Lecture Notes in Mathematics 2062

For further volumes:
http://www.springer.com/series/304

FONDAZIONE
CIME
ROBERTO CONTI
CENTRO INTERNAZIONALE MATEMATICO ESTIVO
INTERNATIONAL MATHEMATICAL SUMMER CENTER

Fondazione C.I.M.E., Firenze

C.I.M.E. stands for *Centro Internazionale Matematico Estivo*, that is, International Mathematical Summer Centre. Conceived in the early fifties, it was born in 1954 in Florence, Italy, and welcomed by the world mathematical community: it continues successfully, year for year, to this day.

Many mathematicians from all over the world have been involved in a way or another in C.I.M.E.'s activities over the years. The main purpose and mode of functioning of the Centre may be summarised as follows: every year, during the summer, sessions on different themes from pure and applied mathematics are offered by application to mathematicians from all countries. A Session is generally based on three or four main courses given by specialists of international renown, plus a certain number of seminars, and is held in an attractive rural location in Italy.

The aim of a C.I.M.E. session is to bring to the attention of younger researchers the origins, development, and perspectives of some very active branch of mathematical research. The topics of the courses are generally of international resonance. The full immersion atmosphere of the courses and the daily exchange among participants are thus an initiation to international collaboration in mathematical research.

C.I.M.E. Director
Pietro ZECCA
Dipartimento di Energetica "S. Stecco"
Università di Firenze
Via S. Marta, 3
50139 Florence
Italy
e-mail: zecca@unifi.it

C.I.M.E. Secretary
Elvira MASCOLO
Dipartimento di Matematica "U. Dini"
Università di Firenze
viale G.B. Morgagni 67/A
50134 Florence
Italy
e-mail: mascolo@math.unifi.it

For more information see CIME's homepage: http://www.cime.unifi.it

CIME activity is carried out with the collaboration and financial support of:

- INdAM (Istituto Nazionale di Alta Matematica)

- MIUR (Ministero dell'Universita' e della Ricerca)

Luigi Ambrosio • Alberto Bressan
Dirk Helbing • Axel Klar • Enrique Zuazua

Modelling and Optimisation of Flows on Networks

Cetraro, Italy 2009

Editors:
Benedetto Piccoli
Michel Rascle

Springer

Fondazione CIME Roberto Conti

Luigi Ambrosio
Scuola Normale Superiore
Department of Mathematics
Pisa, Italy

Alberto Bressan
Penn State University
State College
Department of Mathematics
University Park
PA, USA

Dirk Helbing
ETH Zürich
Swiss Federal Institute of Technology
Zurich, Switzerland

Axel Klar
Technische Universität
Kaiserslautern
Fachbereich Mathematik
Kaiserslautern, Germany

Enrique Zuazua
BCAM Basque Center for
Applied Mathema
Depto. Matematicas
Bilbao, Spain

ISBN 978-3-642-32159-7 ISBN 978-3-642-32160-3 (eBook)
DOI 10.1007/978-3-642-32160-3
Springer Heidelberg New York Dordrecht London

Lecture Notes in Mathematics ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692

Library of Congress Control Number: 2012949580

Mathematics Subject Classification (2010): 35R02, 35L65, 90B20

The contributions co-authored or authored by Helbing have previously been published in

The European Physical Journal B 69(4), 539–548, DOI: 10.1140/epjb/e2009-00192-5(2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ)

The European Physical Journal B 69(4), 549–562, DOI: 10.1140/epjb/e2009-00182-7 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ)

The European Physical Journal B 69(4), 583–598, DOI: 10.1140/epjb/e2009-00140-5 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ)

Networks and Heterogeneous Media 2(2), 193–210 (2007), © American Institute of Mathematical Sciences, Reproduction with kind permission of Networks and Heterogeneous Media (NHM).

The European Physical Journal B 70(2), 257–274, DOI: 10.1140/epjb/e2009-00213-5 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ)

Printed on acid-free paper

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

Preface

The present volume collects notes from lectures delivered for the CIME course on Modelling and optimisation of flows on networks, held in Cetraro in the summer of 2009.

In recent years modelling of flows on networks has been the subject of many investigations leading to an increasing number of research papers. Moreover, a wide set of possible applications, such as vehicular traffic, blood flow, supply chains and others, has directed the attention of mathematicians towards research domains usually populated by engineers, physicists or researchers with other expertise.

The aim of the CIME school was to gather summer courses which could give a wide view of modelling, analysis, numerics and control for dynamic flows on networks. Encompassing all application domains (including irrigation channels, data networks, air traffic management and others) was impossible, thus we focused on mathematical approaches, which are feasible for a number of applications, and a restricted set of specific applications, in particular vehicular traffic and supply chains. The attempt of finding a common ground, for different mathematical techniques to treat flows on networks, was already successful in a number of cases both at the level of research projects (such as the Italian national INDAM project 2005) and editorial initiatives (the foundation in 2006 of a new applied math journal entitled *Networks and Heterogeneous Media*).

The school took place in Cetraro, Italy, on June 15–19 2009. The course subjects were the following:

1. Introduction to conservation laws: Alberto Bressan (PennState)
2. Optimal transportation: Luigi Ambrosio (SNS, Pisa)
3. Pedestrian motions and vehicular traffic: Dirk Helbing (ETH)
4. Control and stabilization of waves on 1-D networks: Enrique Zuazua (BCAM)
5. Modelling and optimization of scalar flows on networks: Axel Klar (Kaiserlautern)
6. Fluid dynamic and kinetic models for supply chains: Christian Ringhofer (Arizona State)

Rationale Behind the Choice of Courses for CIME School

Taking into account the above-mentioned scientific background, courses for the CIME school were chosen in order to give a wide view over main mathematical techniques and their applications in specific contexts.

1. Analysis and control of linear PDEs on networks
2. Analysis of nonlinear PDEs on networks
3. Optimization techniques for complex networks
4. Numerical methods for PDEs on networks

To cover the first topic and last one for the linear PDE aspect, we decided to focus on wave equations on networks of one-dimensional structures and, in particular, on the use of spectral methods. Therefore, the choice was made to contact Enrique Zuazua, Director of the Basque Center for Applied Mathematics and a world leader on the subject. Prof. Zuazua also authored a volume on the subject (SMAI series, Springer-Verlag, 2006).

In many applications it is natural to use conservation laws to model flows on networks, thus for the second course we contacted Alberto Bressan of PennState University, who was one of the major contributors of the theory of systems of conservation laws in last 20 years and author of a well-known monograph (Cambridge University Press, 2000).

The fourth topic for the nonlinear aspect was covered in courses dealing also with applications, and illustrated below, of Klar and Ringhofer. Finally, for the third topic, we individuated optimal transportation as one of the most suited mathematical framework, and thus decided to contact Luigi Ambrosio of Scuola Normale Superiore of Pisa, who authored various recent fundamental papers in the subject and a monograph on the related topics of gradient flows in metric spaces (Birkauser, 2008).

For what concerns applications related to our main theme, Dirk Helbing of ETH of Zurich accepted to deliver a course covering both pedestrian dynamics and vehicular traffic. Helbing was one of the pioneers in providing advanced mathematical modelling for pedestrians with celebrated papers in Nature.

Then we focused on supply chain dynamics and thus contacted Christian Ringhofer of Arizona State University, who coauthored a pioneering paper in 2006 providing the first model of supply chains using PDEs. The course of Ringhofer dealt also with kinetic approaches.

Finally, Axel Klar of Kaiserlautern Technical University provided a course not only dealing with general modelling and numerics of conservation laws on networks but also treating coupled systems of ODEs and PDEs with examples from vehicular traffic, supply chains and sewer systems.

The present volume contains lecture notes from the first five courses of the CIME school. We wish readers a pleasant and fruitful reading.

Camden, NJ Benedetto Piccoli
Nice, France Michel Rascle

Contents

A User's Guide to Optimal Transport

Luigi Ambrosio and Nicola Gigli

Abstract This text is an expanded version of the lectures given by the first author in the 2009 CIME summer school of Cetraro. It provides a quick and reasonably account of the classical theory of optimal mass transportation and of its more recent developments, including the metric theory of gradient flows, geometric and functional inequalities related to optimal transportation, the first and second order differential calculus in the Wasserstein space and the synthetic theory of metric measure spaces with Ricci curvature bounded from below.

1 Introduction

The opportunity to write down these notes on Optimal Transport has been the CIME course in Cetraro given by the first author in 2009. Later on the second author joined to the project, and the initial set of notes has been enriched and made more detailed, in particular in connection with the differentiable structure of the Wasserstein space, the synthetic curvature bounds and their analytic implications. Some of the results presented here have not yet appeared in a book form, with the exception of [44].

It is clear that this subject is expanding so quickly that it is impossible to give an account of all developments of the theory in a few hours, or a few pages. A more modest approach is to give a quick mention of the many aspects of the theory,

L. Ambrosio (✉)
Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy
e-mail: l.ambrosio@sns.it

N. Gigli
Université de Nice, Mathématiques, Parc Valrose, 06108 Nice, France
e-mail: nicola.gigli@unice.fr

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_1,

stimulating the reader's curiosity and leaving to more detailed treatises as [7] (mostly focused on the theory of gradient flows) and the monumental book [80] (for a—much—broader overview on optimal transport).

In chapter "A User's Guide to Optimal Transport" we introduce the optimal transport problem and its formulations in terms of transport maps and transport plans. Then we introduce basic tools of the theory, namely the duality formula, the c-monotonicity and discuss the problem of existence of optimal maps in the model case cost=distance^2.

In chapter "Hyperbolic Conservation Laws: An Illustrated Tutorial" we introduce the Wasserstein distance W_2 on the set $\mathscr{P}_2(X)$ of probability measures with finite quadratic moments and X is a generic Polish space. This distance naturally arises when considering the optimal transport problem with quadratic cost. The connections between geodesics in $\mathscr{P}_2(X)$ and geodesics in X and between the time evolution of Kantorovich potentials and the Hopf–Lax semigroup are discussed in detail. Also, when looking at geodesics in this space, and in particular when the underlying metric space X is a Riemannian manifold M, one is naturally lead to the so-called time-dependent optimal transport problem, where geodesics are singled out by an action minimization principle. This is the so-called Benamou–Brenier formula, which is the first step in the interpretation of $\mathscr{P}_2(M)$ as an infinite-dimensional Riemannian manifold, with W_2 as Riemannian distance. We then further exploit this viewpoint following Otto's seminal work [67].

In chapter "Derivation of Non-local Macroscopic Traffic Equations and Consistent Traffic Pressures from Microscopic Car-Following Models" we make a quite detailed introduction to the theory of gradient flows, borrowing almost all material from [7]. First we present the classical theory, for λ-convex functionals in Hilbert spaces. Then we present some equivalent formulations that involve only the distance, and therefore are applicable (at least in principle) to general metric space. They involve the derivative of the distance from a point (the (EVI) formulation) or the rate of dissipation of the energy (the (EDE) and (EDI) formulations). For all these formulations there is a corresponding discrete version of the gradient flow formulation given by the implicit Euler scheme. We will then show that there is convergence of the scheme to the continuous solution as the time discretization parameter tends to 0. The (EVI) formulation is the stronger one, in terms of uniqueness, contraction and regularizing effects. On the other hand this formulation depends on a compatibility condition between energy and distance; this condition is fulfilled in Non Positively Curved spaces in the sense of Alexandrov if the energy is convex along geodesics. Luckily enough, the compatibility condition holds even for some important model functionals in $\mathscr{P}_2(\mathbb{R}^n)$ (sum of the so-called internal, potential and interaction energies), even though the space is Positively Curved in the sense of Alexandrov.

In chapter "On the Controversy Around Daganzo's Requiem for and Aw–Rascle's Resurrection of Second-Order Traffic Flow Models" we illustrate the power of optimal transportation techniques in the proof of some classical functional/geometric inequalities: the Brunn–Minkowski inequality, the isoperimetric inequality and the Sobolev inequality. Recent works in this area have also shown the possibility to prove by optimal transportation methods optimal effective versions of these inequalities: for instance we can quantify the closedness of E to a ball with the same volume in terms of the vicinity of the isoperimetric ratio of E to the optimal one.

Chapter "Theoretical vs. Empirical Classification and Prediction of Congested Traffic States" is devoted to the presentation of three recent variants of the optimal transport problem, which lead to different notions of Wasserstein distance: the first one deals with variational problems giving rise to branched transportation structures, with a "Y shaped path" opposed to the "V shaped one" typical of the mass splitting occurring in standard optimal transport problems. The second one involves modification in the action functional on curves arising in the Benamou–Brenier formula: this leads to many different optimal transportation distances, maybe more difficult to describe from the Lagrangian viewpoint, but still with quite useful implications in evolution PDE's and functional inequalities. The last one deals with transportation distance between measures with unequal mass, a variant useful in the modeling problems with Dirichlet boundary conditions.

Chapter "Self-organized Network Flows" deals with a more detailed analysis of the differentiable structure of $\mathscr{P}_2(\mathbb{R}^d)$: besides the analytic tangent space arising from the Benamou–Brenier formula, also the "geometric" tangent space, based on constant speed geodesics emanating from a given base point, is introduced. We also present Otto's viewpoint on the duality between Wasserstein space and Arnold's manifolds of measure-preserving diffeomorphisms. A large part of the chapter is also devoted to the second order differentiable properties, involving curvature. The notions of parallel transport along (sufficiently regular) geodesics and Levi–Civita connection in the Wasserstein space are discussed in detail.

Finally, chapter "Operation Regimes and Slower-Is-Faster-Effect in the Control of Traffic Intersections" is devoted to an introduction to the synthetic notions of Ricci lower bounds for metric measure spaces introduced by Lott–Villani and Sturm in recent papers. This notion is based on suitable convexity properties of a dimension-dependent internal energy along Wasserstein geodesics. Synthetic Ricci bounds are completely consistent with the smooth Riemannian case and stable under measured-Gromov–Hausdorff limits. For this reason these bounds, and their analytic implications, are a useful tool in the description of measured-GH-limits of Riemannian manifolds.

2 The Optimal Transport Problem

2.1 Monge and Kantorovich Formulations of the Optimal Transport Problem

Given a Polish space (X, d) (i.e. a complete and separable metric space), we will denote by $\mathscr{P}(X)$ the set of Borel probability measures on X. By support $\mathrm{supp}(\mu)$ of a measure $\mu \in \mathscr{P}(X)$ we intend the smallest closed set on which μ is concentrated.

If X, Y are two Polish spaces, $T : X \to Y$ is a Borel map, and $\mu \in \mathscr{P}(X)$ a measure, the measure $T_\#\mu \in \mathscr{P}(Y)$, called the *push forward of* μ *through* T is defined by

$$T_\#\mu(E) = \mu(T^{-1}(E)), \qquad \forall E \subset Y, \text{ Borel.}$$

The push forward is characterized by the fact that

$$\int f dT_\#\mu = \int f \circ T d\mu,$$

for every Borel function $f : Y \to \mathbb{R} \cup \{\pm\infty\}$, where the above identity has to be understood in the following sense: one of the integrals exists (possibly attaining the value $\pm\infty$) if and only if the other one exists, and in this case the values are equal.

Now fix a Borel *cost function* $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. The Monge version of the transport problem is the following:

Problem 2.1 (Monge's optimal transport problem). Let $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$. Minimize

$$T \mapsto \int_X c\big(x, T(x)\big)\, d\mu(x)$$

among all *transport maps* T from μ to ν, i.e. all maps T such that $T_\#\mu = \nu$. ■

Regardless of the choice of the cost function c, Monge's problem can be ill-posed because:

- No admissible T exists (for instance if μ is a Dirac delta and ν is not).
- The constraint $T_\#\mu = \nu$ is not weakly sequentially closed, w.r.t. any reasonable weak topology.

As an example of the second phenomenon, one can consider the sequence $f_n(x) := f(nx)$, where $f : \mathbb{R} \to \mathbb{R}$ is 1-periodic and equal to 1 on $[0, 1/2)$ and to -1 on $[1/2, 1)$, and the measures $\mu := \mathscr{L}|_{[0,1]}$ and $\nu := (\delta_{-1} + \delta_1)/2$. It is immediate to check that $(f_n)_\#\mu = \nu$ for every $n \in \mathbb{N}$, and yet (f_n) weakly converges to the null function $f \equiv 0$ which satisfies $f_\#\mu = \delta_0 \neq \nu$.

A way to overcome these difficulties is due to Kantorovich, who proposed the following way to relax the problem:

Problem 2.2 (Kantorovich's formulation of optimal transportation). We minimize

$$\gamma \mapsto \int_{X\times Y} c(x, y)\, d\gamma(x, y)$$

in the set $\mathcal{A}dm(\mu, \nu)$ of all *transport plans* $\gamma \in \mathscr{P}(X \times Y)$ from μ to ν, i.e. the set of Borel Probability measures on $X \times Y$ such that

$$\gamma(A \times Y) = \mu(A) \quad \forall A \in \mathscr{B}(X), \qquad \gamma(X \times B) = \nu(B) \quad \forall B \in \mathscr{B}(Y).$$

Equivalently: $\pi^X_\#\gamma = \mu$, $\pi^Y_\#\gamma = \nu$, where π^X, π^Y are the natural projections from $X \times Y$ onto X and Y respectively. ■

Transport plans can be thought of as "multivalued" transport maps: $\gamma = \int \gamma_x \, d\mu(x)$, with $\gamma_x \in \mathscr{P}(\{x\} \times Y)$. Another way to look at transport plans is to observe that for $\gamma \in \mathcal{A}dm(\mu, \nu)$, the value of $\gamma(A \times B)$ is the amount of mass initially in A which is sent into the set B.

There are several advantages in the Kantorovich formulation of the transport problem:

- $\mathcal{A}dm(\mu, \nu)$ is always not empty (it contains $\mu \times \nu$).
- The set $\mathcal{A}dm(\mu, \nu)$ is convex and compact w.r.t. the narrow topology in $\mathscr{P}(X \times Y)$ (see below for the definition of narrow topology and Theorem 2.5), and $\gamma \mapsto \int c \, d\gamma$ is linear.
- Minima always exist under mild assumptions on c (Theorem 2.5).
- Transport plans "include" transport maps, since $T_\# \mu = \nu$ implies that $\gamma := (Id \times T)_\# \mu$ belongs to $\mathcal{A}dm(\mu, \nu)$.

In order to prove existence of minimizers of Kantorovich's problem we recall some basic notions concerning analysis over a Polish space. We say that a sequence $(\mu_n) \subset \mathscr{P}(X)$ *narrowly* converges to μ provided

$$\int \varphi \, d\mu_n \quad \mapsto \quad \int \varphi \, d\mu, \qquad \forall \varphi \in C_b(X),$$

$C_b(X)$ being the space of continuous and bounded functions on X. It can be shown that the topology of narrow convergence is metrizable. A set $\mathscr{K} \subset \mathscr{P}(X)$ is called *tight* provided for every $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset X$ such that

$$\mu(X \setminus K_\varepsilon) \leq \varepsilon, \qquad \forall \mu \in \mathscr{K}.$$

It holds the following important result.

Theorem 2.3 (Prokhorov). *Let (X, d) be a Polish space. Then a family $\mathscr{K} \subset \mathscr{P}(X)$ is relatively compact w.r.t. the narrow topology if and only if it is tight.*

Notice that if $\mathscr{K}$ contains only one measure, one recovers Ulam's theorem: any Borel probability measure on a Polish space is concentrated on a σ-compact set.

Remark 2.4. The inequality

$$\gamma(X \times Y \setminus K_1 \times K_2) \leq \mu(X \setminus K_1) + \nu(Y \setminus K_2), \tag{1}$$

valid for any $\gamma \in \mathcal{A}dm(\mu, \nu)$, shows that if $\mathscr{K}_1 \subset \mathscr{P}(X)$ and $\mathscr{K}_2 \subset \mathscr{P}(Y)$ are tight, then so is the set

$$\left\{ \gamma \in \mathscr{P}(X \times Y) \ : \ \pi^X_\# \gamma \in \mathscr{K}_1, \ \pi^Y_\# \gamma \in \mathscr{K}_2 \right\}$$

■

Existence of minimizers for Kantorovich's formulation of the transport problem now comes from a standard lower-semicontinuity and compactness argument:

Theorem 2.5. *Assume that c is lower semicontinuous and bounded from below. Then there exists a minimizer for Problem 2.2.*

Proof. **Compactness.** Remark 2.4 and Ulam's theorem show that the set $\mathscr{A}dm(\mu,\nu)$ is tight in $\mathscr{P}(X \times Y)$, and hence relatively compact by Prokhorov theorem.

To get the narrow compactness, pick a sequence $(\gamma_n) \subset \mathscr{A}dm(\mu,\nu)$ and assume that $\gamma_n \to \gamma$ narrowly: we want to prove that $\gamma \in \mathscr{A}dm(\mu,\nu)$ as well. Let φ be any function in $C_b(X)$ and notice that $(x,y) \mapsto \varphi(x)$ is continuous and bounded in $X \times Y$, hence we have

$$\int \varphi \, d\pi^X_\# \gamma = \int \varphi(x) \, d\gamma(x,y) = \lim_{n\to\infty} \int \varphi(x) \, d\gamma_n(x,y) = \lim_{n\to\infty} \int \varphi \, d\pi^X_\# \gamma_n = \int \varphi \, d\mu,$$

so that by the arbitrariness of $\varphi \in C_b(X)$ we get $\pi^X_\# \gamma = \mu$. Similarly we can prove $\pi^Y_\# \gamma = \nu$, which gives $\gamma \in \mathscr{A}dm(\mu,\nu)$ as desired.

Lower semicontinuity. We claim that the functional $\gamma \mapsto \int c \, d\gamma$ is l.s.c. with respect to narrow convergence. This is true because our assumptions on c guarantee that there exists an increasing sequence of functions $c_n : X \times Y \to \mathbb{R}$ continuous an bounded such that $c(x,y) = \sup_n c_n(x,y)$, so that by monotone convergence it holds

$$\int c d\gamma = \sup_n \int c_n \, d\gamma.$$

Since by construction $\gamma \mapsto \int c_n \, d\gamma$ is narrowly continuous, the proof is complete. □

We will denote by $\mathcal{O}pt(\mu,\nu)$ the set of *optimal plans* from μ to ν for the Kantorovich formulation of the transport problem, i.e. the set of minimizers of Problem 2.2. More generally, we will say that a plan is optimal, if it is optimal between its own marginals. Observe that with the notation $\mathcal{O}pt(\mu,\nu)$ we are losing the reference to the cost function c, which of course affects the set itself, but the context will always clarify the cost we are referring to.

Once existence of optimal plans is proved, a number of natural questions arise:

- Are optimal plans unique?
- Is there a simple way to check whether a given plan is optimal or not?
- Do optimal plans have any natural regularity property? In particular, are they induced by maps?
- How far is the minimum of Problem 2.2 from the infimum of Problem 2.1?

This latter question is important to understand whether we can really consider Problem 2.2 the relaxation of Problem 2.1 or not. It is possible to prove that if c is continuous and μ is non atomic, then

$$\inf(\text{Monge}) = \min(\text{Kantorovich}), \tag{2}$$

so that transporting with plans can't be strictly cheaper than transporting with maps. We won't detail the proof of this fact.

2.2 Necessary and Sufficient Optimality Conditions

To understand the structure of optimal plans, probably the best thing to do is to start with an example.

Let $X = Y = \mathbb{R}^d$ and $c(x, y) := |x - y|^2/2$. Also, assume that μ, $\nu \in \mathscr{P}(\mathbb{R}^d)$ are supported on finite sets. Then it is immediate to verify that a plan $\gamma \in \mathcal{A}dm(\mu, \nu)$ is optimal if and only if it holds

$$\sum_{i=1}^{N} \frac{|x_i - y_i|^2}{2} \le \sum_{i=1}^{N} \frac{|x_i - y_{\sigma(i)}|^2}{2},$$

for any $N \in \mathbb{N}$, $(x_i, y_i) \in \operatorname{supp}(\gamma)$ and σ permutation of the set $\{1, \dots, N\}$. Expanding the squares we get

$$\sum_{i=1}^{N} \langle x_i, y_i \rangle \ge \sum_{i=1}^{N} \left\langle x_i, y_{\sigma(i)} \right\rangle,$$

which by definition means that the support of γ is cyclically monotone. Let us recall the following theorem:

Theorem 2.6 (Rockafellar). *A set $\Gamma \subset \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone if and only if there exists a convex and lower semicontinuous function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that Γ is included in the graph of the subdifferential of φ.*

We skip the proof of this theorem, because later on we will prove a much more general version. What we want to point out here is that under the above assumptions on μ and ν we have that the following three things are equivalent:

- $\gamma \in \mathcal{A}dm(\mu, \nu)$ is optimal.
- $\operatorname{supp}(\gamma)$ is cyclically monotone.
- There exists a convex and lower semicontinuous function φ such that γ is concentrated on the graph of the subdifferential of φ.

The good news is that the equivalence between these three statements holds in a much more general context (more general underlying spaces, cost functions, measures). Key concepts that are needed in the analysis, are the generalizations of the concepts of cyclical monotonicity, convexity and subdifferential which fit with a general cost function c.

The definitions below make sense for a general Borel and real valued cost.

Definition 2.7 (c-cyclical monotonicity). We say that $\Gamma \subset X \times Y$ is *c-cyclically monotone* if $(x_i, y_i) \in \Gamma$, $1 \le i \le N$, implies

$$\sum_{i=1}^{N} c(x_i, y_i) \le \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}) \qquad \text{for all permutations} \sigma \text{of } \{1, \dots, N\}.$$

Definition 2.8 (c-transforms). Let $\psi : Y \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Its c_+-transform $\psi^{c+} : X \to \mathbb{R} \cup \{-\infty\}$ is defined as

$$\psi^{c+}(x) := \inf_{y \in Y} c(x, y) - \psi(y).$$

Similarly, given $\varphi : X \to \mathbb{R} \cup \{\pm\infty\}$, its c_+-transform is the function $\varphi^{c+} : Y \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$\varphi^{c+}(y) := \inf_{x \in X} c(x, y) - \varphi(x).$$

The c_--transform $\psi^{c-} : X \to \mathbb{R} \cup \{+\infty\}$ of a function ψ on Y is given by

$$\psi^{c-}(x) := \sup_{y \in Y} -c(x, y) - \psi(y),$$

and analogously for c_--transforms of functions φ on X.

Definition 2.9 (c-concavity and c-convexity). We say that $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ is c-concave if there exists $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that $\varphi = \psi^{c+}$. Similarly, $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ is c-concave if there exists $\varphi : Y \to \mathbb{R} \cup \{-\infty\}$ such that $\psi = \varphi^{c+}$.

Symmetrically, $\varphi : X \to \mathbb{R} \cup \{+\infty\}$ is c-convex if there exists $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ such that $\varphi = \psi^{c-}$, and $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ is c-convex if there exists $\varphi : Y \to \mathbb{R} \cup \{+\infty\}$ such that $\psi = \varphi^{c-}$.

Observe that $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ is c-concave if and only if $\varphi^{c+c+} = \varphi$. This is a consequence of the fact that for any function $\psi : Y \to \mathbb{R} \cup \{\pm\infty\}$ it holds $\psi^{c+} = \psi^{c+c+c+}$, indeed

$$\psi^{c+c+c+}(x) = \inf_{\tilde{y} \in Y} \sup_{\tilde{x} \in X} \inf_{y \in Y} c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + c(\tilde{x}, y) - \psi(y),$$

and choosing $\tilde{x} = x$ we get $\psi^{c+c+c+} \geq \psi^{c+}$, while choosing $y = \tilde{y}$ we get $\psi^{c+c+c+} \leq \psi^{c+}$. Similarly for functions on Y and for the c-convexity.

Definition 2.10 (c-superdifferential and c-subdifferential). Let $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ be a c-concave function. The c-superdifferential $\partial^{c+}\varphi \subset X \times Y$ is defined as

$$\partial^{c+}\varphi := \Big\{(x, y) \in X \times Y \ : \ \varphi(x) + \varphi^{c+}(y) = c(x, y)\Big\}.$$

The c-superdifferential $\partial^{c+}\varphi(x)$ at $x \in X$ is the set of $y \in Y$ such that $(x, y) \in \partial^{c+}\varphi$. A symmetric definition is given for c-concave functions $\psi : Y \to \mathbb{R} \cup \{-\infty\}$.

The definition of c-subdifferential ∂^{c-} of a c-convex function $\varphi : X \to \{+\infty\}$ is analogous:

$$\partial^{c-}\varphi := \Big\{(x, y) \in X \times Y \ : \ \varphi(x) + \varphi^{c-}(y) = -c(x, y)\Big\}.$$

Analogous definitions hold for c-concave and c-convex functions on Y.

Remark 2.11 (The base case: $c(x, y) = -\langle x, y\rangle$). Let $X = Y = \mathbb{R}^d$ and $c(x, y) = -\langle x, y\rangle$. Then a direct application of the definitions show that:

- A set is c-cyclically monotone if and only if it is cyclically monotone.
- A function is c-convex (resp. c-concave) if and only if it is convex and lower semicontinuous (resp. concave and upper semicontinuous).
- The c-subdifferential of the c-convex (resp. c-superdifferential of the c-concave) function is the classical subdifferential (resp. superdifferential).
- The c_- transform is the Legendre transform.

Thus in this situation these new definitions become the classical basic definitions of convex analysis. ■

Remark 2.12 (For most applications c-concavity is sufficient). There are several trivial relations between c-convexity, c-concavity and related notions. For instance, φ is c-concave if and only if $-\varphi$ is c-convex, $-\varphi^{c+} = (-\varphi)^{c-}$ and $\partial^{c+}\varphi = \partial^{c-}(-\varphi)$. Therefore, roughly said, every statement concerning c-concave functions can be restated in a statement for c-convex ones. Thus, choosing to work with c-concave or c-convex functions is actually a matter of taste.

Our choice is to work with c-concave functions. Thus all the statements from now on will deal only with these functions. There is only one, important, part of the theory where the distinction between c-concavity and c-convexity is useful: in the study of geodesics in the Wasserstein space (see Sect. 3.2, and in particular Theorem 3.18 and its consequence Corollary 3.24).

We also point out that the notation used here is different from the one in [80], where a less symmetric notion (but better fitting the study of geodesics) of c-concavity and c-convexity has been preferred. ■

An equivalent characterization of the c-superdifferential is the following: $y \in \partial^{c+}\varphi(x)$ if and only if it holds

$$\begin{aligned} \varphi(x) &= c(x, y) - \varphi^{c+}(y), \\ \varphi(z) &\leq c(z, y) - \varphi^{c+}(y), \qquad \forall z \in X, \end{aligned}$$

or equivalently if

$$\varphi(x) - c(x, y) \geq \varphi(z) - c(z, y), \qquad \forall z \in X. \tag{3}$$

A direct consequence of the definition is that the c-superdifferential of a c-concave function is always a c-cyclically monotone set, indeed if $(x_i, y_i) \in \partial^{c+}\varphi$ it holds

$$\sum_i c(x_i, y_i) = \sum_i \varphi(x_i) + \varphi^c(y_i) = \sum_i \varphi(x_i) + \varphi^c(y_{\sigma(i)}) \leq \sum_i c(x_i, y_{\sigma(i)}),$$

for any permutation σ of the indexes.

What is important to know is that actually under mild assumptions on c, *every* c-cyclically monotone set can be obtained as the c-superdifferential of a c-concave function. This result is part of the following important theorem:

Theorem 2.13 (Fundamental theorem of optimal transport). *Assume that* $c : X \times Y \to \mathbb{R}$ *is continuous and bounded from below and let* $\mu \in \mathscr{P}(X)$*,* $\nu \in \mathscr{P}(Y)$ *be such that*

$$c(x,y) \le a(x) + b(y), \tag{4}$$

for some $a \in L^1(\mu)$*,* $b \in L^1(\nu)$*. Also, let* $\gamma \in \mathcal{A}dm(\mu,\nu)$*. Then the following three are equivalent:*

(i) The plan γ *is optimal.*
(ii) The set $\mathrm{supp}(\gamma)$ *is* c*-cyclically monotone.*
(iii) There exists a c*-concave function* φ *such that* $\max\{\varphi, 0\} \in L^1(\mu)$ *and* $\mathrm{supp}(\gamma) \subset \partial^{c+}\varphi$*.*

Proof. Observe that the inequality (4) together with

$$\begin{aligned}\int c(x,y)d\tilde{\gamma}(x,y) &\le \int a(x) + b(y) d\tilde{\gamma}(x,y)\\ &= \int a(x)d\mu(x) + \int b(y)d\nu(y) < \infty, \quad \forall \tilde{\gamma} \in \mathcal{A}dm(\mu,\nu)\end{aligned}$$

implies that for any admissible plan $\tilde{\gamma} \in \mathcal{A}dm(\mu,\nu)$ the function $\max\{c, 0\}$ is integrable. This, together with the bound from below on c gives that $c \in L^1(\tilde{\gamma})$ for any admissible plan $\tilde{\gamma}$.

(i) $\Rightarrow$ **(ii)** We argue by contradiction: assume that the support of γ is not c-cyclically monotone. Thus we can find $N \in \mathbb{N}$, $\{(x_i, y_i)\}_{1 \le i \le N} \subset \mathrm{supp}(\gamma)$ and some permutation σ of $\{1, \dots, N\}$ such that

$$\sum_{i=1}^{N} c(x_i, y_i) > \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}).$$

By continuity we can find neighborhoods $U_i \ni x_i$, $V_i \ni y_i$ with

$$\sum_{i=1}^{N} c(u_i, v_{\sigma(i)}) - c(u_i, v_i) < 0 \qquad \forall (u_i, v_i) \in U_i \times V_i,\ 1 \le i \le N.$$

Our goal is to build a "variation" $\tilde{\gamma} = \gamma + \eta$ of γ in such a way that minimality of γ is violated. To this aim, we need a *signed* measure η with:

(A) $\eta^- \le \gamma$ (so that $\tilde{\gamma}$ is nonnegative).
(B) Null first and second marginal (so that $\tilde{\gamma} \in \mathcal{A}dm(\mu,\nu)$).
(C) $\int c \, d\eta < 0$ (so that γ is not optimal).

Let $\Omega := \Pi_{i=1}^N U_i \times V_i$ and $\mathbf{P} \in \mathscr{P}(\Omega)$ be defined as the product of the measures $\frac{1}{m_i}\gamma|_{U_i \times V_i}$, where $m_i := \gamma(U_i \times V_i)$. Denote by π^{U_i}, π^{V_i} the natural projections of Ω to U_i and V_i respectively and define

$$\eta := \frac{\min_i m_i}{N} \sum_{i=i}^{N} (\pi^{U_i}, \pi^{V_{\sigma(i)}})_\# \mathbf{P} - (\pi^{U_i}, \pi^{V_{(i)}})_\# \mathbf{P}.$$

It is immediate to verify that η fulfills (A), (B), (C) above, so that the thesis is proven.

(ii) $\Rightarrow$ **(iii)** We need to prove that if $\Gamma \subset X \times Y$ is a c-cyclically monotone set, then there exists a c-concave function φ such that $\partial^c \varphi \supset \Gamma$ and $\max\{\varphi, 0\} \in L^1(\mu)$. Fix $(\overline{x}, \overline{y}) \in \Gamma$ and observe that, since we want φ to be c-concave with the c-superdifferential that contains Γ, for any choice of $(x_i, y_i) \in \Gamma$, $i = 1, \dots, N$, we need to have

$$\begin{aligned}
\varphi(x) &\le c(x, y_1) - \varphi^{c+}(y_1) = c(x, y_1) - c(x_1, y_1) + \varphi(x_1)\\
&\le \Big(c(x, y_1) - c(x_1, y_1)\Big) + c(x_1, y_2) - \varphi^{c+}(y_2)\\
&= \Big(c(x, y_1) - c(x_1, y_1)\Big) + \Big(c(x_1, y_2) - c(x_2, y_2)\Big) + \varphi(x_2)\\
&\le \cdots\\
&\le \Big(c(x, y_1) - c(x_1, y_1)\Big) + \Big(c(x_1, y_2) - c(x_2, y_2)\Big) + \cdots + \Big(c(x_N, \overline{y}) - c(\overline{x}, \overline{y})\Big) + \varphi(\overline{x}).
\end{aligned}$$

It is therefore natural to define φ as the infimum of the above expression as $\{(x_i, y_i)\}_{i=1,\dots,N}$ vary among all N-ples in Γ and N varies in $\mathbb{N}$. Also, since we are free to add a constant to φ, we can neglect the addendum $\varphi(\overline{x})$ and define:

$$\varphi(x) := \inf \Big(c(x, y_1) - c(x_1, y_1)\Big) + \Big(c(x_1, y_2) - c(x_2, y_2)\Big) + \cdots + \Big(c(x_N, \overline{y}) - c(\overline{x}, \overline{y})\Big),$$

the infimum being taken on $N \ge 1$ integer and $(x_i, y_i) \in \Gamma$, $i = 1, \dots, N$. Choosing $N = 1$ and $(x_1, y_1) = (\overline{x}, \overline{y})$ we get $\varphi(\overline{x}) \le 0$. Conversely, from the c-cyclical monotonicity of Γ we have $\varphi(\overline{x}) \ge 0$. Thus $\varphi(\overline{x}) = 0$.

Also, it is clear from the definition that φ is c-concave. Choosing again $N = 1$ and $(x_1, y_1) = (\overline{x}, \overline{y})$, using (3) we get

$$\varphi(x) \le c(x, \overline{y}) - c(\overline{x}, \overline{y}) < a(x) + b(\overline{y}) - c(\overline{x}, \overline{y}),$$

which, together with the fact that $a \in L^1(\mu)$, yields $\max\{\varphi, 0\} \in L^1(\mu)$. Thus, we need only to prove that $\partial^{c+}\varphi$ contains Γ. To this aim, choose $(\tilde{x}, \tilde{y}) \in \Gamma$, let $(x_1, y_1) = (\tilde{x}, \tilde{y})$ and observe that by definition of $\varphi(x)$ we have

$$\begin{aligned}
\varphi(x) &\le c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + \inf \Big(c(\tilde{x}, y_2) - c(x_2, y_2)\Big) + \cdots + \Big(c(x_N, \overline{y}) - c(\overline{x}, \overline{y})\Big)\\
&= c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + \varphi(\tilde{x}).
\end{aligned}$$

By the characterization (3), this inequality shows that $(\tilde{x}, \tilde{y}) \in \partial^{c+}\varphi$, as desired.

(iii) $\Rightarrow$ **(i)**. Let $\tilde{\gamma} \in \mathcal{A}dm(\mu, \nu)$ be any transport plan. We need to prove that $\int c\,d\gamma \leq \int c\,d\tilde{\gamma}$. Recall that we have

$$\begin{aligned} \varphi(x) + \varphi^{c+}(y) &= c(x, y), \qquad && \forall (x, y) \in \operatorname{supp}(\gamma) \\ \varphi(x) + \varphi^{c+}(y) &\leq c(x, y), \qquad && \forall x \in X,\ y \in Y, \end{aligned}$$

and therefore

$$\begin{aligned} \int c(x, y)d\gamma(x, y) &= \int \varphi(x) + \varphi^{c+}(y)d\gamma(x, y) = \int \varphi(x)d\mu(x) + \int \varphi^{c+}(y)d\nu(y) \\ &= \int \varphi(x) + \varphi^{c+}(y)d\tilde{\gamma}(x, y) \leq \int c(x, y)d\tilde{\gamma}(x, y). \end{aligned}$$

□

Remark 2.14. Condition (4) is natural in some, but not all, problems. For instance problems with constraints or in Wiener spaces (infinite-dimensional Gaussian spaces) include $+\infty$-valued costs, with a "large" set of points where the cost is not finite. We won't discuss these topics. ■

An important consequence of the previous theorem is that being optimal is a property that depends only on the support of the plan γ, and not on how the mass is distributed in the support itself: if γ is an optimal plan (between its own marginals) and $\tilde{\gamma}$ is such that $\operatorname{supp}(\tilde{\gamma}) \subset \operatorname{supp}(\gamma)$, then $\tilde{\gamma}$ is optimal as well (between its own marginals, of course). We will see in Proposition 3.5 that one of the important consequences of this fact is the *stability of optimality*.

Analogous arguments works for maps. Indeed assume that $T : X \to Y$ is a map such that $T(x) \in \partial^{c+}\varphi(x)$ for some c-concave function φ for all x. Then, for every $\mu \in \mathscr{P}(X)$ such that condition (4) is satisfied for $\nu = T_{\#}\mu$, the map T is optimal between μ and $T_{\#}\mu$. Therefore it makes sense to say that T is an optimal map, without explicit mention to the reference measures.

Remark 2.15. From Theorem 2.13 we know that given $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$ satisfying the assumption of the theorem, for every optimal plan γ there exists a c-concave function φ such that $\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi$. Actually, a stronger statement holds, namely: if $\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi$ for some optimal γ, then $\operatorname{supp}(\gamma') \subset \partial^{c+}\varphi$ for *every* optimal plan γ'. Indeed arguing as in the proof of 2.13 one can see that $\max\{\varphi, 0\} \in L^1(\mu)$ implies $\max\{\varphi^{c+}, 0\} \in L^1(\nu)$ and thus it holds

$$\begin{aligned} \int \varphi d\mu + \int \varphi^{c+} d\nu &= \int \varphi(x) + \varphi^{c+}(y)d\gamma'(x, y) \\ &\leq \int c(x, y)d\gamma'(x, y) = \int c(x, y)d\gamma(x, y) \\ \big(\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi\big) &= \int \varphi(x) + \varphi^{c+}(y)d\gamma(x, y) = \int \varphi d\mu + \int \varphi^{c+} d\nu. \end{aligned}$$

Thus the inequality must be an equality, which is true if and only if for γ'-a.e. (x, y) it holds $(x, y) \in \partial^{c+}\varphi$, hence, by the continuity of c, we conclude $\mathrm{supp}(\gamma') \subset \partial^{c+}\varphi$. ■

2.3 The Dual Problem

The transport problem in the Kantorovich formulation is the problem of minimizing the linear functional $\gamma \mapsto \int c\,d\gamma$ with the affine constraints $\pi^X_\# \gamma = \mu$, $\pi^Y_\# \gamma = \nu$ and $\gamma \geq 0$. It is well known that problems of this kind admit a natural dual problem, where we maximize a linear functional with affine constraints. In our case the dual problem is:

Problem 2.16 (Dual problem). Let $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$. Maximize the value of

$$\int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y),$$

among all functions $\varphi \in L^1(\mu)$, $\psi \in L^1(\nu)$ such that

$$\varphi(x) + \psi(y) \leq c(x, y), \qquad \forall x \in X,\ y \in Y. \tag{5}$$

■

The relation between the transport problem and the dual one consists in the fact that

$$\inf_{\gamma \in \mathcal{A}dm(\mu,\nu)} \int c(x, y) d\gamma(x, y) = \sup_{\varphi,\psi} \int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y),$$

where the supremum is taken among all φ, ψ as in the definition of the problem.

Although the fact that equality holds is an easy consequence of Theorem 2.13 of the previous section (taking $\psi = \varphi^{c+}$, as we will see), we prefer to start with an heuristic argument which shows "why" duality works. The calculations we are going to do are very common in linear programming and are based on the *min-max principle*. Observe how the constraint $\gamma \in \mathcal{A}dm(\mu, \nu)$ "becomes" the functional to maximize in the dual problem and the functional to minimize $\int c\,d\gamma$ "becomes" the constraint in the dual problem.

Start observing that

$$\inf_{\gamma \in \mathcal{A}dm(\mu,\nu)} \int c(x, y) d\gamma(x, y) = \inf_{\gamma \in \mathscr{M}_+(X \times Y)} \int c(x, y) d\gamma + \chi(\gamma), \tag{6}$$

where $\chi(\gamma)$ is equal to 0 if $\gamma \in \mathcal{A}dm(\mu, \nu)$ and $+\infty$ if $\gamma \notin \mathcal{A}dm(\mu, \nu)$, and $\mathscr{M}_+(X \times Y)$ is the set of non negative Borel measures on $X \times Y$. We claim that the function χ may be written as

$$\chi(\gamma) = \sup_{\varphi,\psi} \left\{ \int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y) - \int \varphi(x) + \psi(y) d\gamma(x, y) \right\},$$

where the supremum is taken among all $(\varphi,\psi)\in C_b(X)\times C_b(Y)$. Indeed, if $\gamma\in\mathcal{A}dm(\mu,\nu)$ then $\chi(\gamma)=0$, while if $\gamma\notin\mathcal{A}dm(\mu,\nu)$ we can find $(\varphi,\psi)\in C_b(X)\times C_b(Y)$ such that the value between the brackets is different from 0, thus by multiplying (φ,ψ) by appropriate real numbers we have that the supremum is $+\infty$. Thus from (6) we have

$$\begin{aligned}\inf_{\gamma\in\mathcal{A}dm(\mu,\nu)}\int c(x,y)d\gamma(x,y) = &\inf_{\gamma\in\mathscr{M}_+(X\times Y)}\sup_{\varphi,\psi}\bigg\{\int c(x,y)d\gamma(x,y)\\ &+\int\varphi(x)d\mu(x)+\int\psi(y)d\nu(y)-\int\varphi(x)+\psi(y)d\gamma(x,y)\bigg\}\end{aligned}$$

Call the expression between brackets $F(\gamma,\varphi,\psi)$. Since $\gamma\mapsto F(\gamma,\varphi,\psi)$ is convex (actually linear) and $(\varphi,\psi)\mapsto F(\gamma,\varphi,\psi)$ is concave (actually linear), the min-max principle holds and we have

$$\inf_{\gamma\in\mathcal{A}dm(\mu,\nu)}\sup_{\varphi,\psi}F(\gamma,\varphi,\psi)=\sup_{\varphi,\psi}\inf_{\gamma\in\mathscr{M}_+(X\times Y)}F(\gamma,\varphi,\psi).$$

Thus we have

$$\begin{aligned}\inf_{\gamma\in\mathcal{A}dm(\mu,\nu)}\int c(x,y)d\gamma(x,y) &= \sup_{\varphi,\psi}\inf_{\gamma\in\mathscr{M}_+(X\times Y)}\bigg\{\int c(x,y)d\gamma(x,y)\\ &\quad+\int\varphi(x)d\mu(x)+\int\psi(y)d\nu(y)-\int\varphi(x)+\psi(y)d\gamma(x,y)\bigg\}\\ &=\sup_{\varphi,\psi}\bigg\{\int\varphi(x)d\mu(x)+\int\psi(y)d\nu(y)\\ &\quad+\inf_{\gamma\in\mathscr{M}_+(X\times Y)}\bigg[\int c(x,y)-\varphi(x)-\psi(y)d\gamma(x,y)\bigg]\bigg\}.\end{aligned}$$

Now observe the quantity

$$\inf_{\gamma\in\mathscr{M}_+(X\times Y)}\bigg[\int c(x,y)-\varphi(x)-\psi(y)d\gamma(x,y)\bigg].$$

If $\varphi(x)+\psi(y)\le c(x,y)$ for any (x,y), then the integrand is non-negative and the infimum is 0 (achieved when γ is the null-measure). Conversely, if $\varphi(x)+\psi(y)>c(x,y)$ for some $(x,y)\in X\times Y$, then choose $\gamma:=n\delta_{(x,y)}$ with n large to get that the infimum is $-\infty$.

Thus, we proved that

$$\inf_{\gamma\in\mathcal{A}dm(\mu,\nu)}\int c(x,y)d\gamma(x,y)=\sup_{\varphi,\psi}\int\varphi(x)d\mu(x)+\int\psi(y)d\nu(y),$$

where the supremum is taken among continuous and bounded functions (φ,ψ) satisfying (5).

We now give the rigorous statement and a proof independent of the min-max principle.

Theorem 2.17 (Duality). *Let $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$ and $c : X \times Y \to \mathbb{R}$ a continuous and bounded from below cost function. Assume that (4) holds. Then the minimum of the Kantorovich problem 2.2 is equal to the supremum of the dual problem 2.16.*

Furthermore, the supremum of the dual problem is attained, and the maximizing couple (φ, ψ) is of the form (φ, φ^{c+}) for some c-concave function φ.

Proof. Let $\gamma \in \mathcal{Adm}(\mu, \nu)$ and observe that for any couple of functions $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$ satisfying (5) it holds

$$\int c(x,y)d\gamma(x,y) \geq \int \varphi(x) + \psi(y)d\gamma(x,y) = \int \varphi(x)d\mu(x) + \int \psi(y)d\nu(y).$$

This shows that the minimum of the Kantorovich problem is $\geq$ than the supremum of the dual problem.

To prove the converse inequality pick $\gamma \in \mathcal{Opt}(\mu, \nu)$ and use Theorem 2.13 to find a c-concave function φ such that $\mathrm{supp}(\gamma) \subset \partial^{c+}\varphi$, $\max\{\varphi, 0\} \in L^1(\mu)$ and $\max\{\varphi^{c+}, 0\} \in L^1(\nu)$. Then, as in the proof of $(iii) \Rightarrow (i)$ of Theorem 2.13, we have

$$\int c(x,y)\,d\gamma(x,y) = \int \varphi(x) + \varphi^{c+}(y)\,d\gamma(x,y) = \int \varphi(x)\,d\mu(x) + \int \varphi^{c+}(y)\,d\nu(y),$$

and $\int c d\gamma \in \mathbb{R}$. Thus $\varphi \in L^1(\mu)$ and $\varphi^{c+} \in L^1(\nu)$, which shows that (φ, φ^{c+}) is an admissible couple in the dual problem and gives the thesis. □

Remark 2.18. Notice that a statement stronger than the one of Remark 2.15 holds, namely: under the assumptions of Theorems 2.13 and 2.17, for any c-concave couple of functions (φ, φ^{c+}) maximizing the dual problem and any optimal plan γ it holds

$$\mathrm{supp}(\gamma) \subset \partial^{c+}\varphi.$$

Indeed we already know that for some c-concave φ we have $\varphi \in L^1(\mu)$, $\varphi^{c+} \in L^1(\nu)$ and

$$\mathrm{supp}(\gamma) \subset \partial^{c+}\varphi,$$

for any optimal γ. Now pick another maximizing couple $(\tilde{\varphi}, \tilde{\psi})$ for the dual problem 2.16 and notice that $\tilde{\varphi}(x) + \tilde{\psi}(y) \leq c(x,y)$ for any x, y implies $\tilde{\psi} \leq \tilde{\varphi}^{c+}$, and therefore $(\tilde{\varphi}, \tilde{\varphi}^{c+})$ is a maximizing couple as well. The fact that $\tilde{\varphi}^{c+} \in L^1(\nu)$ follows as in the proof of Theorem 2.17. Conclude noticing that for any optimal plan γ it holds

$$\begin{aligned}\int \tilde{\varphi} d\mu + \int \tilde{\varphi}^{c+} d\nu &= \int \varphi d\mu + \int \varphi^{c+} d\nu = \int \varphi(x) + \varphi^{c+}(y) d\gamma(x,y)\\ &= \int c(x,y) d\gamma \geq \int \tilde{\varphi} d\mu + \int \tilde{\varphi}^{c+} d\nu,\end{aligned}$$

so that the inequality must be an equality. ■

Definition 2.19 (Kantorovich potential). A c-concave function φ such that (φ, φ^{c+}) is a maximizing pair for the dual problem 2.16 is called a c-concave Kantorovich potential, or simply Kantorovich potential, for the couple μ, ν. A c-convex function φ is called c-convex Kantorovich potential if $-\varphi$ is a c-concave Kantorovich potential.

Observe that c-concave Kantorovich potentials are related to the transport problem in the following two different (but clearly related) ways:

- As c-concave functions whose superdifferential contains the support of optimal plans, according to Theorem 2.13.
- As maximizing functions, together with their c_+-transforms, in the dual problem.

2.4 Existence of Optimal Maps

The problem of existence of optimal transport maps consists in looking for optimal plan γ which are induced by a map $T : X \to Y$, i.e. plans γ which are equal to $(Id, T)_\# \mu$, for $\mu := \pi^X_\# \gamma$ and some measurable map T. As we discussed in the first section, in general this problem has no answer, as it may very well be the case when, for given $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$, there is no transport map at all from μ to ν. Still, since we know that (2) holds when μ has no atom, it is possible that under some additional assumptions on the starting measure μ and on the cost function c, optimal transport maps exist.

To formulate the question differently: given μ, ν and the cost function c, is that true that at least one optimal plan γ is induced by a map?

Let us start observing that thanks to Theorem 2.13, the answer to this question relies in a natural way on the analysis of the properties of c-monotone sets, to see how far are they from being graphs. Indeed:

Lemma 2.20. *Let $\gamma \in \mathcal{A}dm(\mu, \nu)$. Then γ is induced by a map if and only if there exists a γ-measurable set $\Gamma \subset X \times Y$ where γ is concentrated, such that for μ-a.e. x there exists only one $y = T(x) \in Y$ such that $(x, y) \in \Gamma$. In this case γ is induced by the map T.*

Proof. The *if* part is obvious. For the *only if*, let Γ be as in the statement of the lemma. Possibly removing from Γ a product $N \times Y$, with N μ-negligible, we can assume that Γ is a graph, and denote by T the corresponding map. By the inner regularity of measures, it is easily seen that we can also assume $\Gamma = \cup_n \Gamma_n$ to be σ-compact. Under this assumption the domain of T (i.e. the projection of Γ on X) is σ-compact, hence Borel, and the restriction of T to the compact set $\pi_X(\Gamma_n)$ is continuous. It follows that T is a Borel map. Since $y = T(x)$ γ-a.e. in $X \times Y$ we conclude that

$$\int \phi(x,y)\,d\gamma(x,y) = \int \phi(x,T(x))d\gamma(x,y) = \int \phi(x,T(x))d\mu(x),$$

so that $\gamma = (Id \times T)_{\#}\mu$. □

Thus the point is the following. We know by Theorem 2.13 that optimal plans are concentrated on c-cyclically monotone sets, still from Theorem 2.13 we know that c-cyclically monotone sets are obtained by taking the c-superdifferential of a c-concave function. Hence from the lemma above what we need to understand is "how often" the c-superdifferential of a c-concave function is single valued.

There is no general answer to this question, but many particular cases can be studied. Here we focus on two special and very important situations:

- $X = Y = \mathbb{R}^d$ and $c(x,y) = |x-y|^2/2$.
- $X = Y = M$, where M is a Riemannian manifold, and $c(x,y) = d^2(x,y)/2$, d being the Riemannian distance.

Let us start with the case $X = Y = \mathbb{R}^d$ and $c(x,y) = |x-y|^2/2$. In this case there is a simple characterization of c-concavity and c-superdifferential:

Proposition 2.21. *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$. Then φ is c-concave if and only if $x \mapsto \overline{\varphi}(x) := |x|^2/2 - \varphi(x)$ is convex and lower semicontinuous. In this case $y \in \partial^{c+}\varphi(x)$ if and only if $y \in \partial^-\overline{\varphi}(x)$.*

Proof. Observe that

$$\begin{aligned}
\varphi(x) = \inf_y \frac{|x-y|^2}{2} - \psi(y) &\Leftrightarrow \varphi(x) = \inf_y \frac{|x|^2}{2} + \langle x, -y\rangle + \frac{|y|^2}{2} - \psi(y)\\
&\Leftrightarrow \varphi(x) - \frac{|x|^2}{2} = \inf_y \langle x, -y\rangle + \left(\frac{|y|^2}{2} - \psi(y)\right)\\
&\Leftrightarrow \overline{\varphi}(x) = \sup_y \langle x, y\rangle - \left(\frac{|y|^2}{2} - \psi(y)\right),
\end{aligned}$$

which proves the first claim. For the second observe that

$$\begin{aligned}
y \in \partial^{c+}\varphi(x) &\Leftrightarrow \begin{cases} \varphi(x) = |x-y|^2/2 - \varphi^{c+}(y), \\ \varphi(z) \le |z-y|^2/2 - \varphi^{c+}(y), \qquad \forall z \in \mathbb{R}^d \end{cases}\\
&\Leftrightarrow \begin{cases} \varphi(x) - |x|^2/2 = \langle x, -y\rangle + |y|^2/2 - \varphi^{c+}(y), \\ \varphi(z) - |z|^2/2 \le \langle z, -y\rangle + |y|^2/2 - \varphi^{c+}(y), \qquad \forall z \in \mathbb{R}^d \end{cases}\\
&\Leftrightarrow \varphi(z) - |z|^2/2 \le \varphi(x) - |x|^2/2 + \langle z - x, -y\rangle \qquad \forall z \in \mathbb{R}^d\\
&\Leftrightarrow -y \in \partial^+(\varphi - |\cdot|^2/2)(x)\\
&\Leftrightarrow y \in \partial^-\overline{\varphi}(x)
\end{aligned}$$

□

Therefore in this situation being concentrated on the c-superdifferential of a c-concave map means being concentrated on the (graph of) the subdifferential of a convex function.

Remark 2.22 (Perturbations of the identity via smooth gradients are optimal). An immediate consequence of the above proposition is the fact that if $\psi \in C_c^\infty(\mathbb{R}^d)$, then there exists $\bar{\varepsilon} > 0$ such that $Id + \varepsilon\nabla\psi$ is an optimal map for any $|\varepsilon| \leq \bar{\varepsilon}$. Indeed, it is sufficient to take $\bar{\varepsilon}$ such that $-Id \leq \bar{\varepsilon}\nabla^2\psi \leq Id$. With this choice, the map $x \mapsto |x|^2/2 + \varepsilon\psi(x)$ is convex for any $|\varepsilon| \leq \bar{\varepsilon}$, and thus its gradient is an optimal map. ■

Proposition 2.21 reduced the problem of understanding when there exists optimal maps reduces to the problem of convex analysis of understanding how the set of non differentiability points of a convex function is made. This latter problem has a known answer; in order to state it, we need the following definition:

Definition 2.23 ($c-c$ hypersurfaces). A set $E \subset \mathbb{R}^d$ is called $c-c$ hypersurface[1] if, in a suitable system of coordinates, it is the graph of the difference of two real valued convex functions, i.e. if there exists convex functions $f, g : \mathbb{R}^{d-1} \to \mathbb{R}$ such that

$$E = \Big\{(y,t) \in \mathbb{R}^d \ : \ y \in \mathbb{R}^{d-1},\ t \in \mathbb{R},\ t = f(y) - g(y)\Big\}.$$

Then it holds the following theorem, which we state without proof:

Theorem 2.24 (Structure of sets of non differentiability of convex functions). *Let $A \subset \mathbb{R}^d$. Then there exists a convex function $\overline{\varphi} : \mathbb{R}^d \to \mathbb{R}$ such that A is contained in the set of points of non differentiability of $\overline{\varphi}$ if and only if A can be covered by countably many $c-c$ hypersurfaces.*

We give the following definition:

Definition 2.25 (Regular measures on $\mathbb{R}^d$). A measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ is called *regular* provided $\mu(E) = 0$ for any $c-c$ hypersurface $E \subset \mathbb{R}^d$.

Observe that absolutely continuous measures and measures which give 0 mass to Lipschitz hypersurfaces are automatically regular (because convex functions are locally Lipschitz, thus a $c-c$ hypersurface is a locally Lipschitz hypersurface).

Now we can state the result concerning existence and uniqueness of optimal maps:

Theorem 2.26 (Brenier). *Let $\mu \in \mathscr{P}(\mathbb{R}^d)$ be such that $\int |x|^2 d\mu(x)$ is finite. Then the following are equivalent:*

(i) For every $\nu \in \mathscr{P}(\mathbb{R}^d)$ with $\int |x|^2 d\nu(x) < \infty$ there exists only one transport plan from μ to ν and this plan is induced by a map T.

[1] Here $c-c$ stands for "convex minus convex" and has nothing to do with the c we used to indicate the cost function.

(ii) μ *is regular.*

If either (i) *or* (ii) *hold, the optimal map* T *can be recovered by taking the gradient of a convex function.*

Proof. **(ii)** $\Rightarrow$ **(i) and the last statement**. Take $a(x) = b(x) = |x|^2$ in the statement of Theorem 2.13. Then our assumptions on μ, ν guarantees that the bound (4) holds. Thus the conclusions of Theorems 2.13 and 2.17 are true as well. Using Remark 2.18 we know that for any c-concave Kantorovich potential φ and any optimal plan $\gamma \in \mathit{Opt}(\mu,\nu)$ it holds $\mathrm{supp}(\gamma) \subset \partial^{c+}\varphi$. Now from Proposition 2.21 we know that $\overline{\varphi} := |\cdot|^2/2 - \varphi$ is convex and that $\partial^c\varphi = \partial^-\overline{\varphi}$. Here we use our assumption on μ: since $\overline{\varphi}$ is convex, we know that the set E of points of non differentiability of $\overline{\varphi}$ is μ-negligible. Therefore the map $\nabla\overline{\varphi} : \mathbb{R}^d \to \mathbb{R}^d$ is well defined μ-a.e. and every optimal plan must be concentrated on its graph. Hence the optimal plan is unique and induced by the gradient of the convex function $\overline{\varphi}$.

(ii) $\Rightarrow$ **(i)**. We argue by contradiction and assume that there is some convex function $\overline{\varphi} : \mathbb{R}^d \to \mathbb{R}$ such that the set E of points of non differentiability of $\overline{\varphi}$ has positive μ measure. Possibly modifying $\overline{\varphi}$ outside a compact set, we can assume that it has linear growth at infinity. Now define the two maps:

$$T(x) := \text{the element of smallest norm in } \partial^-\overline{\varphi}(x),$$
$$S(x) := \text{the element of biggest norm in } \partial^-\overline{\varphi}(x),$$

and the plan

$$\gamma := \frac{1}{2}\big((Id,T)_\#\mu + (Id,S)_\#\mu\big).$$

The fact that $\overline{\varphi}$ has linear growth, implies that $\nu := \pi^Y_\#\gamma$ has compact support. Thus in particular $\int |x|^2 d\nu(x) < \infty$. The contradiction comes from the fact that $\gamma \in \mathcal{A}dm(\mu,\nu)$ is c-cyclically monotone (because of Proposition 2.21), and thus optimal. However, it is not induced by a map, because $T \neq S$ on a set of positive μ measure (Lemma 2.20). □

The question of *regularity* of the optimal map is very delicate. In general it is only of bounded variation (BV in short), since monotone maps always have this regularity property, and discontinuities can occur: just think to the case in which the support of the starting measure is connected, while the one of the arrival measure is not. It turns out that connectedness is not sufficient to prevent discontinuities, and that if we want some regularity, we have to impose a convexity restriction on $\mathrm{supp}\,\nu$. The following result holds:

Theorem 2.27 (Regularity theorem). *Assume* $\Omega_1, \Omega_2 \subset \mathbb{R}^d$ *are two bounded and connected open sets,* $\mu = \rho\mathscr{L}^d|_{\Omega_1}$, $\nu = \eta\mathscr{L}^d|_{\Omega_2}$ *with* $0 < c \le \rho, \eta \le C$ *for some* c, $C \in \mathbb{R}$. *Assume also that* Ω_2 *is convex. Then the optimal transport map* T *belongs to* $C^{0,\alpha}(\Omega_1)$ *for some* $\alpha < 1$. *In addition, the following implication holds:*

$$\rho \in C^{0,\alpha}(\Omega_1),\ \ \eta \in C^{0,\alpha}(\Omega_2) \qquad \Longrightarrow \qquad T \in C^{1,\alpha}(\Omega_1).$$

The convexity assumption on Ω_2 is needed to show that the convex function φ whose gradient provides the optimal map T is a *viscosity* solution of the Monge–Ampere equation

$$\rho^1(x) = \rho^2(\nabla\varphi(x))\det(\nabla^2\varphi(x)),$$

and then the regularity theory for Monge–Ampere, developed by Caffarelli and Urbas, applies.

As an application of Theorem 2.26 we discuss the question of *polar factorization* of vector fields on $\mathbb{R}^d$. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, denote by μ_Ω the normalized Lebesgue measure on Ω and consider the space

$$S(\Omega) := \{\text{Borel map } s : \Omega \to \Omega \ : \ s_\#\mu_\Omega = \mu_\Omega\}.$$

The following result provides a (nonlinear) projection on the (nonconvex) space $S(\Omega)$.

Proposition 2.28 (Polar factorization). *Let $S \in L^2(\mu_\Omega; \mathbb{R}^n)$ be such that $\nu := S_\#\mu$ is regular (Definition 2.25). Then there exist unique $s \in S(\Omega)$ and $\nabla\varphi$, with φ convex, such that $S = (\nabla\varphi)\circ s$. Also, s is the unique minimizer of*

$$\int |S - \tilde{s}|^2 d\mu,$$

among all $\tilde{s} \in S(\Omega)$.

Proof. By assumption, we know that both μ_Ω and ν are regular measures with finite second moment. We claim that

$$\inf_{\tilde{s}\in S(\Omega)} \int |S - \tilde{s}|^2 d\mu = \min_{\gamma\in\mathcal{A}dm(\mu,\nu)} \int |x-y|^2 d\gamma(x,y). \tag{7}$$

To see why, associate to each $\tilde{s} \in S(\Omega)$ the plan $\gamma_{\tilde{s}} := (\tilde{s}, S)_\#\mu$ which clearly belongs to $\mathcal{A}dm(\mu_\Omega, \nu)$. This gives inequality $\geq$. Now let $\overline{\gamma}$ be the unique optimal plan and apply Theorem 2.26 twice to get that

$$\overline{\gamma} = (Id, \nabla\varphi)_\#\mu_\Omega = (\nabla\tilde{\varphi}, Id)_\#\nu,$$

for appropriate convex functions $\varphi, \tilde{\varphi}$, which therefore satisfy $\nabla\varphi\circ\nabla\tilde{\varphi} = Id$ μ-a.e.. Define $s := \nabla\tilde{\varphi}\circ S$. Then $s_\#\mu_\Omega = \mu_\Omega$ and thus $s \in S(\Omega)$. Also, $S = \nabla\varphi\circ s$ which proves the existence of the polar factorization. The identity

$$\int |x-y|^2 d\gamma_s(x,y) = \int |s - S|^2 d\mu_\Omega = \int |\nabla\tilde{\varphi}\circ S - S|^2 d\mu_\Omega = \int |\nabla\tilde{\varphi} - Id|^2 d\nu$$
$$= \min_{\gamma\in\mathcal{A}dm(\mu,\nu)} \int |x-y|^2 d\gamma(x,y),$$

shows inequality $\leq$ in (7) and the uniqueness of the optimal plan ensures that s is the unique minimizer.

To conclude we need to show uniqueness of the polar factorization. Assume that $S = (\nabla\overline{\varphi}) \circ \overline{s}$ is another factorization and notice that $\nabla\overline{\varphi}_{\#}\mu_\Omega = (\nabla\overline{\varphi} \circ \overline{s})_{\#}\mu_\Omega = \nu$. Thus the map $\nabla\overline{\varphi}$ is a transport map from μ_Ω to ν and is the gradient of a convex function. By Proposition 2.21 and Theorem 2.13 we deduce that $\nabla\overline{\varphi}$ is the optimal map. Hence $\nabla\overline{\varphi} = \nabla\varphi$ and the proof is achieved. □

Remark 2.29 (Polar factorization vs. Helmholtz decomposition). The classical Helmholtz decomposition of vector fields can be seen as a linearized version of the polar factorization result, which therefore can be though as a generalization of the former.

To see why, assume that Ω and all the objects considered are smooth (the arguments hereafter are just formal). Let $u : \Omega \to \mathbb{R}^d$ be a vector field and apply the polar factorization to the map $S_\varepsilon := Id + \varepsilon u$ with $|\varepsilon|$ small. Then we have $S_\varepsilon = (\nabla\varphi_\varepsilon) \circ s_\varepsilon$ and both $\nabla\varphi_\varepsilon$ and s_ε will be perturbation of the identity, so that

$$\nabla\varphi_\varepsilon = Id + \varepsilon v + o(\varepsilon),$$
$$s_\varepsilon = Id + \varepsilon w + o(\varepsilon).$$

The question now is: which information is carried on v, w from the properties of the polar factorization? At the level of v, from the fact that $\nabla \times (\nabla\varphi_\varepsilon) = 0$ we deduce $\nabla \times v = 0$, which means that v is the gradient of some function p. On the other hand, the fact that s_ε is measure preserving implies that w satisfies $\nabla \cdot (w\chi_\Omega) = 0$ in the sense of distributions: indeed for any smooth $f : \mathbb{R}^d \to \mathbb{R}$ it holds

$$0 = \frac{d}{d\varepsilon}|_{\varepsilon=0} \int f \, d(s_\varepsilon)_{\#}\mu_\Omega = \frac{d}{d\varepsilon}|_{\varepsilon=0} \int f \circ s_\varepsilon \, d\mu_\Omega = \int \nabla f \cdot w \, d\mu_\Omega.$$

Then from the identity $(\nabla\varphi_\varepsilon) \circ s_\varepsilon = Id + \varepsilon(\nabla p + w) + o(\varepsilon)$ we can conclude that

$$u = \nabla p + w.$$

■

We now turn to the case $X = Y = M$, with M smooth Riemannian manifold, and $c(x, y) = d^2(x, y)/2$, d being the Riemannian distance on M. For simplicity, we will assume that M is compact and with no boundary, but everything holds in more general situations.

The underlying ideas of the foregoing discussion are very similar to the ones of the case $X = Y = \mathbb{R}^d$, the main difference being that there is no more the correspondence given by Proposition 2.21 between c-concave functions and convex functions, as in the Euclidean case. Recall however that the concepts of semiconvexity (i.e. second derivatives bounded from below) and semiconcavity make sense also on manifolds, since these properties can be read locally and changes of coordinates are smooth.

In the next proposition we will use the fact that on a compact and smooth Riemannian manifold, the functions $x \mapsto d^2(x,y)$ are uniformly Lipschitz and uniformly semiconcave in $y \in M$ (i.e. the second derivative along a unit speed geodesic is bounded above by a universal constant depending only on M, see e.g. the third appendix of Chap. 10 of [80] for the simple proof).

Proposition 2.30. *Let M be a smooth, compact Riemannian manifold without boundary. Let $\varphi : M \to \mathbb{R} \cup \{-\infty\}$ be a c-concave function not identically equal to $-\infty$. Then φ is Lipschitz, semiconcave and real valued. Also, assume that $y \in \partial^{c+}\varphi(x)$. Then* $\exp_x^{-1}(y) \subset -\partial^+\varphi(x)$.
Conversely, if φ is differentiable at x, then $\exp_x(-\nabla\varphi(x)) \in \partial^{c+}\varphi(x)$.

Proof. The fact that φ is real valued follows from the fact that the cost function $d^2(x,y)/2$ is uniformly bounded in $x, y \in M$. Smoothness and compactness ensure that the functions $d^2(\cdot,y)/2$ are uniformly Lipschitz and uniformly semiconcave in $y \in M$, this gives that φ is Lipschitz and semiconcave.

Now pick $y \in \partial^{c+}\varphi(x)$ and $\mathrm{v} \in \exp_x^{-1}(y)$. Recall that $-\mathrm{v}$ belongs to the superdifferential of $d^2(\cdot,y)/2$ at x, i.e.

$$\frac{d^2(z,y)}{2} \le \frac{d^2(x,y)}{2} - \left\langle \mathrm{v}, \exp_x^{-1}(z)\right\rangle + o(d(x,z)).$$

Thus from $y \in \partial^{c+}\varphi(x)$ we have

$$\varphi(z) - \varphi(x) \overset{(3)}{\le} \frac{d^2(z,y)}{2} - \frac{d^2(x,y)}{2} \le \left\langle -\mathrm{v}, \exp_x^{-1}(z)\right\rangle + o(d(x,z)),$$

that is $-\mathrm{v} \in \partial^+\varphi(x)$.

To prove the converse implication, it is enough to show that the c-superdifferential of φ at x is non empty. To prove this, use the c-concavity of φ to find a sequence $(y_n) \subset M$ such that

$$\begin{aligned}\varphi(x) &= \lim_{n\to\infty} \frac{d^2(x,y_n)}{2} - \varphi^{c+}(y_n),\\ \varphi(z) &\le \frac{d^2(z,y_n)}{2} - \varphi^{c+}(y_n), \qquad \forall z \in M,\ n \in \mathbb{N}.\end{aligned}$$

By compactness we can extract a subsequence converging to some $y \in M$. Then from the continuity of $d^2(z,\cdot)/2$ and $\varphi^{c+}(\cdot)$ it is immediate to verify that $y \in \partial^{c+}\varphi(x)$. □

Remark 2.31. The converse implication in the previous proposition is *false* if one doesn't assume φ to be differentiable at x: i.e., it is *not* true in general that $\exp_x(-\partial^+\varphi(x)) \subset \partial^{c+}\varphi(x)$. ■

From this proposition, and following the same ideas used in the Euclidean case, we give the following definition:

Definition 2.32 (Regular measures in $\mathscr{P}(M)$). We say that $\mu \in \mathscr{P}(M)$ is regular provided it vanishes on the set of points of non differentiability of ψ for any semiconvex function $\psi : M \to \mathbb{R}$.

The set of points of non differentiability of a semiconvex function on M can be described as in the Euclidean case by using local coordinates. For most applications it is sufficient to keep in mind that absolutely continuous measures (w.r.t. the volume measure) and even measures vanishing on Lipschitz hypersurfaces are regular.

By Proposition 2.30, we can derive a result about existence and characterization of optimal transport maps in manifolds which closely resembles Theorem 2.26:

Theorem 2.33 (McCann). *Let M be a smooth, compact Riemannian manifold without boundary and $\mu \in \mathscr{P}(M)$. Then the following are equivalent:*

(i) For every $\nu \in \mathscr{P}(M)$ there exists only one transport plan from μ to ν and this plan is induced by a map T.
(ii) μ is regular.

If either (i) or (ii) hold, the optimal map T can be written as $x \mapsto \exp_x(-\nabla\varphi(x))$ for some c-concave function $\varphi : M \to \mathbb{R}$.

Proof. **(ii) ⇒ (i) and the last statement**. Pick $\nu \in \mathscr{P}(M)$ and observe that, since $d^2(\cdot,\cdot)/2$ is uniformly bounded, condition (4) surely holds. Thus from Theorem 2.13 and Remark 2.15 we get that any optimal plan $\gamma \in Opt(\mu,\nu)$ must be concentrated on the c-superdifferential of a c-concave function φ. By Proposition 2.30 we know that φ is semiconcave, and thus differentiable μ-a.e. by our assumption on μ. Therefore $x \mapsto T(x) := \exp_x(-\nabla\varphi(x))$ is well defined μ-a.e. and its graph must be of full γ-measure for any $\gamma \in Opt(\mu,\nu)$. This means that γ is unique and induced by T.

(i) ⇒ (ii). Argue by contradiction and assume that there exists a semiconcave function f whose set of points of non differentiability has positive μ measure. Use Lemma 2.34 below to find $\varepsilon > 0$ such that $\varphi := \varepsilon f$ is c-concave and satisfies: $\mathrm{v} \in \partial^+\varphi(x)$ if and only $\exp_x(-\mathrm{v}) \in \partial^{c+}\varphi(x)$. Then conclude the proof as in Theorem 2.26. □

Lemma 2.34. *Let M be a smooth, compact Riemannian manifold without boundary and $\varphi : M \to \mathbb{R}$ semiconcave. Then for $\varepsilon > 0$ sufficiently small the function $\varepsilon\varphi$ is c-concave and it holds $\mathrm{v} \in \partial^+(\varepsilon\varphi)(x)$ if and only $\exp_x(-\mathrm{v}) \in \partial^{c+}(\varepsilon\varphi)(x)$.*

Proof. We start with the following claim: there exists $\varepsilon > 0$ such that for every $x_0 \in M$ and every $\mathrm{v} \in \partial^+\varphi(x_0)$ the function

$$x \mapsto \varepsilon\varphi(x) - \frac{d^2(x, \exp_{x_0}(-\varepsilon\mathrm{v}))}{2}$$

has a global maximum at $x = x_0$.

Use the smoothness and compactness of M to find $r > 0$ such that $d^2(\cdot,\cdot)/2 : \{(x,y) : d(x,y) < r\} \to \mathbb{R}$ is C^∞ and satisfies $\nabla^2 d^2(\cdot,y)/2 \geq cId$, for every $y \in M$, with $c > 0$ independent on y. Now observe that since φ is semiconcave and real valued, it is Lipschitz. Thus, for $\varepsilon_0 > 0$ sufficiently small it holds $\varepsilon_0|\mathrm{v}| < r/3$ for any $\mathrm{v} \in \partial^+\varphi(x)$ and any $x \in M$. Also, since φ is bounded, possibly decreasing the value of ε_0 we can assume that

$$\varepsilon_0|\varphi(x)| \le \frac{r^2}{12}.$$

Fix $x_0 \in M$, $\mathrm{v} \in \partial^+\varphi(x_0)$ and let $y_0 := \exp_{x_0}(-\varepsilon_0\mathrm{v})$. We claim that for ε_0 chosen as above, the maximum of $\varepsilon_0\varphi - d^2(\cdot, y_0)/2$, cannot lie outside $B_r(x_0)$. Indeed, if $d(x, x_0) \ge r$ we have $d(x, y_0) > 2r/3$ and thus:

$$\varepsilon_0\varphi(x) - \frac{d^2(x, y_0)}{2} < \frac{r^2}{12} - \frac{2r^2}{9} = -\frac{r^2}{12} - \frac{r^2}{18} \le \varepsilon_0\varphi(x_0) - \frac{d^2(x_0, y_0)}{2}.$$

Thus the maximum must lie in $B_r(x_0)$. Recall that in this ball, the function $d^2(\cdot, y_0)$ is C^∞ and satisfies $\nabla^2(d^2(\cdot, y_0)/2) \ge cId$, thus it holds

$$\nabla^2\left(\varepsilon_0\varphi(\cdot) - \frac{d^2(\cdot, y_0)}{2}\right) \le (\varepsilon_0\lambda - c)Id,$$

where $\lambda \in \mathbb{R}$ is such that $\nabla^2\varphi \le \lambda Id$ on the whole of M. Thus decreasing if necessary the value of ε_0 we can assume that

$$\nabla^2\left(\varepsilon_0\varphi(\cdot) - \frac{d^2(\cdot, y_0)}{2}\right) < 0 \qquad \text{on } B_r(x_0),$$

which implies that $\varepsilon_0\varphi(\cdot) - d^2(\cdot, y_0)/2$ admits a unique point $x \in B_r(x_0)$ such that $0 \in \partial^+(\varphi - d^2(\cdot, y_0)/2)(x)$, which therefore is the unique maximum. Since $\nabla\frac{1}{2}d^2(\cdot, y_0)(x_0) = \varepsilon_0\mathrm{v} \in \partial^+(\varepsilon_0\varphi)(x_0)$, we conclude that x_0 is the unique global maximum, as claimed.

Now define the function $\psi : M \to \mathbb{R} \cup \{-\infty\}$ by

$$\psi(y) := \inf_{x \in M} \frac{d^2(x, y)}{2} - \varepsilon_0\varphi(x),$$

if $y = \exp_x(-\varepsilon_0\mathrm{v})$ for some $x \in M$, $\mathrm{v} \in \partial^+\varphi(x)$, and $\psi(y) := -\infty$ otherwise. By definition we have

$$\varepsilon_0\varphi(x) \le \frac{d^2(x, y)}{2} - \psi(y), \qquad \forall x, y \in M,$$

and the claim proved ensures that if $y_0 = \exp_{x_0}(-\varepsilon_0\mathrm{v}_0)$ for $x_0 \in M$, $\mathrm{v}_0 \in \partial^+\varphi(x_0)$ the inf in the definition of $\psi(y_0)$ is realized at $x = x_0$ and thus

$$\varepsilon_0\varphi(x_0) = \frac{d^2(x_0, y_0)}{2} - \psi(y_0).$$

Hence $\varepsilon_0\varphi = \psi^{c+}$ and therefore is c-concave. Along the same lines one can easily see that for $y \in \exp_x(-\varepsilon_0\partial^+\varphi(x))$ it holds

$$\varepsilon_0\varphi(x) = \frac{d^2(x,y)}{2} - (\varepsilon_0\varphi)^{c+}(y),$$

i.e. $y \in \partial^{c+}(\varepsilon_0\varphi)(x_0)$. Thus we have $\partial^{c+}(\varepsilon_0\varphi) \supset \exp(-\partial^{+}(\varepsilon\varphi))$. Since the other inclusion has been proved in Proposition 2.30 the proof is finished. □

Remark 2.35. With the same notation of Theorem 2.33, recall that we know that the c-concave function φ whose c-superdifferential contains the graph of any optimal plan from μ to ν is differentiable μ-a.e. (for regular μ). Fix x_0 such that $\nabla\varphi(x_0)$ exists, let $y_0 := \exp_{x_0}(-\nabla\varphi(x_0)) \in \partial^{c+}\varphi(x_0)$ and observe that from

$$\frac{d^2(x,y_0)}{2} - \frac{d^2(x_0,y_0)}{2} \geq \varphi(x) - \varphi(x_0),$$

we deduce that $\nabla\varphi(x_0)$ belongs to the *sub*differential of $d^2(\cdot,y_0)/2$ at x_0. Since we know that $d^2(\cdot,y_0)/2$ always have non empty superdifferential, we deduce that it must be differentiable at x_0. In particular, *there exists only one geodesic connecting x_0 to y_0*. Therefore if μ is regular, not only there exists a unique optimal transport map T, but also for μ-a.e. x there is only one geodesic connecting x to $T(x)$. ■

The question of regularity of optimal maps on manifolds is much more delicate than the corresponding question on $\mathbb{R}^d$, even if one wants to get only the continuity. We won't enter into the details of the theory, we just give an example showing the difficulty that can arise in a curved setting. The example will show a smooth compact manifold, and two measures absolutely continuous with positive and smooth densities, such that the optimal transport map is discontinuous. We remark that similar behaviors occur as soon as M has one point and one sectional curvature at that point which is strictly negative. Also, even if one assumes that the manifold has non negative sectional curvature everywhere, this is not enough to guarantee continuity of the optimal map: what comes into play in this setting is the Ma–Trudinger–Wang tensor, an object which we will not study.

Example 2.36. Let $M \subset \mathbb{R}^3$ be a smooth surface which has the following properties:

- M is symmetric w.r.t. the x axis and the y axis.
- M crosses the line $(x,y) = (0,0)$ at two points, namely O, O'.
- The curvature of M at O is negative.

These assumptions ensure that we can find $a, b > 0$ such that for some z_a, z_b the points

$$A := (a, 0, z_a),$$
$$A' := (-a, 0, z_a),$$
$$B := (0, b, z_b),$$
$$B' := (0, -b, z_b),$$

belong to M and

$$d^2(A,B) > d^2(A,O) + d^2(O,B),$$

d being the intrinsic distance on M. By continuity and symmetry, we can find $\varepsilon > 0$ such that

$$d^2(x,y) > d^2(x,O)+d^2(O,y), \qquad \forall x \in B_\varepsilon(A)\cup B_\varepsilon(A'),\ y \in B_\varepsilon(B)\cup B_\varepsilon(B'). \tag{8}$$

Now let f (resp. g) be a smooth probability density everywhere positive and symmetric w.r.t. the x, y axes such that $\int_{B_\varepsilon(A)\cup B_\varepsilon(A')} f\, d\mathrm{vol} > \frac{1}{2}$ (resp. $\int_{B_\varepsilon(B)\cup B_\varepsilon(B')} g\, d\mathrm{vol} > \frac{1}{2}$), and let T (resp. T') be the optimal transport map from $f\mathrm{vol}$ to $g\mathrm{vol}$ (resp. from $g\mathrm{vol}$ to $f\mathrm{vol}$).

We claim that either T or T' is discontinuous and argue by contradiction. Suppose that both are continuous and observe that by the symmetry of the optimal transport problem it must hold $T'(x) = T^{-1}(x)$ for any $x \in M$. Again by the symmetry of M, f, g, the point $T(O)$ must be invariant under the symmetries around the x and y axes. Thus it is either $T(O) = O$ or $T(O) = O'$, and similarly, $T'(O') \in \{O, O'\}$.

We claim that it must hold $T(O) = O$. Indeed otherwise either $T(O) = O'$ and $T(O') = O$, or $T(O) = O'$ and $T(O') = O'$. In the first case the two couples (O, O') and (O', O) belong to the support of the optimal plan, and thus by cyclical monotonicity it holds

$$d^2(O,O') + d^2(O',O) \leq d^2(O,O) + d^2(O',O') = 0,$$

which is absurdum.

In the second case we have $T'(x) \neq O$ for all $x \in M$, which, by continuity and compactness implies $d(T'(M), O) > 0$. This contradicts the fact that f is positive everywhere and $T'_\#(g\mathrm{vol}) = f\mathrm{vol}$.

Thus it holds $T(O) = O$. Now observe that by construction there must be some mass transfer from $B_\varepsilon(A)\cup B_\varepsilon(A')$ to $B_\varepsilon(B)\cup B_\varepsilon(B')$, i.e. we can find $x \in B_\varepsilon(A)\cup B_\varepsilon(A')$ and $y \in B_\varepsilon(B)\cup B_\varepsilon(B')$ such that (x, y) is in the support of the optimal plan. Since (O, O) is the support of the optimal plan as well, by cyclical monotonicity it must hold

$$d^2(x,y) + d^2(O,O) \leq d^2(x,O) + d^2(O,y),$$

which contradicts (8). ■

2.5 Bibliographical Notes

G. Monge's original formulation of the transport problem [66] was concerned with the case $X = Y = \mathbb{R}^d$ and $c(x,y) = |x - y|$, and L.V. Kantorovich's formulation appeared first in [49].

The equality (2), saying that the infimum of the Monge problem equals the minimum of Kantorovich one, has been proved by W. Gangbo (Appendix A of [41]) and the first author (Theorem 2.1 in [4]) in particular cases, and then generalized by A. Pratelli [68].

In [50] L.V. Kantorovich introduced the dual problem, and later L.V. Kantorovich and G.S. Rubinstein [51] further investigated this duality for the case $c(x, y) = d(x, y)$. The fact that the study of the dual problem can lead to important informations for the transport problem has been investigated by several authors, among others M. Knott and C.S. Smith [52] and S.T. Rachev and L. Rüschendorf [69, 71].

The notions of cyclical monotonicity and its relation with subdifferential of convex function have been developed by Rockafellar in [70]. The generalization to c-cyclical monotonicity and to c-sub/super differential of c-convex/concave functions has been studied, among others, by Rüschendorf [71].

The characterization of the set of non differentiability of convex functions is due to Zajíček ([83], see also the paper by G. Alberti [2] and the one by G. Alberti and the first author [3]).

Theorem 2.26 on existence of optimal maps in $\mathbb{R}^d$ for the cost=distance-squared is the celebrated result of Y. Brenier, who also observed that it implies the polar factorization result Proposition 2.28 [18, 19]. Brenier's ideas have been generalized in many directions. One of the most notable one is R. McCann's Theorem 2.33 concerning optimal maps in Riemannian manifolds for the case cost=squared distance [64]. R. McCann also noticed that the original hypothesis in Brenier's theorem, which was $\mu \ll \mathscr{L}^d$, can be relaxed into "μ gives 0 mass to Lipschitz hypersurfaces". In [42] W. Gangbo and R. McCann pointed out that to get existence of optimal maps in $\mathbb{R}^d$ with $c(x, y) = |x-y|^2/2$ it is sufficient to ask to the measure μ to be regular in the sense of the Definition 2.25. The sharp version of Brenier's and McCann's theorems presented here, where the necessity of the regularity of μ is also proved, comes from a paper of the second author of these notes [46].

Other extensions of Brenier's result are:

- Infinite-dimensional Hilbert spaces (the authors and Savaré—[7]).
- Cost functions induced by Lagrangians, Bernard–Buffoni [13], namely

$$c(x, y) := \inf\left\{ \int_0^1 L(t, \gamma(t), \dot\gamma(t))\, dt \,:\, \gamma(0) = x,\ \gamma(1) = y \right\}.$$

- Carnot groups and sub-Riemannian manifolds, $c = d_{CC}^2/2$: the first author and S. Rigot [6], A. Figalli and L. Rifford [39].
- Cost functions induced by sub-Riemannian Lagrangians A. Agrachev and P. Lee [1].
- Wiener spaces (E, H, γ), D. Feyel–A.S. Üstünel [36].

