.*, **435**, L11, (1994).
121. Hutchings, J.B. & Morris, S.C., *Astron. J.*, **109**, 1541, (1995).
122. Bahcall, J.N., Kirhakos, S., & Schneider, D.F., *Astroph. J.*, **450**, 486, (1995).
123. Bahcall, J.N., Kirhakos, S., & Schneider, D.F., *Astroph. J.*, **457**, 557, (1996).
124. Turnshek, D.A., *et al.*, *Astroph. J.*, **325**, 651–670, (1988).
125. Murray, N., *et al.*, *Astroph. J.*, **451**, 498, (1995).
126. Proga, D., Stone, J., & Kallman, T. R., *Astroph. J.*, **543**, 686, (2000).

127. Proga, D. & Kallman, T. R., *Astroph. J.*, **616**, 688, (2004).
128. Ogle, P., *et al.*, *Astroph. J. Supp Ser.*, **125**, 1–34, (1999).
129. Mushotzky, R.F., Done, C., & Pounds, K.A., *Ann. Rev. Astron. Astroph.*, **31**, 717–761, (1993).
130. Heckman, T., *Astron. Astroph.*, **87**, 152–164, (1980).
131. Ho, L., *Ann. Rev. Astron. Astroph.*, **46**, 475–539, (2008).
132. Genzel, R., Eisenhauer, F., & Gillessen, S., *Rev. Mod. Phys.*, **82**, 3121–3194, (2010).
133. Yuan, F., Markoff, S., & Falcke, H., *Astron. Astroph.*, **383**, 854–863, (2002).
134. Gillessen, S., *et al.*, *Nature*, **481**, 51–54, (2012).
135. Magorrian, J., *et al.*, *Astron. J.*, **115**, 2285–2305, (1998).
136. Ferrarese, L. & Merritt, D., *Astroph. J.*, **539**, L9–L12, (2000).
137. Gebhardt, K., *et al.*, *Astroph. J.*, **539**, L13–L16, (2000).
138. Kormendy, J., *Carnegie Obs. Astroph. Ser.*, **1**, *Coevolution of Black Holes and Galaxies* (Cambridge Univ. Press, Cambridge, 2004), 1–20.
139. Faber, S., *et al.*, *Nearly normal galaxies: From the Planck time to the present*, proc. 8th Santa Cruz Summer Workshop Astron. Astroph. (Springer-Verlag, New York, 1987), 175–183.
140. Dressler, A. *et al.*, *Astroph. J.*, **313**, 42–58, (1987).
141. van der Marel, R. P., *Astron. J.*, **117**, 744–763, (1999).
142. McLure, R., *et al.*, *Mon. Not. R. Astron. Soc.*, **308**, 377, (1999).
143. URL: http://heasarc.nasa.gov/docs/cgro/images/epo/gallery/agns/agn_up_model.gif
144. Urry, C.M. & Padovani, P., *Pub. Astron. Soc. Pacific*, **107**, 803–845, (1995).
145. Boroson, T.A., *Astroph. J.*, **565**, 78–85, (2002).
146. Lacy, M., *et al.*, *Astroph. J.*, **551**, L17–L21, (2001).
147. Dunlop, J. S. & McLure, R. J., in *The Mass of Galaxies at Low and High Redshift*, an ESO Astrophysics Symposium, 268, (2003).

Chapter 3

148. Chandrasekhar, S., *Astroph. J.*, **74**, 81–82, (1931).
149. Giacconi, R., Gursky, H., Paolini, F.R., & Rossi, B.B., *Phys. Rev. Lett.*, **9**, 439–443, (1962).
150. Shaposhnikov, N. & Titarchuk, L., *Astroph. J.*, **663**, 404, (2007).
151. Bowyer, S., Byram, E.T., Chubb, T.A., & Friedman, H., *Nature*, **201**, 1307–1308, (1964).
152. Bolton, J.G., *Nature*, **162**, 141–142, (1948).
153. Hewish, A., *et al.*, *Nature*, **217**, 709–713, (1968).
154. Lorimer, D.R., *Living Rev. Relativity*, **11**, <http://www.livingreviews.org/lrr-2008-8>, (2008).
155. Cordes, J.M. & Lazio, T.J.W., *arXiv:astro-ph/0207156v3*, (2002).
156. Cordes, J.M. & Lazio, T.J.W., *arXiv:astro-ph/0301598v1*, (2003).
157. Manchester, R.N., *Science*, **304**, 542, (2004).
158. Manchester, R.N., *ATNF Newsletter*, June (2006). URL: <http://www.atnf.csiro.au/news/newsletter/jun06/RRATs.htm>
159. Chadwick, J., *Nature*, **129**, 3122, (1932).
160. Baade, W. & Zwicky, F., *Phys. Rev.*, **46**, 76–77, (1934).
161. Baade, W., *Astroph. J.*, **96**, 188–198, (1942).
162. Minkowski, R., *Astroph. J.*, **96**, 199–213, (1942).
163. Gunn, J.E. & Ostriker, J.P., *Astroph. J.*, **160**, 979–1002, (1970).
164. Lyne, A.G., *Pulsars, IAU Symp.*, **95**, 423–436, (1981).
165. Phinney, E.S. & Kulkarni, S.R., *Ann. Rev. Astron. Astroph.*, **32**, 591–639, (1994).
166. Manchester, R.N., *et al.*, *Astron. J.*, **129**, 1993, (2005).
167. Caraveo, P.A. & Mignani, R.P., *Astron. Astroph.*, **344**, 367–370, (1999).
168. De Luca, A., Mignani, R.P., & Caraveo, P.A., *Astron. Astroph.*, **354**, 1011–1013, (2000).
169. Weisskopf, M.C. *et al.*, *Astroph. J.*, **536**, 81, (2000).
170. Hester, J.J. *et al.*, *Astroph. J.*, **577**, 49, (2002).
171. Pavlov, G.G. *et al.*, *Astroph. J.*, **554**, 189, (2002).
172. Mazets, E.P. *et al.*, *Nature*, **282**, 587–589, (1979).

173. Mazets, E.P. *et al.*, *Sov. Astron. Lett.*, **5**, 343–344, (1979).
174. Laros, J.G., *et al.*, *Nature*, **322**, 152–153, (1986).
175. Kouveliotou, C., *et al.*, *Astroph. J. Letters*, **322**, L21–L25, (1987).
176. Duncan, R.C. & Thompson, C., *Astroph. J.*, **392**, L9–L13, (1992).
177. Thompson, C. & Duncan, R.C., *Astroph. J.*, **408**, 194–217, (1993).
178. Thompson, C. & Duncan, R.C., *Mon. Not. R. Astron. Soc.*, **275**, 255–300, (1995).
179. Kouveliotou, C., *et al.*, *Astroph. J. Letters*, **510**, L115–L118, (1999).
180. Haberl, F., *Adv. Sp. Res.*, **33**, 638, (2004).
181. McLaughlin, M. A., *et al.*, *Nature*, **439**, 817, (2006).
182. Fichtel, C.E., *et al.*, *Astroph. J.*, **198**, 163, (1975).
183. Thompson, D.J., *et al.*, *Astroph. J.*, **213**, 252–262, (1977).
184. Halpern, J.P. & Holt, S.S., *Nature*, **357**, 222, (1992).
185. Shklovskii, I.S., *Astroph. J. Letters*, **148**, L1–L4, (1967).
186. Giacconi, R., *et al.*, *Astroph. J. Letters*, **167**, L67–L73, (1971).
187. Tananbaum, H., *et al.*, *Astroph. J. Letters*, **174**, L143–L149, (1972).
188. Hasinger, G. & van der Klis, M., *Astron. Astroph.*, **225**, 79–96, (1989).
189. van der Klis, M., *Astroph. J. Supp. Ser.*, **92**, 511–519, (1994).
190. van der Klis, M., *Lecture Notes in Physics*, **454**, 321–329, (1995).
191. Fender, R., *Lecture Notes in Physics*, **589**, 101–122, (2002).
192. Wijnands, R., *Adv. Sp. Res.*, **28**, 469–479, (2001).
193. Schnerr, R.S. *et al.*, *Astron. Astroph.*, **406**, 221–232, (2003).
194. Liu, Q.Z., van Paradijs, J., & van den Huevel, E.P.J., *Astron. Astroph.*, **455**, 1165–1168, (2006).
195. Liu, Q.Z., van Paradijs, J., & van den Huevel, E.P.J., *Astron. Astroph.*, **469**, 807–810, (2007).
196. Taylor, A.R. & Gregory, P.C., *Astroph. J.*, **255**, 210–216, (1982).
197. Gregory, P.C. & Neish, C., *Astroph. J.*, **580**, 1133–1148, (2002).
198. Hameury, J.M., King, A.R., & Lasota, J.-P., *Astron. Astroph.*, **162**, 71–79, (1986).
199. Tanaka, Y. & Shibazaki, N., *Ann. Rev. Astron. Astroph.*, **34**, 607–644, (1996).
200. Paizis, A., *et al.*, *Astron. Astroph.*, **459**, 187–197, (2006).
201. Falanga, M., *et al.*, *Astron. Astroph.*, **458**, 21–29, (2006).
202. van der Klis, M., *Ann. Rev. Astron. Astroph.*, **27**, 517–553, (1989).
203. van der Klis, M., *Ann. Rev. Astron. Astroph.*, **38**, 717–760, (2000).
204. McClintock, J.E. & Remillard, R.A., in *Compact Stellar X-ray Sources*, eds. W. Lewin and M. van der Klis (Cambridge Univ. Press, Cambridge, 2006).
205. Remillard, R.A. & McClintock, J.E., *Ann. Rev. Astron. Astroph.*, **44**, 49–92, (2006).
206. Margon, B. *et al.*, *Astroph. J.*, **230**, L41–L45, (1979).
207. Margon, B. *et al.*, *Astroph. J.*, **233**, L63–L68, (1979).
208. Dubner, G.M., *et al.*, *Astron. J.*, **116**, 1842–1855, (1998).
209. Blundell, K.M., *et al.*, *Astroph. J.*, **616**, L159–L162, (2004).
210. Backer, D.C., *et al.*, *Nature*, **300**, 615–618, (1982).
211. Hulse, R.A. & Taylor, J.H. *Astroph. J.*, **195**, L51–L53, (1975).
212. Filippenko, A.V., *Ann. Rev. Astron. Astroph.*, **35**, 309–355, (1997).
213. Wang, L. & Wheeler, J.C., *Ann. Rev. Astron. Astroph.*, **46**, 433–474, (2008).
214. Woosley, S.E. & Bloom, J.S., *Ann. Rev. Astron. Astroph.*, **44**, 507–556, (2006).
215. Bennett, D.P., *et al.*, *Astroph. J.*, **579**, 639–659, (2002).
216. Webster, B.L. & Murdin, P., *Nature*, **235**, 37–38, (1972).
217. McConnell, M.L., *et al.*, *Astroph. J.*, **572**, 984, (2002).
218. Coppi, P.S., *High Energy Processes in Accreting Black Holes*, ASP Conf. Ser., **161**, 375, (1999).
219. Gierliński, M., *et al.*, *Mon. Not. R. Astron. Soc.*, **309**, 496, (1999).
220. Zdziarski, A.A. & Gierliński, M., *Prof. Th. Ph. Suppl.*, **155**, 99, (2004).
221. Mirabel, L.F., *et al.*, *Nature*, **358**, 215–217, (1992).
222. Castro-Tirado, A.J., Brandt, S., & Lund, N., *IAU Circ.*, **5590**, 1, (1992).
223. Castro-Tirado, A., *et al.*, *Astroph. J. Supp. Ser.*, **92**, 469–472, (1994).
224. Mirabel, I.F. & Rodriguez, L.F., *Nature*, **371**, 46–48, (1994).

225. Greiner, J., Morgan, E.H., & Remillard, R.A., *Astroph. J.*, **473**, L107–L110, (1996).
226. Belloni, T.M., *Lecture Notes in Physics*, **794**, 53–84, (2009).
227. Mirabel, I.F., *et al.*, *Astron. Astroph.*, **330**, L9–L12, (1998).
228. Marscher, A., *et al.*, *Nature*, **417**, 625–627, (2002).
229. Greiner, J., Cuby, J.G., & McCaughrean, M.J., *Nature*, **414**, 522–525, (2001).
230. Zhang, S.N., *et al.*, *IAU Circ.*, **6046**, 1, (1994).
231. Tingay, S.J., *et al.*, *Nature*, **374**, 141–143, (1995).
232. Harmon, B.A. *et al.*, *Nature*, **374**, 703–706, (1995).
233. Hjellming, R.M. & Rupen, M.P., *Nature*, **375**, 464–468, (1995).
234. Kluzniak, W. & Abramowicz, M.A., *Astroph. Sp. Sci.*, **300**, 143–148, (2005).
235. Abramowicz, M.A. & Kluzniak, W., *Astroph. Sp. Sci.*, **300**, 127–136, (2005).
236. Fender, R.P., Belloni, T.M., & Gallo, E., *Mon. Not. R. Astron. Soc.*, **355**, 1105–1118, (2004).
237. Meier, D.L., in *X-Ray Timing 2003: Rossi and Beyond*, ed. P. Kaaret, F.K. Lamb, & J.H. Swank, pp. 135–142, (2004).
238. Szuszkiewicz, E. & Miller, J.C., *Mon. Not. R. Astron. Soc.*, **328**, 36–44, (2001).
239. Salpeter, E.E., *Astroph. J.*, **121**, 161–167 (1955).
240. van Paradijs, J., *et al.*, *Nature*, **386**, 686–689, (1997).
241. Metzger, M.R. *et al.*, *Nature*, **387**, 878–880, (1997).
242. Fishman, G., *et al.*, *Astroph. J. Supp. Ser.*, **92**, 229–283, (1994).
243. Fishman, G.J. & Meegan, C.A., *Ann. Rev. Astron. Astroph.*, **33**, 415–458, (1995).
244. Paczyński, B., *Astroph. J.*, **308**, L43–L46, (1986).
245. Meszaros, P. & Rees, M.J., *Mon. Not. R. Astron. Soc.*, **257**, 29p–31p (1992).
246. Rees, M.J. & Meszaros, P., *Mon. Not. R. Astron. Soc.*, **258**, 41p–43p (1992).
247. Rhoads, J.E. *Astroph. J.*, **487**, L1–L4, (1997).
248. Castro-Tirado, A.J., *et al.*, *Science*, **283**, 2069–2073, (1999).
249. Piran, T., *Phys. Rep.*, **333–334**, 529–553, (2000).
250. Meszaros, P., *Science*, **291**, 79–83, (2001).
251. Bloom, J.S., *et al.*, *Astroph. J.*, **572**, L45–L49, (2002).
252. MacFadyen, A. & Woosley, S., *Astroph. J.*, **524**, 262–289, (1999).
253. Gehrels, N., *et al.*, *Astroph. J.*, **611**, 1005, (2004).
254. Gehrels, N., *et al.*, *Nature*, **437**, 851, (2005).
255. Fox, D.B., *et al.*, *Nature*, **437**, 845, (2005).
256. Nakar, E., *Ph. Rep.*, **442**, 166, (2007).
257. Berger, E., *et al.*, *Astroph. J.*, **664**, 1000, (2007).
258. Piro, L., *et al.*, *Astroph. J.*, **577**, 680–690, (2002).
259. Soderberg, A.M., *et al.*, *Nature*, **442**, 1014–1017, (2006).
260. Norris, J.P., *et al.*, *Astroph. J.*, **627**, 324–345, (2005).
261. Virgili, F.J., Liang, E.-W., & Zhang, B., *Mon. Not. R. Astron. Soc.*, **392**, 91–103, (2009).
262. Thorsett, S.E., *Astroph. J.*, **444**, L53, (1995).
263. Stanek, K., *et al.*, *Acta Astron.*, **56**, 333, (2006).

Chapter 4

264. Swartz, D.A., *ASP Conf. Ser.*, **423**, 277–281, (2010).
265. Swartz, D.A., *Astron. Nachr.*, **332**, 341–344, (2011).
266. Fabbiano, G., *Ann. Rev. Astron. Astroph.*, **44**, 323–366, (2006).
267. Mushotzky, R., *Prog. Theor. Phys. Supp.*, **155**, 27–44, (2005).
268. Miller, M.C. & Colbert, E.J.M., *Int. J. Mod. Phys.*, **D13**, 1–64, (2004).
269. Long, K., *et al.*, *Astroph. J.*, **246**, L61–L64, (1981).
270. van der Marel, R. P., in *Coevolution of Black Holes and Galaxies*, ed. L. C. Ho (Cambridge, Cambridge Univ. Press, 2004), a Carnegie Observatories Centennial Symposium, p. 37.
271. Ptak, A. & Griffiths, R., *et al.*, *Astroph. J.*, **517**, L85–L89, (1999).
272. Portegies Zwart, S. & McMillan, S. *Astroph. J.*, **576**, 899–907, (2002).
273. Makishima, K., *et al.*, *Astroph. J.*, **535**, 632–643, (2000).
274. King, A.R., *et al.*, *Astroph. J.*, **552**, L109–L112, (2001).

275. Begelman, M., *et al.*, *Astroph. J.*, **568**, L97–L199, (2002).
276. Kajava, J.E. & Poutanen, J., *Mon. Not. R. Astron. Soc.*, **398**, 1450–1460, (2009).
277. Maccarone, T., *et al.*, *Nature*, **445**, 183, (2007).
278. Zepf, S.E., *et al.*, *Astroph. J.*, **683**, L139, (2008).
279. Swartz, D.A., Ghosh, K.K., Tennant, A.F., & Wu, K., *Astroph. J. Supp. Ser.*, **154**, 519–539, (2004).
280. Gebhardt, K., *et al.*, *Astroph. J.*, **578**, L41–L45, (2002).
281. Gerssen, J., *et al.*, *Astron. J.*, **125**, 376–377, (2003).
282. Gerssen, J., *et al.*, *Astron. J.*, **124**, 3270–3288, (2002).
283. Rich, R.M., *et al.*, *Astron. J.*, **111**, 768–776, (1996).
284. Ulvestad, J., *et al.*, *Astroph. J.*, **661**, L151–L154, (1995).
285. Noyola, E., *et al.*, *Astroph. J.*, **676**, 1008–1015, (2008).
286. Baumgardt, H., *et al.*, *Astroph. J.*, **582**, L21–L24, (2003).
287. Hughes, J. & Wallerstein, G., *Astron. J.*, **119**, 1225–1238, (2000).
288. Meylan, G., *et al.*, *Astron. J.*, **122**, 830–851, (2001).
289. Bash, F., *et al.*, *Astron. J.*, **135**, 182–186, (2008).
290. Becklin, E.E. & Neugebauer, G., *Astroph. J.*, **200**, L71–L74, (1975).
291. Ghez, A., *et al.*, *Astroph. J.*, **620**, 744–757, (2005).
292. Gillessen, S., *et al.*, *Astroph. J.*, **692**, 1075–1109, (2009).
293. Hansen, B.M. & Milosavljević, M., *Astroph. J.*, **593**, L77–L80, (2003).
294. Maillard, J.P., *et al.*, *Astron. Astroph.*, **423**, 155–167, (2004).
295. Revnivtsev, M.G., *et al.*, *Astron. Lett.*, **30**, 382–389, (2004).
296. Muno, M.P., *et al.*, *Astroph. J. Supp. Ser.*, **181**, 110–128, (2009).
297. Maccarone, T.J., *Mon. Not. R. Astron. Soc.*, **351**, 1049–1053, (2004).
298. Maccarone, T.J., *et al.*, *Mon. Not. R. Astron. Soc.*, **356**, L17–L22, (2005).

Chapter 5

299. Newton, I., *The Mathematical Principles of Natural Philosophy*, English translation of the 1687 *Philosophiae Naturalis Principia Mathematica*, tr. A. Motte, (1729).
300. Landau, L.D. & Lifshitz, E.M., *Mechanics, Vol. 1 of Course of Theor. Phys.* (Pergamon Press, Oxford, UK, 1969), 2nd edition.
301. Chandrasekhar, S., *An Introduction to the Study of Stellar Structure*; originally published 1939 (Dover, New York, 1958), p. 96.
302. Aller, L.H. & McLaughlin, D.B., *Stellar Structure*, in *Stars and Stellar Systems*, **8** (U. of Chicago Press, Chicago, 1965).
303. Lane, J.H., *Am. J. Sci.; 2nd Ser.*, **1**, 57–74, (1870).
304. Emden, V.R., *Gaskugeln* (Teubner, Leipzig, 1907).

Chapter 6

305. Misner, C.W., Thorne, K.S., & Wheeler, J.A., *Gravitation* (Freeman, San Francisco, 1973) (MTW).
306. Jackson, J.D., *Classical Electrodynamics* (John Wiley & Sons, Inc., New York, 1999).
307. Einstein, A., *Annal. Phys.*, **17**, 891–921, (1905).
308. Einstein, A., *Sitzungsber. Preuss. Akad. Wiss. zu Berlin*, 844–847, (1915).
309. Einstein, A., *Annal. Phys.*, **49**, 284–339, (1916).

Chapter 7

310. Bardeen, J.M., Press, W.H., & Teukolsky, W.H., *Astroph. J.*, **178**, 347–369, (1972).
311. Schwarzschild, K., *Sitzungsber. Preuss. Akad. Wiss.*, *Phys. Math. Kl.*, 189.
312. Hilbert, D., *Nachr. Ges. Wiss. Göttingen, Math. Phys. Kl.*, 53, (1917).
313. Antoci, S., in *Meteor. Geophys. Fluid Dyn.*, ed. W. Schröder, (2004), 343–354.
314. Kerr, R.P., *Phys. Rev. Lett.*, **11**, 237–238, (1963).
315. Boyer, R.H. & Lindquist, R.W., *J. Math. Phys.*, **8**, 265–281, (1967).