  Here E is a Banach space, $\gamma \in \mathscr{P}(E)$ is Gaussian and H is its Cameron–Martin space, namely

$$H := \left\{ h \in E \,:\, (\tau_h)_\sharp \gamma \ll \gamma \right\}.$$

  In this case

$$c(x, y) := \begin{cases} \dfrac{|x-y|_H^2}{2} & \text{if} x - y \in H; \\ +\infty & \text{otherwise.} \end{cases}$$

The issue of regularity of optimal maps would nowadays require a lecture note in its own. A rough statement that one should have in mind is that it is rare to have regular (even just continuous) optimal transport maps. The key Theorem 2.27 is due to L. Caffarelli [21–23].

Example 2.36 is due to G. Loeper [55]. For the general case of cost=squared distance on a compact Riemannian manifold, it turns out that continuity of optimal maps between two measures with smooth and strictly positive density is strictly related to the positivity of the so-called Ma–Trudinger–Wang tensor [59], an object defined taking fourth order derivatives of the distance function. The understanding of the structure of this tensor has been a very active research area in the last years, with contributions coming from X.-N. Ma, N. Trudinger, X.-J. Wang, C. Villani, P. Delanoe, R. McCann, A. Figalli, L. Rifford, H.-Y. Kim and others.

A topic which we didn't discuss at all is the original formulation of the transport problem of Monge: the case $c(x,y) := |x-y|$ on $\mathbb{R}^d$. The situation in this case is much more complicated than the one with $c(x,y) = |x-y|^2/2$ as it is typically not true that optimal plans are unique, or that optimal plans are induced by maps. For example consider on $\mathbb{R}$ any two probability measures μ, ν such that μ is concentrated on the negative numbers and ν on the positive ones. Then one can see that any admissible plan between them is optimal for the cost $c(x,y) = |x-y|$.

Still, even in this case there is existence of optimal maps, but in order to find them one has to use a sort of selection principle. A successful strategy—which has later been applied to a number of different situation—has been proposed by V.N. Sudakov in [77], who used a disintegration principle to reduce the d-dimensional problem to a problem on $\mathbb{R}$. The original argument by V.N. Sudakov was flawed and has been fixed by the first author in [4] in the case of the Euclidean distance. Meanwhile, different proofs of existence of optimal maps have been proposed by L.C. Evans–W. Gangbo [34], Trudinger and Wang [78], and L. Caffarelli, M. Feldman and R. McCann [24].

Later, existence of optimal maps for the case $c(x,y) := \|x-y\|$, $\|\cdot\|$ being any norm has been established, at increasing levels of generality, in [10, 27, 28] (containing the most general result, for any norm) and [25].

3 The Wasserstein Distance W_2

The aim of this chapter is to describe the properties of the Wasserstein distance W_2 on the space of Borel Probability measures on a given metric space (X,d). This amounts to study the transport problem with cost function $c(x,y) = d^2(x,y)$.

An important characteristic of the Wasserstein distance is that it inherits many interesting geometric properties of the base space (X,d). For this reason we split the foregoing discussion into three sections on which we deal with the cases in which X is: a general Polish space, a geodesic space and a Riemannian manifold.

A word on the notation: when considering product spaces like X^n, with $\pi^i : X^n \to X$ we intend the natural projection onto the i-th coordinate, $i = 1,\dots,n$.

Thus, for instance, for $\mu, \nu \in \mathscr{P}(X)$ and $\gamma \in \mathcal{A}dm(\mu, \nu)$ we have $\pi^1_\# \gamma = \mu$ and $\pi^2_\# \gamma = \nu$. Similarly, with $\pi^{i,j} : X^n \to X^2$ we intend the projection onto the i-th and j-th coordinates. And similarly for multiple projections.

3.1 X Polish Space

Let (X, d) be a complete and separable metric space.

The distance W_2 is defined as

$$\begin{aligned} W_2(\mu, \nu) &:= \sqrt{\inf_{\gamma \in \mathcal{A}dm(\mu,\nu)} \int d^2(x, y) d\gamma(x, y)} \\ &= \sqrt{\int d^2(x, y) d\gamma(x, y)}, \qquad \forall \gamma \in Opt(\mu, \nu). \end{aligned}$$

The natural space to endow with the Wasserstein distance W_2 is the space $\mathscr{P}_2(X)$ of Borel Probability measures with finite second moment:

$$\mathscr{P}_2(X) := \left\{ \mu \in \mathscr{P}(X) \ : \ \int d^2(x, x_0) d\mu(x) < \infty \text{ for some, and thus any, } x_0 \in X \right\}.$$

Notice that if either μ or ν is a Dirac delta, say $\nu = \delta_{x_0}$, then there exists only one plan γ in $\mathcal{A}dm(\mu, \nu)$: the plan $\mu \times \delta_{x_0}$, which therefore is optimal. In particular it holds

$$\int d^2(x, x_0) d\mu(x) = W_2^2(\mu, \delta_{x_0}),$$

that is: the second moment is nothing but the squared Wasserstein distance from the corresponding Dirac mass.

We start proving that W_2 is actually a distance on $\mathscr{P}_2(X)$. In order to prove the triangle inequality, we will use the following lemma, which has its own interest:

Lemma 3.1 (Gluing). *Let X, Y, Z be three Polish spaces and let $\gamma^1 \in \mathscr{P}(X \times Y)$, $\gamma^2 \in \mathscr{P}(Y \times Z)$ be such that $\pi^Y_\# \gamma^1 = \pi^Y_\# \gamma^2$. Then there exists a measure $\gamma \in \mathscr{P}(X \times Y \times Z)$ such that*

$$\begin{aligned} \pi^{X,Y}_\# \gamma &= \gamma^1, \\ \pi^{Y,Z}_\# \gamma &= \gamma^2. \end{aligned}$$

Proof. Let $\mu := \pi^Y_\# \gamma^1 = \pi^Y_\# \gamma^2$ and use the disintegration theorem to write $d\gamma^1(x, y) = d\mu(y) d\gamma^1_y(x)$ and $d\gamma^2(y, z) = d\mu(y) d\gamma^2_y(z)$. Conclude defining γ by

$$d\gamma(x, y, z) := d\mu(y) d(\gamma^1_y \times \gamma^2_y)(x, z).$$

□

Theorem 3.2 (**W_2 is a distance**). *W_2 is a distance on $\mathscr{P}_2(X)$.*

Proof. It is obvious that $W_2(\mu,\mu) = 0$ and that $W_2(\mu,\nu) = W_2(\nu,\mu)$. To prove that $W_2(\mu,\nu) = 0$ implies $\mu = \nu$ just pick an optimal plan $\gamma \in \mathit{Opt}(\mu,\nu)$ and observe that $\int d^2(x,y)d\gamma(x,y) = 0$ implies that γ is concentrated on the diagonal of $X \times X$, which means that the two maps π^1 and π^2 coincide γ-a.e., and therefore $\pi^1_\#\gamma = \pi^2_\#\gamma$.

For the triangle inequality, we use the gluing lemma to "compose" two optimal plans. Let $\mu_1,\ \mu_2,\ \mu_3 \in \mathscr{P}_2(X)$ and let $\gamma_1^2 \in \mathit{Opt}(\mu_1,\mu_2)$, $\gamma_2^3 \in \mathit{Opt}(\mu_2,\mu_3)$. By the gluing lemma we know that there exists $\gamma \in \mathscr{P}_2(X^3)$ such that

$$\begin{aligned}\pi^{1,2}_\#\gamma &= \gamma_1^2,\\ \pi^{2,3}_\#\gamma &= \gamma_2^3.\end{aligned}$$

Since $\pi^1_\#\gamma = \mu_1$ and $\pi^3_\#\gamma = \mu_3$, we have $\pi^{1,3}_\#\gamma \in \mathcal{Adm}(\mu_1,\mu_3)$ and therefore from the triangle inequality in $L^2(\gamma)$ it holds

$$\begin{aligned}W_2(\mu_1,\mu_3) &\le \sqrt{\int d^2(x_1,x_3)d\pi^{1,3}_\#\gamma(x_1,x_3)} = \sqrt{\int d^2(x_1,x_3)d\gamma(x_1,x_2,x_3)}\\ &\le \sqrt{\int \big(d(x_1,x_2)+d(x_2,x_3)\big)^2 d\gamma(x_1,x_2,x_3)}\\ &\le \sqrt{\int d^2(x_1,x_2)d\gamma(x_1,x_2,x_3)} + \sqrt{\int d^2(x_2,x_3)d\gamma(x_1,x_2,x_3)}\\ &= \sqrt{\int d^2(x_1,x_2)d\gamma_1^2(x_1,x_2)} + \sqrt{\int d^2(x_2,x_3)d\gamma_2^3(x_2,x_3)}\\ &= W_2(\mu_1,\mu_2) + W_2(\mu_2,\mu_3).\end{aligned}$$

Finally, we need to prove that W_2 is real valued. Here we use the fact that we restricted the analysis to the space $\mathscr{P}_2(X)$: from the triangle inequality we have

$$\begin{aligned}W_2(\mu,\nu) &\le W_2(\mu,\delta_{x_0}) + W_2(\nu,\delta_{x_0})\\ &= \sqrt{\int d^2(x,x_0)d\mu(x)} + \sqrt{\int d^2(x,x_0)d\nu(x)} < \infty.\end{aligned}$$

□

A trivial, yet very useful inequality is:

$$W_2^2(f_\#\mu, g_\#\mu) \le \int d_Y^2(f(x),g(x))d\mu(x), \tag{9}$$

valid for any couple of metric spaces X, Y, any $\mu \in \mathscr{P}(X)$ and any couple of Borel maps $f,\ g : X \to Y$. This inequality follows from the fact that $(f, g)_{\#}\mu$ is an admissible plan for the measures $f_{\#}\mu,\ g_{\#}\mu$, and its cost is given by the right hand side of (9).

Observe that there is a natural isometric immersion of (X, d) into $(\mathscr{P}_2(X), W_2)$, namely the map $x \mapsto \delta_x$.

Now we want to study the topological properties of $(\mathscr{P}_2(X), W_2)$. To this aim, we introduce the notion of *2-uniform integrability*: $\mathscr{K} \subset \mathscr{P}_2(X)$ is 2-uniformly integrable provided for any $\varepsilon > 0$ and $x_0 \in X$ there exists $R_\varepsilon > 0$ such that

$$\sup_{\mu \in \mathscr{K}} \int_{X \setminus B_{R_\varepsilon}(x_0)} d^2(x, x_0)\, d\mu \le \varepsilon.$$

Remark 3.3. Let $(X, d_X), (Y, d_Y)$ be Polish and endow $X \times Y$ with the product distance $d^2\big((x_1, y_1), (x_2, y_2)\big) := d_X^2(x_1, x_2) + d_Y^2(y_1, y_2)$. Then the inequality

$$\begin{aligned}\int\limits_{(B_R(x_0)\times B_R(y_0))^c} d_X^2(x, x_0)\, d\gamma(x, y) &= \int\limits_{(B_R(x_0))^c \times Y} d_X^2(x, x_0)\, d\gamma(x, y) + \int\limits_{B_R(x_0)\times(B_R(y_0))^c} d_X^2(x, x_0)\, d\gamma(x, y)\\ &\le \int\limits_{(B_R(x_0))^c} d_X^2(x, x_0)\, d\mu(x) + \int\limits_{X\times(B_R(y_0))^c} R^2\, d\gamma(x, y)\\ &\le \int\limits_{(B_R(x_0))^c} d_X^2(x, x_0)\, d\mu(x) + \int\limits_{(B_R(y_0))^c} d_Y^2(y, y_0)\, d\nu(y),\end{aligned}$$

valid for any $\gamma \in \mathcal{A}dm(\mu, \nu)$ and the analogous one with the integral of $d_Y^2(y, y_0)$ in place of $d_X^2(x, x_0)$, show that if $\mathscr{K}_1 \subset \mathscr{P}_2(X)$ and $\mathscr{K}_2 \subset \mathscr{P}_2(Y)$ are 2-uniformly integrable, so is the set

$$\Big\{\gamma \in \mathscr{P}(X \times Y) :\ \pi^X_{\#}\gamma \in \mathscr{K}_1,\ \pi^Y_{\#}\gamma \in \mathscr{K}_2\Big\}.$$

■

We say that a function $f : X \to \mathbb{R}$ has quadratic growth provided

$$|f(x)| \le a(d^2(x, x_0) + 1), \tag{10}$$

for some $a \in \mathbb{R}$ and $x_0 \in X$. It is immediate to check that if f has quadratic growth and $\mu \in \mathscr{P}_2(X)$, then $f \in L^1(X, \mu)$.

The concept of 2-uniform integrability (in conjunction with tightness) in relation with convergence of integral of functions with quadratic growth, plays a role similar to the one played by tightness in relation with convergence of integral of bounded functions, as shown in the next proposition.

Proposition 3.4. *Let $(\mu_n) \subset \mathscr{P}_2(X)$ be a sequence narrowly converging to some μ. Then the following three properties are equivalent*

(i) *(μ_n) is 2-uniformly integrable.*
(ii) *$\int f d\mu_n \to \int f d\mu$ for any continuous f with quadratic growth.*
(iii) *$\int d^2(\cdot, x_0) d\mu_n \to \int d^2(\cdot, x_0) d\mu$ for some $x_0 \in X$.*

Proof. **(i)** ⇒ **(ii)**. It is not restrictive to assume $f \geq 0$. Since any such f can be written as supremum of a family of continuous and bounded functions, it clearly holds

$$\int f d\mu \leq \liminf_{n\to\infty} \int f d\mu_n.$$

Thus we only have to prove the limsup inequality. Fix $\varepsilon > 0$, $x_0 \in X$ and find $R_\varepsilon > 1$ such that $\int_{X \setminus B_{R_\varepsilon}(x_0)} d^2(\cdot, x_0) d\mu_n \leq \varepsilon$ for every n. Now let χ be a function with bounded support, values in $[0, 1]$ and identically 1 on B_{R_ε} and notice that for every $n \in \mathbb{N}$ it holds

$$\int f d\mu_n = \int f\chi d\mu_n + \int f(1-\chi) d\mu_n \leq \int f\chi d\mu_n + \int_{X\setminus B_{R_\varepsilon}} f d\mu_n \leq \int f\chi d\mu_n + 2a\varepsilon,$$

a being given by (10). Since $f\chi$ is continuous and bounded we have $\int f\chi d\mu_n \to \int f\chi d\mu$ and therefore

$$\overline{\lim_{n\to\infty}} \int f d\mu_n \leq \int f\chi d\mu + 2a\varepsilon \leq \int f d\mu + 2a\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this part of the statement is proved.
(ii) ⇒ **(iii)**. Obvious.
(iii) ⇒ **(i)**. Argue by contradiction and assume that there exist $\varepsilon > 0$ and $\tilde{x}_0 \in X$ such that for every $R > 0$ it holds $\sup_{n\in\mathbb{N}} \int_{X\setminus B_R(\tilde{x}_0)} d^2(\cdot, \tilde{x}_0) d\mu_n > \varepsilon$. Then it is easy to see that it holds

$$\overline{\lim_{n\to\infty}} \int_{X\setminus B_R(x_0)} d^2(\cdot, x_0) d\mu_n > \varepsilon. \tag{11}$$

For every $R > 0$ let χ_R be a continuous cutoff function with values in $[0, 1]$ supported on $B_R(x_0)$ and identically 1 on $B_{R/2}(x_0)$. Since $d^2(\cdot, x_0)\chi_R$ is continuous and bounded, we have

$$\begin{aligned}
\int d^2(\cdot, x_0)\chi_R d\mu &= \lim_{n\to\infty} \int d^2(\cdot, x_0)\chi_R d\mu_n \\
&= \lim_{n\to\infty} \left(\int d^2(\cdot, x_0) d\mu_n - \int d^2(\cdot, x_0)(1-\chi_R) d\mu_n \right) \\
&= \int d^2(\cdot, x_0) d\mu + \lim_{n\to\infty} - \int d^2(\cdot, x_0)(1-\chi_R) d\mu_n
\end{aligned}$$

$$
\begin{aligned}
&\leq \int d^2(\cdot, x_0)d\mu + \varliminf_{n\to\infty} -\int_{X\setminus B_R(x_0)} d^2(\cdot, x_0)d\mu_n \\
&= \int d^2(\cdot, x_0)d\mu - \varlimsup_{n\to\infty} \int_{X\setminus B_R(x_0)} d^2(\cdot, x_0)d\mu_n \\
&\leq \int d^2(\cdot, x_0)d\mu - \varepsilon,
\end{aligned}
$$

having used (11) in the last step. Since

$$\int d^2(\cdot, x_0)d\mu = \sup_R \int d^2(\cdot, x_0)\chi_R d\mu \leq \int d^2(\cdot, x_0)d\mu - \varepsilon,$$

we got a contradiction. □

Proposition 3.5 (Stability of optimality). *The distance W_2 is lower semicontinuous w.r.t. narrow convergence of measures. Furthermore, if $(\gamma_n) \subset \mathscr{P}_2(X^2)$ is a sequence of optimal plans which narrowly converges to $\gamma \in \mathscr{P}_2(X^2)$, then γ is optimal as well.*

Proof. Let $(\mu_n), (\nu_n) \subset \mathscr{P}_2(X)$ be two sequences of measures narrowly converging to $\mu, \nu \in \mathscr{P}_2(X)$ respectively. Pick $\gamma_n \in Opt(\mu_n, \nu_n)$ and use Remark 2.4 and Prokhorov theorem to get that (γ_n) admits a subsequence, not relabeled, narrowly converging to some $\gamma \in \mathscr{P}(X^2)$. It is clear that $\pi^1_\#\gamma = \mu$ and $\pi^2_\#\gamma = \nu$, thus it holds

$$W_2^2(\mu, \nu) \leq \int d^2(x, y)d\gamma(x, y) \leq \varliminf_{n\to\infty} \int d^2(x, y)d\gamma_n(x, y) = \varliminf_{n\to\infty} W_2^2(\mu_n, \nu_n).$$

Now we pass to the second part of the statement, that is: we need to prove that with the same notation just used it holds $\gamma \in Opt(\mu, \nu)$. Choose $a(x) = b(x) = d^2(x, x_0)$ for some $x_0 \in X$ in the bound (4) and observe that since $\mu, \nu \in \mathscr{P}_2(X)$ Theorem 2.13 applies, and thus optimality is equivalent to c-cyclical monotonicity of the support. The same for the plans γ_n. Fix $N \in \mathbb{N}$ and pick $(x^i, y^i) \in \operatorname{supp}(\gamma)$, $i = 1, \ldots, N$. From the fact that (γ_n) narrowly converges to γ it is not hard to infer the existence of $(x^i_n, y^i_n) \in \operatorname{supp}(\gamma_n)$ such that

$$\lim_{n\to\infty} \Big(d(x^i_n, x^i) + d(y^i_n, y^i) \Big) = 0, \qquad \forall i = 1, \ldots, N.$$

Thus the conclusion follows from the c-cyclical monotonicity of $\operatorname{supp}(\gamma_n)$ and the continuity of the cost function. □

Now we are going to prove that $(\mathscr{P}_2(X), W_2)$ is a Polish space. In order to enable some constructions, we will use (a version of) Kolmogorov's theorem, which we recall without proof (see e.g. [31] Sect. 51).

Theorem 3.6 (Kolmogorov). *Let X be a Polish space and $\mu_n \in \mathscr{P}(X^n)$, $n \in \mathbb{N}$, be a sequence of measures such that*

$$\pi^{1,\ldots,n-1}_{\#}\mu_n = \mu_{n-1}, \qquad \forall n \geq 2.$$

Then there exists a measure $\mu \in X^{\mathbb{N}}$ such that

$$\pi^{1,\ldots,n}_{\#}\mu = \mu_n, \qquad \forall n \in \mathbb{N}.$$

Theorem 3.7 (Basic properties of the space $(\mathscr{P}_2(X), W_2)$). *Let (X,d) be complete and separable. Then*

$$W_2(\mu_n,\mu) \to 0 \qquad \Leftrightarrow \qquad \begin{cases} \mu_n \to \mu & \textit{narrowly} \\ \int d^2(\cdot,x_0)\,d\mu_n \to \int d^2(\cdot,x_0)\,d\mu & \textit{for some } x_0 \in X. \end{cases} \tag{12}$$

Furthermore, the space $(\mathscr{P}_2(X), W_2)$ is complete and separable. Finally, $\mathscr{K} \subset \mathscr{P}_2(X)$ is relatively compact w.r.t. the topology induced by W_2 if and only if it is tight and 2-uniformly integrable.

Proof. We start showing implication $\Rightarrow$ in (12). Thus assume that $W_2(\mu_n,\mu) \to 0$. Then

$$\left| \sqrt{\int d^2(\cdot,x_0)\,d\mu_n} - \sqrt{\int d^2(\cdot,x_0)\,d\mu} \right| = |W_2(\mu_n,\delta_{x_0}) - W_2(\mu,\delta_{x_0})| \leq W_2(\mu_n,\mu) \to 0.$$

To prove narrow convergence, for every $n \in \mathbb{N}$ choose $\gamma_n \in Opt(\mu,\mu_n)$ and[2] use repeatedly the gluing lemma to find, for every $n \in \mathbb{N}$, a measure $\alpha_n \in \mathscr{P}(X \times X^n)$ such that

[2] If closed balls in X are compact, the proof greatly simplifies. Indeed in this case the inequality $R^2\mu(X \setminus B_R(x_0)) \leq \int d^2(\cdot,x_0)\,d\mu$ and the uniform bound on the second moments yields that the sequence $n \mapsto \mu_n$ is tight. Thus to prove narrow convergence it is sufficient to check that $\int f\,d\mu_n \to \int f\,d\mu$ for every $f \in C_c(X)$. Since Lipschitz functions are dense in $C_c(X)$ w.r.t. uniform convergence, it is sufficient to check the convergence of the integral only for Lipschitz f's. This follows from the inequality

$$\begin{aligned} \left| \int f\,d\mu - \int f\,d\mu_n \right| &= \left| \int f(x) - f(y)\,d\gamma_n(x,y) \right| \leq \int |f(x) - f(y)|\,d\gamma_n(x,y) \\ &\leq \mathrm{Lip}(f) \int d(x,y)\,d\gamma_n(x,y) \leq \mathrm{Lip}(f)\sqrt{\int d^2(x,y)\,d\gamma_n(x,y)} \\ &= \mathrm{Lip}(f)W_2(\mu,\mu_n). \end{aligned}$$

$$\pi^{0,n}_{\#}\alpha_n = \gamma_n,$$
$$\pi^{0,1,\ldots,n-1}_{\#}\alpha_n = \alpha_{n-1}.$$

Then by Kolmogorov's theorem we know that there exists a measure $\alpha \in \mathscr{P}(X \times X^{\mathbb{N}})$ such that

$$\pi^{0,1,\ldots,n}_{\#}\alpha = \alpha_n, \qquad \forall n \in \mathbb{N}.$$

By construction we have

$$\|d(\pi^0, \pi^n)\|_{L^2(X\times X^{\mathbb{N}},\alpha)} = \|d(\pi^0, \pi^n)\|_{L^2(X^2,\gamma_n)} = W_2(\mu, \mu_n) \to 0.$$

Thus up to passing to a subsequence, not relabeled, we can assume that $\pi^n(\mathbf{x}) \to \pi^0(\mathbf{x})$ for α-almost any $\mathbf{x} \in X \times X^{\mathbb{N}}$. Now pick $f \in C_b(X)$ and use the dominated convergence theorem to get

$$\lim_{n\to\infty}\int f d\mu_n = \lim_{n\to\infty}\int f \circ \pi^n d\alpha = \int f \circ \pi^0 d\alpha = \int f d\mu.$$

Since the argument does not depend on the subsequence chosen, the claim is proved.

We pass to the converse implication in (12). Pick $\gamma_n \in Opt(\mu, \mu_n)$ and use Remark 2.4 to get that the sequence (γ_n) is tight, hence, up to passing to a subsequence, we can assume that it narrowly converges to some γ. By Proposition 3.5 we know that $\gamma \in Opt(\mu, \mu)$, which forces $\int d^2(x, y)d\gamma(x, y) = 0$. By Proposition 3.4 and our assumption on (μ_n), μ we know that (μ_n) is 2-uniformly integrable, thus by Remark 3.3 again we know that (γ_n) is 2-uniformly integrable as well. Since the map $(x, y) \mapsto d^2(x, y)$ has quadratic growth in X^2 it holds

$$\lim_{n\to\infty} W_2^2(\mu_n, \mu) = \lim_{n\to\infty}\int d^2(x, y)d\gamma_n(x, y) = \int d^2(x, y)d\gamma(x, y) = 0.$$

Now we prove that $(\mathscr{P}_2(X), W_2)$ is complete. Pick a Cauchy sequence (μ_n) and assume,[3] without loss of generality, that $\sum_n W_2(\mu_n, \mu_{n+1}) < \infty$. For every $n \in \mathbb{N}$ choose $\gamma_n \in Opt(\mu_n, \mu_{n+1})$ and use repeatedly the gluing lemma to find, for every $n \in \mathbb{N}$, a measure $\beta_n \in \mathscr{P}_2(X^n)$ such that

$$\pi^{n,n+1}_{\#}\beta_n = \gamma_n,$$
$$\pi^{1,\ldots,n-1}_{\#}\beta_n = \alpha_{n-1}$$

[3] Again, if closed balls in X are compact the argument simplifies. Indeed from the uniform bound on the second moments and the inequality $R^2\mu(X \setminus B_R(x_0)) \leq \int_{X\setminus B_R(x_0)} d^2(\cdot, x_0)d\mu$ we get the tightness of the sequence. Hence up to pass to a subsequence we can assume that (μ_n) narrowly converges to a limit measure μ, and then using the lower semicontinuity of W_2 w.r.t. narrow convergence we can conclude $\overline{\lim}_n W_2(\mu, \mu_n) \leq \overline{\lim}_n \underline{\lim}_m W_2(\mu_m, \mu_n) = 0$.

By Kolmogorov's theorem we get the existence of a measure $\beta \in \mathscr{P}(X^{\mathbb{N}})$ such that $\pi^{1,\ldots,n}_{\#}\beta = \beta_n$ for every $n \in \mathbb{N}$. The inequality

$$\sum_{n=1}^{\infty} \|d(\pi^i, \pi^{i+1})\|_{L^2(X^{\mathbb{N}},\beta)} = \sum_{n=1}^{\infty} \|d(\pi^i, \pi^{i+1})\|_{L^2(X^2,\gamma_i)} = \sum_{n=1}^{\infty} W_2(\mu_i, \mu_{i+1}) < \infty,$$

shows that $n \mapsto \pi^n : X^{\mathbb{N}} \to X$ is a Cauchy sequence in $L^2(\beta, X)$, i.e. the space of maps $f : X^{\mathbb{N}} \to X$ such that $\int d^2(f(y), x_0)d\beta(y) < \infty$ for some, and thus every, $x_0 \in X$ endowed with the distance $\tilde{d}(f, g) := \sqrt{\int d^2(f(y), g(y))d\beta(y)}$. Since X is complete, $L^2(\beta, X)$ is complete as well, and therefore there exists a limit map π^∞ of the Cauchy sequence (π^n). Define $\mu := \pi^\infty_{\#}\beta$ and notice that by (9) we have

$$W_2^2(\mu, \mu_n) \le \int d^2(\pi^\infty, \pi^n)d\beta \to 0,$$

so that μ is the limit of the Cauchy sequence (μ_n) in $(\mathscr{P}_2(X), W_2)$. The fact that $(\mathscr{P}_2(X), W_2)$ is separable follows from (12) by considering the set of finite convex combinations of Dirac masses centered at points in a dense countable set in X with rational coefficients. The last claim now follows. □

Remark 3.8 (On compactness properties of $\mathscr{P}_2(X)$). An immediate consequence of the above theorem is the fact that if X is compact, then $(\mathscr{P}_2(X), W_2)$ is compact as well: indeed, in this case the equivalence (12) tells that convergence in $\mathscr{P}_2(X)$ is equivalent to weak convergence.

It is also interesting to notice that if X is unbounded, then $\mathscr{P}_2(X)$ is not locally compact. Actually, for any measure $\mu \in \mathscr{P}_2(X)$ and any $r > 0$, the closed ball of radius r around μ is not compact. To see this, fix $\overline{x} \in X$ and find a sequence $(x_n) \subset X$ such that $d(x_n, \overline{x}) \to \infty$. Now define the measures $\mu_n := (1 - \varepsilon_n)\mu + \varepsilon_n\delta_{x_n}$, where ε_n is chosen such that $\varepsilon_n d^2(\overline{x}, x_n) = r^2$. To bound from above $W_2^2(\mu, \mu_n)$, leave fixed $(1 - \varepsilon_n)\mu$, move $\varepsilon_n\mu$ to $\overline{x}$ and then move $\varepsilon_n\delta_{\overline{x}}$ into $\varepsilon_n\delta_{x_n}$, this gives

$$W_2^2(\mu, \mu_n) \le \varepsilon_n \left(\int d^2(x, \overline{x})d\mu(x) + d^2(x_n, \overline{x}) \right),$$

so that $\overline{\lim}\, W_2(\mu, \mu_n) \le r$. Conclude observing that

$$\lim_{n\to\infty} \int d^2(x, \overline{x})d\mu_n = \lim_{n\to\infty} (1-\varepsilon_n) \int d^2(x, \overline{x})d\mu + \varepsilon_n d^2(x_n, \overline{x}) = \int d^2(x, \overline{x})d\mu + r^2,$$

thus the second moments do not converge. Since clearly (μ_n) weakly converges to μ, we proved that there is no local compactness. ■

3.2 X Geodesic Space

In this section we prove that if the base space (X, d) is geodesic, then the same is true also for $(\mathscr{P}_2(X), W_2)$ and we will analyze the properties of this latter space.

Let us recall that a curve $\gamma : [0, 1] \to X$ is called *constant speed geodesic* provided

$$d\big(\gamma_t, \gamma_s\big) = |t - s| d\big(\gamma_0, \gamma_1\big), \qquad \forall t, s \in [0, 1], \tag{13}$$

or equivalently if $\leq$ always holds.

Definition 3.9 (Geodesic space). A metric space (X, d) is called *geodesic* if for every $x, y \in X$ there exists a constant speed geodesic connecting them, i.e. a constant speed geodesic such that $\gamma_0 = x$ and $\gamma_1 = y$.

Before entering into the details, let us describe an important example. Recall that $X \ni x \mapsto \delta_x \in \mathscr{P}_2(X)$ is an isometry. Therefore if $t \mapsto \gamma_t$ is a constant speed geodesic on X connecting x to y, the curve $t \mapsto \delta_{\gamma_t}$ is a constant speed geodesic on $\mathscr{P}_2(X)$ which connects δ_x to δ_y. The important thing to notice here is that the natural way to interpolate between δ_x and δ_y is given by this—so called—*displacement interpolation*. Conversely, observe that the classical linear interpolation

$$t \mapsto \mu_t := (1 - t)\delta_x + t\delta_y,$$

produces a curve which has infinite length as soon as $x \neq y$ (because $W_2(\mu_t, \mu_s) = \sqrt{|t - s|} d(x, y)$), and thus is unnatural in this setting.

We will denote by $\mathrm{Geod}(X)$ the metric space of all constant speed geodesics on X endowed with the sup norm. With some work it is possible to show that $\mathrm{Geod}(X)$ is complete and separable as soon as X is (we omit the details). The *evaluation maps* $\mathrm{e}_t : \mathrm{Geod}(X) \to X$ are defined for every $t \in [0, 1]$ by

$$\mathrm{e}_t(\gamma) := \gamma_t. \tag{14}$$

Theorem 3.10. *Let (X, d) be Polish and geodesic. Then $(\mathscr{P}_2(X), W_2)$ is geodesic as well. Furthermore, the following two are equivalent:*

(i) *$t \mapsto \mu_t \in \mathscr{P}_2(X)$ is a constant speed geodesic.*
(ii) *There exists a measure $\mu \in \mathscr{P}_2(\mathrm{Geod}(X))$ such that $(\mathrm{e}_0, \mathrm{e}_1)_\# \mu \in \mathit{Opt}(\mu_0, \mu_1)$ and*

$$\mu_t = (\mathrm{e}_t)_\# \mu. \tag{15}$$

Proof. Choose $\mu^0, \mu^1 \in \mathscr{P}_2(X)$ and find an optimal plan $\gamma \in \mathit{Opt}(\mu, \nu)$. By Lemma 3.11 below and classical measurable selection theorems we know that there exists a Borel map $\mathrm{GeodSel} : X^2 \to \mathrm{Geod}(X)$ such that for any $x, y \in X$ the curve $\mathrm{GeodSel}(x, y)$ is a constant speed geodesic connecting x to y. Define the Borel probability measure $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ by

$$\mu := \mathrm{GeodSel}_\# \gamma,$$

and the measures $\mu_t \in \mathscr{P}(X)$ by $\mu_t := (\mathrm{e}_t)_\# \mu$.

We claim that $t \mapsto \mu_t$ is a constant speed geodesic connecting μ^0 to μ^1. Consider indeed the map $(\mathrm{e}_0, \mathrm{e}_1) : \mathrm{Geod}(X) \to X^2$ and observe that from $(\mathrm{e}_0, \mathrm{e}_1)\Big(\mathrm{GeodSel}(x, y)\Big) = (x, y)$ we get

$$(\mathrm{e}_0, \mathrm{e}_1)_\# \mu = \gamma. \tag{16}$$

In particular, $\mu_0 = (\mathrm{e}_0)_\# \mu = \pi^1_\# \gamma = \mu^0$, and similarly $\mu_1 = \mu^1$, so that the curve $t \mapsto \mu_t$ connects μ^0 to μ^1. The facts that the measures μ_t have finite second moments and (μ_t) is a constant speed geodesic follow from

$$\begin{aligned} W_2^2(\mu_t, \mu_s) &\overset{(15),(9)}{\leq} \int d^2(\mathrm{e}_t(\gamma), \mathrm{e}_s(\gamma)) d\mu(\gamma) \\ &\overset{(13)}{=} (t-s)^2 \int d^2(\mathrm{e}_0(\gamma), \mathrm{e}_1(\gamma)) d\mu(\gamma) \\ &\overset{(16)}{=} (t-s)^2 \int d^2(x, y) d\gamma(x, y) = (t-s)^2 W_2^2(\mu^0, \mu^1). \end{aligned}$$

The fact that (ii) implies (i) follows from the same kind of argument just used. So, we turn to $(i) \Rightarrow (ii)$. For $n \geq 0$ we use iteratively the gluing Lemma 3.1 and the Borel map GeodSel to build a measure $\mu^n \in \mathscr{P}(C([0,1], X))$ such that

$$\left(\mathrm{e}_{i/2^n}, \mathrm{e}_{(i+1)/2^n}\right)_\# \mu^n \in Opt(\mu_{i/2^n}, \mu_{(i+1)/2^n}), \qquad \forall i = 0, \ldots, 2^n - 1,$$

and μ^n-a.e. γ is a geodesic in the intervals $[i/2^n, (i+1)/2^n]$, $i = 0, \ldots, 2^n - 1$. Fix n and observe that for any $0 \leq j < k \leq 2^n$ it holds

$$\begin{aligned} \left\| d\left(\mathrm{e}_{j/2^n}, \mathrm{e}_{k/2^n}\right)\right\|_{L^2(\mu^n)} &\leq \left\| \sum_{i=j}^{k-1} d\left(\mathrm{e}_{i/2^n}, \mathrm{e}_{(i+1)/2^n}\right)\right\|_{L^2(\mu^n)} \\ &\leq \sum_{i=j}^{k-1} \left\| d\left(\mathrm{e}_{i/2^n}, \mathrm{e}_{(i+1)/2^n}\right)\right\|_{L^2(\mu^n)} \\ &= \sum_{i=j}^{k-1} W_2(\mu_{i/2^n}, \mu_{(i+1)/2^n}) = W_2(\mu_{j/2^n}, \mu_{k/2^n}). \end{aligned} \tag{17}$$

Therefore it holds

$$\left(\mathrm{e}_{j/2^n}, \mathrm{e}_{k/2^n}\right)_\# \mu^n \in Opt(\mu_{j/2^n}, \mu_{k/2^n}), \qquad \forall j, k \in \{0, \ldots, 2^n\}.$$

Also, since the inequalities in (17) are equalities, it is not hard to see that for μ^n-a.e. γ the points $\gamma_{i/2^n}$, $i = 0, \ldots, 2^n$, must lie along a geodesic and satisfy $d(\gamma_{i/2^n}, \gamma_{(i+1)/2^n}) = d(\gamma_0, \gamma_1)/2^n$, $i = 0, \ldots, 2^n - 1$. Hence μ^n-a.e. γ is a constant speed geodesic and thus $\mu^n \in \mathscr{P}(\mathrm{Geod}(X))$. Now suppose for a moment that (μ^n) narrowly converges—up to pass to a subsequence—to some $\mu \in \mathscr{P}(\mathrm{Geod}(X))$. Then the continuity of the evaluation maps e_t yields that for any $t \in [0,1]$ the sequence $n \mapsto (\mathrm{e}_t)_\#\mu^n$ narrowly converges to $(\mathrm{e}_t)_\#\mu$ and this, together with the uniform bound (17), easily implies that μ satisfies (15).

Thus to conclude it is sufficient to show that some subsequence of (μ_n) has a narrow limit.[4] We will prove this by showing that $\mu^n \in \mathscr{P}_2(\mathrm{Geod}(X))$ for every $n \in \mathbb{N}$ and that some subsequence is a Cauchy sequence in $(\mathscr{P}_2(\mathrm{Geod}(X)), \mathbf{W}_2)$, $\mathbf{W}_2$ being the Wasserstein distance built over $\mathrm{Geod}(X)$ endowed with the sup distance, so that by Theorem 3.7 we conclude.

We know by Remarks 2.4, 3.3 and Theorem 3.7 that for every $n \in \mathbb{N}$ the set of plans $\alpha \in \mathscr{P}_2(X^{2^n+1})$ such that $\pi^i_\#\alpha = \mu_{i/2^n}$ for $i = 0, \dots, 2^n$, is compact in $\mathscr{P}_2(X^{2^n+1})$. Therefore a diagonal argument tells that possibly passing to a subsequence, not relabeled, we may assume that for every $n \in \mathbb{N}$ the sequence

$$m \mapsto \prod_{i=0}^{2^n} (\mathrm{e}_{i/2^n})_\#\mu^m$$

converges to some plan w.r.t. the distance W_2 on X^{2^n+1}.

Now fix $n \in \mathbb{N}$ and notice that for $t \in [i/2^n, (i+1)/2^n]$ and $\gamma, \tilde{\gamma} \in \mathrm{Geod}(X)$ it holds

$$d\big(\gamma_t, \tilde{\gamma}_t\big) \le d\big(\gamma_{i/2^n}, \tilde{\gamma}_{(i+1)/2^n}\big) + \frac{1}{2^n}\Big(d(\gamma_0, \gamma_1) + d(\tilde{\gamma}_0, \tilde{\gamma}_1)\Big),$$

and therefore squaring and then taking the sup over $t \in [0,1]$ we get

$$\sup_{t \in [0,1]} d^2(\gamma_t, \tilde{\gamma}_t) \le 2\sum_{i=0}^{2^n-1} d^2\big(\gamma_{i/2^n}, \tilde{\gamma}_{(i+1)/2^n}\big) + \frac{1}{2^{n-2}}\Big(d^2(\gamma_0, \gamma_1) + d^2(\tilde{\gamma}_0, \tilde{\gamma}_1)\Big). \tag{18}$$

Choosing $\tilde{\gamma}$ to be a constant geodesic and using (17), we get that $\mu^m \in \mathscr{P}_2(\mathrm{Geod}(X))$ for every $m \in \mathbb{N}$. Now, for any given $\nu, \tilde{\nu} \in \mathscr{P}(\mathrm{Geod}(X))$, by a gluing argument (Lemma 3.12 below with $\nu, \tilde{\nu}$ in place of $\nu, \tilde{\nu}$, $Y = \mathrm{Geod}(X)$, $Z = X^{2^n+1}$) we can find a plan $\beta \in \mathscr{P}([\mathrm{Geod}(X)]^2)$ such that

$$\pi^1_\#\beta = \nu,$$

$$\pi^2_\#\beta = \tilde{\nu},$$

$$\Big(\big(\mathrm{e}_0,\dots,\mathrm{e}_{i/2^n},\dots,\mathrm{e}_1\big)\circ\pi^1, \big(\mathrm{e}_0,\dots,\mathrm{e}_{i/2^n},\dots,\mathrm{e}_1\big)\circ\pi^2\Big)_\#\beta \in Opt(\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\nu, \prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\tilde{\nu})$$

[4] As for Theorem 3.7 everything is simpler if closed balls in X are compact. Indeed, observe that a geodesic connecting two points in $B_R(x_0)$ lies entirely on the compact set $\overline{B_{2R}(x_0)}$, and that the set of geodesics lying on a given compact set is itself compact in $\mathrm{Geod}(X)$, so that the tightness of (μ^n) follows directly from the one of $\{\mu_0, \mu_1\}$.

where optimality between $\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\nu$ and $\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\tilde{\nu}$ is meant w.r.t. the Wasserstein distance on $\mathscr{P}_2(X^{2^n+1})$. Using β to bound from above $\mathbf{W}_2(\nu,\tilde{\nu})$ and using (18) we get that for every couple of measures $\nu,\ \tilde{\nu}\in\mathscr{P}_2(\mathrm{Geod}(X))$ it holds

$$\begin{aligned}\mathbf{W}_2^2(\nu,\tilde{\nu})\ \le\ & 2W_2^2\Big(\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\nu,\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\tilde{\nu}\Big)\\ &+\frac{1}{2^{n-2}}\left(\int d^2(\gamma_0,\gamma_1)d\nu(\gamma)+\int d^2(\tilde{\gamma}_0,\tilde{\gamma}_1)d\nu(\tilde{\gamma})\right)\end{aligned}$$

Plugging $\nu=\mu^m$, $\tilde{\nu}=\mu^{m'}$ and recalling that $W_2\Big(\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\mu^m,\prod_{i=0}^{2^n}(\mathrm{e}_{i/2^n})_\#\mu^{m'}\Big)\to 0$ as $m,\ m'\to+\infty$ for every $n\in\mathbb{N}$ we get that

$$\begin{aligned}\overline{\lim_{m,m'\to\infty}}\mathbf{W}_2^2(\mu^m,\mu^{m'})&\le\frac{1}{2^{n-2}}\left(\int d^2(\gamma_0,\gamma_1)d\mu^m(\gamma)+\int d^2(\tilde{\gamma}_0,\tilde{\gamma}_1)d\mu^{m'}(\tilde{\gamma})\right)\\ &=\frac{1}{2^{n-3}}W_2^2(\mu_0,\mu_1).\end{aligned}$$

Letting $n\to\infty$ we get that $(\mu^m)\subset\mathscr{P}_2(\mathrm{Geod}(X))$ is a Cauchy sequence and the conclusion. □

Lemma 3.11. *The multivalued map from $G:X^2\to\mathrm{Geod}(X)$ which associates to each pair (x,y) the set $G(x,y)$ of constant speed geodesics connecting x to y has closed graph.*

Proof. Straightforward. □

Lemma 3.12 (A variant of gluing). *Let Y,Z be Polish spaces, $\nu,\tilde{\nu}\in\mathscr{P}(Y)$ and $f,g:Y\to Z$ be two Borel maps. Let $\gamma\in\mathcal{Adm}(f_\#\nu,g_\#\tilde{\nu})$. Then there exists a plan $\beta\in\mathscr{P}(Y^2)$ such that*

$$\begin{aligned}\pi^1_\#\beta&=\nu,\\ \pi^2_\#\beta&=\tilde{\nu},\\ (f\circ\pi^1,g\circ\pi^2)_\#\beta&=\gamma.\end{aligned}$$

Proof. Let $\{\nu_z\},\{\tilde{\nu}_{\tilde{z}}\}$ be the disintegrations of $\nu,\tilde{\nu}$ w.r.t. f,g respectively. Then define

$$\beta:=\int_{Z^2}\nu_z\times\tilde{\nu}_{\tilde{z}}\,d\gamma(z,\tilde{z}).$$

□

Remark 3.13 (The Hilbert case). If X is an Hilbert space, then for every $x,\ y\in X$ there exists only one constant speed geodesic connecting them: the curve $t\mapsto(1-t)x+ty$. Thus Theorem 3.10 reads as: $t\mapsto\mu_t$ is a constant speed geodesic if and only if there exists an optimal plan $\gamma\in\mathcal{Opt}(\mu_0,\mu_1)$ such that

$$\mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\# \gamma.$$

If γ is induced by a map T, the formula further simplifies to

$$\mu_t = \big((1-t)Id + tT\big)_\# \mu_0. \tag{19}$$

■

Remark 3.14. A slight modification of the arguments presented in the second part of the proof of Theorem 3.10 shows that if (X, d) is Polish and $(\mathscr{P}_2(X), W_2)$ is geodesic, then (X, d) is geodesic as well. Indeed, given $x, y \in X$ and a geodesic (μ_t) connecting δ_x to δ_y, we can build a measure $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ satisfying (15). Then every $\gamma \in \mathrm{supp}(\mu)$ is a geodesic connecting x to y. ■

Definition 3.15 (Non branching spaces). A geodesic space (X, d) is said non branching if for any $t \in (0, 1)$ a constant speed geodesic γ is uniquely determined by its initial point γ_0 and by the point γ_t. In other words, (X, d) is non branching if the map

$$\mathrm{Geod}(X) \ni \gamma \mapsto (\gamma_0, \gamma_t) \in X^2,$$

is injective for any $t \in (0, 1)$.

Non-branching spaces are interesting from the optimal transport point of view, because for such spaces the behavior of geodesics in $\mathscr{P}_2(X)$ is particularly nice: optimal transport plan from intermediate measures to other measures along the geodesic are unique and induced by maps (it is quite surprising that such a statement is true in this generality—compare the assumption of the proposition below with the ones of Theorems 2.26, 2.33). Examples of non-branching spaces are Riemannian manifolds, Banach spaces with strictly convex norms and Alexandrov spaces with curvature bounded below. Examples of branching spaces are Banach spaces with non strictly convex norms.

Proposition 3.16 (Non branching and interior regularity). *Let (X, d) be a Polish, geodesic, non branching space. Then $(\mathscr{P}_2(X), W_2)$ is non branching as well. Furthermore, if $(\mu_t) \subset \mathscr{P}_2(X)$ is a constant speed geodesic, then for every $t \in (0, 1)$ there exists only one optimal plan in $Opt(\mu_0, \mu_t)$ and this plan is induced by a map from μ_t. Finally, the measure $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ associated to (μ_t) via (15) is unique.*

Proof. Let $(\mu_t) \subset \mathscr{P}_2(X)$ be a constant speed geodesic and fix $t_0 \in (0, 1)$. Pick $\gamma^1 \in Opt(\mu_0, \mu_{t_0})$ and $\gamma^2 \in Opt(\mu_{t_0}, \mu_1)$. We want to prove that both γ^1 and γ^2 are induced by maps from μ_{t_0}. To this aim use the gluing lemma to find a 3-plan $\alpha \in \mathscr{P}_2(X^3)$ such that

$$\pi^{1,2}_\# \alpha = \gamma^1,$$

$$\pi^{2,3}_\# \alpha = \gamma^2,$$

and observe that since (μ_t) is a geodesic it holds

$$\begin{aligned}\|d(\pi^1,\pi^3)\|_{L^2(\alpha)} &\le \|d(\pi^1,\pi^2)+d(\pi^2,\pi^3)\|_{L^2(\alpha)}\\ &\le \|d(\pi^1,\pi^2)\|_{L^2(\alpha)}+\|d(\pi^2,\pi^3)\|_{L^2(\alpha)}\\ &= \|d(\pi^1,\pi^2)\|_{L^2(\gamma^1)}+\|d(\pi^1,\pi^2)\|_{L^2(\gamma^2)}\\ &= W_2(\mu_0,\mu_{t_0})+W_2(\mu_{t_0},\mu_1)\\ &= W_2(\mu_0,\mu_1),\end{aligned}$$

so that $(\pi^1,\pi^3)_\#\alpha \in Opt(\mu_0,\mu_1)$. Also, since the first inequality is actually an equality, we have that $d(x,y)+d(y,z)=d(x,z)$ for α-a.e. (x,y,z), which means that x,y,z lie along a geodesic. Furthermore, since the second inequality is an equality, the functions $(x,y,z)\mapsto d(x,y)$ and $(x,y,z)\mapsto d(y,z)$ are each a positive multiple of the other in supp(α). It is then immediate to verify that for every $(x,y,z)\in$ supp(α) it holds

$$\begin{aligned}d(x,y)&=(1-t_0)d(x,z),\\ d(y,z)&=t_0 d(x,z).\end{aligned}$$

We now claim that for $(x,y,z),(x',y',z')\in$ supp(α) it holds $(x,y,z)=(x',y',z')$ if and only if $y=y'$. Indeed, pick (x,y,z), $(x',y,z')\in$ supp(α) and assume, for instance, that $z\neq z'$. Since $(\pi^1,\pi^3)_\#\alpha$ is an optimal plan, by the cyclical monotonicity of its support we know that

$$\begin{aligned}d^2(x,z)+d^2(x',z')&\le d^2(x,z')+d^2(x',z)\\ &\le \big(d(x,y)+d(y,z')\big)^2+\big(d(x',y)+d(y,z)\big)^2\\ &= \big((1-t_0)d(x,z)+t_0d(x',z')\big)^2+\big((1-t_0)d(x',z')+t_0d(x,z)\big)^2,\end{aligned}$$

which, after some manipulation, gives $d(x,z)=d(x',z')=:D$. Again from the cyclical monotonicity of the support we have $2D^2\le d^2(x,z')+d^2(x',z)$, thus either $d(x',z)$ or $d(x,z')$ is $\ge$ than D. Say $d(x,z')\ge D$, so that it holds

$$D\le d(x,z')\le d(x,y)+d(y,z')=(1-t_0)D+t_0D=D,$$

which means that the triple of points (x,y,z') lies along a geodesic. Since (x,y,z) lies on a geodesic as well, by the non-branching hypothesis we get a contradiction.

Thus the map supp$(\alpha)\ni(x,y,z)\mapsto y$ is injective. This means that there exists two maps $f,\ g:X\to X$ such that $(x,y,z)\in$ supp(α) if and only if $x=f(y)$ and $z=g(y)$. This is the same as to say that γ^1 is induced by f and γ^2 is induced by g.

To summarize, we proved that given $t_0\in(0,1)$, every optimal plan $\gamma\in Opt(\mu_0,\mu_{t_0})$ is induced by a map from μ_{t_0}. Now we claim that the optimal plan is actually unique. Indeed, if there are two of them induced by two different maps, say f and f', then the plan

$$\frac{1}{2}\big((f, Id)_{\#}\mu_{\mu_{t_0}} + (f', Id)_{\#}\mu_{\mu_{t_0}}\big),$$

would be optimal and not induced by a map.

It remains to prove that $\mathscr{P}_2(X)$ is non branching. Choose $\mu \in \mathscr{P}_2(\mathrm{Geod}(X))$ such that (15) holds, fix $t_0 \in (0,1)$ and let γ be the unique optimal plan in $Opt(\mu_0, \mu_{t_0})$. The thesis will be proved if we show that μ depends only on γ. Observe that from Theorem 3.10 and its proof we know that

$$(\mathrm{e}_0, \mathrm{e}_{t_0})_{\#}\mu \in Opt(\mu_0, \mu_{t_0}),$$

and thus $(\mathrm{e}_0, \mathrm{e}_{t_0})_{\#}\mu = \gamma$. By the non-branching hypothesis we know that $(\mathrm{e}_0, \mathrm{e}_{t_0}) : \mathrm{Geod}(X) \to X^2$ is injective. Thus it invertible on its image: letting F the inverse map, we get

$$\mu = F_{\#}\gamma,$$

and the thesis is proved. □

Theorem 3.10 tells us not only that geodesics exists, but provides also a natural way to "interpolate" optimal plans: once we have the measure $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ satisfying (15), an optimal plan from μ_t to μ_s is simply given by $(\mathrm{e}_t, \mathrm{e}_s)_{\#}\mu$. Now, we know that the transport problem has a natural dual problem, which is solved by the Kantorovich potential. It is then natural to ask how to interpolate potentials. In other words, if (φ, φ^{c+}) are $c-$conjugate Kantorovich potentials for (μ_0, μ_1), is there a simple way to find out a couple of Kantorovich potentials associated to the couple μ_t, μ_s? The answer is yes, and it is given—shortly said—by the solution of an Hamilton–Jacobi equation. To see this, we first define the *Hopf–Lax* evolution semigroup H_t^s (which in $\mathbb{R}^d$ produces the viscosity solution of the Hamilton–Jacobi equation) via the following formula:

$$H_t^s(\psi)(x) := \begin{cases} \inf\limits_{y \in X} \dfrac{d^2(x,y)}{s-t} + \psi(y), & \text{if } t < s, \\[2ex] \psi(x), & \text{if } t = s, \\[2ex] \sup\limits_{y \in X} -\dfrac{d^2(x,y)}{s-t} + \psi(y), & \text{if } t > s, \end{cases} \tag{20}$$

To fully appreciate the mechanisms behind the theory, it is better to introduce the *rescaled costs* $c^{t,s}$ defined by

$$c^{t,s}(x,y) := \frac{d^2(x,y)}{s-t}, \qquad \forall t < s,\ x, y \in X.$$

Observe that for $t < r < s$

$$c^{t,r}(x,y) + c^{r,s}(y,z) \geq c^{t,s}(x,z), \qquad \forall x, y, z \in X,$$

and equality holds if and only if there is a constant speed geodesic $\gamma : [t, s] \to X$ such that $x = \gamma_t$, $y = \gamma_r$, $z = \gamma_s$. The notions of $c_+^{t,s}$ and $c_-^{t,s}$ transforms, convexity/concavity and sub/super-differential are defined as in Sect. 2.2, Definitions 2.8–2.10.

The basic properties of the Hopf–Lax formula are collected in the following proposition:

Proposition 3.17 (Basic properties of the Hopf–Lax formula). *We have the following three properties:*

(i) *For any $t, s \in [0, 1]$ the map H_t^s is order preserving, that is $\phi \le \psi \Rightarrow H_t^s(\phi) \le H_t^s(\psi)$.*
(ii) *For any $t < s \in [0, 1]$ it holds*

$$H_s^t\big(H_t^s(\phi)\big) = \phi^{c_-^{t,s} c_-^{t,s}} \le \phi,$$
$$H_t^s\big(H_s^t(\phi)\big) = \phi^{c_+^{t,s} c_+^{t,s}} \ge \phi.$$

(iii) *For any $t, s \in [0, 1]$ it holds*

$$H_t^s \circ H_s^t \circ H_t^s = H_t^s.$$

Proof. The order preserving property is a straightforward consequence of the definition. To prove property (ii) observe that

$$H_s^t\big(H_t^s(\phi)\big)(x) = \sup_y \inf_{x'} \Big(\phi(x') + c^{t,s}(x', y) - c^{t,s}(x, y)\Big),$$

which gives the equality $H_s^t\big(H_t^s(\phi)\big) = \phi^{c_-^{t,s} c_-^{t,s}}$: in particular, choosing $x' = x$ we get the claim (the proof of the other equation is similar). For the last property assume $t < s$ (the other case is similar) and observe that by (i) we have

$$\underbrace{H_t^s \circ H_s^t}_{\ge Id} \circ H_t^s \ge H_t^s$$

and

$$H_t^s \circ \underbrace{H_s^t \circ H_t^s}_{\le Id} \le H_t^s.$$

□

The fact that Kantorovich potentials evolve according to the Hopf–Lax formula is expressed in the following theorem. We remark that in the statement below one must deal at the same time with c-concave and c-convex potentials.

Theorem 3.18 (Interpolation of potentials). *Let (X, d) be a Polish geodesic space, $(\mu_t) \subset \mathscr{P}_2(X)$ a constant speed geodesic in $(\mathscr{P}_2(X), W_2)$ and φ a $c = c^{0,1}$-*convex *Kantorovich potential for the couple (μ_0, μ_1). Then the function $\varphi_s := H_0^s(\varphi)$ is a $c^{t,s}$-*concave *Kantorovich potential for the couple (μ_s, μ_t), for any $t < s$.*

*Similarly, if ϕ is a c-*concave *Kantorovich potential for (μ_1, μ_0), then $H_1^t(\phi)$ is a $c^{t,s}$-*convex *Kantorovich potential for (μ_t, μ_s) for any $t < s$.*

Observe that for $t = 0, s = 1$ the theorem reduces to the fact that $H_0^1(\varphi) = (-\varphi)^{c+}$ is a c-concave Kantorovich potential for μ_1, μ_0, a fact that was already clear by the symmetry of the dual problem discussed in Sect. 2.3.

Proof. We will prove only the first part of the statement, as the second is analogous.

Step 1. We prove that $H_0^s(\psi)$ is a $c^{t,s}$-concave function for any $t < s$ and any $\psi : X \to \mathbb{R} \cup \{+\infty\}$. This is a consequence of the equality

$$c^{0,s}(x, y) = \inf_{z \in X} c^{0,t}(z, y) + c^{t,s}(x, z),$$

from which it follows

$$H_0^s(\psi)(x) = \inf_{y \in X} c^{0,s}(x, y) + \psi(y) = \inf_{z \in X} c^{t,s}(x, z) + \left(\inf_{y \in X} c^{0,t}(z, y) + \psi(y) \right).$$

Step 2. Let $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ be a measure associated to the geodesic (μ_t) via (15). We claim that for every $\gamma \in \mathrm{supp}(\mu)$ and $s \in (0, 1]$ it holds

$$\varphi_s(\gamma_s) = \varphi(\gamma_0) + c^{0,s}(\gamma_0, \gamma_s). \tag{21}$$

Indeed the inequality $\leq$ comes directly from the definition by taking $x = \gamma_0$. To prove the opposite one, observe that since $(\mathrm{e}_0, \mathrm{e}_1)_\# \mu \in \mathit{Opt}(\mu_0, \mu_1)$ and φ is a c-convex Kantorovich potential for μ_0, μ_1, we have from Theorem 2.13 that

$$\varphi^{c-}(\gamma_1) = -c^{0,1}(\gamma_0, \gamma_1) - \varphi(\gamma_0),$$

thus

$$\begin{aligned} \varphi(x) &= \sup_{y \in X} -c^{0,1}(x, y) - \varphi^{c-}(y) \geq -c^{0,1}(x, \gamma_1) - \varphi^{c-}(\gamma_1) \\ &= -c^{0,1}(x, \gamma_1) + c^{0,1}(\gamma_0, \gamma_1) + \varphi(\gamma_0). \end{aligned}$$

Plugging this inequality in the definition of φ_s we get

$$\begin{aligned} \varphi_s(\gamma_s) &= \inf_{x \in X} c^{0,s}(x, \gamma_s) + \varphi(x) \\ &\geq \inf_{x \in X} c^{0,s}(x, \gamma_s) - c^{0,1}(x, \gamma_1) + c^{0,1}(\gamma_0, \gamma_1) + \varphi(\gamma_0) \\ &\geq -c^{s,1}(\gamma_s, \gamma_1) + c^{0,1}(\gamma_0, \gamma_1) - \varphi(\gamma_0) = c^{0,s}(\gamma_0, \gamma_s) + \varphi(\gamma_0). \end{aligned}$$

Step 3. We know that an optimal transport plan from μ_t to μ_s is given by $(\mathrm{e}_t, \mathrm{e}_s)_\# \mu$, thus to conclude the proof we need to show that

$$\varphi_s(\gamma_s) + (\varphi_s)^{c_+^{t,s}}(\gamma_t) = c^{t,s}(\gamma_t, \gamma_s), \qquad \forall \gamma \in \mathrm{supp}(\mu),$$

where $(\varphi_s)^{c_+^{t,s}}$ is the $c^{t,s}$-conjugate of the $c^{t,s}$-concave function φ_s. The inequality $\leq$ follows from the definition of $c^{t,s}$-conjugate. To prove opposite inequality start observing that

$$\begin{aligned}\varphi_s(y) &= \inf_{x\in X} c^{0,s}(x,y)+\varphi(y) \le c^{0,s}(\gamma_0,y)+\varphi(\gamma_0)\\ &\le c^{0,t}(\gamma_0,\gamma_t)+c^{t,s}(\gamma_t,y)+\varphi(\gamma_0),\end{aligned}$$

and conclude by

$$\begin{aligned}\varphi_s^{c^{t,s}_+}(\gamma_t) &= \inf_{y\in X} c^{t,s}(\gamma_t,y)-\varphi_s(y) \ge -c^{0,t}(\gamma_0,\gamma_t)-\varphi(\gamma_0)\\ &= -c^{0,s}(\gamma_0,\gamma_s)+c^{t,s}(\gamma_t,\gamma_s)-\varphi(\gamma_0)\\ &\overset{(21)}{=} c^{t,s}(\gamma_t,\gamma_s)-\varphi_s(\gamma_s).\end{aligned}$$

□

We conclude the section studying some curvature properties of $(\mathscr{P}_2(X), W_2)$. We will focus on spaces *positively/non positively curved* in the sense of Alexandrov, which are the non smooth analogous of Riemannian manifolds having sectional curvature bounded from below/above by 0.

Definition 3.19 (PC and NPC spaces). A geodesic space (X,d) is said to be positively curved (PC) in the sense of Alexandrov if for every constant speed geodesic $\gamma:[0,1]\to X$ and every $z\in X$ the following concavity inequality holds:

$$d^2\big(\gamma_t,z\big) \ge (1-t)d^2\big(\gamma_0,z\big)+td^2\big(\gamma_1,z\big)-t(1-t)d^2\big(\gamma_0,\gamma_1\big). \tag{22}$$

Similarly, X is said to be non positively curved (NPC) in the sense of Alexandrov if the converse inequality always holds.

Observe that in an Hilbert space equality holds in (22).

The result here is that $(\mathscr{P}_2(X), W_2)$ is PC if (X,d) is, while in general it is not NPC if X is.

Theorem 3.20 ($(\mathscr{P}_2(X), W_2)$ is PC if (X,d) is). *Assume that (X,d) is positively curved. Then $(\mathscr{P}_2(X), W_2)$ is positively curved as well.*

Proof. Let (μ_t) be a constant speed geodesic in $\mathscr{P}_2(X)$ and $\nu\in\mathscr{P}_2(X)$. Let $\mu\in\mathscr{P}_2(\mathrm{Geod}(X))$ be a measure such that

$$\mu_t = (\mathrm{e}_t)_\#\mu, \qquad \forall t\in[0,1],$$

as in Theorem 3.10. Fix $t_0\in[0,1]$ and choose $\gamma\in Opt(\mu_{t_0},\nu)$. Using a gluing argument (we omit the details) it is possible to show the existence a measure $\alpha\in\mathscr{P}(\mathrm{Geod}(X)\times X)$ such that

$$\begin{aligned}\pi^{\mathrm{Geod}(X)}_\#\alpha &= \mu,\\ \big(\mathrm{e}_{t_0},\pi^X\big)_\#\alpha &= \gamma,\end{aligned} \tag{23}$$

where $\pi^{\mathrm{Geod}(X)}(\gamma, x) := \gamma \in \mathrm{Geod}(X)$, $\pi^X(\gamma, x) := x \in X$ and $\mathrm{e}_{t_0}(\gamma, x) := \gamma_{t_0} \in X$. Then α satisfies also

$$\begin{aligned} \left(\mathrm{e}_0, \pi^X\right)_{\#}\alpha &\in \mathcal{A}dm(\mu_0, \nu) \\ \left(\mathrm{e}_1, \pi^X\right)_{\#}\alpha &\in \mathcal{A}dm(\mu_1, \nu), \end{aligned} \tag{24}$$

and therefore it holds

$$\begin{aligned} W_2^2(\mu_{t_0}, \nu) &= \int d^2(\mathrm{e}_{t_0}(\gamma), x) d\alpha(\gamma, x) \\ &\overset{(22)}{\geq} \int (1-t_0) d^2\big(\gamma_0, z\big) + t_0 d^2\big(\gamma_1, z\big) - t_0(1-t_0) d^2\big(\gamma_0, \gamma_1\big) d\alpha(\gamma, x) \\ &\overset{(23)}{=} (1-t_0) \int d^2\big(\gamma_0, z\big) d\alpha(\gamma, x) + t_0 \int d^2\big(\gamma_1, z\big) d\alpha(\gamma, x) \\ &\quad - t_0(1-t_0) \int d^2\big(\gamma_0, \gamma_1\big) d\mu(\gamma) \\ &\overset{(24)}{\geq} (1-t_0) W_2^2(\mu_0, \nu) + t_0 W_2^2(\mu_1, \nu) - t_0(1-t_0) W_2^2(\mu_0, \mu_1), \end{aligned}$$

and by the arbitrariness of t_0 we conclude. □

Example 3.21 ($(\mathscr{P}_2(X), W_2)$ may be not NPC if (X, d) is). Let $X = \mathbb{R}^2$ with the Euclidean distance. We will prove that $(\mathscr{P}_2(\mathbb{R}^2), W_2)$ is not NPC. Define

$$\mu_0 := \frac{1}{2}(\delta_{(1,1)} + \delta_{(5,3)}), \quad \mu_1 := \frac{1}{2}(\delta_{(-1,1)} + \delta_{(-5,3)}), \quad \nu := \frac{1}{2}(\delta_{(0,0)} + \delta_{(0,-4)}),$$

then explicit computations show that $W_2^2(\mu_0, \mu_1) = 40$ and $W_2^2(\mu_0, \nu) = 30 = W_2^2(\mu_1, \nu)$. The unique constant speed geodesic (μ_t) from μ_0 to μ_1 is given by

$$\mu_t = \frac{1}{2}\big(\delta_{(1-6t, 1+2t)} + \delta_{(5-6t, 3-2t)}\big),$$

and simple computations show that

$$24 = W_2^2(\mu_{1/2}, \nu) > \frac{30}{2} + \frac{30}{2} - \frac{40}{4}.$$

■

3.3 X Riemannian Manifold

In this section X will always be a compact, smooth Riemannian manifold M without boundary, endowed with the Riemannian distance d.

We study two aspects: the first one is the analysis of some important consequences of Theorem 3.18 about the structure of geodesics in $\mathscr{P}_2(M)$, the second one is the introduction of the so called *weak Riemannian structure* of $(\mathscr{P}_2(M), W_2)$.

Notice that since M is compact, $\mathscr{P}_2(M) = \mathscr{P}(M)$. Yet, we stick to the notation $\mathscr{P}_2(M)$ because all the statements we make in this section are true also for non compact manifolds (although, for simplicity, we prove them only in the compact case).

3.3.1 Regularity of Interpolated Potentials and Consequences

We start observing how Theorem 3.10 specializes to the case of Riemannian manifolds:

Corollary 3.22 (Geodesics in $(\mathscr{P}_2(M), W_2)$). *Let $(\mu_t) \subset \mathscr{P}_2(M)$. Then the following two things are equivalent:*

(i) *(μ_t) is a geodesic in $(\mathscr{P}_2(M), W_2)$.*
(ii) *There exists a plan $\gamma \in \mathscr{P}(TM)$ (TM being the tangent bundle of M) such that*

$$\begin{aligned} \int |\mathrm{v}|^2 d\gamma(x,\mathrm{v}) &= W_2^2(\mu_0,\mu_1), \\ \big(\mathrm{Exp}(t)\big)_\# \gamma &= \mu_t, \end{aligned} \tag{25}$$

$\mathrm{Exp}(t): TM \to M$ *being defined by* $(x,\mathrm{v}) \mapsto \exp_x(t\mathrm{v})$.

Also, for any $\mu, \nu \in \mathscr{P}_2(M)$ such that μ is a regular measure (Definition 2.32), the geodesic connecting μ to ν is unique.

Notice that we cannot substitute the first equation in (25) with $(\pi^M, \exp)_\# \gamma \in Opt(\mu_0,\mu_1)$, because this latter condition is strictly weaker (it may be that the curve $t \mapsto \exp_x(t\mathrm{v})$ is not a globally minimizing geodesic from x to $\exp_x(\mathrm{v})$ for some $(x,\mathrm{v}) \in \operatorname{supp}\gamma$).

Proof. The implication $(i) \Rightarrow (ii)$ follows directly from Theorem 3.10 by taking into account the fact that $t \mapsto \gamma_t$ is a constant speed geodesic on M implies that for some $(x, \mathrm{v} \in TM)$ it holds $\gamma_t = \exp_x(t\mathrm{v})$ and in this case $d(\gamma_0,\gamma_1) = |\mathrm{v}|$.

For the converse implication, just observe that from the second equation in (25) we have

$$\begin{aligned} W_2^2(\mu_t,\mu_s) &\le \int d^2\big(\exp_x(t\mathrm{v}), \exp_x(s\mathrm{v})\big) d\gamma(x,\mathrm{v}) \\ &\le (t-s)^2 \int |\mathrm{v}|^2 d\gamma(x,\mathrm{v}) = (t-s)^2 W_2^2(\mu_0,\mu_1), \end{aligned}$$

having used the first equation in (25) in the last step.

To prove the last claim just recall that by Remark 2.35 we know that for μ-a.e. x there exists a unique geodesic connecting x to $T(x)$, T being the optimal transport map. Hence the conclusion follows from (ii) of Theorem 3.10. □

Now we discuss the regularity properties of Kantorovich potentials which follows from Theorem 3.18.

Corollary 3.23 (Regularity properties of the interpolated potentials). *Let ψ be a $c-$convex potential for (μ_0, μ_1) and let $\varphi := H_0^1(\psi)$. Define $\psi_t := H_0^t(\psi)$, $\varphi_t := H_1^t(\varphi)$ and choose a geodesic (μ_t) from μ_0 to μ_1. Then for every $t \in (0, 1)$ it holds:*

(i) $\psi_t \geq \varphi_t$ *and both the functions are real valued.*
(ii) $\psi_t = \varphi_t$ *on* $\operatorname{supp}(\mu_t)$.
(iii) ψ_t *and* φ_t *are differentiable in the support of* μ_t *and on this set their gradients coincide.*

Proof. For (i) we have

$$\varphi_t = H_1^t(\varphi) = (H_1^t \circ H_0^1)(\psi) = (\underbrace{H_1^t \circ H_t^1}_{\leq Id} \circ H_0^t)\psi \leq H_0^t(\psi) = \psi_t.$$

Now observe that by definition, $\psi_t(x) < +\infty$ and $\varphi_t(x) > -\infty$ for every $x \in M$, thus it holds

$$+\infty > \psi_t(x) \geq \varphi_t(x) > -\infty, \qquad \forall x \in M.$$

To prove (ii), let μ be the unique plan associated to the geodesic (μ_t) via (15) (recall Proposition 3.16 for uniqueness) and pick $\gamma \in \operatorname{supp}(\mu)$. Recall that it holds

$$\psi_t(\gamma_t) = c^{0,t}(\gamma_0, \gamma_t) + \psi(\gamma_0),$$
$$\varphi_t(\gamma_t) = c^{t,1}(\gamma_t, \gamma_1) + \varphi(\gamma_1).$$

Thus from $\varphi(\gamma_1) = c^{0,1}(\gamma_0, \gamma_1) + \psi(\gamma_0)$ we get that $\psi_t(\gamma_t) = \varphi_t(\gamma_t)$. Since $\mu_t = (\mathrm{e}_t)_\# \mu$, the compactness of M gives $\operatorname{supp}(\mu_t) = \{\gamma_t\}_{\gamma \in \operatorname{supp}(\mu)}$, so that (ii) follows.

Now we turn to (iii). With the same choice of $t \mapsto \gamma_t$ as above, recall that it holds

$$\psi_t(\gamma_t) = c^{0,t}(\gamma_0, \gamma_t) + \psi(\gamma_0)$$
$$\psi_t(x) \leq c^{0,t}(\gamma_0, x) + \psi(\gamma_0), \qquad \forall x \in M,$$

and that the function $x \mapsto c^{0,t}(\gamma_0, x) + \psi(\gamma_0)$ is superdifferentiable at $x = \gamma_t$. Thus the function $x \mapsto \psi_t$ is superdifferentiable at $x = \gamma_t$. Similarly, φ_t is subdifferentiable at γ_t. Choose $v_1 \in \partial^+ \psi_t(\gamma_t)$, $v_2 \in \partial^- \varphi_t(\gamma_t)$ and observe that

$$\begin{aligned}\psi_t(\gamma_t) + \left\langle v_1, \exp_{\gamma_t}^{-1}(x)\right\rangle + o(D(x, \gamma_t)) &\geq \psi_t(x)\\ &\geq \varphi_t(x)\\ &\geq \varphi_t(\gamma_t) + \left\langle v_2, \exp_{\gamma_t}^{-1}(x)\right\rangle + o(D(x, \gamma_t)),\end{aligned}$$

which gives $v_1 = v_2$ and the thesis. □

Corollary 3.24 (The intermediate transport maps are locally Lipschitz). *Let $(\mu_t) \subset \mathscr{P}_2(M)$ a constant speed geodesic in $(\mathscr{P}_2(M), W_2)$. Then for every $t \in (0, 1)$ and $s \in [0, 1]$ there exists only one optimal transport plan from μ_t to μ_s, this transport plan is induced by a map, and this map is locally Lipschitz.*

Note: clearly in a compact setting being locally Lipschitz means being Lipschitz. We wrote "locally" because this is the regularity of transport maps in the non compact situation.

Proof. Fix $t \in (0,1)$ and, without loss of generality, let $s = 1$. The fact that the optimal plan from is unique and induced by a map is known by Proposition 3.16. Now let v be the vector field defined on $\operatorname{supp}(\mu_t)$ by $v(x) = \nabla\varphi_t = \nabla\psi_t$ (we are using part (*iii*) of the above corollary, with the same notation). The fact that ψ_t is a $c^{0,t}$-concave potential for the couple μ_t, μ_0 tells that the optimal transport map T satisfies $T(x) \in \partial^{c^{0,t}+}\phi_t(x)$ for μ_t-a.e. x. Using Theorem 2.33, the fact that ψ_t is differentiable in $\operatorname{supp}(\mu_t)$ and taking into account the scaling properties of the cost, we get that T may be written as $T(x) = \exp_x -v(x)$. Since the exponential map is C^∞, the fact that T is Lipschitz will follow if we show that the vector field v on $\operatorname{supp}(\mu_t)$ is, when read in charts, Lipschitz.

Thus, passing to local coordinates and recalling that $d^2(\cdot, y)$ is uniformly semiconcave, the situation is the following. We have a semiconcave function $f : \mathbb{R}^d \to \mathbb{R}$ and a semiconvex function $g : \mathbb{R}^d \to \mathbb{R}$ such that $f \geq g$ on $\mathbb{R}^d$, $f = g$ on a certain closed set K and we have to prove that the vector field $u : K \to \mathbb{R}^d$ defined by $u(x) = \nabla f(x) = \nabla g(x)$ is Lipschitz. Up to rescaling we may assume that f and g are such that $f - |\cdot|^2$ is concave and $g + |\cdot|^2$ is convex. Then for every $x \in K$ and $y \in \mathbb{R}^d$ we have

$$\langle u(x), y - x\rangle - |x-y|^2 \leq g(y) - g(x) \leq f(y) - f(x) \leq \langle u(x), y - x\rangle + |y-x|^2,$$

and thus for every $x \in K$, $y \in \mathbb{R}^d$ it holds

$$|\, f(y) - f(x) - \langle u(x), y - x\rangle \,| \leq |x - y|^2.$$

Picking $x_1, x_2 \in K$ and $y \in \mathbb{R}^d$ we have

$$\begin{gathered} f(x_2) - f(x_1) - \langle u(x_1), x_2 - x_1\rangle \leq |x_1 - x_2|^2, \\ f(x_2 + y) - f(x_2) - \langle u(x_2), y\rangle \leq |y|^2, \\ -f(x_2 + y) + f(x_1) + \langle u(x_1), x_2 + y - x_1\rangle \leq |x_2 + y - x_1|^2. \end{gathered}$$

Adding up we get

$$\langle u(x_1) - u(x_2), y\rangle \leq |x_1 - x_2|^2 + |y|^2 + |x_2 + y - x_1|^2 \leq 3(|x_1 - x_2|^2 + |y|^2).$$

Eventually, choosing $y = (u(x_1) - u(x_2))/6$ we obtain

$$|u(x_1) - u(x_2)|^2 \leq 36|x_1 - x_2|^2.$$

□

It is worth stressing the fact that the regularity property ensured by the previous corollary holds without any assumption on the measures μ_0, μ_1.

Remark 3.25 (A (much) simpler proof in the Euclidean case). The fact that intermediate transport maps are Lipschitz can be proved, in the Euclidean case, via the theory of monotone operators. Indeed if $G : \mathbb{R}^d \to \mathbb{R}^d$ is a—possibly multivalued—monotone map (i.e. satisfies $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$ for every $x_1, x_2 \in \mathbb{R}^d$, $y_i \in G(x_i)$, $i = 1, 2$), then the operator $((1-t)Id + tG)^{-1}$ is single valued, Lipschitz, with Lipschitz constant bounded above by $1/(1-t)$. To prove this, pick $x_1, x_2 \in \mathbb{R}^d$, $y_1 \in G(x_1)$, $y_2 \in G(x_2)$ and observe that

$$\begin{aligned}
&|(1-t)x_1 + ty_1 - (1-t)x_2 + ty_2|^2 \\
&\quad = (1-t)^2|x_1 - x_2|^2 + t^2|y_1 - y_2|^2 + 2t(1-t)\langle x_1 - x_2, y_1 - y_2\rangle \\
&\quad \geq (1-t)^2|x_1 - x_2|^2,
\end{aligned}$$

which is our claim.

Now pick $\mu_0, \mu_1 \in \mathscr{P}_2(\mathbb{R}^d)$, an optimal plan $\gamma \in Opt(\mu_0, \mu_1)$ and consider the geodesic $t \mapsto \mu_t := ((1-t)\pi^1 + t\pi^2)_\#\gamma$ (recall Remark 3.13). From Theorem 2.26 we know that there exists a convex function φ such that $\mathrm{supp}(\gamma) \subset \partial^-\varphi$. Also, we know that the unique optimal plan from μ_0 to μ_t is given by the formula

$$\big(\pi^1, (1-t)\pi^1 + t\pi^2\big)_\#\gamma,$$

which is therefore supported in the graph of $(1-t)Id + t\partial^-\varphi$. Since the subdifferential of a convex function is a monotone operator, the thesis follows from the previous claim.

Considering the case in which μ_1 is a delta and μ_0 is not, we can easily see that the bound $(1-t)^{-1}$ on the Lipschitz constant of the optimal transport map from μ_t to μ_0 is sharp. ■

An important consequence of Corollary 3.24 is the following proposition:

Proposition 3.26 (Geodesic convexity of the set of absolutely continuous measures). *Let M be a Riemannian manifold, $(\mu_t) \subset \mathscr{P}_2(M)$ a geodesic and assume that μ_0 is absolutely continuous w.r.t. the volume measure (resp. gives 0 mass to Lipschitz hypersurfaces of codimension 1). Then μ_t is absolutely continuous w.r.t. the volume measure (resp. gives 0 mass to Lipschitz hypersurfaces of codimension 1) for every $t < 1$. In particular, the set of absolutely continuous measures is geodesically convex (and the same for measures giving 0 mass to Lipschitz hypersurfaces of codimension 1).*

Proof. Assume that μ_0 is absolutely continuous, let $A \subset M$ be of 0 volume measure, $t \in (0, 1)$ and let T_t be the optimal transport map from μ_t to μ_0. Then for every Borel set $A \subset M$ it holds $T_t^{-1}(T_t(A)) \supset A$ and thus

$$\mu_t(A) \leq \mu_t(T_t^{-1}(T_t(A))) = \mu_0(T_t(A)).$$

The claims follow from the fact that T_t is locally Lipschitz. □

Remark 3.27 (The set of regular measures is not *geodesically convex).* It is natural to ask whether the same conclusion of the previous proposition holds for the set of regular measures (Definitions 2.25 and 2.32). The answer is *not*: there are examples of regular measures μ_0, μ_1 in $\mathscr{P}_2(\mathbb{R}^2)$ such that the middle point of the geodesic connecting them is not regular. ■

3.3.2 The Weak Riemannian Structure of $(\mathscr{P}_2(M), W_2)$

In order to introduce the weak differentiable structure of $(\mathscr{P}_2(X), W_2)$, we start with some heuristic considerations. Let $X = \mathbb{R}^d$ and (μ_t) be a constant speed geodesic on $\mathscr{P}_2(\mathbb{R}^d)$ induced by some optimal map T, i.e.:

$$\mu_t = \big((1-t)Id + tT\big)_\# \mu_0.$$

Then a simple calculation shows that (μ_t) satisfies the continuity equation

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0,$$

with $v_t := (T - Id) \circ ((1-t)Id + tT)^{-1}$ for every t, in the sense of distributions. Indeed for $\phi \in C_c^\infty(\mathbb{R}^d)$ it holds

$$\begin{aligned}\frac{d}{dt}\int \phi d\mu_t &= \frac{d}{dt}\int \phi\big((1-t)Id + tT\big)d\mu_0\\ &= \int \langle \nabla\phi\big((1-t)Id + tT\big), T - Id\rangle\, d\mu_0 = \int \langle \nabla\phi, v_t\rangle d\mu_t.\end{aligned}$$

Now, the continuity equation describes the link between the motion of the continuum μ_t and the instantaneous velocity $v_t : \mathbb{R}^d \to \mathbb{R}^d$ of every "atom" of μ_t. It is therefore natural to think at the vector field v_t as the infinitesimal variation of the continuum μ_t.

From this perspective, one might expect that the set of "smooth" curves on $\mathscr{P}_2(\mathbb{R}^d)$ (and more generally on $\mathscr{P}_2(M)$) is somehow linked to the set of solutions of the continuity equation. This is actually the case, as we are going to discuss now.

In order to state the rigorous result, we need to recall the definition of *absolutely continuous curve* on a metric space.

Definition 3.28 (Absolutely continuous curve). Let $(Y, \tilde{d})$ be a metric space and let $[0,1] \ni t \mapsto y_t \in Y$ be a curve. Then (y_t) is said absolutely continuous if there exists a function $f \in L^1(0,1)$ such that

$$\tilde{d}(y_t, y_s) \le \int_t^s f(r)dr, \qquad \forall t < s \in [0,1]. \tag{26}$$

We recall that if (y_t) is absolutely continuous, then for a.e. t the *metric derivative* $|\dot{y}_t|$ exists, given by

$$|\dot{y}_t| := \lim_{h\to 0} \frac{\tilde{d}(y_{t+h}, y_t)}{|h|}, \tag{27}$$

and that $|\dot{y}_t| \in L^1(0,1)$ and is the smallest L^1 function (up to negligible sets) for which inequality (26) is satisfied (see e.g. Theorem 1.1.2 of [7] for the simple proof).

The link between absolutely continuous curves in $\mathscr{P}_2(M)$ and the continuity equation is given by the following theorem:

Theorem 3.29 (Characterization of absolutely continuous curves in $(\mathscr{P}_2(M), W_2)$). *Let M be a smooth complete Riemannian manifold without boundary. Then the following holds.*

*(**A**) For every absolutely continuous curve $(\mu_t) \subset \mathscr{P}_2(M)$ there exists a Borel family of vector fields v_t on M such that $\|v_t\|_{L^2(\mu_t)} \leq |\dot{\mu}_t|$ for a.e. t and the continuity equation*

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{28}$$

holds in the sense of distributions.

*(**B**) If (μ_t, v_t) satisfies the continuity equation (28) in the sense of distributions and $\int_0^1 \|v_t\|_{L^2(\mu_t)} dt < \infty$, then up to redefining $t \mapsto \mu_t$ on a negligible set of times, (μ_t) is an absolutely continuous curve on $\mathscr{P}_2(M)$ and $|\dot{\mu}_t| \leq \|v_t\|_{L^2(\mu_t)}$ for a.e. $t \in [0,1]$.*

Note that we are not assuming any kind of regularity on the μ_t's.

We postpone the (sketch of the) proof of this theorem to the end of the section, for the moment we analyze its consequences in terms of the geometry of $\mathscr{P}_2(M)$.

The first important consequence is that the Wasserstein distance, which was defined via the "static" optimal transport problem, can be recovered via the following "dynamic" Riemannian-like formula:

Proposition 3.30 (Benamou–Brenier formula). *Let $\mu^0, \mu^1 \in \mathscr{P}_2(M)$. Then it holds*

$$W_2(\mu^0, \mu^1) = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t} dt \right\}, \tag{29}$$

where the infimum is taken among all weakly continuous distributional solutions of the continuity equation (μ_t, v_t) such that $\mu_0 = \mu^0$ and $\mu_1 = \mu^1$.

Proof. We start with inequality $\leq$. Let (μ_t, v_t) be a solution of the continuity equation. Then if $\int_0^1 \|v_t\|_{L^2(\mu_t)} = +\infty$ there is nothing to prove. Otherwise we may apply part **B** of Theorem 3.29 to get that (μ_t) is an absolutely continuous curve on $\mathscr{P}_2(M)$. The conclusion follows from

$$W_2(\mu^0, \mu^1) \leq \int_0^1 |\dot{\mu}_t| dt \leq \int_0^1 \|v_t\|_{L^2(\mu_t)} dt,$$

where in the last step we used part (**B**) of Theorem 3.29 again.

To prove the converse inequality it is enough to consider a constant speed geodesic (μ_t) connecting μ^0 to μ^1 and apply part (**A**) of Theorem 3.29 to get the existence of vector fields v_t such that the continuity equation is satisfied and $\|v_t\|_{L^2(\mu_t)} \leq |\dot{\mu}_t| = W_2(\mu^0, \mu^1)$ for a.e. $t \in [0, 1]$. Then we have

$$W_2(\mu^0, \mu^1) \geq \int_0^1 \|v_t\|_{L^2(\mu_t)} dt,$$

as desired. □

This proposition strongly suggests that the scalar product in $L^2(\mu)$ should be considered as the metric tensor on $\mathscr{P}_2(M)$ at μ. Now observe that given an absolutely continuous curve $(\mu_t) \subset \mathscr{P}_2(M)$ in general there is no unique choice of vector field (v_t) such that the continuity equation (28) is satisfied. Indeed, if (28) holds and w_t is a Borel family of vector fields such that $\nabla \cdot (w_t \mu_t) = 0$ for a.e. t, then the continuity equation is satisfied also with the vector fields $(v_t + w_t)$. It is then natural to ask whether there is some natural selection principle to associate uniquely a family of vector fields (v_t) to a given absolutely continuous curve. There are two possible approaches:

Algebraic approach. The fact that for distributional solutions of the continuity equation the vector field v_t acts only on gradients of smooth functions suggests that the v_t's should be taken in the set of gradients as well, or, more rigorously, v_t should belong to

$$\overline{\left\{\nabla\varphi \ : \ \varphi \in C_c^\infty(M)\right\}}^{L^2(\mu_t)} \tag{30}$$

for a.e. $t \in [0, 1]$.

Variational approach. The fact that the continuity equation is linear in v_t and the L^2 norm is strictly convex, implies that there exists a unique, up to negligible sets in time, family of vector fields $v_t \in L^2(\mu_t)$, $t \in [0, 1]$, with minimal norm for a.e. t, among the vector fields compatible with the curve (μ_t) via the continuity equation. In other words, for any other vector field $(\tilde{v}_t)$ compatible with the curve (μ_t) in the sense that (28) is satisfied, it holds $\|\tilde{v}_t\|_{L^2(\mu_t)} \geq \|v_t\|_{L^2(\mu_t)}$ for a.e. t. It is immediate to verify that v_t is of minimal norm if and only if it belongs to the set

$$\left\{v \in L^2(\mu_t) \ : \ \int \langle v, w\rangle \, d\mu_t = 0, \ \forall w \in L^2(\mu_t) \ s.t. \ \nabla \cdot (w\mu_t) = 0\right\}. \tag{31}$$

The important point here is that the sets defined by (30) and (31) are the same, as it is easy to check. Therefore it is natural to give the following

Definition 3.31 (The tangent space). Let $\mu \in \mathscr{P}_2(M)$. Then the tangent space $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)M)$ at $\mathscr{P}_2(M)$ in μ is defined as

$$\begin{aligned}\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)M) :=& \overline{\left\{\nabla\varphi \ : \ \varphi \in C_c^\infty(M)\right\}}^{L^2(\mu)} \\ =& \left\{v \in L^2(\mu) \ : \ \int \langle v, w\rangle \, d\mu = 0, \ \forall w \in L^2(\mu) \ s.t. \ \nabla \cdot (w\mu) = 0\right\}\end{aligned}$$

Thus we now have a definition of tangent space for every $\mu \in \mathscr{P}_2(M)$ and this tangent space is naturally endowed with a scalar product: the one of $L^2(\mu)$. This fact, Theorem 3.29 and Proposition 3.30 are the bases of the so-called weak Riemannian structure of $(\mathscr{P}_2(M), W_2)$.

We now state, without proof, some other properties of $(\mathscr{P}_2(M), W_2)$ which resemble those of a Riemannian manifold. For simplicity, we will deal with the case $M = \mathbb{R}^d$ only and we will assume that the measures we are dealing with are regular (Definition 2.25), but analogous statements hold for general manifolds and general measures.

In the next three propositions (μ_t) is an absolutely continuous curve in $\mathscr{P}_2(\mathbb{R}^d)$ such that μ_t is regular for every t. Also (v_t) is the unique, up to a negligible set of times, family of vector fields such that the continuity equation holds and $v_t \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for a.e. t.

Proposition 3.32 (v_t can be recovered by infinitesimal displacement). *Let (μ_t) and (v_t) as above. Also, let T_t^s be the optimal transport map from μ_t to μ_s (which exists and is unique by Theorem 2.26, due to our assumptions on μ_t). Then for a.e. $t \in [0,1]$ it holds*

$$v_t = \lim_{s\to t} \frac{T_t^s - Id}{s-t},$$

the limit being understood in $L^2(\mu_t)$.

Proposition 3.33 ("Displacement tangency"). *Let (μ_t) and (v_t) as above. Then for a.e. $t \in [0,1]$ it holds*

$$\lim_{h\to 0} \frac{W_2\big(\mu_{t+h}, (Id + hv_t)_\#\mu_t\big)}{h} = 0. \tag{32}$$

Proposition 3.34 (Derivative of the squared distance). *Let (μ_t) and (v_t) as above and $\nu \in \mathscr{P}_2(\mathbb{R}^d)$. Then for a.e. $t \in [0,1]$ it holds*

$$\frac{d}{dt} W_2^2(\mu_t, \nu) = -2 \int \langle v_t, T_t - Id\rangle \, d\mu_t,$$

where T_t is the unique optimal transport map from μ_t to ν (which exists and is unique by Theorem 2.26, due to our assumptions on μ_t).

We conclude the section with a sketch of the proof of Theorem 3.29.

Sketch of the Proof of Theorem 3.29

Reduction to the Euclidean case Suppose we already know the result for the case $\mathbb{R}^d$ and we want to prove it for a compact and smooth manifold M. Use the Nash embedding theorem to get the existence of a smooth map $i : M \to \mathbb{R}^D$ whose differential provides an isometry of T_xM and its image for any $x \in M$. Now notice that the inequality $|i(x) - i(y)| \leq d(x,y)$ valid for any $x, y \in M$ ensures that $W_2(i_\#\mu, i_\#\nu) \leq W_2(\mu,\nu)$ for any $\mu, \nu \in \mathscr{P}_2(M)$. Hence given an absolutely continuous curve $(\mu_t) \subset \mathscr{P}_2(M)$, the curve $(i_\#\mu_t) \subset \mathscr{P}_2(\mathbb{R}^D)$ is absolutely continuous as well, and there exists a family vector fields v_t such that (28) is fulfilled with $i_\# \mu_t$ in place of μ_t and $\|v_t\|_{L^2(i_\#\mu_t)} \leq |\dot{i_\#\mu_t}| \leq |\dot{\mu}_t|$ for a.e. t. Testing the continuity equation with functions constant on $i(M)$ we get that for a.e. t the vector field v_t is tangent to $i(M)$ for $i_\#\mu_t$-a.e. point. Thus the v_t's are the (isometric) image of vector fields on M and part (A) is proved.

Viceversa, let $(\mu_t) \subset \mathscr{P}_2(M)$ be a curve and the v_t's vector fields in M such that $\int_0^1 \|v_t\|_{L^2(\mu_t)} dt < \infty$ and assume that they satisfy the continuity equation. Then the measures $\tilde{\mu}_t := i_\#\mu_t$ and the vector fields $\tilde{v}_t := di(v_t)$ satisfy the continuity equation on $\mathbb{R}^D$. Therefore $(\tilde{\mu}_t)$ is an absolutely continuous curve and it holds $|\dot{\tilde{\mu}}_t| \leq \|\tilde{v}_t\|_{L^2(\tilde{\mu}_t)} = \|v_t\|_{L^2(\mu_t)}$ for a.e. t. Notice that i is bilipschitz and therefore (μ_t) is absolutely continuous as well. Hence to conclude it is sufficient to show that $|\dot{\tilde{\mu}}_t| = |\dot{\mu}_t|$ a.e. t. To prove this, one can notice that the fact that i is bilipschitz and validity of

$$\lim_{r\to 0} \sup_{\substack{x,y\in M\\ d(x,y)<r}} \frac{d(x,y)}{|i(x)-i(y)|} = 1,$$

give that

$$\lim_{r\to 0} \sup_{\substack{\mu,\nu\in \mathscr{P}_2(M)\\ W_2(\mu,\nu)<r}} \frac{W_2(\mu,\nu)}{W_2(i_\#\mu, i_\#\nu)} = 1.$$

We omit the details.

Part A. Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$ and observe that for every $\gamma_t^s \in \mathcal{O}pt(\mu_t,\mu_s)$ it holds

$$\begin{aligned}\left|\int \varphi d\mu_s - \int \varphi d\mu_t\right| &= \left|\int \varphi(y) d\gamma_t^s(x,y) - \int \varphi(x) d\gamma_t^s(x,y)\right| \\ &= \left|\int \varphi(y) - \varphi(x) d\gamma_t^s(x,y)\right| \\ &= \left|\int\int_0^1 \langle \nabla\varphi(x+\lambda(y-x)), y-x\rangle \, d\lambda d\gamma_t^s(x,y)\right| \\ &= \left|\int \langle \nabla\varphi(x), y-x\rangle \, d\gamma_t^s(x,y)\right| + \mathrm{Rem}(\varphi,t,s) \\ &\leq \sqrt{\int |\nabla\varphi(x)|^2 d\gamma_t^s(x,y)}\sqrt{\int |x-y|^2 d\gamma_t^s(x,y)} + \mathrm{Rem}(\varphi,t,s) \\ &= \|\nabla\varphi\|_{L^2(\mu_t)} W_2(\mu_t,\mu_s) + \mathrm{Rem}(\varphi,t,s),\end{aligned} \tag{33}$$

where the remainder term $\mathrm{Rem}(\varphi,t,s)$ can be bounded by

$$|\mathrm{Rem}(\varphi,t,s)| \leq \frac{\mathrm{Lip}(\nabla\varphi)}{2}\int |x-y|^2 d\gamma_t^s(x,y) = \frac{\mathrm{Lip}(\nabla\varphi)}{2} W_2^2(\mu_t,\mu_s).$$

Thus (33) implies that the map $t \mapsto \int \varphi d\mu_t$ is absolutely continuous for any $\varphi \in C_c^\infty(\mathbb{R}^d)$.

Now let $D \subset C_c^\infty(\mathbb{R}^d)$ be a countable set such that $\{\nabla\varphi : \varphi \in D\}$ is dense in $\mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for every $t \in [0,1]$ (the existence of such D follows from the compactness of $\{\mu_t\}_{t\in[0,1]} \subset \mathscr{P}_2(\mathbb{R}^d)$, we omit the details). The above arguments imply that there exists a set $A \subset [0,1]$ of full Lebesgue measure such that $t \mapsto \int \varphi d\mu_t$ is differentiable at $t \in A$ for every $\varphi \in D$; we can also assume that the metric derivative $|\dot{\mu}_t|$ exists for every $t \in A$. Also, by (33) we know that for $t_0 \in A$ the linear functional $L_{t_0} : \{\nabla\varphi : \varphi \in D\} \to \mathbb{R}$ given by

$$\nabla\varphi \mapsto L_{t_0}(\nabla\varphi) := \frac{d}{dt}|_{t=t_0} \int \varphi d\mu_t$$

satisfies

$$|L_{t_0}(\nabla\varphi)| \le \|\nabla\varphi\|_{L^2(\mu_{t_0})} |\dot{\mu}_{t_0}|,$$

and thus it can be uniquely extended to a linear and bounded functional on $\mathrm{Tan}_{\mu_{t_0}}(\mathscr{P}_2(\mathbb{R}^d))$. By the Riesz representation theorem there exists a vector field $v_{t_0} \in \mathrm{Tan}_{\mu_{t_0}}(\mathscr{P}_2(\mathbb{R}^d))$ such that

$$\frac{d}{dt}|_{t=t_0} \int \varphi d\mu_t = L_{t_0}(\nabla\varphi) = \int \langle \nabla\varphi, v_{t_0}\rangle \, d\mu_{t_0}, \qquad \forall \varphi \in D, \tag{34}$$

and whose norm in $L^2(\mu_{t_0})$ is bounded above by the metric derivative $|\dot{\mu}_t|$ at $t = t_0$. It remains to prove that the continuity equation is satisfied in the sense of distributions. This is a consequence of (34), see Theorem 8.3.1 of [7] for the technical details.

Part B. Up to a time reparametrization argument, we can assume that $\|v_t\|_{L^2(\mu_t)} \le L$ for some $L \in \mathbb{R}$ for a.e. t. Fix a Gaussian family of mollifiers ρ^ε and define

$$\mu_t^\varepsilon := \mu_t * \rho^\varepsilon,$$
$$v_t^\varepsilon := \frac{(v_t\mu_t) * \rho^\varepsilon}{\mu_t^\varepsilon}.$$

It is clear that

$$\frac{d}{dt}\mu_t^\varepsilon + \nabla \cdot (v_t^\varepsilon \mu_t^\varepsilon) = 0.$$

Moreover, from Jensen inequality applied to the map $(X,z) \mapsto z|X/z|^2 = |X|^2/z$ $(X = v_t\mu_t)$ it follows that

$$\|v_t^\varepsilon\|_{L^2(\mu_t^\varepsilon)} \le \|v_t\|_{L^2(\mu_t)} \le L. \tag{35}$$

This bound, together with the smoothness of v_t^ε, implies that there exists a unique locally Lipschitz map $\mathbf{T}^\varepsilon(\cdot,\cdot) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, $t \in [0,1]$ satisfying

$$\begin{cases} \dfrac{d}{dt}\mathbf{T}^\varepsilon(t,x) = v_t^\varepsilon\big(\mathbf{T}^\varepsilon(t,x)\big) & \forall x \in \mathbb{R}^d,\ a.e.\ t \in [0,1], \\ \mathbf{T}^\varepsilon(t,x) = x, & \forall x \in \mathbb{R}^d,\ t \in [0,1]. \end{cases}$$

A simple computation shows that the curve $t \mapsto \tilde{\mu}_t^\varepsilon := \mathbf{T}^\varepsilon(t,\cdot)_\# \mu_0^\varepsilon$ solves

$$\frac{d}{dt}\tilde{\mu}_t^\varepsilon + \nabla \cdot (v_t^\varepsilon \tilde{\mu}_t^\varepsilon) = 0, \tag{36}$$

which is the same equation solved by (μ_t^ε). It is possible to show that this fact together with the smoothness of the v_t^ε's and the equality $\mu_0^\varepsilon = \tilde{\mu}_0^\varepsilon$ gives that $\tilde{\mu}_t^\varepsilon = \mu_t^\varepsilon$ for every t, ε (see Proposition 8.1.7 and Theorem 8.3.1 of [7] for a proof of this fact).

Conclude observing that

$$\begin{aligned}
W_2^2(\mu_t^\varepsilon, \mu_s^\varepsilon) &\le \int |\mathbf{T}^\varepsilon(t,x) - \mathbf{T}^\varepsilon(s,x)|^2 d\mu_0^\varepsilon(x) = \int \left| \int_t^s v_r^\varepsilon\big(\mathbf{T}^\varepsilon(r,x)\big)\right|^2 d\mu_0^\varepsilon(x) \\
&\le |t-s| \int \int_t^s \left| v_r^\varepsilon\big(\mathbf{T}^\varepsilon(r,x)\big)\right|^2 dr\, d\mu_0^\varepsilon = |t-s| \int_t^s \left\| v_r^\varepsilon\big(\mathbf{T}^\varepsilon(r,\cdot)\big)\right\|^2_{L^2(\mu_0^\varepsilon)} dr \\
&\le |t-s| \int_t^s \|v_r^\varepsilon\|^2_{L^2(\mu_r^\varepsilon)} dr \overset{(35)}{\le} |t-s|^2 L,
\end{aligned}$$

and that, by the characterization of convergence (12), $W_2(\mu_t^\varepsilon, \mu_t) \to 0$ as $\varepsilon \to 0$ for every $t \in [0,1]$. □

3.4 Bibliographical Notes

To call the distance W_2 the "Wasserstein distance" is quite not fair: a much more appropriate would be Kantorovich distance. Also, the spelling "Wasserstein" is questionable, as the original one was "Vasershtein". Yet, this terminology is nowadays so common that it would be impossible to change it.

The equivalence (12) has been proven by the authors and G. Savaré in [7]. In the same reference Remark 3.8 has been first made. The fact that $(\mathscr{P}_2(X), W_2)$ is complete and separable as soon as (X,d) is belongs to the folklore of the theory, a proof can be found in [7]. Proposition 3.4 was proved by C. Villani in [79], Theorem 7.12.

The terminology *displacement interpolation* was introduced by McCann [63] for probability measures in $\mathbb{R}^d$. Theorem 3.10 appears in this form here for the first time: in [58] the theorem was proved in the compact case, in [80] (Theorem 7.21) this has been extended to locally compact structures and much more general forms of interpolation. The main source of difficulty when dealing with general Polish structure is the potential lack of tightness: the proof presented here is strongly inspired by the work of S. Lisini [54].

Proposition 3.16 and Theorem 3.18 come from [80] (Corollary 7.32 and Theorem 7.36 respectively). Theorem 3.20 and the counterexample 3.21 are taken from [7] (Theorem 3.2 and Example 3.3 respectively).

The proof of Corollary 3.24 is taken from an argument by A. Fathi [35], the paper being inspired by Bernand–Buffoni [13]. Remark 3.27 is due to N. Juillet [48].

The idea of looking at the transport problem as dynamical problem involving the continuity equation is due to J.D. Benamou and Y. Brenier [12], while the fact that $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ can be viewed as a sort of infinite dimensional Riemannian manifold is an intuition by F. Otto [67]. Theorem 3.29 has been proven in [7] (where also Propositions 3.32–3.34 were proven) in the case $M = \mathbb{R}^d$, the generalization to Riemannian manifolds comes from Nash's embedding theorem.

4 Gradient Flows

The aim of this Chapter is twofold: on one hand we give an overview of the theory of Gradient Flows in a metric setting, on the other hand we discuss the important application of the abstract theory to the case of geodesically convex functionals on the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$.

Let us recall that for a smooth function $F : M \to \mathbb{R}$ on a Riemannian manifold, a gradient flow (x_t) starting from $\overline{x} \in M$ is a differentiable curve solving

$$\begin{cases} x_t' = -\nabla F(x_t), \\ x_0 = \overline{x}. \end{cases} \tag{37}$$

Observe that there are two necessary ingredients in this definition: the functional F and the metric on M. The role of the functional is clear. The metric is involved to define ∇F: it is used to identify the cotangent vector dF with the tangent vector ∇F.

4.1 Hilbertian Theory of Gradient Flows

In this section we quickly recall the main results of the theory of Gradient flow for λ-convex functionals on Hilbert spaces. This will deserve as guideline for the analysis that we will make later on of the same problem in a purely metric setting.

Let H be Hilbert and $\lambda \in \mathbb{R}$. A λ-convex functional $F : H \to \mathbb{R} \cup \{+\infty\}$ is a functional satisfying:

$$F\big((1-t)x + ty\big) \leq (1-t)F(x) + tF(y) - \frac{\lambda}{2}t(1-t)|x-y|^2, \qquad \forall x, y \in H,$$

(this corresponds to $\nabla^2 F \geq \lambda Id$ for functionals on $\mathbb{R}^d$). We denote with $D(F)$ the domain of F, i.e. $D(F) := \{x : F(x) < \infty\}$.

The subdifferential $\partial^- F(x)$ of F at a point $x \in D(F)$ is the set of $v \in H$ such that

$$F(x) + \langle v, y-x \rangle + \frac{\lambda}{2}|x-y|^2 \leq F(y), \qquad \forall y \in H.$$

An immediate consequence of the definition is the fact that the subdifferential of F satisfies the *monotonicity inequality*:

$$\langle v - w, x - y \rangle \geq \lambda |x - y|^2 \qquad \forall v \in \partial F(x),\ w \in \partial^- F(y).$$

We will denote by $\nabla F(x)$ the element of minimal norm in $\partial F(x)$, which exists and is unique as soon as $\partial^- F(x) \neq \emptyset$, because $\partial^- F(x)$ is closed and convex.

For convex functions a natural generalization of Definition (37) of Gradient Flow is possible: we say that (x_t) is a Gradient Flow for F starting from $\overline{x} \in H$ if it is a locally absolutely continuous curve in $(0, +\infty)$ such that

$$\begin{cases} x_t' \in -\partial^- F(x_t) & \text{for a.e. } t > 0 \\ \lim_{t \downarrow 0} x_t = \overline{x}. \end{cases} \tag{38}$$

We now summarize without proof the main existence and uniqueness results in this context.

Theorem 4.1 (Gradient Flows in Hilbert spaces—(Brezis, Pazy)). *If $F : H \to \mathbb{R} \cup \{+\infty\}$ is λ-convex and lower semicontinuous, then the following statements hold.*

(i) ***Existence and uniqueness*** *for all $\bar{x} \in \overline{D(F)}$ (38) has a unique solution (x_t).*

(ii) ***Minimal selection and Regularizing effects*** *It holds $\frac{d_+}{dt} x_t = -\nabla F(x_t)$ for every $t > 0$ (that is, the right derivative of x_t always exists and realizes the element of minimal norm in $\partial^- F(x_t)$) and $\frac{d_+}{dt} F \circ x(t) = -|\nabla F(x(t))|^2$ for every $t > 0$. Also*

$$F(x_t) \leq \inf_{v \in D(F)} \left\{ F(v) + \frac{1}{2t} |v - \bar{x}|^2 \right\},$$

$$|\nabla F(x_t)|^2 \leq \inf_{v \in D(\partial F)} \left\{ |\nabla F(v)|^2 + \frac{1}{t^2} |v - \bar{x}|^2 \right\}.$$

(iii) ***Energy Dissipation Equality*** *$|x_t'|, |\nabla F|(x_t) \in L^2_{\mathrm{loc}}(0, +\infty)$, $F(x_t) \in AC_{\mathrm{loc}}(0, +\infty)$ and the following Energy Dissipation Equality holds:*

$$F(x_t) - F(x_s) = \frac{1}{2} \int_t^s |\nabla F(x_r)|^2 \, dr + \frac{1}{2} \int_t^s |x_r'|^2 \, dr \qquad 0 < t \leq s < \infty.$$

(iv) ***Evolution Variational Inequality and contraction*** *(x_t) is the unique solution of the system of differential inequalities*

$$\frac{1}{2} \frac{d}{dt} |\tilde{x}_t - y|^2 + F(x_t) + \frac{\lambda}{2} |\tilde{x}_t - y|^2 \leq F(y), \qquad \forall y \in H,\ a.e.\ t,$$

among all locally absolutely continuous curves $(\tilde{x}_t)$ in $(0, \infty)$ converging to $\overline{x}$ as $t \to 0$. Furthermore, if (y_t) is a solution of (38) starting from $\overline{y}$, it holds

$$|x_t - y_t| \leq e^{-\lambda t} |\overline{x} - \overline{y}|.$$

(v) ***Asymptotic behavior*** *If $\lambda > 0$ then there exists a unique minimum $x_{\min}$ of F and it holds*

$$F(x_t) - F(x_{\min}) \leq \big(F(\bar{x}) - F(x_{\min})\big)e^{-2\lambda t}.$$

In particular, the pointwise energy inequality

$$F(x) \geq F(x_{\min}) + \frac{\lambda}{2}|x - x_{\min}|^2, \qquad \forall x \in H$$

gives

$$|x_t - x_{\min}| \leq \sqrt{\frac{2(F(\overline{x}) - F(x_{\min}))}{\lambda}} e^{-\lambda t}.$$

4.2 The Theory of Gradient Flows in a Metric Setting

Here we give an overview of the theory of Gradient Flows in a purely metric framework.

4.2.1 The Framework

The first thing we need to understand is the meaning of Gradient Flow in a metric setting. Indeed, the system (38) makes no sense in metric spaces, thus we need to reformulate it so that it has a metric analogous. There are several ways to do this, below we summarize the most important ones.

For the purpose of the discussion below, we assume that $H = \mathbb{R}^d$ and that $E : H \to \mathbb{R}$ is λ-convex and of class C^1.

Let us start observing that (38) may be written as: $t \mapsto x_t$ is locally absolutely continuous in $(0, +\infty)$, converges to $\overline{x}$ as $t \downarrow 0$ and it holds

$$\frac{d}{dt}E\big(x_t\big) \leq -\frac{1}{2}|\nabla E|^2\big(x_t\big) - \frac{1}{2}|x_t'|^2, \qquad a.e.\ t \geq 0. \tag{39}$$

Indeed, along *any* absolutely continuous curve y_t it holds

$$\begin{aligned}\frac{d}{dt}E\big(y_t\big) &= \big\langle \nabla E(y_t), y_t' \big\rangle \\ &\geq -|\nabla E|(y_t)|y_t'| \ (= \text{ if and only if } -y_t' \text{ is a positive multiple of } \nabla E(y_t)), \\ &\geq -\frac{1}{2}|\nabla E|^2\big(y_t\big) - \frac{1}{2}|y_t'|^2 \qquad (= \text{ if and only if } |y_t'| = |\nabla E(y_t)|).\end{aligned} \tag{40}$$

Thus in particular (39) may be written in the following integral form

$$E\big(x_s\big) + \frac{1}{2}\int_t^s |x_r'|^2 dr + \frac{1}{2}\int_t^s |\nabla E|^2(x_r)dr \le E(x_t), \qquad a.e.\ t < s \tag{41}$$

which we call *Energy Dissipation Inequality* (EDI in the following).

Since the inequality (40) shows that $\frac{d}{dt}E\big(y_t\big) < -\frac{1}{2}|\nabla E|^2\big(y_t\big) - \frac{1}{2}|y_t'|^2$ never holds, the system (38) may be also written in form of *Energy Dissipation Equality* (EDE in the following) as

$$E\big(x_t\big) + \frac{1}{2}\int_t^s |x_r'|^2 dr + \frac{1}{2}\int_t^s |\nabla E|^2(x_r)dr = E(x_t), \qquad \forall 0 \le t \le s. \tag{42}$$

Notice that the convexity of E does not play any role in this formulation.

A completely different way to rewrite (38) comes from observing that if x_t solves (38) and $y \in H$ is a generic point it holds

$$\frac{1}{2}\frac{d}{dt}|x_t - y|^2 = \big\langle x_t - y, x_t'\big\rangle = \langle y - x_t, \nabla E(x_t)\rangle \le E(y) - E(x_t) - \frac{\lambda}{2}|x_t - y|^2,$$

where in the last inequality we used the fact that E is λ-convex. Since the inequality

$$\langle y - x, v\rangle \le E(y) - E(x) - \frac{\lambda}{2}|x - y|^2, \qquad \forall y \in H,$$

characterizes the elements v of the subdifferential of E at x, we have that an absolutely continuous curve x_t solves (38) if and only if

$$\frac{1}{2}\frac{d}{dt}|x_t - y|^2 + \frac{1}{2}\lambda|x_t - y|^2 + E(x_t) \le E(y), \qquad a.e.\ t \ge 0, \tag{43}$$

holds for every $y \in H$. We will call this system of inequalities the *Evolution Variational Inequality* (EVI).

Thus we got three different characterizations of Gradient Flows in Hilbert spaces: the EDI, the EDE and the EVI. We now want to show that it is possible to formulate these equations also for functionals E defined on a metric space (X, d).

The object $|x_t'|$ appearing in EDI and EDE can be naturally interpreted as the *metric speed* of the absolutely continuous curve x_t as defined in (27). The metric analogous of $|\nabla E|(x)$ is the *slope* of E, defined as:

Definition 4.2 (Slope). Let $E : X \to \mathbb{R} \cup \{+\infty\}$ and $x \in X$ be such that $E(x) < \infty$. Then the slope $|\nabla E|(x)$ of E at x is:

$$|\nabla E|(x) := \overline{\lim_{y \to x}} \frac{(E(x) - E(y))^+}{d(x, y)} = \max\left\{\overline{\lim_{y \to x}} \frac{E(x) - E(y)}{d(x, y)}, 0\right\}.$$

The three definitions of Gradient Flows in a metric setting that we are going to use are:

Definition 4.3 (Energy Dissipation Inequality definition of GF—EDI). Let $E : X \to \mathbb{R} \cup \{+\infty\}$ and let $\overline{x} \in X$ be such that $E(\overline{x}) < \infty$. We say that $[0, \infty) \ni t \mapsto x_t \in X$ is a Gradient Flow in the EDI sense starting at $\overline{x}$ provided it is a locally absolutely continuous curve, $x_0 = \overline{x}$ and

$$\begin{aligned} E(x_s) + \frac{1}{2}\int_0^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_0^s |\nabla E|^2(x_r) dr &\le E(\overline{x}), && \forall s \ge 0, \\ E(x_s) + \frac{1}{2}\int_t^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_t^s |\nabla E|^2(x_r) dr &\le E(x_t), && a.e.\ t > 0,\ \forall s \ge t. \end{aligned} \tag{44}$$

Definition 4.4 (Energy Dissipation Equality definition of GF—EDE). Let $E : X \to \mathbb{R} \cup \{+\infty\}$ and let $\overline{x} \in X$ be such that $E(\overline{x}) < \infty$. We say that $[0, \infty) \ni t \mapsto x_t \in X$ is a Gradient Flow in the EDE sense starting at $\overline{x}$ provided it is a locally absolutely continuous curve, $x_0 = \overline{x}$ and

$$E(x_s) + \frac{1}{2}\int_t^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_t^s |\nabla E|^2(x_r) dr = E(x_t), \qquad \forall 0 \le t \le s. \tag{45}$$

Definition 4.5 (Evolution Variation Inequality definition of GF—EVI). Let $E : X \to \mathbb{R} \cup \{+\infty\}$, $\overline{x} \in \overline{\{E < \infty\}}$ and $\lambda \in \mathbb{R}$. We say that $(0, \infty) \ni t \mapsto x_t \in X$ is a Gradient Flow in the EVI sense (with respect to λ) starting at $\overline{x}$ provided it is a locally absolutely continuous curve in $(0, \infty)$, $x_t \to \overline{x}$ as $t \to 0$ and

$$E(x_t) + \frac{1}{2}\frac{d}{dt} d^2(x_t, y) + \frac{\lambda}{2} d^2(x_t, y) \le E(y), \qquad \forall y \in X,\ a.e.\ t > 0.$$

There are two basic and fundamental things that one needs understand when studying the problem of Gradient Flows in a metric setting:

(1) Although the formulations EDI, EDE and EVI are equivalent for λ-convex functionals on Hilbert spaces, they are *not* equivalent in a metric setting. Shortly said, it holds

$$EVI \qquad \Rightarrow \qquad EDE \qquad \Rightarrow \qquad EDI$$

and typically none of the converse implication holds (see Examples 4.15 and 4.23 below). Here the second implication is clear, for the proof of the first one see Proposition 4.6 below.

(2) Whatever definition of Gradient Flow in a metric setting we use, the main problem is to show existence. The main ingredient in almost all existence proofs is the Minimizing Movements scheme, which we describe after Proposition 4.6.

Proposition 4.6 (EVI implies EDE). *Let $E : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous functional, $\overline{x} \in X$ a given point, $\lambda \in \mathbb{R}$ and assume that (x_t) is a Gradient Flow for E starting from $\overline{x}$ in the EVI sense w.r.t. λ. Then* (45) *holds.*

Proof. First we assume that x_t is locally Lipschitz. The claim will be proved if we show that $t \mapsto E(x_t)$ is locally Lipschitz and it holds

$$-\frac{d}{dt}E(x_t) = \frac{1}{2}|\dot{x}_t|^2 + \frac{1}{2}|\nabla E|^2(x_t), \qquad a.e.\ t > 0.$$

Let us start observing that the triangle inequality implies

$$\frac{1}{2}\frac{d}{dt}d^2(x_t, y) \geq -|\dot{x}_t|d(x_t, y), \qquad \forall y \in X,\ a.e.\ t > 0,$$

thus plugging this bound into the EVI we get

$$-|\dot{x}_t|d(x_t, y) + \frac{\lambda}{2}d^2(x_t, y) + E(x_t) \leq E(y), \qquad \forall y \in X,\ a.e.\ t > 0,$$

which implies

$$|\nabla E|(x_t) = \overline{\lim_{y \to x_t}} \frac{\big(E(x_t) - E(y)\big)^+}{d(x_t, y)} \leq |\dot{x}_t|, \qquad a.e.\ t > 0. \tag{46}$$

Fix an interval $[a, b] \subset (0, \infty)$, let L be the Lipschitz constant of (x_t) in $[a, b]$ and observe that for any $y \in X$ it holds

$$\frac{d}{dt}d^2(x_t, y) \geq -|\dot{x}_t|d(x_t, y) \geq -Ld(x_t, y), \qquad a.e.\ t \in [a, b].$$

Plugging this bound in the EVI we get

$$-Ld(x_t, y) + \frac{\lambda}{2}d^2(x_t, y) + E(x_t) \leq E(y), \qquad a.e.\ t \in [a, b],$$

and by the lower semicontinuity of $t \mapsto E(x_t)$ the inequality holds for every $t \in [a, b]$. Taking $y = x_s$ and then exchanging the roles of x_t, x_s we deduce

$$\begin{aligned}\big|E(x_t) - E(x_s)\big| &\leq Ld(x_t, x_s) - \frac{\lambda}{2}d^2(x_t, x_s)\\ &\leq L|t - s|\left(L + \frac{|\lambda|}{2}L|t - s|\right), \qquad \forall t, s \in [a, b],\end{aligned}$$

thus the map $t \mapsto E(x_t)$ is locally Lipschitz. It is then obvious that it holds

$$\begin{aligned}-\frac{d}{dt}E(x_t) &= \lim_{h \to 0}\frac{E(x_t) - E(x_{t+h})}{h} = \lim_{h \to 0}\frac{E(x_t) - E(x_{t+h})}{d(x_{t+h}, x_t)}\frac{d(x_{t+h}, x_t)}{h}\\ &\leq |\nabla E|(x_t)|\dot{x}_t| \leq \frac{1}{2}|\nabla E|^2(x_t) + \frac{1}{2}|\dot{x}_t|^2, \qquad a.e.\ t.\end{aligned}$$

Thus to conclude we need only to prove the opposite inequality. Integrate the EVI from t to $t+h$ to get

$$\frac{d^2(x_{t+h},y)-d^2(x_t,y)}{2}+\int_t^{t+h}E(x_s)\,ds+\int_t^{t+h}\frac{\lambda}{2}d^2(x_s,y)ds\leq hE(y).$$

Let $y=x_t$ to obtain

$$\begin{aligned}\frac{d^2(x_{t+h},x_t)}{2}&\leq\int_t^{t+h}E(x_t)-E(x_s)\,ds+\frac{|\lambda|}{6}L^2h^3\\&=h\int_0^1 E(x_t)-E(x_{t+hr})\,dr+\frac{|\lambda|}{6}L^2h^3.\end{aligned}$$

Now let $A\subset(0,+\infty)$ be the set of points of differentiability of $t\mapsto E(x_t)$ and where $|\dot{x}_t|$ exists, choose $t\in A\cap(a,b)$, divide by h^2 the above inequality, let $h\to 0$ and use the dominated convergence theorem to get

$$\frac{1}{2}|\dot{x}_t|^2\leq\lim_{h\to 0}\int_0^1\frac{E(x_t)-E(x_{t+hr})}{h}dr=-\frac{d}{dt}E(x_t)\int_0^1 rdr=-\frac{1}{2}\frac{d}{dt}E(x_t).$$

Recalling (46) we conclude with

$$-\frac{d}{dt}E(x_t)\geq|\dot{x}_t|^2\geq\frac{1}{2}|\dot{x}_t|^2+\frac{1}{2}|\nabla E|^2(x_t),\qquad a.e.\ t>0.$$

Finally, we see how the local Lipschitz property of (x_t) can be achieved. It is immediate to verify that the curve $t\mapsto x_{t+h}$ is a Gradient Flow in the EVI sense starting from x_h for all $h>0$. We now use the fact that the distance between curves satisfying the EVI is contractive up to an exponential factor (see the last part of the proof of Theorem 4.25 for a sketch of the argument, and Corollary 4.3.3 of [7] for the rigorous proof). We have

$$d(x_s,x_{s+h})\leq e^{-\lambda(s-t)}d(x_t,x_{t+h}),\qquad\forall s>t.$$

Dividing by h, letting $h\downarrow 0$ and calling $B\subset(0,\infty)$ the set where the metric derivative of x_t exists, we obtain

$$|\dot{x}_s|\leq|\dot{x}_t|e^{-\lambda(s-t)},\qquad\forall s,\ t\in B,\ s>t.$$

This implies that the curve (x_t) is locally Lipschitz in $(0,+\infty)$. □

Let us come back to the case of a convex and lower semicontinuous functional F on an Hilbert space. Pick $\overline{x}\in\overline{D(F)}$, fix $\tau>0$ and define the sequence $n\mapsto x^\tau_{(n)}$ recursively by setting $x^\tau_{(n)}:=\overline{x}$ and defining $x^\tau_{(n+1)}$ as a minimizer of

$$x\qquad\mapsto\qquad F(x)+\frac{|x-x^\tau_{(n)}|^2}{2\tau}.$$

It is immediate to verify that a minimum exists and that it is unique, thus the sequence $n \mapsto x^\tau_{(n)}$ is well defined. The Euler–Lagrange equation of $x^\tau_{(n+1)}$ is:

$$\frac{x^\tau_{(n+1)} - x^\tau_{(n)}}{\tau} \in -\partial^- F(x^\tau_{(n+1)}),$$

which is a time discretization of (38). It is then natural to introduce the rescaled curve $t \mapsto x^\tau_t$ by

$$x^\tau_t := x^\tau_{([t/\tau])},$$

where $[\cdot]$ denotes the integer part, and to ask whether the curves $t \mapsto x^\tau_t$ converge in some sense to a limit curve (x_t) which solves (38) as $\tau \downarrow 0$. This is the case, and this procedure is actually the heart of the proof of Theorem 4.1.

What is important for the discussion we are making now, is that the minimization procedure just described can be naturally posed in a metric setting for a general functional $E : X \to \mathbb{R} \cup \{+\infty\}$: it is sufficient to pick $\overline{x} \in \overline{\{E < \infty\}}$, $\tau > 0$, define $x^\tau_{(0)} := \overline{x}$ and then recursively

$$x^\tau_{(n+1)} \in \operatorname{argmin}\left\{ x \mapsto E(x) + \frac{d^2(x, x^\tau_{(n)})}{2\tau} \right\}. \tag{47}$$

We this give the following definition:

Definition 4.7 (Discrete solution). Let (X, d) be a metric space, $E : X \to \mathbb{R} \cup \{+\infty\}$ lower semicontinuous, $\overline{x} \in \overline{\{E < \infty\}}$ and $\tau > 0$. A *discrete solution* is a map $[0, +\infty) \ni t \mapsto x^\tau_t$ defined by

$$x^\tau_t := x^\tau_{([t/\tau])},$$

where $x^\tau_{(0)} := \overline{x}$ and $x^\tau_{(n+1)}$ satisfies (47).

Clearly in a metric context it is part of the job the identification of suitable assumptions that ensure that the minimization problem (47) admits at least a minimum, so that discrete solutions exist.

We now divide the discussion into three parts, to see under which conditions on the functional E and the metric space X it is possible to prove existence of Gradient Flows in the EDI, EDE and EVI formulation.

4.2.2 General l.s.c. Functionals and EDI

In this section we will make minimal assumptions on the functional E and show how it is possible, starting from them, to prove existence of Gradient Flows in the EDI sense.

Basically, there are two "independent" sets of assumptions that we need: those which ensure the existence of discrete solutions, and those needed to pass to the limit. To better highlight the structure of the theory, we first introduce the hypotheses we need to guarantee the existence of discrete solution and see which properties the discrete solutions have. Then, later on, we introduce the assumptions needed to pass to the limit.

We will denote by $D(E) \subset X$ the domain of E, i.e. $D(E) := \{E < \infty\}$

Assumption 4.8 (Hypothesis for existence of discrete solutions). *(X, d) is a Polish space and $E : X \to \mathbb{R} \cup \{+\infty\}$ be a l.s.c. functional bounded from below. Also, we assume that there exists $\overline{\tau} > 0$ such that for every $0 < \tau < \overline{\tau}$ and $\overline{x} \in \overline{D(E)}$ there exists at least a minimum of*

$$x \quad \mapsto \quad E(x) + \frac{d^2(x, \overline{x})}{2\tau}. \tag{48}$$

Thanks to our assumptions we know that discrete solutions exist for every starting point $\overline{x}$, for τ sufficiently small. The big problem we have to face now is to show that the discrete solutions satisfy a discretized version of the EDI suitable to pass to the limit. The key enabler to do this, is the following result, due to de Giorgi.

Theorem 4.9 (Properties of the variational interpolation). *Let X, E be satisfying the Assumption 4.8. Fix $\overline{x} \in X$, and for any $0 < \tau < \overline{\tau}$ choose x_τ among the minimizers of (48). Then the map $\tau \mapsto E(x_\tau) + \frac{d^2(\overline{x}, x_\tau)}{2\tau}$ is locally Lipschitz in $(0, \overline{\tau})$ and it holds*

$$\frac{d}{d\tau}\left(E(x_\tau) + \frac{d^2(x, x_\tau)}{2\tau}\right) = -\frac{d^2(x, x_\tau)}{2\tau^2}, \qquad a.e. \; \tau \in (0, \overline{\tau}). \tag{49}$$

Proof. Observe that from $E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, x)}{2\tau_0} \le E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, x)}{2\tau_0}$ we deduce

$$\begin{aligned} E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, x)}{2\tau_0} - E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, x)}{2\tau_1} &\le \left(\frac{1}{2\tau_0} - \frac{1}{2\tau_1}\right) d^2(x_{\tau_1}, x) \\ &= \frac{\tau_1 - \tau_0}{2\tau_0\tau_1} d^2(x_{\tau_1}, x). \end{aligned}$$

Arguing symmetrically we see that

$$E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, x)}{2\tau_0} - E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, x)}{2\tau_1} \ge \frac{\tau_1 - \tau_0}{2\tau_0\tau_1} d^2(x_{\tau_0}, x).$$

The last two inequalities show that $\tau \mapsto E(x_\tau) + \frac{d^2(x, x_\tau)}{2\tau}$ is locally Lipschitz and that (49) holds. □

Lemma 4.10. *With the same notation and assumptions as in the previous theorem, $\tau \mapsto d(\overline{x}, x_\tau)$ is non decreasing and $\tau \mapsto E(x_\tau)$ is non increasing. Also, it holds*

$$|\nabla E|(x_\tau) \le \frac{d(x_\tau, \overline{x})}{\tau}. \tag{50}$$

Proof. Pick $0 < \tau_0 < \tau_1 < \overline{\tau}$. From the minimality of x_{τ_0} and x_{τ_1} we get

$$E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, \overline{x})}{2\tau_0} \le E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, \overline{x})}{2\tau_0},$$
$$E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, \overline{x})}{2\tau_1} \le E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, \overline{x})}{2\tau_1}.$$

Adding up and using the fact that $\frac{1}{\tau_0} - \frac{1}{\tau_1} \ge 0$ we get $d(\overline{x}, x_{\tau_0}) \le d(\overline{x}, x_{\tau_1})$. The fact that $\tau \mapsto E(x_\tau)$ is non increasing now follows from

$$E(x_{\tau_1}) + \frac{d^2(x_{\tau_0}, \overline{x})}{2\tau_1} \le E(x_{\tau_1}) + \frac{d^2(x_{\tau_1}, \overline{x})}{2\tau_1} \le E(x_{\tau_0}) + \frac{d^2(x_{\tau_0}, \overline{x})}{2\tau_1}.$$

For the second part of the statement, observe that from

$$E(x_\tau) + \frac{d^2(x_\tau, \overline{x})}{2\tau} \le E(y) + \frac{d^2(y, \overline{x})}{2\tau}, \qquad \forall y \in X$$

we get

$$\frac{E(x_\tau) - E(y)}{d(x_\tau, y)} \le \frac{d^2(y, \overline{x}) - d^2(x_\tau, \overline{x})}{2\tau d(x_\tau, y)} = \frac{\big(d(y, \overline{x}) - d(x_\tau, \overline{x})\big)\big(d(x_\tau, \overline{x}) + d(y, \overline{x},)\big)}{2\tau d(x_\tau, y)}$$
$$\le \frac{d(x_\tau, \overline{x},) + d(y, \overline{x})}{2\tau}.$$

Taking the limsup as $y \to x_\tau$ we get the thesis. □

By Theorem 4.9 and Lemma 4.10 it is natural to introduce the following *variational interpolation* in the Minimizing Movements scheme (as opposed to the classical piecewise constant/affine interpolations used in other contexts):

Definition 4.11 (Variational interpolation). Let X, E be satisfying Assumption 4.8, $\overline{x} \in \overline{D(E)}$ and $0 < \tau < \overline{\tau}$. We define the map $[0, \infty) \ni t \mapsto x_t^\tau$ in the following way:

- $x_0^\tau := \overline{x}$.
- $x_{(n+1)\tau}^\tau$ is chosen among the minimizers of (48) with $\overline{x}$ replaced by $x_{n\tau}^\tau$.
- x_t^τ with $t \in (n\tau, (n+1)\tau)$ is chosen among the minimizers of (48) with $\overline{x}$ and τ replaced by $x_{n\tau}^\tau$ and $t - n\tau$ respectively.

For (x_t^τ) defined in this way, we define the *discrete speed* $\mathrm{Dsp}^\tau : [0, +\infty) \to [0, +\infty)$ and the *Discrete slope* $\mathrm{Dsl}^\tau : [0, +\infty) \to [0, +\infty)$ by:

$$\begin{aligned} \mathrm{Dsp}_t^\tau &:= \frac{d\left(x_{n\tau}^\tau, x_{(n+1)\tau}^\tau\right)}{\tau}, \qquad t \in (n\tau, (n+1)\tau), \\ \mathrm{Dsl}_t^\tau &:= \frac{d\left(x_t^\tau, x_{n\tau}^\tau\right)}{t - n\tau}, \qquad t \in (n\tau, (n+1)\tau). \end{aligned} \tag{51}$$

Although the object Dsl_t^τ does not look like a slope, we chose this name because from (50) we know that $|\nabla E|(x_t^\tau) \leq \mathrm{Dsl}_t^\tau$ and because in the limiting process Dsl^τ will produce the slope term in the EDI (see the proof of Theorem 4.14).

With this notation we have the following result:

Corollary 4.12 (EDE for the discrete solutions). *Let X, E be satisfying Assumption 4.8, $\overline{x} \in \overline{D(E)}$, $0 < \tau < \overline{\tau}$ and (x_t^τ) defined via the variational interpolation as in Definition 4.11 above. Then it holds*

$$E(x_s^\tau) + \frac{1}{2}\int_t^s |\mathrm{Dsp}_r^\tau|^2 dr + \frac{1}{2}\int_t^s |\mathrm{Dsl}_r^\tau|^2 dr = E(x_t^\tau), \tag{52}$$

for every $t = n\tau$, $s = m\tau$, $n < m \in \mathbb{N}$.

Proof. It is just a restatement of (49) in terms of the notation given in (51). □

Thus, at the level of discrete solutions, it is possible to get a discrete form of the Energy Dissipation Equality under the quite general Assumptions 4.8. Now we want to pass to the limit as $\tau \downarrow 0$. In order to do this, we need to add some compactness and regularity assumptions on the functional:

Assumption 4.13 (Coercivity and regularity assumptions). *Assume that $E : X \to \mathbb{R} \cup \{+\infty\}$ satisfies:*

- *E is bounded from below and its sublevels are boundedly compact, i.e. $\{E \leq c\} \cap \overline{B_r(x)}$ is compact for any $c \in \mathbb{R}$, $r > 0$ and $x \in X$.*
- *The slope $|\nabla E| : D(E) \to [0, +\infty]$ is lower semicontinuous.*
- *E has the following continuity property:*

$$x_n \to x, \ \sup_n\{|\nabla E|(x_n), E(x_n)\} < \infty \qquad \Rightarrow \qquad E(x_n) \to E(x).$$

Under these assumptions we can prove the following result:

Theorem 4.14 (Gradient Flows in EDI formulation). *Let (X, d) be a metric space and let $E : X \to \mathbb{R}\cup\{+\infty\}$ be satisfying the Assumptions 4.8 and 4.13. Also, let $\overline{x} \in D(E)$ and for $0 < \tau < \overline{\tau}$ define the discrete solution via the variational interpolation as in Definition 4.11. Then it holds:*

- *The set of curves $\{(x_t^\tau)\}_\tau$ is relatively compact in the set of curves in X w.r.t. local uniform convergence.*
- *Any limit curve (x_t) is a Gradient Flow in the EDI formulation (Definition 4.3).*

Sketch of the Proof

Compactness. By Corollary 4.12 we have

$$d^2(x_t^\tau, \overline{x}) \leq \left(\int_0^T |\mathrm{Dsp}_r^\tau| dr\right)^2 \leq T\int_0^T |\mathrm{Dsp}_\mathrm{r}^\tau|^2 dr \leq 2T\big(E(\overline{x}) - \inf E\big), \qquad \forall t \leq T,$$

for any $T = n\tau$. Therefore for any $T > 0$ the set $\{x_t^\tau\}_{t \le T}$ is uniformly bounded in τ. As this set is also contained in $\{E \le E(\overline{x})\}$, it is relatively compact. The fact that there is relative compactness w.r.t. local uniform convergence follows by an Ascoli–Arzelà-type argument based on the inequality

$$d^2\big(x_t^\tau, x_s^\tau\big) \le \left(\int_t^s |\mathrm{Dsp}_r^\tau| dr\right)^2 \le 2(s-t)\big(E(\overline{x}) - \inf E\big),\ \forall t = n\tau,\ s = m\tau,\ n < m \in \mathbb{N}. \tag{53}$$

Passage to the limit. Let $\tau_n \downarrow 0$ be such that $(x_t^{\tau_n})$ converges to a limit curve x_t locally uniformly. Then by standard arguments based on inequality (53) it is possible to check that $t \mapsto x_t$ is absolutely continuous and satisfies

$$\int_t^s |\dot{x}_r|^2 dr \le \underline{\lim}_{n\to\infty} \int_t^s |\mathrm{Dsp}_r^{\tau_n}|^2 dr \qquad \forall 0 \le t < s. \tag{54}$$

By the lower semicontinuity of $|\nabla E|$ and (50) we get

$$|\nabla E|(x_t) \le \underline{\lim}_{n\to\infty} |\nabla E|(x_t^{\tau_n}) \le \underline{\lim}_{n\to\infty} \mathrm{Dsl}_t^{\tau_n}, \qquad \forall t,$$

thus Fatou's lemma ensures that for any $t < s$ it holds

$$\int_t^s |\nabla E|^2(x_r) dr \le \int_t^s \underline{\lim}_{n\to\infty} |\nabla E|^2(x_r^\tau) dr \le \underline{\lim}_{n\to\infty} \int_t^s |\mathrm{Dsl}_r^{\tau_n}|^2\, dr \le 2T\big(E(\overline{x}) - \inf E\big). \tag{55}$$

Now passing to the limit in (52) written for $t = 0$ we get the first inequality in (44). Also, from (55) we get that the L^2 norm of $f(t) := \underline{\lim}_{n\to\infty} |\nabla E|(x_t^{\tau_n})$ on $[0,\infty)$ is finite. Thus $A := \{f < \infty\}$ has full Lebesgue measure and for each $t \in A$ we can find a subsequence $\tau_{n_k} \downarrow 0$ such that $\sup_k |\nabla E|(x_t^{\tau_{n_k}}) < \infty$. Then the third assumption in 4.13 guarantees that $E(x_t^{\tau_{n_k}}) \to E(x_t)$ and the lower semicontinuity of E that $E(x_s) \le \underline{\lim}_{k\to\infty} E(x_s^{\tau_{n_k}})$ for every $s \ge t$. Thus passing to the limit in (52) as $\tau_{n_k} \downarrow 0$ and using (54) and (55) we get

$$E(x_s) + \frac{1}{2}\int_t^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_t^s |\nabla E|^2(x_r) dr \le E(x_t), \qquad \forall t \in A,\ \forall s \ge t.$$

□

We conclude with an example which shows why in general we cannot hope to have equality in the EDI. Shortly said, the problem is that we don't know whether $t \mapsto E(x_t)$ is an absolutely continuous map.

Example 4.15. Let $X = [0,1]$ with the Euclidean distance, $C \subset X$ a Cantor-type set with null Lebesgue measure and $f : [0,1] \to [1,+\infty]$ a continuous, integrable function such that $f(x) = +\infty$ for any $x \in C$, which is smooth on the complement of C. Also, let $g : [0,1] \to [0,1]$ be a "Devil staircase" built over C, i.e. a continuous, non decreasing function satisfying $g(0) = 0$, $g(1) = 1$ which is constant in each of the connected components of the complement of C. Define the energies $E, \tilde{E} : [0,1] \to \mathbb{R}$ by

$$E(x) := -g(x) - \int_0^x f(y)dy.$$

$$\tilde{E}(x) := -\int_0^x f(y)dy.$$

It is immediate to verify that $E, \tilde{E}$ satisfy all the Assumptions 4.8, 4.13 (the choice of f guarantees that the slopes of $E, \tilde{E}$ are continuous). Now build a Gradient Flow starting from 0: with some work it is possible to check that the Minimizing Movement scheme converges in both cases to absolutely continuous curves (x_t) and $(\tilde{x}_t)$ respectively satisfying

$$x'_t = -|\nabla E|(x_t), \qquad a.e.\ t$$

$$\tilde{x}'_t = -|\nabla \tilde{E}|(\tilde{x}_t), \qquad a.e.\ t.$$

Now, notice that $|\nabla E|(x) = |\nabla \tilde{E}|(x) = f(x)$ for every $x \in [0,1]$, therefore the fact that $f \geq 1$ is smooth on $[0,1] \setminus C$ gives that each of these two equations admit a unique solution. Therefore—this is the key point of the example—(x_t) and $(\tilde{x}_t)$ must coincide. In other words, the effect of the function g is not seen at the level of Gradient Flow. It is then immediate to verify that there is Energy Dissipation Equality for the energy $\tilde{E}$, but there is only the Energy Dissipation Inequality for the energy E. ■

4.2.3 The Geodesically Convex Case: EDE and Regularizing Effects

Here we study gradient flows of so called *geodesically convex* functionals, which are the natural metric generalization of convex functionals on linear spaces.

Definition 4.16 (Geodesic convexity). Let $E : X \to \mathbb{R} \cup \{+\infty\}$ be a functional and $\lambda \in \mathbb{R}$. We say that E is λ-geodesically convex provided for every $x, y \in X$ there exists a constant speed geodesic $\gamma : [0,1] \to X$ connecting x to y such that

$$E(\gamma_t) \leq (1-t)E(x) + tE(y) - \frac{\lambda}{2}t(1-t)d^2(x,y). \tag{56}$$

In this section we will assume that:

Assumption 4.17 (Geodesic convexity hypothesis). *(X,d) is a Polish geodesic space, $E : X \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous, λ-geodesically convex for some $\lambda \in \mathbb{R}$. Also, we assume that the sublevels of E are boundedly compact, i.e. the set $\{E \leq c\} \cap \overline{B_r(x)}$ is compact for any $c \in \mathbb{R}$, $r > 0$, $x \in X$.*

What we want to prove is that for X, E satisfying these assumptions there is existence of Gradient Flows in the formulation EDE (Definition 4.4).

Our first goal is to show that in this setting it is possible to recover the results of the previous section. We start claiming that it holds:

$$|\nabla E|(x) = \sup_{y \neq x} \left(\frac{E(x) - E(y)}{d(x, y)} + \frac{\lambda}{2} d(x, y) \right)^+, \tag{57}$$

so that the $\overline{\lim}$ in the definition of the slope can be replaced by a sup. Indeed, we know that

$$|\nabla E|(x) = \varlimsup_{y \to x} \left(\frac{E(x) - E(y)}{d(x, y)} + \frac{\lambda}{2} d(x, y) \right)^+ \leq \sup_{y \neq x} \left(\frac{E(x) - E(y)}{d(x, y)} + \frac{\lambda}{2} d(x, y) \right)^+.$$

To prove the opposite inequality fix $y \neq x$ and a constant speed geodesic γ connecting x to y for which (56) holds. Then observe that

$$\begin{aligned} |\nabla E|(x) &\geq \varlimsup_{t \downarrow 0} \left(\frac{E(x) - E(\gamma_t)}{d(x, \gamma_t)} \right)^+ = \left(\varlimsup_{t \downarrow 0} \frac{E(x) - E(\gamma_t)}{d(x, \gamma_t)} \right)^+ \\ &\overset{(56)}{\geq} \left(\varlimsup_{t \downarrow 0} \left(\frac{E(x) - E(y)}{d(x, y)} + \frac{\lambda}{2}(1 - t) d(x, y) \right) \right)^+ \\ &= \left(\frac{E(x) - E(y)}{d(x, y)} + \frac{\lambda}{2} d(x, y) \right)^+. \end{aligned}$$

Using this representation formula we can show that all the Assumptions 4.8 and 4.13 hold:

Proposition 4.18. *Suppose that Assumption 4.17 holds. Then Assumptions 4.8 and 4.13 hold as well.*

Sketch of the Proof From the λ-geodesic convexity and the lower semicontinuity assumption it is possible to deduce (we omit the details) that E has at most quadratic decay at infinity, i.e. there exists $\overline{x} \in X$, $a, b > 0$ such that

$$E(x) \geq -a - b d(x, \overline{x}) + \lambda^- d^2(x, \overline{x}), \qquad \forall x \in X.$$

Therefore from the lower semicontinuity again and the bounded compactness of the sublevels of E we immediately get that the minimization problem (48) admits a solution if $\tau < 1/\lambda^-$.

The lower semicontinuity of the slope is a direct consequence of (57) and of the lower semicontinuity of E. Thus, to conclude we need only to show that

$$x_n \to x, \ \sup_n \{|\nabla E|(x_n), E(x_n)\} < \infty \qquad \Rightarrow \qquad \varlimsup_{n \to \infty} E(x_n) \leq E(x). \tag{58}$$

From (57) with $x,\ y$ replaced by $x_n,\ x$ respectively we get

$$E(x) \geq E(x_n) - |\nabla E|(x_n) d(x, x_n) + \frac{\lambda}{2} d^2(x, x_n),$$

and the conclusion follows by letting $n \to \infty$. □

Thus Theorem 4.14 applies directly also to this case and we get existence of Gradient Flows in the EDI formulation. To get existence in the stronger EDE formulation, we need the following result, which may be thought as a sort of weak chain rule (observe that the validity of the proposition below rules out behaviors like the one described in Example 4.15).

Proposition 4.19. *Let E be a λ-geodesically convex and l.s.c. functional. Then for every absolutely continuous curve $(x_t) \subset X$ such that $E(x_t) < \infty$ for every t, it holds*

$$\big|E(x_s) - E(x_t)\big| \leq \int_t^s |\dot{x}_r||\nabla E(x_r)| dr, \qquad \forall t < s. \tag{59}$$

Proof. We may assume that the right hand side of (59) is finite for any $t, s \in [0, 1]$, and, by a reparametrization argument, we may also assume that $|\dot{x_t}| = 1$ for a.e. t (in particular (x_t) is 1-Lipschitz), so that $t \mapsto |\nabla E|(x_t)$ is an L^1 function. Notice that it is sufficient to prove that $t \mapsto E(x_t)$ is absolutely continuous, as then the inequality

$$\begin{aligned}\overline{\lim_{h\uparrow 0}} \frac{E(x_{t+h}) - E(x_t)}{h} &\leq \overline{\lim_{h\uparrow 0}} \frac{(E(x_t) - E(x_{t+h}))^+}{|h|} \\ &\leq \overline{\lim_{h\uparrow 0}} \frac{(E(x_t) - E(x_{t+h}))^+}{d(x_t, x_{t+h})} \overline{\lim_{h\uparrow 0}} \frac{d(x_t, x_{t+h})}{|h|} \leq |\nabla E(x_t)||\dot{x_t}|,\end{aligned}$$

valid for any $t \in [0, 1]$ gives (59).

Define the functions $f, g : [0, 1] \to \mathbb{R}$ by

$$\begin{aligned} f(t) &:= E(x_t), \\ g(t) &:= \sup_{s\neq t} \frac{(f(t) - f(s))^+}{|s - t|} \end{aligned}$$

Let D be the diameter of the compact set $\{x_t\}_{t\in[0,1]}$, use the fact that (x_t) is 1-Lipschitz, formula (57) and the trivial inequality $a^+ \leq (a + b)^+ + b^-$ (valid for any $a,\ b \in \mathbb{R}$) to get

$$g(t) \leq \sup_{s\neq t} \frac{(E(x_t) - E(x_s))^+}{d(x_s, x_t)} \leq |\nabla E|(x_t) + \frac{\lambda^-}{2} D.$$

Therefore the thesis will be proved if we show that:

$$g \in L^1 \qquad \Rightarrow \qquad |f(s) - f(t)| \leq \int_t^s g(r) dr \qquad \forall t < s. \tag{60}$$

Fix $M > 0$ and define $f^M := \min\{f, M\}$. Now fix $\varepsilon > 0$, pick a smooth mollifier $\rho_\varepsilon : \mathbb{R} \to \mathbb{R}$ with support in $[-\varepsilon, \varepsilon]$ and define $f_\varepsilon^M,\ g_\varepsilon^M : [\varepsilon, 1-\varepsilon] \to \mathbb{R}$ by

$$\begin{aligned} f_\varepsilon^M(t) &:= f^M * \rho_\varepsilon(t),\\ g_\varepsilon^M(t) &:= \sup_{s\neq t} \frac{(f_\varepsilon^M(t) - f_\varepsilon^M(s))^+}{|s-t|}. \end{aligned}$$

Since f_ε^M is smooth and $g_\varepsilon^M \geq (f_\varepsilon^M)'$ it holds

$$|f_\varepsilon^M(s) - f_\varepsilon^M(t)| \leq \int_t^s g_\varepsilon^M(r)dr. \tag{61}$$

From the trivial bound $(\int h)^+ \leq \int h^+$ we get

$$\begin{aligned} g_\varepsilon^M(t) &\leq \sup_s \frac{\int (f^M(t-r) - f^M(s-r))^+ \rho_\varepsilon(r)dr}{|s-t|}\\ &\leq \sup_s \frac{\int (f(t-r) - f(s-r))^+ \rho_\varepsilon(r)dr}{|s-t|}\\ &= \sup_s \int \frac{(f(t-r) - f(s-r))^+}{|(s-r)-(t-r)|}\rho_\varepsilon(r)dr \leq \int g(t-r)\rho_\varepsilon(r)dr = g * \rho_\varepsilon(t). \end{aligned} \tag{62}$$

Thus the family of functions $\{g_\varepsilon^M\}_\varepsilon$ is dominated in $L^1(0,1)$. From (61) and (62) it follows that the family of functions $\{f_\varepsilon^M\}$ uniformly converge to some function $\tilde{f}^M$ on $[0,1]$ as $\varepsilon \downarrow 0$ for which it holds

$$|\tilde{f}^M(s) - \tilde{f}^M(t)| \leq \int_t^s g(r)dr.$$

We know that $f^M = \tilde{f}^M$ on some set $A \subset [0,1]$ such that $\mathscr{L}^1([0,1]\setminus A) = 0$, and we want to prove that they actually coincide everywhere. Recall that f^M is l.s.c. and $\tilde{f}^M$ is continuous, hence $f^M \leq \tilde{f}^M$ in $[0,1]$. If by contradiction it holds $f^M(t_0) < c < C < \tilde{f}^M(t_0)$ for some t_0, c, C, we can find $\delta > 0$ such that $\tilde{f}^M(t) > C$ in $t \in [t_0-\delta, t_0+\delta]$. Thus $f^M(t) > C$ for $t \in [t_0-\delta, t_0+\delta] \cap A$ and the contradiction comes from

$$\int_0^1 g(t)dt \geq \int_{[t_0-\delta,t_0+\delta]\cap A} g(t)dt \geq \int_{[t_0-\delta,t_0+\delta]\cap A} \frac{C-c}{|t-t_0|}dt = +\infty.$$

Thus we proved that if $g \in L^1(0,1)$ it holds

$$|f^M(t) - f^M(s)| \leq \int_t^s g(r)dr, \qquad \forall t < s \in [0,1],\ M > 0.$$

Letting $M \to \infty$ we prove (60) and hence the thesis. □

This proposition is the key ingredient to pass from existence of Gradient Flows in the EDI formulation to the one in the EDE formulation:

Theorem 4.20 (Gradient Flows in the EDE formulation). *Let X, E be satisfying Assumption 4.17 and $\overline{x} \in X$ be such that $E(\overline{x}) < \infty$. Then all the results of Theorem 4.14 hold.*

Also, any Gradient Flow in the EDI sense is also a Gradient Flow in the EDE sense (Definition 4.4).

Proof. The first part of the statement follows directly from Proposition 4.18.

By Theorem 4.14 we know that the limit curve is absolutely continuous and satisfies

$$E(x_s) + \frac{1}{2}\int_0^s |\dot{x}|_r^2 dr + \frac{1}{2}\int_0^s |\nabla E|^2(x_r)dr \leq E(\overline{x}), \qquad \forall s \geq 0. \tag{63}$$

In particular, the functions $t \mapsto |\dot{x}_t|$ and $t \mapsto |\nabla E|(x_t)$ belong to $L^2_{loc}(0, +\infty)$. Now we use Proposition 4.19: we know that for any $s \geq 0$ it holds

$$\big|E(\overline{x}) - E(x_s)\big| \leq \int_0^s |\dot{x}_r||\nabla E|(x_r)dr \leq \frac{1}{2}\int_0^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_0^s |\nabla E|^2(x_r)dr. \tag{64}$$

Therefore $t \mapsto E(x_t)$ is locally absolutely continuous and it holds

$$E(x_s) + \frac{1}{2}\int_0^s |\dot{x}_r|^2 dr + \frac{1}{2}\int_0^s |\nabla E|^2(x_r)dr = E(\overline{x}), \qquad \forall s \geq 0.$$

Subtracting from this last equation the same equality written for $s = t$ we get the thesis. □

Remark 4.21. It is important to underline that the hypothesis of λ-geodesic convexity is in general of no help for what concerns the compactness of the sequence of discrete solutions. ■

The λ-geodesic convexity hypothesis, ensures various regularity results for the limit curve, which we state without proof:

Proposition 4.22. *Let X, E be satisfying Assumption 4.17 and let (x_t) be any limit of a sequence of discrete solutions. Then:*

(i) The limit

$$|\dot{x}_t^+| := \lim_{h\downarrow 0} \frac{d(x_{t+h}, x_t)}{h},$$

exists for every $t > 0$.

(ii) The equation

$$\frac{d}{dt_+}E(x_t) = -|\nabla E|^2(x_t) = -|\dot{x}_t^+|^2 = -|\dot{x}_t^+||\nabla E|(x_t),$$

is satisfied at every $t > 0$.

(iii) The map $t \mapsto e^{-2\lambda^- t}E(x_t)$ *is convex, the map* $t \mapsto e^{\lambda t}|\nabla E|(x_t)$ *is non increasing, right continuous and satisfies*

$$\frac{t}{2}|\nabla E|^2(x_t) \leq e^{2\lambda^- t}\Big(E(x_0) - E_t(x_0)\Big),$$

$$t|\nabla E|^2(x_t) \leq (1 + 2\lambda^+ t)e^{-2\lambda t}\Big(E(x_0 - \inf E\Big),$$

where $E_t : X \to \mathbb{R}$ *is defined as*

$$E_t(x) := \inf_y E(y) + \frac{d^2(x,y)}{2t}.$$

(iv) If $\lambda > 0$*, then* E *admits a unique minimum* x_{min} *and it holds*

$$\frac{\lambda}{2}d^2(x_t, x_{min}) \leq E(x_t) - E(x_{min}) \leq e^{-2\lambda t}\big(E(x_0) - E(x_{min})\big).$$

Observe that we didn't state any result concerning the uniqueness (nor about contractivity) of the curve (x_t) satisfying the Energy Dissipation Equality (45). The reason is that if no further assumptions are made on either X or E, in general uniqueness fails, as the following simple example shows:

Example 4.23 (Lack of uniqueness). Let $X := \mathbb{R}^2$ endowed with the L^∞ norm, $E : X \to \mathbb{R}$ be defined by $E(x^1, x^2) := x^1$ and $\overline{x} := (0,0)$. Then it is immediate to verify that $|\nabla E| \equiv 1$ and that any Lipschitz curve $t \mapsto x_t = (x_t^1, x_t^2)$ satisfying

$$x_t^1 = -t, \qquad \forall t \geq 0$$

$$|x_t^{2'}| \leq 1, \qquad a.e.\, t > 0,$$

satisfies also

$$E(x_t) = -t,$$

$$|\dot{x}_t| = 1.$$

This implies that any such (x_t) satisfies the Energy Dissipation Equality (45). ■

4.2.4 The Compatibility of Energy and Distance: EVI and Error Estimates

As the last example of the previous section shows, in general we cannot hope to have uniqueness of the limit curve (x_t) obtained via the Minimizing Movements scheme for a generic λ-geodesically convex functional. If we want to derive properties like uniqueness and contractivity of the flow, we need to have some stronger relation between the Energy functional E and the distance d on X: in this section we will assume the following:

Assumption 4.24 (Compatibility in Energy and distance).
(X,d) is a Polish space. $E : X \to \mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous functional and for any $x_0,\ x_1,\ y \in X$, there exists a curve $t \mapsto \gamma(t)$ such that

$$\begin{aligned} E(\gamma_t) &\le (1-t)E(x_0) + tE(x_1) - \frac{\lambda}{2}t(1-t)d^2(x_0,x_1), \\ d^2(\gamma_t,y) &\le (1-t)d^2(x_0,y) + td^2(x_1,y) - t(1-t)d^2(x_0,x_1), \end{aligned} \tag{65}$$

for every $t \in [0,1]$.

Observe that there is no compactness assumption of the sublevels of E. If X is an Hilbert space (and more generally a NPC space—Definition 3.19) then the second inequality in (65) is satisfied by geodesics. Hence λ-convex functionals are automatically compatible with the metric.

Following the same lines of the previous section, it is possible to show that this assumption implies both Assumption 4.8 and, if the sublevels of E are boundedly compact, Assumption 4.13, so that Theorem 4.14 holds. Also it can be shown that formula (57) is true and thus that Proposition 4.19 holds also in this setting, so that Theorem 4.20 can be proved as well.

However, if Assumption 4.24 holds, it is better not to follow the general theory as developed before, but to restart from scratch: indeed, in this situation much stronger statements hold, also at the level of discrete solutions, which can be proved by a direct use of Assumption 4.24.

We collect the main results achievable in this setting in the following theorem:

Theorem 4.25 (Gradient Flows for compatible E and d: EVI). *Assume that X, E satisfy Assumption 4.24. Then the following hold.*

- *For every $x \in \overline{D(E)}$ and $0 < \tau < 1/\lambda^-$ there exists a unique discrete solution (x_t^τ) as in Definition 4.7.*
- *Let $x \in \overline{D(E)}$ and (x_t^τ) any family of discrete solutions starting from it. Then (x_t^τ) converge locally uniformly to a limit curve (x_t) as $\tau \downarrow 0$ (so that the limit curve is unique). Furthermore, (x_t) is the unique solution of the system of differential inequalities:*

$$\frac{1}{2}\frac{d}{dt}d^2(\tilde{x}_t,y) + \frac{\lambda}{2}d^2(\tilde{x}_t,y) + E(\tilde{x}_t) \le E(y), \qquad a.e.\ t \ge 0,\ \forall y \in X, \tag{66}$$

  among all locally absolutely continuous curves $(\tilde{x}_t)$ converging to $\overline{x}$ as $t \downarrow 0$. I.e. x_t is a Gradient Flow in the EVI formulation—see Definition 4.5.
- *Let $\overline{x}, \overline{y} \in \overline{D(E)}$ and (x_t), (y_t) be the two Gradient Flows in the EVI formulation. Then there is λ-exponential contraction of the distance, i.e.:*

$$d^2(x_t,y_t) \le e^{-\lambda t}d^2(\overline{x},\overline{y}). \tag{67}$$

- *Suppose that $\lambda \geq 0$, that $\overline{x} \in D(E)$ and build x_t^τ, x_t as above. Then the following a priori error estimate holds:*

$$\sup_{t\geq 0} d(x_t, x_t^\tau) \leq 8\sqrt{\tau(E(\overline{x}) - E(x_t))}. \tag{68}$$

Sketch of the Proof We will make the following simplifying assumptions: $E \geq 0$, $\lambda \geq 0$ and $\overline{x} \in D(E)$. Also we will prove just that the sequence of discrete solutions $n \mapsto x_t^{\tau/2^n}$ converges to a limit curve as $n \to \infty$ for any given $\tau > 0$.

Existence and uniqueness of the discrete solution. Pick $x \in X$. We have to prove that there exists a unique minimizer of (48). Let $I \geq 0$ be the infimum of (48). Let (x_n) be a minimizing sequence for (48), fix $n, m \in \mathbb{N}$ and let $\gamma : [0,1] \to X$ be a curve satisfying (65) for $x_0 := x_n$, $x_1 := x_m$ and $y := x$. Using the inequalities (65) at $t = 1/2$ we get

$$\begin{aligned} I &\leq E(\gamma_{1/2}) + \frac{d^2(\gamma_{1/2}, x)}{2\tau} \\ &\leq \frac{1}{2}\left(E(x_n) + \frac{d^2(x_n, x)}{2\tau} + E(x_m) + \frac{d^2(x_m, x)}{2\tau}\right) - \frac{1+\lambda\tau}{8\tau} d^2(x_n, x_m). \end{aligned}$$

Therefore

$$\overline{\lim_{n,m\to\infty}} \frac{1+\lambda\tau}{8\tau} d^2(x_n, x_m) \leq \overline{\lim_{n,m\to\infty}} \frac{1}{2}\left(E(x_n) + \frac{d^2(x_n, x)}{2\tau} + E(x_m) + \frac{d^2(x_m, x)}{2\tau}\right) - I = 0$$

and thus the sequence (x_n) is a Cauchy sequence as soon as $0 < \tau < 1/\lambda^-$. This shows uniqueness, existence follows by the l.s.c. of E.

One step estimates We claim that the following discrete version of the EVI (66) holds: for any $x \in X$,

$$\frac{d^2(x^\tau, y) - d^2(x, y)}{2\tau} + \frac{\lambda}{2} d^2(x^\tau, y) \leq E(y) - E(x^\tau), \qquad \forall y \in X, \tag{69}$$

where x^τ is the minimizer of (48). Indeed, pick a curve γ satisfying (65) for $x_0 := x^\tau$, $x_1 := y$ and $y := x$ and use the minimality of x^τ to get

$$\begin{aligned} E(x^\tau) + \frac{d^2(x, x^\tau)}{2\tau} &\leq E(\gamma_t) + \frac{d^2(x, \gamma_t)}{2\tau} \\ &\leq (1-t)E(x^\tau) + tE(y) - \frac{\lambda}{2} t(1-t) d^2(x^\tau, y) \\ &\quad + \frac{(1-t)d^2(x, x^\tau) + t d^2(x, y) - t(1-t) d^2(x^\tau, y)}{2\tau}. \end{aligned}$$

Rearranging the terms, dropping the positive addend $td^2(x, x^\tau)$ and dividing by $t > 0$ we get

$$\frac{(1-t)d^2(x^\tau, y)}{2\tau} - \frac{d^2(x, y)}{2\tau} + \frac{\lambda}{2}(1-t)d^2(x^\tau, y) \leq E(y) - E(x^\tau),$$

so that letting $t \downarrow 0$ we get (69).

Now we pass to the discrete version of the error estimate, which will also give the full convergence of the discrete solutions to the limit curve. Given $\overline{x}$, $\overline{y} \in D(E)$, and the associate discrete solutions x_t^τ, y_t^τ, we are going to bound the distance $d(x_\tau^{\tau/2}, y_\tau^\tau)$ in terms of the distance $d(\overline{x}, \overline{y})$.

Write two times the discrete EVI (69) for $\tau := \tau/2$ and $y := \overline{y}$: first with $x := \overline{x}$, then with $x := x_{\tau/2}^{\tau/2}$ to get (we use the assumption $\lambda \geq 0$)

$$\frac{d^2(x_{\tau/2}^{\tau/2}, \overline{y}) - d^2(\overline{x}, \overline{y})}{\tau} \leq E(\overline{y}) - E(x_{\tau/2}^{\tau/2}),$$

$$\frac{d^2(x_\tau^{\tau/2}, \overline{y}) - d^2(x_{\tau/2}^{\tau/2}, \overline{y})}{\tau} \leq E(\overline{y}) - E(x_\tau^{\tau/2}).$$

Adding up these two inequalities and observing that $E(x_\tau^{\tau/2}) \leq E(x_{\tau/2}^{\tau/2})$ we obtain

$$\frac{d^2(x_\tau^{\tau/2}, \overline{y}) - d^2(\overline{x}, \overline{y})}{\tau} \leq 2\big(E(\overline{y}) - E(x_\tau^{\tau/2})\big).$$

On the other hand, (69) with $x := \overline{y}$ and $y := x_\tau^{\tau/2}$ reads as

$$\frac{d^2(y_\tau^\tau, x_\tau^{\tau/2}) - d^2(\overline{y}, x_\tau^{\tau/2})}{\tau} \leq 2\big(E(x_\tau^{\tau/2}) - E(y_\tau^\tau)\big).$$

Adding up these last two inequalities we get

$$\frac{d^2(y_\tau^\tau, x_\tau^{\tau/2}) - d^2(\overline{x}, \overline{y})}{\tau} \leq 2\big(E(\overline{y}) - E(y_\tau^\tau)\big). \tag{70}$$

Discrete estimates. Pick $t = n\tau < m\tau = s$, write inequality (69) for $x := x_{i\tau}^\tau$, $i = n, \ldots, m-1$ and add everything up to get

$$\frac{d^2(x_t^\tau, y) - d^2(x_s^\tau, y)}{2(s-t)} + \frac{\lambda\tau}{2(s-t)} \sum_{i=n+1}^{m} d^2(x_{i\tau}^\tau, y) \leq E(y) - \frac{\tau}{s-t} \sum_{i=n+1}^{m} E(x_{i\tau}^\tau). \tag{71}$$

Similarly, pick $t = n\tau$, write inequality (70) for $\overline{x} := x_{i\tau}^{\tau/2}$ and $\overline{y} := y_{i\tau}^\tau$ for $i = 0, \ldots, n-1$ and add everything up to get

$$\frac{d^2(x_t^{\tau/2}, y_t^\tau) - d^2(\overline{x}, \overline{y})}{\tau} \leq 2\big(E(\overline{y}) - E(y_t^\tau)\big).$$

Now let $\overline{y} = \overline{x}$ to get

$$d^2(x_t^{\tau/2}, x_t^{\tau}) \le 2\tau\big(E(\overline{x}) - E(x_t^{\tau})\big) \le 2\tau E(\overline{x}), \tag{72}$$

having used the fact that $E \ge 0$.

Conclusion of passage to the limit. Putting $\tau/2^n$ instead of τ in (72) we get

$$d^2(x_t^{\tau/2^{n+1}}, x_t^{\tau/2^n}) \le \frac{\tau}{2^{n-1}} E(\overline{x}),$$

therefore

$$d^2(x_t^{\tau/2^n}, x_t^{\tau/2^m}) \le \tau(2^{2-n} - 2^{2-m})E(\overline{x}), \qquad \forall n < m \in \mathbb{N},$$

which tells that $n \mapsto x_t^{\tau/2^n}$ is a Cauchy sequence for any $t \ge 0$. Also, choosing $n = 0$ and letting $m \to \infty$ we get the error estimate (68).

We pass to the EVI. Letting $\tau \downarrow 0$ in (71) it is immediate to verify that we get

$$\frac{d^2(x_t, y) - d^2(x_s, y)}{2(s-t)} + \frac{\lambda}{2(s-t)} \int_t^s d^2(x_r, y) \le E(y) - \frac{1}{s-t} \int_t^s E(x_r)dr,$$

which is precisely the EVI (66) written in integral form.

Uniqueness and contractivity. It remains to prove that the solution to the EVI is unique and the contractivity (67). The heuristic argument is the following: pick (x_t) and (y_t) solutions of the EVI starting from $\overline{x}$, $\overline{y}$ respectively. Choose $y = y_t$ in the EVI for (x_t) to get

$$\frac{1}{2}\frac{d}{ds}|_{s=t} d^2(x_s, y_t) + \frac{\lambda}{2} d^2(x_t, y_t) + E(x_t) \le E(y_t).$$

Symmetrically we have

$$\frac{1}{2}\frac{d}{ds}|_{s=t} d^2(x_t, y_s) + \frac{\lambda}{2} d^2(x_t, y_t) + E(y_t) \le E(x_t).$$

Adding up these two inequalities we get

$$\frac{d}{dt} d^2(x_t, y_t) \le -2\lambda d^2(x_t, y_t), \qquad a.e.\ t.$$

The rigorous proof follows this line and uses a doubling of variables argument á la Kruzkhov.

Uniqueness and contraction then follow by the Gronwall lemma. □

4.3 Applications to the Wasserstein Case

The aim of this section is to apply the abstract theory developed in the previous one to the case of functionals on $(\mathscr{P}_2(\mathbb{R}^d), W_2)$. As we will see, various diffusion equations may be interpreted as Gradient Flows of appropriate energy functionals w.r.t. to the Wasserstein distance, and quantitive analytic properties of the solutions can be derived by this interpretation.

Most of what we are going to discuss here is valid in the more general contexts of Riemannian manifolds and Hilbert spaces, but the differences between these latter cases and the Euclidean one are mainly technical, thus we keep the discussion at a level of $\mathbb{R}^d$ to avoid complications that would just obscure the main ideas.

The secton is split in two subsections: in the first one we discuss the definition of subdifferential of a λ-geodesically convex functional on $\mathscr{P}_2(\mathbb{R}^d)$, which is based on the interpretation of $\mathscr{P}_2(\mathbb{R}^d)$ as a sort of Riemannian manifold as discussed in Sect. 3.3.2. In the second one we discuss three by now classical applications, for which the full power of the abstract theory can be used (i.e. we will have Gradient Flows in the EVI formulation).

Before developing this program, we want to informally discuss a fundamental example.

Let us consider the Entropy functional $E : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ defined by

$$E(\mu) := \begin{cases} \displaystyle\int \rho \log(\rho) d\mathscr{L}^d, & \text{if } \mu = \rho\mathscr{L}^d, \\ +\infty & \text{otherwise.} \end{cases}$$

We claim that: *the Gradient Flow of the Entropy in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ produces a solution of the Heat equation*. This can be proved rigorously (see Sect. 4.3.2), but for the moment we want to keep the discussion at the heuristic level.

By what discussed in the previous section, we know that the Minimizing Movements scheme produces Gradient Flows. Let us apply the scheme to this setting. Fix an absolutely continuous measure ρ_0 (here we will make no distinction between an absolutely continuous measure and its density), fix $\tau > 0$ and minimize

$$\mu \qquad \mapsto \qquad E(\mu) + \frac{W_2^2(\mu, \rho_0)}{2\tau}. \tag{73}$$

It is not hard to see that the minimum is attained at some absolutely continuous measure ρ_τ (actually the minimum is unique, but this has no importance). Our claim will be "proved" if we show that for any $\varphi \in C_c^\infty(\mathbb{R}^d)$ it holds

$$\frac{\int \varphi\rho_\tau - \int \varphi\rho_0}{\tau} = \int \Delta\varphi\, \rho_\tau + o(\tau), \tag{74}$$

because this identity tells us that ρ_τ is a first order approximation of the distributional solution of the Heat equation starting from ρ_0 and evaluated at time τ.

To prove (74), fix $\varphi \in C_c^\infty(\mathbb{R}^d)$ and perturb ρ_τ in the following way:

$$\rho^\varepsilon := (Id + \varepsilon\nabla\varphi)_\# \rho_\tau.$$

The density of ρ^ε can be explicitly expressed by

$$\rho^\varepsilon(x + \varepsilon\nabla\varphi(x)) = \frac{\rho_\tau(x)}{\det(Id + \varepsilon\nabla^2\varphi(x))}.$$

Observe that it holds

$$\begin{aligned} E(\rho^\varepsilon) &= \int \rho^\varepsilon \log(\rho^\varepsilon) = \int \rho_\tau \log\big(\rho^\varepsilon \circ (Id + \varepsilon\nabla\varphi)\big) = \int \rho_\tau \log\left(\frac{\rho_\tau}{\det(Id + \varepsilon\nabla^2\varphi)}\right) \\ &= E(\rho_\tau) - \int \rho_\tau \log\big(\det(Id + \varepsilon\nabla^2\varphi)\big) = E(\rho_\tau) - \varepsilon \int \rho_\tau \Delta\varphi + o(\varepsilon), \end{aligned} \tag{75}$$

where we used the fact that $\det(Id + \varepsilon A) = 1 + \varepsilon \mathrm{tr}(A) + o(\varepsilon)$.

To evaluate the first variation of the distance squared, let T be the optimal transport map from ρ_τ to ρ_0, which exists because of Theorem 2.26, and observe that from $T_\#\rho_\tau = \rho_0$, $(Id + \varepsilon\nabla\varphi)_\#\rho_\tau = \rho^\varepsilon$ and inequality (9) we have

$$W_2^2(\rho_0, \rho^\varepsilon) \le \|T - Id - \varepsilon\nabla\varphi\|^2_{L^2(\rho_\tau)},$$

therefore from the fact that equality holds at $\varepsilon = 0$ we get

$$\begin{aligned} W_2^2(\rho_0, \rho^\varepsilon) - W_2^2(\rho_0, \rho_\tau) &\le \|T - Id - \varepsilon\nabla\varphi\|^2_{L^2(\rho_\tau)} - \|T - Id\|^2_{L^2(\rho_\tau)} \\ &= -2\varepsilon \int \langle T - Id, \nabla\varphi\rangle \, \rho_\tau + o(\varepsilon). \end{aligned} \tag{76}$$

From the minimality of ρ_τ for the problem (73) we know that

$$E(\rho^\varepsilon) + \frac{W_2^2(\rho^\varepsilon, \rho_0)}{2\tau} \ge E(\rho_\tau) + \frac{W_2^2(\rho_\tau, \rho_0)}{2\tau}, \qquad \forall \varepsilon,$$

so that using (75) and (76), dividing by ε, rearranging the terms and letting $\varepsilon \downarrow 0$ and $\varepsilon \uparrow 0$ we get following Euler–Lagrange equation for ρ_τ:

$$\int \rho_\tau \Delta\varphi + \int \left\langle \frac{T - Id}{\tau}, \nabla\varphi \right\rangle \rho_\tau = 0. \tag{77}$$

Now observe that from $T_\#\rho_\tau = \rho_0$ we get

$$
\begin{aligned}
\frac{\int \varphi\rho_\tau - \int \varphi\rho_0}{\tau} &= -\frac{1}{\tau}\int \big(\varphi(T(x)) - \varphi(x)\big)\rho_\tau(x)dx \\
&= -\frac{1}{\tau}\iint_0^1 \langle \nabla\varphi((1-t)x + tT(x)), T(x) - x\rangle \, dt \, \rho_\tau(x)\, dx \\
&= -\frac{1}{\tau}\int \langle \nabla\varphi(x), T(x) - x\rangle \, \rho_\tau(x)\, dx + \mathrm{Rem}_\tau \\
&\overset{(77)}{=} \int \Delta\varphi \, \rho_\tau + \mathrm{Rem}_\tau,
\end{aligned}
$$

where the remainder term Rem_τ is bounded by

$$
|\mathrm{Rem}_\tau| \le \frac{\mathrm{Lip}(\nabla\varphi)}{\tau}\iint_0^1 t|T(x) - x|^2 dt \, \rho_\tau(x)\, dx = \frac{\mathrm{Lip}(\nabla\varphi)}{2\tau} W_2^2(\rho_0, \rho_\tau).
$$

Since, heuristically speaking, $W_2(\rho_0, \rho_\tau)$ has the same magnitude of τ, we have $\mathrm{Rem}_\tau = o(\tau)$ and the "proof" is complete.

4.3.1 Elements of Subdifferential Calculus in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$

Recall that we introduced a weak Riemannian structure on the space $(\mathscr{P}_2(M), W_2)$ in Sect. 3.3.2. Among others, this weak Riemannian structure of $(\mathscr{P}_2(M), W_2)$ allows the development of a *subdifferential calculus for geodesically convex functionals*, in the same spirit (and with many formal similarities) of the usual subdifferential calculus for convex functionals on an Hilbert space.

To keep the notation and the discussion simpler, we are going to define the subdifferential of a geodesically convex functional only for the case $\mathscr{P}_2(\mathbb{R}^d)$ and for regular measures (Definition 2.25), but everything can be done also on manifolds (or Hilbert spaces) and for general $\mu \in \mathscr{P}_2(M)$.

Recall that for a λ-convex functional F on an Hilbert space H, the subdifferential $\partial^- F(x)$ at a point x is the set of vectors $v \in H$ such that

$$
F(x) + \langle v, y - x\rangle + \frac{\lambda}{2}|x - y|^2 \le F(y), \qquad \forall y \in H.
$$

Definition 4.26 (Subdifferential in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$). Let $E : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ be a λ-geodesically convex and lower semicontinuous functional, and $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ be a regular measure such that $E(\mu) < \infty$. The set $\partial^W E(\mu) \subset \mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ is the set of vector fields $v \in L^2(\mu, \mathbb{R}^d)$ such that

$$
E(\mu) + \int \Big\langle T_\mu^\nu - Id, v\Big\rangle d\mu + \frac{\lambda}{2} W_2^2(\mu, \nu) \le E(\nu), \qquad \forall \nu \in \mathscr{P}_2(\mathbb{R}^d),
$$

where here and in the following T_μ^ν will denote the optimal transport map from the regular measure μ to ν (whose existence and uniqueness is guaranteed by Theorem 2.26).