316. Thorne, K.S., *Astroph. J.*, **191**, 507–519, (1974).
317. Newman, E.T., *et al.*, *J. Math. Phys.*, **6**, 918–919, (1965).
318. Arnowitt, R.L., Deser, S., & Misner, C.W., *The Dynamics of General Relativity*, arXiv:gr-qc/0405109; originally published in *Gravitation: An Introduction to Current Research* (John Wiley & Sons, Inc., New York, 1962), pp. 227–264.
319. Thorne, K.S., Price, R.H., & Macdonald, D.A., *Black Holes: The Membrane Paradigm* (Yale Univ Press, New Haven, 1986).
320. Koide, S., Shibata, K., Kudoh, T., & Meier, D.L., *J. Korean Astron. Soc.*, **34**, S215–S224, (2001).
321. Koide, S., *Phys. Rev. D*, **67**, 104010, (2003).
322. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **350**, 427–448, (2004).

Chapter 8

323. Einstein, A., *Sitzungsber. Preuss. Akad. Wiss. zu Berlin*, 154–167, (1918).
324. Sathyaprakash, B.S. & Schutz, B.F., *Living Rev. Relativity*, **12**, <http://www.livingreviews.org/lrr-2009-2>, (2009).
325. Landau, L.D. & Lifshitz, E.M., *The Classical Theory of Fields, Vol. 2 of Course of Theor. Phys.* (Pergamon Press, Oxford, UK, 1975), 4th edition.
326. Hughes, S.A., *Ann. Rev. Astron. Astroph.*, **47**, 107–157, (2009).
327. Postnov, K.A. & Yungelson, L.R., *Living Rev. Relativity*, **9**, <http://www.livingreviews.org/lrr-2006-6>, (2006).
328. Shibata, M. & Nakamura, T., *Phys. Rev. D*, **52**, 5428, (1992).
329. Baumgarte, T.W. & Shapiro, S.L., *Phys. Rev. D*, **59**, 024007, (1999).
330. Pretorius, F., *Phys. Rev. Lett.*, **95**, 121101, (2005).
331. Campanelli, M., Lousto, C.O., & Zlochower, Y., *Phys. Rev. D*, **73**, 061501(R), (2006).
332. Campanelli, M., *et al.*, *Phys. Rev. Lett.*, **96**, 111101, (2006).
333. Baker, J.G., *et al.*, *Phys. Rev. Lett.*, **96**, 111102, (2006).
334. Campanelli, M., Lousto, C.O., & Zlochower, Y., *Phys. Rev. D*, **74**, 041501, (2006).
335. Campanelli, M., Lousto, C.O., Mundim, B.C., Nakano, H., Zlochower, Y., & Bischof, H.-P., *Class. Quantum Grav.*, **27**, 084034, (2010).
336. Baker, J.G., *et al.*, *Astroph. J.*, **653**, L93–L96, (2006).
337. Baker, J.G., *et al.*, *Astroph. J.*, **668**, 1170–1144, (2007).
338. van Meter, J.R., *et al.*, *Astroph. J.*, **719**, 1427–1432, (2010).
339. Lousto, C.O. & Zlochower, Y., *Phys. Rev. D*, **83**, 024003, (2011).
340. Natarajan, P. & Pringle, J.E., *Astroph. J.*, **506**, L97–L100, (1998).
341. Dotti, M., *et al.*, *Mon. Not. R. Astron. Soc.*, **402**, 682–690, (2010).
342. Van Riper, K.A., *Astroph. J.*, **232**, 558–571, (1979).
343. Shapiro, S.L. & Teukolsky, S.A., *Astroph. J.*, **235**, 199–215, (1980).
344. Tolman, R.C., *Phys. Rev.*, **55**, 364–373, (1939).
345. Oppenheimer, J.R. & Volkoff, G.M., *Phys. Rev.*, **55**, 374–381, (1939).

Chapter 9

346. Meier, D.L., *Astroph. J.*, **605**, 340–349, (2004).
347. Wang, K., *Statistical Mechanics* (John Wiley & Sons, Inc., New York, 1963).
348. Reif, F., *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965).
349. Clayton, D.D., *Principles of Stellar Evolution and Nucleosynthesis* (McGraw-Hill, New York, 1968).
350. Krall, N.A. & Trivelpiece, A.W., *Principles of Plasma Physics* (McGraw-Hill, New York, 1973).
351. Shakura, N.I. & Sunyaev, R.A., *Astron. Astroph.*, **24**, 337–355, (1973).
352. Pringle, J.E., Rees, M.J., & Pacholczyk, A.G., *Astron. Astroph.*, **29**, 179–184, (1973).
353. Esin, A.A., Narayan, R., Ostriker, E., & Yi, I., *Astroph. J.*, **465**, 312–326, (1996).
354. Mahadevan, R., Narayan, R., & Yi, I., *Astroph. J.*, **465**, 327–337, (1996).
355. Fragile, P.C. & Meier, D.L., *Astroph. J.*, **693**, 771–783, (2009).

356. Sunyaev, R.A. & Titarchuk, L.G., *Astron. Astroph.*, **86**, 121–138, (1980).
357. Koide, S., Shibata, K., & Kudoh, T., *Accretion Phenomena and Related Outflows*, ASP Conf. Ser., **121**, 667–671, (1997).
358. De Villiers, J.-P. & Hawley, J.F., *Astroph. J.*, **589**, 458–480, (2003).
359. Gammie, C.F., McKinney, J.C., & Tóth, G., *Astroph. J.*, **589**, 444–457, (2003).
360. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **350**, 1431–1436, (2004).
361. Anninos, P., Fragile, P.C., & Salmonson, J.D., *Astroph. J.*, **635**, 723–740, (2005).
362. Blandford, R.D., *Lighthouses of the Universe: The Most Luminous Celestial Objects and Their Use for Cosmology*, ESO Astroph. Symp., 381–404, (2002).
363. Lüst, R., & Schlüter, A., *Z. Naturf.*, **12a**, 85, (1957).
364. Shafranov, V.D., *Sov. Phys. – JETP*, **6**, 545, (1958).
365. Grad, H. & Rubin, H., *Hydromag. Equilibria and Force-Free Fields*, UN Conf., **31**, 190, (1958).
366. Bogovalov, S.V., *Mon. Not. R. Astron. Soc.*, **270**, 721–733, (1994).
367. Balbus, S.A. & Hawley, J.F., *Rev. Mod. Phys.*, **70**, 1–53, (1998).
368. Vlahakis, N. & Königl, A., *Astroph. J.*, **596**, 1030–1103, (2003).
369. Polovin, R.V. & Demutskii, V.P., *Fundamentals of Magnetohydrodynamics*, tr. D. ter Haar (Consultants Bureau, New York, 1990).
370. Contopoulos, J., *Astroph. J.*, **460**, 185–198, (1996).

Chapter 10

371. Smith, H.E., <http://cass.ucsd.edu/archive/public/tutorial/StevII.html>, (1999).
372. Woosley, S.E., Heger, A., & Weaver, T.A., *Rev. Mod. Phys.*, **74**, 1015–1071, (2002).
373. LeBlanc, J.M. & Wilson, J.R., *Astroph. J.*, **161**, 541–551, (1970).
374. Wheeler, J.C., Meier, D.L., & Wilson, J.R., *Astroph. J.*, **568**, 807–819, (2002).
375. Fryer, C.L. & Heger, A., *Astroph. J.*, **541**, 1033–1050, (2000).
376. Heger, A., Woosley, S.E., & Spruit, H.C., *Astroph. J.*, **626**, 350–363, (2005).
377. Nelemans, G., & Tout, C.A., *Mon. Not. R. Astron. Soc.*, **356**, 753–764, (2005).
378. Miller, M.A., Gressman, P., & Suen, W.-M., *Phys. Rev. D*, **69**, 064026, (2004).
379. Miller, Mark A., *Phys. Rev. D*, **75**, 024001, (2007).
380. Shibata, M. & Uryu, K., *Phys. Rev. D*, **61**, 064001, (2000).
381. Baumgarte, T.W., Shapiro, S.L., & Shibata, M., *Astroph. J.*, **528**, L29–L32, (2000).
382. Shibata, M., Baumgarte, T.W., & Shapiro, S.L., *Phys. Rev. D*, **61**, 044012, (2000).
383. Höflich, P., Khokhlov, A.M., & Wheeler, J.C., *Astroph. J.*, **444**, 831–847, (1995).
384. Khokhlov, A.M., Oran, E.S., & Wheeler, J.C., *Astroph. J.*, **478**, 678–688, (1997).
385. Mazurek, T.J., Meier, D.L., & Wheeler, J.C., *Astroph. J.*, **213**, 518–526, (1977).
386. Nelemans, G., *et al.*, *Astron. Astroph.*, **365**, 491–507, (2001).
387. Scalzo, J.M., *Fundam. Cosmic Phys.*, **11**, 1, (1986).
388. Abel, A., *et al.*, *Astroph. J.*, **508**, 518–529, (1998).
389. Nakamura, F. & Umemura, M., *Astroph. J.*, **515**, 239–248, (1999).
390. Nakamura, F. & Umemura, M., *Astroph. J.*, **548**, 19–32, (2001).
391. Bromm, V., Coppi, P.S., & Larson, R.B., *Astroph. J.*, **527**, L5–L8, (1999).
392. Bromm, V., Kudritzki, R.P., & Loeb, A., *Astroph. J.*, **552**, 464–472, (2001).
393. King, I.R., *Astron. J.*, **71**, 64–75, (1966).
394. Cohn, H., *Astroph. J.*, **242**, 765–771, (1980).
395. Lynden-Bell, D., & Wood, R., *Mon. Not. R. Astron. Soc.*, **138**, 495–525, (1968).
396. O’Leary, R.M. *et al.*, *Astroph. J.*, **637**, 937–951, (2006).
397. Freitag, M. & Benz, W., *Astron. Astroph.*, **392**, 345–374, (2002).
398. Gürkan, M.A., Freitag, M., & Rasio, F.A., *Astroph. J.*, **604**, 632–652, (2004).
399. Kauffmann, G. & Haehnelt, M., *Mon. Not. R. Astron. Soc.*, **311**, 576–588, (2000).
400. King, A.R., *Astroph. J.*, **596**, L26–L29, (2003).
401. Ebisuzaki, T., *et al.*, *Astroph. J.*, **562**, L19–L22, (2001).
402. Merritt, D. & Milosavljević, M., *Living Rev. Relativity*, **8**, <http://www.livingreviews.org/lrr-2005-8>, (2005).

Chapter 11

403. Hoyle, F. & Lyttleton, R.A., *Mon. Not. R. Astron. Soc.*, **101**, 227–236, (1941).
404. Bondi, H. & Hoyle, F., *Mon. Not. R. Astron. Soc.*, **104**, 273–282, (1944).
405. Bondi, H., *Mon. Not. R. Astron. Soc.*, **112**, 195–204, (1952).
406. Novikov, I.D. & Thorne, K.S., *Black Holes*, proc. Les Houches Summer School 1972, ed. Dewitt, C. & Dewitt, B.S. (1973).
407. Faber, S.M. & Jackson, R.E., *Astroph. J.*, **204**, 668–683, (1976).
408. Young, P.J., Shields, G.A., & Wheeler, J.C., *Astroph. J.*, **212**, 367–382, (1977).
409. Danby, J.M.A. & Camm, G.L., *Mon. Not. R. Astron. Soc.*, **117**, 50–712, (1957).
410. Frank, J. & Rees, M.J., *Mon. Not. R. Astron. Soc.*, **176**, 633, (1976).
411. Lightman, A.P. & Shapiro, S.L., *Astroph. J.*, **211**, 2244, (1977).
412. Cohn, H & Kulsrud, R.M., *Astroph. J.*, **226**, 108, (1978).
413. Merritt, D., *Rep. Prog. Phys.*, **69**, 2513–2579, (2006).
414. Buta, R. & Combes, F., *Fund. Cosmic Phys.*, **17**, 95–281, (1996).
415. Mulchaey, J.S. & Regan, M.W., *Astroph. J.*, **482**, L135–L137, (1997).
416. Hunt, L.K. & Malkan, M.A., *Astroph. J.*, **516**, 660–671, (1999).
417. Knapen, J.H., Shlosman, I., & Peletier, R.F., *Astroph. J.*, **529**, 93–100, (2000).
418. Laurikainen, E., Salo, H., & Rautiainen, P., *Mon. Not. R. Astron. Soc.*, **331**, 880–892, (2002).
419. Jogee, S., *Lecture Notes in Physics*, **693**, 143, (2006).
420. Bains, I., *et al.*, *Mon. Not. R. Astron. Soc.*, **367**, 1609–1628, (2006).
421. Velusamy, T., *et al.*, *Astroph. J.*, **688**, L87–L90, (2008).
422. Murray, N., *Astroph. J.*, **729**, 133–146, (2011).
423. Demircan, O. & Kahraman, G., *Astroph. Sp. Sci.*, **181**, 313–322, (1991).
424. Davies, R.I., *et al.*, *Astroph. J.*, **671**, 1388–1412, (2007).
425. Eggleton, P.P., *Astroph. J.*, **268**, 368–369, (1983).
426. Hughes, S.A. & Blandford, R.D., *Astroph. J.*, **585**, L101–L104, (1995).
427. Heger, A., Woosley, S.E., & Langer, N., *IAU Symp.*, **212**, 357–364, (2003).

Chapter 12

428. Parker, E.N., *Astroph. J.*, **132**, 821–866, (1960).
429. Flammang, R.A., *Mon. Not. R. Astron. Soc.*, **199**, 833–867, (1982).
430. Thorne, K.S., Flammang, R.A., & Żytkow, A.N., *Mon. Not. R. Astron. Soc.*, **194**, 475–484, (1981).
431. Livio, M., Ogilvie, G.I., & Pringle, J.E. *Astroph. J.*, **512**, 100–104, (1999).
432. Meier, D.L., *Astroph. J.*, **548**, L9–L12, (2001).
433. Lightman, A.P. & Eardley, D.M., *Astroph. J.*, **187**, L1–L3, (1974).
434. Shakura, N.I. & Sunyaev, R.A., *Mon. Not. R. Astron. Soc.*, **175**, 613–632, (1976).
435. Paczyński, B. & Wiita, P.J., *Astron. Astroph.*, **88**, 23–31, (1980).
436. Thorne, K.S. & Price, R.H., *Astroph. J.*, **195**, L101–L105, (1975).
437. Shapiro, S.I., Lightman, A.P., & Eardley, D.M., *Astroph. J.*, **204**, 187–199, (1976).
438. Pringle, J.E., *Mon. Not. R. Astron. Soc.*, **177**, 65–71, (1976).
439. Begelman, M.C. & Meier, D.L., *Astroph. J.*, **253**, 873–896, (1982).
440. Ichimaru, S., *Astroph. J.*, **214**, 840–855, (1977).
441. Narayan, R. & Yi, I., *Astroph. J.*, **452**, 710–735, (1995).
442. Narayan, R., Mahadevan, R., & Quataert, E., in *Theory of Black Hole Accretion Disks*, ed. M. Abramowicz, G. Björnsson, & J.E. Pringle, 148–182, (1998).
443. Liu, B.F., *et al.*, *Astroph. J.*, **527**, L17–L20, (1999).
444. Meyer, F. Liu, B.F., Meyer-Hofmeister, E., *Astron. Astroph.*, **354**, L67–L70, (2000).
445. Meyer, F. Liu, B.F., Meyer-Hofmeister, E., *Astron. Astroph.*, **361**, 176–188, (2000).
446. Liu, B.F., *et al.*, *Astroph. J.*, **575**, 117–126, (2002).
447. Czerny, B., Róžańska, A., & Janiuk, A., *Adv. Space Res.*, **28**, 439–443, (2001).
448. Chen, X., Abramowicz, M.A., Lasota, J.-P., Narayan, R., & Yi, I., *Astroph. J.*, **443**, L61–L64, (1995).
449. Esin, A.A., McClintock, J.E., & Narayan, R., *Astroph. J.*, **489**, 865–889, (1997).

450. Cowling, T.G., *Mon. Not. R. Astron. Soc.*, **94**, 39–48, (1934).
451. Hawley, J.F., Gammie, C.F., & Balbus, S.A., *Astroph. J.*, **440**, 742–763, (1995).
452. Hirose, S., Krolik, J.H., & Blaes, O., *Astroph. J.*, **691**, 16–31, (2009).
453. Turner, N.J., *Astroph. J.*, **605**, L45–L48, (2004).
454. Hirose, S., Blaes, O., & Krolik, J.H., *Astroph. J.*, **704**, 781–788, (2009).
455. Meier, D.L., *Astroph. J.*, **233**, 664–684, (1979).
456. De Villiers, J.-P., Hawley, J.F. & Krolik, J.H., *Astroph. J.*, **599**, 1238–1253, (2003).
457. McKinney, J.C., *Mon. Not. R. Astron. Soc.*, **368**, 1561–1582, (2006).

Chapter 13

458. Meier, D.L., *Astroph. J.*, **256**, 681–692, (1982).
459. Meier, D.L., *Astroph. J.*, **256**, 693–705, (1982).
460. Meier, D.L., *Astroph. J.*, **256**, 706–716, (1982).
461. Rees, M.J., talk presented at the NATO Advanced Study Institute “Active Galactic Nuclei”, Univ. Cambridge, UK, (1977).
462. Quinn, T. & Paczyński, B., *Astroph. J.*, **289**, 634–643, (1985).
463. Meier, D.L., Ph. D. thesis, *The Structure and Appearance of Winds from Supercritical Accretion Disks*, University of Texas at Austin (1977).
464. Blandford, R.D. & Begelman, M.C. 1999, *Mon. Not. R. Astron. Soc.*, **303**, L1–L5.
465. Falcke, H. & Markoff, S., *Astron. Astroph.*, **362**, 113–118, (2000).
466. Markoff, S., Falcke, H., & Fender, R., *Astron. Astroph.*, **372**, L25–L28, (2001).
467. Markoff, S., *et al.*, *Astron. Astroph.*, **397**, 645–658, (2003).
468. Markoff, S. & Nowak, M., *Astroph. J.*, **609**, 972–976, (2004).
469. Markoff, S., in *New Horizons in Astronomy*, ASP Conf. Ser., **352**, 129–143, (2005).
470. Falcke, H., *Astroph. J.*, **464**, L67–L70, (1996).
471. Falcke, H. & Biermann, P.L., *Astron. Astroph.*, **293**, 665–682, (1995).

Chapter 14

472. Pacini, F., *Nature*, **219**, 145–146, (1968).
473. Ostriker, J.P. & Gunn, J.E., *Astroph. J.*, **157**, 1395–1417, (1969).
474. Goldreich, P. & Julian, W.H., *Astroph. J.*, **157**, 869–880, (1969).
475. Contopoulos, I., Kazanas, D., & Fendt, C., *Astroph. J.*, **511**, 351–358, (1999).
476. Uzdensky, D.A., *Astroph. J.*, **598**, 446–457, (2003).
477. Spitkovsky, A., *Astroph. J.*, **648**, L51–L54, (2006).
478. Spitkovsky, A., in *Astrophysical Sources of High Energy Particles and Radiation*, ed. T. Bulik, B. Rudak, & G. Madejski, AIP 0-7354-0290-6/05, pp. 253–258, (2005).
479. Ruderman, M.A. & Sutherland, P.G., *Astroph. J.*, **196**, 51–72, (1975).
480. Camenzind, M., in *Rev. Mod. Astron.*, **3**, Klare, G. ed. (Springer, Berlin, 1990), pp. 234–265.
481. Romanova, M.M., *et al.*, *Adv. Sp. Res.*, **38**, 2887–2892, (2006).
482. Romanova, M.M., Kulkarni, A.K., & Lovelace, R.V.E., *Astroph. J.*, **673**, L171–L174, (2008).
483. Ustyugova, G.V., *et al.*, *Astroph. J.*, **646**, 304–318, (2006).
484. Blandford, R.D. & Znajek, R.L., *Mon. Not. R. Astron. Soc.*, **179**, 433–456, (1977).
485. Znajek, R.L., *Mon. Not. R. Astron. Soc.*, **179**, 457–472, (1977).
486. Uzdensky, D.A., *Astroph. J.*, **620**, 889–904, (2005).
487. Michel, F.C., *Astroph. J.*, **180**, L133–L137, (1973).
488. Michel, F.C., *Rev. Mod. Phys.*, **54**, 1–66, (1982).
489. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **326**, L41–L44, (2001).
490. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **336**, 759–766, (2002).
491. McKinney, J.C., *Mon. Not. R. Astron. Soc.*, **367**, 1797–1807, (2006).
492. Wald, R.M., *Phys. Rev. D*, **10**, 1680–1685, (1974).
493. Meissner, W. & Ochsenfeld, R., *Naturwissenschaften*, **44**, 787–788, (1933).
494. Komissarov, S.S. & McKinney, J.C., *Mon. Not. R. Astron. Soc.*, **377**, L49–L53, (2007).
495. Punsly, B. & Coroniti, F.V., *Astroph. J.*, **354**, 583–615, (1990).