Observe that the subdifferential of a λ-geodesically convex functional E has the following monotonicity property (which closely resembles the analogous valid for λ-convex functionals on an Hilbert space):

$$\int \Big\langle v, T_\mu^\nu - Id \Big\rangle \, d\mu + \int \langle w, T_\nu^\mu - Id \rangle \, d\nu \leq -\lambda W_2^2(\mu, \nu), \tag{78}$$

for every couple of regular measures μ, ν in the domain of E, and $v \in \partial^W E(\mu)$, $w \in \partial^W E(\nu)$. To prove (78) just observe that from the definition of subdifferential we have

$$E(\mu) + \int \Big\langle T_\mu^\nu - Id, v \Big\rangle \, d\mu + \frac{\lambda}{2} W_2^2(\mu, \nu) \leq E(\nu),$$
$$E(\nu) + \int \langle T_\nu^\mu - Id, w \rangle \, d\nu + \frac{\lambda}{2} W_2^2(\mu, \nu) \leq E(\mu),$$

and add up these inequalities.

The definition of subdifferential leads naturally to the definition of Gradient Flow: it is sufficient to transpose the definition given with the system (38).

Definition 4.27 (Subdifferential formulation of Gradient Flow). Let E be a λ-geodesically convex functional on $\mathscr{P}_2(\mathbb{R}^d)$ and $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. Then (μ_t) is a Gradient Flow for E starting from μ provided it is a locally absolutely continuous curve, $\mu_t \to \mu$ as $t \to 0$ w.r.t. the distance W_2, μ_t is regular for $t > 0$ and it holds

$$-v_t \in \partial^W E(\mu_t), \qquad a.e.\ t,$$

where (v_t) is the vector field uniquely identified by the curve (μ_t) via

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0,$$
$$v_t \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d)) \qquad a.e.\ t,$$

(recall Theorem 3.29 and Definition 3.31).

Thus we have a total of four different formulations of Gradient Flows of λ-geodesically convex functionals on $\mathscr{P}_2(\mathbb{R}^d)$ based respectively on the Energy Dissipation Inequality, the Energy Dissipation Equality, the Evolution Variational Inequality and the notion of subdifferential.

The important point is that these four formulations are *equivalent* for $\lambda-$geodesically convex functionals:

Proposition 4.28 (Equivalence of the various formulation of GF in the Wasserstein space). *Let E be a λ-geodesically convex functional on $\mathscr{P}_2(\mathbb{R}^d)$ and (μ_t) a curve made of regular measures. Then for (μ_t) the four definitions of Gradient Flow for E (EDI, EDE, EVI and the Subdifferential one) are equivalent.*

Sketch of the Proof We prove only that the EVI formulation is equivalent to the Subdifferential one. Recall that by Proposition 3.34 we know that

$$\frac{1}{2}\frac{d}{dt}W_2^2(\mu_t,\nu) = -\int \Big\langle v_t, T_{\mu_t}^{\nu} - Id \Big\rangle \, d\mu_t, \qquad a.e.t$$

where $T_{\mu_t}^{\nu}$ is the optimal transport map from μ_t to ν. Then we have

$$-v_t \in \partial^W E(\mu_t), \qquad a.e.\ t,$$

$$\Updownarrow$$

$$E(\mu_t) + \int \Big\langle -v_t, T_{\mu_t}^{\nu} - Id \Big\rangle \, d\mu_t + \frac{\lambda}{2} W_2^2(\mu_t,\nu) \le E(\nu), \qquad \forall \nu \in \mathscr{P}_2(\mathbb{R}^d),\ a.e.\ t$$

$$\Updownarrow$$

$$E(\mu_t) + \frac{1}{2}\frac{d}{dt}W_2^2(\mu_t,\nu) + \frac{\lambda}{2} W_2^2(\mu_t,\nu) \le E(\nu), \qquad \forall \nu \in \mathscr{P}_2(\mathbb{R}^d),\ a.e.\ t.$$

□

4.3.2 Three Classical Functionals

We now pass to the analysis of three by now classical examples of Gradient Flows in the Wasserstein space. Recall that in terms of strength, the best theory to use is the one of Sect. 4.2.4, because the compatibility in Energy and distance ensures strong properties both at the level of discrete solutions and for the limit curve obtained. Once we will have a Gradient Flow, the Subdifferential formulation will let us understand which is the PDE associated to it.

Let us recall (Example 3.21) that the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ is *not* Non Positively Curved in the sense of Alexandrov, this means that if we want to check whether a given functional is compatible with the distance or not, we cannot use geodesics to interpolate between points (because we would violate the second inequality in (65)). A priori the choice of the interpolating curves may depend on the functional, but actually in what comes next we will always use the ones defined by:

Definition 4.29 (Interpolating curves). Let $\mu, \nu_0, \nu_1 \in \mathscr{P}_2(\mathbb{R}^d)$ and assume that μ is regular (Definition 2.25). The interpolating curve (ν_t) from ν_0 to ν_1 with base μ is defined as

$$\nu_t := ((1-t)T_0 + tT_1)_{\#}\mu,$$

where T_0 and T_1 are the optimal transport maps from μ to ν_0 and ν_1 respectively. Observe that if $\mu = \nu_0$, the interpolating curve reduces to the geodesic connecting it to ν_1.

Strictly speaking, in order to apply the theory of Sect. 4.2.4 we should define interpolating curves having as base any measure $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, and not just regular ones. This is actually possible, and the foregoing discussion can be applied to the more general definition, but we prefer to avoid technicalities, and just focus on the main concepts.

For an interpolating curve as in the definition it holds:

$$W_2^2(\mu,\nu_t) \le (1-t)W_2^2(\mu,\nu_0) + tW_2^2(\mu,\nu_1) - t(1-t)W_2^2(\nu_0,\nu_1). \tag{79}$$

Indeed the map $(1-t)T_0 + tT_1$ is optimal from μ to ν_t (because we know that T_0 and T_1 are the gradients of convex functions φ_0, φ_1 respectively, thus $(1-t)T_0+tT_1$ is the gradient of the convex function $(1-t)\varphi_0 + t\varphi_1$, and thus is optimal), and we know by inequality (9) that $W_2^2(\nu_0,\nu_1) \le \|T_0 - T_1\|^2_{L^2(\mu)}$, thus it holds

$$\begin{aligned} W_2^2(\mu,\nu_t) &= \|(1-t)T_0 + tT_1\|^2_{L^2(\mu)} \\ &= (1-t)\|T_0 - Id\|^2_{L^2(\mu)} + t\|T_1 - Id\|^2_{L^2(\mu)} - t(1-t)\|T_0 - T_1\|^2_{L^2(\mu)} \\ &\le (1-t)W_2^2(\mu,\nu_0) + tW_2^2(\mu,\nu_1) - t(1-t)W_2^2(\nu_0,\nu_1). \end{aligned}$$

We now pass to the description of the three functionals we want to study.

Definition 4.30 (Potential energy). Let $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous and bounded from below. The potential energy functional $\mathscr{V} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ associated to V is defined by

$$\mathscr{V}(\mu) := \int V d\mu.$$

Definition 4.31 (Interaction energy). Let $W : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous, even and bounded from below. The interaction energy functional $\mathscr{W} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ associated to W is defined by

$$\mathscr{W}(\mu) := \frac{1}{2}\int W(x_1 - x_2) d\mu \times \mu(x_1, x_2).$$

Observe that the definition makes sense also for not even functions W; however, replacing if necessary the function $W(x)$ with $(W(x) + W(-x))/2$ we get an even function leaving the value of the functional unchanged.

Definition 4.32 (Internal energy). Let $u : [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ be a convex function bounded from below such that $u(0) = 0$ and

$$\varliminf_{z\to 0} \frac{u(z)}{z^\alpha} > -\infty, \qquad \text{for some } \alpha > \frac{d}{d+2}, \tag{80}$$

let $u'(\infty) := \lim_{z\to\infty} u(z)/z$. The internal energy functional $\mathscr{E}$ associated to u is

$$\mathscr{E}(\mu) := \int u(\rho)\mathscr{L}^d + u'(\infty)\mu^s(\mathbb{R}^d),$$

where $\mu = \rho\mathscr{L}^d + \mu^s$ is the decomposition of μ in absolutely continuous and singular parts w.r.t. the Lebesgue measure.

Condition (80) ensures that the negative part of $u(\rho)$ is integrable for $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, so that $\mathscr{E}$ is well defined (possibly $+\infty$). Indeed from (80) we have $u^-(z) \leq az + bz^\alpha$ for some $\alpha < 1$ satisfying $2\alpha/(1-\alpha) > d$, and it holds

$$\begin{aligned}\int \rho^\alpha(x) d\mathscr{L}^d(x) &= \int \rho^\alpha(x)(1+|x|)^{2\alpha}(1+|x|)^{-2\alpha} d\mathscr{L}^d(x) \\ &\leq \left(\int \rho(x)(1+|x|)^2 d\mathscr{L}^d(x)\right)^\alpha \left(\int (1+|x|)^{\frac{-2\alpha}{1-\alpha}} \mathscr{L}^d(x)\right)^{1-\alpha} < \infty.\end{aligned}$$

Under appropriate assumptions on V, W and e the above defined functionals are compatible with the distance W_2. As said before we will use as interpolating curves those given in Definition 4.29.

Proposition 4.33. *Let $\lambda \geq 0$. The following holds.*

(i) The functional $\mathscr{V}$ is λ-convex along interpolating curves in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ if and only if V is λ-convex.

(ii) The functional $\mathscr{W}$ is convex along interpolating curves $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ if W is convex.

(iii) The functional $\mathscr{E}$ is convex along interpolating curves $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ provided u satisfies

$$z \quad \mapsto \quad z^d u(z^{-d}) \qquad \text{is convex and non increasing on } (0, +\infty). \tag{81}$$

Proof. Since the second inequality in (65) is satisfied by the interpolating curves that we are considering (inequality (79)) we need only to check the convexity of the functionals.

Let (ν_t) be an interpolating curve with base the regular measure μ, and T_0, T_1 the optimal transport maps from μ to ν_0 and ν_1 respectively.

The *only if* part of (i) follows simply considering interpolation of deltas. For the *if*, observe that[5]

$$\begin{aligned}\mathscr{V}(\nu_t) &= \int V(x) d\nu_t(x) = \int V\big((1-t)T_0(x) + tT_1(x)\big) d\mu(x) \\ &\leq (1-t)\int V(T_0(x)) d\mu(x) + t\int V(T_1(x)) d\mu(x) - \frac{\lambda}{2} t(1-t) \int |T_0(x) - T_1(x)|^2 d\mu(x) \\ &\leq (1-t)\mathscr{V}(\nu_0) + t\mathscr{V}(\nu_1) - \frac{\lambda}{2} t(1-t) W_2^2(\nu_0, \nu_1).\end{aligned} \tag{82}$$

[5] The assumption $\lambda \geq 0$ is necessary to have the last inequality in (82). If $\lambda < 0$, $\lambda-$convexity of $\mathscr{V}$ along interpolating curves is not anymore true, so that we cannot apply directly the results of Sect. 4.2.4. Yet, adapting the arguments, it possible to show that all the results which we will present hereafter are true for general $\lambda \in \mathbb{R}$.

For (ii) we start claiming that $W_2^2(\mu \times \mu, \nu \times \nu) = 2W_2^2(\mu, \nu)$ for any $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$. To prove this, it is enough to check that if $\gamma \in \mathit{Opt}(\mu, \nu)$ then $\tilde{\gamma} := (\pi^1, \pi^1, \pi^2, \pi^2)_\# \gamma \in \mathit{Opt}(\mu \times \mu, \nu \times \nu)$. To see this, let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function such that $\mathrm{supp}(\gamma) \subset \partial^- \varphi$ and define the convex function $\tilde{\varphi}$ on $\mathbb{R}^{2d}$ by $\tilde{\varphi}(x, y) = \varphi(x) + \varphi(y)$. It is immediate to verify that $\mathrm{supp}(\tilde{\gamma}) \subset \partial^- \tilde{\varphi}$, so that $\tilde{\gamma}$ is optimal as well. This argument also shows that if (ν_t) is an interpolating curve with base μ, then $t \mapsto \nu_t \times \nu_t$ is an interpolating curve from $\nu_0 \times \nu_0$ to $\nu_1 \times \nu_1$ with base $\mu \times \mu$. Also, $(x_1, x_2) \mapsto W(x_1 - x_2)$ is convex if W is. The conclusion now follows from case (i).

We pass to (iii). We will make the simplifying assumption that $\mu \ll \mathscr{L}^d$ and that T_0 and T_1 are smooth and satisfy $\det(\nabla T_0)(x) \neq 0$, $\det(\nabla T_1)(x) \neq 0$ for every $x \in \mathrm{supp}(\mu)$ (up to an approximation argument, it is possible to reduce to this case, we omit the details). Then, writing $\mu = \rho \mathscr{L}^d$, from the change of variable formula we get that $\nu_t \ll \mathscr{L}^d$ and for its density $\tilde{\rho}_t$ it holds

$$\tilde{\rho}_t(T_t(x)) = \frac{\rho(x)}{\det(\nabla T_t(x))},$$

where we wrote T_t for $(1-t)T_0 + tT_1$. Thus

$$\mathscr{E}(\nu_t) = \int u(\tilde{\rho}_t(y))d\mathscr{L}^d(y) = \int u\left(\frac{\rho(x)}{\det(\nabla T_t)(x)}\right) \det(\nabla T_t)(x) d\mathscr{L}^d(x).$$

Therefore the proof will be complete if we show that $A \mapsto u(\frac{\rho(x)}{\det(A)}) \det(A)$ is convex on the set of positively defined symmetric matrices for any $x \in \mathrm{supp}(\mu)$. Observe that this map is the composition of the convex and non increasing map $z \mapsto z^d u(\rho(x)/z^d)$ with the map $A \mapsto (\det(A))^{1/d}$. Thus to conclude it is sufficient to show that $A \mapsto (\det(A))^{1/d}$ is concave. To this aim, pick two symmetric and positive definite matrices A_0 and A_1, notice that

$$\big(\det((1-t)A_0 + tA_1)\big)^{1/d} = \big(\det(A_0)\det(Id + tB)\big)^{1/d},$$

where $B = \sqrt{A_0}(A_1 - A_0)\sqrt{A_0}$ and conclude by

$$\begin{aligned}
\frac{d}{dt}\det(Id + tB)^{1/d} &= \frac{1}{d}\big(\det(Id + tB)\big)^{1/d} \mathrm{tr}\big(B\,(Id + tB)^{-1}\big),\\
\frac{d^2}{dt^2}\det(Id + tB)^{1/d} &= \frac{1}{d^2}\mathrm{tr}^2\big(B\,(Id + tB)^{-1}\big) - \frac{1}{d}\mathrm{tr}\big(\big(B\,(Id + tB)^{-1}\big)^2\big) \le 0
\end{aligned}$$

where in the last step we used the inequality $\mathrm{tr}^2(C) \le d\,\mathrm{tr}(C^2)$ for $C = B\,(Id + tB)^{-1}$. □

Important examples of functions u satisfying (80) and (81) are:

$$\begin{aligned}
u(z) &= \frac{z^\alpha - z}{\alpha - 1}, \qquad \alpha \ge 1 - \frac{1}{d},\ \alpha \neq 1\\
u(x) &= z\log(z).
\end{aligned} \tag{83}$$

Remark 4.34 (A dimension free condition on u). We saw that a sufficient condition on u to ensure that $\mathscr{E}$ is convex along interpolating curves is the fact that the map $z \mapsto z^d u(z^{-d})$ is convex and non increasing, so the dimension d of the ambient space plays a role in the condition. The fact that the map is non increasing follows by the convexity of u together with $u(0) = 0$, while by simple computations we see that its convexity is equivalent to

$$z^{-1}u(z) - u'(z) + zu''(z) \geq -\frac{1}{d-1}zu''(z). \tag{84}$$

Notice that the higher d is, the stricter the condition becomes. For applications in infinite dimensional spaces, it is desirable to have a condition on u ensuring the convexity of $\mathscr{E}$ in which the dimension does not enter. As inequality (84) shows, the weakest such condition for which $\mathscr{E}$ is convex in any dimension is:

$$z^{-1}u(z) - u'(z) + zu''(z) \geq 0,$$

and some computations show that this is in turn equivalent to the convexity of the map

$$z \quad \mapsto \quad e^z u(e^{-z}).$$

A key example of map satisfying this condition is $z \mapsto z\log(z)$. ■

Therefore we have the following existence and uniqueness result:

Theorem 4.35. *Let $\lambda \geq 0$ and $\mathscr{F}$ be either $\mathscr{V}$, $\mathscr{W}$, $\mathscr{E}$ (or a linear combination of them with positive coefficients) and λ-convex along interpolating curves. Then for every $\overline{\mu} \in \mathscr{P}_2(\mathbb{R}^d)$ there exists a unique Gradient Flow (μ_t) for $\mathscr{F}$ starting from $\overline{\mu}$ in the EVI formulation. The curve (μ_t) satisfies: is locally absolutely continuous on $(0, +\infty)$, $\mu_t \to \overline{\mu}$ as $t \to 0$ and, if μ_t is regular for every $t \geq 0$, it holds*

$$-v_t \in \partial^W F(\mu_t), \qquad a.e.\ t \in (0, +\infty), \tag{85}$$

where (v_t) is the velocity vector field associated to (μ_t) characterized by

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0,$$
$$v_t \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d)) \qquad a.e.\ t.$$

Proof. Use the existence Theorem 4.25 and the equivalence of the EVI formulation of Gradient Flow and the Subdifferential one provided by Proposition 4.28. □

It remains to understand which kind of equation is satisfied by the Gradient Flow (μ_t). By (85), this corresponds to identify the subdifferentials of $\mathscr{V}$, $\mathscr{W}$, $\mathscr{E}$ at a generic $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. This is the content of the next three propositions. For simplicity, we state and prove them only under some—unneeded—smoothness assumptions. The underlying idea of all the calculations we are going to do is the following equivalence:

$$v \in \partial^W \mathscr{F}(\mu) \quad \overset{\approx}{\Leftrightarrow} \quad \lim_{\varepsilon \to 0} \frac{\mathscr{F}((Id + \varepsilon \nabla\varphi)_\# \mu) - \mathscr{F}(\mu)}{\varepsilon} = \int \langle v, \nabla\varphi \rangle \, , \ \forall \varphi \in C_c^\infty(\mathbb{R}^d), \tag{86}$$

valid for any λ-geodesically convex functional, where we wrote $\overset{\approx}{\Leftrightarrow}$ to intend that this equivalence holds only when everything is smooth. To understand why (86) holds, start assuming that $v \in \partial^W F(\mu)$, fix $\varphi \in C_c^\infty(\mathbb{R}^d)$ and recall that for ε sufficiently small the map $Id + \varepsilon\nabla\varphi$ is optimal (Remark 2.22). Thus by definition of subdifferential we have

$$\mathscr{F}(\mu) + \varepsilon \int \langle v, \nabla\varphi \rangle \, d\mu + \varepsilon^2 \frac{\lambda}{2} \|\nabla\varphi\|^2_{L^2(\mu)} \leq \mathscr{F}((Id + \varepsilon\nabla\varphi)_\# \mu).$$

Subtracting $\mathscr{F}(\mu)$ on both sides, dividing by $\varepsilon > 0$ and $\varepsilon < 0$ and letting $\varepsilon \to 0$ we get the implication $\Rightarrow$. To "prove" the converse one, pick $\nu \in \mathscr{P}_2(\mathbb{R}^d)$, let T be the optimal transport map from μ to ν and recall that T is the gradient of a convex function ϕ. Assume that ϕ is smooth and define $\varphi(x) := \phi(x) - |x|^2/2$. The geodesic (μ_t) from μ to ν can then be written as

$$\mu_t = \big((1-t)Id + tT\big)_\# \mu = \big((1-t)Id + t\nabla\phi\big)_\# \mu = \big(Id + t\nabla\varphi\big)_\# \mu.$$

From the λ-convexity hypothesis we know that

$$\mathscr{F}(\nu) \geq \mathscr{F}(\mu) + \frac{d}{dt}|_{t=0} \mathscr{F}(\mu_t) + \frac{\lambda}{2} W_2^2(\mu, \nu),$$

therefore, since we know that $\frac{d}{dt}|_{t=0}\mathscr{F}(\mu_t) = \int \langle v, \nabla\varphi \rangle \, d\mu$, from the arbitrariness of ν we deduce $v \in \partial^W \mathscr{F}(\mu)$.

Proposition 4.36 (Subdifferential of $\mathscr{V}$). *Let $V : \mathbb{R}^d \to \mathbb{R}$ be λ-convex and C^1, let $\mathscr{V}$ be as in Definition 4.30 and let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ be regular and satisfying $\mathscr{V}(\mu) < \infty$. Then $\partial^W \mathscr{V}(\mu)$ is non empty if and only if $\nabla V \in L^2(\mu)$, and in this case ∇V is the only element in the subdifferential of $\mathscr{V}$ at μ.*

Therefore, if (μ_t) is a Gradient Flow of $\mathscr{V}$ made of regular measures, it solves

$$\frac{d}{dt}\mu_t = \nabla \cdot (\nabla V \mu_t),$$

in the sense of distributions in $\mathbb{R}^d \times (0, +\infty)$.

Sketch of the Proof Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$ and observe that

$$\lim_{\varepsilon \to 0} \frac{\mathscr{V}((Id + \varepsilon\nabla\varphi)_\# \mu) - \mathscr{V}(\mu)}{\varepsilon} = \lim_{\varepsilon \to 0} \int \frac{V \circ (Id + \varepsilon\nabla\varphi) - V}{\varepsilon} d\mu = \int \langle \nabla V, \nabla\varphi \rangle \, d\mu.$$

Conclude using the equivalence (86). □

Proposition 4.37 (Subdifferential of $\mathscr{W}$). *Let $W : \mathbb{R}^d \to \mathbb{R}$ be convex, even and C^1, let $\mathscr{W}$ be defined by Definition 4.31 and μ be regular and satisfying $\mathscr{W}(\mu) < \infty$. Then $\partial^W \mathscr{W}(\mu) \neq \emptyset$ if and only if $(\nabla W) * \mu$ belongs to $L^2(\mu)$ and in this case $(\nabla W) * \mu$ is the only element in the subdifferential of $\mathscr{W}$ at μ.*

Therefore, if (μ_t) is a Gradient Flow of $\mathscr{W}$ made of regular measures, it solves the non local evolution equation

$$\frac{d}{dt}\mu_t = \nabla \cdot ((\nabla W * \mu_t)\mu_t),$$

in the sense of distributions in $\mathbb{R}^d \times (0, +\infty)$.

Sketch of the Proof Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$, let $\mu^\varepsilon := (Id + \varepsilon\nabla\varphi)_\#\mu$ and observe that

$$\begin{aligned}\mathscr{W}(\mu^\varepsilon) &= \frac{1}{2}\int W(x-y)d\mu^\varepsilon(x)d\mu^\varepsilon(y)\\ &= \frac{1}{2}\int W(x-y+\varepsilon(\nabla\varphi(x)-\nabla\varphi(y)))d\mu(x)d\mu(y)\\ &= \frac{1}{2}\int W(x-y)d\mu(x)d\mu(y)\\ &\quad + \frac{\varepsilon}{2}\int \langle \nabla W(x-y), \nabla\varphi(x)-\nabla\varphi(y)\rangle \, d\mu(x)d\mu(y) + o(\varepsilon).\end{aligned}$$

Now observe that

$$\begin{aligned}\int \langle \nabla W(x-y), \nabla\varphi(x)\rangle \, d\mu(x)d\mu(y) &= \int \left\langle \int \nabla W(x-y)d\mu(y), \nabla\varphi(x)\right\rangle d\mu(x)\\ &= \int \langle \nabla W * \mu(x), \nabla\varphi(x)\rangle \, d\mu(x),\end{aligned}$$

and, similarly,

$$\begin{aligned}\int \langle \nabla W(x-y), -\nabla\varphi(y)\rangle \, d\mu(x)d\mu(y) &= \int \langle \nabla W * \mu(y), \nabla\varphi(y)\rangle \, d\mu(y)\\ &= \int \langle \nabla W * \mu(x), \nabla\varphi(x)\rangle \, d\mu(x).\end{aligned}$$

Thus the conclusion follows by applying the equivalence (86). □

Proposition 4.38 (Subdifferential of $\mathscr{E}$). *Let $u : [0, +\infty) \to \mathbb{R}$ be convex, C^2 on $(0, +\infty)$, bounded from below and satisfying conditions (80) and (81). Let $\mu = \rho\mathscr{L}^d \in \mathscr{P}_2(\mathbb{R}^d)$ be an absolutely continuous measure with smooth density. Then $\nabla(u'(\rho))$ is the unique element in $\partial^W \mathscr{E}(\mu)$.*

Therefore, if (μ_t) is a Gradient Flow for $\mathscr{E}$ and μ_t is absolutely continuous with smooth density ρ_t for every $t > 0$, then $t \mapsto \rho_t$ solves the equation

$$\frac{d}{dt}\rho_t = \nabla \cdot (\rho_t \nabla(u'(\rho_t))).$$

Note: this statement is not perfectly accurate, because we are neglecting the integrability issues. Indeed a priori we don't know that $\nabla(u'(\rho))$ belongs to $L^2(\mu)$.

Sketch of the Proof Fix $\varphi \in C_c^\infty(\mathbb{R}^d)$ and define $\mu^\varepsilon := (Id + \varepsilon\nabla\varphi)_\#\mu$. For ε sufficiently small, μ^ε is absolutely continuous and its density ρ^ε satisfies—by the change of variable formula—the identity

$$\rho^\varepsilon(x + \varepsilon\nabla\varphi(x)) = \frac{\rho(x)}{\det(Id + \varepsilon\nabla^2\varphi(x))}.$$

Using the fact that $\frac{d}{d\varepsilon}|_{\varepsilon=0}(\det(Id + \varepsilon\nabla^2\varphi(x))) = \Delta\varphi(x)$ we have

$$\begin{aligned}
\frac{d}{d\varepsilon}|_{\varepsilon=0}\mathscr{E}(\mu^\varepsilon) &= \frac{d}{d\varepsilon}|_{\varepsilon=0}\int u(\rho^\varepsilon(y))dy \\
&= \frac{d}{d\varepsilon}|_{\varepsilon=0}\int u\left(\frac{\rho(x)}{\det(Id + \varepsilon\nabla^2\varphi(x))}\right)\det(Id + \varepsilon\nabla^2\varphi(x))dx \\
&= \int -\rho u'(\rho)\Delta\varphi + u(\rho)\Delta\varphi = \int \left\langle \nabla(\rho u'(\rho) - u(\rho)), \nabla\varphi \right\rangle \\
&= \int \left\langle \nabla(u'(\rho)), \nabla\varphi \right\rangle \rho,
\end{aligned}$$

and the conclusion follows by the equivalence (86). □

As an example, let $u(z) := z\log(x)$, and let V be a λ-convex smooth function on $\mathbb{R}^d$. Since $u'(z) = \log(z) + 1$, we have $\rho\nabla(u'(\rho)) = \Delta\rho$, thus a gradient flow (ρ_t) of $\mathscr{F} = \mathscr{E} + \mathscr{V}$ solves the Fokker–Plank equation

$$\frac{d}{dt}\rho_t = \Delta\rho_t + \nabla \cdot (\nabla V \rho_t).$$

Also, the contraction property (67) in Theorem 4.25 gives that for two gradient flows (ρ_t), $(\tilde{\rho}_t)$ it holds the contractivity estimate

$$W_2(\rho_t, \tilde{\rho}_t) \le e^{-\lambda t} W_2(\rho_0, \tilde{\rho}_0).$$

4.4 Bibliographical Notes

The content of Sect. 4.2 is taken from the first part of [7] (we refer to this book for a detailed bibliographical references on the topic of gradient flows in metric spaces), with the only exception of Proposition 4.6, whose proof has been communicated to us by Savaré (see also [72, 73]).

The study of geodesically convex functionals in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ has been introduced by R. McCann in [63], who also proved that conditions (80) and (81) were sufficient to deduce the geodesic convexity (called by him displacement convexity) of the internal energy functional.

The study of gradient flows in the Wasserstein space began in the seminal paper by R. Jordan et al. [47], where it was proved that the minimizing movements procedure for the functional

$$\rho\mathscr{L}^d \qquad \mapsto \qquad \int \rho \log \rho + V\rho d\mathscr{L}^d,$$

on the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$, produce solutions of the Fokker–Planck equation. Later, F. Otto in [67] showed that the same discretization applied to

$$\rho\mathscr{L}^d \qquad \mapsto \qquad \frac{1}{\alpha - 1}\int \rho^\alpha d\mathscr{L}^d,$$

(with the usual meaning for measures with a singular part) produce solutions of the porous medium equation. The impact of Otto's work on the community of optimal transport has been huge: not only he was able to provide concrete consequences (in terms of new estimates for the rate of convergence of solutions of the porous medium equation) out of optimal transport theory, but he also clearly described what is now called the "weak Riemannian structure" of $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ (see also chapter "Self-organized Network Flows" and Sect. 3.3.2).

Otto's intuitions have been studied and extended by many authors. The rigorous description of many of the objects introduced by Otto, as well as a general discussion about gradient flows of λ-geodesically convex functionals on $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ has been done in the second part of [7] (the discussion made here is taken from this latter reference).

5 Geometric and Functional Inequalities

In this short chapter we show how techniques coming from optimal transport can lead to simple proofs of some important geometric and functional inequalities. None of the results proven here are new, in the sense that they all were well known before the proofs coming from optimal transport appeared. Still, it is interesting to observe how the tools described in the previous sections allow to produce proofs which are occasionally simpler and in any case providing new informations when compared to the "standard" ones.

5.1 Brunn–Minkowski Inequality

Recall that the Brunn–Minkowski inequality in $\mathbb{R}^d$ is:

$$\left(\mathscr{L}^d\left(\frac{A+B}{2}\right)\right)^{1/d} \geq \frac{1}{2}\left(\left(\mathscr{L}^d(A)\right)^{1/d} + \left(\mathscr{L}^d(B)\right)^{1/d}\right),$$

and is valid for any couple of compact sets $A,\ B \subset \mathbb{R}^d$.

To prove it, let $A, B \subset \mathbb{R}^d$ be compact sets and notice that without loss of generality we can assume that $\mathscr{L}^d(A), \mathscr{L}^d(B) > 0$. Define

$$\mu_0 := \frac{1}{\mathscr{L}^d(A)}\mathscr{L}^d|_A \qquad \mu_1 := \frac{1}{\mathscr{L}^d(B)}\mathscr{L}^d|_B,$$

and let (μ_t) be the unique geodesic in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ connecting them.

Recall from (83) that for $u(z) = -d(z^{1-1/d} - z)$ the functional $\mathscr{E}(\rho) := \int u(\rho) d\mathscr{L}^d$ is geodesically convex in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$. Also, simple calculations show that $\mathscr{E}(\mu_0) = -d(\mathscr{L}^d(A)^{1/d} - 1)$, $\mathscr{E}(\mu_1) = -d(\mathscr{L}^d(B)^{1/d} - 1)$. Hence we have

$$\mathscr{E}(\mu_{1/2}) \leq -\frac{d}{2}\left(\left(\mathscr{L}^d(A)\right)^{1/d} + \left(\mathscr{L}^d(B)\right)^{1/d}\right) + d.$$

Now notice that Theorem 3.10 (see also Remark 3.13) ensures that $\mu_{1/2}$ is concentrated on $\frac{A+B}{2}$, thus letting $\tilde{\mu}_{1/2} := (\mathscr{L}^d((A+B)/2))^{-1}\mathscr{L}^d|_{(A+B)/2}$ and applying Jensen's inequality to the convex function u we get

$$\mathscr{E}(\mu_{1/2}) \geq \mathscr{E}(\tilde{\mu}_{1/2}) = -d\left(\mathscr{L}^d\left(\frac{A+B}{2}\right)^{1/d} - 1\right),$$

which concludes the proof.

5.2 Isoperimetric Inequality

On $\mathbb{R}^d$ the isoperimetric inequality can be written as

$$\mathscr{L}^d(E)^{1-\frac{1}{d}} \leq \frac{P(E)}{d\mathscr{L}^d(B)^{\frac{1}{d}}},$$

where E is an arbitrary open set, $P(E)$ its perimeter and B the unitary ball.

We will prove this inequality via Brenier's Theorem 2.26, neglecting all the smoothness issues. Let

$$\mu := \frac{1}{\mathscr{L}^d(E)}\mathscr{L}^d|_E, \qquad \nu := \frac{1}{\mathscr{L}^d(B)}\mathscr{L}^d|_B,$$

and $T : E \to B$ be the optimal transport map (w.r.t. the cost given by the distance squared). The change of variable formula gives

$$\frac{1}{\mathscr{L}^d(E)} = \det(\nabla T(x))\frac{1}{\mathscr{L}^d(B)}, \qquad \forall x \in E.$$

Since we know that T is the gradient of a convex function, we have that $\nabla T(x)$ is a symmetric matrix with non negative eigenvalues for every $x \in E$. Hence the arithmetic-geometric mean inequality ensures that

$$(\det \nabla T(x))^{1/d} \le \frac{\nabla \cdot T(x)}{d}, \qquad \forall x \in E.$$

Coupling the last two equations we get

$$\frac{1}{\mathscr{L}^d(E)^{\frac{1}{d}}} \le \frac{\nabla \cdot T(x)}{d}\frac{1}{\mathscr{L}^d(B)^{\frac{1}{d}}} \qquad \forall x \in E.$$

Integrating over E and applying the divergence theorem we get

$$\mathscr{L}^d(E)^{1-\frac{1}{d}} \le \frac{1}{d\mathscr{L}^d(B)^{1/d}}\int_E \nabla \cdot T(x)dx = \frac{1}{d\mathscr{L}^d(B)^{1/d}}\int_{\partial E}\langle T(x), \nu(x)\rangle \, d\mathscr{H}^{d-1}(x),$$

where $\nu : \partial E \to \mathbb{R}^d$ is the outer unit normal vector. Since $T(x) \in B$ for every $x \in E$, we have $|T(x)| \le 1$ for $x \in \partial E$ and thus $\langle T(x), \nu(x)\rangle \le 1$. We conclude with

$$\mathscr{L}^d(E)^{1-\frac{1}{d}} \le \frac{1}{d\mathscr{L}^d(B)^{1/d}}\int_{\partial E}\langle T(x), \nu(x)\rangle \, d\mathscr{H}^{d-1}(x) \le \frac{P(E)}{d\mathscr{L}^d(B)^{1/d}}.$$

5.3 Sobolev Inequality

The Sobolev inequality in $\mathbb{R}^d$ reads as:

$$\left(\int |f|^{p^*}\right)^{1/p^*} \le C(d,p)\left(\int |\nabla f|^p\right)^{1/p}, \qquad \forall f \in W^{1,p}(\mathbb{R}^d),$$

where $1 \le p < d$, $p^* := \frac{dp}{d-p}$ and $C(d,p)$ is a constant which depends only on the dimension d and the exponent p.

We will prove it via a method which closely resemble the one just used for the isoperimetric inequality. Again, we will neglect all the smoothness issues. Fix d, p and observe that without loss of generality we can assume $f \ge 0$ and $\int |f|^{p*} = 1$, so that our aim is to prove that

$$\left(\int |\nabla f|^p\right)^{1/p} \ge C, \tag{87}$$

for some constant C not depending on f. Fix once and for all a smooth, non negative function $g : \mathbb{R}^d \to \mathbb{R}$ satisfying $\int g = 1$, define the probability measures

$$\mu := f^{p^*}\mathscr{L}^d, \qquad \nu := g\mathscr{L}^d,$$

and let T be the optimal transport map from μ to ν (w.r.t. the cost given by the distance squared). The change of variable formula gives

$$g(T(x)) = \frac{f^{p^*}(x)}{\det(\nabla T(x))}, \qquad \forall x \in \mathbb{R}^d.$$

Hence we have

$$\int g^{1-\frac{1}{d}} = \int g^{-\frac{1}{d}} g = \int (g \circ T)^{-\frac{1}{d}} f^{p^*} = \int \det(\nabla T)^{\frac{1}{d}} (f^{p^*})^{1-\frac{1}{d}}.$$

As for the case of the isoperimetric inequality, we know that T is the gradient of a convex function, thus $\nabla T(x)$ is a symmetric matrix with non negative eigenvalues and the arithmetic-geometric mean inequality gives $(\det(\nabla T(x)))^{1/d} \leq \frac{\nabla\cdot T(x)}{d}$. Thus we get

$$\int g^{1-\frac{1}{d}} \leq \frac{1}{d}\int \nabla \cdot T (f^{p^*})^{1-\frac{1}{d}} = -\frac{p^*}{d}\left(1-\frac{1}{d}\right)\int f^{\frac{p^*}{q}} T \cdot \nabla f,$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Finally, by Hölder inequality we have

$$\begin{aligned}\int g^{1-\frac{1}{d}} &\leq \frac{p^*}{d}\left(1-\frac{1}{d}\right)\left(\int f^{p^*}|T|^q\right)^{\frac{1}{q}}\left(\int |\nabla f|^p\right)^{\frac{1}{p}} \\ &= \frac{p^*}{d}\left(1-\frac{1}{d}\right)\left(\int g(y)|y|^q dy\right)^{\frac{1}{q}}\left(\int |\nabla f|^p\right)^{\frac{1}{p}}.\end{aligned}$$

Since g was a fixed given function, (87) is proved.

5.4 Bibliographical Notes

The possibility of proving Brunn–Minkowski inequality via a change of variable is classical. It has been McCann in his PhD thesis [62] to notice that the use of optimal transport leads to a natural choice of reparametrization. It is interesting to notice that this approach can be generalized to curved and non-smooth spaces having *Ricci curvature bounded below*, see Proposition 8.14.

The idea of proving the isoperimetric inequality via a change of variable argument is due to Gromov [65]: in Gromov's proof it is not used the optimal transport map, but the so called Knothe's map. Such a map has the property that its gradient has non negative eigenvalues at every point, and the reader can easily check that this is all we used of Brenier's map in our proof, so that the argument of Gromov is the same we used here. The use of Brenier's map instead of Knothe's one makes the difference when studying the quantitative version of the isoperimetric problem: Figalli et al. in [38], using tools coming from optimal transport, proved the sharp quantitative isoperimetric inequality in $\mathbb{R}^d$ endowed with any norm (the sharp quantitative isoperimetric inequality for the Euclidean norm was proved earlier by Fusco et al. in [40] by completely different means).

The approach used here to prove the Sobolev inequality has been generalized by Cordero-Erasquin, Nazaret and Villani in [30] to provide a new proof of the sharp Gagliardo–Nirenberg–Sobolev inequality together with the identification of the functions realizing the equality.

6 Variants of the Wasserstein Distance

In this chapter we make a quick overview of some variants of the Wasserstein distance W_2 together with their applications. No proofs will be reported: our goal here is only to show that concepts coming from the transport theory can be adapted to cover a broader range of applications.

6.1 Branched Optimal Transportation

Consider the transport problem with $\mu := \delta_x$ and $\nu := \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$ for the cost given by the distance squared on $\mathbb{R}^d$. Then Theorem 3.10 and Remark 3.13 tell that the unique geodesic (μ_t) connecting μ to ν is given by

$$\mu_t := \frac{1}{2}\Big(\delta_{(1-t)x+ty_1} + \delta_{(1-t)x+ty_2}\Big),$$

so that the geodesic produces a "V-shaped" path.

For some applications, this is unnatural: for instance in real life networks, when one wants to transport the good located in x to the destinations y_1 and y_2 it is preferred to produce a branched structure, where first the good it is transported "on a single truck" to some intermediate point, and only later split into two parts which are delivered to the two destinations. This produces a "Y-shaped" path.

If we want to model the fact that "it is convenient to ship things together", we are lead to the following construction, due to Gilbert. Say that the starting distribution of mass is given by $\mu = \sum_i a_i\delta_{x_i}$ and that the final one is $\nu = \sum_j b_j\delta_{y_j}$, with $\sum_i a_i = \sum_j b_j = 1$. An admissible dynamical transfer is then given by a finite, oriented, weighted graph G, where the weight is a function $w : \{\textit{set of edges of } G\} \to \mathbb{R}$, satisfying the Kirchoff's rule:

$$\sum_{\text{edges}e\text{outgoing from}x_i} w(e) - \sum_{\text{edges}e\text{incoming in}x_i} w(e) = a_i, \qquad \forall i$$

$$\sum_{\text{edges}e\text{outgoing from}y_j} w(e) - \sum_{\text{edges}e\text{incoming in}y_j} w(e) = -b_j, \qquad \forall j$$

$$\sum_{\text{edges}e\text{outgoing from}z} w(e) - \sum_{\text{edges}e\text{incoming in}z} w(e) = 0, \quad \text{for any ``internal'' node}z\text{of}G$$

Then for $\alpha \in [0, 1]$ one minimizes

$$\sum_{\text{edges}e\text{of}G} w^{\alpha}(e) \cdot \text{length}(e),$$

among all admissible graphs G.

Observe that for $\alpha = 0$ this problem reduces to the classical Steiner problem, while for $\alpha = 1$ it reduces to the classical optimal transport problem for *cost* $=$ *distance*.

It is not hard to show the existence of a minimizer for this problem. What is interesting, is that a "continuous" formulation is possible as well, which allows to discuss the minimization problem for general initial and final measure in $\mathscr{P}(\mathbb{R}^d)$.

Definition 6.1 (Admissible continuous dynamical transfer). Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$. An admissible continuous dynamical transfer from μ to ν is given by a countably $\mathscr{H}^1$-rectifiable set Γ, an orientation on it $\tau : \Gamma \to S^{d-1}$, and a weight function $w : \Gamma \to [0, +\infty)$, such that the $\mathbb{R}^d$ valued measure $J_{\Gamma,\tau,w}$ defined by

$$J_{\Gamma,\tau,w} := w\tau\mathscr{H}^1|_{\Gamma},$$

satisfies

$$\nabla \cdot J_{\Gamma,\tau,w} = \nu - \mu,$$

(which is the natural generalization of the Kirchoff rule).

Given $\alpha \in [0, 1]$, the cost function associated to (Γ, τ, w) is defined as

$$\mathscr{E}_{\alpha}(J_{\Gamma,\tau,w}) := \int_{\Gamma} w^{\alpha} \, d\mathscr{H}^1.$$

Theorem 6.2 (Existence). *Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ with compact support. Then for all $\alpha \in [0, 1)$ there exists a minimizer of the cost in the set of admissible continuous dynamical transfers connecting μ to ν. If $\mu = \delta_z$ and $\nu = \mathscr{L}^d|_{[0,1]^d}$, the minimal cost is finite if and only if $\alpha > 1 - 1/d$.*

The fact that $1 - 1/d$ is a limit value to get a finite cost, can be heuristically understood by the following calculation. Suppose we want to move a Delta mass δ_x into the Lebesgue measure on a unit cube whose center is x. Then the first thing one wants to try is: divide the cube into 2^d cubes of side length $1/2$, then split the delta into 2^d masses and let them move onto the centers of these 2^d cubes. Repeat the process by dividing each of the 2^d cubes into 2^d cubes of side length $1/4$ and so on. The total cost of this dynamical transfer is proportional to:

$$\sum_{i=1}^{\infty} \underbrace{2^{id}}_{\substack{\text{number of segments}\\ \text{at the step } i}} \underbrace{\frac{1}{2^i}}_{\substack{\text{length of each}\\ \text{segment at the step } i}} \underbrace{\frac{1}{2^{\alpha i d}}}_{\substack{\text{weighted mass on each}\\ \text{segment at the step } i}} = \sum_{i=1}^{\infty} 2^{i(d-1-\alpha d)},$$

which is finite if and only if $d-1-\alpha d < 0$, that is, if and only if $\alpha > 1 - \frac{1}{d}$.

A regularity result holds for $\alpha \in (1-1/d, 1)$ which states that far away from the supports of the starting and final measures, any minimal transfer is actually a finite tree:

Theorem 6.3 (Regularity). *Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ with compact support, $\alpha \in (1 - 1/n, 1)$ and let (Γ, τ, w) be a continuous tree with minimal α-cost between μ and ν. Then Γ is locally a finite tree in $\mathbb{R}^d \setminus (\operatorname{supp}\mu \cup \operatorname{supp}\nu)$.*

6.2 Different Action Functional

Let us recall that the Benamou–Brenier formula (Proposition 3.30) identifies the squared Wasserstein distance between $\mu^0 = \rho^0\mathscr{L}^d, \mu^1 := \rho^1\mathscr{L}^d \in \mathscr{P}_2(\mathbb{R}^d)$ by

$$W_2^2(\mu^0, \mu^1) = \inf \int_0^1 \int |v_t|^2(x)\rho_t(x)d\mathscr{L}^d(x)dt,$$

where the infimum is taken among all the distributional solutions of the continuity equation

$$\frac{d}{dt}\rho_t + \nabla \cdot (v_t\rho_t) = 0,$$

with $\rho_0 = \rho^0$ and $\rho_1 = \mu^1$.

A natural generalization of the distance W_2 comes by considering a new action, modified by putting a weight on the density, that is: given a smooth function $h : [0, \infty) \to [0, \infty)$ we define

$$W_h^2(\rho^0\mathscr{L}^d, \rho^1\mathscr{L}^d) = \inf \int_0^1 \int |v_t|^2(x)h(\rho_t(x))d\mathscr{L}^d(x)dt, \tag{88}$$

where the infimum is taken among all the distributional solutions of the *non linear* continuity equation

$$\frac{d}{dt}\rho_t + \nabla \cdot (v_t h(\rho_t)) = 0, \tag{89}$$

with $\rho_0 = \rho^0$ and $\rho_1 = \rho^1$.

The key assumption that leads to the existence of an action minimizing curve is the concavity of h, since this leads to the joint convexity of

$$(\rho, J) \mapsto h(\rho) \left| \frac{J}{h(\rho)} \right|^2,$$

so that using this convexity with $J = vh(\rho)$, one can prove existence of minima of (88). Particularly important is the case given by $h(z) := z^\alpha$ for $\alpha < 1$ from which we can build the distance $\tilde{W}_\alpha$ defined by

$$\tilde{W}_\alpha(\rho^0 \mathscr{L}^d, \rho^1 \mathscr{L}^d) := \left(\inf \int_0^1 \int |v_t|^2(x) \rho_t^{2-\alpha}(x) d\mathscr{L}^d(x) dt \right)^{\frac{1}{\alpha}}, \tag{90}$$

the infimum being taken among all solutions of (89) with $\rho_0 = \rho^0$ and $\rho_1 = \rho^1$. The following theorem holds:

Theorem 6.4. *Let $\alpha > 1 - \frac{1}{d}$. Then the infimum in (90) is always reached and, if it is finite, the minimizer is unique. Now fix a measure $\mu \in \mathscr{P}(\mathbb{R}^d)$. The set of measures ν with $\tilde{W}_\alpha(\mu, \nu) < \infty$ endowed with $\tilde{W}_\alpha$ is a complete metric space and bounded subsets are narrowly compact.*

We remark that the behavior of action minimizing curves in this setting is, in some very rough sense, "dual" of the behavior of the branched optimal transportation discussed in the previous section. Indeed, in this problem the mass tends to spread out along an action minimizing curve, rather than to glue together.

6.3 An Extension to Measures with Unequal Mass

Let us come back to the Heat equation seen as Gradient Flow of the entropy functional $E(\rho) = \int \rho \log(\rho)$ with respect to the Wasserstein distance W_2, as discussed at the beginning of Sect. 4.3 and in Sect. 4.3.2. We discussed the topic for arbitrary probability measures in $\mathbb{R}^d$, but actually everything could have been done for probability measures concentrated on some open bounded set $\Omega \subset \mathbb{R}^d$ with smooth boundary, that is: consider the metric space $(\mathscr{P}(\Omega), W_2)$ and the entropy functional $E(\rho) = \int \rho \log(\rho)$ for absolutely continuous measures and $E(\mu) = +\infty$ for measures with a singular part. Now use the Minimizing Movements scheme to build up a family of discrete solutions ρ_t^τ starting from some given measure $\overline{\rho} \in \mathscr{P}(\Omega)$. It is then possible to see that these discrete families converge as $\tau \downarrow 0$ to the solution of the Heat equation with *Neumann boundary condition*:

$$\begin{cases} \frac{d}{dt}\rho_t = \Delta \rho_t, & \text{in } \Omega \times (0, +\infty), \\ \rho_t \to \overline{\rho}, & \text{weakly as } t \to 0 \\ \nabla \rho_t \cdot \eta = 0, & \text{in } \partial\Omega \times (0, \infty), \end{cases}$$

where η is the outward pointing unit vector on $\partial\Omega$.

The fact that the boundary condition is the Neumann's one, can be heuristically guessed by the fact that working in $\mathscr{P}(\Omega)$ enforces the mass to be constant, with no flow of the mass through the boundary.

It is then natural to ask whether it is possible to modify the transportation distance in order to take into account measures with unequal masses, and such that the Gradient Flow of the entropy functional produces solutions of the Heat equation in Ω with Dirichlet boundary conditions. This is actually doable, as we briefly discuss now.

Let $\Omega \subset \mathbb{R}^d$ be open and bounded. Consider the set $\mathscr{M}_2(\Omega)$ defined by

$$\mathscr{M}_2(\Omega) := \left\{\text{measures } \mu \text{ on } \Omega \text{ such that } \int d^2(x, \partial\Omega)d\mu(x) < \infty\right\},$$

and for any $\mu,\ \nu \in \mathscr{M}_2(\Omega)$ define the set of admissible transfer plans $\mathrm{Adm}_b(\mu, \nu)$ by: $\gamma \in \mathrm{Adm}_b(\mu, \nu)$ if and only if γ is a measure on $(\overline{\Omega})^2$ such that

$$\pi^1_\# \gamma|_\Omega = \mu, \qquad \pi^2_\# \gamma|_\Omega = \nu.$$

Notice the difference w.r.t. the classical definition of transfer plan: here we are requiring the first (respectively, second) marginal to coincide with μ (respectively ν) only inside the open set Ω. This means that in transferring the mass from μ to ν we are free to take/put as much mass as we want from/to the boundary. Then one defines the *cost* $C(\gamma)$ of a plan γ by

$$C(\gamma) := \int |x - y|^2 d\gamma(x, y),$$

and then the distance Wb_2 by

$$Wb_2(\mu, \nu) := \inf \sqrt{C(\gamma)},$$

where the infimum is taken among all $\gamma \in \mathrm{Adm}_b(\mu, \nu)$.

The distance Wb_2 shares many properties with the Wasserstein distance W_2.

Theorem 6.5 (Main properties of Wb_2). *The following hold:*

- *Wb_2 is a distance on $\mathscr{M}_2(\Omega)$ and the metric space $(\mathscr{M}_2(\Omega), Wb_2)$ is Polish and geodesic.*
- *A sequence $(\mu_n) \subset \mathscr{M}_2(\Omega)$ converges to μ w.r.t. Wb_2 if and only if μ_n converges weakly to μ in duality with continuous functions with compact support in Ω and $\int d^2(x, \partial\Omega)d\mu_n \to \int d^2(x, \partial\Omega)d\mu$ as $n \to \infty$.*
- *Finally, a plan $\gamma \in Adm_b(\mu, \nu)$ is optimal (i.e. it attains the minimum cost among admissible plans) if and only there exists a c-concave function φ* which is identically 0 on $\partial\Omega$ *such that* $\mathrm{supp}(\gamma) \subset \partial^c\varphi$ *(here $c(x, y) = |x - y|^2$).*

Observe that $(\mathscr{M}_2(\Omega), Wb_2)$ is always a geodesic space (while from Theorem 3.10 and Remark 3.14 we know that $(\mathscr{P}(\Omega), W_2)$ is geodesic if and only if Ω is, that is, if and only if Ω is convex).

It makes perfectly sense to extend the entropy functional to the whole $\mathscr{M}_2(\Omega)$: the formula is still $E(\mu) = \int \rho \log(\rho)$ for $\mu = \rho \mathscr{L}^d|_{\Omega}$, and $E(\mu) = +\infty$ for measures not absolutely continuous. The Gradient Flow of the entropy w.r.t. Wb_2 produces solutions of the Heat equation with Dirichlet boundary conditions in the following sense:

Theorem 6.6. *Let $\mu \in \mathscr{M}_2(\Omega)$ be such that $E(\mu) < \infty$. Then:*

- *For every $\tau > 0$ there exists a unique discrete solution ρ_t^τ starting from μ and constructed via the Minimizing Movements scheme as in Definition 4.7.*
- *As $\tau \downarrow 0$, the measures ρ_t^τ converge to a unique measure ρ_t in $(\mathscr{M}_2(\Omega), Wb_2)$ for any $t > 0$.*
- *The map $(x,t) \mapsto \rho_t(x)$ is a solution of the Heat equation*

$$\begin{cases} \frac{d}{dt}\rho_t = \Delta \rho_t, & \text{in } \Omega \times (0,+\infty), \\ \quad \rho_t \to \mu, & \text{weakly as } t \to 0, \end{cases}$$

subject to the Dirichlet boundary condition $\rho_t(x) = e^{-1}$ in $\partial\Omega$ for every $t > 0$ (that is, $\rho_t - e^{-1}$ belongs to $H_0^1(\Omega)$ for every $t > 0$).

The fact that the boundary value is given by e^{-1} can be heuristically guessed by the fact that the entropy has a global minimum in $\mathscr{M}_2(\Omega)$: such minimum is given by the measure with constant density e^{-1}, i.e. the measure whose density is everywhere equal to the minimum of $z \mapsto z\log(z)$.

On the bad side, the entropy E is *not* geodesically convex in $(\mathscr{M}_2(\Omega), Wb_2)$, and this implies that it is not clear whether the strong properties of Gradient Flows w.r.t. W_2 as described in Sect. 4.3—Theorem 4.35 and Proposition 4.38 are satisfied also in this setting. In particular, it is not clear whether there is contractivity of the distance or not:

Open Problem 6.7. *Let ρ_t^1, ρ_t^2 two solutions of the Heat equation with Dirichlet boundary condition $\rho_t^i = e^{-1}$ in $\partial\Omega$ for every $t > 0$, $i = 1, 2$. Prove or disprove that*

$$Wb_2(\rho_s^1, \rho_s^2) \leq Wb_2(\rho_t^1, \rho_t^2), \qquad \forall t > s.$$

The question is open also for convex and smooth open sets Ω.

6.4 Bibliographical Notes

The connection of branched transport and transport problem as discussed in Sect. 6.1 was first pointed out by Q. Xia in [81]. An equivalent model was proposed by F. Maddalena et al. in [61]. In [60, 81] and [15] the existence of an optimal branched transport (Theorem 6.2) was also provided. Later, this result has been extended in several directions, see for instance the works A. Brancolini et al. [16] and Bianchini–Brancolini [15]. The interior regularity result (Theorem 6.3) has been proved By

Q. Xia in [82] and M. Bernot et al. in [14]. Also, we remark that L. Brasco, G. Buttazzo and F. Santambrogio proved a kind of Benamou–Brenier formula for branched transport in [17].

The content of Sect. 6.2 comes from J. Dolbeault, B. Nazaret and G. Savaré [33] and [26] of J. Carrillo, S. Lisini, G. Savaré and D. Slepcev.

Section 6.3 is taken from a work of the second author and A. Figalli [37].

7 More on the Structure of $(\mathscr{P}_2(M), W_2)$

The aim of this Chapter is to give a comprehensive description of the structure of the "Riemannian manifold" $(\mathscr{P}_2(\mathbb{R}^d), W_2)$, thus the content of this part of the work is the natural continuation of what we discussed in Sect. 3.3.2. For the sake of simplicity, we are going to stick to the Wasserstein space on $\mathbb{R}^d$, but the reader should keep in mind that the discussions here can be generalized with only little effort to the Wasserstein space built over a Riemannian manifold.

7.1 "Duality" Between the Wasserstein and the Arnold Manifolds

The content of this section is purely formal and directly comes from the seminal paper of Otto [67]. We won't even try to provide a rigorous background for the discussion we will do here, as we believe that dealing with the technical problems would lead the reader far from the geometric intuition. Also, we will not use the "results" presented here later on: we just think that these concepts are worth of mention. Thus for the purpose of this section just think that "each measure is absolutely continuous with smooth density", that "each L^2 function is C^∞", and so on.

Let us recall the definition of Riemannian submersion. Let M, N be Riemannian manifolds and let $f : M \to N$ a smooth map. f is a submersion provided the map:

$$df : \mathrm{Ker}^{\perp}\big(df(x)\big) \to T_{f(x)}N,$$

is a surjective isometry for any $x \in M$. A trivial example of submersion is given in the case $M = N \times L$ (for some Riemannian manifold L, with M endowed with the product metric) and $f : M \to N$ is the natural projection. More generally, if f is a Riemannian submersion, for each $y \in N$, the set $f^{-1}(y) \subset M$ is a smooth Riemannian submanifold.

The "duality" between the Wasserstein and the Arnold Manifolds consists in the fact that there exists a Big Manifold BM which is flat and a natural Riemannian submersion from BM to $\mathscr{P}_2(\mathbb{R}^d)$ whose fibers are precisely the Arnold Manifolds.

Let us define the objects we are dealing with. Fix once and for all a reference measure $\overline{\rho} \in \mathscr{P}_2(\mathbb{R}^d)$ (recall that we are "assuming" that all the measures are absolutely continuous with smooth densities—so that we will use the same notation for both the measure and its density).

- The Big Manifold BM is the space $L^2(\overline{\rho})$ of maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ which are L^2 w.r.t. the reference measure $\overline{\rho}$. The tangent space at some map $T \in$ BM is naturally given by the set of vector fields belonging to $L^2(\overline{\rho})$, where the perturbation of T in the direction of the vector field u is given by $t \mapsto T + tu$.
- The target manifold of the submersion is the Wasserstein "manifold" $\mathscr{P}_2(\mathbb{R}^d)$. We recall that the tangent space $\mathrm{Tan}_\rho(\mathscr{P}_2(\rho)\mathbb{R}^d)$ at the measure ρ is the set
$$\mathrm{Tan}_\rho(\mathscr{P}_2(\rho)\mathbb{R}^d) := \Big\{\nabla\varphi \ : \ \varphi \in C_c^\infty(\mathbb{R}^d)\Big\},$$
endowed with the scalar product of $L^2(\rho)$ (we neglect to take the closure in $L^2(\rho)$ because we want to keep the discussion at a formal level). The perturbation of a measure ρ in the direction of a tangent vector $\nabla\varphi$ is given by $t \mapsto (Id + t\nabla\varphi)_\#\rho$.
- The Arnold Manifold $\mathrm{Arn}(\rho)$ associated to a certain measure $\rho \in \mathscr{P}_2(\mathbb{R}^d)$ is the set of maps $S : \mathbb{R}^d \to \mathbb{R}^d$ which preserve ρ:
$$\mathrm{Arn}(\rho) := \Big\{S : \mathbb{R}^d \to \mathbb{R}^d \ : \ S_\#\rho = \rho\}.$$
We endow $\mathrm{Arn}(\rho)$ with the L^2 distance calculated w.r.t. ρ. To understand who is the tangent space at $\mathrm{Arn}(\rho)$ at a certain map S, pick a vector field v on $\mathbb{R}^d$ and consider the perturbation $t \mapsto S + tv$ of S in the direction of v. Then v is a tangent vector if and only if $\frac{d}{dt}|_{t=0}(S + tv)_\#\rho = 0$. Observing that
$$\begin{aligned}\frac{d}{dt}|_{t=0}(S + tv)_\#\rho &= \frac{d}{dt}|_{t=0}(Id + tv \circ S^{-1})_\#(S_\#\rho)\\ &= \frac{d}{dt}|_{t=0}(Id + tv \circ S^{-1})_\#\rho = \nabla \cdot (v \circ S^{-1}\rho),\end{aligned}$$
we deduce
$$\mathrm{Tan}_S\mathrm{Arn}(\rho) = \Big\{\text{vector fields } v \text{ on } \mathbb{R}^d \text{ such that } \nabla \cdot (v \circ S^{-1}\rho) = 0\Big\},$$
which is naturally endowed with the scalar product in $L^2(\rho)$.

  We are calling the manifold $\mathrm{Arn}(\rho)$ an Arnold Manifold, because if ρ is the Lebesgue measure restricted to some open, smooth and bounded set Ω, this definition reduces to the well known definition of Arnold manifold in fluid mechanics: the geodesic equation in such space is—formally—the Euler equation for the motion of an incompressible and inviscid fluid in Ω.
- Finally, the "Riemannian submersion" Pf from BM to $\mathscr{P}_2(\mathbb{R}^d)$ is the push forward map:
$$\begin{aligned}\mathrm{Pf} : \mathrm{BM} &\to \mathscr{P}_2(\mathbb{R}^d),\\ T &\mapsto T_\#\overline{\rho},\end{aligned}$$

We claim that Pf is a Riemannian submersion and that the fiber $\mathrm{Pf}^{-1}(\rho)$ is isometric to the manifold $\mathrm{Arn}(\rho)$.

We start considering the fibers. Fix $\rho \in \mathscr{P}_2(\mathbb{R}^d)$. Observe that

$$\mathrm{Pf}^{-1}(\rho) = \Big\{T \in \mathrm{BM} \ : \ T_\# \overline{\rho} = \rho\Big\},$$

and that the tangent space $\mathrm{Tan}_T \mathrm{Pf}^{-1}(\rho)$ is the set of vector fields u such that $\frac{d}{dt}|_{t=0}(T + tu)_\# \overline{\rho} = 0$, so that from

$$\begin{aligned}\frac{d}{dt}|_{t=0}(T + tu)_\# \overline{\rho} &= \frac{d}{dt}|_{t=0}(Id + tu \circ T^{-1})_\#(T_\# \overline{\rho}) \\ &= \frac{d}{dt}|_{t=0}(Id + tu \circ T^{-1})_\# \rho = \nabla \cdot (u \circ T^{-1} \rho),\end{aligned}$$

we have

$$\mathrm{Tan}_T \mathrm{Pf}^{-1}(\rho) = \Big\{\text{vector fields } u \text{ on } \mathbb{R}^d \text{ such that } \nabla \cdot (u \circ T^{-1} \rho) = 0\Big\},$$

and the scalar product between two vector fields in $\mathrm{Tan}_T \mathrm{Pf}^{-1}(\rho)$ is the one inherited by the one in BM, i.e. is the scalar product in $L^2(\overline{\rho})$.

Now choose a distinguished map $T^\rho \in \mathrm{Pf}^{-1}(\rho)$ and notice that the right composition with T^ρ provides a natural bijective map from $\mathrm{Arn}(\rho)$ into $\mathrm{Pf}^{-1}(\rho)$, because

$$S_\# \rho = \rho \qquad \Leftrightarrow \qquad (S \circ T^\rho)_\# \overline{\rho} = \rho.$$

We claim that this right composition also provides an isometry between the "Riemannian manifolds" $\mathrm{Arn}(\rho)$ and $\mathrm{Pf}^{-1}(\rho)$: indeed, if $v \in \mathrm{Tan}_S \mathrm{Arn}(\rho)$, then the perturbed maps $S + tv$ are sent to $S \circ T^\rho + tv \circ T^\rho$, which means that the perturbation v of S is sent to the perturbation $u := v \circ T^\rho$ of $S \circ T^\rho$ by the differential of the right composition. The conclusion follows from the change of variable formula, which gives

$$\int |v|^2 d\rho = \int |u|^2 d\overline{\rho}.$$

Clearly, the kernel of the differential $d\mathrm{Pf}$ of Pf at T is given by $\mathrm{Tan}_T \mathrm{Pf}^{-1}\big(\mathrm{Pf}(T)\big)$, thus it remains to prove that its orthogonal is sent isometrically onto $\mathrm{Tan}_{\mathrm{Pf}(T)}(\mathscr{P}_2(\mathbb{R}^d))$ by $d\mathrm{Pf}$. Fix $T \in \mathrm{BM}$, let $\rho := \mathrm{Pf}(T) = T_\# \overline{\rho}$ and observe that

$$\begin{aligned}&\mathrm{Tan}_T^\perp\big(\mathrm{Pf}^{-1}(\rho)\big) \\ &= \Big\{\text{vector fields } w \ : \ \int \langle w, u\rangle \, d\overline{\rho} = 0, \ \forall u \text{ s.t. } \nabla \cdot (u \circ T^{-1} \rho) = 0\Big\} \\ &= \Big\{\text{vector fields } w \ : \ \int \Big\langle w \circ T^{-1}, u \circ T^{-1}\Big\rangle d\rho = 0, \ \forall u \text{ s.t. } \nabla \cdot (u \circ T^{-1} \rho) = 0\Big\} \\ &= \Big\{\text{vector fields } w \ : \ w \circ T^{-1} = \nabla \varphi \text{ for some } \varphi \in C_c^\infty(\mathbb{R}^d)\Big\}.\end{aligned}$$

Now pick $w \in \mathrm{Tan}_T^{\perp}\big(\mathrm{Pf}^{-1}(\rho)\big)$, let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be such that $w \circ T^{-1} = \nabla\varphi$ and observe that

$$\begin{aligned}\frac{d}{dt}|_{t=0}\mathrm{Pf}(T + tw) &= \frac{d}{dt}|_{t=0}(T + tw)_\#\overline{\rho}\\ &= \frac{d}{dt}|_{t=0}(Id + tw \circ T^{-1})_\#(T_\#\overline{\rho}) = \frac{d}{dt}|_{t=0}(Id + t\nabla\varphi)_\#\rho,\end{aligned}$$

which means, by definition of $\mathrm{Tan}_\rho(\mathscr{P}_2(\mathbb{R}^d))$ and the action of tangent vectors, that the differential $d\mathrm{Pf}(T)(w)$ of Pf calculated at T along the direction w is given by $\nabla\varphi$. The fact that this map is an isometry follows once again by the change of variable formula

$$\int |w|^2 d\overline{\rho} = \int |w \circ T^{-1}|^2 d\rho = \int |\nabla\varphi|^2 d\rho.$$

7.2 On the Notion of Tangent Space

Aim of this section is to quickly discuss the definition of tangent space of $\mathscr{P}_2(\mathbb{R}^d)$ at a certain measure μ from a purely geometric perspective. We will see how this perspective is related to the discussion made in Sect. 3.3.2, where we defined tangent space as

$$\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d) := \overline{\Big\{\nabla\varphi \ : \ \varphi \in C_c^\infty(\mathbb{R}^d)\Big\}}^{L^2(\mathbb{R}^d,\mathbb{R}^d;\mu)}.$$

Recall that this definition came from the characterization of absolutely continuous curves on $\mathscr{P}_2(\mathbb{R}^d)$ (Theorem 3.29 and the subsequent discussion).

Yet, there is a completely different and purely geometrical approach which leads to a definition of tangent space at μ. The idea is to think the tangent space at μ as the "space of directions", or, which is the same, as the set of constant speed geodesics emanating from μ. More precisely, let the set $\mathcal{Geod}_\mu$ be defined by:

$$\mathcal{Geod}_\mu := \left\{\begin{matrix}\text{constant speed geodesics starting from } \mu\\ \text{and defined on some interval of the kind } [0,T]\end{matrix}\right\} \Big/ \approx,$$

where we say that $(\mu_t) \approx (\mu'_t)$ provided they coincide on some right neighborhood of 0. The natural distance D on $\mathcal{Geod}_\mu$ is:

$$D\big((\mu_t),(\mu'_t)\big) := \overline{\lim_{t\downarrow 0}}\frac{W_2(\mu_t,\mu'_t)}{t}. \qquad (91)$$

The *Geometric Tangent space* $\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ is then defined as the completion of $\mathcal{Geod}_\mu$ w.r.t. the distance D.

The natural question here is: what is the relation between the "space of gradients" $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ and the "space of directions" $\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$?

Recall that from Remark 2.22 we know that given $\varphi \in C_c^\infty(\mathbb{R}^d)$, the map $t \mapsto (Id + t\nabla\varphi)_\# \mu$ is a constant speed geodesic on a right neighborhood of 0. This means that there is a natural map ι_μ from the set $\{\nabla\varphi : \varphi \in C_c^\infty\}$ into $\mathscr{G}eod_\mu$, and therefore into $\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$, which sends $\nabla\varphi$ into the (equivalence class of the) geodesic $t \mapsto (Id + t\nabla\varphi)_\#\mu$. The main properties of the Geometric Tangent space and of this map are collected in the following theorem, which we state without proof.

Theorem 7.1 (The tangent space). *Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. Then:*

- *The $\overline{\lim}$ in (91) is always a limit.*
- *The metric space $(\boldsymbol{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d), D)$ is complete and separable.*
- *The map $\iota_\mu : \{\nabla\varphi\} \to \boldsymbol{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ is an injective isometry, where on the source space we put the L^2 distance w.r.t. μ. Thus, ι_μ always extends to a natural isometric embedding of* $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ *into $\boldsymbol{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$.*

Furthermore, the following statements are equivalent:

(i) The space $(\boldsymbol{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d), D)$ is an Hilbert space.
(ii) The map ι_μ : $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d) \to \boldsymbol{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ *is surjective.*
(iii) The measure μ is regular (Definition 2.25).

We comment on the second part of the theorem. The first thing to notice is that the "space of directions" $\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ can be strictly larger than "the space of gradients" $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$. This is actually not surprising if one thinks to the case in which μ is a Dirac mass. Indeed in this situation the space $(\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d), D)$ coincides with the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ (this can be checked directly from the definition), however, the space $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ is actually isometric to $\mathbb{R}^d$ itself, and is therefore much smaller.

The reason is that geodesics are not always induced by maps, that is, they are not always of the form $t \mapsto (Id + tu)_\#\mu$ for some vector field $u \in L^2_\mu$. To some extent, here we are facing the same problem we had to face when starting the study of the optimal transport problem: maps are typically not sufficient to produce (optimal) transports. From this perspective, it is not surprising that if the measure we are considering is regular (that is, if for any $\nu \in \mathscr{P}_2(\mathbb{R}^d)$ there exists a unique optimal plan, and this plan is induced by a map), then the "space of directions" coincides with the "space of directions induced by maps".

7.3 Second Order Calculus

Now we pass to the description of the second order analysis over $\mathscr{P}_2(\mathbb{R}^d)$. The concepts that now enter into play are: Covariant Derivative, Parallel Transport and Curvature. To some extent, the situation is similar to the one we discussed in Sect. 3.3.2 concerning the first order structure: the metric space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$

is not a Riemannian manifold, but if we are careful in giving definitions and in the regularity requirements of the objects involved we will be able to perform calculations very similar to those valid in a genuine Riemannian context.

Again, we are restricting the analysis to the Euclidean case only for simplicity: all of what comes next can be generalized to the analysis over $\mathscr{P}_2(M)$, for a generic Riemannian manifold M.

On a typical course of basic Riemannian geometry, one of the first concepts introduced is that of Levi–Civita connection, which identifies the only natural ("natural" here means: "compatible with the Riemannian structure") way of differentiating vector fields on the manifold. It would therefore be natural to set up our discussion on the second order analysis on $\mathscr{P}_2(\mathbb{R}^d)$ by giving the definition of Levi–Civita connection in this setting. However, this cannot be done. The reason is that we don't have a notion of smoothness for vector fields, therefore not only we don't know how to covariantly differentiate vector fields, but we don't know either which are the vector fields regular enough to be differentiated. In a purely Riemannian setting this problem does not appear, as a Riemannian manifold borns as smooth manifold on which we define a scalar product on each tangent space; but the space $\mathscr{P}_2(\mathbb{R}^d)$ does not have a smooth structure (there is no diffeomorphism of a small ball around the origin in $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ onto a neighborhood of μ in $\mathscr{P}_2(\mathbb{R}^d)$). Thus, we have to proceed in a different way, which we describe now:

Regular curves. First of all, we drop the idea of defining a smooth vector field on the whole "manifold". We will rather concentrate on finding an appropriate definition of smoothness for vector fields defined along curves. We will see that to do this, we will need to work with a particular kind of curves, which we call *regular*, see Definition 7.2.

Smoothness of vector fields. We will then be able to define the smoothness of vector fields defined along regular curves (Definition 7.5). Among others, a notion of smoothness of particular relevance is that of *absolutely continuous* vector fields: for this kind of vector fields we have a natural notion of *total derivative* (not to be confused with the covariant one, see Definition 7.6).

Levi–Civita connection. At this point we have all the ingredients we need to define the covariant derivative and to prove that it is the Levi–Civita connection on $\mathscr{P}_2(\mathbb{R}^d)$ (Definition 7.8 and discussion thereafter).

Parallel transport. This is the main existence result on this subject: we prove that along regular curves the parallel transport always exists (Theorem 7.15). We will also discuss a counterexample to the existence of parallel transport along a non-regular geodesic (Example 7.16). This will show that the definition of regular curve is not just operationally needed to provide a definition of smoothness of vector fields, but is actually intrinsically related to the geometry of $\mathscr{P}_2(\mathbb{R}^d)$.

Calculus of derivatives. Using the technical tools developed for the study of the parallel transport, we will be able to explicitly compute the total and covariant derivatives of basic examples of vector fields.

Curvature. We conclude the discussion by showing how the concepts developed can lead to a rigorous definition of the curvature tensor on $\mathscr{P}_2(\mathbb{R}^d)$.

We will write $\|v\|_\mu$ and $\langle v, w\rangle_\mu$ for the norm of the vector field v and the scalar product of the vector fields v, w in the space $L^2(\mu)$ (which we will denote by L^2_μ), respectively.

We now start with the definition of regular curve. All the curves we will consider are defined on $[0, 1]$, unless otherwise stated.

Definition 7.2 (Regular curve). Let (μ_t) be an absolutely continuous curve and let (v_t) be its velocity vector field, that is (v_t) is the unique vector field—up to equality for a.e. t—such that $v_t \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for a.e. t and the continuity equation

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0,$$

holds in the sense of distributions (recall Theorem 3.29 and Definition 3.31). We say that (μ_t) is regular provided

$$\int_0^1 \|v_t\|^2_{\mu_t}\, dt < \infty, \tag{92}$$

and

$$\int_0^1 \mathrm{Lip}(v_t) dt < \infty. \tag{93}$$

Observe that the validity of (93) is independent on the parametrization of the curve, thus if it is fulfilled it is always possible to reparametrize the curve (e.g. with constant speed) in order to let it satisfy also (92).

Now assume that (μ_t) is regular. Then by the classical Cauchy–Lipschitz theory we know that there exists a unique family of maps $\mathbf{T}(t, s, \cdot) : \mathrm{supp}(\mu_t) \to \mathrm{supp}(\mu_s)$ satisfying

$$\begin{cases} \dfrac{d}{ds}\mathbf{T}(t, s, x) = v_s(\mathbf{T}(t, s, x)), & \forall t \in [0, 1],\ x \in \mathrm{supp}(\mu_t),\ a.e.\ s \in [0, 1], \\ \mathbf{T}(t, t, x) = x, & \forall t \in [0, 1],\ x \in \mathrm{supp}(\mu_t). \end{cases} \tag{94}$$

Also it is possible to check that these maps satisfy the additional properties

$$\begin{aligned} \mathbf{T}(r, s, \cdot) \circ \mathbf{T}(t, r, \cdot) &= \mathbf{T}(t, s, \cdot) \ \forall t, r, s \in [0, 1], \\ \mathbf{T}(t, s, \cdot)_\# \mu_t &= \mu_s, \qquad \forall t, s \in [0, 1]. \end{aligned}$$

We will call this family of maps the *flow maps* of the curve (μ_t). Observe that for any couple of times $t, s \in [0, 1]$, the right composition with $\mathbf{T}(t, s, \cdot)$ provides a bijective isometry from $L^2_{\mu_s}$ to $L^2_{\mu_t}$. Also, notice that from condition (92) and the inequalities

$$\|\mathbf{T}(t,s,\cdot)-\mathbf{T}(t,s',\cdot)\|^2_{\mu_t} \leq \int \left(\int_s^{s'} v_r(\mathbf{T}(t,r,x))dr\right)^2 d\mu_t(x)$$
$$\leq |s'-s| \int_s^{s'} \|v_r(x)\|^2_{\mu_r(x)} dr$$

we get that for fixed $t \in [0,1]$, the map $s \mapsto \mathbf{T}(t,s,\cdot) \in L^2_{\mu_t}$ is absolutely continuous.

It can be proved that the set of regular curves is dense in the set of absolutely continuous curves on $\mathscr{P}_2(\mathbb{R}^d)$ with respect to uniform convergence plus convergence of length. We omit the technical proof of this fact and focus instead on the important case of geodesics:

Proposition 7.3 (Regular geodesics). *Let* (μ_t) *be a constant speed geodesic on* $[0,1]$. *Then its restriction to any interval* $[\varepsilon, 1-\varepsilon]$, *with* $\varepsilon > 0$, *is regular. In general, however, the whole curve* (μ_t) *may be not regular on* $[0,1]$.

Proof. To prove that (μ_t) may be not regular just consider the case of $\mu_0 := \delta_x$ and $\mu_1 := \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$: it is immediate to verify that for the velocity vector field (v_t) it holds $\mathrm{Lip}(v_t) = t^{-1}$.

For the other part, recall from Remark 3.25 (see also Proposition 3.16) that for $t \in (0,1)$ and $s \in [0,1]$ there exists a unique optimal map T_t^s from μ_t to μ_s. It is immediate to verify from formula (19) that these maps satisfy

$$\frac{T_t^s - Id}{s-t} = \frac{T_t^{s'} - Id}{s'-t}, \qquad \forall t \in (0,1),\ s \in [0,1].$$

Thus, thanks to Proposition 3.32, we have that v_t is given by

$$v_t = \lim_{s \to t} \frac{T_t^s - Id}{s-t} = \frac{Id - T_t^0}{t}. \tag{95}$$

Now recall that Remark 3.25 gives $\mathrm{Lip}(T_0^t) \leq (1-t)^{-1}$ to obtain

$$\mathrm{Lip}(v_t) \leq t^{-1}((1-t)^{-1}+1) = \frac{2-t}{t(1-t)}.$$

Thus $t \mapsto \mathrm{Lip}(v_t)$ is integrable on any interval of the kind $[\varepsilon, 1-\varepsilon]$, $\varepsilon > 0$. □

Definition 7.4 (Vector fields along a curve). A vector field along a curve (μ_t) is a Borel map $(t,x) \mapsto u_t(x)$ such that $u_t \in L^2_{\mu_t}$ for a.e. t. It will be denoted by (u_t).

Observe that we are considering also non tangent vector fields, that is, we are not requiring $u_t \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for a.e. t.

To define the (time) smoothness of a vector field (u_t) defined along a regular curve (μ_t) we will make an essential use of the flow maps: notice that the main problem in considering the smoothness of (u_t) is that for different times, the vectors belong to different spaces. To overcome this obstruction we will define the smoothness of $t \mapsto u_t \in L^2_{\mu_t}$ in terms of the smoothness of $t \mapsto u_t \circ \mathbf{T}(t_0, t, \cdot) \in L^2_{\mu_{t_0}}$:

Definition 7.5 (Smoothness of vector fields). Let (μ_t) be a regular curve, $\mathbf{T}(t, s, \cdot)$ its flow maps and (u_t) a vector field defined along it. We say that (u_t) is absolutely continuous (or C^1, or C^n, ..., or C^∞ or analytic) provided the map

$$t \mapsto u_t \circ \mathbf{T}(t_0, t, \cdot) \in L^2_{\mu_{t_0}}$$

is absolutely continuous (or C^1, or C^n, ..., or C^∞ or analytic) for every $t_0 \in [0, 1]$.

Since $u_t \circ \mathbf{T}(t_1, t, \cdot) = u_t \circ \mathbf{T}(t_0, t, \cdot) \circ \mathbf{T}(t_1, t_0, \cdot)$ and the composition with $\mathbf{T}(t_1, t_0, \cdot)$ provides an isometry from $L^2_{\mu_{t_0}}$ to $L^2_{\mu_{t_1}}$, it is sufficient to check the regularity of $t \mapsto u_t \circ \mathbf{T}(t_0, t, \cdot)$ for *some* $t_0 \in [0, 1]$ to be sure that the same regularity holds for every t_0.

Definition 7.6 (Total derivative). With the same notation as above, assume that (u_t) is an absolutely continuous vector field. Its total derivative is defined as:

$$\frac{\boldsymbol{d}}{dt} u_t := \lim_{h \to 0} \frac{u_{t+h} \circ \mathbf{T}(t, t+h, \cdot) - u_t}{h},$$

where the limit is intended in $L^2_{\mu_t}$.

Observe that we are not requiring the vector field to be tangent, and that the total derivative is in general a non tangent vector field, even if (u_t) is.

The identity

$$\begin{aligned}\lim_{h \to 0} \frac{u_{t+h} \circ \mathbf{T}(t, t+h, \cdot) - u_t}{h} &= \left(\lim_{h \to 0} \frac{u_{t+h} \circ \mathbf{T}(t_0, t+h, \cdot) - u_t \circ \mathbf{T}(t_0, t, \cdot)}{h} \right) \circ \mathbf{T}(t, t_0, \cdot) \\ &= \left(\frac{d}{dt} \big(u_t \circ \mathbf{T}(t_0, t, \cdot) \big) \right) \circ \mathbf{T}(t, t_0, \cdot),\end{aligned}$$

shows that the total derivative is well defined for a.e. t and that is an L^1 vector field, in the sense that it holds

$$\int_0^1 \left\| \frac{\boldsymbol{d}}{dt} u_t \right\|_{\mu_t} dt < \infty.$$

Notice also the inequality

$$\|u_s \circ \mathbf{T}(t, s, \cdot) - u_t\|_{\mu_t} \leq \int_t^s \left\| \frac{d}{dt} (u_r \circ \mathbf{T}(t, r, \cdot)) \right\|_{\mu_t} dr = \int_t^s \left\| \frac{\boldsymbol{d}}{dt} u_r \right\|_{\mu_r} dr.$$

An important property of the total derivative is the *Leibnitz rule*: for any couple of absolutely continuous vector fields (u_t^1), (u_t^2) along the same regular curve (μ_t) the map $t \mapsto \langle u_t^1, u_t^2 \rangle_{\mu_t}$ is absolutely continuous and it holds

$$\frac{d}{dt}\langle u_t^1, u_t^2\rangle_{\mu_t} = \left\langle \frac{\boldsymbol{d}}{dt}u_t^1, u_t^2\right\rangle_{\mu_t} + \left\langle u_t^1, \frac{\boldsymbol{d}}{dt}u_t^2\right\rangle_{\mu_t}, \qquad a.e.\ t. \tag{96}$$

Indeed, from the identity

$$\langle u_t^1, u_t^2\rangle_{\mu_t} = \langle u_t^1 \circ \mathbf{T}(t_0, t, \cdot), u_t^2 \circ \mathbf{T}(t_0, t, \cdot)\rangle_{\mu_{t_0}},$$

it follows the absolute continuity, and the same expression gives

$$\begin{aligned}
\frac{d}{dt}\langle u_t^1, u_t^2\rangle_{\mu_t} &= \frac{d}{dt}\langle u_t^1 \circ \mathbf{T}(t_0, t, \cdot), u_t^2 \circ \mathbf{T}(t_0, t, \cdot)\rangle_{\mu_{t_0}} \\
&= \left\langle \frac{d}{dt}\big(u_t^1 \circ \mathbf{T}(t_0, t, \cdot)\big), u_t^2 \circ \mathbf{T}(t_0, t, \cdot)\right\rangle_{\mu_{t_0}} \\
&\quad + \left\langle u_t^1 \circ \mathbf{T}(t_0, t, \cdot), \frac{d}{dt}\big(u_t^2 \circ \mathbf{T}(t_0, t, \cdot)\big)\right\rangle_{\mu_{t_0}} \\
&= \left\langle \frac{\boldsymbol{d}}{dt}u_t^1, u_t^2\right\rangle_{\mu_t} + \left\langle u_t^1, \frac{\boldsymbol{d}}{dt}u_t^2\right\rangle_{\mu_t}.
\end{aligned}$$

Example 7.7 (The smooth case). Let $(x, t) \mapsto \xi_t(x)$ be a C_c^∞ vector field on $\mathbb{R}^d$, (μ_t) a regular curve and (v_t) its velocity vector field. Then the inequality

$$\begin{aligned}
\|\xi_s \circ \mathbf{T}(t, s, \cdot) - \xi_t\|_{\mu_t} &\le \|\xi_s - \xi_t\|_{\mu_s} + \|\xi_t \circ \mathbf{T}(t, s, \cdot) - \xi_t\|_{\mu_t} \\
&\le C|s - t| + C'\|\mathbf{T}(t, s, \cdot) - Id\|_{\mu_t},
\end{aligned}$$

with $C := \sup_{t,x} |\partial_t \xi_t(x)|$, $C' := \sup_{t,x} |\xi_t(x)|$, together with the fact that $s \mapsto \mathbf{T}(t, s, \cdot) \in L^2(\mu_t)$ is absolutely continuous, gives that (ξ_t) is absolutely continuous along (μ_t).

Then a direct application of the definition gives that its total derivative is given by

$$\frac{\boldsymbol{d}}{dt}\xi_t = \partial_t \xi_t + \nabla \xi_t \cdot v_t, \qquad a.e.\ t, \tag{97}$$

which shows that the total derivative is nothing but the *convective derivative* well known in fluid dynamics. ■

For $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, we denote by $\mathrm{P}_\mu : L^2_\mu \to \mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ the orthogonal projection, and we put $\mathrm{P}^\perp_\mu := Id - \mathrm{P}_\mu$.

Definition 7.8 (Covariant derivative). Let (u_t) be an absolutely continuous and *tangent* vector field along the regular curve (μ_t). Its covariant derivative is defined as

$$\frac{\boldsymbol{D}}{dt}u_t := \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t\right). \tag{98}$$

The trivial inequality

$$\left\| \frac{\boldsymbol{D}}{dt} u_t \right\|_{\mu_t} \leq \left\| \frac{\boldsymbol{d}}{dt} u_t \right\|_{\mu_t}$$

shows that the covariant derivative is an L^1 vector field.

In order to prove that the covariant derivative we just defined is the Levi–Civita connection, we need to prove two facts: *compatibility with the metric* and *torsion free identity*. Recall that on a standard Riemannian manifold, these two conditions are respectively given by:

$$\frac{d}{dt}\langle X(\gamma_t), Y(\gamma_t)\rangle = \left\langle (\nabla_{\gamma_t'} X)(\gamma_t), Y(\gamma_t)\right\rangle + \left\langle X(\gamma_t), (\nabla_{\gamma_t'} Y)(\gamma_t)\right\rangle$$
$$[X, Y] = \nabla_X Y - \nabla_Y X,$$

where X, Y are smooth vector fields and γ is a smooth curve on M.

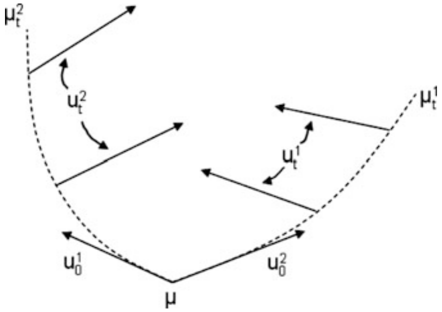
The compatibility with the metric follows immediately from the Leibnitz rule (96), indeed if (u_t^1), (u_t^2) are tangent absolutely continuous vector fields we have:

$$\begin{aligned} \frac{d}{dt}\langle u_t^1, u_t^2\rangle_{\mu_t} &= \left\langle \frac{\boldsymbol{d}}{dt} u_t^1, u_t^2 \right\rangle_{\mu_t} + \left\langle u_t^1, \frac{\boldsymbol{d}}{dt} u_t^2 \right\rangle_{\mu_t} \\ &= \left\langle \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt} u_t^1\right), u_t^2 \right\rangle_{\mu_t} + \left\langle u_t^1, \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt} u_t^2\right)\right\rangle_{\mu_t} \\ &= \left\langle \frac{\boldsymbol{D}}{dt} u_t^1, u_t^2 \right\rangle_{\mu_t} + \left\langle u_t^1, \frac{\boldsymbol{D}}{dt} u_t^2 \right\rangle_{\mu_t}. \end{aligned} \tag{99}$$

To prove the torsion-free identity, we need first to understand how to calculate the Lie bracket of two vector fields. To this aim, let μ_t^i, $i = 1, 2$, be two regular curves such that $\mu_0^1 = \mu_0^2 =: \mu$ and let $u_t^i \in \mathrm{Tan}_{\mu_t^i}(\mathscr{P}_2(\mathbb{R}^d))$ be two C^1 vector fields satisfying $u_0^1 = v_0^2$, $u_0^2 = v_0^1$, where v_t^i are the velocity vector fields of μ_t^i. We assume that the velocity fields v_t^i of μ_t^i are continuous in time (in the sense that the map $t \mapsto v_t^i \mu_t^i$ is continuous in the set of vector valued measure with the weak topology and $t \mapsto \|v_t^i\|_{\mu_t^i}$ is continuous as well), to be sure that (97) holds for *all* t with $v_t = v_t^i$ and the initial condition makes sense. With these hypotheses, it makes sense to consider the covariant derivative $\frac{\boldsymbol{D}}{dt} u_t^2$ along (μ_t^2) at $t = 0$: for this derivative we write $\boldsymbol{\nabla}_{u_0^1} u_t^2$. Similarly for (u_t^1).