496. Takahashi, M., *et al.*, *Astroph. J.*, **363**, 206–217, (1990).
497. Hirose, S., Krolik, J.H., De Villiers, J.-P., & Hawley, J.F., *Astroph. J.*, **606**, 1083–1097, (2004).
498. Tomimatsu, A. & Takahashi, M., *Astroph. J.*, **552**, 710–717, (2001).
499. Uzdensky, D.A., *Astroph. J.*, **603**, 652–662, (2004).
500. Koide, S., Shibata, K., & Kudoh, T., *Astroph. J.*, **522**, 727–752, (1999).
501. Garofalo, D.A., *Astroph. J.*, **699**, 400–408, (2009).
502. Hirotani, K., *et al.*, *Astroph. J.*, **386**, 455–463, (1992).
503. McKinney, J.C. & Narayan, R., *Mon. Not. R. Astron. Soc.*, **375**, 531–547, (2007).
504. Koide, S., Shibata, K., Kudoh, T., & Meier, D.L., *Science*, **295**, 1688–1691, (2002).
505. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **359**, 801–808, (2005).
506. Koide, S., *Astroph. J.*, **606**, L45–L48, (2004).
507. Blandford, R.D. & Payne, D.G., *Mon. Not. R. Astron. Soc.*, **199**, 883–903, (1982).
508. Uchida, Y., *et al.*, *IAU Symp.*, **195**, 213–222, (2000).
509. Uchida, Y. & Shibata, K., *Publ. Astron. Soc. Japan*, **37**, 515–535, (1985).
510. Meier, D.L., Koide, S., & Uchida, Y., *Science*, **291**, 84–92, (2001).
511. Ustyugova, G.V., *et al.*, *Astroph. J.*, **439**, L39–L42, (1995).
512. Lynden-Bell, D., *Mon. Not. R. Astron. Soc.*, **279**, 389–401, (1996).
513. Uzdensky, D.A. & MacFadyen, A.I., *Astroph. J.*, **647**, 1192–1212, (2006).
514. Meier, D.L., Edgington, P.G., Godon, P., Payne, D.G., & Lind, K.R., *Nature*, **388**, 350–352, (1997).
515. Romanova, M.M., *et al.*, *Astroph. J.*, **500**, 703–713, (1998).
516. Ustyugova, G.V., *et al.*, *Astroph. J.*, **541**, L21–L24, (2000).
517. Kudoh, T., Matsumoto, R., & Shibata, K., *Publ. Astron. Soc. Japan*, **54**, 267–274, (2002).
518. Koide, S., Kudoh, T., & Shibata, K., *Phys. Rev. D*, **74**, 044005, (2006).
519. Ogilvie, G.L. & Livio, M., *Astroph. J.*, **499**, 329–339, (1998).
520. De Villiers, J.-P., Hawley, J.F., Krolik, J.H., & Hirose, S., *Astroph. J.*, **620**, 878–888, (2005).

Chapter 15

521. Lovelace, R.V.E., *et al.*, *Astroph. J. Supp. Ser.*, **62**, 1–37, (1986).
522. Lovelace, R.V.E., Wang, J.C.L., & Sulkanen, M.E., *Astroph. J.*, **315**, 504–535, (1987).
523. Krasnopolsky, R., Li, Z.-Y., & Blandford, R.D., *Astroph. J.*, **595**, 631–642, (2003).
524. Contopoulos, J. & Lovelace, R.V.E., *Astroph. J.*, **429**, 139–152, (1994).
525. Meier, D.L., *Astroph. J.*, **459**, 185–192, (1996).
526. Papaloizou, J.C.B. & Pringle, J.E., *Mon. Not. R. Astron. Soc.*, **208**, 721, (1984).
527. Papaloizou, J.C.B. & Pringle, J.E., *Mon. Not. R. Astron. Soc.*, **213**, 799, (1985).
528. Antonucci, R., *Ann. Rev. Astron. Astroph.*, **31**, 473–521, (1993).
529. Siemiginowska, A. & Elvis, M., *Astroph. J. Supp. Ser.*, **92**, 603–605, (1994).
530. Blandford, R.D. & Levinson, A., *Astroph. J.*, **441**, 79–95, (1995).
531. Wilson, A.S., *ASP Energy Transport in Radio Galaxies and Quasars*, ASP Conf. Ser. **100**, 9–24, (1996).
532. Levinson, A. & Blandford, R.D., *Astroph. J.*, **449**, 86–92, (1995).
533. Sikora, M., *et al.*, *Mon. Not. R. Astron. Soc.*, **280**, 781–796, (1996).
534. Camenzind, M., *Astron. Astroph.*, **162**, 12–44, (1986).
535. Camenzind, M., *Astron. Astroph.*, **184**, 341–360, (1987).
536. Camenzind, M., *Rev. Mod. Astron.*, **8**, 201–233, (1995).
537. Li, Z.-Y., Chiueh, T., & Begelman, M.C., *Astroph. J.*, **394**, 459–471, (1992).
538. Contopoulos, J., *Astroph. J.*, **446**, 67–74, (1995).
539. Contopoulos, J., *Astroph. J.*, **432**, 508–517, (1994).
540. Longair, M.S., Ryle, M., & Scheuer, P.A.G., *Mon. Not. R. Astron. Soc.*, **164**, 243–270, (1973).
541. Scheuer, P.A.G., *Mon. Not. R. Astron. Soc.*, **166**, 513–528, (1974).
542. Norman, M.L., Smarr, L., Winkler, K.-H. A., & Smith, M.D., *Astron. Astroph.*, **113**, 285–302, (1982).

543. Clarke, D.A., Norman, M.L., & Burns, J.O., *Astroph. J.*, **311**, L63–L67, (1986).
544. Lind, K.R., Payne, D.G., Meier, D.L., & Blandford, R.D., *Astroph. J.*, **344**, 89–103, (1989).
545. Hughes, P.A., Aller, H.D., & Aller, M.F., *Astroph. J.*, **341**, 68–79, (1989).
546. Gabuzda, D.C., *Proc. Natl. Acad. Sci.*, **92**, 11393–11398, (1995).
547. Vitrichchak, V.M., Pashchenko, I.N., & Gabuzda, D.C., *Astron. Rep.*, **54**, 269–276, (2010).
548. De Sterck, H., Low, B.C., & Poedts, S., *Phys. Plasmas*, **6**, 954–969, (1999).
549. Vlahakis, N., Tsinganos, K., Sauty, C., & Trussoni, E., *Mon. Not. R. Astron. Soc.*, **318**, 417–428, (2000).
550. Ustyugova, G.V., Koldoba, A.V., Romanova, M.M., Chechetkin, V.M. & Lovelace, R.V.E., *Astroph. J.*, **516**, 221–235, (1999).
551. Polko, P., Meier, D.L., & Markoff, S., *Astroph. J.*, **723**, 1343–1350, (2010).
552. Polko, P., Meier, D.L., & Markoff, S., preprint (2011).
553. Komissarov, S.S., Barkov, M.V., Vlahakis, N., & Königl, A., *Mon. Not. R. Astron. Soc.*, **380**, 51–70, (2007).
554. Duncan, G.C. & Hughes, P.A., *Astroph. J.*, **436**, L119–L122, (1994).
555. Rosen, A., Hughes, P.A., Duncan, G.C., & Hardee, P.E., *Astroph. J.*, **516**, 729–743, (1999).
556. Hardee, P.E., Clarke, D.A., & Howell, D.A., *Astroph. J.*, **441**, 644–664, (1995).
557. Clarke, D.A., *Energy Transport in Radio Galaxies and Quasars*, ASP Conf. Ser. **100**, 311–317, (1996).
558. Nishikawa, K.-I., *et al.*, *Astroph. J.*, **498**, 166–169, (1998).
559. Aloy, M.A., Ibáñez, Marti, J.M., & Müller, E., *Astroph. J. Supp.*, **122**, 151–166, (1999).
560. Hughes, P.A., Miller, M.A., & Duncan, G.C., *Astroph. J.*, **572**, 713–728, (2002).
561. Bicknell, G.V., *Pro. ASA*, **6**, 130–137, (1985).
562. Norman, M.L., Burns, J.O., & Sulkanen, M.E., *Nature*, **335**, 146–149, (1988).
563. Meier, D.L., Sadun, A.C., & Lind, K.R., *Astroph. J.*, **379**, 141–156, (1991).
564. Komissarov, S.S., *Mon. Not. R. Astron. Soc.*, **308**, 1069–1076, (1999).
565. Nakamura, M., Li, H., & Li, S., *Astroph. J.*, **652**, 1059–1067, (2006).
566. Nakamura, M., Li, H., & Li, S., *Astroph. J.*, **656**, 721–732, (2007).
567. Nakamura, M., *et al.*, *Astroph. J.*, **686**, 843–850, (2002).
568. Nakamura, M., Ph. D. thesis, *Formation of the Wiggled Structure of Radio Jets from Active Galactic Nuclei*, Tokyo University of Science (2001).
569. Nakamura, M. & Meier, D.L., *Astroph. J.*, **617**, 123–154, (2004).
570. Mizuno, Y., Hardee, P.E., & Nishikawa, K.-I., *Astroph. J.*, **734**, 19–36, (2011).
571. Mizuno, Y., Hardee, P., & Nishikawa, K.-I., *Astroph. J.*, **662**, 835–850, (2007).
572. McKinney, J.C. & Blandford, R.D., *Mon. Not. R. Astron. Soc.*, **394**, L126–L130, (2009).

Chapter 16

573. Meier, D.L., *Cosmic Explosions in Three Dimensions: Asymmetries in Supernovae and Gamma-Ray Bursts*, ed. P. Höflich, P. Kumar, & J.C. Wheeler, pp. 219–230, (2004).
574. Akiyama, S., Wheeler, J.C., Meier, D.L., & Lichtenstadt, I., *Astroph. J.*, **584**, 954–970, (2003).
575. Meier, D.L., Epstein, R.I., Arnett, W.D., & Schramm, D.N., *Astroph. J.*, **204**, 869–878, (1976).
576. Chen, K. & Ruderman, M., *Astroph. J.*, **402**, 264–270, (1993).
577. Woods, P.M. & Thompson, C., in *Compact Stellar X-Ray Sources*, Lewin, W. & van der Klis, M. ed. (Cambridge Univ. Press, Cambridge, UK, 2004), pp. 547–586.
578. Meliani, M.T., *Pub. Astron. Soc. Aust.*, **16**, 175–205, (1999).
579. Reig, P., Martinez-Nunez, S., & Reglero, V., *Astron. Astroph.*, **449**, 703–710, (2006).
580. Leahy, D. A., *Astron. Astroph. Supp.*, **113**, 21–29, (1995).
581. Migliari, S. & Fender, R.P., *Mon. Not. R. Astron. Soc.*, **366**, 79–91, (2006).
582. Merloni, A., Heinz, S., & Di Matteo, T., *Mon. Not. R. Astron. Soc.*, **345**, 1057–1076, (2003).
583. Merloni, A., Heinz, S., & Di Matteo, T., *Astroph. Sp. Sci.*, **300**, 45–53, (2005).
584. Falcke, H., Kording, E., & Markoff, S., *Astron. Astroph.*, **414**, 895–903, (2004).
585. Kording, E., Falcke, H., & Corbel, S., *Astron. Astroph.*, **456**, 439–450, (2006).

586. Brinkmann, W., *et al.*, *Astron. Astroph.*, **196**, 313–326, (1988).
587. Panferov, A.A., *Astron. Astroph.*, **351**, 156–160, (1999).
588. Fender, R.P., Gallo, E., & Russel, D.O., *Mon. Not. R. Astron. Soc.*, **406**, 1425–1434, (2010).
589. Comment attributed to Martin Rees in Berendzen, R. 1972, ed. *Life Beyond Earth & the Mind of Man*, Proc. of a Boston Univ. symposium, NASA SP-328, p. 1.
590. Mirabel, I.F., *et al.*, *Nature*, **413**, 139–141, (2001).
591. Mirabel, I.F., *et al.*, *Astron. Astroph.*, **395**, 595–599, (2002).
592. Soleri, P. & Fender, R.P.1, *Jets at all Scales, IAU Symp.*, **275**, 265–268, (2011).
593. Muno, M.P., Morgan, E.H., & Remillard, R.A., *Astroph. J.*, **527**, 321–340, (1995).
594. Belloni, T., Mendez, M., King, A.R., van der Klis, M. & van Paradijs, J., *Astroph. J.*, **479**, L145, (1997).
595. Taam, R.E., Chen, X., & Swank, J., *Astroph. J.*, **485**, L83, (1997).
596. Pence, W.D., *et al.*, *Astroph. J.*, **561**, 189–202, (2001).
597. Swartz, D.A., *et al.*, *Astroph. J.*, **574**, 382–397, (2002).
598. Bitzaraki, O.M., *et al.*, *Astron. Astroph.*, **416**, 263–280, (2004).
599. Mukai, K. Pence, W.D. Snowden, S.L., *Astroph. J.*, **582**, 184–189, (2003).
600. Crowther, P.A., *et al.*, *Mon. Not. R. Astron. Soc.*, **408**, 731–751, (2010).
601. Kaspi, S., *et al.*, *Astroph. J.*, **533**, 631–649, (2000).
602. Elitzur, M. & Ho, L.C., *Astroph. J.*, **701**, L91–L94, (2009).
603. Tombesi, F., *et al.*, *Astroph. J.*, **719**, 700–715, (2010).
604. Tombesi, F., *et al.*, *Astron. Astroph.*, **521**, A57, (2010).
605. Grupe, D., *et al.*, *Astron. Astroph.*, **300**, L21–L24, (1995).
606. Grupe, D., *et al.*, *Astron. J.*, **133**, 1998–1994, (2007).
607. Grupe, D., *et al.*, *Astron. J.*, **136**, 2343–2349, (2008).
608. Marscher, A.P., *et al.*, *Nature*, **452**, 966–969, (2008).
609. King, A.L., *et al.*, *Astroph. J.*, **729**, 19–26, (2011).
610. Jones, S., *et al.*, *Mon. Not. R. Astron. Soc.*, **412**, 2641–2652, (2011).
611. McHardy, I.M., *et al.*, *Mon. Not. R. Astron. Soc.*, **348**, 783–801, (2004).
612. Kormendy, J., Bender, R., & Cornell, E., *Nature*, **469**, 374–376, (2011).
613. Kormendy, J. & Bender, R., *Nature*, **469**, 377–380, (2011).
614. Walter, F., *et al.*, *Nature*, **424**, 406–408, (2003).
615. King, A.R., *Mon. Not. R. Astron. Soc.*, **408**, L95–L98, (2010).
616. Malin, D.F., Quinn, P.J., & Graham, J.A., *Astroph. J.*, **272**, L5–L7, (1983).
617. Quinn, P.J., *Astroph. J.*, **279**, 596–609, (1984).
618. Bennert, N.V., *et al.*, arXiv:1102.1975, (2011).
619. Yu, Q. & Tremaine, S., *Mon. Not. R. Astron. Soc.*, **335**, 965–976, (2002).
620. Elvis, M., Risaliti, G., & Zamorani, G., *Astroph. J. Letters*, **565**, L75–L77, (2002).
621. Fabian, A., *Astron. Geophys.*, **50**, 3.18–3.24, (2009).
622. Iwasawa, K., *et al.*, *Mon. Not. R. Astron. Soc.*, **282**, 1038–1048, (1996).
623. Brenneman, L.W. & Reynolds, C.S., *Astroph. J.*, **652**, 1028–1043, (2006).
624. Fabian, A., *et al.*, *Nature*, **459**, 540–542, (2009).
625. Katoaka, J., *et al.*, *Publ. Astron. Soc. Japan*, **59**, 279–297, (2007).
626. Sambruna, R.M., *Astroph. J.*, **700**, 1473–1487, (2009).
627. Evans, D.A., *et al.*, *Astroph. J.*, **710**, 859–868, (2010).
628. White, R.L., *et al.*, *Astroph. J. Supp.*, **126**, 133–207, (2000).
629. Garofalo, D., Evans, D.A., & Sambruna, R.M., *Mon. Not. R. Astron. Soc.*, **406**, 975–986, (2010).
630. Volonteri, M., Sikora, M., & Lasota, J.-P., *Astroph. J.*, **667**, 704–713, (2007).
631. Junor, W., Biretta, J.A., & Livio, M., *Nature*, **401**, 891–892, (1999).
632. Asada, K., Nakamura, M., Doi, A., Nagai, H., & Inoue, M., *Proc. IAU S275*, **6**, 198–199, (2011).
633. Cheung, C.C., Harris, D.E., & Stawarz, Ł., *Astroph. J.*, **663**, L65–L68, (2007).
634. Madrid, J.P., *Astron. J.*, **137**, 3864–3868, (2009).
635. Nakamura, M., Garofalo, D., & Meier, D.L., *Astroph. J.*, **721**, 1783–1789, (2010).
636. Xu, Y.-D., Cao, X., & Wu, Q., *Astroph. J.*, **694**, L107–L110, (2009).
637. Lyubarsky, Y.E., *Mon. Not. R. Astron. Soc.*, **402**, 353–361, (2010).

Appendix C

638. Misner, C.W. & Sharp, D.H., *Phys. Rev.*, **136**, B571–B576, (1964).
639. May, M.M. & White, R.H., *Phys. Rev.*, **141**, 1232–1241, (1966).

Appendix E

640. Breitmoser, E. & Camenzind, M., *Astron. Astroph.*, **361**, 207–225, (2000).

Appendix F

641. Mestel, L., *Mon. Not. R. Astron. Soc.*, **122**, 473–478, (1961).

Glossary

3C	Third Cambridge catalog (of radio sources)	21
3CR	Third Cambridge catalog (Revised)	20
3CRR	Third Cambridge catalog (Revised a second time)	21
AAAS	American Association for the Advancement of Science	106
AAS	American Astronomical Society	9
ADAF	Advection-Dominated Accretion Flow	520
ADIOS	Advection-Dominated Inflow–Outflow Solutions (for accretion disk winds)	567
ADM formalism	Arnowitt–Deser–Misner formalism for expressing Einstein’s equations	246
AGN	Active Galactic Nucleus	13
AIGRMHD	Adiabatic Ideal General Relativistic MagnetoHydroDynamics	335
AP	Alfvén Point (on the MHD Alfvén surface)	657
APS	American Physical Society	273
AS	Alfvén Surface	368
ASCA	Advanced Satellite for Cosmology and Astrophysics	41
ASIAA	Astronomica Sinica Institute of Astronomy and Astrophysics (Taiwan)	819
ASJ	Astronomical Society of Japan	651
ASP	Astronomical Society of the Pacific	48
ATNF	Australia Telescope National Facility	94
AU	Astronomical Unit	8
AUI	Associated Universities, Inc.	19
AXP	Anomalous X-ray Pulsar	70
BAL	Broad Absorption Line (QSO)	49

BATSE	Burst And Transient Source Experiment, on the Compton Gamma-Ray Observatory	101
BEL	Broad Emission Line (QSO)	566
BH	Black Hole	109
BL	Boyer–Lindquist (coordinate system)	235
blazar	Generic name for highly variable AGN (BL Lacertae objects, OVV quasars, <i>etc.</i>)	32
BLR	Broad-Line Region (of AGN)	16
BLRG	Broad-Line Radio Galaxy	22
blue blazar	see HBL	32
BP	Blandford and Payne 1982 paper on non-relativistic MHD winds from accretion disk [507]	645
BSO	Blue Stellar Object (early name for QSO)	43
BSSN method	Baumgarte–Shapiro–Shibata–Nakamura method for solving Einstein’s equations	272
BZ	Blandford and Znajek 1977 paper on model for black hole magnetospheres [484]	616
CD	Current-Driven	717
CDI	Current-Driven Instability	717
CfA	Center for Astrophysics (Harvard University)	820
CGRO	Compton Gamma-Ray Observatory	101
CGS	Centimeter–Gram–Second (system of units)	513
Chandra	Chandra X-ray mission	120
CITA	Canadian Institute for Theoretical Astrophysics	49
CND	Circum-Nuclear Disk	54
CNO	Carbon–Nitrogen–Oxygen (nuclear burning cycle)	384
CO	Carbon Monoxide	456
Compton depth (y)	measure of photon optical depth and energy transfer by Compton electron scattering	333, 513
CR	Co-Rotation (disk radius)	605
CS	Cusp Surface	367
CSIRO	Commonwealth Scientific and Industrial Research Organisation, Australia	67
CUP	Cambridge University Press	729
CV	Cataclysmic Variable (binary star)	405
DD	Doubly-Degenerate (binary star)	402
EBBH	Equal-mass Binary Black Hole	268
ED	ElectroDynamics	656
Eddington ratio	ratio of a source’s total radiative luminosity to the Eddington luminosity for its mass	60
EGRET	Energetic Gamma-Ray Experiment Telescope, on the Compton Gamma-Ray Observatory	89