Let us consider vector fields as derivations, and the functional $\mu \mapsto F_\varphi(\mu) := \int \varphi d\mu$, for given $\varphi \in C_c^\infty(\mathbb{R}^d)$. By the continuity equation, the derivative of F_φ along u_t^2 is equal to $\langle \nabla\varphi, u_t^2\rangle_{\mu_t^2}$, therefore the compatibility with the metric (99) gives:

$$\begin{aligned} u^1(u^2(F_\varphi))(\mu) &= \frac{d}{dt}\langle \nabla\varphi, u_t^2\rangle_{\mu_t^2}|_{t=0} = \langle \nabla^2\varphi \cdot v_0^2, u_0^2\rangle_\mu + \left\langle \nabla\varphi, \boldsymbol{\nabla}_{u_0^1} u_t^2\right\rangle_\mu \\ &= \langle \nabla^2\varphi \cdot u_0^1, u_0^2\rangle_\mu + \left\langle \nabla\varphi, \boldsymbol{\nabla}_{u_0^1} u_t^2\right\rangle_\mu. \end{aligned}$$



Subtracting the analogous term $u^2(u^1(F_\varphi))(\mu)$ and using the symmetry of $\nabla^2\varphi$ we get

$$[u^1, u^2](F_\varphi)(\mu) = \left\langle \nabla\varphi, \boldsymbol{\nabla}_{u_0^1} u_t^2 - \boldsymbol{\nabla}_{u_0^2} u_t^1 \right\rangle_\mu .$$

Given that the set $\{\nabla\varphi\}_{\varphi\in C_c^\infty}$ is dense in $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$, the above equation characterizes $[u^1, u^2]$ as:

$$[u^1, u^2] = \boldsymbol{\nabla}_{u_0^1} u_t^2 - \boldsymbol{\nabla}_{u_0^2} u_t^1, \tag{100}$$

which proves the torsion-free identity for the covariant derivative.

Example 7.9 (The velocity vector field of a geodesic). Let (μ_t) be the restriction to $[0, 1]$ of a geodesic defined in some larger interval $(-\varepsilon, 1 + \varepsilon)$ and let (v_t) be its velocity vector field. Then we know by Proposition 7.3 that (μ_t) is regular. Also, from formula (95) it is easy to see that it holds

$$v_s \circ \mathbf{T}(t, s, \cdot) = v_t, \qquad \forall t, s \in [0, 1],$$

and thus (v_t) is absolutely continuous and satisfies $\frac{\boldsymbol{d}}{dt} v_t = 0$ and a fortiori $\frac{\boldsymbol{D}}{dt} v_t = 0$.

Thus, as expected, the velocity vector field of a geodesic has zero covariant derivative, in analogy with the standard Riemannian case. Actually, it is interesting to observe that not only the covariant derivative is 0 in this case, but also the total one. ■

Now we pass to the question of parallel transport. The definition comes naturally:

Definition 7.10 (Parallel transport). Let (μ_t) be a regular curve. A tangent vector field (u_t) along it is a parallel transport if it is absolutely continuous and

$$\frac{\boldsymbol{D}}{dt} u_t = 0, \qquad a.e.\ t.$$

It is immediate to verify that the scalar product of two parallel transports is preserved in time, indeed the compatibility with the metric (99) yields

$$\frac{d}{dt}\langle u_t^1, u_t^2\rangle_{\mu_t} = \left\langle \frac{\boldsymbol{D}}{dt} u_t^1, u_t^2 \right\rangle_{\mu_t} + \left\langle u_t^1, \frac{\boldsymbol{D}}{dt} u_t^2 \right\rangle_{\mu_t} = 0, \quad a.e.\ t,$$

for any couple of parallel transports. In particular, this fact and the linearity of the notion of parallel transport give uniqueness of the parallel transport itself, in the sense that for any $u^0 \in \mathrm{Tan}_{\mu_0}(\mathscr{P}_2(\mathbb{R}^d))$ there exists at most one parallel transport (u_t) along (μ_t) satisfying $u_0 = u^0$.

Thus the problem is to show the existence. There is an important analogy, which helps understanding the proof, that we want to point out: we already know that the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ looks like a Riemannian manifold, but actually it has also stronger similarities with a Riemannian manifold M embedded in some bigger space (say, on some Euclidean space $\mathbb{R}^D$), indeed in both cases:

- We have a natural presence of non tangent vectors: elements of $L^2_\mu \setminus \mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ for $\mathscr{P}_2(\mathbb{R}^d)$, and vectors in $\mathbb{R}^D$ non tangent to the manifold for the embedded case.
- The scalar product in the tangent space can be naturally defined also for non tangent vectors: scalar product in L^2_μ for the space $\mathscr{P}_2(\mathbb{R}^d)$, and the scalar product in $\mathbb{R}^D$ for the embedded case. This means in particular that there are natural orthogonal projections from the set of tangent and non tangent vectors onto the set of tangent vectors: $\mathrm{P}_\mu : L^2_\mu \to \mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ for $\mathscr{P}_2(\mathbb{R}^d)$ and $P_x : \mathbb{R}^D \to T_x M$ for the embedded case.
- The Covariant derivative of a tangent vector field is given by projecting the "time derivative" onto the tangent space. Indeed, for the space $\mathscr{P}_2(\mathbb{R}^d)$ we know that the covariant derivative is given by formula (98), while for the embedded manifold it holds:
$$\nabla_{\dot\gamma_t} u_t = P_{\gamma_t}\left(\frac{d}{dt}u_t\right), \tag{101}$$
where $t \mapsto \gamma_t$ is a smooth curve and $t \mapsto u_t \in T_{\gamma_t}M$ is a smooth tangent vector field.

Given these analogies, we are going to proceed as follows: first we give a proof of the existence of the parallel transport along a smooth curve in an embedded Riemannian manifold, then we will see how this proof can be adapted to the Wasserstein case: this approach should help highlighting what's the geometric idea behind the construction.

Thus, say that M is a given smooth Riemannian manifold embedded on $\mathbb{R}^D$, $t \mapsto \gamma_t \in M$ a smooth curve on $[0, 1]$ and $u^0 \in T_{\gamma_0}M$ is a given tangent vector. Our goal is to prove the existence of an absolutely continuous vector field $t \mapsto u_t \in T_{\gamma_t}M$ such that $u_0 = u^0$ and
$$P_{\gamma_t}\left(\frac{d}{dt}u_t\right) = 0, \qquad a.e.\ t.$$

For any $t, s \in [0, 1]$, let $\mathrm{tr}_t^s : T_{\gamma_t}\mathbb{R}^D \to T_{\gamma_s}\mathbb{R}^D$ be the natural translation map which takes a vector with base point γ_t (tangent or not to the manifold) and gives back the translated of this vector with base point γ_s. Notice that an effect of the curvature of the manifold and the chosen embedding on $\mathbb{R}^D$, is that $\mathrm{tr}_t^s(u)$ may be not tangent to M even if u is. Now define $P_t^s : T_{\gamma_t}\mathbb{R}^D \to T_{\gamma_s}M$ by

$$P_t^s(u) := P_{\gamma_s}(\mathrm{tr}_t^s(u)), \qquad \forall u \in T_{\gamma_t}\mathbb{R}^D.$$

An immediate consequence of the smoothness of M and γ are the two inequalities:

$$|\mathrm{tr}_t^s(u) - P_t^s(u)| \le C|u||s-t|, \qquad \forall t,\ s \in [0,1] \text{ and } u \in T_{\gamma_t}M, \tag{102a}$$

$$|P_t^s(u)| \le C|u||s-t|, \qquad \forall t,\ s \in [0,1] \text{ and } u \in T_{\gamma_t}^{\perp}M, \tag{102b}$$

where $T_{\gamma_t}^{\perp}M$ is the orthogonal complement of $T_{\gamma_t}M$ in $T_{\gamma_t}\mathbb{R}^D$. These two inequalities are all we need to prove existence of the parallel transport. The proof will be constructive, and is based on the identity:

$$\nabla_{\dot\gamma_t} P_0^t(u)|_{t=0} = 0, \quad \forall u \in T_{\gamma(0)}M, \tag{103}$$

which tells that the vectors $P_0^t(u)$ are a first order approximation at $t = 0$ of the parallel transport. Taking (101) into account, (103) is equivalent to

$$|P_t^0(\mathrm{tr}_0^t(u) - P_0^t(u))| = o(t), \qquad u \in T_{\gamma(0)}M. \tag{104}$$

Equation (104) follows by applying inequalities (102) (note that $\mathrm{tr}_0^t(u) - P_0^t(u) \in T_{\gamma_t}^{\perp}M$):

$$|P_t^0(\mathrm{tr}_0^t(u) - P_0^t(u))| \le Ct|\mathrm{tr}_0^t(u) - P_0^t(u)| \le C^2t^2|u|.$$

Now, let $\mathfrak{P}$ be the direct set of all the partitions of $[0,1]$, where, for $\mathscr{P}, \mathscr{Q} \in \mathfrak{P}$, $\mathscr{P} \ge \mathscr{Q}$ if $\mathscr{P}$ is a refinement of $\mathscr{Q}$. For $\mathscr{P} = \{0 = t_0 < t_1 < \cdots < t_N = 1\} \in \mathfrak{P}$ and $u \in T_{\gamma_0}M$ define $\mathscr{P}(u) \in T_{\gamma_1}M$ as:

$$\mathscr{P}(u) := P_{t_{N-1}}^{t_N}(P_{t_{N-2}}^{t_{N-1}}(\cdots(P_0^{t_1}(u)))).$$

Our first goal is to prove that the limit $\mathscr{P}(u)$ for $\mathscr{P} \in \mathfrak{P}$ exists. This will naturally define a curve $t \to u_t \in T_{\gamma_t}M$ by taking partitions of $[0,t]$ instead of $[0,1]$: the final goal is to show that this curve is actually the parallel transport of u along the curve γ.

The proof is based on the following lemma.

Lemma 7.11. *Let* $0 \le s_1 \le s_2 \le s_3 \le 1$ *be given numbers. Then it holds:*

$$\left|P_{s_1}^{s_3}(u) - P_{s_2}^{s_3}(P_{s_1}^{s_2}(u))\right| \le C^2|u||s_1 - s_2||s_2 - s_3|, \quad \forall u \in T_{\gamma_{s_1}}M.$$

Proof. From $P_{s_1}^{s_3}(u) = P_{\gamma_{s_3}}(\mathrm{tr}_{s_1}^{s_3}(u)) = P_{\gamma_{s_3}}(\mathrm{tr}_{s_2}^{s_3}(\mathrm{tr}_{s_1}^{s_2}(u)))$ we get

$$P_{s_1}^{s_3}(u) - P_{s_2}^{s_3}(P_{s_1}^{s_2}(u)) = P_{s_2}^{s_3}(\mathrm{tr}_{s_1}^{s_2}(u) - P_{s_1}^{s_2}(u))$$

Since $u \in T_{\gamma_{s_1}}M$ and $\mathrm{tr}_{s_1}^{s_2}(u) - P_{s_1}^{s_2}(u) \in T_{\gamma_{s_2}}^{\perp}M$, the proof follows applying inequalities (102). □

From this lemma, an easy induction shows that for any $0 \leq s_1 < \cdots < s_N \leq 1$ and $u \in T_{\gamma_{s_1}} M$ we have

$$\begin{aligned}
&\left|P_{s_1}^{s_N}(u) - P_{s_{N-1}}^{s_N}(P_{s_{N-2}}^{s_{N-1}}(\cdots(P_{s_1}^{s_2}(u))))\right| \\
&\quad \leq \left|P_{s_1}^{s_N}(u) - P_{s_{N-1}}^{s_N}(P_{s_1}^{s_{N-1}}(u))\right| + \left|P_{s_{N-1}}^{s_N}(P_{s_1}^{s_{N-1}}(u)) - P_{s_{N-1}}^{s_N}(P_{s_{N-2}}^{s_{N-1}}(\cdots(P_{s_1}^{s_2}(u))))\right| \\
&\quad \leq C^2|u||s_{N_1} - s_1|!s_N - s_{N-1}| + \left|P_{s_1}^{s_{N-1}}(u) - P_{s_{N-2}}^{s_{N-1}}(\cdots(P_{s_1}^{s_2}(u)))\right| \\
&\quad \leq \cdots \\
&\quad \leq C^2|u| \sum_{i=2}^{N-1} |s_1 - s_i||s_i - s_{i+1}| \leq C^2|u||s_1 - s_N|^2.
\end{aligned} \tag{105}$$

With this result, we can prove existence of the limit of $P(u)$ as P varies in $\mathfrak{P}$.

Theorem 7.12. *For any $u \in T_{\gamma_0} M$ there exists the limit of $\mathscr{P}(u)$ as $\mathscr{P}$ varies in $\mathfrak{P}$.*

Proof. We have to prove that, given $\varepsilon > 0$, there exists a partition $\mathscr{P}$ such that

$$|\mathscr{P}(u) - \mathscr{Q}(u)| \leq |u|\varepsilon, \quad \forall \mathscr{Q} \geq \mathscr{P}. \tag{106}$$

In order to do so, it is sufficient to find $0 = t_0 < t_1 < \cdots < t_N = 1$ such that $\sum_i |t_{i+1} - t_i|^2 \leq \varepsilon / C^2$, and repeatedly apply (105) to all partitions induced by $\mathscr{Q}$ in the intervals (t_i, t_{i+1}). □

Now, for $s \leq t$ we can introduce the maps $T_t^s : T_{\gamma_t} M \to T_{\gamma_s} M$ which associate to the vector $u \in T_{\gamma_t} M$ the limit of the process just described taking into account partitions of $[s, t]$ instead of those of $[0, 1]$.

Theorem 7.13. *For any $t_1 \leq t_2 \leq t_3 \in [0, 1]$ it holds*

$$T_{t_2}^{t_3} \circ T_{t_1}^{t_2} = T_{t_1}^{t_3}. \tag{107}$$

Moreover, for any $u \in T_{\gamma_0} M$ the curve $t \to u_t := T_0^t(u) \in T_{\gamma_t} M$ is the parallel transport of u along γ.

Proof. For the group property, consider those partitions of $[t_1, t_3]$ which contain t_2 and pass to the limit first on $[t_1, t_2]$ and then on $[t_2, t_3]$. To prove the second part of the statement, we prove first that (u_t) is absolutely continuous. To see this, pass to the limit in (105) with $s_1 = t_0$ and $s_N = t_1$, $u = u_{t_0}$ to get

$$|P_{t_0}^{t_1}(u_{t_0}) - u_{t_1}| \leq C^2|u_{t_0}|(t_1 - t_0)^2 \leq C^2|u|(t_1 - t_0)^2, \tag{108}$$

so that from (102a) we get

$$|\mathrm{tr}_{t_0}^{t_1}(u_{t_0}) - u_{t_1}| \leq |\mathrm{tr}_{t_0}^{t_1}(u_{t_0}) - P_{t_0}^{t_1}(u_{t_0})| + |P_{t_0}^{t_1}(u_{t_0}) - u_{t_1}| \leq C|u||t_1 - t_0|(1 + C|t_1 - t_0|),$$

which shows the absolute continuity. Finally, due to (107), it is sufficient to check that the covariant derivative vanishes at 0. To see this, put $t_0 = 0$ and $t_1 = t$ in (108) to get $|P_0^t(u) - u_t| \leq C^2|u|t^2$, so that the thesis follows from (103). □

Now we come back to the Wasserstein case. To follow the analogy with the Riemannian case, keep in mind that the analogous of the translation map tr_t^s is the right composition with $\mathbf{T}(s,t,\cdot)$, and the analogous of the map P_t^s is

$$\mathscr{P}_t^s(u) := \mathrm{P}_{\mu_s}(u \circ \mathbf{T}(s,t,\cdot)),$$

which maps $L^2_{\mu_t}$ onto $\mathrm{Tan}_{\mu_s}(\mathscr{P}_2(\mathbb{R}^d))$ We saw that the key to prove the existence of the parallel transport in the embedded Riemannian case are inequalities (102). Thus, given that we want to imitate the approach in the Wasserstein setting, we need to produce an analogous of those inequalities. This is the content of the following lemma.

We will denote by $\mathrm{Tan}^\perp_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ the orthogonal complement of $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ in L^2_μ.

Lemma 7.14 (Control of the angles between tangent spaces). *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ be any Borel map satisfying $T_\#\mu = \nu$. Then it holds:*

$$\|v \circ T - \mathrm{P}_\mu(v \circ T)\|_\mu \leq \|v\|_\nu \mathrm{Lip}(T - Id), \qquad \forall v \in \mathrm{Tan}_\nu(\mathscr{P}_2(\nu)\mathbb{R}^d),$$

and, if T is invertible, it also holds

$$\|\mathrm{P}_\mu(w \circ T)\|_\mu \leq \|w\|_\nu \mathrm{Lip}(T^{-1} - Id), \qquad \forall w \in \mathrm{Tan}^\perp_\nu(\mathscr{P}_2(\nu)\mathbb{R}^d).$$

Proof. We start with the first inequality, which is equivalent to

$$\|\nabla\varphi \circ T - P_\mu(\nabla\varphi \circ T)\|_\mu \leq \|\nabla\varphi\|_\nu \mathrm{Lip}(T - Id), \qquad \forall \varphi \in C_c^\infty(\mathbb{R}^d). \quad (109)$$

Let us suppose first that $T - Id \in C_c^\infty(\mathbb{R}^d)$. In this case the map $\varphi \circ T$ is in $C_c^\infty(\mathbb{R}^d)$, too, and therefore $\nabla(\varphi \circ T) = \nabla T \cdot (\nabla\varphi) \circ T$ belongs to $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$. From the minimality properties of the projection we get:

$$\begin{aligned}
\|\nabla\varphi \circ T - P_\mu(\nabla\varphi \circ T)\|_\mu &\leq \|\nabla\varphi \circ T - \nabla T \cdot (\nabla\varphi) \circ T\|_\mu \\
&= \left(\int |(I - \nabla T(x)) \cdot \nabla\varphi(T(x))|^2 d\mu(x)\right)^{1/2} \\
&\leq \left(\int |\nabla\varphi(T(x))|^2 \|\nabla(Id - T)(x)\|^2_{op} d\mu(x)\right)^{1/2} \\
&\leq \|\nabla\varphi\|_\nu \mathrm{Lip}(T - Id),
\end{aligned}$$

where I is the identity matrix and $\|\nabla(Id - T)(x)\|_{op}$ is the operator norm of the linear functional from $\mathbb{R}^d$ to $\mathbb{R}^d$ given by $v \mapsto \nabla(Id - T)(x) \cdot v$.

Now turn to the general case, and we can certainly assume that T is Lipschitz. Then, it is not hard to see that there exists a sequence $(T_n - Id) \subset C_c^\infty(\mathbb{R}^d)$ such that $T_n \to T$ uniformly on compact sets and $\overline{\lim}_n \operatorname{Lip}(T_n - Id) \le \operatorname{Lip}(T - Id)$. It is clear that for such a sequence it holds $\|T - T_n\|_\mu \to 0$, and we have

$$\begin{aligned}\|\nabla\varphi \circ T - P_\mu(\nabla\varphi \circ T)\|_\mu &\le \|\nabla\varphi \circ T - \nabla(\varphi \circ T_n)\|_\mu \\ &\le \|\nabla\varphi \circ T - \nabla\varphi \circ T_n\|_\mu + \|\nabla\varphi \circ T_n - \nabla(\varphi \circ T_n)\|_\mu \\ &\le \operatorname{Lip}(\nabla\varphi)\|T - T_n\|_\mu + \|\nabla\varphi \circ T_n\|_\mu \operatorname{Lip}(T_n - Id).\end{aligned}$$

Letting $n \to +\infty$ we get the thesis.

For the second inequality, just notice that

$$\begin{aligned}\|\mathrm{P}_\mu(w \circ T)\|_\mu &= \sup_{\substack{v \in \operatorname{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d) \\ \|v\|_\mu = 1}} \langle w \circ T, v\rangle_\mu = \sup_{\substack{v \in \operatorname{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d) \\ \|v\|_\mu = 1}} \left\langle w, v \circ T^{-1}\right\rangle_\nu \\ &= \sup_{\substack{v \in \operatorname{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d) \\ \|v\|_\mu = 1}} \left\langle w, v \circ T^{-1} - \mathrm{P}_\nu(v \circ T^{-1})\right\rangle_\nu \le \|w\|_\nu \operatorname{Lip}(T^{-1} - Id)\end{aligned}$$

□

From this lemma and the inequality

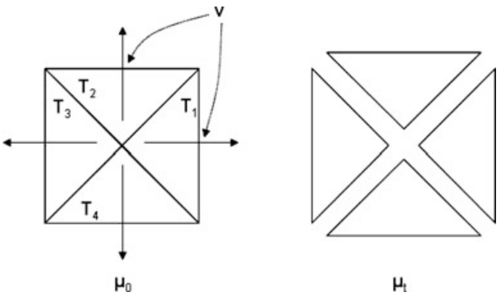
$$\operatorname{Lip}\Big(\mathbf{T}(s,t,\cdot) - Id\Big) \le e^{\left|\int_t^s \operatorname{Lip}(v_r)dr\right|} - 1 \le C\left|\int_t^s \operatorname{Lip}(v_r)dr\right|, \qquad \forall t, s \in [0,1],$$

(whose simple proof we omit), where $C := e^{\int_0^1 \operatorname{Lip}(v_r)dr} - 1$, it is immediate to verify that it holds:

$$\begin{aligned}\|u \circ \mathbf{T}(s,t,\cdot) - \mathscr{P}_t^s(u)\|_{\mu_s} &\le C\|u\|_{\mu_t}\left|\int_t^s \operatorname{Lip}(v_r)dr\right|, \qquad u \in \operatorname{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d)), \\ \|\mathscr{P}_t^s(u)\|_{\mu_s} &\le C\|u\|_{\mu_t}\left|\int_t^s \operatorname{Lip}(v_r)dr\right|, \qquad u \in \operatorname{Tan}_{\mu_t}^\perp(\mathscr{P}_2(\mathbb{R}^d)).\end{aligned} \tag{110}$$

These inequalities are perfectly analogous to the (102) (well, the only difference is that here the bound on the angle is L^1 in t, s while for the embedded case it was L^∞, but this does not really change anything). Therefore the arguments presented before apply also to this case, and we can derive the existence of the parallel transport along regular curves:

Theorem 7.15 (Parallel transport along regular curves). *Let (μ_t) be a regular curve and $u^0 \in \operatorname{Tan}_{\mu_0}(\mathscr{P}_2(\mathbb{R}^d))$. Then there exists a parallel transport (u_t) along (μ_t) such that $u_0 = u^0$.*



Now, we know that the parallel transport exists along regular curves, and we know also that regular curves are dense, it is therefore natural to try to construct the parallel transport along any absolutely continuous curve via some limiting argument. However, this cannot be done, as the following counterexample shows:

Example 7.16 (Non existence of parallel transport along a non regular geodesic). Let $Q = [0,1] \times [0,1]$ be the unit square in $\mathbb{R}^2$ and let T_i, $i = 1,2,3,4$, be the four open triangles in which Q is divided by its diagonals. Let $\mu_0 := \chi_Q \mathscr{L}^2$ and define the function $v : Q \to \mathbb{R}^2$ as the gradient of the convex map $\max\{|x|,|y|\}$, as in the figure. Set also $w = v^\perp$, the rotation by $\pi/2$ of v, in Q and $w = 0$ out of Q. Notice that $\nabla \cdot (w\mu_0) = 0$.

Set $\mu_t := (Id + tv)_\# \mu_0$ and observe that, for positive t, the support Q_t of μ_t is made of four connected components, each one the translation of one of the sets T_i, and that $\mu_t = \chi_{Q_t} \mathscr{L}^2$.

It is immediate to check that (μ_t) is a geodesic in $[0,\infty)$, so that from Proposition 7.3 we know that the restriction of μ_t to any interval $[\varepsilon,1]$ with $\varepsilon > 0$ is regular. Fix $\varepsilon > 0$ and note that, by construction, the flow maps of μ_t in $[\varepsilon,1]$ are given by

$$\mathbf{T}(t,s,\cdot) = (Id + sv) \circ (Id + tv)^{-1}, \quad \forall t,s \in [\varepsilon,1].$$

Now, set $w_t := w \circ \mathbf{T}(t,0,\cdot)$ and notice that w_t is tangent at μ_t (because w_t is constant in the connected components of the support of μ_t, so we can define a C_c^∞ function to be affine on each connected component and with gradient given by w_t, and then use the space between the components themselves to rearrange smoothly the function). Since $w_{t+h} \circ \mathbf{T}(t,t+h,\cdot) = w_t$, we have $\frac{\boldsymbol{d}}{dt} w_t = 0$ and a fortiori $\frac{\boldsymbol{D}}{dt} w_t = 0$. Thus (w_t) is a parallel transport in $[\varepsilon,1]$. Furthermore, since $\nabla \cdot (w\mu_0) = 0$, we have $w_0 = w \notin \mathrm{Tan}_{\mu_0}(\mathscr{P}_2(\mathbb{R}^2))$. Therefore there is no way to extend w_t to a continuous *tangent* vector field on the whole $[0,1]$. In particular, there is no way to extend the parallel transport up to $t = 0$. ■

Now we pass to the calculus of total and covariant derivatives. Let (μ_t) be a fixed regular curve and let (v_t) be its velocity vector field. Start observing that, if (u_t) is absolutely continuous along (μ_t), then $(\mathrm{P}_{\mu_t}(u_t))$ is absolutely continuous as well, as it follows from the inequality

$$
\begin{aligned}
&\left\|\big(\mathrm{P}_{\mu_s}(u_s)\big)\circ\mathbf{T}(t,s,\cdot)-\mathrm{P}_{\mu_t}(u_t)\right\|_{\mu_t}\\
&\quad\le\left\|\big(\mathrm{P}_{\mu_s}(u_s)\big)\circ\mathbf{T}(t,s,\cdot)-\mathrm{P}_{\mu_t}\Big(\big(\mathrm{P}_{\mu_s}(u_s)\big)\circ\mathbf{T}(t,s,\cdot)\Big)\right\|_{\mu_t}\\
&\qquad+\left\|\mathrm{P}_{\mu_t}\Big(\big(\mathrm{P}_{\mu_s}(u_s)\big)\circ\mathbf{T}(t,s,\cdot)\Big)-\mathrm{P}_{\mu_t}\big(u_s\circ\mathbf{T}(t,s,\cdot)\big)\right\|_{\mu_t}\\
&\qquad+\left\|\mathrm{P}_{\mu_t}(u_s\circ\mathbf{T}(t,s,\cdot))-\mathrm{P}_{\mu_t}(u_t)\right\|_{\mu_t}\\
&\quad\le\left\|\mathrm{P}^{\perp}_{\mu_t}\Big(\mathrm{P}_{\mu_s}(u_s)\circ\mathbf{T}(t,s,\cdot)\Big)\right\|_{\mu_t}\\
&\qquad+\left\|\mathrm{P}_{\mu_t}\Big(\mathrm{P}^{\perp}_{\mu_s}(u_s)\circ\mathbf{T}(t,s,\cdot)\Big)\right\|_{\mu_t}\\
&\qquad+\|u_s\circ\mathbf{T}(t,s,\cdot)-u_t\|_{\mu_t}\\
&\quad\overset{(110)}{\le}2SC\int_t^s\mathrm{Lip}(v_r)dr+\int_t^s\left\|\frac{\boldsymbol{d}}{dr}u_r\right\|_{\mu_r}dr,
\end{aligned}
\tag{111}
$$

valid for any $t\le s$, where $S:=\sup_t\|u_t\|_{\mu_t}$. Thus $(\mathrm{P}_{\mu_t}(u_t))$ has a well defined covariant derivative for a.e. t. The question is: can we find a formula to express this derivative?

To compute it, apply the Leibniz rule for the total and covariant derivatives ((96) and (99)), to get that for a.e. $t\in[0,1]$ it holds

$$
\begin{aligned}
\frac{d}{dt}\left\langle\mathrm{P}_{\mu_t}(u_t),\nabla\varphi\right\rangle_{\mu_t}&=\left\langle\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t),\nabla\varphi\right\rangle_{\mu_t}+\left\langle\mathrm{P}_{\mu_t}(u_t),\frac{\boldsymbol{D}}{dt}\nabla\varphi\right\rangle_{\mu_t},\\
\frac{d}{dt}\left\langle u_t,\nabla\varphi\right\rangle_{\mu_t}&=\left\langle\frac{\boldsymbol{d}}{dt}u_t,\nabla\varphi\right\rangle_{\mu_t}+\left\langle u_t,\frac{\boldsymbol{d}}{dt}\nabla\varphi\right\rangle_{\mu_t}.
\end{aligned}
$$

Since $\nabla\varphi\in\mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for any t, it holds $\left\langle\mathrm{P}_{\mu_t}(u_t),\nabla\varphi\right\rangle_{\mu_t}=\langle u_t,\nabla\varphi\rangle_{\mu_t}$ for any $t\in[0,1]$, and thus the left hand sides of the previous equations are equal for a.e. t. Recalling formula (97) we have $\frac{\boldsymbol{d}}{dt}\nabla\varphi=\nabla^2\varphi\cdot v_t$ and $\frac{\boldsymbol{D}}{dt}\nabla\varphi=\mathrm{P}_{\mu_t}(\nabla^2\varphi\cdot v_t)$, thus from the equality of the right hand sides we obtain

$$
\begin{aligned}
\left\langle\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t),\nabla\varphi\right\rangle_{\mu_t}&=\left\langle\frac{\boldsymbol{d}}{dt}u_t,\nabla\varphi\right\rangle_{\mu_t}+\left\langle u_t,\nabla^2\varphi\cdot v_t\right\rangle_{\mu_t}-\left\langle\mathrm{P}_{\mu_t}(u_t),\mathrm{P}_{\mu_t}(\nabla^2\varphi\cdot v_t)\right\rangle_{\mu_t}\\
&=\left\langle\frac{\boldsymbol{d}}{dt}u_t,\nabla\varphi\right\rangle_{\mu_t}+\left\langle\mathrm{P}^{\perp}_{\mu_t}(u_t),\mathrm{P}^{\perp}_{\mu_t}(\nabla^2\varphi\cdot v_t)\right\rangle_{\mu_t}.
\end{aligned}
\tag{112}
$$

This formula characterizes the scalar product of $\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t)$ with any $\nabla\varphi$ when φ varies on $C_c^\infty(\mathbb{R}^d)$. Since the set $\{\nabla\varphi\}$ is dense in $\mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$ for any $t\in[0,1]$, the formula actually identifies $\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t)$.

However, from this expression it is unclear what is the value of $\left\langle \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t), w\right\rangle_{\mu_t}$ for a general $w \in \mathrm{Tan}_{\mu_t}(\mathscr{P}_2(\mathbb{R}^d))$, because some regularity of $\nabla\varphi$ seems required to compute $\nabla^2\varphi \cdot v_t$. In order to better understand what the value of $\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t)$ is, fix $t \in [0,1]$ and assume for a moment that $v_t \in C_c^\infty(\mathbb{R}^d)$. Then compute the gradient of $x \mapsto \langle \nabla\varphi(x), v_t(x)\rangle$ to obtain

$$\nabla \langle \nabla\varphi, v_t\rangle = \nabla^2\varphi \cdot v_t + \nabla v_t^{\mathrm{t}} \cdot \nabla\varphi,$$

and consider this expression as an equality between vector fields in $L^2_{\mu_t}$. Taking the projection onto the Normal space we derive

$$\mathrm{P}^\perp_{\mu_t}(\nabla^2\varphi \cdot v_t) + \mathrm{P}^\perp_{\mu_t}(\nabla v_t^{\mathrm{t}} \cdot \nabla\varphi) = 0.$$

Plugging the expression for $\mathrm{P}^\perp_{\mu_t}(\nabla^2\varphi \cdot v_t)$ into the formula for the covariant derivative we get

$$\begin{aligned} \left\langle \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t), \nabla\varphi\right\rangle_{\mu_t} &= \left\langle \frac{\boldsymbol{d}}{dt}u_t, \nabla\varphi\right\rangle_{\mu_t} - \left\langle \mathrm{P}^\perp_{\mu_t}(u_t), \mathrm{P}^\perp_{\mu_t}(\nabla v_t^{\mathrm{t}} \cdot \nabla\varphi)\right\rangle_{\mu_t} \\ &= \left\langle \frac{\boldsymbol{d}}{dt}u_t, \nabla\varphi\right\rangle_{\mu_t} - \left\langle \nabla v_t \cdot \mathrm{P}^\perp_{\mu_t}(u_t), \nabla\varphi\right\rangle_{\mu_t}, \end{aligned}$$

which identifies $\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t)$ as

$$\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t) = \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t - \nabla v_t \cdot \mathrm{P}^\perp_{\mu_t}(u_t)\right). \tag{113}$$

We found this expression assuming that v_t was a smooth vector field, but given that we know that $\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t)$ exists for a.e. t, it is realistic to believe that the expression makes sense also for general Lipschitz v_t's. The problem is that the object ∇v_t may very well be not defined μ_t-a.e. for arbitrary μ_t and Lipschitz v_t (Rademacher's theorem is of no help here, because we are not assuming the measures μ_t to be absolutely continuous w.r.t. the Lebesgue measure). To give a meaning to formula (113) we need to introduce a new tensor.

Definition 7.17 (The Lipschitz non Lipschitz space). Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. The set $\mathrm{LNL}_\mu \subset [L^2_\mu]^2$ is the set of couples of vector fields (u, v) such that $\min\{\mathrm{Lip}(u), \mathrm{Lip}(v)\} < \infty$, i.e. the set of couples of vectors such that at least one of them is Lipschitz.

We say that a sequence $(u_n, v_n) \in \mathrm{LNL}_\mu$ converges to $(u, v) \in \mathrm{LNL}_\mu$ provided $\|u_n - u\|_\mu \to 0$, $\|v_n - v\|_\mu \to 0$ and

$$\sup_n \min\{\mathrm{Lip}(u_n), \mathrm{Lip}(v_n)\} < \infty.$$

The following theorem holds:

Theorem 7.18 (The Normal tensor). *Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. The map*

$$\begin{aligned} \mathscr{N}_\mu(u,v) : [C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)]^2 \quad &\to \mathrm{Tan}_\mu^\perp(\mathscr{P}_2(\mu)\mathbb{R}^d), \\ (u,v) \quad &\mapsto \mathrm{P}_\mu^\perp(\nabla u^{\mathrm{t}} \cdot v) \end{aligned}$$

extends uniquely to a sequentially continuous bilinear and antisymmetric map, still denoted by $\mathscr{N}_\mu$, from $\mathrm{L{\scriptsize N}L}_\mu$ in $\mathrm{Tan}_\mu^\perp(\mathscr{P}_2(\mu)\mathbb{R}^d)$ for which the bound

$$\|\mathscr{N}_\mu(u,v)\|_\mu \le \min\{\mathrm{Lip}(u)\|v\|_\mu, \mathrm{Lip}(v)\|u\|_\mu\}, \tag{114}$$

holds.

Proof. For $u,\ v \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$ we have $\nabla\langle u,v\rangle = \nabla u^{\mathrm{t}} \cdot v + \nabla v^{\mathrm{t}} \cdot u$ so that taking the projections on $\mathrm{Tan}_\mu^\perp(\mathscr{P}_2(\mu)\mathbb{R}^d)$ we get

$$\mathscr{N}_\mu(u,v) = -\mathscr{N}_\mu(v,u) \qquad \forall u,v \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d).$$

In this case, the bound (114) is trivial.

To prove existence and uniqueness of the sequentially continuous extension, it is enough to show that for any given sequence $n \mapsto (u_n,v_n) \in [C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)]^2$ converging to some $(u,v) \in \mathrm{L{\scriptsize N}L}_\mu$, the sequence $n \mapsto \mathscr{N}_\mu(u_n,v_n) \in \mathrm{Tan}_\mu^\perp(\mathscr{P}_2(\mu)\mathbb{R}^d)$ is a Cauchy sequence. Fix such a sequence (u_n,v_n), let $L := \sup_n \min\{\mathrm{Lip}(u_n),\mathrm{Lip}(v_n)\}$, $I \subset \mathbb{N}$ be the set of indexes n such that $\mathrm{Lip}(u_n) \le L$ and fix two smooth vectors $\tilde{u},\tilde{v} \in C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$.

Notice that for $n,m \in I$ it holds

$$\begin{aligned} \|\mathscr{N}_\mu(u_n,v_n) - \mathscr{N}_\mu(u_m,v_m)\|_\mu &\le \|\mathscr{N}_\mu(u_n,v_n-\tilde{v})\|_\mu + \|\mathscr{N}_\mu(u_n-u_m,\tilde{v})\|_\mu \\ &\quad + \|\mathscr{N}_\mu(u_m,\tilde{v}-v_m)\|_\mu \\ &\le L\|v_n-\tilde{v}\|_\mu + \mathrm{Lip}(\tilde{v})\|u_n-u_m\|_\mu + L\|v_m-\tilde{v}\|_\mu, \end{aligned}$$

and thus

$$\varlimsup_{\substack{n,m\to\infty \\ n,m\in I}} \|\mathscr{N}_\mu(u_n,v_n) - \mathscr{N}_\mu(u_m,v_m)\|_\mu \le 2L\|v-\tilde{v}\|_\mu,$$

(this expression being vacuum if I is finite). If $n \in I$ and $m \notin I$ we have $\mathrm{Lip}(v_m) \le L$ and

$$\begin{aligned} &\|\mathscr{N}_\mu(u_n,v_n) - \mathscr{N}_\mu(u_m,v_m)\|_\mu \\ &\quad \le \|\mathscr{N}_\mu(u_n,v_n-\tilde{v})\|_\mu + \|\mathscr{N}_\mu(u_n-\tilde{u},\tilde{v})\|_\mu + \|\mathscr{N}_\mu(\tilde{u},\tilde{v}-v_m)\|_\mu + \|\mathscr{N}_\mu(\tilde{u}-u_m,v_m)\|_\mu \\ &\quad \le L\|v_n-\tilde{v}\|_\mu + \mathrm{Lip}(\tilde{v})\|u_n-\tilde{u}\|_\mu + \mathrm{Lip}(\tilde{u})\|\tilde{v}-v_m\|_\mu + L\|u_m-\tilde{u}\|_\mu, \end{aligned}$$

which gives

$$\varlimsup_{\substack{n,m\to\infty \\ n\in I,\ m\notin I}} \|\mathscr{N}_\mu(u_n,v_n) - \mathscr{N}_\mu(u_m,v_m)\|_\mu \le L\|v-\tilde{v}\|_\mu + L\|u-\tilde{u}\|_\mu.$$

Exchanging the roles of the u's and the v's in these inequalities for the case in which $n \notin I$ we can conclude

$$\varlimsup_{n,m\to\infty} \|\mathscr{N}_\mu(u_n, v_n) - \mathscr{N}_\mu(u_m, v_m)\|_\mu \le 2L\|v - \tilde{v}\|_\mu + 2L\|u - \tilde{u}\|_\mu.$$

Since $\tilde{u}, \tilde{v}$ are arbitrary, we can let $\tilde{u} \to u$ and $\tilde{v} \to v$ in L^2_μ and conclude that $n \mapsto \mathscr{N}_\mu(u_n, v_n)$ is a Cauchy sequence, as requested.

The other claims follow trivially by the sequential continuity. □

Definition 7.19 (The operators $\mathscr{O}_v(\cdot)$ and $\mathscr{O}^*_v(\cdot)$). Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ and $v \in L^2_\mu$ with $\mathrm{Lip}(v) < \infty$. Then the operator $u \mapsto \mathscr{O}_v(u)$ is defined by

$$\mathscr{O}_v(u) := \mathscr{N}_\mu(v, u).$$

The operator $u \mapsto \mathscr{O}^*_v(u)$ is the adjoint of $\mathscr{O}_v(\cdot)$, i.e. it is defined by

$$\left\langle \mathscr{O}^*_v(u), w \right\rangle_\mu = \langle u, \mathscr{O}_v(w)\rangle_\mu, \qquad \forall w \in L^2_\mu.$$

It is clear that the operator norm of $\mathscr{O}_v(\cdot)$ and $\mathscr{O}^*_v(\cdot)$ is bounded by $\mathrm{Lip}(v)$. Observe that in writing $\mathscr{O}_v(u)$, $\mathscr{O}^*_v(u)$ we are losing the reference to the base measure μ, which certainly plays a role in the definition; this simplifies the notation and hopefully should create no confusion, as the measure we are referring to should always be clear from the context. Notice that if $v \in C^\infty_c(\mathbb{R}^d, \mathbb{R}^d)$ these operators read as

$$\mathscr{O}_v(u) = \mathrm{P}^\perp_\mu(\nabla v^{\mathrm{t}} \cdot u),$$
$$\mathscr{O}^*_v(u) = \nabla v \cdot \mathrm{P}^\perp_\mu(u).$$

The introduction of the operators $\mathscr{O}_v(\cdot)$ and $\mathscr{O}^*_v(\cdot)$ allows to give a precise meaning to formula (113) for general regular curves:

Theorem 7.20 (Covariant derivative of $\mathrm{P}_{\mu_t}(u_t)$). *Let (μ_t) be a regular curve, (v_t) its velocity vector field and let (u_t) be an absolutely continuous vector field along it. Then $(\mathrm{P}_{\mu_t}(u_t))$ is absolutely continuous as well and for a.e. t it holds*

$$\frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t) = \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t - \mathscr{O}^*_{v_t}(u_t)\right). \tag{115}$$

Proof. The fact that $(\mathrm{P}_{\mu_t}(u_t))$ is absolutely continuous has been proved with inequality (111). To get the thesis, start from (112) and conclude noticing that for a.e. t it holds $\mathrm{Lip}(v_t) < \infty$ and thus

$$\mathrm{P}^\perp_{\mu_t}(\nabla^2\varphi \cdot v_t) = \mathscr{N}_\mu(\nabla\varphi, v_t) = -\mathscr{N}_\mu(v_t, \nabla\varphi) = -\mathscr{O}_{v_t}(\nabla\varphi).$$

□

Corollary 7.21 (Total derivatives of $\mathrm{P}_{\mu_t}(u_t)$ and $\mathrm{P}^{\perp}_{\mu_t}(u_t)$). *Let (μ_t) be a regular curve, let (v_t) be its velocity vector field and let (u_t) be an absolutely continuous vector field along it. Then $(\mathrm{P}^{\perp}_{\mu_t}(u_t))$ is absolutely continuous and it holds*

$$\begin{aligned} \frac{\boldsymbol{d}}{dt}\mathrm{P}_{\mu_t}(u_t) &= \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t\right) - \mathrm{P}_{\mu_t}\left(\mathscr{O}^*_{v_t}(u_t)\right) - \mathscr{O}_{v_t}\left(\mathrm{P}_{\mu_t}(u_t)\right), \\ \frac{\boldsymbol{d}}{dt}\mathrm{P}^{\perp}_{\mu_t}(u_t) &= \mathrm{P}^{\perp}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t\right) + \mathrm{P}_{\mu_t}\left(\mathscr{O}^*_{v_t}(u_t)\right) + \mathscr{O}_{v_t}\left(\mathrm{P}_{\mu_t}(u_t)\right). \end{aligned} \tag{116}$$

Proof. The absolute continuity of $(\mathrm{P}^{\perp}_{\mu_t}(u_t))$ follows from the fact that both (u_t) and $(\mathrm{P}_{\mu_t}(u_t))$ are absolutely continuous. Similarly, the second formula in (116) follows immediately from the first one noticing that $u_t = \mathrm{P}_{\mu_t}(u_t) + \mathrm{P}^{\perp}_{\mu_t}(u_t)$ yields $\frac{\boldsymbol{d}}{dt}u_t = \frac{\boldsymbol{d}}{dt}\mathrm{P}_{\mu_t}(u_t) + \frac{\boldsymbol{d}}{dt}\mathrm{P}^{\perp}_{\mu_t}(u_t)$. Thus we have only to prove the first equality in (116). To this aim, let (w_t) be an arbitrary absolutely continuous vector field along (μ_t) and observe that it holds

$$\begin{aligned} \frac{d}{dt}\left\langle \mathrm{P}_{\mu_t}(u_t), w_t\right\rangle_{\mu_t} &= \left\langle \frac{\boldsymbol{d}}{dt}\mathrm{P}_{\mu_t}(u_t), w_t\right\rangle_{\mu_t} + \left\langle \mathrm{P}_{\mu_t}(u_t), \frac{\boldsymbol{d}}{dt}w_t\right\rangle_{\mu_t}, \\ \frac{d}{dt}\left\langle \mathrm{P}_{\mu_t}(u_t), \mathrm{P}_{\mu_t}(w_t)\right\rangle_{\mu_t} &= \left\langle \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t), \mathrm{P}_{\mu_t}(w_t)\right\rangle_{\mu_t} + \left\langle \mathrm{P}_{\mu_t}(u_t), \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(w_t)\right\rangle_{\mu_t}. \end{aligned}$$

Since the left hand sides of these expression are equal, the right hand sides are equal as well, thus we get

$$\begin{aligned} \left\langle \frac{\boldsymbol{d}}{dt}\mathrm{P}_{\mu_t}(u_t) - \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t), w_t\right\rangle_{\mu_t} &= -\left\langle \mathrm{P}_{\mu_t}(u_t), \frac{\boldsymbol{d}}{dt}w_t - \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(w_t)\right\rangle_{\mu_t} \\ &= -\left\langle \mathrm{P}_{\mu_t}(u_t), \mathrm{P}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}w_t\right) - \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(w_t)\right\rangle_{\mu_t} \\ &\overset{(115)}{=} -\left\langle \mathrm{P}_{\mu_t}(u_t), \mathscr{O}^*_{v_t}(w_t)\right\rangle_{\mu_t} \\ &= -\left\langle \mathscr{O}_{v_t}\left(\mathrm{P}_{\mu_t}(u_t)\right), w_t\right\rangle_{\mu_t}, \end{aligned}$$

so that the arbitrariness of (w_t) gives

$$\frac{\boldsymbol{d}}{dt}\mathrm{P}_{\mu_t}(u_t) = \frac{\boldsymbol{D}}{dt}\mathrm{P}_{\mu_t}(u_t) - \mathscr{O}_{v_t}\left(\mathrm{P}_{\mu_t}(u_t)\right),$$

and the conclusion follows from (115). □

Along the same lines, the total derivative of $(\mathscr{N}_{\mu_t}(u_t, w_t))$ for given absolutely continuous vector fields (u_t), (w_t) along the same regular curve (μ_t) can be calculated. The only thing the we must take care of, is the fact that $\mathscr{N}_{\mu_t}$ is not defined on the whole $[L^2_{\mu_t}]^2$, so that we need to make some assumptions on (u_t), (w_t) to be sure that $(\mathscr{N}_{\mu_t}(u_t, w_t))$ is well defined and absolutely continuous. Indeed, observe that from a purely formal point of view, we expect that the total derivative of $(\mathscr{N}_{\mu_t}(u_t, w_t))$ is something like

$$\begin{aligned}\frac{\boldsymbol{d}}{dt}\mathscr{N}_{\mu_t}(u_t, w_t) &= \mathscr{N}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t, w_t\right) + \mathscr{N}_{\mu_t}\left(u_t, \frac{\boldsymbol{d}}{dt}w_t\right)\\ &\quad + \begin{pmatrix}\text{some tensor - which we may think}\\ \text{as the derivative of } \mathscr{N}_{\mu_t} \text{ - applied to the couple } (u_t, w_t)\end{pmatrix}.\end{aligned}$$

Forget about the last object and look at the first two addends: given that the domain of definition of $\mathscr{N}_{\mu_t}$ is not the whole $[L^2_{\mu_t}]^2$, in order for the above formula to make sense, we should ask that in each of the couples $(\frac{\boldsymbol{d}}{dt}u_t, w_t)$ and $(u_t, \frac{\boldsymbol{d}}{dt}w_t)$, at least one vector is Lipschitz. Under the assumption that $\{\int_0^1 \mathrm{Lip}(u_t)dt < \infty$ and $\int_0^1 \mathrm{Lip}(\frac{\boldsymbol{d}}{dt}u_t)dt < +\infty$ }, it is possible to prove the following theorem (whose proof we omit).

Theorem 7.22. *Let (μ_t) be an absolutely continuous curve, let (v_t) be its velocity vector field and let (u_t), (w_t) be two absolutely continuous vector fields along it. Assume that $\int_0^1 \mathrm{Lip}(u_t)dt < \infty$ and $\int_0^1 \mathrm{Lip}(\frac{\boldsymbol{d}}{dt}u_t)dt < +\infty$. Then $(\mathscr{N}_{\mu_t}(u_t, w_t))$ is absolutely continuous and it holds*

$$\begin{aligned}\frac{\boldsymbol{d}}{dt}\mathscr{N}_{\mu_t}(u_t, w_t) &= \mathscr{N}_{\mu_t}\left(\frac{\boldsymbol{d}}{dt}u_t, w_t\right) + \mathscr{N}_{\mu_t}\left(u_t, \frac{\boldsymbol{d}}{dt}w_t\right)\\ &\quad - \mathscr{O}_{v_t}\left(\mathscr{N}_{\mu_t}(u_t, w_t)\right) + \mathrm{P}_{\mu_t}\left(\mathscr{O}^*_{v_t}\left(\mathscr{N}_{\mu_t}(u_t, w_t)\right)\right).\end{aligned} \tag{117}$$

Corollary 7.23. *Let (μ_t) be a regular curve and assume that its velocity vector field (v_t) satisfies:*

$$\int_0^1 \mathrm{Lip}\left(\frac{\boldsymbol{d}}{dt}v_t\right)dt < \infty. \tag{118}$$

Then for every absolutely continuous vector field (u_t) both $(\mathscr{O}_{v_t}(u_t))$ and $(\mathscr{O}^*_{v_t}(u_t))$ are absolutely continuous and their total derivatives are given by:

$$\begin{aligned}\frac{\boldsymbol{d}}{dt}\mathscr{O}_{v_t}(u_t) &= \mathscr{O}_{\frac{\boldsymbol{d}}{dt}v_t}(u_t) + \mathscr{O}_{v_t}\left(\frac{\boldsymbol{d}}{dt}u_t\right) - \mathscr{O}_{v_t}\left(\mathscr{O}_{v_t}(u_t)\right) + \mathrm{P}_{\mu_t}\left(\mathscr{O}^*_{v_t}\left(\mathscr{O}_{v_t}(u_t)\right)\right)\\ \frac{\boldsymbol{d}}{dt}\mathscr{O}^*_{v_t}(u_t) &= \mathscr{O}^*_{\frac{\boldsymbol{d}}{dt}v_t}(u_t) + \mathscr{O}^*_{v_t}\left(\frac{\boldsymbol{d}}{dt}u_t\right) - \mathscr{O}^*_{v_t}\left(\mathscr{O}^*_{v_t}(u_t)\right) + \mathscr{O}^*_{v_t}\left(\mathscr{O}_{v_t}\left(\mathrm{P}_{\mu_t}(u_t)\right)\right)\end{aligned} \tag{119}$$

Proof. The first formula follows directly from Theorem 7.22, the second from the fact that $\mathscr{O}^*_{v_t}(\cdot)$ is the adjoint of $\mathscr{O}_{v_t}(\cdot)$. □

An important feature of (117) and (119) is that to express the derivatives of $(\mathcal{N}_{\mu_t}(u_t, w_t))$, $(\mathcal{O}_{v_t}(u_t))$ and $(\mathcal{O}^*_{v_t}(u_t))$ no "new operators appear". This implies that we can recursively calculate derivatives of any order of the vector fields $(\mathrm{P}_{\mu_t}(u_t))$, $(\mathrm{P}^\perp_{\mu_t}(u_t))$, $\mathcal{O}_{v_t}(u_t)$ and $\mathcal{O}^*_{v_t}(u_t)$, provided—of course—that we make appropriate regularity assumptions on the vector field (u_t) and on the velocity vector field (v_t). An example of result which can be proved following this direction is that the operator $t \mapsto \mathrm{P}_{\mu_t}(\cdot)$ is analytic along (the restriction of) a geodesic:

Proposition 7.24 (Analyticity of $t \mapsto \mathrm{P}_{\mu_t}(\cdot)$). *Let (μ_t) be the restriction to $[0,1]$ of a geodesic defined in some larger interval $[-\varepsilon, 1+\varepsilon]$. Then the operator $t \mapsto \mathrm{P}_{\mu_t}(\cdot)$ is analytic in the following sense. For any $t_0 \in [0,1]$ there exists a sequence of bounded linear operators $A_n : L^2_{\mu_{t_0}} \to L^2_{\mu_{t_0}}$ such that the following equality holds in a neighborhood of t_0*

$$\mathrm{P}_{\mu_t}(u) = \sum_{n\in\mathbb{N}} \frac{(t-t_0)^n}{n!} A_n\big(u \circ \mathbf{T}(t_0, t, \cdot)\big) \circ \mathbf{T}(t, t_0, \cdot), \qquad \forall u \in L^2_{\mu_t}. \tag{120}$$

Proof. From the fact that (μ_t) is the restriction of a geodesic we know that $L := \sup_{t\in[0,1]} \mathrm{Lip}(v_t) < \infty$ and that $\frac{\boldsymbol{d}}{dt} v_t = 0$ (recall Example 7.9). In particular condition (118) is fulfilled.

Fix $t_0 \in [0,1]$, $u \in L^2_{\mu_{t_0}}$ and define $u_t := u \circ \mathbf{T}(t, t_0, \cdot)$, so that $\frac{\boldsymbol{d}}{dt} u_t = 0$. From (116) and (119) and by induction it follows that $(\mathrm{P}_{\mu_t}(u_t))$ is C^∞. Also, $\frac{\boldsymbol{d}^n}{dt^n}\mathrm{P}_{\mu_t}(u_t)$ is the sum of addends each of which is the composition of projections onto the tangent or normal space and up to n operators $\mathcal{O}_{v_t}(\cdot)$ and $\mathcal{O}^*_{v_t}(\cdot)$, applied to the vector u_t. Since the operator norm of $\mathcal{O}_{v_t}(\cdot)$ and $\mathcal{O}^*_{v_t}(\cdot)$ is bounded by L, we deduce that

$$\left\| \frac{\boldsymbol{d}^n}{dt^n} \mathrm{P}_{\mu_t}(u_t) \right\|_{\mu_t} \le \|u_t\|_{\mu_t} L^n = \|u\|_{\mu_{t_0}} L^n, \qquad \forall n \in \mathbb{N},\ t \in [0,1].$$

Defining the curve $t \mapsto U_t := \mathrm{P}_{\mu_t}(u_t) \circ \mathbf{T}(t_0, t, \cdot) \in L^2_{\mu_{t_0}}$, the above bound can be written as

$$\left\| \frac{d^n}{dt^n} U_t \right\|_{\mu_{t_0}} \le \|U_{t_0}\|_{\mu_{t_0}} L^n, \qquad \forall n \in \mathbb{N},\ t \in [0,1],$$

which implies that the curve $t \mapsto U_t \in L^2_{\mu_{t_0}}$ is analytic. This means that for t close to t_0 it holds

$$\mathrm{P}_{\mu_t}(u_t) \circ \mathbf{T}(t_0, t, \cdot) = \sum_{n\ge 0} \frac{(t-t_0)^n}{n!} \frac{\boldsymbol{d}^n}{dt^n}|_{t=t_0} (\mathrm{P}_{\mu_t}(u_t)).$$

Now notice that (116) and (119) and the fact that $\frac{\boldsymbol{d}}{dt} u_t \equiv 0$ ensure that $\frac{\boldsymbol{d}^n}{dt^n}|_{t=t_0}$ $(\mathrm{P}_{\mu_t}(u_t)) = A_n(u)$, where $A_n : L^2_{\mu_{t_0}} \to L^2_{\mu_{t_0}}$ is bounded. Thus the thesis follows by the arbitrariness of $u \in L^2_{\mu_{t_0}}$. □

Now we have all the technical tools we need in order to study the curvature tensor of the "manifold" $\mathscr{P}_2(\mathbb{R}^d)$.

Following the analogy with the Riemannian case, we are lead to define the curvature tensor in the following way: given three vector fields $\mu \mapsto \nabla\varphi^i_\mu \in \mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$, $i = 1, \dots, 3$, the curvature tensor $\mathbf{R}$ calculated on them at the measure μ is defined as:

$$\mathbf{R}(\nabla\varphi^1_\mu, \nabla\varphi^2_\mu)(\nabla\varphi^3_\mu) := \boldsymbol{\nabla}_{\nabla\varphi^2_\mu}(\boldsymbol{\nabla}_{\nabla\varphi^1_\mu}\nabla\varphi^3_\mu) - \boldsymbol{\nabla}_{\nabla\varphi^1_\mu}(\boldsymbol{\nabla}_{\nabla\varphi^2_\mu}\nabla\varphi^3_\mu) + \boldsymbol{\nabla}_{[\nabla\varphi^1_\mu, \nabla\varphi^2_\mu]}\nabla\varphi^3_\mu,$$

where the objects like $\boldsymbol{\nabla}_{\nabla\varphi_\mu}(\nabla\psi_\mu)$, are, heuristically speaking, the covariant derivative of the vector field $\mu \mapsto \nabla\psi_\mu$ along the vector field $\mu \mapsto \nabla\varphi_\mu$.

However, in order to give a precise meaning to the above formula, we should be sure, at least, that the derivatives we are taking exist. Such an approach is possible, but heavy: indeed, consider that we should define what are C^1 and C^2 vector fields, and in doing so we cannot just consider derivatives along curves. Indeed we would need to be sure that "the partial derivatives have the right symmetries", otherwise there won't be those cancellations which let the above operator be a tensor.

Instead, we adopt the following strategy:

- First we calculate the curvature tensor for some very specific kind of vector fields, for which we are able to do and justify the calculations. Specifically, we will consider vector fields of the kind $\mu \mapsto \nabla\varphi$, where the function $\varphi \in C^\infty_c(M)$ does not depend on the measure μ.
- Then we prove that the object found is actually a tensor, i.e. that its value depends only on the μ−a.e. value of the considered vector fields, and not on the fact that we obtained the formula assuming that the functions φ's were independent on the measure.
- Finally, we discuss the minimal regularity requirements for the object found to be well defined.

Pick $\varphi, \psi \in C^\infty_c(\mathbb{R}^d)$ and observe that a curve of the kind $t \mapsto (Id + t\nabla\varphi)_\#\mu$ is a regular geodesic on an interval $[-T, T]$ for T sufficiently small (Remark 2.22 and Proposition 7.3). It is then immediate to verify that a vector field of the kind $(\nabla\psi)$ along it is C^∞. Its covariant derivative calculated at $t = 0$ is given by $\mathrm{P}_\mu(\nabla^2\psi \cdot \nabla\varphi)$. Thus we write:

$$\boldsymbol{\nabla}_{\nabla\varphi}\nabla\psi := \mathrm{P}_\mu(\nabla^2\psi \cdot \nabla\varphi) \qquad \forall \varphi, \psi \in C^\infty_c(\mathbb{R}^d). \tag{121}$$

Proposition 7.25. *Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ and $\varphi_1, \varphi_2, \varphi_3 \in C^\infty_c(\mathbb{R}^d)$. The curvature tensor $\mathbf{R}$ in μ calculated for the 3 vector fields $\nabla\varphi_i$, $i = 1, 2, 3$ is given by*

$$\begin{aligned}\mathbf{R}(\nabla\varphi_1, \nabla\varphi_2)\nabla\varphi_3 =& \mathrm{P}_\mu\Bigg(\mathscr{O}^*_{\nabla\varphi_2}\big(\mathscr{N}_\mu(\nabla\varphi_1, \nabla\varphi_3)\big) \\ & - \mathscr{O}^*_{\nabla\varphi_1}\big(\mathscr{N}_\mu(\nabla\varphi_2, \nabla\varphi_3)\big) + 2\mathscr{O}^*_{\nabla\varphi_3}\big(\mathscr{N}_\mu(\nabla\varphi_1, \nabla\varphi_2)\big)\Bigg).\end{aligned} \tag{122}$$

Proof. We start computing the value of $\boldsymbol{\nabla}_{\nabla\varphi_2}\boldsymbol{\nabla}_{\nabla\varphi_1}\nabla\varphi_3$. Let $\mu_t := (Id + t\nabla\varphi_2)_\#\mu$ and observe, as just recalled, that (μ_t) is a regular geodesic in some symmetric interval $[-T, T]$. The vector field $\nabla^2\varphi_3 \cdot \nabla\varphi_1$ is clearly C^∞ along it, thus by Proposition 7.24 also the vector field $u_t := \mathrm{P}_{\mu_t}(\nabla^2\varphi_3 \cdot \nabla\varphi_1) = \boldsymbol{\nabla}_{\nabla\varphi_1}\nabla\varphi_3(\mu_t)$ is C^∞. The covariant derivative at $t = 0$ of (u_t) along (μ_t) is, by definition, the value of $\boldsymbol{\nabla}_{\nabla\varphi_2}\boldsymbol{\nabla}_{\nabla\varphi_1}\nabla\varphi_3$ at μ. Applying formula (115) we get

$$\boldsymbol{\nabla}_{\nabla\varphi_2}\boldsymbol{\nabla}_{\nabla\varphi_1}\nabla\varphi_3 = \mathrm{P}_\mu\left(\nabla(\nabla^2\varphi_3 \cdot \nabla\varphi_1) \cdot \nabla\varphi_2 - \nabla^2\varphi_2 \cdot \mathrm{P}^\perp_\mu(\nabla^2\varphi_3 \cdot \nabla\varphi_1)\right). \tag{123}$$

Symmetrically, it holds

$$\boldsymbol{\nabla}_{\nabla\varphi_1}\boldsymbol{\nabla}_{\nabla\varphi_2}\nabla\varphi_3 = \mathrm{P}_\mu\left(\nabla(\nabla^2\varphi_3 \cdot \nabla\varphi_2) \cdot \nabla\varphi_1 - \nabla^2\varphi_1 \cdot \mathrm{P}^\perp_\mu(\nabla^2\varphi_3 \cdot \nabla\varphi_2)\right). \tag{124}$$

Finally, from the torsion free identity (100) we have

$$[\nabla\varphi_1, \nabla\varphi_2] = \mathrm{P}_\mu(\nabla^2\varphi_1 \cdot \nabla\varphi_2 - \nabla^2\varphi_2 \cdot \nabla\varphi_1),$$

and thus

$$\boldsymbol{\nabla}_{[\nabla\varphi_1,\nabla\varphi_2]}\nabla\varphi_3 = \mathrm{P}_\mu\Big(\nabla^2\varphi_3 \cdot \Big(\mathrm{P}_\mu(\nabla^2\varphi_1 \cdot \nabla\varphi_2 - \nabla^2\varphi_2 \cdot \nabla\varphi_1)\Big)\Big). \tag{125}$$

Subtracting (125) and (124) from (123) and observing that

$$\nabla(\nabla^2\varphi_3\cdot\nabla\varphi_1)\cdot\nabla\varphi_2 - \nabla(\nabla^2\varphi_3 \cdot \nabla\varphi_2) \cdot \nabla\varphi_1 = \nabla^2\varphi_3\cdot\nabla^2\varphi_1\cdot\nabla\varphi_2 - \nabla^2\varphi_3 \cdot \nabla^2\varphi_2\cdot\nabla\varphi_1,$$

we get the thesis. □

Observe that (122) is equivalent to

$$\begin{aligned}\langle\mathbf{R}(\nabla\varphi_1, \nabla\varphi_2)\nabla\varphi_3, \nabla\varphi_4\rangle_\mu = &\big\langle\mathcal{N}_\mu(\nabla\varphi_1, \nabla\varphi_3), \mathcal{N}_\mu(\nabla\varphi_2, \nabla\varphi_4)\big\rangle_\mu \\ &- \big\langle\mathcal{N}_\mu(\nabla\varphi_2, \nabla\varphi_3), \mathcal{N}_\mu(\nabla\varphi_1, \nabla\varphi_4)\big\rangle_\mu \\ &+ 2\big\langle\mathcal{N}_\mu(\nabla\varphi_1, \nabla\varphi_2), \mathcal{N}_\mu(\nabla\varphi_3, \nabla\varphi_4)\big\rangle_\mu,\end{aligned} \tag{126}$$

for any $\varphi_4 \in C^\infty_c(M)$. From this formula it follows immediately that the operator $\mathbf{R}$ is actually a tensor:

Proposition 7.26. *Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. The curvature operator, given by formula* (126)*, is a tensor on* $[\{\nabla\varphi\}]^4$*, i.e. its value depends only on the μ–a.e. value of the four vector fields.*

Proof. Clearly the left hand side of (126) is a tensor w.r.t. the fourth entry. The conclusion follows from the symmetries of the right hand side. □

We remark that from (126) it follows that $\mathbf{R}$ has all the expected symmetries.

Concerning the domain of definition of the curvature tensor, the following statement holds, whose proof follows from the properties of the normal tensor $\mathscr{N}_\mu$:

Proposition 7.27. *Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$. Then the curvature tensor, thought as map from $[\{\nabla\varphi\}]^4$ to $\mathbb{R}$ given by (126), extends uniquely to a sequentially continuous map on the set of 4-ples of vector fields in L^2_μ in which at least three vector fields are Lipschitz, where we say that $(v^1_n, v^2_n, v^3_n, v^4_n)$ is converging to (v^1, v^2, v^3, v^4) if there is convergence in L^2_μ on each coordinate and*

$$\sup_n \mathrm{Lip}(v^i_n) < \infty,$$

for at least three indexes i.

Thus, in order for the curvature tensor to be well defined we need at least three of the four vector fields involved to be Lipschitz. However, for some related notion of curvature the situation simplifies. Of particular relevance is the case of sectional curvature:

Example 7.28 (The sectional curvature). If we evaluate the curvature tensor $\mathbf{R}$ on a 4-ple of vectors of the kind (u, v, u, v) and we recall the antisymmetry of $\mathscr{N}_\mu$ we obtain

$$\langle \mathbf{R}(u,v)u, v\rangle_\mu = 3 \left\| \mathscr{N}_\mu(u,v) \right\|^2_\mu .$$

Thanks to the simplification of the formula, the value of $\langle \mathbf{R}(u,v)u, v\rangle_\mu$ is well defined as soon as either u or v is Lipschitz. That is, $\langle \mathbf{R}(u,v)u, v\rangle_\mu$ is well defined for $(u,v) \in \mathrm{L{\scriptstyle N}L}_\mu$. In analogy with the Riemannian case we can therefore define the sectional curvature $\mathbf{K}(u,v)$ at the measure μ along the directions u, v by

$$\mathbf{K}(u,v) := \frac{\langle \mathbf{R}(u,v)u, v\rangle_\mu}{\|u\|^2_\mu \|v\|^2_\mu - \langle u, v\rangle^2_\mu} = \frac{3 \left\| \mathscr{N}_\mu(u,v) \right\|^2_\mu}{\|u\|^2_\mu \|v\|^2_\mu - \langle u, v\rangle^2_\mu}, \qquad \forall (u,v) \in \mathrm{L{\scriptstyle N}L}_\mu .$$

This expression confirms the fact that the sectional curvatures of $\mathscr{P}_2(\mathbb{R}^d)$ are positive (coherently with Theorem 3.20), and provides a rigorous proof of the analogous formula already appeared in [67] and formally computed using O'Neill formula. ■

7.4 Bibliographical Notes

The idea of looking at the Wasserstein space as a sort of infinite dimensional Riemannian manifold is due to F. Otto and given in his seminal paper [67]. The whole discussion in Sect. 7.1 is directly taken from there.

The fact that the "tangent space made of gradients" $\mathrm{Tan}_\mu(\mathscr{P}_2(\mu)\mathbb{R}^d)$ was not sufficient to study all the aspects of the "Riemannian geometry" of $(\mathscr{P}_2(\mathbb{R}^d), W_2)$

has been understood in [7] in connection with the definition of subdifferential of a geodesically convex functional, in particular concerning the issue of having a closed subdifferential. In the appendix of [7] the concept of Geometric Tangent space discussed in Sect. 7.2 has been introduced. Further studies on the properties of $\mathbf{Tan}_\mu(\mathscr{P}_2(\mu)M)$ have been made in [43]. Theorem 7.1 has been proved in [46].

The first work in which a description of the covariant derivative and the curvature tensor of $(\mathscr{P}_2(M), W_2)$, M being a compact Riemannian manifold has been given (beside the formal calculus of the sectional curvature via O'Neill formula done already in [67]) is the paper of J. Lott [56]: rigorous formulas are derived for the computation of such objects on the "submanifold" $\mathscr{P}_{C^\infty}(M)$ made of absolutely continuous measures with density C^∞ and bounded away from 0. In the same paper Lott shows that if M has a Poisson structure, then the same is true for $\mathscr{P}_{C^\infty}(M)$ (a topic which has not been addressed in these notes).

Independently on Lott's work, the second author built the parallel transport on $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ in his PhD thesis [43], along the same lines provided in Sect. 7.3. The differences with Lott's work are the fact that the analysis was carried out on $\mathbb{R}^d$ rather than on a compact Riemannian manifold, that no assumptions on the measures were given, and that both the existence Theorem 7.15 for the parallel transport along a regular curve and counterexamples to its general existence (the Example 7.16) were provided. These results have been published by the authors of these notes in [5]. Later on, after having been aware of Lott's results, the second author generalized the construction to the case of Wasserstein space built over a manifold in [44]. Not all the results have been reported here: we mention that it is possible to push the analysis up show the differentiability properties of the exponential map and the existence of Jacobi fields.

8 Ricci Curvature Bounds

Let us start recalling what is the Ricci curvature for a Riemannian manifold M (which we will always consider smooth and complete). Let R be the Riemann curvature tensor on M, $x \in M$ and $u, v \in T_xM$. Then the Ricci curvature $\mathrm{Ric}(u, v) \in \mathbb{R}$ is defined as

$$\mathrm{Ric}(u, v) := \sum_i \langle R(u, e_i)v, e_i \rangle ,$$

where $\{e_i\}$ is any orthonormal basis of T_xM. An immediate consequence of the definition and the symmetries of R is the fact that $\mathrm{Ric}(u, v) = \mathrm{Ric}(v, u)$.

Another, more geometric, characterization of the Ricci curvature is the following. Pick $x \in M$, a small ball B around the origin in T_xM and let μ be the Lebesgue measure on B. The exponential map $\exp_x : B \to M$ is injective and smooth, thus the measure $(\exp_x)_\#\mu$ has a smooth density w.r.t. the volume measure Vol on M. For any $u \in B$, let $f(u)$ be the density of $(\exp_x)_\#\mu$ w.r.t. Vol at the point $\exp_x(u)$.

Then the function f has the following Taylor expansion:

$$f(u) = 1 + \frac{1}{2}\mathrm{Ric}(u,u) + o(|u|^2). \tag{127}$$

It is said that the Ricci curvature is bounded below by $\lambda \in \mathbb{R}$ provided

$$\mathrm{Ric}(u,u) \geq \lambda |u|^2,$$

for every $x \in M$ and $u \in T_x M$.

Several important geometric and analytic inequalities are related to bounds from below on Ricci curvature, we mention just two of them.

- *Brunn–Minkowski*. Suppose that M has non negative Ricci curvature, and for any $A_0, A_1 \subset M$ compact, let

  $$A_t := \Big\{\gamma_t \ : \ \gamma \text{ is a constant speed geodesic s.t. } \gamma_0 \in A_0,\ \gamma_1 \in A_1\Big\}, \qquad \forall t \in [0,1].$$

  Then it holds

  $$\big(\mathrm{Vol}(A_t)\big)^{1/n} \geq (1-t)\big(\mathrm{Vol}(A_0)\big)^{1/n} + t\big(\mathrm{Vol}(A_1)\big)^{1/n}, \qquad \forall t \in [0,1], \tag{128}$$

  where n is the dimension of M.
- *Bishop-Gromov*. Suppose that M has Ricci curvature bounded from below by $(n-1)k$, where n is the dimension of M and k a real number. Let $\tilde{M}$ be the simply connected, n-dimensional space with constant curvature, having Ricci curvature equal to $(n-1)k$ (so that $\tilde{M}$ is a sphere if $k > 0$, a Euclidean space if $k = 0$ and an hyperbolic space if $k < 0$). Then for every $x \in M$ and $\tilde{x} \in \tilde{M}$ the map

  $$(0,\infty) \ni r \quad \mapsto \quad \frac{\mathrm{Vol}(B_r(x))}{\widetilde{\mathrm{Vol}}(B_r(\tilde{x}))}, \tag{129}$$

  is non increasing, where Vol and $\widetilde{\mathrm{Vol}}$ are the volume measures on M, $\tilde{M}$ respectively.

A natural question is whether it is possible to formulate the notion of Ricci bound from below also for metric spaces, analogously to the definition of Alexandrov spaces, which are a metric analogous of Riemannian manifolds with bounded (either from above or from below) sectional curvature. What became clear over time, is that the correct non-smooth object where one could try to give a notion of Ricci curvature bound is not a metric space, but rather a metric *measure* space, i.e. a metric space where a reference non negative measure is also given. When looking to the Riemannian case, this fact is somehow hidden, as a natural reference measure is given by the volume measure, which is a function of the distance.

There are several viewpoints from which one can see the necessity of a reference measure (which can certainly be the Hausdorff measure of appropriate dimension, if available). A first (cheap) one is the fact that in most of identities/inequalities where the Ricci curvature appears, also the reference measures appears (e.g. (127)–(129) above). A more subtle point of view comes from studying stability issues: consider a sequence (M_n, g_n) of Riemannian manifolds and assume that it converges to a smooth Riemannian manifold (M, g) in the Gromov–Hausdorff sense. Assume that the Ricci curvature of (M_n, g_n) is uniformly bounded below by some $K \in R$. Can we deduce that the Ricci curvature of (M, g) is bounded below by K? The answer is *no* (while the same question with sectional curvature in place of Ricci one has affirmative answer). It is possible to see that when Ricci bounds are not preserved in the limiting process, it happens that the volume measures of the approximating manifolds are not converging to the volume measure of the limit one.

Another important fact to keep in mind is the following: if we want to derive useful analytic/geometric consequences from a weak definition of Ricci curvature bound, we should also known what is the dimension of the metric measure space we are working with: consider for instance the Brunn–Minkowski and the Bishop–Gromov inequalities above, both make sense if we know the dimension of M, and not just that its Ricci curvature is bounded from below. This tells that the natural notion of bound on the Ricci curvature should be a notion speaking both about the *curvature* and the *dimension* of the space. Such a notion exists and is called $CD(K, N)$ condition, K being the bound from below on the Ricci curvature, and N the bound from above on the dimension. Let us tell in advance that we will focus only on two particular cases: the curvature dimension condition $CD(K, \infty)$, where no upper bound on the dimension is specified, and the curvature-dimension condition $CD(0, N)$, where the Ricci curvature is bounded below by 0. Indeed, the general case is much more complicated and there are still some delicate issues to solve before we can say that the theory is complete and fully satisfactory.

Before giving the definition, let us highlight which are the qualitative properties that we expect from a weak notion of curvature-dimension bound:

Intrinsicness. The definition is based only on the property of the space itself, that is, is not something like "if the space is the limit of smooth spaces...."

Compatibility. If the metric-measure space is a Riemannian manifold equipped with the volume measure, then the bound provided by the abstract definition coincides with the lower bound on the Ricci curvature of the manifold, equipped with the Riemannian distance and the volume measure.

Stability. Curvature bounds are stable w.r.t. the natural passage to the limit of the objects which define it.

Interest. Geometrical and analytical consequences on the space can be derived from curvature-dimension condition.

In the next section we recall some basic concepts concerning convergence of metric measure spaces (which are key to discuss the stability issue), while in the following one we give the definition of curvature-dimension condition and analyze its properties.

All the metric measure spaces (X, d, m) that we will consider satisfy the following assumption:

Assumption 8.1. *(X, d) is Polish, the measure m is a Borel probability measure and $m \in \mathscr{P}_2(X)$.*

8.1 Convergence of Metric Measure Spaces

We say that two metric measure spaces (X, d_X, m_X) and (Y, d_Y, m_Y) are *isomorphic* provided there exists a bijective isometry $f : \text{supp}(m_X) \to \text{supp}(m_Y)$ such that $f_\# m_X = m_Y$. This is the same as to say that "we don't care about the behavior of the space (X, d_X) where there is no mass". This choice will be important in discussing the stability issue.

Definition 8.2 (Coupling between metric measure spaces). Given two metric measure spaces (X, d_X, m_X), (Y, d_Y, m_Y), we consider the product space $(X \times Y, D_{XY})$, where D_{XY} is the distance defined by

$$D_{XY}\big((x_1, y_1), (x_2, y_2)\big) := \sqrt{d_X^2(x_1, x_2) + d_Y^2(y_1, y_2)}.$$

We say that a couple (d, γ) is an admissible coupling between (X, d_X, m_X) and (Y, d_Y, m_Y), we write $(d, \gamma) \in \mathcal{A}dm((d_X, m_X), (d_Y, m_Y))$ if:

- d is a pseudo distance on $\text{supp}\, m_X \sqcup \text{supp}\, m_Y$ (i.e. it may be zero on two different points) which coincides with d_X (resp. d_Y) when restricted to $\text{supp}\, m_X \times \text{supp}\, m_X$ (resp. $\text{supp}\, m_Y \times \text{supp}\, m_Y$).
- A Borel (w.r.t. the Polish structure given by D_{XY}) measure γ on $\text{supp}\, m_X \times \text{supp}\, m_Y$ such that $\pi^1_\# \gamma = m_X$ and $\pi^2_\# \gamma = m_Y$.

It is not hard to see that the set of admissible couplings is always non empty.

The *cost* $C(d, \gamma)$ of a coupling is given by

$$C(d, \gamma) := \int_{\text{supp ú}m_X \times \text{supp ú}m_Y} d^2(x, y)\, d\gamma(x, y).$$

The distance $\mathbb{D}\big((X, d_X, m_X), (Y, d_Y, m_Y)\big)$ is then defined as

$$\mathbb{D}\big((X, d_X, m_X), (Y, d_Y, m_Y)\big) := \inf \sqrt{C(d, \gamma)}, \tag{130}$$

the infimum being taken among all couplings (d, γ) of (X, d_X, m_X) and (Y, d_Y, m_Y).

A trivial consequence of the definition is that if (X, d_X, m_X) and $(\tilde{X}, d_{\tilde{X}}, m_{\tilde{X}})$ (resp. (Y, d_Y, m_Y) and $(\tilde{Y}, d_{\tilde{Y}}, m_{\tilde{Y}})$) are isomorphic, then

$$\mathbb{D}\Big((X, d_X, m_X), (Y, d_Y, m_Y)\Big) = \mathbb{D}\Big((\tilde{X}, d_{\tilde{X}}, m_{\tilde{X}}), (\tilde{Y}, d_{\tilde{Y}}, m_{\tilde{Y}})\Big),$$

so that $\mathbb{D}$ is actually defined on isomorphism classes of metric measure spaces.

In the next proposition we collect, without proof, the main properties of $\mathbb{D}$.

Proposition 8.3 (Properties of $\mathbb{D}$). *The* inf *in* (130) *is realized, and a coupling realizing it will be called* optimal.

Also, let $\mathbb{X}$ be the set of isomorphism classes of metric measure spaces satisfying Assumption 8.1. Then $\mathbb{D}$ is a distance on $\mathbb{X}$, and in particular $\mathbb{D}$ is 0 only on couples of isomorphic metric measure spaces.

Finally, the space $(\mathbb{X}, \mathbb{D})$ is complete, separable and geodesic.

Proof. See Sect. 3.1 of [74]. □

We will denote by $\mathcal{O}pt((d_X, m_X), (d_Y, m_Y))$ the set of optimal couplings between (X, d_X, m_X) and (Y, d_Y, m_Y), i.e. the set of couplings where the inf in (130) is realized.

Given a metric measure space (X, d, m) we will denote by $\mathscr{P}_2^a(X) \subset \mathscr{P}(X)$ the set of measures which are absolutely continuous w.r.t. m.

To any coupling (d, γ) of two metric measure spaces (X, d_X, m_X) and (Y, d_Y, m_Y), it is naturally associated a map $\gamma_\# : \mathscr{P}_2^a(X) \to \mathscr{P}_2^a(Y)$ defined as follows:

$$\mu = \rho m_X \quad \mapsto \quad \gamma_\#\mu := \eta m_Y, \quad \text{where } \eta \text{ is defined by } \quad \eta(y) := \int \rho(x) d\gamma_y(x), \tag{131}$$

where $\{\gamma_y\}$ is the disintegration of γ w.r.t. the projection on Y. Similarly, there is a natural map $\gamma_\#^{-1} : \mathscr{P}_2^a(Y) \to \mathscr{P}_2^a(X)$ given by:

$$\nu = \eta m_Y \quad \mapsto \quad \gamma_\#^{-1}\nu := \rho m_X, \quad \text{where } \rho \text{ is defined by } \quad \rho(x) := \int \eta(y) d\gamma_x(y),$$

where, obviously, $\{\gamma_x\}$ is the disintegration of γ w.r.t. the projection on X.

Notice that $\gamma_\# m_X = m_Y$ and $\gamma_\#^{-1} m_Y = m_X$ and that in general $\gamma_\#^{-1}\gamma_\#\mu \neq \mu$. Also, if γ is induced by a map $T : X \to Y$, i.e. if $\gamma = (Id, T)_\# m_X$, then $\gamma_\#\mu = T_\#\mu$ for any $\mu \in \mathscr{P}_2^a(X)$.

Our goal now is to show that if $(X_n, d_n, m_n) \overset{\mathbb{D}}{\to} (X, d, m)$ of the *internal energy* kind on $(\mathscr{P}_2^a(X_n), W_2)$ Mosco-converge to the corresponding functional on $(\mathscr{P}_2^a(X), W_2)$. Thus, fix a convex and continuous function $u : [0, +\infty) \to \mathbb{R}$, define

$$u'(\infty) := \lim_{z \to +\infty} \frac{u(z)}{z},$$

and, for every compact metric space (X, d), define the functional $\mathscr{E} : [\mathscr{P}(X)]^2 \to \mathbb{R} \cup \{+\infty\}$ by

$$\mathscr{E}(\mu|\nu) := \int u(\rho) d\nu + u'(\infty)\mu^s(X), \tag{132}$$

where $\mu = \rho\nu + \mu^s$ is the decomposition of μ in absolutely continuous $\rho\nu$ and singular part μ^s w.r.t. to ν.

Lemma 8.4 (**$\mathscr{E}$ decreases under $\gamma_\#$**). *Let (X, d_X, m_X) and (Y, d_Y, m_Y) be two metric measure space and (d, γ) a coupling between them. Then it holds*

$$\begin{aligned} \mathscr{E}(\gamma_\#\mu|m_Y) &\leq \mathscr{E}(\mu|m_X), \qquad \forall \mu \in \mathscr{P}_2^a(X), \\ \mathscr{E}(\gamma_\#^{-1}\nu|m_X) &\leq \mathscr{E}(\nu|m_Y), \qquad \forall \nu \in \mathscr{P}_2^a(Y). \end{aligned}$$

Proof. Clearly it is sufficient to prove the first inequality. Let $\mu = \rho m_X$ and $\gamma_\#\mu = \eta m_Y$, with η given by (131). By Jensen's inequality we have

$$\begin{aligned} \mathscr{E}(\gamma_\#\mu|m_Y) &= \int u(\eta(y))dm_Y(y) = \int u\left(\int \rho(x)d\gamma_y(x)\right)dm_Y(y) \\ &\leq \int\int u(\rho(x))d\gamma_y(x)dm_Y(y) = \int u(\rho(x))d\gamma(x,y) \\ &= \int u(\rho(x))dm_X(x) = \mathscr{E}(\mu|m_X) \end{aligned}$$

□

Proposition 8.5 ("Mosco" convergence of internal energy functionals). *Let $(X_n, d_n, m_n) \overset{\mathbb{D}}{\to} (X, d, m)$ and $(d_n, \gamma_n) \in \mathcal{O}pt((d_n, m_n), (d, m))$. Then the following two are true:*

Weak $-\underline{\lim}$. *For any sequence $n \mapsto \mu_n \in \mathscr{P}_2^a(X_n)$ such that $n \mapsto (\gamma_n)_\#\mu_n$ narrowly converges to some $\mu \in \mathscr{P}(X)$ it holds*

$$\underline{\lim}_{n\to\infty} \mathscr{E}(\mu_n|m_n) \geq \mathscr{E}(\mu|m).$$

Strong $-\overline{\lim}$. *For any $\mu \in \mathscr{P}_2^a(X)$ with bounded density there exists a sequence $n \mapsto \mu_n \in \mathscr{P}_2^a(X_n)$ such that $W_2((\gamma_n)_\#\mu_n, \mu) \to 0$ and*

$$\overline{\lim_{n\to\infty}} \mathscr{E}(\mu_n|m_n) \leq \mathscr{E}(\mu|m).$$

Note: we put the apexes in *Mosco* because we prove the $\Gamma - \overline{\lim}$ inequality only for measures with bounded densities. This will be enough to prove the stability of Ricci curvature bounds (see Theorem 8.12).

Proof. For the first statement we just notice that by Lemma 8.4 we have

$$\mathscr{E}(\mu_n|m_n) \geq \mathscr{E}((\gamma_n)_\#\mu_n|m),$$

and the conclusion follows from the narrow lower semicontinuity of $\mathscr{E}(\cdot|m)$.

For the second one we define $\mu_n := (\gamma_n^{-1})_\#\mu$. Then applying Lemma 8.4 twice we get

$$\mathscr{E}(\mu|m) \geq \mathscr{E}(\mu_n|m_n) \geq \mathscr{E}((\gamma_n)_\#\mu_n|m),$$

from which the $\Gamma - \overline{\lim}$ inequality follows. Thus to conclude we need to show that $W_2((\gamma_n)_\# \mu_n, \mu) \to 0$. To check this, we use the Wassertein space built over the (pseudo-)metric space $(X_n \sqcup X, d_n)$: let $\mu = \rho m_X$ and for any $n \in \mathbb{N}$ define the plan $\tilde{\gamma}_n \in \mathscr{P}(X_n \times X)$ by $d\tilde{\gamma}_n(y,x) := \rho(x) d\gamma_n(y,x)$ and notice that $\tilde{\gamma}_n \in \mathcal{A}dm(\mu_n, \mu)$. Thus

$$W_2(\mu_n, \mu) \le \sqrt{\int d_n^2(x,y) d\tilde{\gamma}_n(y,x)} \le \sqrt{\int d_n^2(x,y)\rho(x) d\gamma_n(y,x)} \le \sqrt{M}\sqrt{C(d_n, \gamma_n)},$$

where M is the essential supremum of ρ. By definition, it is immediate to check that the density η_n of μ_n is also bounded above by M. Introduce the plan $\overline{\gamma}_n$ by $d\overline{\gamma}_n(y,x) := \eta_n(y) d\gamma_n(y,x)$ and notice that $\overline{\gamma}_n \in \mathcal{A}dm(\mu_n, (\gamma_n)_\# \mu_n)$, so that, as before, we have

$$\begin{aligned} W_2(\mu_n, (\gamma_n)_\# \mu_n) &\le \sqrt{\int d_n^2(x,y) d\overline{\gamma}_n(y,x)} \\ &\le \sqrt{\int d_n^2(x,y) \eta_n(y) d\gamma_n(y,x)} \le \sqrt{M}\sqrt{C(d_n, \gamma_n)}. \end{aligned}$$

In conclusion we have

$$W_2(\mu, (\gamma_n)_\# \mu_n) \le W_2(\mu_n, (\gamma_n)_\# \mu_n) + W_2(\mu_n, \mu) \le 2\sqrt{M}\sqrt{C(d_n, \gamma_n)},$$

which gives the thesis. □

8.2 *Weak Ricci Curvature Bounds: Definition and Properties*

Define the functions u_N, $N > 1$, and u_∞ on $[0, +\infty)$ as

$$u_N(z) := N(z - z^{1-\frac{1}{N}}),$$

and

$$u_\infty(z) := z \log(z).$$

Then given a metric measure space (X, d, m) we define the functionals $\mathscr{E}_N, \mathscr{E}_\infty : \mathscr{P}(X) \to \mathbb{R} \cup \{+\infty\}$ by

$$\mathscr{E}_N(\mu) := \mathscr{E}(\mu|m),$$

where $\mathscr{E}(\cdot|\cdot)$ is given by formula (132) with $u := u_N$; similarly for $\mathscr{E}_\infty$.

The definitions of weak Ricci curvature bounds are the following:

Definition 8.6 (Curvature $\geq K$ and no bound on dimension—$CD(K,\infty)$). We say that a metric measure space (X,d,m) has Ricci curvature bounded from below by $K\in\mathbb{R}$ provided the functional

$$\mathscr{E}_\infty : \mathscr{P}(X)\to\mathbb{R}\cup\{+\infty\},$$

is K-geodesically convex on $(\mathscr{P}_2^a(X),W_2)$. In this case we say that (X,d,m) satisfies the curvature dimension condition $CD(K,\infty)$ or that (X,d,m) is a $CD(K,\infty)$ space.

Definition 8.7 (Curvature ≥ 0 and dimension $\leq N$ - $CD(0,N)$). We say that a metric measure space (X,d,m) has nonnegative Ricci curvature and dimension bounded from above by N provided the functionals

$$\mathscr{E}_{N'} : \mathscr{P}(X)\to\mathbb{R}\cup\{+\infty\},$$

are geodesically convex on $(\mathscr{P}_2^a(X),W_2)$ for every $N'\geq N$. In this case we say that (X,d,m) satisfies the curvature dimension condition $CD(0,N)$, or that (X,d,m) is a $CD(0,N)$ space.

Note that $N>1$ is not necessarily an integer.

Remark 8.8. Notice that geodesic convexity is required on $\mathscr{P}_2(\text{supp}(m_X))$ and not on $\mathscr{P}_2(X)$. This makes no difference for what concerns $CD(K,\infty)$ spaces, as $\mathscr{E}_\infty$ is $+\infty$ on measures having a singular part w.r.t. m, but is important for the case of $CD(0,N)$ spaces, as the functional $\mathscr{E}_N$ has only real values, and requiring geodesic convexity on the whole $\mathscr{P}_2(X)$ would lead to a notion not invariant under isomorphism of metric measure spaces.

Also, for the $CD(0,N)$ condition one requires the geodesic convexity of all $\mathscr{E}_{N'}$ to ensure the following compatibility condition: if X is a $CD(0,N)$ space, then it is also a $CD(0,N')$ space for any $N'>N$. Using Proposition 3.16 it is not hard to see that such compatibility condition is automatically satisfied on non branching spaces. ■

Remark 8.9 (How to adapt the definitions to general bounds on curvature the dimension).

It is pretty natural to guess that the notion of bound from below on the Ricci curvature by $K\in\mathbb{R}$ and bound from above on the dimension by N can be given by requiring the functional $\mathscr{E}_N$ to be K-geodesically convex on $(\mathscr{P}(X),W_2)$. However, this is *wrong*, because such condition is not compatible with the Riemannian case. The hearth of the definition of $CD(K,N)$ spaces still concerns the properties of $\mathscr{E}_N$, but a different and more complicated notion of "convexity" is involved. ■

Let us now check that the definitions given have the qualitative properties that we discussed in the introduction of this chapter.

Intrinsicness. This property is clear from the definition.

Compatibility. To give the answer we need to do some computations on Riemannian manifolds:

Lemma 8.10 (Second derivative of the internal energy). *Let M be a compact and smooth Riemannian manifold, m its normalized volume measure, $u : [0,+\infty)$ be convex, continuous and C^2 on $(0,+\infty)$ with $u(0) = 0$ and define the "pressure" $p : [0,+\infty) \to \mathbb{R}$ by*

$$p(z) := zu'(z) - u(z), \qquad \forall z > 0,$$

and $p(0) := 0$. Also, let $\mu = \rho m \in \mathscr{P}_2^a(M)$ with $\rho \in C^\infty(M)$, pick $\varphi \in C_c^\infty(M)$, and define $T_t : M \to M$ by $T_t(x) := \exp_x(t\nabla\varphi(x))$. Then it holds:

$$\frac{d^2}{dt^2}|_{t=0}\mathscr{E}((T_t)_\#\mu) = \int p'(\rho)\,\rho\,(\Delta\varphi)^2 - p(\rho)\Big((\Delta\varphi)^2 - \big|\nabla^2\varphi\big|^2 - \mathrm{Ric}\big(\nabla\varphi,\nabla\varphi\big)\Big)\,dm,$$

where by $\big|\nabla^2\varphi(x)\big|^2$ we mean the trace of the linear map $(\nabla^2\varphi(x))^2 : T_xM \to T_xM$ (in coordinates, this reads as $\sum_{ij}(\partial_{ij}\varphi(x))^2$).

Proof (Computation of the second derivative). Let $D_t(x) := \det(\nabla T_t(x))$, $\mu_t := (T_t)_\#\mu = \rho_t\mathrm{Vol}$. By compactness, for t sufficiently small T_t is invertible with smooth inverse, so that $D_t, \rho_t \in C^\infty(M)$. For small t, the change of variable formula gives

$$\rho_t(T_t(x)) = \frac{\rho(x)}{\det(\nabla T_t(x))} = \frac{\rho(x)}{D_t(x)}.$$

Thus we have (all the integrals being w.r.t. m):

$$\frac{d}{dt}\int u(\rho_t) = \frac{d}{dt}\int u\left(\frac{\rho}{D_t}\right)D_t = \int -u'\left(\frac{\rho}{D_t}\right)\frac{\rho D_t'}{D_t^2}D_t + u\left(\frac{\rho}{D_t}\right)D_t' = -\int p\left(\frac{\rho}{D_t}\right)D_t',$$

and

$$\frac{d^2}{dt^2}|_{t=0}\int u(\rho_t) = -\frac{d}{dt}|_{t=0}\int p\left(\frac{\rho}{D_t}\right)D_t' = \int p'(\rho)\rho(D_0')^2 - p(\rho)D_0'',$$

having used the fact that $D_0 \equiv 1$.

(Evaluation of D_0' and D_0''). We want to prove that

$$\begin{aligned} D_0'(x) &= \Delta\varphi(x), \\ D_0''(x) &= (\Delta\varphi(x))^2 - \big|\nabla^2\varphi(x)\big|^2 - \mathrm{Ric}\big(\nabla\varphi(x),\nabla\varphi(x)\big). \end{aligned} \tag{133}$$

For $t \geq 0$ and $x \in M$, let $J_t(x)$ be the operator from T_xM to $T_{\exp_x(t\nabla\varphi(x))}M$ given by:

$$J_t(x)(v) := \begin{cases} \text{the value at } s = t \text{ of the Jacobi field } j_s \text{ along the geodesic} \\ s \mapsto \exp_x(s\nabla\varphi(x)), \text{having the initial conditions } j_0 := v,\ j_0' := \nabla^2\varphi \cdot v, \end{cases}$$

(where here and in the following the apex' on a vector/tensor field stands for covariant differentiation), so that in particular we have

$$\begin{aligned} J_0 &= Id, \\ J_0' &= \nabla^2\varphi. \end{aligned} \tag{134}$$

The fact that Jacobi fields are the differential of the exponential map reads, in our case, as:

$$\nabla T_t(x)\cdot v = J_t(x)\cdot v,$$

therefore we have

$$D_t = \det(J_t). \tag{135}$$

Also, Jacobi fields satisfy the Jacobi equation, which we write as

$$J_t'' + A_t J_t = 0, \tag{136}$$

where $A_t(x) : T_{\exp_x(t\nabla\varphi(x))}M \to T_{\exp_x(t\nabla\varphi(x))}M$ is the map given by

$$A_t(x)\cdot v := R(\dot\gamma_t, v)\dot\gamma_t,$$

where $\gamma_t := \exp_x(t\nabla\varphi(x))$. Recalling the rule $(\det B_t)' = \det(B_t)\mathrm{tr}(B_t' B_t^{-1})$, valid for a smooth curve of linear operators, we obtain from (135) the validity of

$$D_t' = D_t \mathrm{tr}(J_t' J_t^{-1}). \tag{137}$$

Evaluating this identity at $t = 0$ and using (134) we get the first of (133). Recalling the rule $(B_t^{-1})' = -B_t^{-1}B_t'B_t^{-1}$, valid for a smooth curve of linear operators, and differentiating in time equation (137) we obtain

$$\begin{aligned} D_t'' &= D_t\big(\mathrm{tr}(J_t'J_t^{-1})\big)^2 + D_t\mathrm{tr}(J_t''J_t^{-1} - J_t'J_t^{-1}J_t'J_t^{-1}) \\ &= D_t\Big(\big(\mathrm{tr}(J_t'J_t^{-1})\big)^2 - \mathrm{tr}\big(A_t + J_t'J_t^{-1}J_t'J_t^{-1}\big)\Big), \end{aligned}$$

having used the Jacobi equation (136). Evaluate this expression at $t = 0$, use (134) and observe that

$$\mathrm{tr}(A_0) = \mathrm{tr}\Big\{v \mapsto R(\nabla\varphi, v)\nabla\varphi\Big\} = \mathrm{Ric}(\nabla\varphi, \nabla\varphi),$$

to get the second of (133). □

Theorem 8.11 (Compatibility of weak Ricci curvature bounds). *Let M be a compact Riemannian manifold, d its Riemannian distance and m its normalized volume measure. Then:*

(i) The functional $\mathscr{E}_\infty$ is K-geodesically convex on $(\mathscr{P}_2(M), W_2)$ if and only if M has Ricci curvature uniformly bounded from below by K.

(ii) The functional $\mathscr{E}_N$ is geodesically convex on $(\mathscr{P}_2(M), W_2)$ if and only if M has non negative Ricci curvature and $\dim(M) \leq N$.

Sketch of the Proof We will give only a formal proof, neglecting all the issues which arise due to the potential non regularity of the objects involved.

We start with (i). Assume that $\mathrm{Ric}(v, v) \geq K|v|^2$ for any v. Pick a geodesic $(\rho_t m) \subset \mathscr{P}_2(M)$ and assume that $\rho_t \in C^\infty$ for any $t \in [0, 1]$. By Theorem 2.33 we know that there exists a function $\varphi : M \to \mathbb{R}$ differentiable $\rho_0 m$-a.e. such that $\exp(\nabla\varphi)$ is the optimal transport map from $\rho_0 m$ to $\rho_1 m$ and

$$\rho_t m = \big(\exp(t\nabla\varphi)\big)_\# \rho_0 m.$$

Assume that φ is C^∞. Then by Lemma 8.10 with $u := u_\infty$ we know that

$$\frac{d^2}{dt^2}\mathscr{E}_\infty(\rho_t m) = \int \Big(\big|\nabla^2\varphi\big|^2 + \mathrm{Ric}(\nabla\varphi, \nabla\varphi)\Big)\rho_0 \, dm \geq K \int |\nabla\varphi|^2 \rho_0 \, dm.$$

Since $\int |\nabla\varphi|^2 \rho_0 dm = W_2^2(\rho_0, \rho_1)$, the claim is proved.

The converse implication follows by an explicit construction: if $\mathrm{Ric}(v, v) < K|v|^2$ for some $x \in M$ and $v \in T_x M$, then for $\varepsilon \ll \delta \ll 1$ define $\mu_0 := c_0 m|_{B_\varepsilon(x)}$ (c_0 being the normalizing constant) and $\mu_t := (T_t)_\# \mu_0$ where $T_t(y) := \exp_y(t\delta\nabla\varphi(y))$ and $\varphi \in C^\infty$ is such that $\nabla\varphi(x) = v$ and $\nabla^2\varphi(x) = 0$. Using Lemma 8.10 again and the hypothesis $\mathrm{Ric}(v, v) < K|v|^2$ it is not hard to prove that $\mathscr{E}_\infty$ is not λ-geodesically convex along (μ_t). We omit the details.

Now we turn to (ii). Let $(\rho_t m)$ and φ as in the first part of the argument above. Assume that M has non negative Ricci curvature and that $\dim(M) \leq N$. Observe that for $u := u_N$ Lemma 8.10 gives

$$\begin{aligned}&\frac{d^2}{dt^2}|_{t=0}\mathscr{E}_N(\rho_t)\\ &\quad = \int\!\!\left(1-\frac{1}{N}\right)\rho^{1-\frac{1}{N}}(\Delta\varphi)^2 - \rho^{1-\frac{1}{N}}\Big((\Delta\varphi)^2 - \big|\nabla^2\varphi\big|^2 - \frac{1}{2}\mathrm{Ric}(\nabla\varphi, \nabla\varphi)\Big)\, dm.\end{aligned}$$

Using the hypothesis on M and the fact that $(\Delta\varphi)^2 \leq N\big|\nabla^2\varphi\big|^2$ we get $\frac{d^2}{dt^2}|_{t=0}$ $\mathscr{E}_N(\rho_t) \geq 0$, i.e. the geodesic convexity of $\mathscr{E}_N$. For the converse implication it is possible to argue as above, we omit the details also in this case. □

Now we pass to the **stability**:

Theorem 8.12 (Stability of weak Ricci curvature bound). *Assume that $(X_n, d_n, m_n) \overset{\mathbb{D}}{\to} (X, d, m)$ and that for every $n \in \mathbb{N}$ the space (X_n, d_n, m_n) is $CD(K, \infty)$ (resp. $CD(0, N)$). Then (X, d, m) is a $CD(K, \infty)$ (resp. $CD(0, N)$) space as well.*

Sketch of the Proof Pick $\mu_0, \mu_1 \in \mathscr{P}_2^a(X)$ and assume they are both absolutely continuous with bounded densities, say $\mu_i = \rho_i m$, $i = 0, 1$. Choose $(\tilde{d}_n, \gamma_n) \in Opt((d_n, m_n), (d, m))$. Define $\mu_i^n := (\gamma_n^{-1})_\# \mu_i \in \mathscr{P}_2^a(X_n)$, $i = 0, 1$. Then by assumption there is a geodesic $(\mu_t^n) \subset \mathscr{P}_2^a(X_n)$ such that

$$\mathscr{E}_\infty(\mu_t^n) \leq (1-t)\mathscr{E}_\infty(\mu_0^n) + t\mathscr{E}_\infty(\mu_1^n) - \frac{K}{2}t(1-t)W_2^2(\mu_0^n, \mu_1^n). \tag{138}$$

Now let $\sigma_t^n := (\gamma_n)_\# \mu_t^n \in \mathscr{P}_2^a(X)$, $t \in [0, 1]$. From Proposition 8.5 and its proof we know that $W_2(\mu_i, \sigma_i^n) \to 0$ as $n \to \infty$, $i = 0, 1$. Also, from (138) ad Lemma 8.4, we know that $\mathscr{E}_\infty(\sigma_t^n)$ is uniformly bounded in n, t. Thus for every fixed t the sequence $n \mapsto \sigma_t^n$ is tight, and we can extract a subsequence, not relabeled, such that σ_t^n narrowly converges to some $\sigma_t \in \mathscr{P}_2(\mathrm{supp}(m))$ for every rational t. By an equicontinuity argument it is not hard to see that then σ_t^n narrowly converges to some σ_t for any $t \in [0, 1]$ (we omit the details). We claim that (σ_t) is a geodesic, and that the K-convexity inequality is satisfied along it. To check that it is a geodesic just notice that for any partition $\{t_i\}$ of $[0, 1]$ we have

$$\begin{aligned} W_2(\mu_0, \mu_1) &= \lim_{n\to\infty} W_2(\sigma_0^n, \sigma_1^n) = \lim_{n\to\infty} \sum_i W_2(\sigma_{t_i}^n, \sigma_{t_{i+1}}^n) \\ &\geq \sum_i \varliminf_{n\to\infty} W_2(\sigma_{t_i}^n, \sigma_{t_{i+1}}^n) \geq \sum_i W_2(\sigma_{t_i}, \sigma_{t_{i+1}}). \end{aligned}$$

Passing to the limit in (138), recalling Proposition 8.5 to get that $\mathscr{E}_\infty(\mu_i^n) \to \mathscr{E}_\infty(\mu_i)$, $i = 0, 1$, and that $\varliminf_{n\to\infty} \mathscr{E}_\infty(\mu_t^n) \geq \varliminf_{n\to\infty} \mathscr{E}_\infty(\sigma_t^n) \geq \mathscr{E}_\infty(\sigma_t)$ we conclude.

To deal with general μ_0, μ_1, we start recalling that the sublevels of $\mathscr{E}_\infty$ are tight, indeed using first the bound $z\log(z) \geq -\frac{1}{e}$ and then Jensen's inequality we get

$$\frac{1}{e} + C \geq \frac{m(X \setminus E)}{e} + \mathscr{E}_\infty(\mu) \geq \int_E \rho \log(\rho) dm \geq \mu(E) \log\left(\frac{\mu(E)}{m(E)}\right),$$

for any $\mu = \rho m$ such that $\mathscr{E}_\infty(\mu) \leq C$ and any Borel $E \subset X$. This bound gives that if $m(E_n) \to 0$ then $\mu(E_n) \to 0$ uniformly on the set of μ's such that $\mathscr{E}_\infty(\mu) \leq C$. This fact together with the tightness of m gives the claimed tightness of the sublevels of $\mathscr{E}_\infty$.

Now the conclusion follows by a simple truncation argument using the narrow compactness of the sublevels of $\mathscr{E}_\infty$ and the lower semicontinuity of $\mathscr{E}_\infty$ w.r.t. narrow convergence.

For the stability of the $CD(0, N)$ condition, the argument is the following: we first deal with the case of μ_0, μ_1 with bounded densities with exactly the same ideas used for $\mathscr{E}_\infty$. Then to pass to the general case we use the fact that if (X, d, m) is a $CD(0, N)$ space, then $(\mathrm{supp}(m), d, m)$ is a doubling space (Proposition 8.15 below—notice that $\mathscr{E}_{N'} \leq N'$ and thus it is not true that sublevels of $\mathscr{E}_{N'}$ are tight) and therefore boundedly compact. Then the inequality

$$R^2 \mu(\mathrm{supp}(m) \setminus B_R(x_0)) \leq \int d^2(\cdot, x_0) d\mu,$$

shows that the set of μ's in $\mathscr{P}_2^a(X)$ with bounded second moment is tight. Hence the conclusion follows, as before, using this narrow compactness together with the lower semicontinuity of $\mathscr{E}_{N'}$ w.r.t. narrow convergence. □

It remains to discuss the **interest**: from now on we discuss some of the geometric and analytic properties of spaces having a weak Ricci curvature bound.

Proposition 8.13 (Restriction and rescaling). *Let (X, d, m) be a $CD(K, \infty)$ space (resp. $CD(0, N)$ space). Then:*

(i) ***Restriction****. If $Y \subset X$ is a closed totally convex subset (i.e. every geodesic with endpoints in Y lies entirely inside Y) such that $m(Y) > 0$, then the space $(Y, d, m(Y)^{-1}m|_Y)$ is a $CD(K, \infty)$ space (resp. $CD(0, N)$ space).*
(ii) ***Rescaling****. For every $\alpha > 0$ the space $(X, \alpha d, m)$ is a $CD(\alpha^{-2}K, \infty)$ space (resp. $CD(0, N)$ space).*

Proof. (**i**). Pick $\mu_0, \mu_1 \in \mathscr{P}(Y) \subset \mathscr{P}(X)$ and a constant speed geodesic $(\mu_t) \subset \mathscr{P}(X)$ connecting them such that

$$\mathscr{E}_\infty(\mu_t) \le (1-t)\mathscr{E}_\infty(\mu_0) + t\mathscr{E}_\infty(\mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0, \mu_1),$$

(resp. satisfying the convexity inequality for the functional $\mathscr{E}_{N'}$, $N' \ge N$).

We claim that $\mathrm{supp}(\mu_t) \subset Y$ for any $t \in [0, 1]$. Recall Theorem 3.10 and pick a measure $\mu \in \mathscr{P}(\mathrm{Geod}(X))$ such that

$$\mu_t = (\mathrm{e}_t)_\#\mu,$$

where e_t is the evaluation map defined by (14). Since $\mathrm{supp}(\mu_0), \mathrm{supp}(\mu_1) \subset Y$ we know that for any geodesic $\gamma \in \mathrm{supp}(\mu)$ it holds $\gamma_0, \gamma_1 \in Y$. Since Y is totally convex, this implies that $\gamma_t \in Y$ for any t and any $\gamma \in \mathrm{supp}(\mu)$, i.e. $\mu_t = (\mathrm{e}_t)_\#\mu \in \mathscr{P}(Y)$. Therefore (μ_t) is a geodesic connecting μ_0 to μ_1 in (Y, d). Conclude noticing that for any $\mu \in \mathscr{P}_2(Y)$ it holds

$$\int \frac{d\mu}{dm_Y}\log\left(\frac{d\mu}{dm_Y}\right)dm_Y = \log(m(Y)) + \int \frac{d\mu}{dm}\log\left(\frac{d\mu}{dm}\right)dm,$$

$$\int \left(\frac{d\mu}{dm_Y}\right)^{1-\frac{1}{N'}} dm_Y = m(Y)^{-\frac{1}{N'}} \int \left(\frac{d\mu}{dm}\right)^{1-\frac{1}{N'}} dm,$$

where we wrote m_Y for $m(Y)^{-1}m|_Y$.

(**ii**). Fix $\alpha > 0$ and let $\tilde{d} := \alpha d$ and $\tilde{W}_2$ be the Wasserstein distance on $\mathscr{P}(X)$ induced by the distance $\tilde{d}$. It is clear that a plan $\gamma \in \mathcal{A}dm(\mu, \nu)$ is optimal for the distance W_2 if and only if it is optimal for $\tilde{W}_2$, thus $\tilde{W}_2 = \alpha W_2$. Now pick μ_0, $\mu_1 \in \mathscr{P}(X)$ and let $(\mu_t) \subset \mathscr{P}(X)$ be a constant speed geodesic connecting them such that

$$\mathscr{E}_\infty(\mu_t) \le (1-t)\mathscr{E}(\mu_0) + t\mathscr{E}(\mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0, \mu_1),$$

then it holds

$$\mathscr{E}_\infty(\mu_t) \le (1-t)\mathscr{E}(\mu_0) + t\mathscr{E}(\mu_1) - \frac{K}{2\alpha^2}t(1-t)\tilde{W}_2^2(\mu_0,\mu_1),$$

and the proof is complete. A similar argument applies for the case $CD(0,N)$. □

For $A_0, A_1 \subset X$, we define $[A_0, A_1]_t \subset X$ as:

$$[A_0, A_1]_t := \Big\{\gamma(t) \ : \ \gamma \text{ is a constant speed geodesic such that } \gamma(0) \in A_0,\ \gamma(1) \in A_1\Big\}.$$

Observe that if A_0, A_1 are open (resp. compact) $[A_0, A_1]_t$ is open (resp. compact), hence Borel.

Proposition 8.14 (Brunn–Minkowski). *Let (X,d,m) be a metric measure space and $A_0,\ A_1 \subset \mathrm{supp}(m)$ compact subsets. Then:*

(i) *If (X,d,m) is a $CD(K,\infty)$ space it holds:*

$$\log(m([A_0,A_1]_t)) \ge (1-t)\log(m(A_0)) + t\log(m(A_1)) + \frac{K}{2}t(1-t)D_K^2(A_0,A_1), \tag{139}$$

where $D_K(A_0,A_1)$ is defined as $\sup_{\substack{x_0\in A_0\\ x_1\in A_1}} d(x_0,x_1)$ if $K<0$ and as $\inf_{\substack{x_0\in A_0\\ x_1\in A_1}} d^2(x_0,x_1)$ if $K>0$.

(ii) *If (X,d,m) is a $CD(0,N)$ space it holds:*

$$m\big([A_0,A_1]_t\big)^{1/N} \ge (1-t)m(A_0)^{1/N} + tm(A_1)^{1/N}. \tag{140}$$

Proof. We start with (i). Suppose that A_0, A_1 are open satisfying $m(A_0), m(A_1) > 0$. Define the measures $\mu_i := m(A_i)^{-1}m|_{A_i}$ for $i = 0,1$ and find a constant speed geodesic $(\mu_t) \subset \mathscr{P}(X)$ such that

$$\mathscr{E}_\infty(\mu_t) \le (1-t)\mathscr{E}_\infty(\mu_0) + t\mathscr{E}_\infty(\mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0,\mu_1).$$

Arguing as in the proof of the previous proposition, it is immediate to see that μ_t is concentrated on $[A_0, A_1]_t$ for any $t \in [0,1]$.

In particular $m([A_0,A_1]_t) > 0$, otherwise $\mathscr{E}_\infty(\mu_t)$ would be $+\infty$ and the convexity inequality would fail. Now let $\nu_t := m([A_0,A_1]_t)^{-1}m|_{[A_0,A_1]_t}$: an application of Jensen inequality shows that $\mathscr{E}_\infty(\mu_t) \ge \mathscr{E}_\infty(\nu_t)$, thus we have

$$\mathscr{E}_\infty(\nu_t) \le (1-t)\mathscr{E}_\infty(\mu_0) + t\mathscr{E}_\infty(\mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0,\mu_1).$$

Notice that for a general μ of the form $m(A)^{-1}m|_A$ it holds

$$\mathscr{E}_\infty(\mu) = \log\big(m(A)^{-1}\big) = -\log\big(m(A)\big),$$

and conclude using the trivial inequality

$$\inf_{\substack{x_0\in A_0\\ x_1\in A_1}} d^2(x_0,x_1) \le W_2^2(\mu_0,\mu_1) \le \sup_{\substack{x_0\in A_0\\ x_1\in A_1}} d^2(x_0,x_1).$$

The case of A_0, A_1 compact now follows by a simple approximation argument by considering the ε-neighborhood $A_i^\varepsilon := \{x : d(x, A_i) < \varepsilon\}$, $i = 0, 1$, noticing that $[A_0, A_1]_t = \cap_{\varepsilon>0}[A_0^\varepsilon, A_1^\varepsilon]_t$, for any $t \in [0, 1]$ and that $m(A_i^\varepsilon) > 0$ because $A_i \subset \operatorname{supp}(m)$, $i = 0, 1$.

Part (*ii*) follows along the same lines taking into account that for a general μ of the form $m(A)^{-1}m|_A$ it holds

$$\mathscr{E}_N(\mu) = N(1 - m(A)^{1/N}),$$

and that, as before, if $m(A_0), m(A_1) > 0$ it cannot be $m([A_0, A_1]_t) = 0$ or we would violate the convexity inequality. □

A consequence of Brunn–Minkowski is the Bishop–Gromov inequality.

Proposition 8.15 (Bishop–Gromov). *Let* (X, d, m) *be a* $CD(0, N)$ *space. Then it holds*

$$\frac{m(B_r(x))}{m(B_R(x))} \ge \left(\frac{r}{R}\right)^N, \qquad \forall x \in \operatorname{supp}(m). \tag{141}$$

In particular, $(\operatorname{supp}(m), d, m)$ *is a doubling space.*

Proof. Pick $x \in \operatorname{supp}(m)$ and assume that $m(\{x\}) = 0$. Let $v(r) := m(\overline{B_r(x)})$. Fix $R > 0$ and apply the Brunn–Minkowski inequality to $A_0 = \{x\}$, $A_1 = B_R(x)$ observing that $[A_0, A_1]_t \subset \overline{B_{tR}(x)}$ to get

$$v^{1/N}(tR) \ge m\big([A_0, A_1]_t\big)^{1/N} \ge t v^{1/N}(R), \qquad \forall 0 \le t \le 1.$$

Now let $r := tR$ and use the arbitrariness of R, t to get the conclusion.

It remains to deal with the case $m(\{x\}) \ne 0$. We can also assume $\operatorname{supp}(m) \ne \{x\}$, otherwise the thesis would be trivial: under this assumption we will prove that $m(\{x\}) = 0$ for any $x \in X$.

A simple consequence of the geodesic convexity of $\mathscr{E}_N$ tested with delta measures is that $\operatorname{supp}(m)$ is a geodesically convex set, therefore it is uncountable. Then there must exist some $x' \in \operatorname{supp}(m)$ such that $m(\{x'\}) = 0$. Apply the previous argument with x' in place of x to get that

$$\frac{v(r)}{v(R)} \ge \left(\frac{r}{R}\right)^N, \qquad \forall 0 \le r < R, \tag{142}$$

where now $v(r)$ is the volume of the closed ball of radius r around x'. By definition, v is right continuous; letting $r \uparrow R$ we obtain from (142) that v is also left continuous. Thus it is continuous, and in particular the volume of the spheres $\{y : d(y, x') = r\}$ is 0 for any $r \ge 0$. In particular $m(\{y\}) = 0$ for any $y \in X$ and the proof is concluded. □

An interesting geometric consequence of the Brunn–Minkowski inequality in conjunction with the non branching hypothesis is the fact that the "cut-locus" is negligible.

Proposition 8.16 (Negligible cut-locus). *Assume that (X, d, m) is a $CD(0, N)$ space and that it is non branching. Then for every $x \in \operatorname{supp}(m)$ the set of y's such that there is more than one geodesic from x to y is m-negligible. In particular, for $m \times m$-a.e. (x, y) there exists only one geodesic $\gamma^{x,y}$ from x to y and the map $X^2 \ni (x, y) \mapsto \gamma^{x,y} \in \operatorname{Geod}(X)$ is measurable.*

Proof. Fix $x \in \operatorname{supp}(m)$, $R > 0$ and consider the sets $A_t := [\{x\}, B_R(x)]_t$. Fix $t < 1$ and $y \in A_t$. We claim that there is only one geodesic connecting it to x. By definition, we know that there is some $z \in B_R(x)$ and a geodesic γ from z to x such that $\gamma_t = y$. Now argue by contradiction and assume that there are two geodesics γ^1, γ^2 from y to x. Then starting from z, following γ for time $1 - t$, and then following each of γ^1, γ^2 for the rest of the time we find two different geodesics from z to x which agree on the non trivial interval $[0, 1 - t]$. This contradicts the non-branching hypothesis.

Clearly $A_t \subset A_s \subset B_R(x)$ for $t \leq s$, thus $t \mapsto m(A_t)$ is non decreasing. By (140) and the fact that $m(\{x\}) = 0$ (proved in Proposition 8.15) we know that $\lim_{t \to 1} m(A_t) = m(B_R(x))$ which means that m-a.e. point in $B_R(x)$ is connected to x by a unique geodesic. Since R and x are arbitrary, uniqueness is proved.

The measurability of the map $(x, y) \mapsto \gamma^{x,y}$ is then a consequence of uniqueness, of Lemma 3.11 and classical measurable selection results, which ensure the existence of a measurable selection of geodesics: in our case there is $m \times m$-almost surely no choice, so the unique geodesic selection is measurable. □

Corollary 8.17 (Compactness). *Let $N, D < \infty$. Then the family $\mathscr{X}(N, D)$ of (isomorphism classes of) metric measure spaces (X, d, m) satisfying the condition $CD(0, N)$, with diameter bounded above by D is compact w.r.t. the topology induced by $\mathbb{D}$.*

Sketch of the Proof Using the Bishop–Gromov inequality with $R = D$ we get that

$$m(\overline{B_\varepsilon(x)}) \geq \left(\frac{\varepsilon}{D}\right)^N, \qquad \forall (X, d, m) \in \mathscr{X}(N, D),\ x \in \operatorname{supp}(m_X). \tag{143}$$

Thus there exists $n(N, D, \varepsilon)$ which does not depend on $X \in \mathscr{X}(N, D)$, such that we can find at most $n(N, D, \varepsilon)$ disjoint balls of radius ε in X. Thus $\operatorname{supp}(m_X)$ can be covered by at most $n(N, D, \varepsilon)$ balls of radius 2ε. This means that the family $\mathscr{X}(N, D)$ is uniformly totally bounded, and thus it is compact w.r.t. Gromov–Hausdorff convergence (see e.g. Theorem 7.4.5 of [20]).

Pick a sequence $(X_n, d_n, m_n) \in \mathscr{X}(N, D)$. By what we just proved, up to pass to a subsequence, not relabeled, we may assume that $(\operatorname{supp}(m_n), d_n)$ converges in the Gromov–Hausdorff topology to some space (X, d). It is well known that in this situation there exists a compact space (Y, d_Y) and a family of isometric embeddings $f_n : \operatorname{supp}(m_n) \to Y$, $f : X \to Y$, such that the Hausdorff distance between $f_n(\operatorname{supp}(m_n))$ and $f(X)$ goes to 0 as $n \to \infty$.

The space $(f_n(\mathrm{supp}(m_n), d_Y, (f_n)_\# m_n))$ is isomorphic to (X_n, d_n, m_n) by construction for every $n \in \mathbb{N}$, and $(f(X), d_Y)$ is isometric to (X, d), so we identify these spaces with the respective subspaces of (Y, d_Y). Since (Y, d_Y) is compact, the sequence (m_n) admits a subsequence, not relabeled, which weakly converges to some $m \in \mathscr{P}(Y)$. It is immediate to verify that actually $m \in \mathscr{P}(X)$. Also, again by compactness, weak convergence is equivalent to convergence w.r.t. W_2, which means that there exists plans $\gamma_n \in \mathscr{P}(Y^2)$ admissible for the couple (m, m_n) such that

$$\int d_Y^2(x, \tilde{x}) d\gamma_n(x, \tilde{x}) \to 0.$$

Therefore $n \mapsto (d_Y, \gamma_n)$ is a sequence of admissible couplings for (X, d, m) and (X_n, d_n, m_n) whose cost tends to zero. This concludes the proof. □

Now we prove the HWI (which relates the entropy, often denoted by H, the Wasserstein distance W_2 and the Fisher information I) and the log-Sobolev inequalities. To this aim, we introduce the Fisher information functional $I : \mathscr{P}(X) \to [0, \infty]$ on a general metric measure space (X, d, m) as the squared slope of the entropy $\mathscr{E}_\infty$:

$$I(\mu) := \begin{cases} \overline{\lim\limits_{\nu \to \mu}} \dfrac{\left((\mathscr{E}_\infty(\mu) - \mathscr{E}_\infty(\nu))^+\right)^2}{W_2^2(\mu, \nu)}, & \text{if } \mathscr{E}_\infty(\mu) < \infty, \\ +\infty, & \text{otherwise.} \end{cases}$$

The functional I is called Fisher information because its value on $(\mathbb{R}^d, |\cdot - \cdot|, \mathscr{L}^d)$ is given by

$$I(\rho \mathscr{L}^d) = \int \frac{|\nabla \rho|^2}{\rho} d\mathscr{L}^d,$$

and the object on the right hand side is called Fisher information on $\mathbb{R}^d$. It is possible to prove that a formula like the above one is writable and true on general $CD(K, \infty)$ spaces (see [8]), but we won't discuss this topic.

Proposition 8.18 (HWI inequality). *Let (X, d, m) be a metric measure space satisfying the condition $CD(K, \infty)$. Then*

$$\mathscr{E}_\infty(\mu) \le \mathscr{E}_\infty(\nu) + W_2(\mu, \nu)\sqrt{I(\mu)} - \frac{K}{2} W_2^2(\mu, \nu), \qquad \forall \mu, \nu \in \mathscr{P}(X). \tag{144}$$

In particular, choosing $\nu = m$ it holds

$$\mathscr{E}_\infty(\mu) \le W_2(\mu, m)\sqrt{I(\mu)} - \frac{K}{2} W_2^2(\mu, m), \qquad \forall \mu \in \mathscr{P}(X). \tag{145}$$

Finally, if $K > 0$ the log-Sobolev inequality with constant K holds:

$$\mathscr{E}_\infty \le \frac{I}{2K}. \tag{146}$$

Proof. Clearly to prove (144) it is sufficient to deal with the case $\mathscr{E}_\infty(\nu)$, $\mathscr{E}_\infty(\mu) < \infty$. Let (μ_t) be a constant speed geodesic from μ to ν such that

$$\mathscr{E}_\infty(\mu_t) \le (1-t)\mathscr{E}_\infty(\mu) + t\mathscr{E}_\infty(\nu) - \frac{K}{2}t(1-t)W_2^2(\mu,\nu).$$

Then from $\sqrt{I(\mu)} \ge \overline{\lim}_{t\downarrow 0}(\mathscr{E}_\infty(\mu) - \mathscr{E}_\infty(\mu_t))/W_2(\mu,\mu_t)$ we get the thesis.

Equation (146) now follows from (145) and the trivial inequality

$$ab - \frac{1}{2}a^2 \le \frac{1}{2}b^2,$$

valid for any $a, b \ge 0$. □

The log-Sobolev inequality is a notion of *global* Sobolev-type inequality, and it is known that it implies a global Poincaré inequality (we omit the proof of this fact). When working on metric measure spaces, however, it is often important to have at disposal a *local* Poincaré inequality (see e.g. the analysis done by Cheeger in [29]).

Our final goal is to show that in non-branching $CD(0,N)$ spaces a local Poincaré inequality holds. The importance of the non-branching assumption is due to the following lemma.

Lemma 8.19. *Let (X,d,m) be a non branching $CD(0,N)$ space, $B \subset X$ a closed ball of positive measure and $2B$ the closed ball with same center and double radius. Define the measures $\mu := m(B)^{-1}m|_B$ and $\mu := \gamma_\#^{\cdot,\cdot}(\mu \times \mu) \in \mathscr{P}(\mathrm{Geod}(X))$, where $(x,y) \mapsto \gamma^{x,y}$ is the map which associates to each x, y the unique geodesic connecting them (such a map is well defined for $m \times m$-a.e. x, y by Proposition 8.16). Then*

$$(\mathrm{e}_t)_\#\mu \le \frac{2^N}{m(B)} m|_{2B}, \qquad \forall t \in [0,1].$$

Proof. Fix $x \in B$, $t \in (0,1)$ and consider the "homothopy" map $B \ni y \mapsto Hom_t^x(y) := \gamma_t^{x,y}$. By Proposition 8.16 we know that this map is well defined for m-a.e. y and that (using the characterization of geodesics given in Theorem 3.10) $t \mapsto \mu_t^x := (Hom_t^x)_\#\mu$ is the unique geodesic connecting δ_x to μ. We have

$$\mu_t^x(E) = \mu\big((Hom_t^x)^{-1}(E)\big) = \frac{m\big((Hom_t^x)^{-1}(E)\big)}{m(B)}, \qquad \forall E \subset X \text{ Borel.}$$

The non branching assumption ensures that Hom_t^x is invertible, therefore from the fact that $[\{x\}, (Hom_t^x)^{-1}(E)]_t = Hom_t^x\big((Hom_t^x)^{-1}(E)\big) = E$, the Brunn–Minkowski inequality and the fact that $m(\{x\}) = 0$ we get

$$m(E) \ge t^N m\big((Hom_t^x)^{-1}(E)\big),$$

and therefore $\mu_t^x(E) \le \frac{m(E)}{t^N m(B)}$. Given that E was arbitrary, we deduce

$$\mu_t^x \le \frac{m}{t^N m(B)}. \tag{147}$$

Notice that the expression on the right hand side is independent on x.

Now pick μ as in the hypothesis, and define $\mu_t := (\mathrm{e}_t)_\# \mu$. The equalities

$$\int_X \varphi d\mu_t = \int_{\mathrm{Geod}(X)} \varphi(\gamma_t) d\mu(\gamma) = \int_{X^2} \varphi(\gamma_t^{x,y}) d\mu(x) d\mu(y),$$
$$\int_X \varphi d\mu_t^x = \int_X \varphi(\gamma_t^{x,y}) d\mu(y),$$

valid for any $\varphi \in C_b(X)$, show that

$$\mu_t = \int \mu_t^x \, d\mu(x),$$

and therefore, by (147), we have

$$\mu_t \le \frac{m}{t^N m(B)}.$$

All these arguments can be repeated symmetrically with $1-t$ in place of t (because the push forward of μ via the map which takes γ and gives the geodesic $t \mapsto \gamma_{1-t}$, is μ itself), thus we obtain

$$\mu_t \le \min\left\{ \frac{m}{t^N m(B)}, \frac{m}{(1-t)^N m(B)} \right\} \le \frac{2^N m}{m(B)}, \qquad \forall t \in (0,1).$$

To conclude, it is sufficient to prove that μ_t is concentrated on $2B$ for all $t \in (0,1)$. But this is obvious, as μ_t is concentrated on $[B,B]_t$ and a geodesic whose endpoints lie on B cannot leave $2B$. □

As we said, we will use this lemma (together with the doubling property, which is a consequence of the Bishop–Gromov inequality) to prove a local Poincaré inequality. For simplicity, we stick to the case of Lipschitz functions and their local Lipschitz constant, although everything could be equivalently stated in terms of generic Borel functions and their upper gradients.

For $f : X \to \mathbb{R}$ Lipschitz, the local Lipschitz constant $|\nabla f| : X \to R$ is defined as

$$|\nabla f|(x) := \overline{\lim_{y \to x}} \frac{|f(x) - f(y)|}{d(x,y)}.$$

For any ball B such that $m(B) > 0$, the number $\langle f \rangle_B$ is the average value of f on B:

$$\langle f \rangle_B := \frac{1}{m(B)} \int_B f \, dm.$$

Proposition 8.20 (Local Poincaré inequality). *Assume that* (X, d, m) *is a non-branching* $CD(0, N)$ *space. Then for every ball* B *such that* $m(B) > 0$ *and any Lipschitz function* $f : X \to \mathbb{R}$ *it holds*

$$\frac{1}{m(B)}\int_B |f(x) - \langle f\rangle_B\,|\,dm(x) \le r\frac{2^{2N+1}}{m(2B)}\int_{2B}|\nabla f|dm,$$

r *being the radius of* B.

Proof. Notice that

$$\begin{aligned}\frac{1}{m(B)}\int_B |f(x) - \langle f\rangle_B\,|dm(x) &\le \frac{1}{m(B)^2}\int_{B\times B}|f(x) - f(y)|\,dm(x)dm(y)\\ &= \int_{\mathrm{Geod}(X)}|f(\gamma_0) - f(\gamma_1)|\,d\mu(\gamma),\end{aligned}$$

where μ is defined as in the statement of Lemma 8.19. Observe that for any geodesic γ, the map $t \mapsto f(\gamma_t)$ is Lipschitz and its derivative is bounded above by $d(\gamma_0, \gamma_1)|\nabla f|(\gamma_t)$ for a.e. t. Hence, since any geodesic γ whose endpoints are in B satisfies $d(\gamma_0, \gamma_1) \le 2r$, we have

$$\begin{aligned}\int_{\mathrm{Geod}(X)}|f(\gamma_0) - f(\gamma_1)|\,d\mu(\gamma) &\le 2r\int_0^1\int_{\mathrm{Geod}(X)}|\nabla f|(\gamma_t)\,d\mu(\gamma)dt\\ &= 2r\int_0^1\int_X |\nabla f|d(\mathrm{e}_t)_\#\mu dt.\end{aligned}$$

By Lemma 8.19 we obtain

$$2r\int_0^1\int_X|\nabla f|d(\mathrm{e}_t)_\#\mu dt \le \frac{2^{N+1}r}{m(B)}\int_{2B}|\nabla f|dm.$$

By the Bishop–Gromov inequality we know that $m(2B) \le 2^N m(B)$ and thus

$$\frac{2^{N+1}r}{m(B)}\int_{2B}|\nabla f|dm \le \frac{2^{2N+1}r}{m(2B)}\int_{2B}|\nabla f|dm,$$

which is the conclusion. □

8.3 Bibliographical Notes

The content of this chapter is taken from the works of Lott and Villani on one side [57, 58] and of Sturm [74, 75] on the other.

The first link between K-geodesic convexity of the relative entropy functional in $(\mathscr{P}_2(M), W_2)$ and the bound from below on the Ricci curvature is has been given by Sturm and von Renesse in [76]. The works [74,75] and [58] have been developed independently. The main difference between them is that Sturm provides the general definition of $CD(K, N)$ bound (which we didn't speak about, with the exception of the quick citation in Remark 8.9), while Lott and Villani focused on the cases $CD(K, \infty)$ and $CD(0, N)$. Apart from this, the works are strictly related and the differences are mostly on the technical side. We mention only one of these. In giving the definition of $CD(0, N)$ space we followed Sturm and asked only the functionals $\rho m \mapsto N' \int (\rho - \rho^{1-1/N'})dm$, $N' \geq N$, to be geodesically convex. Lott and Villani asked for something more restrictive, namely they introduced the *displacement convexity* classes DC_N as the set of functions $u : [0, \infty) \to \mathbb{R}$ continuous, convex and such that

$$z \qquad \mapsto \qquad z^N u(z^{-N}),$$

is convex. Notice that $u(z) := N'(z - z^{1-1/N'})$ belongs to DC_N. Then they say that a space is $CD(0, N)$ provided

$$\rho m \qquad \mapsto \qquad \int u(\rho)dm,$$

(with the usual modifications for a measure which is not absolutely continuous) is geodesically convex for any $u \in DC_N$. This notion is still compatible with the Riemannian case and stable under convergence. The main advantage one has in working with this definition is the fact that for a $CD(0, N)$ space in this sense, for any couple of absolutely continuous measures there exists a geodesic connecting them which is made of absolutely continuous measures.

The distance $\mathbb{D}$ that we used to define the notion of convergence of metric measure spaces has been defined and studied by Sturm in [74]. This is not the only possible notion of convergence of metric measure spaces: Lott and Villani used a different one, see [58] or Chap. 27 of [80]. A good property of the distance $\mathbb{D}$ is that it pleasantly reminds the Wasserstein distance W_2: to some extent, the relation of $\mathbb{D}$ to W_2 is the same relation that there is between Gromov–Hausdorff distance and Hausdorff distance between compact subsets of a given metric space. A bad property is that it is not suitable to study convergence of metric measure spaces which are endowed with infinite reference measures (well, the definition can easily be adapted, but it would lead to a too strict notion of convergence—very much like the Gromov–Hausdorff distance, which is not used to discuss convergence of non compact metric spaces). The only notion of convergence of Polish spaces endowed with σ-finite measures that we are aware of, is the one discussed by Villani in Chap. 27 of [80] (Definition 27.30). It is interesting to remark that this notion of convergence does *not* guarantee uniqueness of the limit (which can be though of as a negative point of the theory), yet, bounds from below on the Ricci curvature are stable w.r.t. such convergence (which in turn is a positive point, as it tells that these bounds are "even more stable").

The discussion on the local Poincaré inequality and on Lemma 8.19 is extracted from [57].

There is much more to say about the structure and the properties of spaces with Ricci curvature bounded below. This is an extremely fast evolving research area, and to give a complete discussion on the topic one would probably need a book nowadays. Two things are worth to be quickly mentioned.

The first one is the most important open problem on the subject: is the property of being a $CD(K,N)$ space a local notion? That is, suppose we have a metric measure space (X,d,m) and a finite open cover $\{\Omega_i\}$ such that $(\Omega_i,d,m(\Omega_i)^{-1}m|_{\Omega_i})$ is a $CD(K,N)$ space for every i. Can we deduce that (X,d,m) is a $CD(K,N)$ space as well? One would like the answer to be affirmative, as any notion of curvature should be local. For $K=0$ or $N=\infty$, this is actually the case, at least under some technical assumptions. The general case is still open, and up to now we only know that the Conjecture 30.34 in [80] is *false*, being disproved by Deng and Sturm in [32] (see also [11]).

The second, and final, thing we want to mention is the case of Finsler manifolds, which are differentiable manifolds endowed with a norm—possibly not coming from an inner product—on each tangent space, which varies smoothly with the base point. A simple example of Finsler manifolds is the space $(\mathbb{R}^d,\|\cdot\|)$, where $\|\cdot\|$ is any norm. It turns out that for any choice of the norm, the space $(\mathbb{R}^d,\|\cdot\|,\mathscr{L}^d)$ is a $CD(0,N)$ space. Various experts have different opinion about this fact: namely, there is no agreement on the community concerning whether one really wants or not Finsler geometries to be included in the class of spaces with Ricci curvature bounded below. In any case, it is interesting to know whether there exists a different, more restrictive, notion of Ricci curvature bound which rules out the Finsler case. Progresses in this direction have been made in [9], where the notion of spaces with *Riemannian Ricci* bounded below is introduced: shortly said, these spaces are the subclass of $CD(K,N)$ spaces where the heat flow (studied in [8,45,53]) is linear.

Acknowledgements Work partially supported by a MIUR PRIN2008 grant.

References

1. A. Agrachev, P. Lee, Optimal transportation under nonholonomic constraints. Trans. Am. Math. Soc. **361**, 6019–6047 (2009)
2. G. Alberti, On the structure of singular sets of convex functions. Calc. Var. Partial Differ. Equat. **2**, 17–27 (1994)
3. G. Alberti, L. Ambrosio, A geometrical approach to monotone functions in $\mathbf{R}^n$. Math. Z. **230**, 259–316 (1999)
4. L. Ambrosio, Lecture notes on optimal transport problem, in *Mathematical Aspects of Evolving Interfaces*, vol. 1812, ed. by P. Colli, J. Rodrigues. CIME summer school in Madeira (Pt) (Springer, Berlin, 2003), pp. 1–52
5. L. Ambrosio, N. Gigli, Construction of the parallel transport in the Wasserstein space. Meth. Appl. Anal. **15**, 1–29 (2008)
6. L. Ambrosio, S. Rigot, Optimal mass transportation in the Heisenberg group. J. Funct. Anal. **208**, 261–301 (2004)

7. L. Ambrosio, N. Gigli, G. Savaré, in *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd edn. Lectures in Mathematics ETH Zürich (Birkhäuser, Basel, 2008)
8. L. Ambrosio, N. Gigli, G. Savaré, Calculus and heat flows in metric measure spaces with ricci curvature bounded below, Comm. Pure and Applied Math. (2011)
9. L. Ambrosio, N. Gigli, G. Savaré, Spaces with riemannian ricci curvature bounded below, Comm. Pure and Applied Math. (2011)
10. L. Ambrosio, B. Kirchheim, A. Pratelli, Existence of optimal transport maps for crystalline norms. Duke Math. J. **125** 207–241 (2004)
11. K. Bacher, K.T. Sturm, Localization and tensorization properties of the curvature-dimension condition for metric measure spaces. J. Funct. Anal. **259**, 28–56 (2010)
12. J.-D. Benamou, Y. Brenier, A numerical method for the optimal time-continuous mass transport problem and related problems, in *Monge Ampère Equation: Applications to Geometry and Optimization* (Deerfield Beach, FL, 1997). Contemporary Mathematics, vol. 226 (American Mathematical Society, Providence, 1999), pp. 1–11
13. P. Bernard, B. Buffoni, Optimal mass transportation and Mather theory. J. Eur. Math. Soc. (JEMS), **9**, 85–127 (2007)
14. M. Bernot, V. Caselles, J.-M. Morel, The structure of branched transportation networks. Calc. Var. Partial Differ. Equat. **32**, 279–317 (2008)
15. S. Bianchini, A. Brancolini, Estimates on path functionals over Wasserstein spaces. SIAM J. Math. Anal. **42**, 1179–1217 (2010)
16. A. Brancolini, G. Buttazzo, F. Santambrogio, Path functionals over Wasserstein spaces. J. Eur. Math. Soc. (JEMS), **8**, 415–434 (2006)
17. L. Brasco, G. Buttazzo, F. Santambrogio, A Benamou-Brenier approach to branched transport. SIAM J. Math. Anal. **43**(2), 1023–1040 (2011). doi:10.1137/10079286X
18. Y. Brenier, Décomposition polaire et réarrangement monotone des champs de vecteurs. C. R. Acad. Sci. Paris I Math. **305**, 805–808 (1987)
19. Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. **44**, 375–417 (1991)
20. D. Burago, Y. Burago, S. Ivanov, in *A Course in Metric Geometry*. Graduate Studies in Mathematics, vol. 33 (American Mathematical Society, Providence, 2001)
21. L.A. Caffarelli, Boundary regularity of maps with convex potentials. Comm. Pure Appl. Math. **45**, 1141–1151 (1992)
22. L.A. Caffarelli, The regularity of mappings with a convex potential. J. Am. Math. Soc. **5**, 99–104 (1992)
23. L.A. Caffarelli, Boundary regularity of maps with convex potentials, II. Ann. Math. (2) **144**, 453–496 (1996)
24. L.A. Caffarelli, M. Feldman, R.J. McCann, Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. J. Am. Math. Soc. **15**, 1–26 (2002) (electronic)
25. L. Caravenna, A proof of Sudakov theorem with strictly convex norms. Math. Z. **268**(1–2), 371–407 (2011) doi:10.1007/s00209-010-0677-6
26. J.A. Carrillo, S. Lisini, G. Savaré, D. Slepcev, Nonlinear mobility continuity equations and generalized displacement convexity. J. Funct. Anal. **258**, 1273–1309 (2010)
27. T. Champion, L. De Pascale, The Monge problem in $\mathbb{R}^d$. Duke Math. J. **157**(3), 551–572 (2011). doi:10.1215/00127094-1272939
28. T. Champion, L. De Pascale, The Monge problem for strictly convex norms in $\mathbb{R}^d$. J. Eur. Math. Soc. (JEMS), **12**, 1355–1369 (2010)
29. J. Cheeger, Differentiability of Lipschitz functions on metric measure spaces. Geom. Funct. Anal. **9**, 428–517 (1999)
30. D. Cordero-Erausquin, B. Nazaret, C. Villani, A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. Adv. Math. **182**, 307–332 (2004)
31. C. Dellacherie, P.-A. Meyer, in *Probabilities and Potential*. North-Holland Mathematics Studies, vol. 29 (North-Holland, Amsterdam, 1978)

32. Q. Deng, K.-T. Sturm, Localization and tensorization properties of the curvature-dimension condition for metric measure spaces, II. J. Funct. Anal. **260**(12), 3718–3725 (2011). doi:10.1016/j.jfa.2011.02.026
33. J. Dolbeault, B. Nazaret, G. Savaré, On the Bakry-Emery criterion for linear diffusions and weighted porous media equations. Comm. Math. Sci **6**, 477–494 (2008)
34. L.C. Evans, W. Gangbo, Differential equations methods for the Monge-Kantorovich mass transfer problem. Mem. Am. Math. Soc. **137**, viii+66 (1999)
35. A. Fathi, A. Figalli, Optimal transportation on non-compact manifolds. Isr. J. Math. **175**, 1–59 (2010)
36. D. Feyel, A.S. Üstünel, Monge-Kantorovitch measure transportation and Monge-Ampère equation on Wiener space. Probab. Theor. Relat. Fields **128**, 347–385 (2004)
37. A. Figalli, N. Gigli, A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. J. Math. Pures Appl. (9), **94**(2), 107–130 (2010). doi:10.1016/j.matpur.2009.11.005
38. A. Figalli, F. Maggi, A. Pratelli, A mass transportation approach to quantitative isoperimetric inequalities. Invent. Math. **182**, 167–211 (2010)
39. A. Figalli, L. Rifford, Mass transportation on sub-Riemannian manifolds. Geom. Funct. Anal. **20**, 124–159 (2010)
40. N. Fusco, F. Maggi, A. Pratelli, The sharp quantitative isoperimetric inequality. Ann. Math. (2) **168**, 941–980 (2008)
41. W. Gangbo, The Monge mass transfer problem and its applications, in *Monge Ampère Equation: Applications to Geometry and Optimization*, (Deerfield Beach, FL, 1997). Contemporary Mathematics, vol. 226 (American Mathematical Society, Providence, 1999), pp. 79–104
42. W. Gangbo, R.J. McCann, The geometry of optimal transportation. Acta Math. **177**, 113–161 (1996)
43. N. Gigli, On the geometry of the space of probability measures in R^n endowed with the quadratic optimal transport distance, Thesis (Ph.D.)–Scuola Normale Superiore, 2008
44. N. Gigli, Second order analysis on $(P_2(M), W_2)$. Memoir. Am. Math. Soc. **216**(1018), xii+154 (2012). doi:10.1090/S0065-9266-2011-00619-2
45. N. Gigli, On the heat flow on metric measure spaces: existence, uniqueness and stability. Calc. Var. Partial Differential Equations **39**(1–2), 101–120 (2010). doi:10.1007/s00526-009-0303-9
46. N. Gigli, On the inverse implication of Brenier-McCann theorems and the structure of $(P_2(M), W_2)$. Methods Appl. Anal. **18**(2), 127–158 (2011)
47. R. Jordan, D. Kinderlehrer, F. Otto, The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. **29**, 1–17 (1998) (electronic)
48. N. Juillet, On displacement interpolation of measures involved in Brenier's theorem. Proc. Am. Math. Soc. **139**(10), 3623–3632 (2011). doi:10.1090/S0002-9939-2011-10891-8
49. L.V. Kantorovich, On an effective method of solving certain classes of extremal problems. Dokl. Akad. Nauk. USSR **28**, 212–215 (1940)
50. L.V. Kantorovich, On the translocation of masses. Dokl. Akad. Nauk. USSR **37**, 199–201 (1942). English translation in J. Math. Sci. **133**(4), 1381–1382 (2006)
51. L.V. Kantorovich, G.S. Rubinshtein, On a space of totally additive functions. Vestn. Leningr. Univ. **7**(13), 52–59 (1958)
52. M. Knott, C.S. Smith, On the optimal mapping of distributions. J. Optim. Theor. Appl. **43**, 39–49 (1984)
53. K. Kuwada, N. Gigli, S.-I. Ohta, Heat flow on alexandrov spaces, Comm. Pure and Applied Math. (2010)
54. S. Lisini, Characterization of absolutely continuous curves in Wasserstein spaces. Calc. Var. Partial Differ. Equat. **28**, 85–120 (2007)
55. G. Loeper, On the regularity of solutions of optimal transportation problems. Acta Math. **202**, 241–283 (2009)
56. J. Lott, Some geometric calculations on Wasserstein space. Comm. Math. Phys. **277**, 423–437 (2008)

57. J. Lott, C. Villani, Weak curvature conditions and functional inequalities. J. Funct. Anal. **245**(1), 311–333 (2007). doi:10.1016/j.jfa.2006.10.018
58. J. Lott, C. Villani, Ricci curvature for metric-measure spaces via optimal transport. Ann. Math. **169**(2), 903–991 (2009)
59. X.-N. Ma, N.S. Trudinger, and X.-J. Wang, Regularity of potential functions of the optimal transportation problem. Arch. Ration. Mech. Anal. **177**, 151–183 (2005)
60. F. Maddalena, S. Solimini, Transport distances and irrigation models. J. Convex Anal. **16**, 121–152 (2009)
61. F. Maddalena, S. Solimini, J.-M. Morel, A variational model of irrigation patterns. Interfaces Free Bound. **5**, 391–415 (2003)
62. R.J. Mccann, A convexity theory for interacting gases and equilibrium crystals. Ph.D. Thesis, Princeton University. ProQuest LLC, Ann Arbor (1994)
63. R.J. McCann, A convexity principle for interacting gases. Adv. Math. **128**, 153–179 (1997)
64. R.J. McCann, Polar factorization of maps on riemannian manifolds. Geom. Funct. Anal. **11**, 589–608 (2001)
65. V.D. Milman, G. Schechtman, in *Asymptotic Theory of Finite-Dimensional Normed Spaces*. Lecture Notes in Mathematics, vol. 1200 (Springer, Berlin, 1986). With an appendix by M. Gromov
66. G. Monge, Mémoire sur la théorie des d'eblais et des remblais. Histoire de lÕAcadémie Royale des Sciences de Paris (1781), pp. 666–704
67. F. Otto, The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differ. Equat. **26**, 101–174 (2001)
68. A. Pratelli, On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. Ann. l'Institut Henri Poincare B Probab. Stat. **43**, 1–13 (2007)
69. S.T. Rachev, L. Rüschendorf, Mass *Transportation Problems, vol. I. Probability and Its Applications* (Springer, New York, 1998), pp. xxvi+508 (Theory)
70. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
71. L. Rüschendorf, S.T. Rachev, A characterization of random variables with minimum L^2-distance. J. Multivariate Anal. **32**, 48–54 (1990)
72. G. Savaré, Gradient flows and diffusion semigroups in metric spaces under lower curvature bounds. C. R. Math. Acad. Sci. Paris **345**, 151–154 (2007)
73. G. Savaré, Gradient flows and evolution variational inequalities in metric spaces (2010) (in preparation)
74. K.-T. Sturm, On the geometry of metric measure spaces, I. Acta Math. **196**, 65–131 (2006)
75. K.-T. Sturm, On the geometry of metric measure spaces, II. Acta Math. **196**, 133–177 (2006)
76. K.-T. Sturm, M.-K. von Renesse, Transport inequalities, gradient estimates, entropy, and Ricci curvature. Comm. Pure Appl. Math. **58**, 923–940 (2005)
77. V.N. Sudakov, Geometric problems in the theory of infinite-dimensional probability distributions. Proc. Steklov Inst. Math. (2), i–v, 1–178 (1979) (Cover to cover translation of Trudy Mat. Inst. Steklov **141** (1976))
78. N.S. Trudinger, X.-J. Wang, On the Monge mass transfer problem. Calc. Var. Partial Differ. Equat. **13**, 19–31 (2001)
79. C. Villani, in *Topics in Optimal Transportation*. Graduate Studies in Mathematics, vol. 58 (American Mathematical Society, Providence, 2003)
80. C. Villani, *Optimal Transport, Old and New* (Springer, Berlin, 2008)
81. Q. Xia, Optimal paths related to transport problems. Comm. Contemp. Math. **5**, 251–279 (2003)
82. Q. Xia, Interior regularity of optimal transport paths. Calc. Var. Partial Differ. Equat. **20**, 283–299 (2004)
83. L. Zajı̂˜ cek, On the differentiability of convex functions in finite and infinite dimensional spaces. Czechoslovak Math. J. **29**, 340–348 (1979)

Hyperbolic Conservation Laws: An Illustrated Tutorial

Alberto Bressan

Abstract These notes provide an introduction to the theory of hyperbolic systems of conservation laws in one space dimension. The various chapters cover the following topics: (1) Meaning of a conservation equation and definition of weak solutions. (2) Hyperbolic systems. Explicit solutions in the linear, constant coefficients case. Nonlinear effects: loss of regularity and wave interactions. (3) Shock waves: Rankine–Hugoniot equations and admissibility conditions. (4) Genuinely nonlinear and linearly degenerate characteristic fields. Centered rarefaction waves. The general solution of the Riemann problem. Wave interaction estimates. (5) Weak solutions to the Cauchy problem, with initial data having small total variation. Approximations generated by the front-tracking method and by the Glimm scheme. (6) Continuous dependence of solutions w.r.t. the initial data, in the $\mathbf{L}^1$ distance. (7) Characterization of solutions which are limits of front tracking approximations. Uniqueness of entropy-admissible weak solutions. (8) Vanishing viscosity approximations. (9) Extensions and open problems. The survey is concluded with an Appendix, reviewing some basic analytical tools used in the previous sections.

Throughout the exposition, technical details are mostly left out. The main goal of these notes is to convey basic ideas, also with the aid of a large number of figures.

A. Bressan
Department of Mathematics, Penn State University,
University Park, PA 16802, USA
e-mail: bressan@math.psu.edu

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_2,

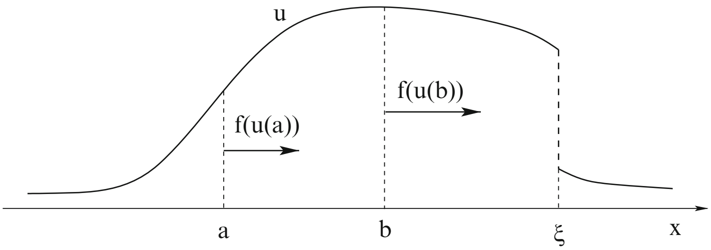


Fig. 1 Flow across two points

1 Conservation Laws

1.1 The Scalar Conservation Law

A *scalar conservation law* in one space dimension is a first order partial differential equation of the form

$$u_t + f(u)_x = 0. \tag{1}$$

Here $u = u(t,x)$ is called the *conserved quantity*, while f is the *flux*. The variable t denotes time, while x is the one-dimensional space variable. Integrating (1) over a given interval $[a,b]$ one obtains

$$\begin{aligned} \frac{d}{dt}\int_a^b u(t,x)\,dx &= \int_a^b u_t(t,x)\,dx \;=\; -\int_a^b f\big(u(t,x)\big)_x\,dx \\ &= f\big(u(t,a)\big) - f\big(u(t,b)\big) \;=\; [\text{inflow at } a] - [\text{outflow at } b]. \end{aligned} \tag{2}$$

According to (2), the quantity u is neither created nor destroyed: the total amount of u contained inside any given interval $[a,b]$ can change only due to the flow of u across the two boundary points (Fig. 1).

Using the chain rule, (1) can be written in the quasilinear form

$$u_t + a(u)u_x = 0, \tag{3}$$

where $a = f'$ is the derivative of f. For smooth solutions, the two (1) and (3) are entirely equivalent. However, if u has a jump at a point ξ, the left hand side of (3) will contain the product of a discontinuous function $a(u)$ with the distributional derivative u_x, which in this case contains a Dirac mass at the point ξ. In general, such a product is not well defined. Hence (3) is meaningful only within a class of continuous functions. On the other hand, working with the equation in divergence form (1) allows us to consider discontinuous solutions as well, interpreted in distributional sense.

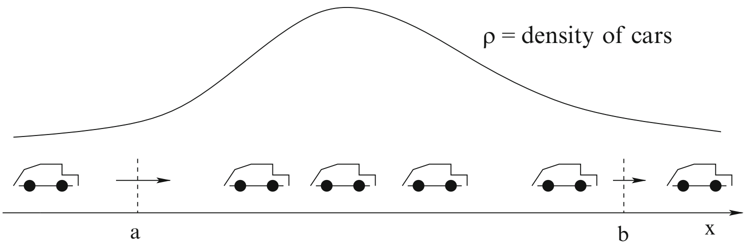


Fig. 2 The density of cars can be described by a conservation law

A function $u = u(t,x)$ will be called a *weak solution* of (1) provided that

$$\iint \left\{u\phi_t + f(u)\phi_x\right\} dxdt \;=\; 0 \tag{4}$$

for every continuously differentiable function with compact support $\phi \in \mathscr{C}^1_c$. Notice that (4) is meaningful as soon as both u and $f(u)$ are *locally integrable* in the t-x plane.

Example 1 (traffic flow). Let $\rho(t,x)$ be the density of cars on a highway, at the point x at time t. For example, u may be the number of cars per kilometer (Fig. 2). In the classic Lighthill–Witham model [33,43], one assumes that the velocity v of the cars depends only on their density, say

$$v = v(\rho), \qquad\qquad \text{with} \qquad \frac{dv}{d\rho} < 0.$$

Given any two points a, b on the highway, the number of cars between a and b therefore varies according to the law

$$\begin{aligned}\int_a^b \rho_t(t,x)\,dx &= \frac{d}{dt}\int_a^b \rho(t,x)\,dx \;=\; [\text{inflow ata}] - \;[\text{outflow at}b]\\ &=\; v\big(\rho(t,a)\big)\cdot\rho(t,a) - v\big(\rho(t,b)\big)\cdot\rho(t,b) = -\int_a^b \big[v(\rho)\,\rho\big]_x\,dx.\end{aligned} \tag{5}$$

Since (5) holds for all a, b, this leads to the conservation law

$$\rho_t + \big[v(\rho)\,\rho\big]_x = 0,$$

where ρ is the conserved quantity and $f(\rho) = v(\rho)\rho$ is the flux function.

1.2 Strictly Hyperbolic Systems

The main object of our study will be the $n \times n$ *system of conservation laws*

$$\begin{cases} \dfrac{\partial}{\partial t} u_1 + \dfrac{\partial}{\partial x} f_1(u_1, \ldots, u_n) = 0, \\ \qquad\qquad \cdots \\ \dfrac{\partial}{\partial t} u_n + \dfrac{\partial}{\partial x} f_n(u_1, \ldots, u_n) = 0. \end{cases} \tag{6}$$

To shorten notation, it is convenient to write this system also in the form (1). However, one should keep in mind that now $u = (u_1, \ldots, u_n)$ is a vector in $I\!R^n$ while $f = (f_1, \ldots, f_n)$ is a map from $I\!R^n$ into $I\!R^n$. Calling

$$A(u) \doteq Df(u) = \begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ & \cdots & \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix},$$

the $n \times n$ Jacobian matrix of the map f at the point u, the system (6) can be written in the quasilinear form

$$u_t + A(u)u_x = 0. \tag{7}$$

A $\mathscr{C}^1$ function $u = u(t, x)$ provides a classical solution to (6) if and only if it solves (7). In addition, for the conservative system (6) one can also consider weak solutions $u \in \mathbf{L}^1_{loc}$ in distributional sense, according (4).

In order to achieve the well-posedness of the initial value problem, a basic algebraic property will now be introduced.

Definition 1 (strictly hyperbolic system). The system of conservation laws (6) is *strictly hyperbolic* if, for every u, the Jacobian matrix $A(u) = Df(u)$ has n real, distinct eigenvalues: $\lambda_1(u) < \cdots < \lambda_n(u)$.

If the matrix $A(u)$ has real distinct eigenvalues, one can find bases of left and right eigenvectors, denoted by $l_1(u), \ldots, l_n(u)$ and $r_1(u), \ldots, r_n(u)$. The left eigenvectors are regarded as row vectors, while right eigenvectors are column vectors. For every $u \in I\!R^n$ and $i = 1, \ldots, n$, we thus have

$$A(u)r_i(u) = \lambda_i(u)r_i(u), \qquad\qquad l_i(u)A(u) = \lambda_i(u)l_i(u).$$

It is convenient to choose dual bases of left and right eigenvectors, so that

$$|r_i| = 1, \qquad\qquad l_i \cdot r_j = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases} \tag{8}$$

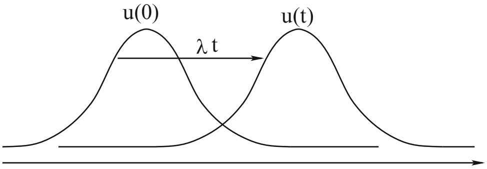


Fig. 3 A traveling wave solution to the linear, scalar Cauchy problem (10)–(11)

Example 2 (gas dynamics). The Euler equations describing the evolution of a non viscous gas take the form

$$\begin{cases} \rho_t + (\rho v)_x = 0 & \text{(conservation of mass)} \\ (\rho v)_t + (\rho v^2 + p)_x = 0 & \text{(conservation of momentum)} \\ (\rho E)_t + (\rho E v + p v)_x = 0 & \text{(conservation of energy)} \end{cases}$$

Here ρ is the mass density, v is the velocity while $E = e + v^2/2$ is the energy density per unit mass. The system is closed by a *constitutive relation* of the form $p = p(\rho, e)$, giving the pressure as a function of the density and the internal energy. The particular form of p depends on the gas under consideration. Denoting by $u = (u_1, u_2, u_3) = (\rho, \rho v, \rho E)$ the vector of conserved quantities, one checks that for physically meaningful functions $p = p(\rho, e)$ the above system is strictly hyperbolic [26, 57].

1.3 Linear Systems

We describe here two elementary cases where the solution of the initial value problem can be written explicitly.

Consider the initial value problem for a scalar conservation law

$$u_t + f(u)_x = 0, \tag{9}$$

$$u(0, x) = \bar{u}(x). \tag{10}$$

In the special case where the flux f is an affine function, say $f(u) = \lambda u + c$, the (9) reduces to

$$u_t + \lambda u_x = 0. \tag{11}$$

The Cauchy problem (10)–(11) admits an explicit solution, namely

$$u(t, x) = \bar{u}(x - \lambda t). \tag{12}$$

As shown in Fig. 3, this has the form of a traveling wave, with speed $\lambda = f'(u)$. If $\bar{u} \in \mathscr{C}^1$, the function $u = u(t, x)$ defined by (12) is a classical solution. On the other

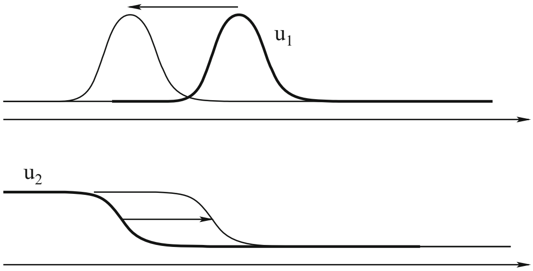


Fig. 4 The solution to the linear hyperbolic system (13) is obtained as the superposition of n traveling waves

hand, if the initial condition $\bar{u}$ is not differentiable and we only have $\bar{u} \in \mathbf{L}^1_{loc}$, the above function u can still be interpreted as a weak solution in distributional sense.

Next, consider the linear homogeneous system with constant coefficients

$$u_t + Au_x = 0, \qquad\qquad u(0,x) = \bar{u}(x), \tag{13}$$

where A is a $n \times n$ hyperbolic matrix, with real eigenvalues $\lambda_1 < \cdots < \lambda_n$ and right and left eigenvectors $r_i,\ l_i$, chosen as in (8).

Call $u_i \doteq l_i \cdot u$ the coordinates of a vector $u \in I\!R^n$ w.r.t. the basis of right eigenvectors $\{r_1, \cdots, r_n\}$. Multiplying (13) on the left by $l_1, \ldots, l_n$ we obtain

$$\begin{aligned}(u_i)_t + \lambda_i(u_i)_x = (l_i u)_t + \lambda_i(l_i u)_x \ &= \ l_i u_t + l_i A u_x \ = \ 0,\\ u_i(0,x) = l_i \bar{u}(x) \ &\doteq \ \bar{u}_i(x).\end{aligned}$$

Therefore, (13) decouples into n scalar Cauchy problems, which can be solved separately in the same way as (10)–(11). The function

$$u(t,x) \ = \ \sum_{i=1}^{n} \bar{u}_i(x - \lambda_i t)\, r_i \tag{14}$$

now provides the explicit solution to (13), because

$$u_t(t,x) \ = \ \sum_{i=1}^{n} -\lambda_i \big(l_i \cdot \bar{u}_x(x - \lambda_i t)\big) r_i \ = \ -Au_x(t,x).$$

Observe that in the scalar case (11) the initial profile is shifted with constant speed $\lambda = f'(u)$. For the system (13), the initial profile is decomposed as a sum of n waves (Fig. 4), each traveling with one of the characteristic speeds $\lambda_1, \ldots, \lambda_n$.

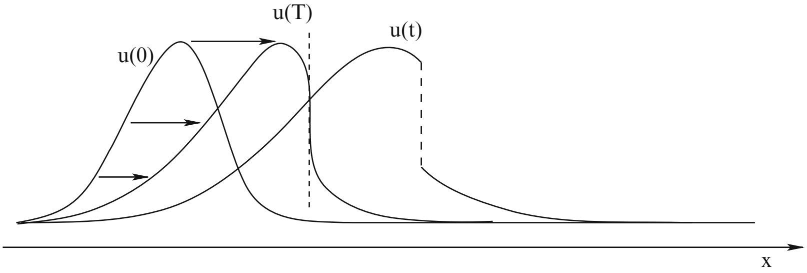


Fig. 5 If the wave propagation speed depends on u, the profile of the solution changes in time, eventually leading to shock formation at a finite time T

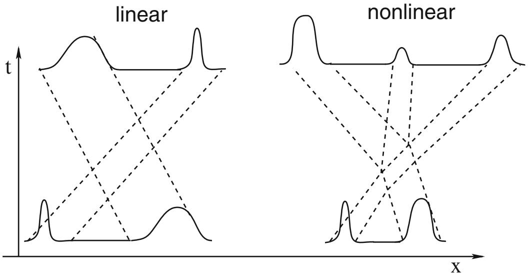


Fig. 6 *Left*: for the linear hyperbolic system (13), the solution is a simple superposition of traveling waves. *Right*: For the general non-linear system (6), waves of different families have nontrivial interactions

1.4 Nonlinear Effects

In the general case where the matrix A depends on the state u, new features will appear in the solutions.

(a) Since the eigenvalues λ_i now depend on u, the shape of the various components in the solution will vary in time (Fig. 5). Rarefaction waves will decay, and compression waves will become steeper, possibly leading to shock formation in finite time.

(b) Since the eigenvectors r_i also depend on u, nontrivial interactions between different waves will occur (Fig. 6). The strength of the interacting waves may change, and new waves of different families can be created, as a result of the interaction.

The strong nonlinearity of the equations and the lack of regularity of solutions, also due to the absence of second order terms that could provide a smoothing effect, account for most of the difficulties encountered in a rigorous mathematical analysis of the system (1). It is well known that the main techniques of abstract functional analysis do not apply in this context. Solutions cannot be represented as fixed points of continuous transformations, or in variational form, as critical points of suitable functionals. Dealing with vector valued functions, comparison principles based on upper or lower solutions cannot be used. Moreover, the theory of accretive operators and contractive nonlinear semigroups works well in the scalar case [25], but does not apply to systems. For the above reasons, the theory of hyperbolic conservation laws has largely developed by *ad hoc* methods, along two main lines.

1. The BV setting, considered by J. Glimm [34]. Solutions are here constructed within a space of functions with bounded variation, controlling the BV norm by a wave interaction functional.
2. The $\mathbf{L}^\infty$ setting, considered by L. Tartar and R. DiPerna [29], based on weak convergence and a compensated compactness argument.

Both approaches yield results on the global existence of weak solutions. However, the method of compensated compactness appears to be suitable only for 2×2 systems. Moreover, it is only in the BV setting that the well-posedness of the Cauchy problem could recently be proved, as well as the stability and convergence of vanishing viscosity approximations. In these lecture we thus restrict ourselves to the analysis of BV solutions, referring to [29] or [50,56] for the alternative approach based on compensated compactness.

1.5 Loss of Regularity

A basic feature of nonlinear systems of the form (1) is that, even for smooth initial data, the solution of the Cauchy problem may develop discontinuities in finite time. To achieve a global existence result, it is thus essential to work within a class of discontinuous functions, interpreting the (1) in their distributional sense (4).

The loss of regularity can be seen already in the solution to a scalar equation with nonlinear flux. Consider the scalar Cauchy problem

$$u_t + f(u)_x = 0 \qquad\qquad u(0,x) = \phi(x). \tag{15}$$

In the case of smooth solutions, the equation can be written in quasilinear form

$$u_t + f'(u)u_x = 0. \tag{16}$$

Geometrically, this means that the directional derivative of $u(t,x)$ in the direction of the vector $(1\,,\ f'(u))$ vanishes. Hence u is constant on each line of the form $\Big\{(t,x)\,;\ \ x = x_0 + tf'(u(x_0))\Big\}$. For each $x_0 \in I\!R$ we thus have

$$u\Big(t,\ x_0 + t\ f'(\phi(x_0))\Big)\ =\ \phi(x_0). \tag{17}$$

This is indeed the solution to the first order PDE (16) provided by the classical method of characteristics, see for example [31]. In general, beyond a finite time T, the map

$$x_0 \mapsto x_0 + t\ f'(\phi(x_0))$$

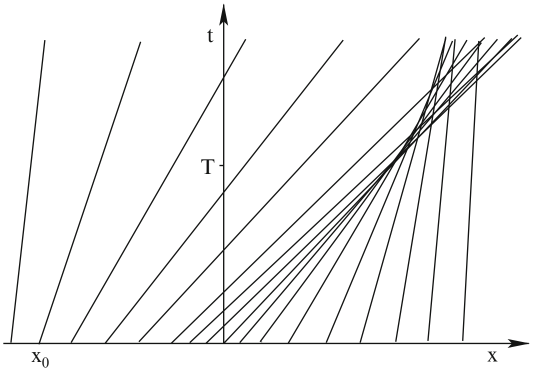


Fig. 7 At time T when characteristics start to intersect, a shock is produced

is no longer one-to-one, and the implicit (17) does not define a single valued function $u = u(t, x)$. At time T a shock is formed, and the solution can be extended for $t > T$ in the weak sense, as in (4).

Example 3 (shock formation in Burgers' equation). Consider the scalar conservation law (inviscid Burgers' equation)

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \tag{18}$$

with initial condition

$$u(0, x) = \bar{u}(x) = \frac{1}{1 + x^2}.$$

For $t > 0$ small the solution can be found by the method of characteristics. Indeed, if u is smooth, (18) is equivalent to

$$u_t + uu_x = 0. \tag{19}$$

By (19) the directional derivative of the function $u = u(t, x)$ along the vector $(1, u)$ vanishes. Therefore, u must be constant along the characteristic lines in the t-x plane:

$$t \mapsto \big(t,\ x + t\bar{u}(x)\big) = \left(t,\ x + \frac{t}{1 + x^2}\right).$$

For $t < T \doteq 8/\sqrt{27}$, these lines do not intersect (Fig. 7). The solution to our Cauchy problem is thus given implicitly by

$$u\left(t,\ x + \frac{t}{1 + x^2}\right) = \frac{1}{1 + x^2}. \tag{20}$$

On the other hand, when $t > T$, the characteristic lines start to intersect. As a result, the map

$$x \;\mapsto\; x + \frac{t}{1+x^2}$$

is not one-to-one and (20) no longer defines a single valued solution of our Cauchy problem.

An alternative point of view is the following (Fig. 5). As time increases, points on the graph of $u(t,\cdot)$ move horizontally with speed u, equal to their distance from the x-axis. This determines a change in the profile of the solution. As t approaches the critical time $T \doteq 8/\sqrt{27}$, one has

$$\lim_{t\to T-} \left\{ \inf_{x\in I\!R} \; u_x(t,x)\right\} \;=\; -\infty,$$

and no classical solution exists beyond time T. The solution can be prolonged for all times $t \geq 0$ only within a class discontinuous functions.

1.6 Wave Interactions

Consider the quasilinear, strictly hyperbolic system

$$u_t \;=\; -A(u)u_x. \tag{21}$$

If the matrix A is independent of u, then the solution can be obtained as a superposition of traveling waves. On the other hand, if A depends on u, these waves can interact with each other, producing additional waves. To understand this nonlinear effect, define the i-th component of the gradient u_x as

$$u_x^i \;\doteq\; l_i\cdot u_x. \tag{22}$$

We regard u_x^i as the i-th component of the gradient u_x w.r.t. the basis of eigenvectors $\{r_1(u),\ldots,r_n(u)\}$. Equivalently, one can also think of u_x^i as the density of i-waves in the solution u. From (22) and (8), (21) it follows

$$u_x \;=\; \sum_{i=1}^n u_x^i r_i(u) \qquad\qquad u_t \;=\; -\sum_{i=1}^n \lambda_i(u)u_x^i r_i(u)$$

Differentiating the first equation w.r.t. t and the second one w.r.t. x, then equating the results, one obtains a system of evolution equations for the scalar components u_x^i, namely

$$(u_x^i)_t + (\lambda_i u_x^i)_x \;=\; \sum_{j>k}(\lambda_j-\lambda_k)\Big(l_i\cdot[r_j,r_k]\Big)u_x^j u_x^k. \tag{23}$$

See [8] or [42] for details. Notice that the left hand side of (23) is in conservation form. However, here the total amount of waves can increase in time, due to the source terms on the right hand side. The source term

$$S_{ijk} \doteq (\lambda_j - \lambda_k)\Big(l_i \cdot [r_j, r_k]\Big) u_x^j u_x^k$$

describes the amount of i-waves produced by the interaction of j-waves with k-waves. Here

$$\begin{aligned}
\lambda_j - \lambda_k &= \text{[difference in speed]} \\
&= \text{[rate at which} j - \text{waves and} k - \text{waves cross each other]} \\[1em]
u_x^j u_x^k &= \text{[density of} j - \text{waves]} \times \text{[density of} k - \text{waves]} \\[1em]
[r_j, r_k] &= (Dr_k) r_j - (Dr_j) r_k \quad \text{(Lie bracket)} \\
&= \text{[directional derivative of } r_k \text{ in the direction of } r_j] \\
&\quad - \text{[directional derivative of } r_j \text{ in the direction of } r_k].
\end{aligned}$$

Finally, the product $l_i \cdot [r_j, r_k]$ gives the i-th component of the Lie bracket $[r_j, r_k]$ along the basis of eigenvectors $\{r_1, \dots, r_n\}$.

2 Weak Solutions

A basic feature of nonlinear hyperbolic systems is the possible loss of regularity: solutions which are initially smooth may become discontinuous within finite time. In order to construct solutions globally in time, we are thus forced to work in a space of discontinuous functions, and interpret the conservation equations in a distributional sense.

Definition 2 (weak solution). Let $f : I\!R^n \mapsto I\!R^n$ be a smooth vector field. A measurable function $u = u(t, x)$, defined on an open set $\Omega \subseteq I\!R \times I\!R$ and with values in $I\!R^n$, is a *weak solution* of the system of conservation laws

$$u_t + f(u)_x = 0 \tag{24}$$

if, for every $\mathscr{C}^1$ function $\phi : \Omega \mapsto I\!R$ with compact support, one has

$$\iint_\Omega \left\{ u\,\phi_t + f(u)\,\phi_x \right\} dxdt = 0. \tag{25}$$

Observe that no continuity assumption is made on u. To make sense of the integral in (25) we only need that u and $f(u)$ be locally integrable in Ω. Notice also that weak solutions are defined up to $\mathbf{L}^1$ equivalence. A solution is not affected by changing its values on a set of measure zero in the t-x plane. An easy consequence of the above definition is the closure of the set of solutions w.r.t. convergence in $\mathbf{L}^1_{\mathrm{loc}}$.

Lemma 1. *Let $(u_m)_{m\geq 1}$ be a uniformly bounded sequence of distributional solutions of (24). If $u_m \to u$ and $f(u_m) \to f(u)$ in $\mathbf{L}^1_{\mathrm{loc}}$ then the limit function u is also a weak solution.*

Indeed, for every $\phi \in \mathscr{C}^1_c$ one has

$$\iint_\Omega \left\{u\,\phi_t + f(u)\,\phi_x\right\}\,dxdt \;=\; \lim_{m\to\infty}\iint_\Omega \left\{u_m\,\phi_t + f(u_m)\,\phi_x\right\}\,dxdt \;=\; 0.$$

□

We observe that, in particular, the assumptions of the lemma are satisfied if $u_m \to u$ in $\mathbf{L}^1_{loc}$ and the flux function f is bounded.

In the following, we shall be mainly interested in solutions defined on a strip $[0,T]\times I\!R$, with an assigned initial condition

$$u(0,x) = \bar{u}(x). \tag{26}$$

Here $\bar{u} \in \mathbf{L}^1_{\mathrm{loc}}(I\!R)$. To treat the initial value problem, it is convenient to require some additional regularity w.r.t. time.

Definition 3 (weak solution to the Cauchy problem). A function $u : [0,T]\times I\!R \mapsto I\!R^n$ is a *weak solution* of the Cauchy problem (24), (26) if u is continuous as a function from $[0,T]$ into $\mathbf{L}^1_{\mathrm{loc}}$, the initial condition (26) holds and the restriction of u to the open strip $]0,T[\,\times I\!R$ is a distributional solution of (24).