EM	ElectroMagnetic	192
EMRIBH	Extreme Mass-Ratio Inspiral Black Hole	268
ESAC	European Space Astronomy Centre	741
ESO	European Southern Observatory	117
EUV	Extreme UltraViolet (radiation)	806
FB	Flaring Branch (of neutron star Z sources)	74
FBG diagram	Fender–Belloni–Gallo diagram; see HID	95
FFDE	Force-Free Degenerate Electrodynamics	335
FIDO	FIDucial Observer (coordinate system)	225
FIX	FIXed (FIDO) coordinate system	224
FMS	Fast Magnetosonic Surface; classical fast surface	368
FMSS	Fast Magnetosonic Separatrix Surface	682
FR	Fanaroff & Riley radio source morphological classification, 1974 paper [38]	21
FSRQ	Flat Spectrum Radio Quasar	25
GHz	Giga-Hertz (billion cycles per second)	36
GJ	Goldreich–Julian (pulsar magnetosphere model)	594
GLAST	Gamma-ray Large Area Space Telescope; a.k.a. Fermi gamma-ray telescope	89
GRB	Gamma-Ray Burst	65
GRHD	General Relativistic HydroDynamics	335
GRMHD	General Relativistic MagnetoHydroDynamics	297
GRO	Gamma-Ray Observatory; see CGRO	94
GRS	GRanat Source	92
GSFC	Goddard Space Flight Center	99
GSS	Grad–Schlüter–Shafranov (general relativistic magnetosphere equation)	345
GW	Gravitational Wave	256
Gyr	Giga-year (billion years)	36
H I	atomic Hydrogen	456
H II	ionized Hydrogen	53
HartRAO	Hartebeesthoek Radio Astronomy Observatory	36
HB	Horizontal Branch (of neutron star Z sources)	74
HBL	High-frequency BL Lacertae object	32
HD	HydroDynamics	359
HF QPO	High Frequency Quasi-Periodic Oscillation	80
HiBAL	High-ionization Broad Absorption Line QSOs	49
HID	Hardness–Intensity Diagram	788
HLX	Hyper-Luminous X-ray source	117
HMXB	High-Mass X-ray Binary	75

horizon, (magneto)sonic	locus of points in space beyond which component of flow characteristics along streamlines all have the same sign	681
horizon, black hole	Surface surrounding the collapsed star interior to which events cannot affect the outside universe	3, 224, 236
HP	Horizon-Penetrating (coordinate system)	234
HR diagram	Hertzsprung–Russell (color–magnitude) diagram	96
HST	Hubble Space Telescope	19
HyLIRG	Hyper-Luminous InfraRed Galaxy	45
HYMOR	HYbrid MORphology (FR I/II) radio source	22
ICC	Interstellar Cloud Core	321
IDV	Intra-Day Variable blazar	34
IGM	InterGalactic Medium	22
IKI	Institut Kosmicheskix Issledovanii (Space Research Institute, Moscow)	334
IMBH	Intermediate Mass Black Hole	13
IMF	Initial Mass Function (for newborn stars)	99
INTEGRAL	INTErnational Gamma-Ray Astrophysics Laboratory	127
IR	InfraRed (radiation)	xx
IRAS	Infrared Astronomy Satellite	44
IRS	InfraRed Source	125
IS	Island State (of neutron star atoll sources)	74
ISCO	Innermost Stable Circular Orbit (of a black hole)	231
ISM	InterStellar Medium	21
IXO	Intermediate X-ray luminosity Object; ULX	117
JCMT	James Clerk Maxwell Telescope	45
JPL	Jet Propulsion Laboratory	xxi
JWST	James Webb Space Telescope	111
KACST	King Abdulaziz City for Science and Technology	36
KER	KERr (metric)	235
KER–NEW	KERr–NEWman (metric)	244
KFD	Kinetic- (energy) Flux-Dominated	663
KH	Kelvin–Helmholtz	717
KHI	Kelvin–Helmholtz Instability	717
KITP	Kavli Institute for Theoretical Physics	598
LB	Lower Banana state (of neutron star atoll sources)	74
LBL	Low-frequency BL Lacertae object	32
LBV	Luminous Blue Variable (star)	474
LCB	Li, Chiueh, and Begelman 1992 paper on cold, relativistic MHD winds [537]	673

LF QPO	Low Frequency Quasi-Periodic Oscillation	81
LGRB	Long-duration Gamma-Ray Burst	102
LIGO	Laser Interferometer Gravitational wave Observa- tory	255
LINER	Low-Ionization Nuclear Emission-line Region	53
LIRG	Luminous InfraRed Galaxy	44
LISA	Laser Interferometer Space Antenna	255
LLAGN	Low-Luminosity AGN	52
LLNL	Lawrence Livermore National Laboratory (formerly Lawrence Radiation Laboratory)	391
LMC	Large Magellanic Cloud	87
LMXB	Low-Mass X-ray Binary	75
LNRF	Local Non-rotating Reference Frame (FIDO coordi- nate system)	237
LoBAL	Low-ionization Broad Absorption Line QSOs	49
LSU	Louisiana State University	708
MACHO	MAssive Compact Halo Object	87
MBH	Massive Black Hole	790
MCG	Morphological Galaxy Catalog	41
MDAF	Magnetically-Dominated Accretion Flow	531
MERLIN	Multi-Element Radio Linked Interferometry Net- work	29
MFP	Modified Fast Point (on the MHD modified fast sur- face)	657
MFS	Modified Fast Surface; FMSS	661
MHD	MagnetoHydroDynamics	xx
MHz	Mega-Hertz (million cycles per second)	21
microquasar	Binary black hole system, usually with a jet; μ QSR	13
MIT	Massachusetts Institute of Technology	93
MK	Mega-Kelvin; million kelvins	78
MOV	MOVing-body (coordinate system)	224
MPG	Max-Planck-institute for Gravitational physics	258
MRI	Magneto-Rotational shearing Instability	374
MSFC	Manned Space Flight Center	117
MSP	Modified Slow Point (on the MHD modified slow surface)	684
MSS	Modified Slow Surface; SMSS	684
MSSSO	Mount Stromlo & Siding Springs Observatories	87
Myr	Mega-year (million years)	54
N-galaxy	Galaxy (generally elliptical) with a bright Nucleus; NLRG, BLRG	23
NAOJ	National Astronomical Observatory of Japan	246
NB	Normal Branch (of neutron star Z sources)	74

neutrinosphere	the surface of last neutrino absorption in a source	387
NGC	New General Catalog (of Nebulae and Clusters of Stars)	14
NLR	Narrow-Line Region (of AGN)	16
NLRG	Narrow-Line Radio Galaxy	22
NLSy1	Narrow-Line Seyfert (Type 1) galaxy	42
NOAO	National Optical Astronomy Observatory	60
NRAF	Non-Radiative Accretion Flow (computational approximation to RIAF/ADAF)	549
NRAO	National Radio Astronomy Observatory	22
NRHD	Non-Relativistic HydroDynamics	355
NRMHD	Non-Relativistic MagnetoHydroDynamics	335
NS	Neutron Star	109
NSF	National Science Foundation, USA	19
NuSTAR	Nuclear Spectroscopic Telescope ARray	46
OGLE	Optical Gravitational Lensing Experiment	87
OIS	Observer-at-Infinity/Synchronous (coordinate system)	238
ONeMg	Oxygen–Neon–Magnesium (stellar core)	386
peribarathron	minimum distance of star orbiting a BH	447
PDF	Poynting- (energy) Flux-Dominated	663
photosphere	the surface of last photon absorption in a source	484, 569
PNS	Proto-Neutron Star	108
PP	Papaloizou–Pringle (instability)	670
PWN	pulsar wind nebula	69
QPO	Quasi-Periodic Oscillation	80, 90
QSO	Quasi-Stellar Object; optically-identified quasar (usually RQQ)	43
QSR	Quasi-Stellar Radio source; quasar	13
quasar	QUAsi-StellAr Radio source	13
red blazar	see LBL	33
RIAF	Radiatively-Inefficient Accretion Flow; ADAF	55
RIKEN	RIKEN science institute (Japan)	680
RLQ	Radio Loud Quasar; classical quasar	13
RMS	Root Mean Square	325
ROSAT	ROentgen SATellite	72
RQQ	Radio Quiet Quasar	47
RRAT	Rotating Radio Transient	71
RSG	Red SuperGiant star	395
RXTE	Rossi X-ray Timing Explorer	93

SAS-1	First Small Astronomy Satellite (Uhuru)	73
SAS-2	Second Small Astronomy Satellite	72
SASI	Standing Accretion Shock Instability	391
SBH	Stellar-mass Black Hole	88
scattersphere	the surface of last photon scattering in a source	567
SCH	SCHwarzschild (metric)	222
SCUBA	Submillimeter Common-User Bolometer Array	45
SDSS	Sloan Digital Sky Survey	48
SEW	Super-Eddington Wind	582
Seyfert	Galaxy (generally spiral) with a bright nucleus, originally discovered by Carl Seyfert; AGN	16
SGR	Soft Gamma-ray Repeater	70
SGRB	Short-duration Gamma-Ray Burst	102
SH	Schwarzschild–Hilbert (coordinate system)	222
SHB	Short Hard Burst; SGRB	102
SIM	Space Interferometer Mission	435
SIS	Singular Isothermal Sphere (distribution of stars)	417
SISSA	Scuola Internazionale Superiore di Studi Avanzati	508
SLE	Shapiro–Lightman–Eardley (accretion disk solution)	512
SMBH	SuperMassive Black Hole	3
SMC	Small Magellanic Cloud	91
SMG	SubMillimeter Galaxy	45
SMS	Slow Magnetosonic Surface	368
SMSS	Slow Magnetosonic Separatrix Surface	682
SN	SuperNova	65
SNR	SuperNova Remnant	118
SPH	Smooth Particle Hydrodynamics	155
SPL	Steep Power-Law (X-ray binary accretion state)	777
SRMHD	Special Relativistic MagnetoHydroDynamics	335
SS	Shakura–Sunyaev (accretion disk solutions)	496
STScI	Space Telescope Science Institute	117
surface, critical	locus of points in space where flow characteristics appear or disappear	362, 681
surface, separatrix	locus of points in space where component of flow characteristics along a streamline changes sign	368, 681
surface, singular	locus of points in space where denominator of a mathematical accretion/wind equation vanishes	488, 682
SXT	Soft X-ray Transient source	77
TOV equation	Tolman–Oppenheimer–Volkoff relativistic stellar structure equation	287
traceless-Lorenz gauge	see TT gauge	256
TT gauge	Transverse–Traceless gauge for the Einstein equations	256

UB	Upper Banana state (of neutron star atoll sources)	74
UC	University of California	287
UCLA	University of California Los Angeles	46
UCSB	University of California Santa Barbara	552
UCSC	University of California Santa Cruz	xix
UFO	Ultra-Fast Outflow (from AGN central engines)	794
ULIRG	Ultra-Luminous InfraRed Galaxy	45
ULX	Ultra-Luminous X-ray source	117
UNAM	Universidad Nacional Autónoma de México	91
UT	University of Texas	122
UV	UltraViolet (radiation)	xx
UVOIR	UltraViolet–Optical–InfraRed (radiation)	24
VH	Very High (X-ray binary accretion state)	777
VHS	Very High State (for X-ray binaries)	90
VK	Vlahakis and Königl 2003 paper on warm, relativistic MHD winds [368]	692
VLA	Very Large Array	19
VLBA	Very Long Baseline Array	819
VLBI	Very Long Baseline Interferometry	17
VMS	Very Massive Star	118, 410
VTST	Vlahakis, Tsinganos, Sauty, and Trussoni 2000 paper on warm, non-relativistic MHD winds [549]	685
WC	Wolf–Rayet star with strong Carbon emission lines	395
WD	White Dwarf	402
WLRG	Weak-Lined Radio Galaxy	53
WN	Wolf–Rayet star with strong Nitrogen emission lines	395
WNE	see WN	396
WO	Wolf–Rayet star with strong Oxygen emission lines	395
WPVS	Wamsteker, Prieto, Vitores, Schuster <i>et al.</i> H α galaxy survey	796
XDIN	X-ray Dim Isolated Neutron star	71
XMM	X-ray Multi-mirror Mission; a.k.a., XMM Newton	120
XRb	X-ray binary	7
XTE	X-ray Timing Explorer; see RXTE	95
ZAMO	Zero-Angular-Momentum Observer (FIDO coordinate system)	237
ZAMS	Zero-Age Main Sequence (main locus of stars in HR diagram)	785

Index of Names

- Abel, Tom, 414
Abramowicz, Marek, 95, 515
Akiyama, Shizuka, 730
Alfvén, Hannes O. G., 364
Aloy, Miguel Ángel, 708
Andrew, Bryan H., 30
Anninos, Peter, 336
Antonucci, Robert, 38, 39, 671
Argelander, Friedrich W. A., 30
Arnowitt, Richard L., 246
- Baade, W. H. Walter, 68, 84
Backer, Donald, 83
Bahcall, John N., 48
Baker, John, 272, 276
Balbus, Steven A., 374, 547, 551
Bash, Frank N., 123
Baumgardt, Holger, 123
Baumgarte, Thomas W., 272
Becklin, Eric E., 124, 125
Begelman, Mitchell C., 119, 391, 492,
515, 555, 582, 656, 673, 674,
684, 782
Bell, S. Jocelyn, 64
Belloni, Tomaso, 93, 95, 776
Berger, Edo, 110
Bicknell, Geoffrey, 22, 710, 822
Biermann, Peter L., 589
Blaes, Omer M., 552
- Blandford, Roger D., 17, 26, 39, 479,
582, 610, 614, 624, 636, 642,
655, 656, 660, 664, 666, 672,
685, 701, 714, 721, 759, 761
Bogovalov, Sergey V., 656, 680, 687
Bolton, John G., 64
Bondi, Sir Herman, 439, 442, 487
Boroson, Todd A., 60, 792
Boyer, Robert H., 235
Bromm, Volker, 414
Browne, Ian W. A., 29
Burns, Jack, 713
Burrows, Adam, 391
- Camenzind, Max, 672
Camm, G. Leslie, 448
Campanelli, Manuela, 272, 274
Campbell, W. Wallace, 14
Castro-Tirado, Alberto J., 92, 105
Celotti, Annalisa, 32
Centrella, Joan M., 272
Chadwick, Sir James, 68
Chandrasekhar, Subrahmanyan, 63, 161
Chen, Xingming, 525
Cheung, C. C., 820
Chiang, James, 49, 564
Chiu, Hong-Yee, 24
Chiueh, Tzihong, 673, 674, 684
Clarke, David, 706, 710, 713
Cohen, Marshall H., 27

- Cohn, Haldan, 448
 Colgate, Stirling, 650
 Comastri, Andrea, 32
 Combes, Françoise, 453
 Contopoulos, Ioannis (John), 597, 628,
 656, 665, 673, 675, 678, 680,
 692
 Conway, John, 29
 Coroniti, Ferdinand V., 620, 636
 Couette, Maurice M. A., 371
 Cowling, Thomas G., 551
 Curtis, Heber D., 14, 19
 Czerny, Božena, 515, 525
- Danby, J. M. Anthony, 448
 De Villiers, Jean-Pierre, 336, 556, 625,
 653
 Deser, Stanley, 246
 Di Matteo, Tiziana, 751
 Doppler, Christian A., 189
 Dreyer, John L. E., 14
 Duncan, G. Comer, 705, 708, 721
 Duncan, Robert C., 70
 Dunlop, James S., 59
- Eardley, Douglas M., 507, 512, 514,
 519
 Ebisuzaki, Toshikazu, 431, 464, 808
 Eddington, Sir Arthur, 5, 10
 Edgington, Samantha, 648
 Eggleton, Peter P., 472
 Einstein, Albert, 165, 177, 209, 210,
 217, 273
 Eisenhardt, Peter M. E., 45
 Elitzur, Moshe, 792
 Elvis, Martin, 816
 Emden, J. Robert, 157
 Esin, Ann A., 527
 Euclid of Alexandria, 166, 176
 Euler, Leonhard, 155
- Faber, Sandra M., 429, 443
 Falcke, Heino, 587, 589, 759
 Fanaroff, Bernard L., 21
 Fath, Edward A., 14
- Fender, Robert, 74, 95, 741, 747, 751,
 755, 766, 773
 Fendt, Christian, 597
 FitzGerald, George F., 179
 Flammang, Richard A., 489
 Fossati, Giovanni, 32, 34
 Fowler, William A., 63
 Fragile, P. Christopher, 336, 558, 650
 Frank, Juhan, 448
 Freitag, Marc D., 422
 Friedrichs, Kurt O., 364
 Fryer, Chris L., 393, 730
- Gallo, Elena, 95, 766
 Gammie, Charles F., 336, 339
 Garofalo, David, 632, 672, 817, 820,
 830
 Gebhardt, Karl, 121
 Gehrels, Neil, 110
 Genzel, Reinhard, 54, 125
 Gerssen, Joris, 122
 Ghez, Andrea, 8, 125
 Ghisellini, Gabriele, 32
 Giacconi, Riccardo, 63, 64
 Gillessen, Stefan, 54, 55
 Godon, Patrick, 648
 Goldreich, Peter, 594, 597, 599
 Grad, H., 344
 Green, George, 145
 Greiner, Jochen, 93
 Griffiths, Richard E., 118
 Groot, Paul J., 407
 Gunn, James E., 68
 Gursky, Herbert, 63
- Haehnelt, Martin G., 429
 Halpern, Jules P., 42, 72
 Hameury, Jean-Marie, 77
 Hansen, Brad M. S., 125
 Hardee, Philip, 706, 719
 Harding, Alice K., 603
 Harmon, B. Alan, 94
 Harris, Daniel E., 820
 Harrison, Fiona A., 46
 Hawley, John F., 336, 374, 547, 551,
 625

- Hazard, Cyril, 23
 Heckman, Timothy M., 53
 Heger, Alexander, 393, 397, 481, 730, 744
 Heinz, Sebastian, 751
 Hewish, Antony, 64
 Hirose, Shigenobu, 552, 625, 653
 Hirotani, Kouichi, 634, 639
 Hjellming, Robert M., 94
 Ho, Luis C., 57, 123, 792
 Hoffmeister, Cuno, 30
 Höflich, Peter, 406
 Holt, Stephen S., 72
 Honma, Fumio, 508
 Hoyle, Sir Fred, 439, 442
 Hubble, Edwin P., 14
 Hughes, Philip A., 705, 708–710, 721
 Hughes, Scott A., 479
 Hulse, Russell A., 84
 Hutchings, John B., 25, 48

 Ichimaru, Setsuo, 519

 Jackson, John D., 165
 Jackson, Robert E., 429, 443
 Janiuk, Agnieszka, 525
 Jeans, Sir James H., 412
 Jogee, Shardha, 454, 456
 Julian, William H., 594, 597, 599

 Kafka, Peter, 35
 Kato, Shoji, 508
 Kauffmann, Guinevere A. M.-I., 429
 Kaufman, Michele, 45
 Kazanas, Demosthenes, 597
 Kerr, Roy P., 234
 Khachikian, Edward Ye., 16
 Khokhlov, Alexei, 406
 King, Andrew, 77, 119, 430, 805, 812
 King, Ivan, 417
 van der Klis, Michiel, 74, 737, 739, 742, 826
 Kluźniak, Włodek, 95
 Koide, Shinji, xx, 238, 240, 246, 250, 336, 620, 629, 639, 641, 651
 Koldoba, Aleksander V., 605
 Komissarov, Serguei S., xx, 246, 335, 336, 339, 614, 617, 619, 628, 639, 641, 700, 716
 Königl, Arie, 26, 39, 656, 685, 692, 694, 699
 Kormendy, John, 801
 Krasnopolsky, Ruben, 664
 Krolik, Julian H., 552, 625
 Kudoh, Takahiro, 246, 629, 639, 650
 Kulsrud, Russell M., 448

 Lacy, Mark, 59
 Lagrange, Joseph-Louis, 148
 Lane, Jonathan H., 157
 Larson, Richard B., 432
 Lasota, Jean-Pierre, 77, 91, 515
 LeBlanc, James M., 391, 704
 Li, Hui, 650
 Li, Zhi-Yun, 664, 673, 674, 684
 Lightman, Alan P., 448, 507, 512, 514, 519
 Lind, Kevin, 648, 710, 714
 Lindquist, Richard W., 235
 Liu, Bifang, 525
 Livio, Mario, 498, 642, 653
 Long, Knox S., 117
 Longair, Malcolm S., 701
 Lorentz, Hendrik A., 185, 194
 Lorenz, Ludwig V., 194
 Lousto, Carlos O., 272
 Lovelace, Richard V. E., 605, 655, 656, 665, 673, 690
 Lynden-Bell, Donald, 419, 646, 650
 Lyttleton, Raymond A., 439

 Maccarone, Thomas J., 120, 128
 Macdonald, Douglas A., 246
 MacFadyen, Andrew I., 108, 646
 MacLeod, John M., 30
 Mahadevan, Rohan, 520
 Maillard, Jean-Pierre, 126
 Malin, David, 15
 Maraschi, Laura, 32
 van der Marel, Roeland P., 57
 Margon, Bruce H., 81
 Markarian, Benik E., 16