Remark 1 (classical solutions). Let u be a weak solution of (24). If u is continuously differentiable restricted to an open domain $\widetilde{\Omega} \subseteq \Omega$, then at every point $(t,x) \in \widetilde{\Omega}$, the function u must satisfy the quasilinear system

$$u_t + A(u)u_x = 0, \tag{27}$$

with $A(u) \doteq Df(u)$. Indeed, from (25) an integration by parts yields

$$\iint \left[u_t + A(u)u_x\right]\phi dxdt = 0.$$

Since this holds for every $\phi \in \mathscr{C}^1_c(\widetilde{\Omega})$, the identity (27) follows.

2.1 Rankine–Hugoniot Conditions

Next, we look at a discontinuous solution and derive some conditions which must be satisfied at points of jump. Consider first the simple case of a piecewise constant function, say

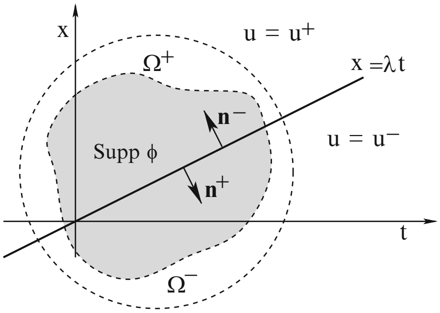


Fig. 8 Deriving the Rankine–Hugoniot equations. Here the *shaded area* describes the support of the test function ϕ

$$U(t,x) = \begin{cases} u^+ & \text{if} \qquad x > \lambda t, \\ u^- & \text{if} \qquad x < \lambda t, \end{cases} \tag{28}$$

for some $u^-, u^+ \in I\!R^n$, $\lambda \in I\!R$.

Lemma 2. *If the function U in (2.5) is a weak solution of the system of conservation laws (2.1), then*

$$\lambda\,(u^+ - u^-) = f(u^+) - f(u^-). \tag{29}$$

Proof. Let $\phi = \phi(t,x)$ be any continuously differentiable function with compact support. Let Ω be an open disc containing the support of ϕ and consider the two domains

$$\Omega^+ \doteq \Omega \cap \{x > \lambda t\}, \qquad\qquad \Omega^- \doteq \Omega \cap \{x < \lambda t\},$$

as in Fig. 8. Introducing the vector field $\mathbf{v} \doteq \big(U\phi,\ f(U)\phi\big)$, and recalling that U is constant separately on Ω_- and on Ω_+, we write the identity (25) as

$$\iint_{\Omega^+\cup\Omega^-} \Big\{U\phi_t + f(U)\phi_x\Big\}\,dxdt \;=\; \left(\iint_{\Omega^+} + \iint_{\Omega^-}\right) \operatorname{div}\mathbf{v}\,dxdt \;=\; 0. \tag{30}$$

We now apply the divergence theorem separately on the two domains Ω^+, Ω^-. Call $\mathbf{n}^+, \mathbf{n}^-$ the outer unit normals to Ω^+, Ω^-, respectively. Observe that $\phi = 0$ on the boundary $\partial\Omega$. Therefore, the only portion of the boundaries $\partial\Omega_-, \partial\Omega_+$ where $\mathbf{v} \neq 0$ is the line where $x = \lambda t$. Denoting by ds the differential of the arc-length, along the line $\{x = \lambda t\}$ we have

$$\mathbf{n}^+\,ds = (\lambda,\ -1)\,dt \qquad\qquad \mathbf{n}^-\,ds = (-\lambda,\ 1)\,dt\,,$$

$$\begin{aligned} 0 = \iint_{\Omega^+\cup\Omega^-} \operatorname{div}\mathbf{v}\,dxdt \;&=\; \int_{\partial\Omega^+} \mathbf{n}^+\cdot\mathbf{v}\,ds + \int_{\partial\Omega^-} \mathbf{n}^-\cdot\mathbf{v}\,ds \\ = \int \big[\lambda u^+ - f(u^+)\big]\,\phi(t,\lambda t)\,dt &+ \int \big[-\lambda u^- + f(u^-)\big]\,\phi(t,\lambda t)\,dt. \end{aligned}$$

Therefore, the identity

$$\int \big[\lambda(u^+ - u^-) - f(u^+) + f(u^-)\big]\,\phi(t,\lambda t)\,dt \;=\; 0$$

must hold for every function $\phi \in \mathscr{C}_c^1$. This implies (29). □

The vector equations (29) are the famous *Rankine–Hugoniot conditions*. They form a set of n scalar equations relating the right and left states $u^+, u^- \in I\!R^n$ and the speed λ of the discontinuity, namely:

$$\text{[speed of the shock]} \times \text{[jump in the state]} \;=\; \text{[jump in the flux]}.$$

An alternative way of writing these conditions is as follows. Denote by $A(u) = Df(u)$ the $n \times n$ Jacobian matrix of f at u. For any $u, v \in I\!R^n$, define the averaged matrix

$$A(u,v) \doteq \int_0^1 A\big(\theta v + (1-\theta)u\big)\,d\theta \tag{31}$$

and call $\lambda_i(u,v)$, $i = 1,\ldots,n$, its eigenvalues. We observe that $A(u,v) = A(v,u)$ and $A(u,u) = A(u)$. Equation (29) can now be written in the equivalent form

$$\begin{aligned}\lambda\,(u^+ - u^-) \;&= f(u^+) - f(u^-) = \int_0^1 Df\big(\theta u^+ + (1-\theta)u^-\big)\cdot(u^+ - u^-)\,d\theta \\ &= A(u^-,u^+)\cdot(u^+ - u^-).\end{aligned} \tag{32}$$

In other words, the Rankine–Hugoniot conditions hold if and only if the jump $u^+ - u^-$ is an eigenvector of the averaged matrix $A(u^-,u^+)$ and the speed λ coincides with the corresponding eigenvalue.

Remark 2. In the scalar case, one arbitrarily assign the left and right states $u^-, u^+ \in I\!R$ and determine the shock speed as

$$\lambda \;=\; \frac{f(u^+) - f(u^-)}{u^+ - u^-} \;=\; \frac{1}{u^+ - u^-}\int_{u^-}^{u^+} f'(s)\,ds. \tag{33}$$

A geometric interpretation of these identities (see Fig. 9) is that

$$\begin{aligned}\text{[speed of the shock]} &= \text{[slope of secant line through } u^-, u^+ \text{ on the graph of } f\text{]} \\ &= \text{[average of the characteristic speeds between } u^- \text{ and } u^+\text{]}.\end{aligned}$$

We now consider a more general solution $u = u(t,x)$ of (24) and show that the Rankine–Hugoniot equations are still satisfied at every point (τ,ξ) where u has an approximate jump, in the following sense [32].

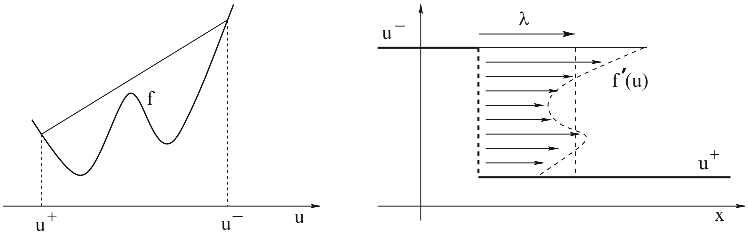


Fig. 9 The Rankine–Hugoniot equation in the scalar case

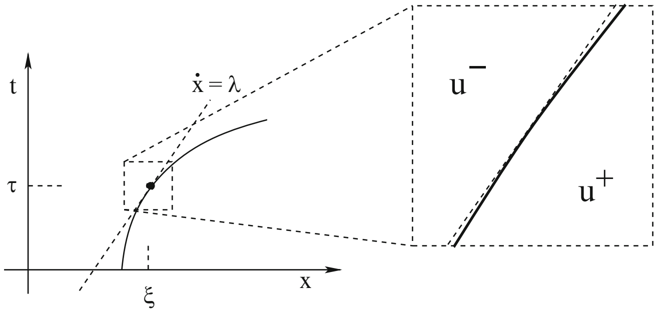


Fig. 10 A point of approximate jump. Looking through a microscope, i.e. rescaling the variables t, x in a neighborhood of the point (τ, ξ), the function u becomes arbitrarily close (in an integral sense) to the piecewise constant function U in (28)

Definition 4 (approximate jump). We say that a function $u = u(t,x)$ has an *approximate jump discontinuity* at the point (τ, ξ) if there exists vectors $u^+ \neq u^-$ and a speed λ such that, defining U as in (28), there holds

$$\lim_{r \to 0+} \frac{1}{r^2} \int_{-r}^{r} \int_{-r}^{r} \Big| u(\tau + t,\ \xi + x) - U(t,x) \Big| \, dx dt \;=\; 0. \tag{34}$$

Moreover, we say that u is *approximately continuous* at the point (τ, ξ) if the above relations hold with $u^+ = u^-$ (and λ arbitrary).

Observe that the above definitions depend only on the $\mathbf{L}^1$ equivalence class of u. Indeed, the limit in (34) is unaffected if the values of u are changed on a set $\mathcal{N} \subset I\!R^2$ of Lebesgue measure zero.

Example 4 (a piecewise smooth function). Let $g^-, g^+ : I\!R^2 \mapsto I\!R^n$ be any two continuous functions and let $x = \gamma(t)$ be a smooth curve, with derivative $\dot{\gamma}(t) \doteq \frac{d}{dt}\gamma(t)$. Define the function (see Fig. 10)

$$u(t,x) \doteq \begin{cases} g^-(t,x) & \text{if} \qquad x < \gamma(t), \\ g^+(t,x) & \text{if} \qquad x > \gamma(t). \end{cases}$$

At a point (τ, ξ), with $\xi = \gamma(\tau)$, call $u^- \doteq g^-(\tau, \xi)$, $u^+ \doteq g^+(\tau, \xi)$. If $u^+ = u^-$, then u is continuous at (τ, ξ), hence also approximately continuous. On the other hand, if $u^+ \neq u^-$, then u has an approximate jump at (τ, ξ). Indeed, writing $\dot\gamma(t) = \frac{d\gamma}{dt}$, the limit (34) holds with $\lambda = \dot\gamma(\tau)$ and U as in (28).

We now prove the Rankine–Hugoniot conditions in the more general case of a point of approximate jump.

Theorem 1 (Rankine–Hugoniot equations). *Let u be a bounded distributional solution of (24) having an approximate jump at a point (τ, ξ). In other words, assume that (34) holds, for some states u^-, u^+ and a speed λ, with U as in (28). Then the Rankine–Hugoniot equations (29) hold.*

Proof. For any given $\theta > 0$, the rescaled function

$$u^\theta(t, x) \doteq u(\tau + \theta t,\ \xi + \theta x)$$

is also a solution to the system of conservation laws. We claim that, as $\theta \to 0$, the convergence $u^\theta \to U$ holds in $\mathbf{L}^1_{\mathrm{loc}}(I\!R^2;\ I\!R^n)$. Indeed, for any $R > 0$ one has

$$\begin{aligned}
&\lim_{\theta\to 0} \int_{-R}^{R}\int_{-R}^{R} \left|u^\theta(t, x) - U(t, x)\right| dxdt \\
&\quad = \lim_{\theta\to 0} \frac{1}{\theta^2} \int_{-\theta R}^{\theta R}\int_{-\theta R}^{\theta R} \left|u(\tau + t,\ \xi + x) - U(t, x)\right| dxdt = 0
\end{aligned}$$

because of (34). Lemma 1 now implies that U itself is a distributional solution of (24), hence by Lemma 2 the Rankine–Hugoniot equations (29) hold. □

2.2 Construction of Shock Curves

In this section we consider the following problem. Given $u_0 \in I\!R^n$, find the states $u \in I\!R^n$ which, for some speed λ, satisfy the Rankine–Hugoniot equations

$$\lambda(u - u_0) \;=\; f(u) - f(u_0) \;=\; A(u_0, u)(u - u_0). \tag{35}$$

Trivially, the (35) are satisfied by setting $u = u_0$, with $\lambda \in I\!R$ arbitrary. Our aim is to construct non-trivial solutions with u close to u_0, relying on the implicit function theorem. Since this goal cannot be achieved by looking directly at the system (35), we adopt an alternative formulation.

Fix $i \in \{1, \ldots, n\}$. By a classical result in linear algebra, the jump $u - u_0$ is a right i-eigenvector of the averaged matrix $A(u_0, u)$ if and only if it is orthogonal to all left eigenvectors $l_j(u_0, u)$ of $A(u_0, u)$, for every $j \neq i$. This means

$$\psi_j(u) \;\doteq\; l_j(u_0, u)\cdot(u - u_0) \;=\; 0 \qquad\qquad \text{for all}\ \ j \neq i. \tag{36}$$

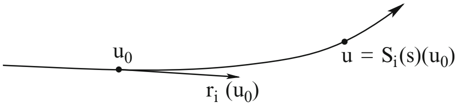


Fig. 11 Parameterization of the i-th shock curve through a point u_0

Instead of the system (35) of n equations in the $n+1$ variables $(u,\lambda) = (u_1,\dots,u_n,\lambda)$, we thus look at the system (36), consisting of $n-1$ equations for the n variables $(u_1,\dots,u_n)$.

The point $u = u_0$ is of course a solution. Moreover, the definition (31) trivially implies $A(u_0,u_0) = A(u_0)$, hence $l_j(u_0,u_0) = l_j(u_0)$ for all j. Linearizing the system (36) at $u = u_0$ we obtain the linear system of $n-1$ equations

$$l_j(u_0)\cdot(u-u_0) = 0 \qquad\qquad j \neq i. \tag{37}$$

Since the left eigenvectors $l_j(u_0)$ are linearly independent, this has maximum rank.

We can thus apply the implicit function theorem to the nonlinear system (36) and conclude that, for each $i \in \{1,\dots,n\}$, there exists a curve $s \mapsto S_i(s)(u_0)$ of points that satisfy (36). At the point u_0, this curve has to be perpendicular to all vectors $l_j(u_0)$, for $j \neq i$. Therefore, it must be tangent to the i-th eigenvector $r_i(u_0)$ (Fig. 11).

2.3 *Admissibility Conditions*

To motivate the following discussion, we first observe that the concept of weak solution is usually not stringent enough to achieve uniqueness for a Cauchy problem. In some cases, infinitely many weak solutions can be found, all with the same initial condition.

Example 5 (multiple weak solutions). For Burgers' equation

$$u_t + (u^2/2)_x = 0\,, \tag{38}$$

consider the Cauchy problem with initial data

$$u(0,x) = \begin{cases} 1 & \text{if} \qquad x \geq 0, \\ 0 & \text{if} \qquad x < 0. \end{cases}$$

As shown in Fig. 12, for every $0 < \alpha < 1$, a weak solution is

$$u_\alpha(t,x) = \begin{cases} 0 & \text{if} \qquad x < \alpha t/2, \\ \alpha & \text{if} \qquad \alpha t/2 \leq x < (1+\alpha)t/2, \\ 1 & \text{if} \qquad x \geq (1+\alpha)t/2. \end{cases} \tag{39}$$

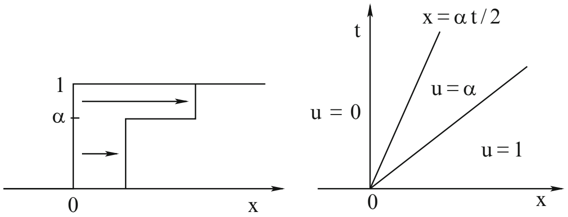


Fig. 12 For every $\alpha \in [0, 1]$ one obtains a different weak solution of Burgers' equation, always with the same initial data

Indeed, the piecewise constant function u_α trivially satisfies the equation outside the jumps. Moreover, the Rankine–Hugoniot conditions hold along the two lines of discontinuity $\{x = \alpha t/2\}$ and $\{x = (1 + \alpha)t/2\}$, for all $t > 0$.

From the previous example it is clear that, in order to achieve the uniqueness of solutions and their continuous dependence on the initial data, the notion of weak solution must be supplemented with further "admissibility conditions". Three main approaches can be followed.

I: Singular limits.

Assume that, by physical considerations, the system of conservation laws (24) can be regarded as an approximation to a more general system, say

$$u_t + f(u)_x = \varepsilon \Lambda(u), \tag{40}$$

for some $\varepsilon > 0$ small. Here $\Lambda(u)$ is typically a higher order differential operator.

We then say that a weak solution $u = u(t, x)$ of the system of conservation laws (24) is "admissible" if there exists a sequence of solutions u^ε to the perturbed (40) which converges to u in $\mathbf{L}^1_{loc}$, as $\varepsilon \to 0+$.

A natural choice is to take the diffusion operator $\Lambda(u) \doteq u_{xx}$. This leads to

Admissibility Condition 1 (vanishing viscosity). A weak solution u of (24) is *admissible in the vanishing viscosity sense* if there exists a sequence of smooth solutions u^ε to

$$u^\varepsilon_t + f(u^\varepsilon)_x \;=\; \varepsilon u^\varepsilon_{xx} \tag{41}$$

which converge to u in $\mathbf{L}^1_{\text{loc}}$ as $\varepsilon \to 0+$.

The main drawback of this approach is that it is very difficult to provide a priori estimates on general solutions to the higher order system (40), and characterize the corresponding limits as $\varepsilon \to 0+$. For the vanishing viscosity approximations (41), this goal has been reached only recently in [7], within the class of solutions with small total variation. From the above condition, however, one can deduce other conditions which can be more easily verified in practice.

II: Entropy conditions.

An alternative approach relies on the concept of entropy.

Definition 5 (entropy and entropy flux). A continuously differentiable function $\eta : I\!R^n \mapsto I\!R$ is called an *entropy* for the system of conservation laws (24), with *entropy flux* $q : I\!R^n \mapsto I\!R$, if for all $u \in I\!R^n$ there holds

$$D\eta(u)\cdot Df(u) = Dq(u). \tag{42}$$

An immediate consequence of (42) is that, if $u = u(t,x)$ is a $\mathscr{C}^1$ solution of (24), then

$$\eta(u)_t + q(u)_x = 0. \tag{43}$$

Indeed,

$$\eta(u)_t + q(u)_x \;=\; D\eta(u)u_t + Dq(u)u_x \;=\; D\eta(u)\big(-Df(u)u_x\big) + Dq(u)u_x \;=\; 0.$$

In other words, for a smooth solution u, not only the quantities $u_1,\dots,u_n$ are conserved but the additional conservation law (43) holds as well. However one should be aware that, when u is discontinuous, the quantity $\eta(u)$ may not be conserved.

Example 6. Consider Burgers' equation (38). Here the flux is $f(u) = u^2/2$. Taking $\eta(u) = u^3$ and $q(u) = (3/4)u^4$, one checks that the (42) is satisfied. Hence η is an entropy and q is the corresponding entropy flux. We observe that the function

$$u(0,x) \;=\; \begin{cases} 1 & \text{if} \qquad x < t/2, \\ 0 & \text{if} \qquad x \ge t/2, \end{cases}$$

is a (discontinuous) weak solution of (38). However, it does not satisfy (43) in distribution sense. Indeed, calling $u^- = 1$, $u^+ = 0$ the left and right states, and $\lambda = 1/2$ the speed of the shock, one has

$$\frac{3}{4} \;=\; q(u^+) - q(u^-) \;\neq\; \lambda\Big[\eta(u^+) - \eta(u^-)\Big] \;=\; \frac{1}{2}.$$

We now study how a convex entropy behaves in the presence of a small diffusion term. Assume $\eta, q \in \mathscr{C}^2$, with η convex. Multiplying both sides of (41) on the left by $D\eta(u^\varepsilon)$ and using (42) one finds

$$\big[\eta(u^\varepsilon)\big]_t + \big[q(u^\varepsilon)\big]_x \;=\; \varepsilon D\eta(u^\varepsilon)u^\varepsilon_{xx} \;=\; \varepsilon\Big\{\big[\eta(u^\varepsilon)\big]_{xx} - D^2\eta(u^\varepsilon)\cdot\big(u^\varepsilon_x\otimes u^\varepsilon_x\big)\Big\}. \tag{44}$$

Observe that the last term in (44) satisfies

$$D^2\eta(u^\varepsilon)\big(u^\varepsilon_x\otimes u^\varepsilon_x\big) \;=\; \sum_{i,j=1}^{n} \frac{\partial^2\eta(u^\varepsilon)}{\partial u_i\,\partial u_j}\cdot\frac{\partial u^\varepsilon_i}{\partial x}\,\frac{\partial u^\varepsilon_j}{\partial x} \ge 0,$$

because η is convex, hence its second derivative at any point u^ε is a positive semidefinite quadratic form. Multiplying (44) by a nonnegative smooth function φ with compact support and integrating by parts, we thus have

$$\iint \big\{\eta(u^\varepsilon)\varphi_t + q(u^\varepsilon)\varphi_x\big\}\,dxdt \;\ge\; -\varepsilon\iint \eta(u^\varepsilon)\varphi_{xx}\,dxdt.$$

If $u^\varepsilon \to u$ in $\mathbf{L}^1$ as $\varepsilon \to 0$, the previous inequality yields

$$\iint \left\{\eta(u)\varphi_t + q(u)\varphi_x\right\} dxdt \ \geq \ 0 \tag{45}$$

whenever $\varphi \in \mathscr{C}_c^1$, $\varphi \geq 0$. The above can be restated by saying that $\eta(u)_t + q(u)_x \leq 0$ in distribution sense. The previous analysis leads to:

Admissibility Condition 2 (entropy inequality). A weak solution u of (24) is *entropy-admissible* if

$$\eta(u)_t + q(u)_x \ \leq \ 0 \tag{46}$$

in the sense of distributions, for every pair (η, q), where η is a convex entropy for (24) and q is the corresponding entropy flux.

For the piecewise constant function U in (28), an application of the divergence theorem shows that $\eta(U)_t + q(U)_x \leq 0$ in distribution if and only if

$$\lambda\left[\eta(u^+) - \eta(u^-)\right] \ \geq \ q(u^+) - q(u^-). \tag{47}$$

More generally, let $u = u(t,x)$ be a bounded function which satisfies the conservation law (24). Assume that u has an approximate jump at (τ, ξ), so that (34) holds with U as in (28). Then, by the rescaling argument used in the proof of Theorem 1, one can show that the inequality (47) must again hold.

We remark that the above admissibility condition can be useful only if some nontrivial convex entropy for the system (24) is known. For $n \times n$ systems of conservation laws, the (42) can be regarded as a first order system of n equations for the two scalar variables η, q, namely

$$\left(\frac{\partial\eta}{\partial u_1} \cdots \frac{\partial\eta}{\partial u_n}\right)\begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_n} & \\ & \cdots & & \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_n} & = \end{pmatrix}\left(\frac{\partial q}{\partial u_1} \cdots \frac{\partial q}{\partial u_n}\right).$$

When $n \geq 3$, this system is overdetermined. In general, one should thus expect to find solutions only in the case $n \leq 2$. However, there are important physical examples of larger systems which admit a nontrivial entropy function.

III: Stability conditions.

Admissibility conditions on shocks can also be derived purely from stability consideration, without any reference to physical models.

We consider first the scalar case. Let $U = U(t,x)$ be the piecewise constant solution introduced in (28), with left and right states u^-, u^+. Let us slightly perturb the initial data by inserting an intermediate state $u^* \in [u^-, u^+]$, as in Fig. 13. The original shock is thus split in two smaller shocks, whose speeds are determined by the Rankine–Hugoniot equations.

To ensure that the $\mathbf{L}^1$ distance between the original solution and the perturbed one does not increase in time, we need:

$$\text{[speed of jump behind]} \ \geq \ \text{[speed of jump ahead]}.$$

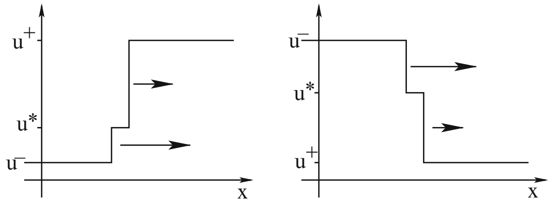


Fig. 13 In both cases $u^- < u^+$ or $u^- > u^+$, the solution is stable if the speed of the shock behind is greater or equal than the speed of the one ahead

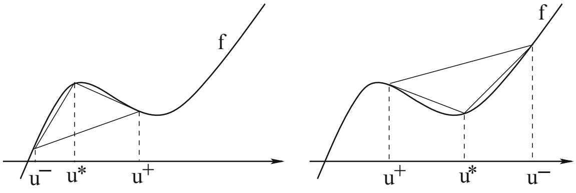


Fig. 14 Geometric interpretation of the stability conditions (48). In both cases, the jump with left state u^- and right state u^+ is admissible

By (33), this is the case if and only if

$$\frac{f(u^*) - f(u^-)}{u^* - u^-} \geq \frac{f(u^+) - f(u^*)}{u^+ - u^*} \qquad \text{for all} \quad u^* \in [u^-, u^+]. \tag{48}$$

From (48) we thus obtain the following stability conditions (see Fig. 14).

1. If $u^- < u^+$, on the interval $[u^-, u^+]$ the graph of f should remain above the secant line.
2. If $u^+ < u^-$, on the interval $[u^+, u^-]$ the graph of f should remain below the secant line.

Next, we seek a generalization of this stability conditions, valid also for $n \times n$ hyperbolic systems. Observe that, still in the scalar case, the condition (48) is equivalent to

$$\frac{f(u^*) - f(u^-)}{u^* - u^-} \geq \frac{f(u^+) - f(u^-)}{u^+ - u^-} \qquad \text{for all} \quad u^* \in [u^-, u^+]. \tag{49}$$

In other words, the speed of the original shock (u^-, u^+) should be not greater than the speed of any intermediate shock (u^-, u^*), where $u^* \in [u^-, u^+]$ is any intermediate state (Fig. 15).

Next, we consider $n \times n$ hyperbolic systems. As in Sect. 2.2, we let $s \mapsto S_i(s)(u^-)$ describe the i-shock curve through u^-. This is the curve of all states u that can be connected to u^- by a shock of the i-th family (Fig. 16).

Observe that, if $u^+ = S_i(\sigma)(u^-)$ and $u^* = S_i(s)(u^-)$ are two points on the i-shock curve through u^-, in general it is not true that the two states u^+ and u^* can be connected by a shock. For this reason, a straightforward generalization of the condition (48) to systems is not possible. However, the equivalent condition (49) has a natural extension to the vector valued case, namely:

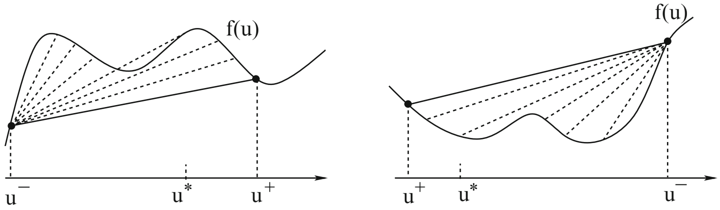


Fig. 15 Geometric interpretation of the stability conditions (49)

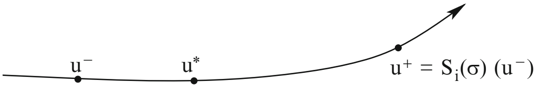


Fig. 16 The i-shock (u^-, u^+) satisfies the Liu admissibility conditions if its speed satisfies $\lambda_i(u^-, u^+) \leq \lambda_i(u^-, u^*)$ for every intermediate state u^* along the i-shock curve through u^-

Admissibility Condition 3 (Liu condition). Let $u^+ = S_i(\sigma)(u^-)$ for some $\sigma \in I\!R$. We say that the shock with left and right states u^-, u^+ *satisfies the Liu admissibility condition* provided that its speed is less or equal to the speed of every smaller shock, joining u^- with an intermediate state $u^* = S_i(s)(u^-)$, $s \in [0, \sigma]$.

This condition was introduced by T.P. Liu in [45]. Much later, the paper [7] showed that, among solutions with small total variation, the Liu condition completely characterizes the ones which can be obtained as vanishing viscosity limits.

We conclude this section by mentioning another admissibility condition, introduced by Lax in [40] and widely used in the literature.

Admissibility Condition 4 (Lax condition). A shock of the i-th family, connecting the states u^-, u^+ and traveling with speed $\lambda = \lambda_i(u^-, u^+)$, *satisfies the Lax admissibility condition* if

$$\lambda_i(u^-) \geq \lambda_i(u^-, u^+) \geq \lambda_i(u^+). \tag{50}$$

To appreciate the geometric meaning of this condition, consider a piecewise smooth solution, having a discontinuity along the line $x = \gamma(t)$, where the solution jumps from a left state u^- to a right state u^+ (see Fig. 17). According to (32), this discontinuity must travel with a speed $\lambda \doteq \dot{\gamma} = \lambda_i(u^-, u^+)$ equal to the i-eigenvalue of the averaged matrix $A(u^-, u^+)$, for some $i \in \{1, \ldots, n\}$. If we now look at the *i-characteristics*, i.e. at the solutions of the O.D.E.

$$\dot{x} = \lambda_i\big(u(t, x)\big),$$

we see that the Lax condition requires that these lines run into the shock, from both sides.

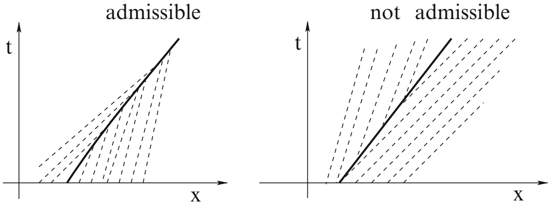


Fig. 17 *Left*: a shock satisfying the Lax condition. As time increases, characteristics run toward the shock, from both sides. *Right*: a shock violating this condition

3 The Riemann Problem

In this chapter we construct the solution to the *Riemann problem*, consisting of the system of conservation laws

$$u_t + f(u)_x = 0 \tag{51}$$

together with the simple, piecewise constant initial data

$$u(0,x) = \bar{u}(x) \doteq \begin{cases} u^- & \text{if} \qquad x < 0, \\ u^+ & \text{if} \qquad x > 0. \end{cases} \tag{52}$$

This will provide the basic building block toward the solution of the Cauchy problem with more general initial data.

This problem was first studied by B. Riemann in [52], in connection with the 2×2 system of isentropic gas dynamics. In [40], P. Lax constructed solutions to the Riemann problem for a wide class of $n \times n$ strictly hyperbolic systems. Further results were provided by T. P. Liu in [44], dealing with systems under generic assumptions. The paper [6] by S. Bianchini provides a fully general construction, valid even for systems not in conservation form. In this case, "solutions" are interpreted as limits of vanishing viscosity approximations.

The central role played by the Riemann problem, within the general theory of conservation laws, can be explained in terms of symmetries. We observe that, if $u = u(t,x)$ is a weak solution of (51), then for every $\theta > 0$ the rescaled function

$$u^\theta(t,x) \doteq u(\theta t,\, \theta x) \tag{53}$$

provides yet another solution. Among all solutions to a system of conservation laws, the Riemann problems yield precisely those weak solutions which are invariant w.r.t. the above rescaling: $u^\theta = u$ for every $\theta > 0$ (see Fig. 18).

3.1 Some Examples

We begin by describing the explicit solution of the Riemann problem (51)–(52) in a few elementary cases.

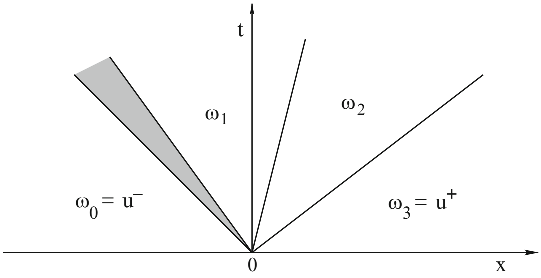


Fig. 18 The solution to a Riemann problem is constant along rays through the origin, in the t-x plane. Hence it is invariant w.r.t. the symmetry transformation (53)

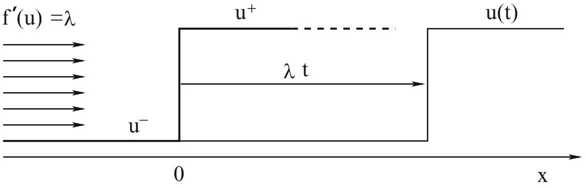


Fig. 19 A contact discontinuity. Here the characteristic speed $f'(u) \equiv \lambda$ is constant, for all values of $u \in [u^-, u^+]$

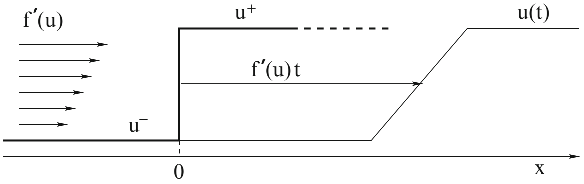


Fig. 20 The centered rarefaction wave defined at (54)

Example 7. Consider a scalar conservation law with linear flux $f(u) = \lambda u + c$.

As shown in Fig. 19, the solution of the Riemann problem is

$$u(t,x) = \begin{cases} u^- & \text{if} \qquad x < \lambda t, \\ u^+ & \text{if} \qquad x > \lambda t. \end{cases}$$

It consists of a single jump, called a *contact discontinuity*, traveling with speed λ.

Example 8. Consider a scalar conservation law with strictly convex flux, so that $u \mapsto f'(u)$ is strictly increasing. Moreover, assume that $u^+ > u^-$.

The solution is then a *centered rarefaction wave*, obtained by the method of characteristics (Fig. 20).

$$u(t,x) = \begin{cases} u^- & \text{if} \qquad \frac{x}{t} < f'(u^-), \\ \\ u^+ & \text{if} \qquad \frac{x}{t} > f'(u^+), \\ \\ \omega & \text{if} \qquad \frac{x}{t} = f'(\omega) \text{ for some } \omega \in [u^-, u^+]. \end{cases} \tag{54}$$

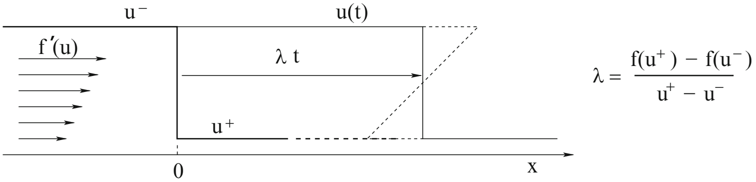


Fig. 21 A shock satisfying the admissibility conditions

Since the mapping $\omega \mapsto f'(\omega)$ is strictly increasing, for $\frac{x}{t} \in [f'(u^-),\ f'(u^+)]$ there exists a unique value $\omega \in [u^-, u^+]$ such that $\frac{x}{t} = f'(\omega)$. The above function u is thus well defined.

Example 9. Consider again a scalar conservation law with strictly convex flux. However, we now assume that $u^+ < u^-$.

The solution consists of a single *shock*:

$$u(t,x) = \begin{cases} u^- & \text{if} \qquad x < \lambda t, \\ u^+ & \text{if} \qquad x > \lambda t, \end{cases} \tag{55}$$

As usual, the shock speed is determined by the Rankine–Hugoniot equations (33). We observe that this shock satisfies both the Liu and the Lax admissibility conditions.

Remark 3. The formula (55) defines a weak solution to the Riemann problem also in Example 8. However, if $u^- < u^+$, this solution does not satisfy the Liu admissibility condition. The Lax condition fails as well.

On the other hand, if $u^+ < u^-$, the formula (54) does not define a single valued function (Fig. 21). Hence it cannot provide a solution in Example 9.

Example 10. Consider the Riemann problem for a linear system:

$$u_t + Au_x = 0 \qquad\qquad u(0,x) = \begin{cases} u^- & \text{if} \qquad x < 0, \\ u^+ & \text{if} \qquad x > 0. \end{cases}$$

For linear systems, the general solution to the Cauchy problem was already constructed in (14).

For this particular initial data, the solution can be obtained as follows. Write the vector $u^+ - u^-$ as a linear combination of eigenvectors of A, i.e.

$$u^+ - u^- = \sum_{j=1}^{n} c_j r_j\,.$$

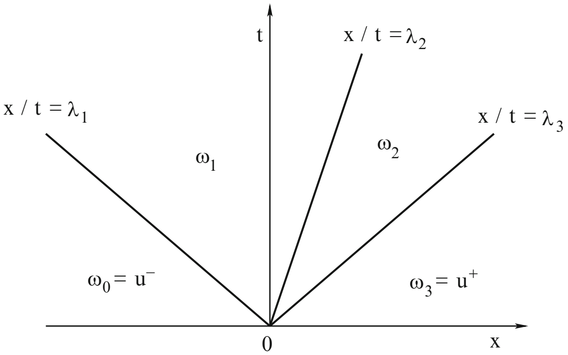


Fig. 22 Solution to the Riemann problem for a linear system

Define the intermediate states

$$\omega_i \doteq u^- + \sum_{j \le i} c_j r_j, \qquad i = 0, \dots, n.$$

The solution then takes the form

$$u(t,x) = \begin{cases} \omega_0 = u^- & for \quad x/t \ < \lambda_1, \\ & \dots \\ \omega_i & for \quad \lambda_i < x/t < \lambda_{i+1}, \\ & \dots \\ \omega_n = u^+ & for \quad x/t \ > \lambda_n. \end{cases} \tag{56}$$

Notice that, in this linear case, the general solution to the Riemann problem consists of n jumps. The i-th jump: $\omega_i - \omega_{i-1} = c_i r_i$ is parallel to the i-eigenvector of the matrix A and travels with speed λ_i, given by the corresponding eigenvalue (Fig. 22).

3.2 A Class of Hyperbolic Systems

We shall consider hyperbolic systems which satisfy the following simplifying assumption, introduced by P. Lax [40].

(H) For each $i = 1, \dots, n$, the i-th field is either *genuinely nonlinear*, so that $D\lambda_i(u) \cdot r_i(u) > 0$ for all u, or *linearly degenerate*, with $D\lambda_i(u) \cdot r_i(u) = 0$ for all u.

We recall that $D\lambda_i$ denotes the gradient of the scalar function $u \mapsto \lambda_i(u)$. Hence $D\lambda_i(u) \cdot r_i(u)$ is the directional derivative of λ_i in the direction of the vector r_i. Notice that, in the genuinely nonlinear case, the i-th eigenvalue λ_i is strictly increasing along each integral curve of the corresponding field of eigenvectors r_i.

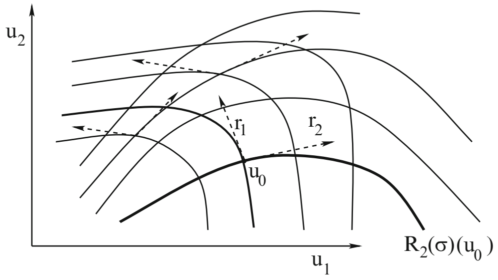


Fig. 23 Integral curves of the vector fields $r_1(u)$, $r_2(u)$

In the linearly degenerate case, on the other hand, the eigenvalue λ_i is constant along each such curve (see Fig. 23). With the above assumption (H), we are ruling out the possibility that, along some integral curve of an eigenvector r_i, the corresponding eigenvalue λ_i may partly increase and partly decrease, having several local maxima and minima.

Example 11 (isentropic gas dynamics). Denote by ρ the density of a gas, by $v = \rho^{-1}$ its specific volume and by u its velocity. A simple model for isentropic gas dynamics (in Lagrangian coordinates) is then provided by the so-called "p-system"

$$\begin{cases} v_t - u_x = 0\,, \\ u_t + p(v)_x = 0. \end{cases} \tag{57}$$

Here $p = p(v)$ is a function which determines the pressure in terms of of the specific volume. An appropriate choice is $p(v) = kv^{-\gamma}$, with $1 \leq \gamma \leq 3$. In the region where $v > 0$, the Jacobian matrix of the system is

$$A \doteq Df = \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix}.$$

The eigenvalues and eigenvectors are found to be

$$\lambda_1 = -\sqrt{-p'(v)}\,, \qquad \lambda_2 = \sqrt{-p'(v)}\,, \tag{58}$$

$$r_1 = \begin{pmatrix} 1 \\ \\ \sqrt{-p'(v)} \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -1 \\ \\ \sqrt{-p'(v)} \end{pmatrix}. \tag{59}$$

It is now clear that the system is strictly hyperbolic provided that $p'(v) < 0$ for all $v > 0$. Moreover, observing that

$$D\lambda_1 \cdot r_1 = \frac{p''(v)}{2\sqrt{-p'(v)}} = D\lambda_2 \cdot r_2\,,$$

we conclude that both characteristic fields are genuinely nonlinear if $p''(v) > 0$ for all $v > 0$.

As we shall see in the sequel, if the assumption (H) holds, then the solution of the Riemann problem has a simple structure consisting of the superposition of n elementary waves: shocks, rarefactions or contact discontinuities. This considerably simplifies all further analysis. On the other hand, for strictly hyperbolic systems that do not satisfy the condition (H), basic existence and stability results can still be obtained, but at the price of heavier technicalities [44].

3.3 Elementary Waves

Fix a state $u_0 \in I\!R^n$ and an index $i \in \{1, \dots, n\}$. As before, let $r_i(u)$ be an i-eigenvector of the Jacobian matrix $A(u) = Df(u)$. The integral curve of the vector field r_i through the point u_0 is called the *i-rarefaction curve* through u_0. It is obtained by solving the Cauchy problem in state space:

$$\frac{du}{d\sigma} = r_i(u), \qquad u(0) = u_0. \tag{60}$$

We shall denote this curve as

$$\sigma \mapsto R_i(\sigma)(u_0). \tag{61}$$

Clearly, the parametrization depends on the choice of the eigenvectors r_i. In particular, if we impose the normalization $\big|r_i(u)\big| \equiv 1$, then the rarefaction curve (61) will be parameterized by arc-length. In the genuinely nonlinear case, we always choose the orientation so that the eigenvalue $\lambda_i(u)$ increases as the parameter σ increases along the curve.

Next, for a fixed $u_0 \in I\!R^n$ and $i \in \{1, \dots, n\}$, we consider the i-shock curve through u_0. This is the set of states u which can be connected to u_0 by an i-shock. As in Sect. 2.2, this curve will be parameterized as

$$\sigma \mapsto S_i(\sigma)(u_0). \tag{62}$$

Using a suitable parametrization (say, by arclength), one can show that the two curves R_i, S_i have a second order contact at the point u_0 (see Fig. 24). More precisely, the following estimates hold.

$$\begin{cases} R_i(\sigma)(u_0) = u_0 + \sigma r_i(u_0) + \mathcal{O}(1) \cdot \sigma^2, \\ S_i(\sigma)(u_0) = u_0 + \sigma r_i(u_0) + \mathcal{O}(1) \cdot \sigma^2, \end{cases} \tag{63}$$

$$\Big|R_i(\sigma)(u_0) - S_i(\sigma)(u_0)\Big| = \mathcal{O}(1) \cdot \sigma^3, \tag{64}$$

$$\lambda_i\Big(S_i(\sigma)(u_0),\ u_0\Big) = \lambda_i(u_0) + \frac{\sigma}{2} D\lambda_i(u_0) \cdot r_i(u_0) + \mathcal{O}(1) \cdot \sigma^2. \tag{65}$$

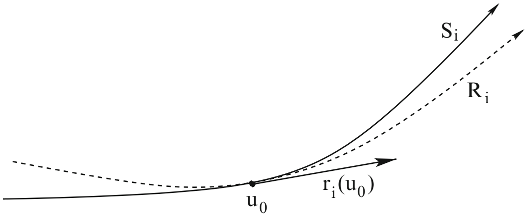


Fig. 24 The i-shock curve and the i-rarefaction curve through a point u_0

Here and throughout the following, the Landau symbol $\mathcal{O}(1)$ denotes a quantity whose absolute value satisfies a uniform bound, depending only on the system (51).

Toward the general solution of the Riemann problem (51)–(52), we first study three special cases.

1. Centered Rarefaction Waves. Let the i-th field be genuinely nonlinear, and assume that u^+ lies on the positive i-rarefaction curve through u^-, i.e. $u^+ = R_i(\sigma)(u^-)$ for some $\sigma > 0$. For each $s \in [0, \sigma]$, define the characteristic speed

$$\lambda_i(s) = \lambda_i\big(R_i(s)(u^-)\big).$$

Observe that, by genuine nonlinearity, the map $s \mapsto \lambda_i(s)$ is strictly increasing. Hence, for every $\lambda \in \big[\lambda_i(u^-),\ \lambda_i(u^+)\big]$, there is a unique value $s \in [0, \sigma]$ such that $\lambda = \lambda_i(s)$. For $t \geq 0$, we claim that the function

$$u(t,x) = \begin{cases} u^- & if \quad x/t < \lambda_i(u^-), \\ R_i(s)(u^-) & if \quad x/t = \lambda_i(s) \in \big[\lambda_i(u^-),\ \lambda_i(u^+)\big], \\ u^+ & if \quad x/t > \lambda_i(u^+), \end{cases} \tag{66}$$

is a piecewise smooth solution of the Riemann problem, continuous for $t > 0$. Indeed, from the definition it follows

$$\lim_{t \to 0+} \big\| u(t,\cdot) - \bar{u} \big\|_{\mathbf{L}^1} = 0.$$

Moreover, the (51) is trivially satisfied in the sectors where $x < t\lambda_i(u^-)$ or $x > t\lambda_i(u^+)$, because here $u_t = u_x = 0$. Next, assume $x = t\lambda_i(s)$ for some $s \in]0, \sigma[$. Since u is constant along each ray through the origin $\{x/t = c\}$, we have

$$u_t(t,x) + \frac{x}{t}\, u_x(t,x) = 0. \tag{67}$$

We now observe that the definition (66) implies $x/t = \lambda_i\big(u(t,x)\big)$. By construction, the vector u_x has the same direction as $r_i(u)$, hence it is an eigenvector of the Jacobian matrix $A(u) \doteq Df(u)$ with eigenvalue $\lambda_i(u)$. On the sector of the t-x plane where $\lambda_i(u^-) < x/t < \lambda_i(u^+)$ we thus have

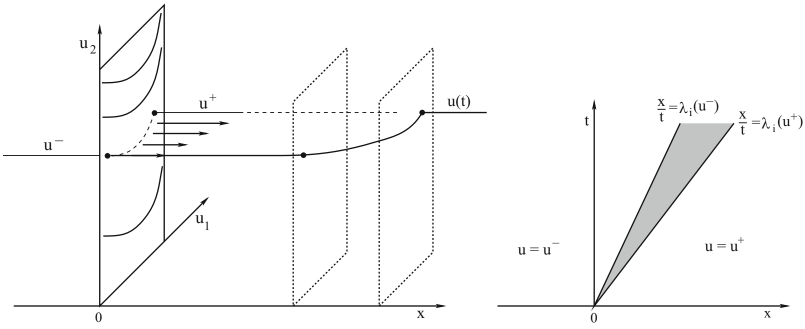


Fig. 25 A solution to the Riemann problem consisting of centered rarefaction wave. *Left*: the profile of the solution at a fixed time t, in the x-u space. *Right*: the values of u in the t-x plane

$$u_t + A(u)u_x \;=\; u_t + \lambda_i(u)u_x \;=\; 0\,,$$

proving our claim. As shown in Fig. 25, at a fixed time $t > 0$, the profile $x \mapsto u(t,x)$ is obtained as follows. Consider the rarefaction curve R_i joining u^- with u^+, on the hyperplane where $x = 0$. Move each point of this curve horizontally, in the amount $t\,\lambda_i(u)$. The new curve yields the graph of $u(t,\cdot)$. Notice that the assumption $\sigma > 0$ is essential for the validity of this construction. In the opposite case $\sigma < 0$, the definition (66) would yield a triple-valued function in the region where $x/t \in \big[\lambda_i(u^+)\,,\ \lambda_i(u^-)\big]$.

2. Shocks. Assume again that the i-th family is genuinely nonlinear and that the state u^+ is connected to the right of u^- by an i-shock, i.e. $u^+ = S_i(\sigma)(u^-)$. Then, calling $\lambda \doteq \lambda_i(u^+, u^-)$ the Rankine–Hugoniot speed of the shock, the function

$$u(t,x) \;=\; \begin{cases} u^- & if \qquad x < \lambda t, \\ u^+ & if \qquad x > \lambda t, \end{cases} \tag{68}$$

(Fig. 26) provides a piecewise constant solution to the Riemann problem. Observe that, if $\sigma < 0$, than this solution is entropy admissible in the sense of Lax. Indeed, since the speed is monotonically increasing along the shock curve, recalling (65) we have

$$\lambda_i(u^+) \;<\; \lambda_i(u^-, u^+) \;<\; \lambda_i(u^-). \tag{69}$$

Hence the Lax admissibility conditions (50) hold. In the case $\sigma > 0$, however, one has $\lambda_i(u^-) < \lambda_i(u^+)$ and the conditions (50) are violated.

3. Contact discontinuities. Assume that the i-th field is linearly degenerate and that the state u^+ lies on the i-th rarefaction curve through u^-, i.e. $u^+ = R_i(\sigma)(u^-)$ for some σ. By assumption, the i-th characteristic speed λ_i is constant along this curve. Choosing $\lambda = \lambda(u^-)$, the piecewise constant function (68) then provides a

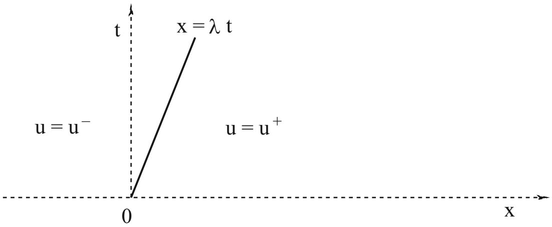


Fig. 26 A solution consisting of a single shock, or a contact discontinuity

solution to our Riemann problem. Indeed, the Rankine–Hugoniot conditions hold at the point of jump:

$$\begin{aligned} f(u^+) - f(u^-) = & \int_0^\sigma Df\ (R_i(s)(u^-))\ r_i\ (R_i(s)(u^-))\ ds \\ & = \int_0^\sigma \lambda(u^-)\, r_i\big(R_i(s)(u^-)\big)\, ds \ = \ \lambda_i(u^-)\cdot\Big(R_i(\sigma)(u^-) - u^-\Big). \end{aligned} \tag{70}$$

In this case, the Lax entropy condition holds regardless of the sign of σ. Indeed,

$$\lambda_i(u^+) \ = \ \lambda_i(u^-, u^+) \ = \ \lambda_i(u^-). \tag{71}$$

Observe that, according to (70), for linearly degenerate fields the shock and rarefaction curves actually coincide: $S_i(\sigma)(u_0) = R_i(\sigma)(u_0)$ for all σ.

The above results can be summarized as follows. For a fixed left state u^- and $i \in \{1, \ldots, n\}$ define the mixed curve

$$\Psi_i(\sigma)(u^-) \ = \ \begin{cases} R_i(\sigma)(u^-) & if \qquad \sigma \geq 0, \\ S_i(\sigma)(u^-) & if \qquad \sigma < 0. \end{cases} \tag{72}$$

In the special case where $u^+ = \Psi_i(\sigma)(u^-)$ for some σ, the Riemann problem can then be solved by an elementary wave: a rarefaction, a shock or a contact discontinuity.

3.4 General Solution of the Riemann Problem

Relying on the previous analysis, the solution of the general Riemann problem (51)–(52) can now be obtained by finding intermediate states $\omega_0 = u^-, \omega_1, \ldots, \omega_n = u^+$ such that each pair of adjacent states ω_{i-1}, ω_i can be connected by an elementary wave, i. e.

$$\omega_i \ = \ \Psi_i(\sigma_i)(\omega_{i-1}) \qquad\qquad i = 1, \ldots, n. \tag{73}$$

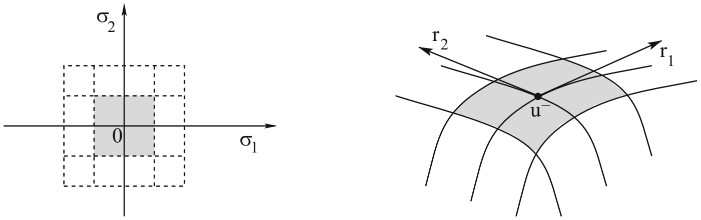


Fig. 27 The range of the map $(\sigma_1, \sigma_2) \mapsto \Psi_2(\sigma_2) \circ \Psi_1(\sigma_1)(u^-)$ covers a whole neighborhood of u^-

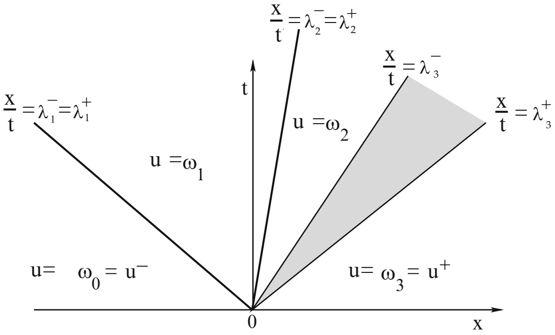


Fig. 28 A solution to the Riemann problem, consisting of a 1-shock, a 2-contact, and a 3-rarefaction

This can be done whenever u^+ is sufficiently close to u^-. Indeed, consider the map

$$\Lambda(\sigma_1, \dots, \sigma_n) \;=\; \Psi_n(\sigma_n) \circ \cdots \circ \Psi_1(\sigma_1)(u^-).$$

Taking a first order Taylor expansion at the point $(\sigma_1, \dots, \sigma_n) = (0, \dots, 0)$ we obtain the affine map

$$(\sigma_1, \dots, \sigma_n) \;\mapsto\; u^- + \sum_{i=1}^{n} \sigma_i r_i(u^-).$$

Since $\{r_1, \dots, r_n\}$ is a basis of the space $I\!R^n$, the above map has full rank (it is one-to-one and surjective). We can thus apply the implicit function theorem and conclude that the nonlinear mapping Λ is a continuous bijection of a neighborhood of the origin in $I\!R^n$ onto a neighborhood of u^- (Fig. 27).

Therefore, for u^+ sufficiently close to u^-, there exist unique wave strengths $\sigma_1, \dots \sigma_n$ such that

$$u^+ \;=\; \Psi_n(\sigma_n) \circ \cdots \circ \Psi_1(\sigma_1)(u^-). \tag{74}$$

In turn, these determine the intermediate states ω_i in (73). The complete solution is now obtained by piecing together the solutions of the n Riemann problems (Fig. 28)

$$u_t + f(u)_x = 0, \qquad u(0,x) = \begin{cases} \omega_{i-1} & if \quad x < 0, \\ \omega_i & if \quad x > 0, \end{cases} \tag{75}$$

on different sectors of the t-x plane. By construction, each of these problems has an entropy-admissible solution consisting of a simple wave of the i-th characteristic family. More precisely:

CASE 1: The i-th characteristic field is genuinely nonlinear and $\sigma_i > 0$. Then the solution of (75) consists of a centered rarefaction wave. The i-th characteristic speeds range over the interval $[\lambda_i^-,\ \lambda_i^+]$, defined as

$$\lambda_i^- \doteq \lambda_i(\omega_{i-1}), \qquad \lambda_i^+ \doteq \lambda_i(\omega_i).$$

CASE 2: Either the i-th characteristic field is genuinely nonlinear and $\sigma_i \leq 0$, or else the i-th characteristic field is linearly degenerate (with σ_i arbitrary). Then the solution of (75) consists of an admissible shock or a contact discontinuity, traveling with Rankine–Hugoniot speed

$$\lambda_i^- \doteq \lambda_i^+ \doteq \lambda_i(\omega_{i-1}, \omega_i).$$

The solution to the original problem (51)–(52) can now be constructed by piecing together the solutions of the n Riemann problems (75), $i = 1, \ldots, n$. Indeed, for $\sigma_1, \ldots, \sigma_n$ sufficiently small, the speeds λ_i^-, λ_i^+ introduced above remain close to the corresponding eigenvalues $\lambda_i(u^-)$ of the matrix $A(u^-)$. By strict hyperbolicity and continuity, we can thus assume that the intervals $[\lambda_i^-, \lambda_i^+]$ are disjoint, i.e.

$$\lambda_1^- \leq \lambda_1^+ < \lambda_2^- \leq \lambda_2^+ < \cdots < \lambda_n^- \leq \lambda_n^+.$$

Therefore, a piecewise smooth solution $u : [0,\infty) \times I\!R \mapsto I\!R^n$ is well defined by the assignment

$$u(t,x) = \begin{cases} u^- = \omega_0 & \text{if} \quad x/t \in \left]-\infty,\ \lambda_1^-\right[, \\ R_i(s)(\omega_{i-1}) & \text{if} \quad x/t = \lambda_i\big(R_i(s)(\omega_{i-1})\big) \in \left[\lambda_i^-,\ \lambda_i^+\right[, \\ \omega_i & \text{if} \quad x/t \in \left[\lambda_i^+,\ \lambda_{i+1}^-\right[, \\ u^+ = \omega_n & \text{if} \quad x/t \in \left[\lambda_n^+,\ \infty\right[. \end{cases} \tag{76}$$

Observe that this solution is self-similar, having the form $u(t,x) = \psi(x/t)$, with $\psi : I\!R \mapsto I\!R^n$ possibly discontinuous.

3.5 The Riemann Problem for the p-System

Example 12 (the p-system). Consider again the equations for isentropic gas dynamics (in Lagrangian coordinates)

$$\begin{cases} v_t - u_x = 0\,, \\ u_t + p(v)_x = 0. \end{cases} \tag{77}$$

Writing $U = (v, u)$, the Riemann problem takes the form

$$U(0,x) = \begin{cases} U^- = (v^-, u^-) & if \quad x < 0, \\ U^+ = (v^+, u^+) & if \quad x > 0. \end{cases} \tag{78}$$

Here u^-, u^+ are the velocities to the left and to the right of the initial jump, while $v^-, v^+ > 0$ are the specific volumes.

By (59), the 1-rarefaction curve through U^- is obtained by solving the Cauchy problem

$$\frac{du}{dv} = \sqrt{-p'(v)}, \qquad\qquad u(v^-) = u^-.$$

This yields the curve

$$R_1 = \Big\{(v,u); \quad u - u^- = \int_{v^-}^{v} \sqrt{-p'(y)}\, dy\Big\}. \tag{79}$$

Similarly, the 2-rarefaction curve through the point U^- is

$$R_2 = \Big\{(v,u); \quad u - u^- = -\int_{v^-}^{v} \sqrt{-p'(y)}\, dy\Big\}. \tag{80}$$

The shock curves S_1, S_2 through the left state U^- are obtained from the Rankine–Hugoniot conditions

$$\lambda(v - v^-) = -(u - u^-), \qquad\qquad \lambda(u - u^-) = p(v) - p(v^-). \tag{81}$$

One can use the first equation in (81) to obtain the shock speed λ. From the second equation, the shock curves are then computed as

$$S_1 = \left\{(v,u); \qquad -(u-u^-)^2 = (v - v^-)\big(p(v) - p(v^-)\big), \quad \lambda \doteq -\frac{u-u^-}{v-v^-} < 0\right\}, \tag{82}$$

$$S_2 = \left\{(v,u); \qquad -(u-u^-)^2 = (v - v^-)\big(p(v) - p(v^-)\big), \quad \lambda \doteq -\frac{u-u^-}{v-v^-} > 0\right\}. \tag{83}$$

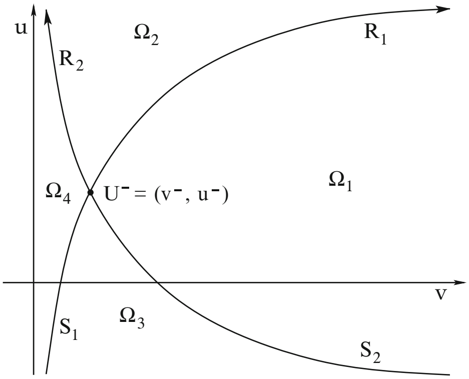


Fig. 29 Shocks and rarefaction curves through the point $U^- = (v^-, u^-)$

By (58)–(59) and the assumptions $p'(v) < 0$, $p''(v) > 0$, the directional derivatives of the eigenvalues λ_1, λ_2 in the direction of the corresponding eigenvectors r_1, r_2 are found to be

$$(D\lambda_1)r_1 \ = \ (D\lambda_2)r_2 \ = \ \frac{p''(v)}{2\sqrt{-p'(v)}} \ > \ 0. \tag{84}$$

Therefore, the Riemann problem (77)–(78) admits a solution in the form of a centered rarefaction wave provided that $U^+ \in R_1$, $v^+ > v^-$, or else $U^+ \in R_2$, $v^+ < v^-$. On the other hand, a shock connecting U^- with U^+ will be admissible if either $U^+ \in S_1$ and $v^+ < v^-$, or else $U^+ \in S_2$ and $v^+ > v^-$.

Taking the above admissibility conditions into account, we thus obtain four curves originating from the point $U^- = (v^-, u^-)$. Namely, the two rarefaction curves

$$\sigma \mapsto R_1(\sigma),\ R_2(\sigma) \qquad\qquad \sigma \geq 0,$$

and the two shock curves

$$\sigma \mapsto S_1(\sigma),\ S_2(\sigma) \qquad\qquad \sigma \leq 0.$$

In turn, these curves divide a neighborhood of U^- into four regions (Fig. 29):

$$\begin{array}{ll} \Omega_1: \text{ bounded by } R_1, S_2, & \qquad \Omega_2: \text{ bounded by } R_1, R_2, \\ \Omega_3: \text{ bounded by } S_1, S_2, & \qquad \Omega_4: \text{ bounded by } S_1, R_2. \end{array}$$

For $U^+ = (v^+, u^+)$ sufficiently close to $U^- = (v^-, u^-)$, the structure of the general solution to the Riemann problem is now determined by the location of the state U^+, with respect to the curves R_i, S_i (Fig. 30).

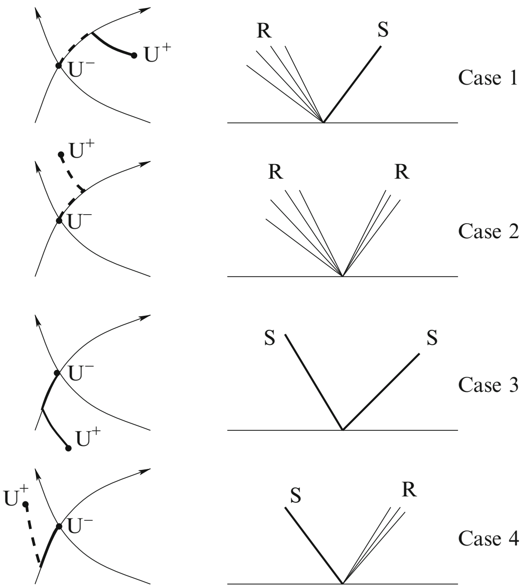


Fig. 30 Solution to the Riemann problem for the p-system. The four different cases

CASE 1: $U^+ \in \Omega_1$. The solution consists of a 1-rarefaction wave and a 2-shock.
CASE 2: $U^+ \in \Omega_2$. The solution consists of two centered rarefaction waves.
CASE 3: $U^+ \in \Omega_3$. The solution consists of two shocks.
CASE 4: $U^+ \in \Omega_4$. The solution consists of a 1-shock and a 2-rarefaction wave.

Remark 4. Consider a 2×2 strictly hyperbolic system of conservation laws. Assume that the i-th characteristic field is genuinely nonlinear. The relative position of the i-shock and the i-rarefaction curve through a point u_0 can be determined as follows (Fig. 24). Let $\sigma \mapsto R_i(\sigma)$ be the i-rarefaction curve, parameterized so that $\lambda_i\big(R_i(\sigma)\big) = \lambda_i(u_0) + \sigma$. Assume that, for some constant α, the point

$$S_i(\sigma) = R_i(\sigma) + \big(\alpha\sigma^3 + o(\sigma^3)\big)r_j(u_0) \tag{85}$$

lies on the i-shock curve through u_0, for all σ. Here the Landau symbol $o(\sigma^3)$ denotes a higher order infinitesimal, as $\sigma \to 0$. The wedge product of two vectors in $I\!R^2$ is defined as $\begin{pmatrix} a \\ b \end{pmatrix} \wedge \begin{pmatrix} c \\ d \end{pmatrix} \doteq ad - bc$. We then have

$$\begin{aligned} &\Psi(\sigma) \\ &\quad \doteq \Big[R_i(\sigma) + \big(\alpha\sigma^3 + o(\sigma^3)\big)r_j(u_0) - u_0\Big] \wedge \Big[f\Big(R_i(\sigma) + \big(\alpha\sigma^3 + o(\sigma^3)\big)r_j(u_0)\Big) - f(u_0)\Big] \\ &\quad \doteq A(\sigma) \wedge B(\sigma) \equiv 0. \end{aligned}$$

Indeed, the Rankine–Hugoniot equations imply that the vectors $A(\sigma)$ and $B(\sigma)$ are parallel. According to Leibnitz' rule, the fourth derivative is computed by

$$\frac{d^4}{d\sigma^4}\Psi = \left(\frac{d^4}{d\sigma^4}A\right)\wedge B + 4\left(\frac{d^3}{d\sigma^3}A\right)\wedge\left(\frac{d}{d\sigma}B\right) + 6\left(\frac{d^2}{d\sigma^2}A\right)\wedge\left(\frac{d^2}{d\sigma^2}B\right)$$
$$+4\left(\frac{d}{d\sigma}A\right)\wedge\left(\frac{d^3}{d\sigma^3}B\right) + A\wedge\left(\frac{d^4}{d\sigma^4}B\right)$$

By the choice of the parametrization, $\frac{d}{d\sigma}\lambda_i\big(R_i(\sigma)\big) \equiv 1$. Hence

$$\begin{aligned}
\frac{d}{d\sigma}f\big(R_i(\sigma)\big) &= \lambda_i\big(R_i(\sigma)\big)\frac{d}{d\sigma}R_i(\sigma)\,,\\
\frac{d^2}{d\sigma^2}f\big(R_i(\sigma)\big) &= \frac{d}{d\sigma}R_i(\sigma) + \lambda_i\big(R_i(\sigma)\big)\frac{d^2}{d\sigma^2}R_i(\sigma)\,,\\
\frac{d^3}{d\sigma^3}f\big(R_i(\sigma)\big) &= 2\frac{d^2}{d\sigma^2}R_i(\sigma) + \lambda_i\big(R_i(\sigma)\big)\frac{d^3}{d\sigma^3}R_i(\sigma).
\end{aligned}$$

For convenience, we write $r_i \bullet r_j \doteq (Dr_j)r_i$ to denote the directional derivative of r_j in the direction of r_i. At $\sigma = 0$ we have

$$A = B = 0\,, \qquad \frac{d}{d\sigma}R_i = r_i(u_0)\,, \qquad \frac{d^2}{d\sigma^2}R_i = (r_i \bullet r_i)(u_0).$$

Using the above identities and the fact that the wedge product is anti-symmetric, we conclude

$$\frac{d^4}{d\sigma^4}\Psi\Big|_{\sigma=0} = 4\left(\frac{d^3}{d\sigma^3}R_i + 6\alpha r_j\right)\wedge\left(\lambda_i\frac{d}{d\sigma}R_i\right) + 6\left(\frac{d^2}{d\sigma^2}R_i\right)\wedge\left(\frac{d}{d\sigma}R_i + \lambda_i\frac{d^2}{d\sigma^2}R_i\right)$$
$$+4\left(\frac{d}{d\sigma}R_i\right)\wedge\left(2\frac{d^2}{d\sigma^2}R_i + \lambda_i\frac{d^3}{d\sigma^3}R_i + 6\alpha\lambda_j r_j\right)$$
$$= 24\alpha(\lambda_i - \lambda_j)(r_j \wedge r_i) - 2(r_i \bullet r_i)\wedge r_i = 0.$$

The i-shock curve through u_0 is thus traced by points $S_i(\sigma)$ at (85), with

$$\alpha = \frac{(r_i \bullet r_i)\wedge r_i}{12(\lambda_i - \lambda_j)(r_j \wedge r_i)}. \tag{86}$$

The sign of α in (86) gives the position of the i-shock curve, relative to the i-rarefaction curve, near the point u_0. In particular, if $(r_i \bullet r_i) \wedge r_i \neq 0$, it is clear that these two curves do not coincide.

3.6 Error and Interaction Estimates

In this final section we provide two types of estimates, which will play a key role in the analysis of front tracking approximations.

Fix a left state u^-, a right state u^+, and a speed λ. If these satisfy the Rankine–Hugoniot equations, we have

$$\lambda(u^+ - u^-) - [f(u^+) - f(u^-)] \;=\; 0.$$

On the other hand, if these values are chosen arbitrarily, the only available estimate is

$$\lambda(u^+ - u^-) - [f(u^+) - f(u^-)] \;=\; \mathscr{O}(1) \cdot |u^+ - u^-|. \tag{87}$$

Here an throughout the sequel, the Landau symbol $\mathscr{O}(1)$ denotes a quantity which remains uniformly bounded as all variables $u^-, u^+, \lambda, \sigma \ldots$ range on bounded sets. The next lemma describes by how much the Rankine–Hugoniot equation fail to be satisfied, if the point u^+ lies on the i-rarefaction curve through u^- and we choose λ to be the i-th characteristic speed at the point u^-.

Lemma 3 (error estimate). *For $\sigma > 0$ small, one has the estimate*

$$\lambda_k(u^-)\Big[R_k(\sigma)(u^-) - u^-\Big] - \Big[f\big(R_k(\sigma)(u^-)\big) - f(u^-)\Big] \;=\; \mathscr{O}(1) \cdot \sigma^2. \tag{88}$$

Proof. Call $E(\sigma)$ the left hand side of (88). Clearly $E(0) = 0$. Differentiating w.r.t. σ at the point $\sigma = 0$ and recalling that $dR_k/d\sigma = r_k$, we find

$$\left.\frac{dE}{d\sigma}\right|_{\sigma=0} = \lambda_k(u^-) r_k(u^-) - Df(u^-) r_k(u^-) = 0.$$

Since E varies smoothly with u^- and σ, the estimate (88) follows by Taylor's formula. □

Next, consider a left state u^l, a middle state u^m and a right state u^r (Fig. 31, left). Assume that the pair (u^l, u^m) is connected by a j-wave of strength σ', while the pair (u^m, u^r) is connected by an i-wave of strength σ'', with $i < j$. We are interested in the strength of the waves $(\sigma_1, \ldots, \sigma_n)$ in the solution of the Riemann problem where $u^- = u^l$ and $u^+ = u^r$. Roughly speaking, these are the waves determined by the

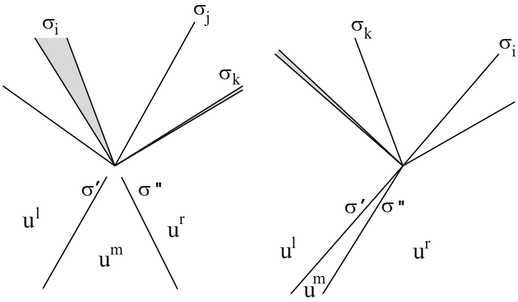


Fig. 31 Wave interactions. Strengths of the incoming and outgoing waves

interaction of the σ' and σ''. The next lemma shows that $\sigma_i \approx \sigma''$, $\sigma_j \approx \sigma'$ while $\sigma_k \approx 0$ for $k \neq i, j$.

A different type of interaction is considered in Fig. 31, right. Here the pair (u^l, u^m) is connected by an i-wave of strength σ', while the pair (u^m, u^r) is connected by a second i-wave, say of strength σ''. In this case, the strengths $(\sigma_1, \ldots, \sigma_n)$ of the outgoing waves satisfy $\sigma_i \approx \sigma' + \sigma''$ while $\sigma_k \approx 0$ for $k \neq i$. As usual, $\mathcal{O}(1)$ will denote a quantity which remains uniformly bounded as u^-, σ', σ'' range on bounded sets.

Lemma 4 (interaction estimates). *Consider the Riemann problem (51)–(52).*

(i) Recalling (72), assume that the right state is given by

$$u^+ = \Psi_i(\sigma'') \circ \Psi_j(\sigma')(u^-). \tag{89}$$

Let the solution consist of waves of size $(\sigma_1, \ldots, \sigma_n)$*, as in (74). Then*

$$|\sigma_i - \sigma''| + |\sigma_j - \sigma'| + \sum_{k \neq i,j} |\sigma_k| \;=\; \mathcal{O}(1) \cdot |\sigma'\sigma''|. \tag{90}$$

(ii) Next, assume that the right state is given by

$$u^+ = \Psi_i(\sigma'') \circ \Psi_i(\sigma')(u^-), \tag{91}$$

Then the waves $(\sigma_1, \ldots, \sigma_n)$ *in the solution of the Riemann problem are estimated by*

$$|\sigma_i - \sigma' - \sigma''| + \sum_{k \neq i} |\sigma_k| \;=\; \mathcal{O}(1) \cdot |\sigma'\sigma''|\big(|\sigma'| + |\sigma''|\big). \tag{92}$$

For a proof we refer to [11].

4 Global Solutions to the Cauchy Problem

In this chapter we study the global existence of weak solutions to the general Cauchy problem

$$u_t + f(u)_x = 0, \tag{93}$$

$$u(0,x) = \bar{u}(x). \tag{94}$$

Here the flux function $f : I\!R^n \mapsto I\!R^n$ is smooth, defined on a neighborhood of the origin. We always assume that the system is strictly hyperbolic, and that the assumption (H) introduced in the previous chapter holds.

A fundamental result proved by Glimm [34] provides the global existence of an entropy weak solution, for all initial data with suitably small total variation.

Theorem 2 (Global existence of weak solutions). *Assume that the system (93) is strictly hyperbolic, and that each characteristic field is either linearly degenerate or genuinely nonlinear.*

Then there exists a constant $\delta_0 > 0$ such that, for every initial condition $\bar{u} \in \boldsymbol{L}^1(I\!R;\ I\!R^n)$ with

$$Tot.Var.\{\bar{u}\} \leq \delta_0\,, \tag{95}$$

the Cauchy problem (93)–(94) has a weak solution $u = u(t,x)$ defined for all $t \geq 0$.

In addition, one can prove the existence of a global solution satisfying all the admissibility conditions introduced in Sect. 2.3. A proof of Theorem 2 requires two main steps:

(a) Construct a sequence of approximate solutions u_ν.
(b) Show that a subsequence converges in L^1_{loc} to a weak solution u of the Cauchy problem.

Approximate solutions can be constructed by piecing together solutions to several Riemann problems. Two techniques have been developed in the literature:

- In the *Glimm scheme* (Fig. 40) one considers a fixed grid of points $(t_j, x_k) = (j\,\Delta t\,,\ k\,\Delta x)$ in the t-x plane, and solves a Riemann problem at each node of the grid.
- In a *front tracking approximation*, one constructs a piecewise constant approximate solution $u = u(t,x)$, whose jumps are located along a finite number of segments in the t-x plane (Fig. 33). A new Riemann problem is solved at each point where two fronts interact. These points depend on the particular solution being constructed.

Having constructed a sequence of approximate solutions $(u_\nu)_{\nu\geq 1}$ (Fig. 32), one needs to extract a subsequence converging to some limit $u = u(t,x)$ in $\mathbf{L}^1_{loc}$. By Helly's compactness theorem, this can be achieved by establishing an a priori bound on the total variation Tot.Var.$\{u_\nu(t,\cdot)\}$, uniformly valid for $t > 0$ and $\nu \geq 1$.

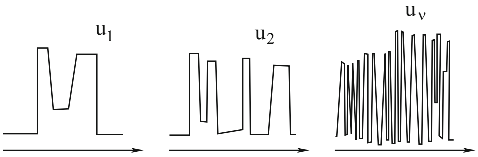


Fig. 32 Without a bound on the total variation, a sequence of approximate solutions may oscillate more and more, without admitting any convergent subsequence

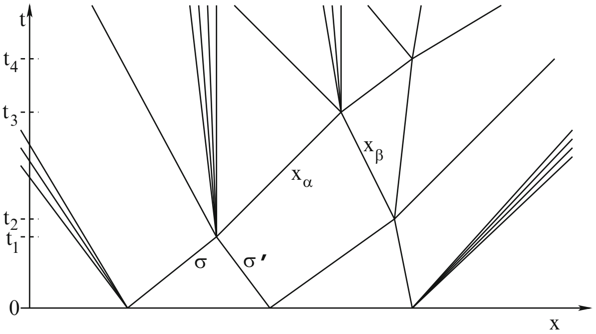


Fig. 33 An approximate solution obtained by front tracking

4.1 *Front Tracking Approximations*

In this section we describe the construction of front tracking approximations. This method was developed in [26, 28], and in [9] respectively for scalar conservation laws, for 2 × 2 systems, and for general $n \times n$ systems satisfying the assumptions (H). Further versions of this algorithm can also be found in [5,37,55]. An extension to fully general $n \times n$ systems, without the assumptions (H), is provided in [3].

Let the initial condition $\bar{u}$ be given and fix $\varepsilon > 0$. We now describe an algorithm which produces a piecewise constant approximate solution to the Cauchy problem (93)–(94). The construction (Fig. 33) starts at time $t = 0$ by taking a piecewise constant approximation $u(0,\cdot)$ of $\bar{u}$, such that

$$\text{Tot.Var.}\big\{u(0,\cdot)\big\} \leq \text{Tot.Var.}\{\bar{u}\}, \qquad \int \big|u(0,x) - \bar{u}(x)\big|\, dx \leq \varepsilon. \tag{96}$$

Let $x_1 < \cdots < x_N$ be the points where $u(0,\cdot)$ is discontinuous. For each $\alpha = 1,\ldots,N$, the Riemann problem generated by the jump $\big(u(0,x_\alpha-),\ u(0,x_\alpha+)\big)$ is approximately solved on a forward neighborhood of $(0,x_\alpha)$ in the t-x plane by a piecewise constant function, according to the following procedure.

Accurate Riemann Solver. Consider the general Riemann problem at a point $(\bar{t},\bar{x})$,

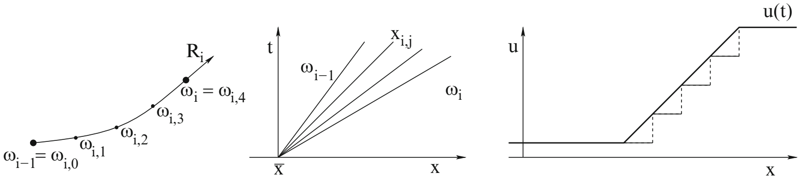


Fig. 34 Replacing a centered rarefaction wave by a rarefaction fan

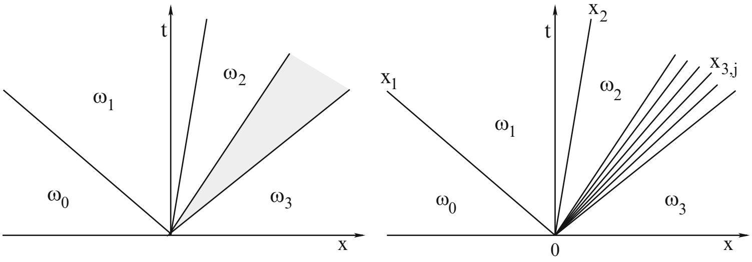


Fig. 35 *Left*: the exact solution to a Riemann problem. *Right*: a piecewise constant approximation. The centered rarefaction wave of the 3-d family has been replaced by a rarefaction fan

$$v_t + f(v)_x = 0, \qquad v(\bar t, x) = \begin{cases} u^- & if\ x < \bar x, \\ u^+ & if\ x > \bar x, \end{cases} \tag{97}$$

Recalling (72), let $\omega_0, \dots, \omega_n$ be the intermediate states and $\sigma_1, \dots, \sigma_n$ be the strengths of the waves in the solution, so that

$$\omega_0 = u^-, \;\; \omega_n = u^+, \qquad \omega_i = \Psi_i(\sigma_i)(\omega_{i-1}) \qquad i = 1, \dots, n. \tag{98}$$

If all jumps (ω_{i-1}, ω_i) were shocks or contact discontinuities, then this solution would be already piecewise constant. In general, the exact solution of (97) is not piecewise constant, because of the presence of centered rarefaction waves. These will be approximated by piecewise constant rarefaction fans, inserting additional states $\omega_{i,j}$ as follows.

If the i-th characteristic field is genuinely nonlinear and $\sigma_i > 0$, we divide the centered i-rarefaction into a number p_i of smaller i-waves, each with strength σ_i/p_i. Here we choose the integer p_i big enough so that $\sigma/p_i < \varepsilon$. For $j = 1, \dots, p_i$, we now define the intermediate states and wave-fronts (Fig. 34)

$$\omega_{i,j} = R_i(j\sigma_i/p_i)(\omega_{i-1}), \qquad x_{i,j}(t) = \bar x + (t - \bar t)\lambda_i(\omega_{i,j-1}). \tag{99}$$

Replacing each centered rarefaction wave with a rarefaction fan, we thus obtain a piecewise constant approximate solution to the Riemann problem (Fig. 35).

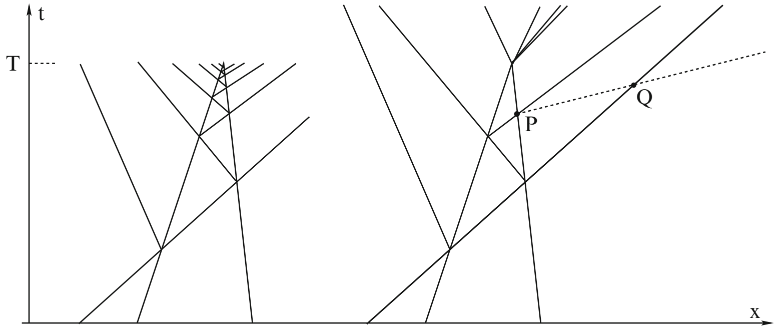


Fig. 36 *Left*: the number of wave fronts can become infinite in finite time. *Right*: by using the simplified Riemann solver at two interaction points P and Q, the total number of fronts remains bounded

We now resume the construction of a front tracking solution to the original Cauchy problem (93)–(94). Having solved all the Riemann problems at time $t = 0$, the approximate solution u can be prolonged until a first time t_1 is reached, when two wave-fronts interact (Fig. 33). Since $u(t_1, \cdot)$ is still a piecewise constant function, the corresponding Riemann problems can again be approximately solved within the class of piecewise constant functions. The solution u is then continued up to a time t_2 where a second interaction takes place, etc... We remark that, by an arbitrary small change in the speed of one of the wave fronts, it is not restrictive to assume that at most two incoming fronts collide, at each given time $t > 0$. This will considerably simplify all subsequent analysis, since we don't need to consider the case where three or more incoming fronts meet together.

The above construction can be continued for all times $t > 0$, as long as

(a) The total variation Tot.Var.$\{u(t, \cdot)\}$ remains small enough. This guarantees that all jumps $u(t, x-), u(t, x+)$ are small, hence the corresponding Riemann problems admit a solution.

(b) The total number of fronts remains finite.

Bounds on the total variations will be discussed in the next section. Here we observe that a naive implementation of the front tracking algorithm can produce an infinite number of fronts within finite time (Fig. 36).

As shown in [9], this can be avoided by occasionally implementing a *Simplified Riemann Solver*, which introduces one single additional front (Fig. 37). In this case, the solution is continued by means of two outgoing fronts of exactly the same strength as the incoming one. All other waves resulting from the interaction are lumped together in a single front, traveling with a constant speed $\hat{\lambda}$, strictly larger than all characteristic speeds.

In the end, for a given $\varepsilon > 0$, this modified front tracking algorithm generates a piecewise constant ε-approximate solution $u = u(t, x)$, defined as follows.

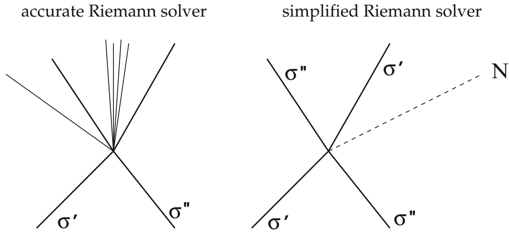


Fig. 37 *Left*: the solution to a Riemann problem obtained by the Accurate Riemann Solver introduces several new wave fronts. *Right*: the Simplified Riemann solver produces two outgoing fronts of the same strength as the incoming ones, plus a small *Non-Physical* front

Definition 6 (front tracking approximate solution). A piecewise constant function $u = u(t, x)$, defined for $t \geq 0$, $x \in R$, is called an *ε-approximate front tracking solution* to the Cauchy problem (93)–(94) provided that

(i) The initial condition is approximately attained, namely $\|u(0, \cdot) - \bar{u}\|_{\mathbf{L}^1} \leq \varepsilon$.
(ii) All shock fronts and all contact discontinuities satisfy the Rankine–Hugoniot equations, as well as the admissibility conditions.
(iii) Each rarefaction front has strength $\leq \varepsilon$.
(iv) At each time $t > 0$, the total strength of all non-physical fronts in $u(t, \cdot)$ is $\leq \varepsilon$.
(v) The total variation of $u(t, \cdot)$ satisfies a uniform bound, depending only on Tot.Var.$\{\bar{u}\}$.

– By a *shock front* we mean a jump whose right and left states satisfy $u^+ = S_i(\sigma)(u^-)$ for some $\sigma \in I\!R$ and $i \in \{1, \ldots, n\}$. This travels with Rankine–Hugoniot speed $\lambda = \lambda_i(u^-, u^+) = \frac{f(u^+)-f(u^-)}{u^+-u^-}$.
– By a *rarefaction front* we mean a jump whose right and left states satisfy $u^+ = R_i(\sigma)(u^-)$ for some σ, i. This travels with speed $\lambda = \lambda_i(u^+)$, i.e. with the characteristic speed of its right state.
– By a *non-physical front* we mean a jump whose right and left states u^+, u^- are arbitrary. This travels with a fixed speed $\hat{\lambda}$, strictly greater than all characteristic speeds.

4.2 Bounds on the Total Variation

In this section we derive bounds on the total variation of a front tracking approximation $u(t, \cdot)$, uniformly valid for all $t \geq 0$. These estimates will be obtained from Lemma 4, using an interaction functional.

We begin by introducing some notation. At a fixed time t, let x_α, $\alpha = 1, \ldots, N$, be the locations of the fronts in $u(t, \cdot)$. Moreover, let $|\sigma_\alpha|$ be the strength of the wave-front at x_α, say of the family $k_\alpha \in \{1, \ldots, n\}$. Following [34], consider the two functionals

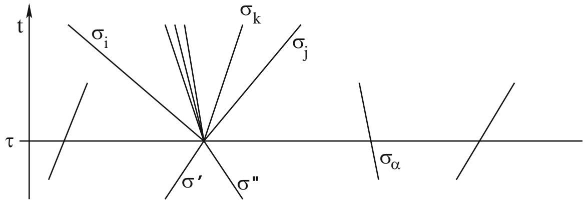


Fig. 38 Estimating the change in the total variation at a time where two fronts interact

$$V(t) \doteq V\big(u(t)\big) \doteq \sum_{\alpha} |\sigma_\alpha|\,, \tag{100}$$

measuring the *total strength of waves* in $u(t,\cdot)$, and

$$Q(t) \doteq Q\big(u(t)\big) \doteq \sum_{(\alpha,\beta)\in\mathscr{A}} |\sigma_\alpha\sigma_\beta|\,, \tag{101}$$

measuring the *wave interaction potential*. In (101), the summation ranges over the set $\mathscr{A}$ of all couples of approaching wave-fronts:

Definition 7 (approaching fronts). Two fronts, located at points $x_\alpha < x_\beta$ and belonging to the characteristic families $k_\alpha, k_\beta \in \{1,\dots,n\}$ respectively, are *approaching* if $k_\alpha > k_\beta$ or else if $k_\alpha = k_\beta$ and at least one of the wave-fronts is a shock of a genuinely nonlinear family.

Roughly speaking, two fronts are approaching if the one behind has the larger speed (and hence it can collide with the other, at some future time).