- Markoff, Sera, 587, 695, 696, 757, 759, 818, 821
 Marscher, Alan, 821
 Matsumoto, Ryoji, 508, 650
 May, Michael M., 841
 Mazurek, Ted J., 406
 McClintock, Jeffrey E., 762, 775, 777, 801
 McKinney, Jonathan C., 336, 339, 558, 617, 619, 628, 638, 698, 700, 721
 McLure, Ross J., 59
 McMillan, Stephen L. W., 118, 422, 784
 Meier, David L., 22, 309, 391, 480, 515, 558, 568, 648, 650, 670, 695, 710, 714, 717, 782, 847
 Meissner, F. Walther, 619
 Merloni, Andrea, 751
 Merritt, David, 432, 448
 Messier, Charles, 14, 64
 Mészáros, Peter, 103
 Metzger, Mark R., 101
 Meyer, Friedrich, 525
 Meyer-Hofmeister, Emmi, 525
 Michel, F. Curtis, 614
 Migliari, Simone, 741, 747, 751, 773
 Miller, John C., 508
 Miller, Joseph S., 38, 39
 Miller, Mark A., 405, 708, 721
 Milosavljević, Miloš, 125, 432
 Minkowski, Hermann, 68, 178
 Minkowski, Rudolph L. B., 68
 Mirabel, I. Felix, 91, 93, 94
 Misner, Charles W., 165, 246, 841
 Mizuno, Yosuke, 719
 Moore, Joseph H., 14
 Morgan, Edward H., 93, 776
 Muno, Michael P., 776
 Murdin, Paul G., 88
 Murphy, David W., 29
 Murray, Norman, 49, 564
 Nakamura, Fumitaka, 414
 Nakamura, Masanori, 356, 650, 710, 712, 717, 718, 820, 822
 Nakamura, Takashi, 272
 Narayan, Ramesh, 520, 583, 638
 Natarajan, Priyamvada, 479
 Nelemans, Gijs, 400, 407
 Neugebauer, Gerald, 124, 125
 Newman, Ezra T., 244
 Newton, Sir Isaac, 133, 166
 Nishikawa, Ken-Ichi, 708, 719
 Nordström, Gunnar, 210
 Norman, Michael L., 704, 710, 713
 Noyola, Eva, 122
 O'Leary, Ryan M., 421
 Ogilvie, Gordon, 498, 642, 653
 Oppenheimer, J. Robert, 287
 Osmer, Patrick S., 43
 Osterbrock, Donald, xix, 42
 Ostriker, Jeremiah P., 68
 Owen, Frazer N., 22
 Paczyński, Bohdan, 103, 510
 Padovani, Paolo, 60
 Paolini, Frank, 63
 van Paradijs, Johannes (Jan), 101
 Parker, Eugene N., 486, 487
 Payne, David G., 642, 648, 656, 660, 666, 672, 685, 714, 759
 Pearson, Timothy J., 27
 Perley, Richard A., 29
 Phinney, E. Sterl, 636
 Pineda, Jorge L., 452
 Piran, Tsvi, 106
 Pogge, Richard, 42
 Poisson, Siméon-Denis, 145
 Polko, Peter, 695, 697, 759
 Portegies Zwart, Simon F., 118, 126, 422, 784
 Postnov, Konstantin A., 399, 401
 Pretorius, Frans, 272
 Price, Richard M., 246, 512
 Pringle, James, 479, 498, 514
 Proga, Daniel, 49
 Ptak, Andrew, 118
 Punsly, Brian, 620, 636

- Pythagoras of Samos, 166
- Quataert, Eliot, 520
- Readhead, Anthony C. S., 17, 18, 29
- Rees, Sir Martin J., 17, 27, 103, 448, 576, 701
- Remillard, Ronald A., 93, 762, 775–777, 801
- Rhoads, James E., 105
- Ricci-Curbastro, Gregorio, 218
- Rich, R. Michael, 122
- Riemann, G. F. Bernhard, 213
- Riley, Julia M., 21
- Roche, Édouard A., 446
- Rodriguez, Luis F., 91, 93
- Romanova, Marina M., 605, 606, 608, 650, 690, 737
- Rosen, Alexander, 705
- Rossi, Bruno B., 63
- Róžańska, Agata, 525
- Ruderman, Malvin A., 601
- Rupen, Michael P., 94
- Russell, David, 766
- Ryle, Sir Martin J., 17, 64, 701
- Sadun, Alberto, 710
- Salmonson, Jay D., 336
- Salpeter, Edwin E., 99, 803
- Sathyaprakash, B. S., 258, 267
- Sauty, Christophe, 680, 685
- Scalo, John M., 411
- Scheuer, Peter A. G., 701
- Schlüter, A., 344
- Schmidt, Maarten, 23, 36, 52, 818
- Schmitt, John L., 30
- Schwarzschild, Karl, 222
- Schwarzschild, Martin, 222
- Seyfert, Carl K., 16
- Shafranov, Vitalii D., 344
- Shakura, Nikolay I., 494, 496, 507, 551
- Shapiro, Stuart L., 272, 286, 448, 512, 514
- Sharp, David H., 841
- Shibata, Kazunari, 246, 629, 639, 644, 645, 648, 650
- Shibata, Masaru, 272
- Shields, Gregory A., 447, 478
- Shklovskii, Iosif S., 73
- Slipher, Vesto M., 14
- Smarr, Larry, 704
- Smith, Harding E., 383
- Smith, Malcolm G., 43
- Smith, Michael D., 704
- Soleri, Paolo, 766
- Spitkovsky, Anatoly, 599, 733
- Stanek, Krzysztof Z., 114
- Stawarz, Łukasz, 820
- Stern, Daniel, 45
- Sunyaev, Rashid A., 334, 494, 496, 507, 551, 795
- Sutherland, Peter G., 601
- Szuskiewicz, Ewa, 508, 515
- Taam, Ron, 776
- Takahashi, Masaaki, 620, 626, 628, 631, 635
- Tan, Loke K., 122
- Tanaka, Yasuo, 41
- Tassis, Konstantinos, 452, 454
- Taylor, Joseph H., 84
- Teukolsky, Saul A., 286
- Thompson, Christopher, 70
- Thorne, Kip S., 165, 242, 246, 269, 489, 512
- Thorsett, Stephen E., 113
- Tingay, Steven, 94
- Titarchuk, Lev G., 334, 795
- Tolman, Richard C., 287
- Tomimatsu, Akira, 626, 628, 631, 635
- Tóth, Gábor, 336
- Tout, Christopher A., 400
- Tremaine, Scott, 816
- Trussoni, Edoardo, 680, 685
- Tsinganos, Kanaris, 680, 685
- Turner, Neal J., 552
- Turnshek, David, 50, 51
- Uchida, Yutaka, 644, 645, 648
- Ulvestad, James S., 122

- Umemura, Masayuki, 414
Urry, C. Megan, 60
Ustyugova, Galina V., 605, 608, 645,
650, 656, 690
Uzdensky, Dmitri A., 598, 628, 630,
633, 646, 855

Van Riper, Kenneth, 286
Velusamy, Thangasamy, 452
Vlahakis, Nektarios, 656, 685, 692,
694, 699
Volkoff, George M., 287

Wald, Robert M., 618
Walker, R. Craig, 15
Wang, Lifan, 86, 729
Weaver, Thomas A., 397, 744
Webster, B. Louise, 88
Weedman, Daniel W., 16
Werner, Michael J., 43
Wheeler, J. Craig, 391, 406, 447, 478,
480, 729
Wheeler, John A., 24, 48, 165
White, Richard H., 841
Wiita, Paul J., 510
Wilson, James R., 391, 480, 704
Winkler, Karl-Heinz A., 704
Wood, Roger, 419
Woosley, Stanford E., 108, 383, 393,
397, 730, 744
Wright, Edward L., 45

Yi, Insu, 520, 583
Young, Peter J., 447, 478
Yu, Qingjuan, 816

Zepf, Stephen E., 120
Zhang, S. Nan, 94
Znajek, Roman, 610, 614, 636, 761
Zwicky, Fritz, 68, 84
Żytkow, Anna N., 489

Subject Index

- Accretion of angular momentum, 437, 475–481, 537–541, 558
 - in binary systems, 476–477
 - in black hole mergers, 268–277, 479–480
 - spin–orbit coupling, 274
 - in collapsing supernova cores, 480–481
 - onto central BHs in AGN/globular clusters, 477–480
 - onto central BHs in galaxy mergers, 478–480, 817–818, 830
 - accretion disk/hole alignment, 479, 480, 830
 - BH spinup and spindown, 476, 478–480, 616, 635, 818
 - galaxy core spinup, 478
 - spin–orbit coupling, 277, 479
- Accretion of matter, carburetion of fuel
 - in binary systems
 - Roche/tidal stripping of companion star, 471–474
 - stars in galactic centers, 444–448
 - BH mass limit for tidal disruption, 447
 - loss cone, angle, 448
 - loss cone, depletion, 434, 448
 - loss cone, refilling, 434, 448
 - tidal disruption accretion rate, 448–451
 - tidal disruption of stars by BHs, 446–448
 - tidal orbit, 447
 - tidal radius/Roche limit, 446, 461–463, 465, 466, 468, 483
- Accretion of matter, efficiency
 - compared with nuclear burning, 438
 - onto black holes, 231, 239, 438
 - onto Jupiter, 438
 - onto neutron stars, 439
 - onto the earth, 438
 - onto the Sun, 438
 - onto white dwarfs, 438
- Accretion of matter, general, 437–475
 - critical surfaces, 698, 699
 - in binary systems, 470–475
 - in collapsing supernova cores, 475
 - onto BHs in the ISM, 439–442
 - onto central BHs in AGN/globular clusters, 442–470
 - onto central BHs in galaxy mergers, 790, 802–806, 812–813
 - stagnation surface, 698, 699
 - wind, ingoing, 698, 699
 - wind, outgoing, 699

- Accretion of matter, magnetically-dominated disk theory, 529–547
 - magnetically-advective (transitional) accretion, 532–536, 768
 - physical structure, 532–533
 - stability, 535
 - thermal structure, 533
- magnetically-dominated accretion
 - flow (MDAF), 535–547, 768, 825
 - ADAF vs. MDAF, 546–547
 - example MDAF disk, 542–544
 - MDAF “end-game” accretion, 544–546
 - physical structure (magneto-centrifugal accretion), 535–542
 - singular (Alfvén) surface, 538
 - singular and non-singular solutions, 537–541
 - thermal structure, 542
 - observational appearance, 766–771
 - simulations, 558–560
- Accretion of matter, numerical simulations
 - general MRI simulations, 547–560
 - limitations, 548–549
 - power and promise, 547–548
 - global MRI simulations, 555–560
 - non-radiative, with jets, 555–558
 - radiative, 558–560
 - local MRI simulations, 549–553
 - Cowling anti-dynamo theorem, 549, 551
 - Maxwell stress, 549–551
 - Reynolds stress, 549–551
 - shearing box approximation, 549
 - semi-local MRI simulations, 551–555
 - radiation-pressure-dominated, 552–555
- Accretion of matter, sources of fuel
 - gas in early-type galaxy mergers
 - cluster inspiral does not fuel SMBHs, 469–470, 811–812
 - interstellar gas from mergers, 812–813
 - gas in late-type galaxy centers, 451–470
 - bar accretion process, 452–455
 - cloud collisions vs. tidal disruption, 458–459
 - cluster tidal stripping, 465–468
 - dynamical friction of H I clumps, 460–464
 - dynamical friction process, 455–460
 - dynamical friction time, 460, 461, 465
 - nuclear star-formation ring/shock, 454, 461–465
 - star cluster inspiral, 462, 464–470
 - gas in the ISM, 439–442
 - accretion by a moving star, 440–442
 - accretion by a stationary star, 440–441
 - generic Bondi–Hoyle–Lyttleton accretion, 442
 - mass transfer in binary systems, 470–475
 - Roche lobe overflow, 472–474
 - dynamical, 473
 - secular, 474
 - thermal, 474
 - wind from companion, 474–475
 - stars in galactic centers, 442–451
 - mean free path for star–star encounters, 446
 - two-body relaxation, 445–446
- Accretion of matter, spherical wind/accretion theory
 - thermal, sub-Eddington (Bondi) accretion, 485–492

- accretion equation, polytropic, non-relativistic, 486–489
- accretion equation, polytropic, relativistic, 489–490
- general solutions, 486
- isothermal, 488
- regular solutions, 486–488
- singular/critical (sonic) surface, 486–488
- thermal, super-Eddington (Begelman) accretion, 490–492
- advection of photons, 490–491
- Bondi *vs.* Begelman accretion, 491–492
- photon mean free path, 492
- trapping radius, 490–491
- Accretion of matter, turbulent disk theory, 492–529
- instabilities, 504–508
 - secular (viscous), “inner” region, 507, 509
 - thermal, “inner” region, 507–509
 - thermal, SLE model, 515
- numerical models, 508–511, 519, 521, 525–529
- observational appearance, 494–496, 764–766, 771–784
- physical structure, 493–498
 - alpha-model, 496–498
 - effective temperature, 495
 - frictional heating, optically thick spectrum, 494–496
 - one-zone approximation, 494
 - stability, 507
 - vertical disk magnetic field, 498
- simulations, 549–558
- theoretical models, 492–529, 536, 546, 765, 768, 774, 776, 780, 781, 783
- thermal structure, general, 498–529
 - parameters, 500–501
 - Σ - $\dot{m}$ plane, 506, 526
 - stability, 507–508
 - thermal structure, geometrically thick
 - “slim” (super-Eddington) disks, 515–519, 780, 781, 783, 825
 - ADAF (very sub-Eddington) disks, 519–525, 765, 768, 825
 - SLE disks (unstable), 511–515
 - thermal structure, geometrically thin
 - SS disks, “inner” region, 503–504, 776, 780, 781, 783, 825
 - SS disks, “middle” region, 502–503, 505, 765, 768, 774, 776, 780, 781, 783, 825
 - SS disks, “outer” region, 499–502, 505, 765, 768, 774, 776, 825
- Active galactic nuclei, xix, 13–55, *see also* Black holes, massive; Black holes, supermassive
- BALs, 48–50, 61, 450, 451, 566, 567, 671, 755, 796–798, 806, 812, 826, 829
- HiBALs *vs.* LoBALs, 49, 51
- blazars, 30–37, 60, 679
 - BL Lacertae objects, 30, 32, 37, 679, 821–822
 - blue blazars/HBLs, 32–34, 37
 - IDVs, 33–35
 - radio selected, 32
 - red blazars/LBLs, 32–34, 37
 - X-ray selected, 32
- BSOs, 42–43
- components
 - BLR, 16, 22, 24, 791–796, 798, 806, 828
 - central engine, xix, 7, 16, 24, 25, 38, 53–55, 106, 478, 790–801, 805, 807–809, 811, 812, 819
 - dusty torus, 38–42, 46, 49, 53, 60, 478, 791
 - light cusp, 9, 57, 121, 445, 449, 821, 822

- NLR, 16, 22, 24, 39, 46, 791
- UFO, 776, 794–795, 797, 829
- warm absorber, 41–42
- evolution, 35–37, 52
- H II nuclei, 46, 52–53
- host galaxies, 14, 17, 32, 47–48, 59, 60, 276, 449, 789, 790, 796, 807
- “inactive” galactic nuclei, 55–58, 122, 425, 443, 807
- LINERs, 46, 52–53, 800
- LLAGN, 52–55, 93, 127, 425, 449, 450, 775, 789, 790, 800, 807
- radio cores, 53
- models, *see also* Black holes, massive (MBHs): accretion models
- central ionizing source, 791, 792, 795, 828
- broad line clouds, 791–794
- narrow line clouds, 38–40, 791
- ultra-fast outflows, 794–795, 797, 829
- radio properties of AGN, 790, 796–801, 805–806, 810, 815–823
- QSOs, *see* Active galactic nuclei: quasars: RQQs
- QSRs, *see* Active galactic nuclei: quasars: RLQs
- quasars, xix, 4–6, 10, 13, 23–37, 52, 55, 64, 65, 74, 91–94, 97, 101, 108, 115, 117, 121, 180, 264, 425, 429–430, 443, 447, 449–451, 469, 470, 483, 496, 523, 529, 561, 564–567, 747, 760, 789–801, 803, 810–812, *see also* Black holes, supermassive
- FR I quasars, 24
- FR II quasars, 21, 24, 29, 37, 50, 59, 796, 797
- FSRQs, 25–27, 29, 32–34, 36, 37, 119, 679
- RLQs/QSRs, 13, 23–37, 42, 45, 47–51, 59–61, 63, 425, 766, 774, 791, 796, 810, 815–823, 826, 829, 830
- RQQs/QSOs, 41–52, 58–61, 425, 754, 766, 774, 776, 790–801, 804, 816, 828
- radio galaxies, xix, 6, 13, 17–23, 40, 52, 53, 60, 74, 425, 447, 529, 701, 710, 818, *see also* Black holes, supermassive
- BLRGs, 22, 24–25
- FR I radio sources, 21, 22, 24, 36–37, 51, 679, 695, 708–725, 775, 818–823, 830
- FR I–II break, 22, 29, 816, 823
- FR I–II dichotomy, 819, 830
- FR II radio sources, 21, 22, 24, 29, 36–37, 50, 51, 59, 60, 679, 695, 701–712, 723, 725, 796, 797, 818, 822–823, 830
- HYMORs, 22
- N-galaxies, 22–24, 47, 53, 790
- NLRGs, 22
- WLRGs, 53
- Seyfert galaxies, 6, 13–17, 22, 38–44, 47, 49, 52, 53, 59–61, 85, 94, 113, 425, 443, 449–451, 454, 511, 529, 564, 689, 760, 789–801, 804, 806, 808, 828, 829, *see also* Black holes, massive
- dwarf Seyfert galaxies, 52–53
- NLSy1s, 42, 61, 450, 795–796, 798, 800, 801, 806, 826, 829
- Type 1, *see* Active galactic nuclei: Type 1: Seyferts
- Type 2, *see* Active galactic nuclei: Type 2: Seyferts
- transition-type, 52–53
- Type 1
- QSOs, 42–43, 47–52, 790–801, 828

- radio galaxies, *see* Active galactic nuclei: radio galaxies: BLRGs
- Seyferts, 14, 16, 22, 38–44, 51, 790–801, 828
- Type 2
 - HyLIRGs, 45, 46, 53, 425, 434, 785, 798, 811, 812
 - LIRGs, 44–46, 53, 458, 459
 - QSOs, 43–46, 49, 804
 - radio galaxies, *see* Active galactic nuclei: radio galaxies: NLRGs
 - Seyferts, 16, 22, 38–42, 44, 53, 791, 804
 - SMGs, 45
 - ULIRGs, 45, 46, 111, 425, 434, 458, 459, 785, 798, 811, 812, 824
- Type 3
 - AGN/QSOs, 46–47
 - unification of AGN, 23–35, 37–40, 42–43, 48–50, 58–61, 789–823, 828–830
- Binary black holes (BBHs), *see different mass classes, e.g.* Black holes, stellar-mass (SBHs): binary
- Binary stars
 - black hole, *see different mass classes, e.g.* Black holes, stellar-mass (SBHs): binary
 - close, types
 - contact, 472
 - detached, 472
 - semi-detached, 472
 - doubly-degenerate, 402, 407–408
 - effective binary potential, 471, 643, 644
 - evolution, *see* Stellar evolution: binary stars
 - gravitational waves from BH ring-down, 269, 270
 - kick velocities, 275–277
 - gravitational waves from inspiraling binaries, 263–270
 - chirp mass, 266
 - chirp signal, 265, 266
 - coalescence time, 265–266, 268
 - EBBHs, 267, 268
 - EMRIBHs, 267, 268
 - LIGO/LISA predictions, 266–268
 - gravitational waves from merging binaries, 268–277
 - numerical relativity, 268–277
 - orbital hangup, 273–275, 404
 - perturbation methods, 269
 - post-Newtonian methods, 269
 - mass function, 7, 142
 - neutron star, *see* Neutron stars(NSs): binary
 - orbits
 - circular, 140–142
 - general, 135–137, 142–143
 - Keplerian frequency, 142
 - Lagrangian points, 472
 - orbital phase angle, 140, 141
 - Trojan points, 472
 - X-ray (XRBs), xx, 4, 13, 42, 142, 334, 397, 472, 476, 580, 670, 696
 - jets, 827
 - masses, 7
 - theoretical models, 746–760
 - X-ray novae/SXTs, 77–78, 91–92, 519
- Black body
 - multi-colored, 496, 747, 789
 - Planckian distribution, 315, 318, 329, 496
 - radiation, 6, 45, 78, 495
 - radius, 6
- Black hole astrophysics summary
 - accretion power, *see also* Accretion of matter, ...
 - low $\dot{m}$, 746–747, 750–752
 - intermediate/hard-state $\dot{m}$, 746–747, 750–753

- intermediate-to-high $\dot{m}$, 747, 752–753
- high sub-Eddington $\dot{m}$, 747–748, 754
- moderate super-Eddington $\dot{m}$, “slim” disk accretion, 748–750, 754–755
- high super-Eddington $\dot{m}$, super-Eddington wind, 748–750, 754–755
- jet launching
 - by rotating accretion disks, 755–757
 - by rotating BHs, 755–757
- jet acceleration and collimation
 - MHD jet models, 758–760
 - thermal jet models, 757–758
- jet power
 - observed radio luminosity estimates, 751
 - total jet luminosity, 751
- jet suppression
 - by super-Eddington accretion, 671, 754
 - in the ADAF-to-MDAF transition, 753
 - in the geometrically thick-to-thin disk transition, 753
- Black hole binaries (BHBs), *see different mass classes, e.g. Black holes, stellar-mass (SBHs): binary*
- Black hole engine
 - combustion chamber, 291, 481–562
 - energy generation by gravitational accretion, 127, 437–439, 490–492, 522, 546, 571, 736, 746–750, 816
 - engine block, 381–435
 - exhaust systems, angular momentum
 - electromagnetic winds and jets, 291, 589–725
 - gravitational waves, 254–277
 - viscous transport, 291, 493–498, 585
 - exhaust systems, energy
 - loss down the hole, 291, 522, 640, 777
 - photon loss, 291, 498–529
 - thermal- and radiation-driven wind loss, 291, 562–589
 - fueling and carburetion, 291, 435–481
- Black holes
 - Galactic population, 99–100, 760, *see also Sources in the Milky Way: black holes*
 - horizon
 - definition, 3
 - formation, 270–289, 405
 - magnetic field, 609–641
 - Schwarzschild, 223–224, 229–233
 - Kerr, 236, 699, 756
 - Kerr–Newman, 245
 - of SBHs, 743
 - of IMBHs, 784
 - of MBHs, 790
 - of SMBHs, 810
 - magnetospheres, accreting, 621–641
 - black hole spindown, 615–616, 818
 - BZ process/effect, 638, 640, 742, 755–757
 - closed vs. open field solutions, 630–634
 - field line types, 624–625
 - GJ space charge density, role of, 625–630
 - magnetic Penrose process, 635, 639–640
 - negative energy at infinity, 639–640
 - sources of black hole magnetic field, 622–625
 - very weak intrinsic magnetic fields, 622–623

- when disk does not reach horizon, 625–635
 - when disk reaches into ergosphere, 635–640
 - wind, ingoing, 634–635, 698, 699
 - wind, outgoing, 634–635, 699
- magnetospheres, isolated, 609–621
 - BZ process/effect, 615–616, 619–620, 634
 - force free model, 614–616
 - frame dragging, 611, 621
 - GJ space charge density, role of, 619–620
 - GSS equation and singular surfaces, 611–614, 627, 853–858, 864
 - numerical models, 616–621
 - prograde *vs.* retrograde BHs, 612–613
 - weak intrinsic magnetic fields, 610–611
 - wind, ingoing, 620–621
 - wind, outgoing, 621
- Black holes, accretion states
 - hard state (with jet), 762–771
 - ADAF model, 764–766
 - fluctuation spectrum, 763, 766
 - photon spectrum, 763–766
 - ADAF steady jet radio power, 766
 - in SBHs, 764–766, 827
 - in IMBHs, 786–789, 828
 - in MBHs, 807
 - in SMBHs, 807
 - intermediate hard/outlier state, 766–771
 - MDAF model, 766–771
 - band-limited source fluctuation spectrum, 770–771
 - band-limited source photon spectrum, 769–770
 - MDAF steady jet radio power, 767–769
 - in SBHs, 766–771, 827
 - in IMBHs, 786–787, 828
 - in MBHs, 806
 - in SMBHs, 806
 - soft state, 771–775
 - SS disk “middle” region model, 773–775
 - fluctuation spectrum, 772–773
 - photon spectrum, 772–773
 - suppression of MHD disk jet in the soft state, 773–775
 - in SBHs, 771–775, 827
 - in IMBHs, 787–788, 828
 - not in MBHs or SMBHs, 789
 - unstable state, 775–777
 - SS disk “inner” region model, 775–777
 - “middle” region-like (soft) substate, 775–776
 - ADAF-like (hard) substate, 775–776
 - limit cycle behavior, 775–776
 - unsteady jet ejection, 775–777
 - in SBHs, 775–777, 827
 - in IMBHs, 788, 828
 - in MBHs, 790–795, 806–809, 828–829
 - in SMBHs, 807–809
 - SPL state, 777–781
 - “slim” disk model, 777–781
 - fluctuation spectrum, 777–780
 - photon spectrum, 777, 778
 - suppression of MHD disk jet in the SPL state, 779, 780
 - in SBHs, 777–781, 826, 827
 - in IMBHs, 788, 828
 - in MBHs, 795–798, 805–809, 829
 - in SMBHs, 812–813
 - super-soft state, 781–784
 - super-Eddington wind model, 781–784
 - in SBHs, 781–784, 826, 827
 - in IMBHs, 788–789, 828
 - in MBHs, 803–805, 807–809

- in SMBHs, 812–813
- Black holes, formation
 - SBHs, *see* Black holes, stellar-mass (SBHs): formation
 - IMBHs, *see* Black holes, intermediate mass (IMBHs): formation
 - MBHs, *see* Black holes, massive (MBHs): formation
 - SMBHs, *see* Black holes, supermassive (SMBHs): formation
- Black holes, intermediate mass (IMBHs),
 - 13, 57, 115–129, 163, 411, 427, 430–435, 464, 784–789, 823, 828
 - definition, 115–116, 784
 - accretion models, *see* Black hole astrophysics summary: accretion power
 - binary, 786–789
 - predicted hard state sources, 786–787
 - ULXs as soft-state super-micro-quasars, 787–788
 - predicted unstable state sources, 788
 - predicted SPL state sources, 788
 - predicted super-Eddington wind sources, 788–789
 - formation, 408–423, 784–786, 828
 - from many SBHs in star clusters, 421–422
 - in globular cluster core collapse?, 418–420
 - in Pop III VMSs and supermassive stars, 409–414, 785–786
 - in young Pop I/II star clusters and VMSs, 118, 414–423, 784–785, 828
 - horizon, 784
 - isolated
 - predicted hard state sources in ISM, 786, 828
 - jet production, *see* Black hole astrophysics summary: jet...
- Black holes, massive (MBHs), 13–16, 39–42, 52–55, 789–809, 828–829, *see also* Active galactic nuclei: Type 1, 2; Seyfert galaxies; LINERs; LLAGN
 - definition, 13, 790
 - accretion models, 802–809, 824, *see also* Active galactic nuclei: models
 - LLAGN as the hard state, 807, 824
 - outliers(?) as intermediate states, 806, 824
 - Type 1/2 AGN as the unstable state, 806, 824
 - NLSy1s and BAL AGN as the SPL state, 805–806, 824
 - hidden and unveiled AGN as the super-Eddington wind state, 803–805, 824
- binary, 432–435
 - formation, growth, and fueling, 801–809
 - in galaxy mergers (hidden and unveiled AGN), 802–806
 - with cluster inspiral, 807–809
 - with interstellar gas, 451–470
 - with stars, 442–451, 807
 - without galaxy mergers, 806–809
 - horizon, 790
 - jet production, *see* Black holes, supermassive (SMBHs): jet production
- Black holes, stellar-mass (SBHs), 87–114, 117, 127, 129, 588, 743–784, 791, 799, 805, 807, 815, 819, 821, 826–829
 - definition, 63, 743

- accretion models, *see* Black hole astrophysics summary: accretion power
- binary, 88–100, 397–408, 471, 761–784, 827
 - HMXBs, 90–92, 98–100, 402, 745, 762
 - LMXBs, 90–100, 402, 745, 762
 - microquasars, 98–100, 119, 121, 562, 761–784, 827–828
 - X-ray novae, 91–92, 95, 100
- formation, 743–746, *see also* γ -ray bursts (GRB)
 - direct collapse of stellar core, 393, 410, 729, 743
 - failed core-collapse supernovae, 108, 393, 394, 397, 410, 729, 743
 - in binary star systems, 109–110, 399, 401–405
 - in LGRBs, 108–109, 399
 - in massive stars, 106, 108–109, 394–399, 403, 408, 409, 729, 743, 744
 - in Pop I/II VMSs, 396
 - in SGRBs, 109–110, 405
 - in single stars, 394–397, 409
- horizon, 743
- isolated, 87–88, 471, 609–621, 827
 - detection, 760–761
 - OGLE/MACHO 1999, 11, 88
- jet production, *see* Black hole astrophysics summary: jet...
- Black holes, supermassive (SMBHs), 16–37, 42–53, 55–58, 809–823, 829–830, *see also* Active galactic nuclei: BALs; blazars; RLQs, RQQs, WLRGs, *etc.*
 - definition, 13, 810
 - accretion models, *see* Black holes, massive (MBHs): accretion models
 - binary, 432–435
 - formation, growth, and fueling, 423–435, *see also* Black holes, massive (MBHs): formation, growth, and fueling
 - horizon, 810
 - jet production, 815–823
 - FR I vs. FR II dichotomy, 20–22, 29, 700–725, 816, 822–823
 - recollimation features in radio-loud AGN, 694–700, 818–823
 - rotating SMBH power, 615–616, 633–634, 815–817
 - why are gE galaxies so radio-loud?, 817–818
- Blazars, *see* Active galactic nuclei: blazars
- Catalogs and surveys
 - optical sources
 - Hubble deep field, 48
 - Markarian catalog, 16
 - Messier catalog, 14, 64
 - NGC catalog, 14
 - SDSS, 48
 - radio sources
 - 3CR survey, 20, 23, 27
- Causality, EM
 - light cone, 181–182
 - singular surfaces, 611–614
 - causality limit surfaces, 612–614
 - light surfaces, 612
- Causality, HD, 360–363
 - characteristics, 362–363
 - critical (sonic) surface, 362–363
 - Mach cone, 361–362
 - sonic future, 361–362
- Causality, MHD, 360, 366–368, 679–685
 - characteristics, 366–369, 681
 - critical surfaces, 681
 - cusp surface (CS), 366, 367, 369, 681

- fast magnetosonic surface (FMS), 368, 369, 681
- slow magnetosonic surface (SMS), 368, 369, 681
- separatrix surfaces, 681–682
 - Alfvén surface (AS), 368, 369, 681, 682
 - fast magnetosonic separatrix surface (FMSS), 681, 682
 - slow magnetosonic separatrix surface (SMSS), 681, 682
- singular surfaces, 682
- Causality, RaHD, 573–578
 - critical surface, 574, 576–578
- Chandrasekhar mass/limit
 - definition, 161
 - in core-collapse SN, 63, 86, 162, 386, 389, 407
 - in Type Ia SN, 385, 406, 407, 731
- Charge dynamics, non-relativistic
 - generalized Ohm's law, 310
- Charge dynamics, relativistic
 - conservation of beamed current, *see* generalized Ohm's law
 - beamed spatial current, 307–310, 852
 - charge-current tensor, 307, 852
 - charge-weighted thermodynamic quantities, 308, 852–853
 - conservation of charge, 306–307, 852
 - current beaming factor, 308
 - current four-vector, 307, 851
 - equations of state, *see* Equations of state: charge dynamic
 - general relativistic Boltzmann equation, 296, 847
 - generalized Ohm's law, 307–312, 852
 - Hall effect and Hall MHD, 310
 - Lorentz effect and ideal MHD, 309
 - resistive losses and resistive MHD, 310–311
 - spatial current four-vector, 307, 308, 852
- Conservation laws
 - from the Boltzmann equation, 297–298, 308, 846–853
 - generated by gauge symmetries, *see* Gauge field theory: Bianchi identities/symmetries: conservation laws from
- Newtonian
 - conservation of mass, 147–149, 155, 352
 - conservation of energy and momentum, 149–153, 155, 352
 - conservation of charge, 192
 - conservation of current (Ohm's law), 310
- relativistic
 - conservation of rest mass, 299, 850
 - conservation of energy-momentum, 299–306, 850, 851
 - conservation of rest charge, 306–307, 851
 - conservation of beamed current (Ohm's law), 307–312, 851
- Core
 - galaxy, 9, 57, 430, 443, 449, 478–480
 - cusp, 9, 57, 121, 445, 449, 821
 - mass, 58, 123
 - radius, 9, 57, 58, 123, 417, 418, 421, 443, 445, 469
 - spinup, 478
 - molecular cloud, 456
 - star cluster, 121–123, 163, 416, 423–428
 - core collapse, 163, 415, 418–423, 428, 431, 789
 - stellar
 - C/O, 385–386, 395, 405–408, 410

- core collapse, 85, 388–397, 408, 411, 475, 480–481, 728, 826, 827
- iron/pre-supernova, 85, 389, 475, 480–481, 728, 743
- Doppler
 - boosting, *see* Relativistic motion/beaming
 - effect/shift, 23, 41, 189, 190
 - factor, 28, 189, 190
- Dynamics
 - charge dynamics, *see* Charge dynamics
 - electrodynamics (ED), *see* Electrodynamics
 - fluid dynamics, *see* Hydrodynamics and MHD
 - galaxy dynamics, *see* Globular cluster and galaxy structure & dynamics
 - globular cluster dynamics, *see* Globular cluster and galaxy structure & dynamics
 - hydrodynamics (HD), *see* Hydrodynamics and MHD
 - magnetohydrodynamics (MHD), *see* Hydrodynamics and MHD
 - radiation hydrodynamics (RaHD), *see* Hydrodynamics and MHD: RaHD
 - stellar dynamics, *see* Stellar structure & dynamics
 - thermodynamics, *see* Thermodynamics
- Eddington accretion rate
 - definition, 439, 501
 - in disk accretion, 515–519, 562, 567, 580–582, 669–671, 738, 740, 747–750, 754–755, 775, 779, 782
 - in spherical accretion, 490–492
- Eddington luminosity/limit
 - as a lower limit on BH mass, 5–6, 116–117
 - definition, 5
- Eddington ratio, 60, 439, 449, 450, 564, 791–792
- in disk accretion, 517, 561, 670, 744, 748, 754, 775
- in spherical accretion, 492
- in winds, 563–582, 782, 805
- in NSs, 740
- in SBHs, 521, 781–784, 827
- in IMBHs, 787–789, 828
- in MBHs, 791–792, 803–806, 808, 829
- in SMBHs, 430, 791–792, 803–806, 811, 829
- Electrodynamics
 - Minkowskian 3+1, 191–200
 - conservation laws and gauge freedom, 193–194
 - conservation of charge, 191–192
 - conservation of current (Ohm's law), 309–312
 - Lorentz force and Joule heating, 193
 - Maxwell's field equations, potential form, 193–194
 - Maxwell's field equations, vector form, 191
 - Minkowskian covariant
 - conservation laws and gauge freedom, 204
 - conservation of rest charge, 202
 - conservation of beamed current (Ohm's law), 307–310
 - Lorentz force and Joule heating, 203
 - Maxwell's field equations, potential form, 204
 - Maxwell's field equations, tensor form, 202
 - Maxwell's field equations, vector form, 205
 - ADM 3+1 (stationary metrics)
 - Lorentz force and Joule heating, 193, 341

- Maxwell's field equations, vector form, 248–250
- equations of state, *see* Equations of state: electrodynamic
- force-free/FFDE, *see* Hydrodynamics and MHD: FFDE
- Engine, black hole, *see* Black hole engine
- Equations of state
 - charge dynamic
 - charge-weighted internal energy and pressure, 308, 852–853
 - Hall coefficient, 309, 853
 - plasma frequency, 309, 853
 - resistivity/conductivity, 309, 853
 - electrodynamic
 - electric permittivity, 191, 250
 - magnetic permeability, 191, 250
 - in conservation laws, 312
 - radiative opacity and heat transport, *see also* Radiative processes (short wavelength)
 - bound–bound, 328, 564–565
 - bound–free, 326–328
 - electron scattering opacity, 326
 - free–free, 326–328
 - thermal conductivity, 322–323
 - importance of, 323
 - thermodynamic, internal energy/pressure
 - adiabatic index, 152, 153, 159, 161, 287, 315, 318, 583, 588, 705, 706, 714
 - degenerate electrons, 319
 - degenerate neutrons, 320
 - nonthermal gases, 320–322
 - perfect fluid/gas, 315
 - photons, 318
 - polytropic, 152–153, 156, 157, 161, 278, 287, 300, 315, 317, *see also* Polytropes
 - thermal gases, 314, 315, 317
 - thermodynamic, other
 - average molecular weight per electron, 161, 316
 - average molecular weight per gas particle, 315–316
 - chemical potential, 150, 313–314
 - Fermi temperature, 319
 - specific heats, 151–152, 315, 317, 322
 - viscosity
 - bulk, 301–302
 - importance of, 324
 - particle, 323–324
 - shear, 301–303
 - turbulent, 324–325, 496–498
- Fluid mechanics, *see* Hydrodynamics and MHD
- Frames of reference/coordinate systems
 - binary star
 - center-of-mass, 135–137, 140, 141, 143, 264, 471, 472
 - inertial, 136, 137, 143, 471
 - star-centered (non-inertial), 136
 - Cartesian, 139, 141, 146–148, 167–170, 172, 174, 175, 187, 212, 223, 342, 353
 - circumferential, 213, 222
 - curvilinear, 187, 212, 214, 217, 226, 251, 342, 626, 834, 837, 839
 - cylindrical, 176, 351, 371, 493, 505, 520, 532, 536, 658, 663, 689, 860, 861, 864
 - Galactic, 68
 - mass coordinates (“MOV”), 153–154, 282, 283, 841–842, 845
 - non-relativistic
 - Eulerian (“FIX”), 147, 154–156
 - Lagrangian (“MOV”), 146–154, 844–846
 - phase space, 295–297, 415, 847, 848

- polar, 168–170, 172, 175, 176, 216, 834
- relativistic
 - Boyer–Lindquist (BL), 209, 234–242, 244–246, 248, 250, 305, 338, 339, 618, 639, 672, 693, 855, 858
 - drifting/shifting, 233, 234, 242–243, 245, 247, 251, 841
 - harmonic, 272
 - horizon-penetrating (HP), 234, 242–243, 245, 339, 639
 - locally-Lorentz fixed/FIDO/ZAMO (FIX), 224, 225, 237, 238, 305
 - moving-body (MOV), 224, 225, 238, 305
 - observer-at-infinity/synchronous (OIS), 237–239, 305
 - Schwarzschild–Hilbert (SH), 222, 224–226, 228, 229, 231, 232, 235, 271, 280, 283, 626, 845
- spherical-polar, 146, 148, 176, 212, 222, 223, 233, 234, 248, 658, 684, 692
- Free fall
 - collapse parameter, 138, 282, 283, 285
 - free-fall time, 456, 570, 650, 653, 748
 - free-fall/escape speed, 229, 277, 392, 422, 538, 543, 558, 649, 650, 673, 677, 814
 - in Newtonian gravity, *see* Gravity: Newton’s law/theory of: free fall
 - in stationary Schwarzschild metric, *see* Geometry/metrics: Schwarzschild: free fall
 - in evolving Schwarzschild metric, 281–286
 - in Kerr metric, *see* Geometry/metrics: Kerr: free fall
- Galaxies
 - AGN hosts, *see* Active galactic nuclei: host galaxies
 - ellipticals and spheroidals, 9, 15, 22, 25, 48, 53, 56, 57, 59, 60, 63, 85, 109, 110, 116, 120–121, 123, 416, 424–425, 428, 432, 443–444, 451, 452, 454, 458, 480, 786, 790, 801, 807, 809–823, 830
 - irregulars, 15, 85, 109, 114, 811
 - radio, *see* Active galactic nuclei: radio galaxies
 - Seyfert, *see* Active galactic nuclei: Seyfert galaxies
 - spirals, 15, 23, 25, 48, 56, 59, 66, 85, 109, 114, 116, 117, 122, 123, 451–470, 478, 789–809
- Galaxy components
 - globular clusters, 7, 115, 120–123, 128, 162–163, 276, 403, 416–420, 422, 442, 464, 477, 785
 - H I clumps, 412–413, 454–455, 460–466, 470, 808
 - ISM, 22, 34, 39, 382–394, 410–414, 439–442, 451–471, 701, 702, 705, 710, 728, 731–732, 760–761, 773, 786, 803–806, 808, 811–813, 829
 - molecular clouds, 54, 68, 111, 114, 413, 414, 451–460, 463–466, 477–478, 728, 756, 807, 814
 - star clusters, 118, 128–129, 414–423, 430–435, 443–444, 450, 464–470, 784–785, 787, 807, 808, 813, 828
 - stars, 415–423, 442–451
- γ -ray bursts (GRBs), 100–114
 - beaming break, 103, 105
 - long-duration γ -ray bursts (LGRBs), 102–109
- mechanisms
 - collapsar, 106, 108–109

- neutron star merger, 109–110, 404–405
 - relativistic fireball, 102–104
 - short-duration γ -ray bursts (SGRBs), 102, 109–110
 - stages, 105–107
 - supernova connection, 107–108
- Gauge field theory
 - Bianchi identities/symmetries
 - algebraic symmetries, 214–215
 - conservation laws from, 191–192, 202, 205, 218, 261
 - differential symmetries, 202, 205, 210, 215–219, 261
 - need for gauge constraints, 191–192, 202, 218
 - constrained transport, 192, 202
 - gauge conditions, 191, 194–196, 199, 205, 206, 255, 257, 271, 272
 - gauge freedom, 193–200, 204, 207, 220
 - gauge invariance, 194, 202
 - gauge transformation, 194, 195, 220
 - general description
 - conservation laws, 192, 203–206, 218, 298
 - equations of state, 312
 - field equations, 191, 192, 203–206, 217, 220, 221
 - gauge constraints, 191, 192, 203–206, 220
 - transverse (Lorenz) gauges, 194–196, 199, 205, 206, 256, 257
- Geometry/metrics, general
 - 1-forms/covariant vectors, 170–172, 176, 183–187, 201, 203–206, 225, 833, 834, 838
 - 2-forms, 173, 201, 203, 204, 249, 305, 835, 838
 - 2-tensors, 173, 210, 217–219, 838
 - 4-tensors, 165, 212–218, 835
 - basis 1-forms, 171
 - basis vectors, 171
 - contravariant tensors, 173, 201, 219, 254–256, 834, 835
 - coordinate transformations
 - boosts, 185, 201, 238
 - coordinate change, 168, 170–171, 174–175, 225, 233, 242
 - general, 171–173, 184–185, 201
 - global absolute, 167, 174, 175
 - global differential, 168, 170–171, 174, 175, 233, 242
 - local differential, 174–175, 185, 201, 225, 238, 259
 - rotation, 167, 168, 259
 - translation, 167, 168
 - covariant tensors, 173, 256, 834, 835
 - curvature, 209–221, 247, 621, 837, 840, 842
 - tensors and 2-forms, 172–173, 187–188, 200–204
 - vectors, 171
- Geometry/metrics, non-stationary
 - perturbed Minkowskian, 219, 254–268
 - metric/line element, 254
 - evolving Schwarzschild, 277–289, 840–846
 - gravitational collapse, 277–289
 - metric/line element, 841
- ADM 3+1 spacetime split, 247–250, 268–277
 - drift vector, 247, 251
 - lapse function, 247, 251
 - line element, 247
 - metric, 247
 - shift vector, 247
 - three-metric, 247–249
- Geometry/metrics, stationary
 - Pythagorean, 166–175, *see also* Vector and tensor physics:
 - two-vectors
 - coordinate transformations, 167–175
 - coordinates, Cartesian, 167–170, 172, 174, 175

- coordinates, polar, 168–170, 172, 175
- distance, 166–168, 172
- metric/line element in Cartesian coordinates, 166, 169
- metric/line element in polar coordinates, 169, 170
- plane geometry, 166–167
- tensors and 2-forms, 172–173
- vectors and 1-forms, 170–172
- Euclidean, 175–176, *see also* Vector and tensor physics: three-vectors
- coordinate transformations, 176
- coordinates, Cartesian, 175
- coordinates, cylindrical, 176
- coordinates, spherical-polar, 176
- distance, 175
- metric/line element in Cartesian coordinates, 175
- metric/line element in cylindrical coordinates, 176
- metric/line element in spherical-polar coordinates, 176
- tensors and 2-forms, 176
- vectors and 1-forms, 176
- Minkowskian, 176–183, *see also* Vector and tensor physics: four-vectors
 - coordinates, Cartesian, 187
 - events, 178, 182–183
 - FitzGerald contraction, 179–180
 - geodesics, 182–183
 - light cone, 181–184
 - Lorentz transformations (general), 184–185
 - metric/line element, 178, 183–184
 - null surface, 178, 181, 182
 - proper distance, 178–183
 - proper time, 182–183
 - rest mass, 185, 188
 - tensors and 2-forms, 187–188, 200–204
 - time dilation, 179–180
 - vectors and 1-forms, 183–187, 201, 203–206, 225
- Schwarzschild, 221–234
 - coordinates, horizon-penetrating (HP), 234
 - coordinates, locally-Lorentz fixed (FIX), 224, 225
 - coordinates, locally-Lorentz moving (MOV), 224, 225
 - coordinates, mass, 282, 283, 841–842, 845
 - coordinates, Schwarzschild–Hilbert (SH), 222, 224–226, 228, 229, 231, 232, 235, 271, 280, 283, 626, 845
 - free fall, 229
 - laws of motion, 226–228
 - metric/line element in ingoing HP coordinates, 233
 - metric/line element in SH coordinates, 222
 - orbits, ISCO, 231
 - orbits, marginally bound, 232
 - orbits, particle, 230–232
 - orbits, photon, 229–230
 - Schwarzschild radius, 222–224
 - time warp, 227–228
- Kerr, 234–243
 - dimensionless angular momentum parameter j , 235
 - angular velocity of space, 235, 237
 - coordinates, Boyer–Lindquist (BL), 209, 234, 242, 246, 248, 250, 305, 338, 339, 618, 639, 672, 693, 855, 858
 - coordinates, horizon-penetrating (HP), 242–243, 339, 639
 - coordinates, locally-Lorentz fixed/FIDO/ZAMO (FIX), 237, 238, 305
 - coordinates, observer-at-infinity/synchronous (OIS), 237, 239, 305

- equatorial ergosphere radius, 236, 242, 243
- equatorial ISCO radius, 241–243
- free fall, 240
- gravitational radius, 235
- horizon radius, 236, 242, 243
- in 3+1 notation, 248, 250–251
- laws of motion, 239
- metric/line element in BL coordinates, 234
- metric/line element in ingoing HP coordinates, 243
- negative-energy trajectories, 239
- orbits, equatorial marginally bound, 241–243
- orbits, equatorial particle, 240–241
- orbits, equatorial photon, 241–243
- Penrose process, 240
- retrograde/prograde BH spin, 235, 241–243
- Kerr–Newman, 243–245
 - dimensionless charge parameter q , 244
 - coordinates, Boyer–Lindquist (BL), 244, 245
 - coordinates, horizon-penetrating (HP), 245
 - metric/line element in BL coordinates, 244
 - metric/line element in ingoing HP coordinates, 245
- ADM 3+1 spacetime split
 - electrodynamics, 248–250
 - orthonormal vectors in three-space, 250–251
- Globular cluster structure & dynamics, *see* Star cluster and spheroidal galaxy structure & dynamics
- Gravity
 - Newton’s law/theory of, 7, 118, 133, 135–146, 209–212, 219–220, 229, 269, 282, 287, 294, 349, 351–353, 471–472, 486, 570, 642–643, 859, 863
 - circular orbits, 139–142
 - free fall, 137–139
 - general potential theory, 144–146
 - gravitational collapse, 138, 162
 - Green’s function solution, 145
 - Poisson’s equation, 145–146, 155, 211
 - Roche potential, 471, 643, 644
- pseudo-Newtonian, 510, 862
- Einstein’s theory of, *see also* Geometry/metrics
 - connection to Newton’s theory, 209–212, 219–220
 - Einstein tensor, 217–221
 - field equations, 210–212, 217–221, 254–256, 270, 272, 276, 278, 335, 404, 841–845, 851
 - general relativity, 84, 207–289, 336–340, 346–348, 405, 846–858
- Schwarzschild, *see* Geometry/metrics, stationary: Schwarzschild
- evolving Schwarzschild, *see* Geometry/metrics, non-stationary: evolving Schwarzschild
- Kerr, *see* Geometry/metrics, stationary: Kerr
- Kerr–Newman, *see* Geometry/metrics, stationary: Kerr–Newman
- ADM 3+1 spacetime split, *see* Geometry/metrics, non-stationary: ADM 3+1 spacetime split
- Green’s functions, 145, 197, 260
- Horizon
 - event
 - Schwarzschild, 223–224, 229–233
 - Kerr, 236, 699, 756
 - Kerr–Newman, 245
 - general, 270–289, 405

- magnetosonic (FMSS), 539, 679–692, 694, 696–700, 724, 758–760, 821
- Hydrodynamics and MHD
 - Electrodynamics, 305–306
 - Eulerian frame, 154–156
 - FFDE, 340–346
 - alternate form, 346–347
 - cross-field equation, 345, 858
 - current equation, 342, 858
 - drift current, 342
 - force-free condition, 341, 346, 599
 - GSS equation, 344–346
 - pulsar equation, 346
 - standard form, 341–342
 - stationary and axisymmetric, 344–346, 853–858
 - general
 - conserved variables, 337–340, 349
 - evolution equations, conservative form, 155, 337, 339, 348
 - gas particle heat flux and viscosity, 300–303
 - gas pressure and internal energy, 300, 303, 314, 317
 - primitive variables, 297, 337, 340
 - pseudo-forces, centrifugal, 187, 211, 227, 230, 240, 339, 471, 557, 837
 - pseudo-forces, Coriolis, 187, 211, 339, 372, 837
 - pseudo-forces, frame dragging, 339
 - pseudo-forces, gravity, 209, 211, 212, 220, 227
 - stress-energy-momentum tensor, 303
 - GRHD, 335, 347–348
 - GRMHD, 846–853
 - Hall MHD, 310
 - ideal MHD, 311–312
 - metric shear, 338, 339, 654
 - multi-fluid, 293, 847–850
 - one-fluid, 293, 850–853
 - resistive MHD, 310–311
 - Lagrangian frame, 146–154, 844–846
 - NRHD, 143–163
 - accretion disk structure, 355, 493
 - Couette flow, 370–372
 - stellar structure & dynamics, 153, 156, 355
 - vortex flow, 372
 - NRMHD, 335, 351–355
 - accretion flow structure, 536
 - MRI analysis, 372
 - RaHD, 303–305, 569–570, 574–578
 - general radiation stress-energy-momentum tensor, 305
 - radiation heat flux and viscosity, 304–305
 - radiation pressure and internal energy, 303, 304, 318
 - SRMHD, general, 348–349
 - SRMHD, stationary and axisymmetric, 349–351, 858–864
 - cross-field equation, 351, 864
 - field-parallel equation, 349–351, 859–864
 - frozen-in condition, 350, 859–862
 - mass-loading parameter k , 350, 864
 - relation to stationary and axisymmetric FFDE, 349, 859, 861
- Inactive galactic nuclei, *see* Active galactic nuclei: “inactive” galactic nuclei
- Initial mass function (IMF), 99, 411, 745
- Instabilities
 - accretion disk

- secular/Lightman–Eardley/viscous, 506–509, 514, 526
- thermal, of SLE hot disk model, 506, 514–515, 526
- thermal, of SS “inner” region, 507–509
- HD
 - Kelvin–Helmholtz (KH), 705–708, 717, 721, 723
 - Rayleigh criterion, 371–372, 376
 - rotational shearing, 370–372
- in black hole feeding
 - cluster inspiral instability, 808–809
- in supernovae
 - pair instability, 409–411, 785
 - photodisintegration instability, 411
 - standing accretion shock instability (SASI), 391
- MHD, helical-kink/current-driven (CD), 717–721, 723, 725, 823
- MHD, magneto-rotational shearing/MRI, 372–377
 - strong-field, 375–377
 - suppression, 375–377, 535
 - weak-field, 373–375
- MHD, Parker, 497, 542
- Irreducible mass, 235–237, 476
 - accretion of, *see* Accretion of matter, ...
- Jet propagation, HD jets and FR II sources, 701–711, 823
 - one-D, supersonic jets, 702–704
 - contact discontinuity, 703
 - forward shock, 703
 - reverse shock, 703
 - two-D, supersonic jets, 704–706
 - ambient medium, 704, 705
 - bow shock, 704, 705
 - cocoon, 704, 705
 - relativistic vs. non-relativistic, 705–706
 - working surface/Mach disk, 704, 705
- three-D, supersonic jets, 705–709
 - KH instability, 706–708
 - relativistic vs. non-relativistic, 708–709
- birth of the jet concept, 701–702
- models for FR II jet formation, 725, 823
 - magnetic field dissipation, 725, 823
 - recollimation shock, 724–725, 821–823, 830
- Jet propagation, MHD jets and FR I sources, 711–723, 818–823
 - one-D simulations, super-slow-magnetosonic jets, 712–713
 - contact discontinuity, 712, 713
 - forward fast-mode (bow) shock, 712, 713
 - forward slow-mode shock, 712, 713
 - reverse slow-mode shock, 712, 713
 - two-D simulations, super-fast-magnetosonic jets, 713–717
 - magnetic chamber, 715, 716
 - nose cone, 714–716
 - relativistic vs. non-relativistic, 716–717
 - reverse fast-mode shock, 715, 716
 - three-D simulations, super-slow-magnetosonic jets, 717–722
 - CD instability/helical kink, 717–721
 - KH instability, 721–722
 - relativistic vs. non-relativistic, 721
 - models for FR I jet formation, 708–711, 724–725, 822–823
 - HD (disrupted FR II flows), 708–711, 822–823

- MHD (failed FR II formation), 724–725, 818–823
- Jets
 - acceleration and collimation, *see* Wind and jet acceleration theory
 - current-carrying, *see* Jet propagation, MHD jets
 - kinetic-flux-dominated (KFD), *see* Jet propagation, HD jets
 - launching, *see* Wind and jet launching...
 - Poynting-flux-dominated (PFD), *see* Jet propagation, MHD jets
 - propagation, *see* Jet propagation...
- King star cluster model, *see* Star cluster and spheroidal galaxy structure & dynamics: King model core, *see also* Core: galaxy
- Lane–Emden equation, 157–163
- Lensing, gravitational
 - MACHO project, 11, 87
 - microlensing, 10–11, 87, 760
 - OGLE project, 11, 87
- Lorenz (transverse) gauge/condition, 194–196, 199, 205, 206, 255, 257
- Macroquasars, *see* Active galactic nuclei...
- Magnetars, 70–71, *see also* Neutron stars (NSs): isolated: AXPs or SGRs
- Magnetospheres
 - black hole, *see* Black holes: magnetospheres...
 - general
 - current sheet, 309, 342, 595, 600, 615, 617, 618, 627, 629
 - field angular velocity, 593, 611, 614, 619, 620, 648, 662, 666, 674, 694, 860–861, 864
 - magnetosphere condition, 342–344, 621
 - pulsar, *see* Pulsars: magnetospheres...
 - stationary electrodynamics, *see* Hydrodynamics and MHD: FFDE: stationary and axisymmetric
- Mass function
 - binary, *see* Binary stars: mass function
 - stellar initial, *see* Initial mass function
- Microquasars, 13, 61–114, 762
 - black hole, *see* Black holes, stellar-mass (SBHs): binary: microquasars
 - classical, 92–98
 - formation, 403
 - galactic orbits, 760
 - neutron star, *see* Neutron stars (NSs): binary: microquasars
- Miniquasars, 115–129, *see also* Black holes, intermediate mass
- Neutron stars (NSs), 65–87, 727–743, 826
 - binary, 72–84, 397–408, 566, 735–743, 826, *see also* Binary stars: X-ray (XRBs)
 - atoll sources, 73–75, 79, 738–743, 826
 - HMXBs, 74–76, 402, 735
 - LMXBs, 74–76, 402, 735–739
 - microquasars, 73–75, 81, 737–743, 826
 - recycled (millisecond) binary pulsars, 83, 403
 - X-ray novae, 77–78, 100
 - X-ray pulsars, 75, 78, 603–609, 736–737
 - Z sources, 73–75, 79, 736–743, 824, 826
 - formation, 84–87, 727–731, *see also* Supernovae

- in binary star systems, 399, 401–405, 731
 - in core-collapse supernovae, 386–394
 - in intermediate-mass stars, 386–388
 - in massive stars, 84–87, 388–394, 398, 403, 409, 729, 744, 826
 - in Pop III stars, 409, 410
 - in single stars, 386–394, 396, 409
 - neutron star window, 396–397, 744
- Galactic population, 68, 72, 99, 100, *see also* Sources in the Milky Way: neutron stars
- isolated, 65–72, 731–735, 826, *see also* Pulsars: radio AXP, 67, 70
- detection, 732, 734
- radio pulsars, *see* Pulsars: radio
- RRATs, 67, 71–72
- SGRs, 67, 70, 734
- XDINs, 71–72
- jets in binary systems
 - MHD disk wind-driven jets, 739–742
 - pulsar propeller-driven jets, 735, 739–742, 826
 - radio power at low accretion rate, 742–743
 - suppression of MHD disk jets, 739–742
- jets in isolated systems, *see* Pulsars: radio
- magnetospheres, *see* Pulsars: magnetospheres...
- Nobel Prize, 273
 - 1935, discovery of the neutron, 68
 - 1970, magnetohydrodynamics and plasma physics, 364
 - 1974, radio astronomy imaging techniques and discovery of pulsars, 17, 64
 - 1983, theoretical work on stars and nucleosynthesis, 63
 - 1993, discovery of the binary pulsar and applications to gravitation, 84
 - 2002, discovery of cosmic neutrinos and X-ray sources, 64
- Orbits
 - in Newtonian gravity
 - binary, *see* Binary stars: orbits
 - circular, *see* Gravity: Newton's law/theory of: circular orbits
 - in Schwarzschild metric, *see* Geometry/metrics: Schwarzschild: orbits
 - in Kerr metric, *see* Geometry/metrics: Kerr: orbits
- Particle distribution functions
 - gases, nonthermal, 320–322
 - gases, thermal, 313–320
 - general, 313–315
 - Fermi–Dirac, general, 318–320
 - Fermi–Dirac, non-relativistic, 319
 - Fermi–Dirac, relativistic, 319
 - Maxwellian, general, 316–317
 - Maxwellian, non-relativistic, 314–316
 - Maxwellian, relativistic, 317
 - Planckian, 317–318
- Poisson's equation, 145–146, 155, 211
- Polytropes
 - analytic solution ($n = 1$), 159–160, 287
 - general, 152–153, 157–159
 - isentropic perfect gas ($n = 3/2$), 152, 159–161, 287, 315
 - relativistic gas ($n = 3$), 152, 159–162, 287, 304

- singular isothermal sphere (SIS; $n = \infty$), 54, 159, 162–163, 413, 415–418, 430, 443–444, 449, 459, 464–466, 468, 469
- Pulsars
 - magnetospheres, accreting, 603–609
 - accretion regime, 605–607
 - co-rotation radius, 605, 608
 - magnetic radius, 604, 608, 741
 - propeller spindown time, 609
 - propeller/wind turbine regime, 605, 607–609, 735, 739–742, 826
 - unstable magnetic field at high $\dot{m}$, 606–607
 - magnetospheres, isolated, 592–603
 - aligned rotator FFDE models, 596–598
 - curvature radiation, 602
 - force-free model, 594–603
 - Goldreich–Julian charge density, 594
 - light cylinder radius R_L , 346, 593, 596, 674, 696, 732
 - oblique rotator FFDE simulations, 598–600
 - particle creation, 601–602
 - pulsar emission mechanism, 600–603
 - pulsar equation and singular surface, 346, 597
 - pulsar spindown time, 66
 - vacuum breakdown and spark gaps, 601–603
 - wind, outgoing, 595–596
 - radio, 65–70
 - binary, 84
 - dead, 68, 72, 603, 731–732, 734, 742, 826
 - interstellar dispersion, 66
 - proper motion/kicks, 68–70, 394, 730, 733
 - PWN, 69, 732
 - rotating magnetic dipole radiation, 66–67, 592–594, 601, 734
 - X-ray, *see* Neutron stars (NSs): binary: X-ray pulsars
 - X-ray, anomalous, *see* Neutron stars (NSs): isolated: AXPs
- Quasars
 - compact, *see* Active galactic nuclei: quasars: FSRQs
 - extended, *see* Active galactic nuclei: quasars: RLQs
 - hidden, *see* Active galactic nuclei: Type 2, 3
 - QSOs, *see* Active galactic nuclei: quasars: RQQs/QSOs
 - QSRs, *see* Active galactic nuclei: quasars: RLQs/QSRs
- Radiation hydrodynamics (RaHD), *see* Hydrodynamics and MHD: RaHD
- Radiative processes (long wavelength), *see* Waves and radiation (long wavelength)
- Radiative processes (short wavelength)
 - bound–bound
 - bound–bound absorption, 328, 564–565, 794
 - force multiplier, 564–566, 574, 575, 744
 - iron $K\alpha$, 39–41
 - line emission, 14, 16, 22, 23, 43, 44, 52, 82
 - P Cygni lines, 49
 - bound–free
 - bound–free absorption, 326–328, 331, 499, 504, 521
 - recombination emission, 327
 - electron scattering
 - Compton cooling, 33, 326, 332–334, 513, 671, 755
 - Compton heating, 79, 326, 566–567

- Compton parameter y , 333, 334, 513, 795
- Compton reflection, 39–41, 588, 817
- Compton thick media, 334, 789, 792, 804
- Compton thin media, 334, 514
- Klein–Nishina cross-section, 326
- opacity, 5, 326, 501, 564
- Thomson cross-section, 326, 865
- free–free
 - Bremsstrahlung emission, 329–330, 332, 333, 504, 765
 - free–free absorption, 326–328, 331, 499, 504, 521
- synchrotron
 - nonthermal, 19, 588
 - thermal, 330–332, 521, 559, 755
- Reducible mass, 235–237, 476, 639
 - accretion of, *see* Accretion of angular momentum
- Relativistic motion/beaming, 11, 26–36, 47, 105, 112–114, 119, 188–190, 227, 308
- Relativity
 - general, 207–251
 - numerical, 268–277
 - 3+1 ADM method, 271–272
 - BSSN modifications, 272, 273
 - gauge conditions (slicing), 271, 272
 - harmonic coordinates, 272
 - history, 271–273
 - initial data, 272, 275
 - singularity excision/punctures, 271, 273
 - special, 163–207
- Sources in active galactic nuclei
 - 1823+568, 29, 31
 - 1H 0707-495, 817
 - 3C 9, 23
 - 3C 47, 21, 24, 29
 - 3C 48, 23
 - 3C 120, 94, 510, 511, 562, 791, 794, 799, 817
 - 3C 273, 23, 25–27, 31, 33, 794
 - 3C 279, 33
 - 3C 345, 25
 - 3C 371, 22
 - 3C 390.3, 21, 22, 817
 - Andromeda (M31), 121, 787
 - BL Lac, 30, 37, 821–822
 - Centaurus A (NGC 5128), 17, 25, 787, 810
 - Cygnus A, 17–19, 21, 25, 701, 704, 705
 - Hercules A, 17
 - M51, 478
 - M81, 52, 588
 - M84 (3C 272.1, NGC 4374), 21, 52
 - M87/Virgo A, 4, 9, 14, 15, 17, 19–21, 30, 53, 57, 445, 751, 810, 818–823
 - M104, 52
 - MCG 6-30-15, 41, 817
 - Milky Way (Galaxy/Sgr A*), *see* Sources in the Milky Way: black holes: Sgr A*
 - Mkn 421, 33
 - Mkn 501, 33
 - NGC 1068, 14, 16, 38, 39
 - NGC 2787, 809
 - NGC 4051, 799–801
 - NGC 4151, 14, 16
 - NGC 4258, 52, 478
 - NGC 4261, 39, 40, 478
 - NGC 4472, 120, 783
 - NGC 5728, 39, 40
 - NGC 6251, 17, 18, 20, 31
 - Q1413+113, 50
- Sources in globular clusters and dwarf ellipticals
 - G1 (Mayall II), 121–123
 - M15, 122, 123
 - Omega Centauri (NGC 5139), 122, 123

- RZ 2109 in NGC 4472, 120, 783
- Sources in other galaxies
 - M82 ULX-1, 117, 118
 - N1 in M81, 781
 - P098 in M101, 569, 781–784, 789, 794, 826
- Sources in the Milky Way, 63–115, 121–128, 268, 746, 760, 786, 828
- black holes
 - 1E 1740.7-2942, 91–93
 - A0620-00, 92
 - Cygnus X-1 (HDE 226868), 64, 88, 89, 91, 92, 99, 394, 399, 475, 512, 515, 521, 530, 747, 769, 781, 782, 794
 - GRO J1655-40, 94–95, 670, 751, 756, 763, 764, 768, 770
 - GRS 1915+105, 92–95, 97, 121, 394, 510–512, 553, 560, 562, 756, 762–764, 768–770, 776, 777, 787, 788, 801, 829
 - GS 1124-684, 92
 - GS 2000+251, 92
 - GX 339-4, 91, 95, 553, 740, 756, 764
 - IRS 13E, 125–129
 - J0422+32, 92
 - LMC X-1, 91
 - LMC X-3, 91
 - Sgr A*, 4, 8, 52–55, 530, 588, 695, 790, 825
 - XTE J1550-564, 95, 756, 764, 801
 - XTE J1859+226, 95
- neutron stars
 - Aquila X-1, 77, 740
 - Centaurus X-3, 73, 76, 88, 399
 - Circinus X-1, 77, 566, 567
 - Crab Pulsar, 65–70, 73, 84, 108, 594, 730, 733
 - Geminga, 72, 731
 - Hercules X-1, 73, 75, 736
 - LSI +61 303, 77
 - millisecond pulsar, 83
 - rapid burster, 77, 111
 - Scorpius X-1, 64, 65, 73, 75
 - Vela Pulsar, 69, 730, 733
 - other, *see also* Supernovae: remnants
 - IRS 16, 124–126
 - SS433, 81–83, 99, 400, 671, 755, 781, 798, 826, 827
- Spheroidal galaxy structure & dynamics, *see* Star cluster and spheroidal galaxy structure & dynamics
- Star cluster and spheroidal galaxy structure & dynamics, *see also*
 - Core: galaxy
 - analytic, singular isothermal sphere/ $n = \infty$ polytrope model, 54, 415–418, 430, 443–444, 449, 459, 464–466, 468, 469
 - core density, 417, 443, 444, 449
 - core galaxy, 57, 58, 417–418, 443, 444, 449, 450
 - core mass, 58, 417, 444
 - cusp/power-law galaxy, 57, 416–417, 444, 445, 450
 - mass deficit, 418, 444
 - core collapse, equal-mass stars, 418–420
 - core collapse, mass-segregated, 420–421
 - Fokker–Planck formalism, 415, 418
 - King model, 416–419, 444
 - core (King) radius, 417, 418, 421, 443, 445, 469
 - tidal radius, 162, 417
 - loss cone
 - depletion, 434, 448
 - refilling, 434, 448
 - spheroidal cluster as a single star
 - binary burning, 419–420
 - binary hardening, 421, 427, 432–434

- ensemble of stars as a gas, 415–423
- gravo-thermal catastrophe, 418–420
- heat conduction, 419–420
- isothermal velocity distribution, 417
- temperature/velocity dispersion, 415
- Stars
 - binary, *see* Binary stars
 - black holes, *see* Black holes...
 - evolution, *see* Stellar evolution...
 - interiors, *see* Stellar structure & dynamics
 - lifetimes, 384
 - luminosities, 384
 - mass classes
 - brown dwarfs, 87, 383, 790
 - low-mass, 75, 77, 383–384, 396
 - solar-type, 24, 384–385, 396, 402, 403, 406, 447–450, 729, 807, 810
 - moderate-mass, 385–386, 396, 406
 - intermediate-mass, 386–390, 403, 406, 408
 - massive, 68, 76, 84–87, 106–109, 388–398, 403, 406, 408, 409, 420, 422–423, 434, 474, 481, 729, 743, 744, 803, 826, 828
 - very massive (VMSs), 118, 409–414, 422–423, 784
 - supermassive, 6, 411, 414, 785, 789
 - neutron, *see* Neutron stars (NSs)
 - Population I, 118, 412, 414–423, 784–785, 828
 - Population II, 118, 412, 414–423, 784–786, 803, 828
 - Population III, 116, 408–414, 423, 425–428, 785–786
 - red giants, 112, 114, 384, 386, 395, 398, 400, 419, 420, 447, 449, 451, 729, 790, 807
 - red supergiants, 54, 85, 386, 395, 397, 398, 400, 403, 447, 449–451, 729, 783, 790, 810, 812, 828
 - structure, *see* Stellar structure & dynamics
 - variable, 30
 - white dwarfs
 - C/O, 385–386, 407, 408, 731
 - doubly-degenerate binaries, 407–408
 - ONeMg, 407, 408, 731
- Stellar evolution, binary stars
 - γ -ray bursts mechanisms
 - neutron star merger, 109–110, 404–405
 - close binaries, 397–408
 - cataclysmic variable stars, 402, 405–407, 731
 - common envelope stage/phase, 399–403, 408
 - formation of HMXBs, 398–402
 - formation of LMXBs, 401, 403
 - mass transfer Cases A–D, 398, 403–404
 - merger rates of compact object binaries, 402, 403
 - spinup of millisecond pulsars, 83, 403, 477
 - super-Eddington wind phase(Cyg X-1 on steroids), 399, 403, 783
 - supernova explosion mechanisms
 - carbon deflagration/detonation, 405–407, 731
- Stellar evolution, single stars, 381–394
 - γ -ray bursts mechanisms
 - collapsar, 106, 108–109
 - core-collapse supernova stage, 386–388

- dissociation of iron, 387
- electron capture, 387
- neutrino emission and neutrinosphere, 387–388
- PNS (proto-neutron star) phase, 387
- formation, 462–464
 - initial mass function, 411–414
 - initial metallicity, 396, 409, 410, 412, 414
 - Jeans mass, 412–414
 - Pop I stars, 411–413
 - Pop III stars, 411–414
- main sequence stages, 383–386
 - hydrogen-burning, proton–proton chain, 384
 - hydrogen-burning, CNO cycle, 384, 386, 389
 - helium-burning, 384, 389, 395
 - lifetimes, 384
 - luminosities, 384
- red giant stages, 384
 - He core, 384, 389
 - hydrogen shell burning, 384, 386, 389
 - C/O core, 385, 389
 - helium shell burning, 385–386, 389
 - carbon-burning, degenerate, 385
 - planetary nebula formation, 385
 - carbon shell burning, 386
- red supergiant stages, 386, 389
 - carbon-burning, non-degenerate, 386, 389
 - ONeMg core, 386, 389
 - carbon shell burning, 389
 - neon-burning, 389
 - OMgSi core, 389
 - neon shell burning, 389
 - oxygen-burning, 389
 - SiS core, 389
 - oxygen shell burning, 389
 - silicon-burning, 389
 - FeNi core, 389
 - silicon shell burning, 389
 - stellar wind mass loss, 395–396
 - Wolf–Rayet phase, 395–396
- supernova explosion mechanisms
 - core-collapse, magnetic/MHD bipolar outflow, 391–393
 - core-collapse, neutrino heating, 388
 - pair production, 414
 - synthesis of heavy elements/metals, 381–394
- Stellar structure & dynamics
 - non-relativistic stars, *see also* Hydrodynamics and MHD: NRHD:
 - stellar structure & dynamics
 - Chandrasekhar mass limit, 63, 86, 161, 162, 385, 386, 389, 406, 407, 731
 - collapse parameter, 138
 - free-fall time, 138, 570
 - Lane–Emden equation, 157–163
 - mass coordinates, 153–154
 - polytropes, 156–163
 - relativistic stars
 - black hole/horizon formation, 277–289
 - collapse of relativistic pressureless dust cloud, 281–286
 - collapse of relativistic star with pressure, 286, 288
 - collapse parameter, 282, 283, 285
 - free-fall time, 282
 - mass coordinates, 282, 283, 841–842, 845
 - static structure, 286–288
 - thermodynamic equation of state, 286–287
- Supernovae
 - core-collapse, 84–87, 99, 107, 386–388, 727–731, 743
 - explosion types

- Type Ia, 84, 99, 391, 402, 405–408, 731
- Type Ib, 84–86, 107, 112, 390, 396, 405, 406, 408
- Type Ic, 84–86, 107, 112, 390, 395, 396, 405, 406, 408
- Type Ic-BL (“Id”, “hypernovae”), 86, 87, 107, 111
- Type II, 85–86, 390, 395, 396, 406
- Type “IIa”, 85
- Type IIb, 85, 86, 395
- mechanisms, binary stars, *see* Stellar evolution, binary stars: supernova explosion mechanisms
- mechanisms, single stars, *see* Stellar evolution, single stars: supernova explosion mechanisms
- observed explosions
 - Crab supernova, 85, 388
- rates, 87, 113
 - core-collapse, 72, 99, 108, 397, 744
 - Type Ia, 407–408
- remnants
 - Crab Nebula (M1), 64, 68, 69
 - Vela Nebula, 69
 - W50, 81, 82
- Telescopes and techniques
 - ground
 - Einstein (gravitational wave) Telescope, 267
 - JCMT, 45
 - LIGO, 109, 110, 255, 267, 268, 273, 402, 405
 - MERLIN, 26, 29
 - Ryle 5-km telescope, 18
 - SCUBA, 45
 - VLA, 19, 29, 31, 47, 82, 93, 679, 760, 818
 - VLBA, 15, 819–821
 - VLBI, 17, 18, 25–27, 29, 30, 53, 630, 679, 705, 815, 818, 819
 - space
 - Aerobee sounding rocket, 63
 - ASCA, 41, 118, 119
 - Beppo-Sax, 101
 - CGRO, 89, 94, 101, 102
 - Chandra, 42, 69, 118, 120, 127
 - Einstein X-ray observatory, 42, 91, 117
 - Fermi/GLAST, 33, 89
 - GRANAT, 92
 - HST, 9, 15, 19, 26, 39, 40, 44, 48, 55, 122, 478
 - INTEGRAL, 127
 - IRAS, 44
 - JWST, 111, 435, 824
 - LISA, 255, 267, 268, 273, 402, 435
 - NuSTAR, 46, 127, 731, 760, 786, 803, 804, 826
 - ROSAT, 72, 109, 118
 - RXTE, 93, 94
 - SIM, 435
 - Spitzer Space Telescope, 43–46, 803
 - Swift γ -ray burst mission, 109, 110
 - Uhuru (SAS-1), 73, 88, 92
 - Vela nuclear monitoring satellites, 65
 - XMM, 120
- Thermodynamics
 - adiabatic index, *see* Equations of state: thermodynamic, internal energy/pressure: adiabatic index
 - equations of state, *see* Equations of state: thermodynamic
 - first law (conservation of energy), 150–152, 845
 - specific heat coefficients, *see* Equations of state: thermodynamic, other: specific heats

Time warps, 227–228

Vector and tensor physics, *see also*
Geometry/metrics ...

notation

two-dimensional and three-dimensional, 833–834
four-dimensional, 834–835

two-vectors

inner/scalar/dot product, 169,
171–172

three-vectors

inner/scalar/dot product, 155,
176
outer/dyadic product, 155, 337
unit, 135, 139, 141, 366
velocity, 134, 833
momentum, 134–136, 295–297,
337–338, 347, 348, 833, 863
force, 133–135, 143–146, 149,
156, 192–193, 296
current density, 191–193, 249,
338, 858
electric field, 191–193, 249,
337, 857
electric displacement, 191–193,
249, 338, 857
magnetic induction, 191–193,
249, 337, 856
magnetic field, 191–193, 249,
338, 857
magnetic vector/three-potential,
193–200, 856

four-vectors

outer/dyadic product, 255
unit, 187, 838
velocity, 183–184, 188, 202,
224–227, 238, 257, 296, 300–
305, 307, 834, 847–849
momentum, 185–187, 202, 226,
294, 296–297, 834, 847–849
force, 187, 202–203, 221, 226–
228, 294, 296, 849
current density, 202, 204–205,
210, 249, 307, 309, 834, 851

electromagnetic field, 204–205,
210
magnetic potential, 203–206,
254–256, 834, 835
spatial current density, 307–
310, 852
beamed spatial current density,
307–310, 852

tensors

identity, 156, 169, 198, 203,
347, 834
projection, 301, 310, 850
shear, 302, 353, 495
stress-energy-momentum, 188,
203, 217, 221, 256, 261, 299–
301, 303, 305, 835, 850
charge-current, 307, 852
Faraday 2-form/tensor, 201, 203,
249, 305, 835
Maxwell 2-form/tensor, 201,
204, 835
metric, 169–170, 175, 222, 234,
244, 247, 254, 841
Riemann 4-tensor, 212–218, 835
Riemann symmetries/Bianchi
identities, 202, 205, 210, 215–
219
Ricci 2-tensor, 217–219
Ricci scalar, 217, 218
Einstein 2-tensor, 217–221
inner/scalar/double-dot prod-
uct, 203, 218, 256, 260
vector and tensor derivatives
simple gradient, 251, 838
Christoffel symbols/connection
coefficients, 214, 219, 226,
254, 837
covariant gradient, 837–838
curl, 251
divergence, 251, 838–839
metric has no gradient or di-
vergence, 839–840

Waves and radiation (long wavelength)
electromagnetic

- dipole, electric, 197–200
- dipole, magnetic, 66–67
- equations, Maxwell's in covariant form, 202, 205
- equations, Maxwell's in stationary 3+1 metrics, 191, 193, 249
- equations, wave, 194, 205
- gauge, Lorenz transverse, 194, 205
- in a vacuum, 194–197, 205–206
- in media, 191, 250
- polarization, 196, 206
- Poynting energy flux/power, 199–200
- velocity of, 195, 206
- wave generation, 197–200
- general
 - dispersion relations, 357–358
 - Friedrich's polar diagrams, group, 359, 360, 364, 366
 - Friedrich's polar diagrams, phase, 359, 360, 364, 365
 - velocity, group, 357, 358
 - velocity, phase, 357, 358
 - wave number, 356, 357
- gravitational, 254–277, *see also*
 - Binary stars: gravitational waves
 - equations, Einstein's in covariant form, 220
 - equations, wave, 256
 - gauge is the coordinate system, 257–258
 - gauge, transverse traceless (Lorenz), 257
 - in a vacuum, 256–260
 - polarization, 258–260
 - Poynting energy flux/power, 260–263
 - quadrupole, 260–263
 - stress-energy-momentum of, 256
 - wave generation, close binaries, 263–268
 - wave generation, general, 260–277
 - wave generation, star mergers, 268–277
- HD/sound, non-relativistic
 - characteristics and sonic causality, 362–363
 - epicyclic modes, 371
 - in subsonic (elliptic) flow, 361–362
 - in supersonic (hyperbolic) flow, 361–363
 - Mach cone, 361–362
 - velocity of, adiabatic, 359
 - velocity of, isothermal, 361
- HD/sound, relativistic
 - velocity of, adiabatic, 369, 693
- MHD, non-relativistic
 - Alfvén mode, 363–364
 - magneto-acoustic modes (fast/slow), 364–366
 - velocity, Alfvén, 364
 - velocity, cusp, 366
 - velocity, fast magneto-acoustic, 365
 - velocity, magnetosound, 365
 - velocity, slow magneto-acoustic, 365
- MHD, relativistic
 - velocity, Alfvén, 368, 692
 - velocity, cusp, 369
 - velocity, fast magneto-acoustic, 368, 370
 - velocity, magnetosound, 369, 693
 - velocity, slow magneto-acoustic, 368, 370
- Waves and radiation (short wavelength), *see* Radiative processes (short wavelength)
- Wind and jet acceleration theory, 655–700
 - general prescription
 - critical surfaces, 680–682

- energy equations, 486, 536, 570, 574, 659, 674
- momentum equations, 486, 536, 570, 661
- separatrix surfaces, 680–682
- singular surfaces, 680–682
- stagnation surface, 609, 621, 630, 636, 637, 680, 698, 699
- resulting wind/accretion equations, 486, 490, 537, 565, 574, 661, 673, 685
- wind, ingoing, 609, 620–621, 630, 634–637, 680, 698, 699
- wind, outgoing, 595–596, 609, 621, 630, 634–637, 680, 699
- magnetically-driven, cold, non-relativistic (BP) bipolar winds/jets, 656–672
- r -self-similar models, 663–664
- Alfvén radius, 662
- black hole spin-dependence, 671–672
- Bernoulli energy equation, 659, 660
- Mach number, Alfvén $\mathcal{M}_{\text{NR}}$, 660, 663
- Mach number, magnetosonic t_{NR} , 661
- BP momentum (cross-field) equation, 661
- singular points/surfaces, 661–663
- terminal velocity, 666
- BP MHD wind equation, 659–661
- solutions, 663–665
- simulations, 664–665
- wind loss rates (energy, mass, angular momentum), 665–672
- magnetically-driven, cold, relativistic (LCB) bipolar winds/jets, 672–678
- r -self-similar models, 675–676
- Mach number, relativistic Alfvén $\mathcal{M}_{\text{R}}$, 673
- Mach number, relativistic magnetosonic t_{R} , 674
- magnetization parameter σ_{M} , 675, 694, 696, 697
- models with large initial toroidal field, 677–678
- relativistic Alfvén radius x_{Alf} , 674, 675, 694, 696
- singular points/surfaces, 674–675
- terminal Lorentz factor, 676
- LCB MHD wind equation, 673–675
- solutions, 675–677
- wind loss rates (energy, mass, angular momentum), 676–677
- magnetically-driven, warm, non-relativistic (VTST) bipolar winds/jets, 678–692
- observational insight into jet acceleration, 678–679
- theoretical insight and causality, 679–685
- plasma parameter β_{p} , 685
- VTST MHD wind equation, 685–686
- solutions, 686–690
- simulations, 690–692
- magnetically-driven, warm, relativistic (VK) bipolar winds/jets, 692–700
- proper Alfvén speed, 692
- proper gas sound speed, 693
- proper magnetosound speed, 693
- singular points/surfaces (SMSS, AS, FMSS), 694
- warmth parameter $\mathcal{Q}$, 694, 696, 697
- VK MHD wind equation, 692–694
- solutions, 694–696

- simulations, 696–700
- radiation continuum-driven, super-Eddington winds, 567–582, 780, 781
- “slim” disk connection, 580–582
- adiabatic radius, 575
- injection radius, 568–569, 573, 580
- photosphere radius, 569, 579, 580
- scattersphere radius, 569, 573, 578, 580
- singular/critical radius, 568–569, 578, 580
- structure, physical, 568–578
- structure, thermal, 578–580
- terminal velocity, 580
- trapping radius, 574–575
- wind equation, adiabatic, 574
- wind equation, radiative, 574
- wind equations, general, 569–570
- radiation line-driven, sub-Eddington winds, 564–567
- force multiplier, 564–566, 574, 575, 744
- hitchhiking gas, 49, 566–567, 793
- line opacity, 564, 565
- terminal velocity, 566
- wind equation, isothermal, 565
- thermally-driven (ADIOS) outflow, 582–586
- equations, 583–585
- models, 585–586
- terminal velocity, 586
- thermally-driven (Parker) winds, 485–492
- adiabatic, 488
- general solutions, 486–487
- isothermal, 488
- regular solutions, 486–488
- singular/critical (sonic) point/surface, 486–488
- wind equation, non-relativistic, 486–489
- wind equation, relativistic, 489–490
- thermally-driven bipolar winds/jets, 586–589
- broad-band jet spectrum, 587, 589
- bulk acceleration equation, 587–588
- Compton reflection, 587, 588
- particle acceleration, 588
- strong shock in jet, 587, 588
- synchrotron, 587, 589
- Wind and jet launching, from accretion disks, 641–654, 750–757
- gas pressure launching (slow mode), 645–646
- magnetocentrifugal (BP) launching (Alfvén mode) [fling], 642–644
- black hole spin-dependence, 632, 671–672
- Roche potential, 643, 644
- magnetic pressure launching (fast mode) [spring], 643–654
- closed field lines (magnetic tower), 646–654
- external pressure role, 646–648
- magnetic switch, 649, 651–653
- open field lines, 643–646, 649
- Wind and jet launching, from rotating black holes, 609–641, 755–757
- binary, 621–641, *see also* Black holes: magnetospheres, accreting
- black hole spin-dependence, 633–634, 815–817
- isolated, 609–621, *see also* Black holes: magnetospheres, isolated

- Wind and jet launching, from rotating neutron stars, 592–609, 732–733, 739–743
- binary, 603–609, 739–743, *see also* Pulsars: magnetospheres, accreting
- isolated, 592–603, 732–733, *see also* Pulsars: magnetospheres, isolated
- X-ray binaries, *see also* Binary stars:
- X-ray
 - black hole, *see also* Black holes, stellar-mass (SBHs): binary
 - high mass (HMXB), 74–76, 399, 400, 402
 - Roche-lobe-overflow mass transfer, 76, 399, 472–474, 762, 782
 - wind mass transfer, 76, 399, 474–475, 735
 - low mass (LMXB), 74–77, 740
 - Roche-lobe-overflow mass transfer, 75, 77, 472–474, 735
 - neutron star, *see also* Neutron stars(NSs): binary
 - persistent, 76–78
 - sources, *see* Sources in the Milky Way; Sources in other galaxies
 - transient, 76–78