Now consider the approximate solution $u = u(t,x)$ constructed by the front tracking algorithm. It is clear that the quantities $V\big(u(t)\big)$, $Q\big(u(t)\big)$ remain constant except at times where an interaction occurs. At a time τ where two fronts of strength $|\sigma'|, |\sigma''|$ collide, the interaction estimates (90) or (92) yield

$$\Delta V(\tau) \doteq V(\tau+) - V(\tau-) = \mathscr{O}(1)\cdot|\sigma'\sigma''|, \tag{102}$$

$$\Delta Q(\tau) \doteq Q(\tau+) - Q(\tau-) = -|\sigma'\sigma''| + \mathscr{O}(1)\cdot|\sigma'\sigma''|\cdot V(\tau-). \tag{103}$$

Indeed (Fig. 38), after time τ the two colliding fronts σ', σ'' are no longer approaching. Hence the product $|\sigma'\sigma''|$ is no longer counted within the summation (101). On the other hand, the new waves emerging from the interaction (having strength $\mathscr{O}(1)\cdot|\sigma'\sigma''|$) can approach all the other fronts not involved in the interaction (which have total strength $\leq V(\tau-)$).

If V remains sufficiently small, so that $\mathscr{O}(1)\cdot V(\tau-) \leq 1/2$, from (103) it follows

$$Q(\tau+) - Q(\tau-) \leq -\frac{|\sigma'\sigma''|}{2}. \tag{104}$$

By (102) and (104) we can thus choose a constant C_0 large enough so that the quantity

$$\Upsilon(t) \doteq V(t) + C_0 Q(t)$$

decreases at every interaction time, provided that V remains sufficiently small.

We now observe that the total strength of waves is an equivalent way of measuring the total variation. Indeed, for some constant C one has

$$\text{Tot.Var.}\big\{u(t)\big\} \leq V\big(u(t)\big) \leq C \cdot \text{Tot.Var.}\big\{u(t)\big\}. \tag{105}$$

Moreover, the definitions (100)–(101) trivially imply $Q \leq V^2$. If the total variation of the initial data $u(0,\cdot)$ is sufficiently small, the previous estimates show that the quantity $V + C_0 Q$ is nonincreasing in time. Therefore

$$\text{Tot.Var.}\big\{u(t)\big\} \leq V\big(u(t)\big) \leq V\big(u(0)\big) + C_0 Q\big(u(0)\big). \tag{106}$$

This provides a uniform bound on the total variation of $u(t,\cdot)$ valid for all times $t \geq 0$.

An important consequence of the bound (106) is that, at every time τ where two fronts interact, the corresponding Riemann problem can always be solved. Indeed, the left and right states differ by the quantity

$$|u^+ - u^-| \leq \text{Tot.Var.}\big\{u(\tau)\big\},$$

which remains small.

Another consequence of the bound on the total variation is the continuity of $t \mapsto u(t,\cdot)$ as a function with values in $\mathbf{L}^1_{\text{loc}}$. More precisely, there exists a Lipschitz constant L' such that

$$\int_{-\infty}^{\infty} \big|u(t,x) - u(t',x)\big|\, dx \leq L'|t-t'| \qquad \text{for all} \quad t,t' \geq 0. \tag{107}$$

Indeed, if no interaction occurs inside the interval $[t,t']$, the left hand side of (107) can be estimated simply as

$$\begin{aligned} \big\|u(t) - u(t')\big\|_{\mathbf{L}^1} &\leq |t-t'| \textstyle\sum_\alpha |\sigma_\alpha|\, |\dot{x}_\alpha| \\ &\leq |t-t'| \cdot [\text{total strength of all wave fronts}] \cdot [\text{maximum speed}] \\ &\leq L' \cdot |t-t'|, \end{aligned} \tag{108}$$

for some uniform constant L'. The case where one or more interactions take place within $[t,t']$ is handled in the same way, observing that the map $t \mapsto u(t,\cdot)$ is continuous across interaction times.

4.3 Convergence to a Limit Solution

Given any sequence $\varepsilon_\nu \to 0+$, by the front tracking algorithm we obtain a sequence of piecewise constant functions u_ν, where each $u_n u$ is an ε_ν-approximate solution to the Cauchy problem (93)–(94).

By (107) the maps $t \mapsto u_\nu(t,\cdot)$ are uniformly Lipschitz continuous w.r.t. the $\mathbf{L}^1$ distance. We can thus apply Helly's compactness theorem (see Theorem A.1 in the Appendix) and extract a subsequence which converges to some limit function u in $\mathbf{L}^1_{loc}$, also satisfying (107).

By the second relation in (96), as $\varepsilon_\nu \to 0$ we have $u_\nu(0) \to \bar{u}$ in $\mathbf{L}^1_{\text{loc}}$. Hence the initial condition (94) is clearly attained. To prove that u is a weak solution of the Cauchy problem, it remains to show that, for every $\phi \in \mathscr{C}^1_c$ with compact support contained in the open half plane where $t > 0$, one has

$$\int_0^\infty \int_{-\infty}^\infty \phi_t(t,x)u(t,x) + \phi_x(t,x)f(u(t,x))\,dxdt = 0. \tag{109}$$

Since the u_ν are uniformly bounded and f is uniformly continuous on bounded sets, it suffices to prove that

$$\lim_{\nu\to 0} \int_0^\infty \int_{-\infty}^\infty \Big\{\phi_t(t,x)u_\nu(t,x) + \phi_x(t,x)f\big(u_\nu(t,x)\big)\Big\}\,dxdt = 0. \tag{110}$$

Choose $T > 0$ such that $\phi(t,x) = 0$ whenever $t \notin [0,T]$. For a fixed ν, at any time t call $x_1(t) < \cdots < x_N(t)$ the points where $u_\nu(t,\cdot)$ has a jump, and set

$$\Delta u_\nu(t,x_\alpha) \doteq u_\nu(t,x_\alpha+) - u_\nu(t,x_\alpha-),$$

$$\Delta f\big(u_\nu(t,x_\alpha)\big) \doteq f\big(u_\nu(t,x_\alpha+)\big) - f\big(u_\nu(t,x_\alpha-)\big).$$

Observe that the polygonal lines $x = x_\alpha(t)$ subdivide the strip $[0,T] \times I\!R$ into finitely many regions Γ_j where u_ν is constant (Fig. 39). Introducing the vector

$$\Phi \doteq \big(\phi \cdot u_\nu\,,\ \phi \cdot f(u_\nu)\big),$$

by the divergence theorem the double integral in (110) can be written as

$$\sum_j \iint_{\Gamma_j} \text{div}\,\Phi(t,x)\,dtdx = \sum_j \int_{\partial\Gamma_j} \Phi \cdot \mathbf{n}\,d\sigma. \tag{111}$$

Here $\partial\Gamma_j$ is the oriented boundary of Γ_j, while $\mathbf{n}$ denotes an outer normal. Observe that $\mathbf{n}d\sigma = \pm(\dot{x}_\alpha, -1)dt$ along each polygonal line $x = x_\alpha(t)$, while $\phi(t,x) = 0$ along the lines $t = 0, t = T$. By (111) the expression within square brackets in (110) is computed by

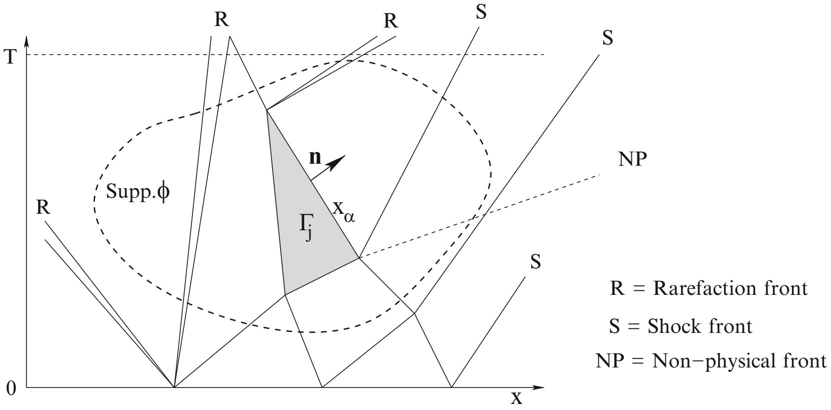


Fig. 39 Estimating the error in an approximate solution obtained by front tracking

$$\int_0^T \sum_\alpha \Big[\dot{x}_\alpha(t)\cdot \Delta u_\nu(t,x_\alpha) - \Delta f\big(u_\nu(t,x_\alpha)\big)\Big]\phi\big(t,x_\alpha(t)\big)\,dt. \tag{112}$$

Here, for each $t \in [0,T]$, the sum ranges over all fronts of $u_\nu(t,\cdot)$. To estimate the above integral, let σ_α be the signed strength of the wave-front at x_α. If this wave is a shock or or contact discontinuity, by construction the Rankine–Hugoniot equations are satisfied exactly, i.e.

$$\dot{x}_\alpha(t)\cdot \Delta u_\nu(t,x_\alpha) - \Delta f\big(u_\nu(t,x_\alpha)\big) = 0. \tag{113}$$

On the other hand, if the wave at x_α is a rarefaction front, its strength will satisfy $\sigma_\alpha \in]0,\varepsilon_\nu[$. Therefore, the error estimate (88) yields

$$\Big|\dot{x}_\alpha(t)\cdot \Delta u_\nu(t,x_\alpha) - \Delta f\big(u_\nu(t,x_\alpha)\big)\Big| \;=\; \mathscr{O}(1)\cdot|\sigma_\alpha|^2 \;=\; \mathscr{O}(1)\cdot\varepsilon_\nu|\sigma_\alpha|. \tag{114}$$

Finally, if the jump at x_α is a non-physical front of strength $|\sigma_\alpha| \doteq |u_\nu(x_\alpha+) - u_\nu(x_\alpha-)|$, by (87) we have the estimate

$$\Big|\dot{x}_\alpha(t)\cdot \Delta u_\nu(t,x_\alpha) - \Delta f\big(u_\nu(t,x_\alpha)\big)\Big| \;=\; \mathscr{O}(1)\cdot|\sigma_\alpha|. \tag{115}$$

Summing over all wave-fronts and recalling that the total strength of waves in $u_\nu(t,\cdot)$ satisfies a uniform bound independent of t,ν, we obtain

$$\begin{aligned}
&\limsup_{\nu\to\infty}\left|\sum_{\alpha}\Big[\dot{x}_\alpha(t)\cdot\Delta u_\nu(t,x_\alpha)-\Delta f\big(u_\nu(t,x_\alpha)\big)\Big]\phi\big(t,x_\alpha(t)\big)\right| \\
&\quad\le\ \Big(\max_{t,x}\big|\phi(t,x)\big|\Big)\cdot\limsup_{\nu\to\infty}\left\{\mathcal{O}(1)\cdot\sum_{\alpha\in\mathcal{R}}\varepsilon_\nu|\sigma_\alpha|+\mathcal{O}(1)\cdot\sum_{\alpha\in\mathcal{N}\mathcal{P}}|\sigma_\alpha|\right\} \\
&\quad=\ 0.
\end{aligned}\tag{116}$$

The limit (110) is now a consequence of (116). This shows that u is a weak solution to the Cauchy problem. For all details we refer to [11].

5 The Glimm Scheme

The fundamental paper of Glimm [34] contained the first rigorous proof of existence of global weak solutions to hyperbolic systems of conservation laws. For several years, the Glimm approximation scheme has provided the foundation for most of the theoretical results on the subject. We shall now describe this algorithm in a somewhat simplified setting, for systems where all characteristic speeds remain inside the interval $[0, 1]$. This is not a restrictive assumption. Indeed, consider any hyperbolic system of the form

$$u_t + A(u)u_x = 0,$$

and assume that all eigenvalues of A remain inside the interval $[-M, M]$. Performing the linear change of independent variables

$$y = x + Mt, \qquad \tau = 2Mt,$$

we obtain a new system

$$u_\tau + A^*(u)u_y = 0, \qquad A^*(u) \doteq \frac{1}{2M}A(u) + \frac{1}{2}I$$

where all eigenvalues of the matrix A^* now lie inside the interval $[0, 1]$.

To construct an approximate solution to the Cauchy problem

$$u_t + f(u)_x = 0, \qquad u(0,x) = \bar{u}(x), \tag{117}$$

we start with a grid in the t-x plane having step size $\Delta t = \Delta x$, with nodes at the points

$$P_{jk} = (t_j, x_k) \doteq (j\Delta t,\ k\Delta x) \qquad j,k \in \mathbb{Z}.$$

Moreover, we shall need a sequence of real numbers $\theta_1, \theta_2, \theta_3, \ldots$ *uniformly distributed* over the interval $[0,1]$. This means that, for every $\lambda \in [0,1]$, the percentage of points θ_i, $1 \leq i \leq N$ which fall inside $[0,\lambda]$ should approach λ as $N \to \infty$, i.e.:

$$\lim_{N\to\infty} \frac{\#\big\{j\ ;\ 1 \leq j \leq N,\ \ \theta_j \in [0,\lambda]\big\}}{N} = \lambda \qquad \text{for each } \lambda \in [0,1]. \qquad (118)$$

By $\# I$ we denote here the cardinality of a set I.

At time $t = 0$, the Glimm algorithm starts by taking an approximation of the initial data $\bar{u}$, which is constant on each interval of the form $\big]x_{k-1},\ x_k\big[$, and has jumps only at the nodal points $x_k \doteq k\,\Delta x$. To fix the ideas, we shall take

$$u(0,x) = \bar{u}(x_k) \qquad \text{for all } \ x \in [x_k,\ x_{k+1}[. \qquad (119)$$

For times $t > 0$ sufficiently small, the solution is then obtained by solving the Riemann problems corresponding to the jumps of the initial approximation $u(0,\cdot)$ at the nodes x_k. Since by assumption all waves speeds are contained in $[0,1]$, waves generated from different nodes remain separated at least until the time $t_1 = \Delta t$. The solution can thus be prolonged on the whole time interval $[0,\ \Delta t[$. For bigger times, waves emerging from different nodes may cross each other, and the solution would become extremely complicated. To prevent this, a restarting procedure is adopted. Namely, at time $t_1 = \Delta t$ the function $u(t_1-,\cdot)$ is approximated by a new function $u(t_1+,\cdot)$ which is piecewise constant, having jumps exactly at the nodes $x_k \doteq k\,\Delta x$. Our approximate solution u can now be constructed on the further time interval $[\Delta t,\ 2\Delta t[$, again by piecing together the solutions of the various Riemann problems determined by the jumps at the nodal points x_k. At time $t_2 = 2\Delta t$, this solution is again approximated by a piecewise constant function, etc…

A key aspect of the construction is the restarting procedure. At each time $t_j \doteq j\,\Delta t$, we need to approximate $u(t_j-,\cdot)$ with a a piecewise constant function $u(t_j+,\cdot)$, having jumps precisely at the nodal points x_k. This is achieved by a random sampling technique. More precisely, we look at the number θ_j in our uniformly distributed sequence. On each interval $[x_{k-1},\ x_k[$, the old value of our solution at the intermediate point $x_k^* = \theta_j x_k + (1-\theta_j)x_{k-1}$ becomes the new value over the whole interval:

$$u(t_j+,\ x) = u\big(t_j-,\ \ \theta_j x_k + (1-\theta_j)x_{k-1}\big) \qquad \text{for all } \ x \in [x_{k-1},\ x_k[. \quad (120)$$

An approximate solution constructed in this way is shown in Fig. 40. The asterisks mark the points where the function is sampled. For sake of illustration, we choose $\theta_1 = 1/2$, $\theta_2 = 1/3$.

For a strictly hyperbolic system of conservation laws, satisfying the hypotheses (H) in Sect. 3, the fundamental results of J. Glimm [34] and T.P. Liu [46] have established that

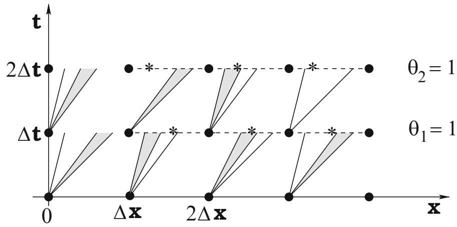


Fig. 40 An approximate solution constructed by the Glimm scheme

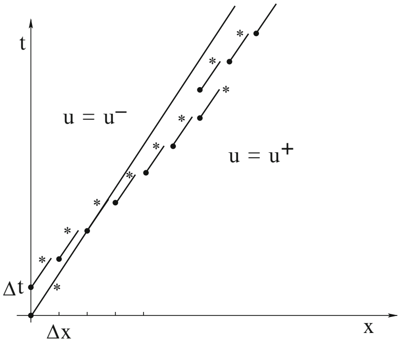


Fig. 41 Applying the Glimm scheme to a solution consisting of a single shock

1. If the initial data $\bar{u}$ has small total variation, then an approximate solution can be constructed by the above algorithm for all times $t \geq 0$. The total variation of $u(t, \cdot)$ remains small.
2. Letting the grid size $\Delta t = \Delta x$ tend to zero and using always the same sequence of numbers $\theta_j \in [0, 1]$, one obtains a sequence of approximate solutions u_ν. By Helly's compactness theorem, one can extract a subsequence that converges to some limit function $u = u(t, x)$ in $\mathbf{L}^1_{\text{loc}}$.
3. If the numbers θ_j are uniformly distributed over the interval $[0, 1]$, i.e. if (118) holds, then the limit function u provides a weak solution to the Cauchy problem (117).

The importance of the sequence θ_j being uniformly distributed can be best appreciated in the following example.

Example 10. Consider a Cauchy problem of the form (117). Assume that the exact solution consists of exactly one single shock, traveling with speed $\lambda \in [0, 1]$, say

$$U(t, x) = \begin{cases} u^+ & if \qquad x > \lambda t, \\ u^- & if \qquad x < \lambda t. \end{cases}$$

Consider an approximation of this solution obtained by implementing the Glimm algorithm (Fig. 41). By construction, at each time $t_j \doteq j\Delta t$, the position of the shock in this approximate solution must coincide with one of the nodes of the grid.

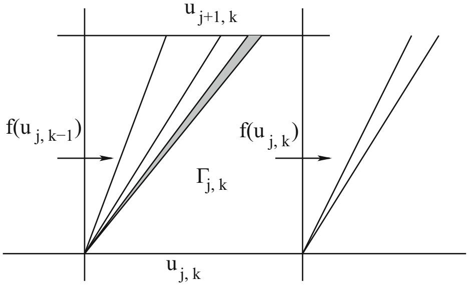


Fig. 42 Approximations leading to the Godunov scheme

Observe that, passing from t_{j-1} to t_j, the position $x(t)$ of the shock remains the same if the j-th sampling point lies on the left, while it moves forward by Δx if the j-th sampling point lies on the right. In other words,

$$x(t_j) \;=\; \begin{cases} x(t_{j-1}) & if \qquad \theta_j \in]\lambda, 1], \\ x(t_{j-1}) + \Delta x & if \qquad \theta_j \in [0, \lambda]. \end{cases} \tag{121}$$

Let us fix a time $T > 0$, and take $\Delta t \doteq T/N$. From (121) it now follows

$$\begin{aligned} x(T) &= \#\big\{j\ ;\ \ 1 \le j \le N,\ \ \theta_j \in [0, \lambda]\,\big\} \cdot \Delta t \\ &= \frac{\#\big\{j\ ;\ \ 1 \le j \le N,\ \ \theta_j \in [0, \lambda]\,\big\}}{N} \cdot T. \end{aligned}$$

It is now clear that the assumption (118) on the uniform distribution of the sequence $\{\theta_j\}_{j\ge 1}$ is precisely what is needed to guarantee that, as $N \to \infty$ (equivalently, as $\Delta t \to 0$), the location $x(T)$ of the shock in the approximate solution converges to the exact value λT.

Remark 7. At each restarting time t_j we need to approximate the BV function $u(t_j-, \cdot)$ with a new function which is piecewise constant on each interval $[x_{k-1},\ x_k[$. Instead of the sampling procedure (120), an alternative method consists of taking average values:

$$u(t_j+,\ x) \;\doteq\; \frac{1}{\Delta x}\int_{x_{k-1}}^{x_k} u(t_j-,\ y)\,dy \qquad\qquad \text{for all}\ \ x \in [x_{k-1},\ x_k[. \tag{122}$$

Calling u_{jk} the constant value of $u(t_j+)$ on the interval $[x_{k-1},\ x_k[$, an application of the divergence theorem on the square Γ_{jk} (Fig. 42) yields

$$u_{j+1,k} \;=\; u_{j,k} + \big[f(u_{j,k-1}) - f(u_{j,k})\big] \tag{123}$$

Indeed, all wave speeds are in [0, 1], hence

$$u(t, x_{k-1}) = u_{j,k-1}, \qquad u(t, x_k) = u_{j,k} \qquad \text{for all } t \in [t_j, t_{j+1}[.$$

The finite difference scheme (122) is the simplest version of the Godunov (upwind) scheme. It is very easy to implement numerically, since it does not require the solution of any Riemann problem. Unfortunately, as shown in [22], in general it is not possible to obtain a priori bounds on the total variation of solutions constructed by the Godunov method. Proving the convergence of these approximations remains an outstanding open problem.

The remaining part of this chapter will be concerned with error bounds, for solutions generated by the Glimm scheme.

Observe that, at each restarting time $t_j = j\,\Delta t$, the replacement of $u(t_j-)$ with the piecewise constant function $u(t_j+)$ produces an error measured by

$$\big\| u(t_j+) - u(t_j-) \big\|_{\mathbf{L}^1}$$

As the time step $\Delta t = T/N$ approaches zero, the total sum of all these errors does not converge to zero, in general. This can be easily seen in Example 10, where we have

$$\begin{aligned} \sum_{j=1}^{N} \big\| u(t_j+) - u(t_j-) \big\|_{\mathbf{L}^1} &\geq \sum_{j=1}^{N} |u^+ - u^-| \cdot \Delta t \cdot \min\big\{(1-\lambda),\ \lambda\big\} \\ &= |u^+ - u^-| \cdot T \cdot \min\big\{(1-\lambda),\ \lambda\big\}. \end{aligned}$$

This fact makes it difficult to obtain sharp error estimates for solutions generated by the Glimm scheme. Roughly speaking, the approximate solutions converge to the correct one not because the total errors become small, but because, by the randomness of the sampling choice, small errors eventually cancel each other in the limit.

Clearly, the speed of convergence of the Glimm approximate solutions as $\Delta t, \Delta x \to 0$ strongly depends on how well the sequence $\{\theta_i\}$ approximates a uniform distribution on the interval [0, 1]. In this connection, let us introduce

Definition 8. Let a sequence of numbers $\theta_j \in [0,1]$ be given. For fixed integers $0 \leq m < n$, the *discrepancy* of the set $\{\theta_m, \ldots, \theta_{n-1}\}$ is defined as

$$D_{m,n} \doteq \sup_{\lambda \in [0,1]} \left| \lambda - \frac{\#\big\{ j\ ;\ m \leq j < n,\ \ \theta_j \in [0,\lambda] \big\}}{n-m} \right|. \tag{124}$$

We now describe a simple method for defining the numbers θ_j, so that the corresponding discrepancies $D_{m,n}$ approach zero as $n - m \to \infty$, at a nearly optimal rate. Write the integer k in decimal digits, then invert the order of the digits and put a zero in front:

$$\theta_1 = 0.1\,, \quad \ldots \quad, \quad \theta_{759} = 0.957\,, \quad \ldots \quad, \quad \theta_{39022} = 0.22093\,, \quad \ldots \tag{125}$$

For the sequence (125) one can prove that the discrepancies satisfy

$$D_{m,n} \;\leq\; C \cdot \frac{1 + \ln(n-m)}{n-m} \qquad\qquad \text{for all } \; n > m \geq 0, \tag{126}$$

for some constant C. For approximate solutions constructed in terms of the above sequences (θ_j), using the restarting procedures (119)–(120), the following estimates were proved in [18].

Theorem 3 (Error estimates for the Glimm scheme). *Given any initial data $\bar{u} \in \mathbf{L}^1$ with small total variation, call $u^{\text{exact}}(t,\cdot) = S_t\bar{u}$ the exact solution of the Cauchy problem (117). Moreover, let $u^{\text{Glimm}}(t,\cdot)$ be the approximate solution generated by the Glimm scheme, in connection with a grid of size $\Delta t = \Delta x$ and a fixed sequence $(\theta_j)_{j\geq 0}$ satisfying (126). For every fixed time $T \geq 0$, letting the grid size tend to zero, one has the error estimate*

$$\lim_{\Delta x \to 0} \frac{\big\| u^{\text{Glimm}}(T,\cdot) - u^{\text{exact}}(T,\cdot)\big\|_{\mathbf{L}^1}}{\sqrt{\Delta x} \cdot |\ln \Delta x|} \;=\; 0. \tag{127}$$

In other words, the $\mathbf{L}^1$ error tends to zero faster then $\sqrt{\Delta x} \cdot |\ln \Delta x|$, i.e. just slightly slower than the square root of the grid size.

To prove Theorem 6, using a fundamental lemma of T.P. Liu [46], one first constructs a front tracking approximate solution $u = u(t,x)$ that coincides with u^{Glimm} at the initial time $t = 0$ and at the terminal time $t = T$. The $\mathbf{L}^1$ distance between $u(T,\cdot)$ and the exact solution $S_T\bar{u}$ can then be estimated using the error formula (7). For all details we refer to [18]. See also the recent paper [4] for a more general result.

6 Continuous Dependence on the Initial Data

Consider again the Cauchy problem (93)–(94), for a strictly hyperbolic system of conservation laws, satisfying the assumptions (H). Given two solutions u, v, in order to estimate the difference $\|u(t) - v(t)\|_{\mathbf{L}^1}$ one could try to follow a standard approach. Namely, set $w = u - v$, derive an evolution equation for w, and show that

$$\frac{d}{dt}\,\|w(t)\| \;\leq\; C\,\|w(t)\|. \tag{128}$$

By Gronwall's lemma, this implies

$$\|u(t) - v(t)\| \;\leq\; e^{Ct}\|u(0) - v(0)\|.$$

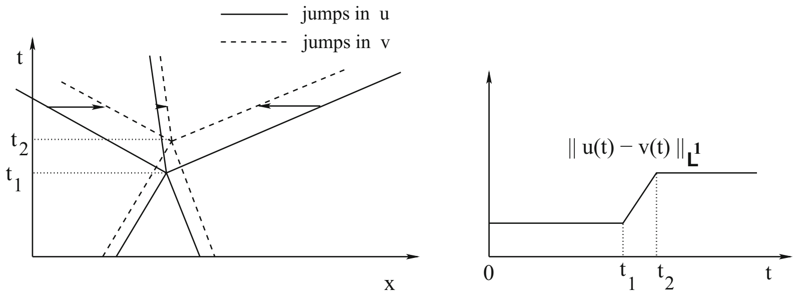


Fig. 43 *Left*: the solutions u and v differ only in the location of the shocks, and for the time of interaction. *Right*: even if u and v are very close, during the short time interval between interaction times, the distance $\|u - v\|_{\mathbf{L}^1}$ can increase rapidly

In particular, if $u(0) = v(0)$, then $u(t) = v(t)$ for all $t > 0$, proving the uniqueness of the solution to the Cauchy problem.

The above approach works well for smooth solutions of the hyperbolic system (93), but fails in the presence of shocks. Indeed, for two solutions u, v of a hyperbolic system containing shocks, the $\mathbf{L}^1$ distance can increase rapidly during short time intervals (Fig. 43).

6.1 Unique Solutions to the Scalar Conservation Law

In the case of a scalar conservation law, the fundamental works of A.I. Volpert [59] and S. Kruzhkov [39] have established:

Theorem 4 (Well posedness for the scalar Cauchy problem). *Let $f : I\!R \mapsto I\!R$ be any smooth flux. Then, for any initial data $\bar{u} \in \boldsymbol{L}^\infty$, the Cauchy problem (93)–(94) has a unique entropy-admissible weak solution, defined for all times $t \geq 0$. The corresponding flow is contractive in the $\boldsymbol{L}^1$ distance. Namely, for any two admissible solutions, one has*

$$\|u(t) - v(t)\|_{L^1} \leq \|u(0) - v(0)\|_{L^1} \qquad \text{for all } t \geq 0. \tag{129}$$

For a proof in the one-dimensional case, see [11]. We observe that the $\mathbf{L}^1$ distance between two solutions u, v remains constant in time, as long as shocks do not appear. An intuitive way to understand this fact, shown in Fig. 44, is as follows. Think of the x-u plane as filled by an incompressible fluid, moving horizontally with speed $(\dot{x}, \dot{u}) = (f'(u), 0)$. Consider the fluid particles that at time $t = 0$ lie in the region enclosed between the graphs of $u(0, \cdot)$ and $v(0, \cdot)$ (the shaded areas in Fig. 44). As long as these solutions remain continuous, the method of characteristics shows that at any positive time t these same particles of fluid will have moved to the region enclosed between the graphs of $u(t, \cdot)$ and $v(t, \cdot)$. Hence the area of these region remains constant in time.

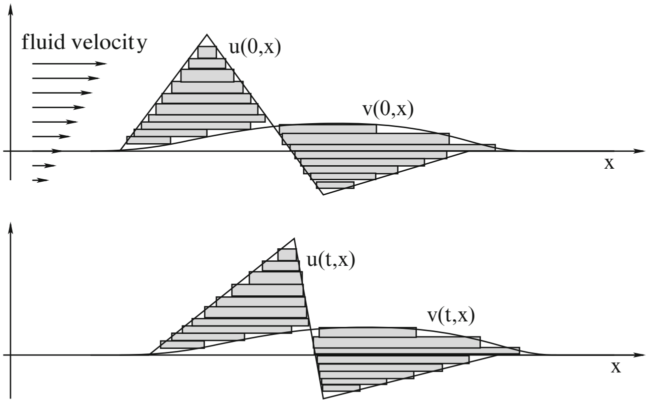


Fig. 44 The $\mathbf{L}^1$ distance between two continuous solutions remains constant in time

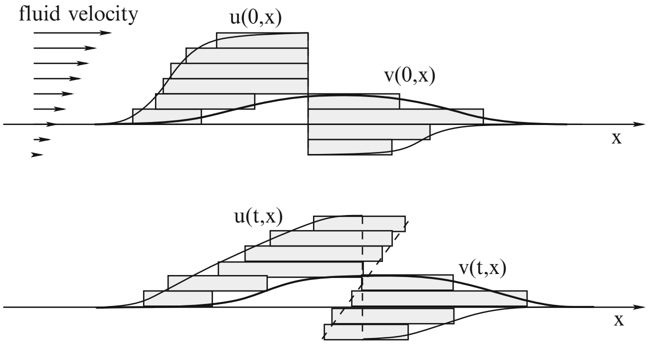


Fig. 45 The $\mathbf{L}^1$ distance decreases when a shock in one solution crosses the graph of the other solution

On the other hand, if a shock in one of the solutions crosses the graph of the other solution, then the $\mathbf{L}^1$ distance $\|u - v\|_{\mathbf{L}^1}$ decreases in time (Fig. 45).

6.2 Linear Hyperbolic Systems

We consider here another special case, where the system is linear with constant coefficients.

$$u_t + Au_x = 0 \qquad\qquad u \in I\!R^n. \tag{130}$$

Let $\{l_1, \ldots, l_n\}$ and $\{r_1, \ldots, r_n\}$ be dual bases of left and right eigenvectors of the matrix A, as in (8). Instead of the norm

$$\|u\|_{\mathbf{L}^1} \doteq \int |u(x)|\,dx$$

where $|u|$ is the Euclidean norm of a vector $u = (u_1, \ldots, u_n) \in I\!R^n$, one can use the equivalent norm

$$\|u\|_A \doteq \sum_{i=1}^{n} \int |l_i \cdot u(x)|\,dx. \tag{131}$$

By linearity, for any two solutions u, v, the difference $w = u - v$ satisfies still the same equation:

$$w_t + Aw_x = 0.$$

From the explicit representation (14), it now follows that

$$\|w(t)\|_A = \|w(0)\|_A \qquad \text{for all } t \in I\!R.$$

In other words, the flow generated by the linear homogeneous equation (130) is a group of isometries w.r.t. the distance $\|u - v\|_A$, namely

$$\|u(t) - v(t)\|_A = \|u(0) - v(0)\|_A \qquad \text{for all } t \in I\!R.$$

6.3 Nonlinear Systems

We always assume that the system (93) is strictly hyperbolic, and satisfies the hypotheses (H), so that each characteristic field is either linearly degenerate or genuinely nonlinear. The analysis in the previous chapter has shown the existence of a global entropy weak solution of the Cauchy problem for every initial data with sufficiently small total variation. More precisely, recalling the definitions (100)–(101), consider a domain of the form

$$\mathscr{D} \doteq cl\Big\{u \in \mathbf{L}^1(I\!R; I\!R^n);\ u \text{ is piecewise constant, } \Upsilon(u) \doteq V(u) + C_0 \cdot Q(u) < \delta_0\Big\}, \tag{132}$$

where cl denotes closure in $\mathbf{L}^1$. With a suitable choice of the constants C_0 and $\delta_0 > 0$, for every $\bar{u} \in \mathscr{D}$, one can construct a sequence of ε-approximate front tracking solutions converging to a weak solution u taking values inside $\mathscr{D}$. Observe that, since the proof of convergence relied on a compactness argument, no information was obtained on the uniqueness of the limit. The main goal of the section is to show that this limit is unique and depends continuously on the initial data.

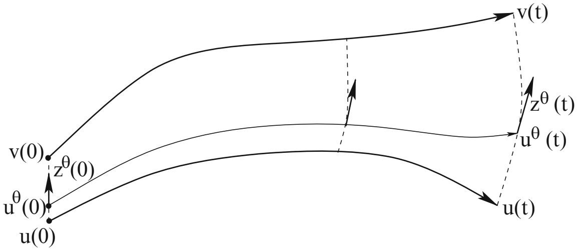


Fig. 46 Estimating the distance between two solutions by a homotopy method

Theorem 5. *For every $\bar{u} \in \mathscr{D}$, as $\varepsilon \to 0$ every sequence of ε-approximate solutions $u_\varepsilon : [0,\infty[\mapsto \mathscr{D}$ of the Cauchy problem (93)–(94), obtained by the front tracking method, converges to a unique limit solution $u : [0,\infty[\mapsto \mathscr{D}$. The map $(\bar{u},t) \mapsto u(t,\cdot) \doteq S_t\bar{u}$ is a uniformly Lipschitz semigroup, i.e.:*

$$S_0\bar{u} = \bar{u}, \qquad S_s(S_t\bar{u}) = S_{s+t}\bar{u}, \tag{133}$$

$$\left\| S_t\bar{u} - S_s\bar{v} \right\|_{\mathbf{L}^1} \leq L \cdot \big(\|\bar{u}-\bar{v}\|_{\mathbf{L}^1} + |t-s| \big) \qquad \text{for all } \bar{u},\bar{v} \in \mathscr{D}, \ \ s,t \geq 0. \tag{134}$$

This result was first proved in [14] for 2×2 systems, then in [21] for general $n \times n$ systems, using a (lengthy and technical) homotopy method. Here the idea is to consider a path of initial data $\gamma_0 : \theta \mapsto u^\theta(0)$ connecting $u(0)$ with $v(0)$. Then one constructs the path $\gamma_t : \theta \mapsto u^\theta(t)$, parameterized by $\theta \in [0,1]$, connecting the corresponding solutions at time t. By careful estimates on the tangent vector $z^\theta(t) \doteq du^\theta(t)/d\theta$, one shows that the length of γ_t can be uniformly bounded in terms of the length of the initial path γ_0 (Fig. 46).

Relying on ideas introduced by T.P. Liu and T. Yang in [48, 49], the paper [20] provided a much simpler proof of the continuous dependence result, which will be described here. An extension of the above result to initial-boundary value problems for hyperbolic conservation laws has recently appeared in [30]. All of the above results deal with solutions having small total variation. The existence of solutions, and the well posedness of the Cauchy problem for large BV data was studied respectively in [54] and in [41].

To prove the uniqueness of the limit of front tracking approximations, we need to estimate the distance between any two ε-approximate solutions u, v of (93). For this purpose we introduce a functional $\Phi = \Phi(u,v)$, uniformly equivalent to the $\mathbf{L}^1$ distance, which is "almost decreasing" along pairs of solutions. Recalling the construction of shock curves at (62), given two piecewise constant functions $u, v : I\!R \mapsto R^n$, we consider the scalar functions q_i defined implicitly by

$$v(x) = S_n\big(q_n(x)\big) \circ \cdots \circ S_1\big(q_1(x)\big)\big(u(x)\big). \tag{135}$$

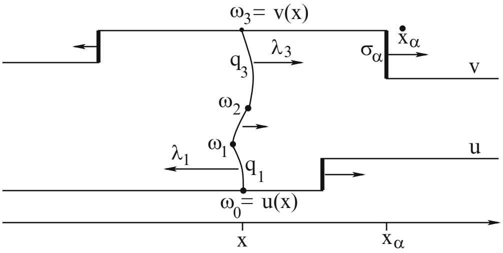


Fig. 47 Decomposing a jump $(u(x),\ v(x))$ in terms of n (possibly non-admissible) shocks

Remark 5. If we wanted to solve the Riemann problem with data $u^- = u(x)$ and $u^+ = v(x)$ only in terms of shock waves (possibly not entropy-admissible), then the corresponding intermediate states would be

$$\omega_0(x) = u(x), \qquad \omega_i(x) = S_i\big(q_i(x)\big) \circ \ \cdots \ \circ S_1\big(q_1(x)\big)\big(u(x)\big) \qquad i = 1, \ldots, n. \tag{136}$$

Moreover, $q_1(x), \ldots, q_n(x)$ would be the sizes of these shocks (Fig. 47). Since the pair of states (ω_{i-1}, ω_i) is connected by a shock, the corresponding speed $\lambda_i(u^-, u^+)$ is well defined. In particular, one can determine whether the i-shock q_i located at x is approaching a j-wave located at some other point x'. It is useful to think of $q_i(x)$ as the strength of the i-th component in the jump $\big(u(x),\ v(x)\big)$. In the linear case (130) we would simply have $q_i = l_i \cdot (v - u)$, and our functional would eventually reduce to (131).

If the shock curves are parameterized by arc-length, on a compact neighborhood of the origin one has

$$\big|v(x) - u(x)\big| \ \leq \ \sum_{i=1}^{n} \big|q_i(x)\big| \ \leq \ C\ \big|v(x) - u(x)\big| \tag{137}$$

for some constant C. We now consider the functional

$$\Phi(u, v) \ \doteq \ \sum_{i=1}^{n} \int_{-\infty}^{\infty} \big|q_i(x)\big| W_i(x)\, dx, \tag{138}$$

where the weights W_i are defined by setting:

$$\begin{aligned} &W_i(x) \\ &\doteq\ 1+\kappa_1\cdot\big[\text{total strength of waves in } u \text{ and in } v \text{ which approach the } i-\text{wave } q_i(x)\big] \\ &\qquad +\kappa_2\cdot\big[\text{wave interaction potentials of } u \text{ and of } v\big] \\ &\doteq\ 1 + \kappa_1 A_i(x) + \kappa_2\big[Q(u) + Q(v)\big]. \end{aligned} \tag{139}$$

Since these weights remain uniformly bounded as u ranges in the domain $\mathscr{D}$, from (137)–(139) it follows

$$\|u-v\|_{\mathbf{L}^1} \ \leq\ \Phi(u,v)\ \leq\ C_1\cdot\|v-u\|_{\mathbf{L}^1} \tag{140}$$

for some constant C_1 and all $u, v \in \mathscr{D}$. A key estimate proved in [20] shows that, for any two ε-approximate front tracking solutions $u, v : [0,T] \mapsto \mathscr{D}$, there holds

$$\frac{d}{dt}\Phi(u(t),v(t))\ \leq\ C_2\varepsilon, \tag{141}$$

for some constant C_2.

Relying on this estimate, we now prove Theorem 5. Let $\bar{u} \in \mathscr{D}$ be given. Consider any sequence $(u_\nu)_{\nu\geq 1}$, such that each u_ν is an ε_ν-approximate front tracking solution of the Cauchy problem (93)–(94). For every $\mu, \nu \geq 1$ and $t \geq 0$, by (140) and (141) it now follows

$$\begin{aligned}\big\|u_\mu(t)-u_\nu(t)\big\|_{\mathbf{L}^1} &\leq\ \Phi\big(u_\mu(t),\ u_\nu(t)\big)\\ &\leq\ \Phi\big(u_\mu(0),\ u_\nu(0)\big) + C_2 t\cdot\max\{\varepsilon_\mu,\ \varepsilon_\nu\}\\ &\leq\ C_1\big\|u_\mu(0)-u_\nu(0)\big\|_{\mathbf{L}^1} + C_2 t\cdot\max\{\varepsilon_\mu,\ \varepsilon_\nu\}.\end{aligned} \tag{142}$$

Since the right hand side of (142) approaches zero as $\mu,\nu \to \infty$, the sequence is Cauchy and converges to a unique limit. The semigroup property (133) is an immediate consequence of uniqueness. Finally, let $\bar{u},\bar{v} \in \mathscr{D}$ be given. For each $\nu \geq 1$, let u_ν, v_ν be ε_ν-approximate front tracking solutions of the Cauchy problem, with initial data $\bar{u}$ and $\bar{v}$, respectively. Using again (140) and (141) we deduce

$$\begin{aligned}\big\|u_\nu(t)-v_\nu(t)\big\|_{\mathbf{L}^1} &\leq\ \Phi\big(u_\nu(t),\ v_\nu(t)\big)\\ &\leq\ \Phi\big(u_\nu(0),\ v_\nu(0)\big) + C_2 t\varepsilon_\nu\\ &\leq\ C_1\Big(\|u_\nu(0)-\bar{u}\|_{\mathbf{L}^1} + \|\bar{u}-\bar{v}\|_{\mathbf{L}^1} + \big\|\bar{v}-v_\nu(0)\big\|_{\mathbf{L}^1}\Big) + C_2 t\varepsilon_\nu.\end{aligned}$$

Letting $\nu \to \infty$ we obtain $\big\|u(t)-v(t)\big\|_{\mathbf{L}^1} \leq C_1\cdot\|\bar{u}-\bar{v}\|_{\mathbf{L}^1}$, proving the Lipschitz continuous dependence w.r.t. the initial data.

7 Uniqueness of Solutions

According to the analysis in the previous chapters, the solution of the Cauchy problem (93)–(94) obtained as limit of front tracking approximations is unique and depends Lipschitz continuously on the initial data, in the $\mathbf{L}^1$ norm. This basic result, however, leaves open the question whether other weak solutions may exist, possibly constructed by different approximation algorithms. We will show that this is not the case: indeed, every entropy admissible solution, satisfying some minimal regularity

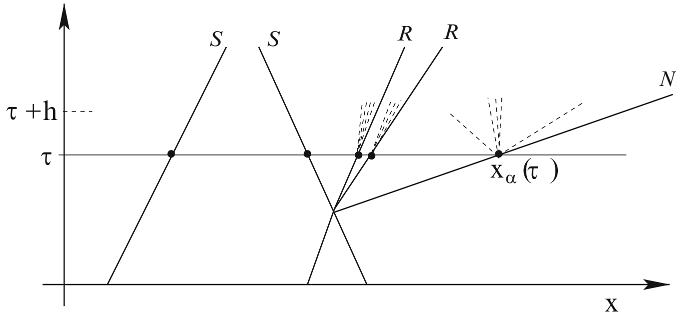


Fig. 48 The exact solution (*dotted lines*) which, at time τ, coincides with the value of a piecewise constant front tracking approximation

assumptions, necessarily coincides with the one obtained as limit of front tracking approximations.

7.1 An Error Estimate for Front Tracking Approximations

As a first step, we estimate the distance between an approximate solution, obtained by the front tracking method, and the exact solution of the Cauchy problem (93)–(94), given by the semigroup trajectory $t \mapsto u(t,\cdot) = S_t\bar{u}$. Let $u^\varepsilon : [0,T] \mapsto \mathscr{D}$ be an ε-approximate front tracking solution, according to Definition 6. We claim that the corresponding error can then be estimated as

$$\left\| u^\varepsilon(T,\cdot) - S_T\bar{u} \right\|_{\mathbf{L}^1} = \mathscr{O}(1)\cdot\varepsilon(1+T). \tag{143}$$

To see this, we first estimate the limit

$$\lim_{h\to 0+} \frac{\left\| u^\varepsilon(\tau+h) - S_h u^\varepsilon(\tau) \right\|_{\mathbf{L}^1}}{h}$$

at any time $\tau \in [0,T]$ where no wave-front interaction takes place. Let $u^\varepsilon(\tau,\cdot)$ have jumps at points $x_1 < \cdots < x_N$.

For each α, call ω_α the self-similar solution of the Riemann problem with data $u^\pm = u(\tau,\ x_\alpha\pm)$. We observe that, for $h > 0$ small enough, the semigroup trajectory $h \mapsto S_h u(\tau)$ is obtained by piecing together the solutions of these Riemann problems (Fig. 48). Splitting the set of all wave-fronts into shocks, rarefactions, and non-physical fronts, we estimate

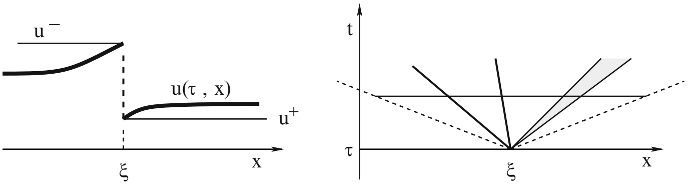


Fig. 49 In a forward neighborhood of a point (τ,ξ) where u has a jump, the admissible solution u should be asymptotically equivalent to the solution of a Riemannn problem

$$\begin{aligned}
&\lim_{h\to 0+}\frac{\left\|u^\varepsilon(\tau+h)-\widetilde{S}_h u^\varepsilon(\tau)\right\|_{\mathbf{L}^1}}{h}\\
&\quad=\sum_{\alpha\in\mathscr{R}\cup\mathscr{S}\cup\mathscr{N}\mathscr{P}}\left(\lim_{h\to 0+}\frac{1}{h}\int_{x_\alpha-\rho}^{x_\alpha+\rho}\left|u^\varepsilon(\tau+h,x)-\omega_\alpha(h,\ x-x_\alpha)\right|dx\right)\\
&\quad=\sum_{\alpha\in\mathscr{R}}\mathscr{O}(1)\cdot\varepsilon\,|\sigma_\alpha|+\sum_{\alpha\in\mathscr{N}\mathscr{P}}\mathscr{O}(1)\cdot|\sigma_\alpha|\ =\ \mathscr{O}(1)\cdot\varepsilon.
\end{aligned}\tag{144}$$

Here ρ can be any suitably small positive number. From the bound (144) and the error formula (7) in the Appendix, we finally obtain

$$\begin{aligned}
\left\|u^\varepsilon(T,\cdot)-S_T\bar{u}\right\|_{\mathbf{L}^1}&\le\left\|S_Tu^\varepsilon(0,\cdot)-S_T\bar{u}\right\|_{\mathbf{L}^1}+\left\|u^\varepsilon(T,\cdot)-S_Tu^\varepsilon(0,\cdot)\right\|_{\mathbf{L}^1}\\
&\le L\cdot\left\|u^\varepsilon(0,\cdot)-\bar{u}\right\|_{\mathbf{L}^1}+L\cdot\int_0^T\left\{\liminf_{h\to 0+}\frac{\left\|u^\varepsilon(\tau+h)-S_hu^\varepsilon(\tau)\right\|_{\mathbf{L}^1}}{h}\right\}d\tau\\
&=\ \mathscr{O}(1)\cdot\varepsilon+\mathscr{O}(1)\cdot\varepsilon T.
\end{aligned}$$

7.2 Characterization of Semigroup Trajectories

In this section, we describe a set of conditions which, among all weak solutions of the system (93) characterizes precisely the ones obtained as limits of front tracking approximations. These conditions, introduced in [10], are obtained by locally comparing a given solution with two types of approximations.

1. Comparison with solutions to a Riemann problem.

Let $u=u(t,x)$ be a weak solution. Fix a point (τ,ξ). Define $U^\sharp=U^\sharp_{(\tau,\xi)}$ as the solution of the Riemann problem corresponding to the jump at (τ,ξ) (Fig. 49):

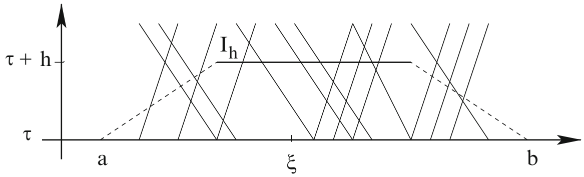


Fig. 50 The solution to a linearized hyperbolic system

$$w_t + f(w)_x = 0, \qquad\qquad w(\tau,x) = \begin{cases} u^+ \doteq u(\tau,\xi+) & if \quad x > \xi \\ u^- \doteq u(\tau,\xi-) & if \quad x < \xi \end{cases}$$

We expect that, if u satisfies the admissibility conditions, then u will be asymptotically equal to $U^\sharp$ in a forward neighborhood of the point (τ,ξ). More precisely, for every $\hat\lambda > 0$, one should have

$$\lim_{h\to 0+} \frac{1}{h}\int_{\xi-h\hat\lambda}^{\xi+h\hat\lambda} \Big| u(\tau+h,\ x) - U^\sharp_{(\tau,\xi)}(\tau+h,\ x)\Big|\, dx \ = \ 0. \tag{E1}$$

2. Comparison with solutions to a linear hyperbolic problem.

Fix again a point (τ,ξ), and choose $\hat\lambda > 0$ larger than all wave speeds. Define $U^\flat = U^\flat_{(\tau,\xi)}$ as the solution of the linear Cauchy problem (Fig. 50)

$$w_t + \widetilde{A} w_x = 0 \qquad\qquad w(\tau,x) = u(\tau,x)$$

with "frozen" coefficients: $\widetilde{A} \doteq A\big(u(\tau,\xi)\big)$. Then, for $a < \xi < b$ and $h > 0$, we expect that the difference between these two solutions should be estimated by

$$\frac{1}{h}\int_{a+\hat\lambda h}^{b-\hat\lambda h} \Big| u(\tau+h,\ x) - U^\flat(\tau+h,\ x)\Big|\, dx \ = \ \mathcal{O}(1)\cdot\Big(\text{Tot.Var.}\,\{u(\tau,\cdot);\ ]a,\, b[\,\}\Big)^2 \tag{E2}$$

A heuristic motivation for the above estimate is as follows. The functions u, w satisfy

$$u_t = -A(u)u_x\,, \qquad\qquad w_t = -\widetilde{A}w_x\,, \qquad\qquad u(\tau) = w(\tau).$$

Hence

$$\int_{a+\hat\lambda h}^{b-\hat\lambda h} \Big| u(\tau+h,x) - U^\flat(\tau+h,x)\Big| dx \approx \int_\tau^{\tau+h}\int_{J(t)} \Big| A(u(t,x))u_x - A(u(\tau,\xi))w_x \Big| dxdt, \tag{145}$$

where $J(t) \doteq]a + (t-\tau)\hat{\lambda}\,,\ \ b - (t-\tau)\hat{\lambda}\,[$. We now have

$$\int_{J(t)} \Big(|u_x(t,x)| + |w_x(t,x)| \Big)\, dx = \mathscr{O}(1) \cdot \text{Tot.Var.}\Big\{ u(\tau,\cdot);\ ]a,\, b[\Big\}\,,$$

$$\sup_{\tau<t<\tau+h,\ \ x\in J(t)} \Big| A(u(t,x)) - A(u(\tau,\xi)) \Big| = \mathscr{O}(1) \cdot \text{Tot.Var.}\big\{ u(\tau,\cdot);\ ]a,\, b[\big\}.$$

Therefore, for each time $t \in [\tau,\ \tau+h]$, the integrand on the right hand side of (145) is of the same order of magnitude as the square of the total variation. This yields (E2).

It can be proved that all solutions obtained as limits of front tracking approximations satisfy the estimates (E1)–(E2), for every τ, ξ, a, b. The following theorem, proved in [10], shows that the estimates (E1)–(E2) completely characterize semigroup trajectories, among all Lipschitz continuous functions $u : [0,T] \mapsto \mathbf{L}^1$ with values in the domain $\mathscr{D}$ defined at (132) .

Theorem 6 (Characterization of semigroup trajectories). *Let* $u : [0,T] \mapsto \mathscr{D}$ *be Lipschitz continuous w.r.t. the* $\mathbf{L}^1$ *distance. Then* u *is a weak solution to the system of conservation laws*

$$u_t + f(u)_x = 0$$

obtained as limit of front tracking approximations if and only if the estimates (E1)–(E2) are satisfied for a.e. $\tau \in [0,T]$, *at every* $\xi \in I\!R$.

The proof is based on the fact that the two estimates (E1) and (E2) together imply that

$$\lim_{h\to 0+} \frac{\|u(\tau+h) - S_h u(\tau)\|_{\mathbf{L}^1}}{h} = 0 \qquad \text{for a.e. } \tau. \tag{146}$$

Hence, by the error formula (7) in the Appendix,

$$\|u(t) - S_t u(0)\|_{\mathbf{L}^1} \leq L \cdot \int_0^T \left\{ \liminf_{h\to 0+} \frac{\big\|u(\tau+h) - S_h u(\tau)\big\|_{\mathbf{L}^1}}{h} \right\} d\tau = 0$$

for all $t \geq 0$.

In order to prove (146), choose points x_i such that $\text{Tot.Var.}\Big\{ u(\tau)\,;\ ]x_{i-1}\,,\ x_i[\Big\} < \varepsilon$ for every i. For $h > 0$ small, we split an integral over the entire real line into a sum of integrals over different intervals, as shown in Fig. 51:

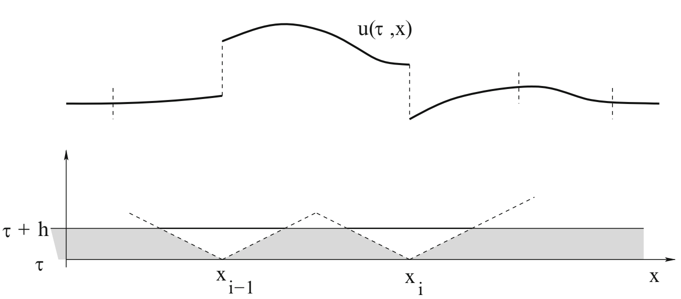


Fig. 51 Proving the asymptotic error estimate (146)

$$
\begin{aligned}
&\frac{1}{h}\int_{-\infty}^{\infty}\Big|u(\tau+h,x)-S_h u(\tau)(x)\Big|\,dx\\
&\quad=\sum_i \frac{1}{h}\int_{x_i-\hat\lambda h}^{x_i+\hat\lambda h}\Big\{\Big|u(\tau+h,x)-U_i^\sharp(\tau+h,x)\Big|+\Big|S_h u(\tau)(x)-U_i^\sharp(\tau+h,x)\Big|\Big\}\,dx\\
&\qquad+\sum_i \frac{1}{h}\int_{x_{i-1}+\hat\lambda h}^{x_i-\hat\lambda h}\Big\{\Big|u(\tau+h,x)-U_i^\flat(\tau+h,x)\Big|+\Big|S_h u(\tau)(x)-U_i^\flat(\tau+h,x)\Big|\Big\}\,dx\\
&\quad=\sum_i A_i+\sum_i B_i\,.
\end{aligned}
$$

The estimate (E1) implies $A_i \to 0$ as $h \to 0$, while the estimate (E2) implies $B_i \leq \varepsilon \cdot \text{Tot.Var.}\Big\{u(\tau)\,;\ \,]x_{i-1}\,,\ x_i[\Big\}$, and hence

$$\sum_i B_i \ \leq\ \varepsilon \cdot \text{Tot.Var.}\Big\{u(\tau)\,;\ I\!R\Big\} \ =\ \mathscr{O}(\varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, this proves (146).

7.3 *Uniqueness Theorems*

Relying on Theorem 6, there is a natural strategy in order to prove uniqueness of solutions to the Cauchy problem:

1. Introduce a suitable set of admissibility + regularity assumptions.
2. Show that these assumptions imply the estimates (E1) and (E2).

For sake of clarity, a complete set of assumptions is listed below.

(A1) **(Conservation Equations)** The function $u = u(t,x)$ is a weak solution of the Cauchy problem (93)–(94), taking values within the domain $\mathscr{D}$ of a semigroup S. More precisely, $u : [0,T] \mapsto \mathscr{D}$ is continuous w.r.t. the $\mathbf{L}^1$ distance. The identity $u(0,\cdot) = \bar{u}$ holds in $\mathbf{L}^1$, and moreover

$$\iint \big(u\varphi_t + f(u)\varphi_x\big)\, dxdt = 0 \tag{147}$$

for every $\mathscr{C}^1$ function φ with compact support contained inside the open strip $]0,T[\times I\!R$.

(A2) **(Lax Admissibility Conditions)** Let u have an approximate jump discontinuity at some point $(\tau,\xi) \in]0,T[\times I\!R$. More precisely, assume that there exists states $u^-, u^+ \in I\!R^n$ and a speed $\lambda \in I\!R$ such that, calling

$$U(t,x) \doteq \begin{cases} u^- & if \qquad x < \lambda t, \\ u^+ & if \qquad x > \lambda t, \end{cases} \tag{148}$$

there holds

$$\lim_{r\to 0+} \frac{1}{r^2}\int_{-r}^{r}\int_{-r}^{r} \Big|u(\tau+t,\ \xi+x) - U(t,x)\Big|\, dxdt = 0. \tag{149}$$

By Theorem 1, the piecewise constant function U must be a weak solution to the system of conservation laws, satisfying the Rankine–Hugoniot equations (29). In particular, the jump $u^+ - u^-$ should be an eigenvector of the averaged matrix $A(u^-,u^+)$, say of the i-th family, for some $i \in \{1,\ldots,n\}$. In this case, we assume that the following shock admissibility conditions hold:

$$\lambda_i(u^-) \ \geq\ \lambda\ \geq\ \lambda_i(u^+). \tag{150}$$

(A3) **(Tame Oscillation Condition)** For some constants C, $\hat{\lambda}$ the following holds. For every point $x \in I\!R$ and every $t, h > 0$ one has

$$\Big|u(t+h,x) - u(t,x)\Big| \ \leq\ C \cdot \text{Tot.Var.}\Big\{u(t,\cdot)\,;\ \ [x-\hat{\lambda}h,\ x+\hat{\lambda}h]\Big\}. \tag{151}$$

(A4) **(Bounded Variation Condition)** There exists $\delta > 0$ such that, for every space-like curve $\big\{t = \tau(x)\big\}$ with $|d\tau/dx| \leq \delta$ a.e., the function $x \mapsto u\big(\tau(x),x\big)$ has locally bounded variation.

Remark 6. The condition (A3) restricts the oscillation of the solution. An equivalent, more intuitive formulation is the following (see Fig. 52). For some constant $\hat{\lambda}$ larger than all characteristic speeds, given any interval $[a,b]$ and $t \geq 0$, the oscillation of u on the triangle $\Delta \doteq \big\{(s,y) :\ \ s \geq t,\ a+\hat{\lambda}(s-t) < y < b-\hat{\lambda}(s-t)\big\}$, defined as

$$\text{Osc}\big\{u;\ \Delta\big\} \doteq \sup_{(s,y),(s',y')\in\Delta} \big|u(s,y) - u(s',y')\big|,$$

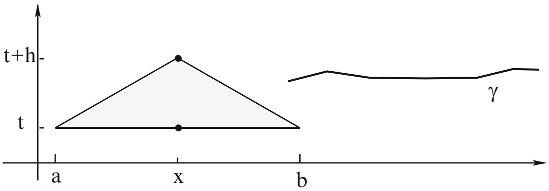


Fig. 52 Illustrating the tame oscillation and the bounded variation condition

is bounded by a constant multiple of the total variation of $u(t,\cdot)$ on $[a,b]$.

The assumption (A4) simply requires that, for some fixed $\delta > 0$, the function u has bounded variation along every space-like curve γ which is "almost horizontal" (Fig. 52). Indeed, the condition is imposed only along curves of the form $\big\{t = \tau(x);\ x \in [a,b]\big\}$ with

$$\big|\tau(x) - \tau(x')\big| \leq \delta|x - x'| \qquad \text{for all } x, x' \in [a,b].$$

One can prove that all of the above assumptions are satisfied by weak solutions obtained as limits of Glimm or wave-front tracking approximations [11]. The following result shows that the entropy weak solution of the Cauchy problem (93)–(94) is unique within the class of functions that satisfy either the additional regularity condition (A3), or (A4).

Theorem 7. *Assume that the function $u : [0,T] \mapsto \mathscr{D}$ is continuous (w.r.t. the $\mathbf{L}^1$ distance), taking values in the domain of the semigroup S generated by the system (93). If* (A1), (A2) *and* (A3) *hold, then*

$$u(t,\cdot) = S_t\bar{u} \qquad \textit{for all } t \in [0,T]. \tag{152}$$

In particular, the weak solution that satisfies these conditions is unique. The same conclusion holds if the assumption (A3) *is replaced by* (A4).

The first part of this theorem was proved in [15], the second part in [17]. Both of these papers extend the result in [16], where this approach to uniqueness was first developed.

8 The Vanishing Viscosity Approach

In view of the previous uniqueness and stability results, one expects that the entropy-admissible weak solutions of the hyperbolic system

$$u_t + f(u)_x = 0 \tag{153}$$

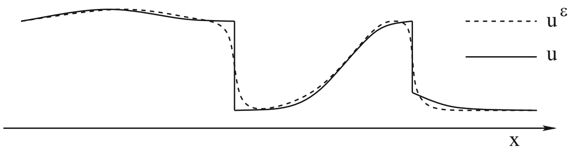


Fig. 53 A discontinuous solution to the hyperbolic system and a viscous approximation

should coincide with the unique limits of solutions to the parabolic system

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon\, u_{xx}^\varepsilon \tag{154}$$

letting the viscosity coefficient $\varepsilon \to 0$. For smooth solutions, this convergence is easy to show. However, one should keep in mind that a weak solution of the hyperbolic system (153) in general is only a function with bounded variation, possibly with a countable number of discontinuities. In this case, as the smooth functions u^ε approach the discontinuous solution u, near points of jump their gradients u_x^ε tend to infinity (Fig. 53), while their second derivatives u_{xx}^ε become even more singular. Therefore, establishing the convergence $u^\varepsilon \to u$ is a highly nontrivial matter. In earlier literature, results in this direction relied on three different approaches:

1. Comparison principles for parabolic equations. For a scalar conservation law, the existence, uniqueness and global stability of vanishing viscosity solutions was first established by Oleinik [51] in one space dimension. The famous paper by Kruzhkov [39] covers the more general class of $\mathbf{L}^\infty$ solutions and is also valid in several space dimensions.

2. Singular perturbations. This technique was developed by Goodman and Xin [36], and covers the case where the limit solution u is piecewise smooth, with a finite number of non-interacting, entropy admissible shocks. See also [58] and [53], for further results in this direction.

3. Compensated compactness. With this approach, introduced by Tartar and DiPerna [29], one first considers a weakly convergent subsequence $u^\varepsilon \rightharpoonup u$. For a class of 2×2 systems, one can show that this weak limit u actually provides a distributional solution to the nonlinear system (153). The proof relies on a compensated compactness argument, based on the representation of the weak limit in terms of Young measures, which must reduce to a Dirac mass due to the presence of a large family of entropies.

Since the hyperbolic Cauchy problem is known to be well posed within a space of functions with small total variation, it is natural to develop a theory of vanishing viscosity approximations within the same space BV. This was indeed accomplished in [7], in the more general framework of nonlinear hyperbolic systems not necessarily in conservation form. The only assumptions needed here are the strict hyperbolicity of the system and the small total variation of the initial data.

Theorem 8 (BV estimates and convergence of vanishing viscosity approximations). *Consider the Cauchy problem for the hyperbolic system with viscosity*

$$u_t^\varepsilon + A(u^\varepsilon)u_x^\varepsilon \;=\; \varepsilon\, u_{xx}^\varepsilon \qquad\qquad u^\varepsilon(0,x) = \bar{u}(x). \tag{155}$$

Assume that the matrices $A(u)$ are strictly hyperbolic (i.e., they have real, distinct eigenvalues), and depend smoothly on u in a neighborhood of the origin. Then there exist constants C, L, L' and $\delta > 0$ such that the following holds. If

$$Tot.Var.\{\bar{u}\} \ < \ \delta\,, \qquad\qquad \|\bar{u}\|_{L^\infty} \ < \ \delta, \tag{156}$$

then for each $\varepsilon > 0$ the Cauchy problem (155)$_\varepsilon$ has a unique solution u^ε, defined for all $t \geq 0$. Adopting a semigroup notation, this will be written as $t \mapsto u^\varepsilon(t,\cdot) \doteq S^\varepsilon_t \bar{u}$.

In addition, one has:

$$\textbf{\textit{BV bounds :}} \qquad\qquad Tot.Var.\left\{S^\varepsilon_t \bar{u}\right\} \ \leq \ C\ Tot.Var.\{\bar{u}\}. \tag{157}$$

$$\boldsymbol{L^1}\ \textbf{\textit{stability :}} \qquad\qquad \left\|S^\varepsilon_t \bar{u} - S^\varepsilon_t \bar{v}\right\|_{L^1} \ \leq \ L\ \left\|\bar{u} - \bar{v}\right\|_{L^1}, \tag{158}$$

$$\left\|S^\varepsilon_t \bar{u} - S^\varepsilon_s \bar{u}\right\|_{L^1} \ \leq \ L'\left(|t-s| + \left|\sqrt{\varepsilon t} - \sqrt{\varepsilon s}\,\right|\right). \tag{159}$$

Convergence: *As $\varepsilon \to 0+$, the solutions u^ε converge to the trajectories of a semigroup S such that*

$$\left\|S_t \bar{u} - S_s \bar{v}\right\|_{L^1} \ \leq \ L\ \|\bar{u} - \bar{v}\|_{L^1} + L'\,|t-s|. \tag{160}$$

These vanishing viscosity limits can be regarded as the unique vanishing viscosity solutions *of the hyperbolic Cauchy problem*

$$u_t + A(u)u_x \ = \ 0, \qquad\qquad u(0,x) = \bar{u}(x). \tag{161}$$

In the conservative case $A(u) = Df(u)$, every vanishing viscosity solution is a weak solution of

$$u_t + f(u)_x \ = \ 0, \qquad\qquad u(0,x) = \bar{u}(x)\,, \tag{162}$$

satisfying the Liu admissibility conditions.

Assuming, in addition, that each characteristic field is genuinely nonlinear or linearly degenerate, the vanishing viscosity solutions coincide with the unique limits of Glimm and front tracking approximations.

In the genuinely nonlinear case, an estimate on the rate of convergence of these viscous approximations was provided in [19]:

Theorem 9 (Convergence rate). *For the strictly hyperbolic system of conservation laws (162), assume that every characteristic field is genuinely nonlinear. At any time $t > 0$, the difference between the corresponding solutions of (155) and (162) can be estimated as*

$$\left\|u^\varepsilon(t,\cdot) - u(t,\cdot)\right\|_{L^1} \ = \ \mathscr{O}(1)\cdot(1+t)\sqrt{\varepsilon}|\ln\varepsilon|\ Tot.Var.\{\bar{u}\}.$$

In the following sections we outline the main ideas of the proof of Theorem 8. For details, see [7] or the lecture notes [12].

8.1 Local Decomposition by Traveling Waves

As a preliminary, observe that u^ε is a solution of (155) if and only if the rescaled function $u(t,x) \doteq u^\varepsilon(\varepsilon t, \varepsilon x)$ is a solution of the parabolic system with unit viscosity

$$u_t + A(u)u_x \;=\; u_{xx}\,, \tag{163}$$

with initial data $u(0,x) = \bar{u}(\varepsilon x)$. Clearly, the stretching of the space variable has no effect on the total variation. Notice however that the values of u^ε on a fixed time interval $[0,T]$ correspond to the values of u on the much longer time interval $[0,\ T/\varepsilon]$. To obtain the desired BV bounds for the viscous solutions u^ε, it suffices to study solutions of (163). However, we need estimates uniformly valid for all times $t \geq 0$, depending only on the total variation of the initial data $\bar{u}$.

To provide a uniform estimate on Tot.Var.$\{u(t,\cdot)\} = \|u_x(t,\cdot)\|_{\mathbf{L}^1}$, we decompose the gradient u_x along a basis of unit vectors $\tilde{r}_1,\ldots,\tilde{r}_n$, say

$$u_x \;=\; \sum_i v_i \tilde{r}_i\,. \tag{164}$$

We then derive an evolution equation for these gradient components, of the form

$$v_{i,t} + (\tilde{\lambda}_i v_i)_x - v_{i,xx} \;=\; \phi_i \qquad\qquad i = 1,\ldots,n\,, \tag{165}$$

Since the left hand side of (165) is in conservation form, we have

$$\|u_x(t)\| \;\leq\; \sum_{i=1}^n \|v_i(t,\cdot)\|_{\mathbf{L}^1} \;\leq\; \sum_i \left(\|v_i(0,\cdot)\|_{\mathbf{L}^1} + \int_0^t \|\phi_i(s,\cdot)\|_{\mathbf{L}^1}\,ds \right). \tag{166}$$

A crucial point in the entire analysis is the choice of the unit vectors $\tilde{r}_i$. A natural guess would be to take $\tilde{r}_i \;=\; r_i(u)$, the i-th eigenvector of the hyperbolic matrix $A(u)$. This was indeed the decomposition used in Sect. 1.6. As in (22), we thus write

$$u_x \;=\; \sum_i u_x^i r_i \qquad\qquad u_x^i \doteq l_i \cdot u_x\,, \tag{167}$$

so that (163) takes the form

$$u_t \;=\; -\sum_i \lambda_i u_x^i r_i + \sum_i (u_x^i r_i)_x\,. \tag{168}$$

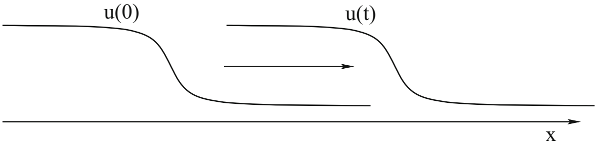


Fig. 54 For a viscous traveling wave, the source terms ϕ_i are usually not integrable

Differentiating the first equation in (167) w.r.t. t and the equation in (168) w.r.t. x, and equating the results, we obtain an evolution equation for the gradient components u_x^i, namely

$$(u_x^i)_t + (\lambda_i u_x^i)_x - (u_x^i)_{xx} = \phi_i(u, u_x^1, \dots, u_x^n) \doteq l_i \cdot \sum_{j<k} \lambda_k [r_k, r_j] u_x^j u_x^k$$

$$+ l^i \cdot \left\{ 2\sum_{j,k} (r_k \bullet r_j)(u_x^j)_x u_x^k + \sum_{j,k,\ell} \Big(r_\ell \bullet (r_k \bullet r_j) - (r_\ell \bullet r_k) \bullet r_j \Big) u_x^j u_x^k u_x^\ell \right\}. \tag{169}$$

Here $r_k \bullet r_j \doteq (Dr_j)r_k$ denotes the directional derivative of r_j along r_k, while $[r_k, r_j] \doteq (Dr_j)r_k - (Dr_k)r_j$ is the Lie bracket of the two vector fields. Relying on the above formula, in order to achieve BV bounds uniformly valid for $t \in [0, \infty[$, we would need $\int_0^\infty \int |\phi_i| \, dxdt < \infty$. Unfortunately this does not hold, in general. Indeed, for a typical solution having the form of a traveling wave $u(t,x) = \bar{u}(x - \lambda t)$, as in Fig. 54, the source terms do not vanish identically: $\phi_i \not\equiv 0$. Therefore

$$\int_0^t \int |\phi_i(\tau, x)| \, dxd\tau = t \cdot \int |\phi_i(0,x)| \, dx \to \infty \qquad \text{as } t \to \infty$$

To readdress this situation, a key idea is to decompose u_x not along the eigenvectors $r_1, \dots, r_n$ of $A(u)$, but along a basis $\{\tilde{r}_1, \dots \tilde{r}_n\}$ of **gradients of viscous traveling waves**.

We recall that a traveling wave solution of the viscous hyperbolic system (163) is a solution of the form

$$u(t,x) = U(x - \sigma t). \tag{170}$$

Here the constant $\sigma = -U_t / U_x$ is the speed of the wave. Inserting (170) in (163), we see that the function U should satisfy the second order O.D.E.

$$U'' = \big(A(U) - \sigma\big) U'. \tag{171}$$

As shown in Fig. 55, we wish to decompose $u_x = \sum_i U_i'$ locally as sum of gradients of traveling waves. More precisely, given (u, u_x, u_{xx}) at a point x, we seek traveling wave profiles $U_1, \dots, U_n$ such that

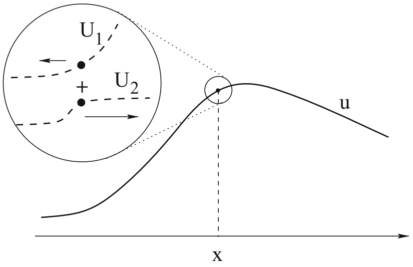


Fig. 55 Decomposing the function *u* as the superposition of two viscous traveling profiles, in a neighborhood of a point *x*

$$U_i'' = \big(A(U_i) - \sigma_i\big)U_i', \qquad U_i(x) = u(x) \qquad i = 1, \dots, n, \tag{172}$$

$$\sum_i U_i'(x) = u_x(x), \qquad \sum_i U_i''(x) = u_{xx}(x). \tag{173}$$

Observe that, having fixed $u(x)$, the system (172)–(173) yields

- $n + n$ scalar equations.
- $n^2 + n$ free parameters: the vectors $U_1'(x), \dots, U_n'(x) \in I\!R^n$, describing the first derivatives of the traveling waves, and the scalars $\sigma_1, \dots, \sigma_n$, describing the speeds.

For $n > 1$, the system is under-determined. To achieve a unique decomposition, further restrictions must thus be imposed on the choice of the traveling wave profiles. Indeed, for each given state $u \in I\!R^n$ and $i = 1, \dots, n$, we should select a two-parameter family of traveling waves through u. This is done using the center manifold theorem [13].

To begin with, we replace the second order O.D.E. (171) describing traveling waves with an equivalent first order system:

$$\begin{cases} \dot u = v, \\ \dot v = \big(A(u) - \sigma\big)v, \\ \dot\sigma = 0. \end{cases} \tag{174}$$

This consists of $n + n + 1$ O.D.E's. Notice that the last equation simply says that the speed σ is a constant. Fix a state $u^* \in I\!R^n$. Linearizing (171) at the equilibrium point $P^* = (u^*, 0, \lambda_i(u^*))$, one obtains the system

$$\begin{pmatrix} \dot u \\ \dot v \\ \dot\sigma \end{pmatrix} = \begin{pmatrix} 0 & I & 0 \\ 0 & A(u^*) - \lambda_i(u^*)I & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ \sigma \end{pmatrix} \in I\!R^{n+n+}. \tag{175}$$

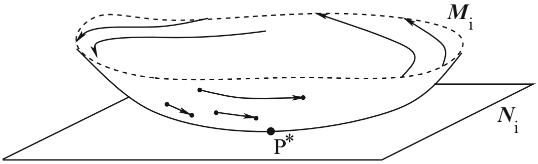


Fig. 56 The linear subspace $\mathscr{N}_i$ and the center manifold $\mathscr{M}_i$ tangent to $\mathscr{N}_i$ at the equilibrium point P^*

Recalling that $A(u^*)$ is a $n \times n$ matrix with real and distinct eigenvalues, one checks that the center subspace $\mathscr{N}_i$ for the $(2n+1) \times (2n+1)$ matrix in (175) (i.e., the invariant subspace corresponding to all generalized eigenvalues with zero real part) has dimension $n+2$.

By the center manifold theorem, for each $i = 1, \ldots, n$, the nonlinear system (174) has a center manifold $\mathscr{M}_i$ of dimension $n+2$, tangent to the center subspace $\mathscr{N}_i$ at P^* (Fig. 56).

A more detailed analysis shows that on $\mathscr{M}_i$ we can choose coordinates $(u, v^i, \sigma_i) \in I\!R^{n+1+1}$. Here v^i is the signed strength of the traveling wave profile through u, and σ_i is its speed. In other words, at any given point $\bar{x}$, for every (u, v^i, σ_i) in a neighborhood of $(u^*, 0, \lambda_i(u^*))$, there exists a unique solution to (171) such that

$$U_i(\bar{x}) = u, \qquad U_i'' = (A(U_i) - \sigma_i)U_i', \qquad U_i'(\bar{x}) = v^i \tilde{r}_i$$

for some unit vector $\tilde{r}_i = \tilde{r}_i(u, v^i, \sigma_i)$.

The previous construction in terms of center manifold trajectories provides a decomposition of u_x along a basis of *generalized eigenvectors*: $\tilde{r}_i(u, v^i, \sigma_i)$. These are unit vectors, close to the usual eigenvectors $r_i(u)$ of the matrix $A(u)$, which depend on two additional parameters.

Defining the corresponding *generalized eigenvalues* in terms of a scalar product:

$$\tilde{\lambda}_i(u, v^i, \sigma_i) \doteq \big\langle \tilde{r}_i\,,\ A(u)\tilde{r}_i \big\rangle,$$

one can prove the key identity

$$(A(u) - \tilde{\lambda}_i)\tilde{r}_i = v^i \big(\tilde{r}_{i,u}\tilde{r}_i + \tilde{r}_{i,v}(\tilde{\lambda}_i - \sigma_i)\big). \tag{176}$$

This replaces the standard identity

$$\big(A(u) - \lambda_i\big) r_i = 0 \tag{177}$$

satisfied by the eigenvectors and eigenvalues of $A(u)$. The additional terms on the right hand side of (176) play a crucial role, achieving a cancellation in the source terms ϕ_i in (165). Eventually, this allows us to prove that these source terms are globally integrable, in t and x.

8.2 Evolution of Gradient Components

Let $(u, u_x, u_{xx}) \in I\!R^{3n}$ be given, in a neighborhood of the origin. For convenience, instead of the decomposition (172)–(173), it is convenient to set $u_t = u_{xx} - A(u)u_x$ and seek a decomposition of the form

$$\begin{cases} u_x = \sum v^i \tilde{r}_i(u, v^i, \sigma_i) \\ u_t = \sum w^i \tilde{r}_i(u, v^i, \sigma_i) \end{cases} \qquad \text{with} \quad \sigma_i \approx -\frac{w^i}{v^i}.$$

After a lengthy computation, one finds that these components satisfy a system of evolution equations of the form

$$\begin{cases} v^i_t + (\tilde{\lambda}_i v^i)_x - v^i_{xx} = \phi_i \\ w^i_t + (\tilde{\lambda}_i w^i)_x - w^i_{xx} = \psi_i \end{cases} \tag{178}$$

A detailed analysis of the right hand sides of (178) shows that these source terms can be estimated as

$$\phi_i,\ \psi_i = \mathcal{O}(1) \cdot \sum\nolimits_j |w^j + \sigma_j v^j| \cdot \Big(|v^j w^j| + |v^j_x| + |w^j_x| \Big) \qquad \textbf{(wrong speed)}$$

$$+\mathcal{O}(1) \cdot \sum\nolimits_j |w^j_x v^j - v^j_x w^j| \qquad \textbf{(change in speed, linear)}$$

$$+\mathcal{O}(1) \cdot \sum\nolimits_j \left| v^j \left(\frac{w^j}{v^j} \right)_x \right|^2 \qquad \textbf{(change in speed, quadratic)}$$

$$+\mathcal{O}(1) \cdot \sum\nolimits_{j \neq k} \Big(|v^j v^k| + |v^j_x v^k| + |v^j w^k| + |v^j_x w^k| + |w^j w^k| \Big)$$

(interaction of waves of different families)

See [7] for detailed computations. Here we can only give an intuitive motivation for how these source terms arise. If u is precisely a j-traveling wave profile on the center manifold $\mathcal{M}_j$, say $u(t, x) = U_j(x - \sigma_j t)$, then by the key identity (176) it follows that all source terms vanish identically (Fig. 57). In essence, the size of these source terms is determined by how much the second order jet (u, u_x, u_{xx}) in our solution u differs from the jet of a traveling wave profile (Fig. 58).

Wrong speed. In a traveling wave profile $u(t, x) = U(t - \sigma t)$, the speed is the constant value $\sigma = -U_t/U_x$. However, near a point x_0 where $u_x = 0$, the speed of a traveling wave would be $\sigma = -u_t/u_x \to \infty$. Since we want $\sigma_i \approx \lambda_i(u^*)$, i.e., close to the i-th characteristic speed, a cut-off function must be used. These source terms describe by how much the identity $\sigma_i = -w^i/v^i$ is violated.

Change in wave speed. These terms account for local interactions of waves of the same family. Think of the viscous traveling j-wave that best approximates u at a point x, and at a nearby point x'. In general, these two profiles will not be the same, hence some local interaction between them will occur. A measure of how much the

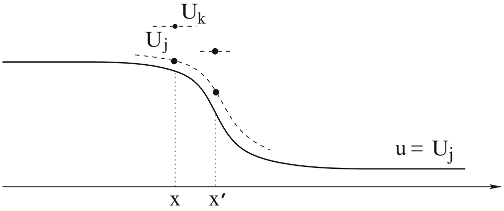


Fig. 57 If u coincides with a traveling wave profile, say of the j-th family, then all source terms vanish identically

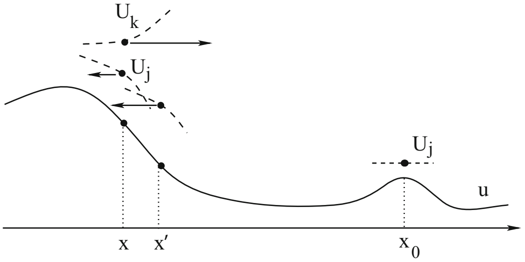


Fig. 58 Source terms arise because of (1) Interactions of j-waves with k-waves, (2) Interactions between waves of the same j-th family, if their speed varies with x, (3) Points x_0 where the decomposition in traveling profiles cannot be performed exactly

j-traveling profile changes, as the point x varies, is provided by the change in speed: $(\sigma_j)_x$

Assuming that the speed satisfies $\sigma_j = -w^j/v^j$, one has

$$\left|(\sigma_j)_x\right| = \frac{|w_x^j v^j - v_x^j w^j|}{|v^j|^2}.$$

The terms related to change in wave speeds can thus be written as products:

$$[\text{strength of the wave}]^2 \times [\text{rate of change of the speed}]^\alpha$$

with $\alpha = 1, 2$. More precisely,

$$\mathscr{O}(1) \cdot \sum_{j=1}^{n} |v^j|^2 \left|(\sigma_j)_x\right| + \mathscr{O}(1) \cdot \sum_{j=1}^{n} |v^j|^2 \left|(\sigma_j)_x\right|^2.$$

Transversal wave interactions. In general, at a given point x, waves of distinct families $j \neq k$ are present. These terms model interactions between these different waves.

8.3 Lyapunov Functionals

We seek uniform bounds on the norms $\|v^i(t)\|_{\mathbf{L}^1}$, $\|w^i(t)\|_{\mathbf{L}^1}$, independent of time. Since the left hand sides of (178) are in conservation form, it suffices to show that all source terms are uniformly integrable in both variables t, x. To prove that

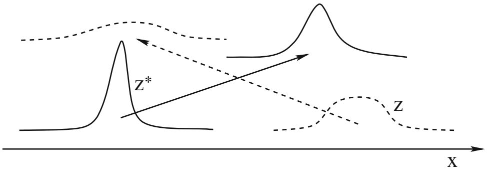


Fig. 59 Interaction of two viscous waves of different families

$$\int_0^\infty \|\phi_i(\tau)\|_{\mathbf{L}^1}\,d\tau \;<\; \infty\,, \qquad\qquad \int_0^\infty \|\psi_i(\tau)\|_{\mathbf{L}^1}\,d\tau \;<\; \infty\,,$$

we construct suitable Lyapunov functionals $\Psi(u) \geq 0$ such that

$$\|\phi_i(t)\|_{\mathbf{L}^1}, \|\psi_i(t)\|_{\mathbf{L}^1} \;\leq\; -\frac{d}{dt}\Psi(u(t))$$

In other words, at each time t, the $\mathbf{L}^1$ norm of source terms should be controlled by the rate of decrease of the functional. A summary of the basic estimates is as follows:

Wrong speed $\Longrightarrow$ **Parabolic energy estimates**
Change in wave speed, linear $\Longrightarrow$ **Area functional**
Change in wave speed, quadratic $\Longrightarrow$ **Curve length functional**
Interaction of waves of different families $\Longrightarrow$ **Wave interaction potential**

In the remainder of this section we describe the main ideas involved in the construction of these functionals.

1. Lyapunov functionals for a pair of linear parabolic equations.

Consider the system of two linear, scalar parabolic equations

$$\begin{cases} z_t + \big[\lambda(t,x)\,z\big]_x - z_{xx} \;=\; 0,\\ z_t^* + \big[\lambda^*(t,x)\,z^*\big]_x - z_{xx}^* \;=\; 0. \end{cases}$$

Assume that the propagation speeds λ and λ^* are strictly different:

$$\inf_{t,x}\lambda^*(t,x) - \sup_{t,x}\lambda(t,x) \;\geq\; c \;>\; 0.$$

It is useful to think of $z(\cdot)$ as the density of waves with slow speed λ, while $z^*(\cdot)$ is the density of waves with fast speed λ^*. The *instantaneous amount of interaction* between z and z^* is defined as (Fig. 59)

$$I(t) \;\doteq\; \int \big|z(t,x)\big|\cdot\big|z^*(t,x)\big|\,dx.$$

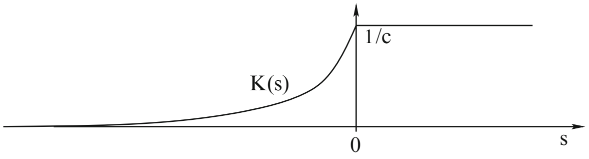


Fig. 60 The interaction kernel K defined at (180)

In order to bound the total amount of interaction, we introduce a *potential for transversal wave interactions* with :

$$Q(z,z^\sharp) \doteq \iint K(x-y)\,\big|z(x)\big|\,\big|z\sharp(y)\big|\,dxdy\,, \tag{179}$$

with (Fig. 60)

$$K(s) \doteq \begin{cases} 1/c & if \quad s \geq 0, \\ e^{cs/2}/c & if \quad s < 0. \end{cases} \tag{180}$$

Computing the distributional derivatives of the kernel K, one checks that $cK' - 2K''$ is precisely the Dirac distribution, i.e. a unit mass at the origin. We now compute

$$\frac{d}{dt}\,Q\big(z(t),z^\sharp(t)\big) = \frac{d}{dt}\iint K(x-y)\big|z(x)\big|\,\big|z^\sharp(y)\big|\,dxdy$$

$$= \iint K(x-y)\Big\{\Big(z_{xx}-(\lambda z)_x\Big)\text{sgn}z(x)\big|z^\sharp(y)\big| + \big|z(x)\big|\Big(z^\sharp_{yy}-(\lambda^\sharp z^\sharp)_y\Big)\text{sgn}z^\sharp(y\Big\}\,dxdy$$

$$\leq \iint K'(x-y)\Big\{\lambda\big|z(x)\big|\,\big|z^\sharp(y)\big| - \lambda^\sharp\big|z(x)\big|\,\big|z^\sharp(y)\big|\Big\}\,dxdy$$
$$+\iint K''(x-y)\Big\{\big|z(x)\big|\,\big|z^\sharp(y)\big| + \big|z(x)\big|\,\big|z^\sharp(y)\big|\Big\}\,dxdy$$

$$\leq -\iint \big(cK'-2K''\big)\big|z(x)\big|\,\big|z^\sharp(y)\big|\,dxdy = -\int \big|z(x)\big|\,\big|z^\sharp(x)\big|\,dx$$

Therefore, since $Q \geq 0$, for every $T \geq 0$ we have

$$\begin{aligned} \textstyle\int_0^T\!\int \big|z(t,x)\big|\,\big|z^\sharp(t,x)\big|\,dxdt &\leq Q\,\big(z(0),z^\sharp(0)\big) - Q\,\big(z(T),z^\sharp(T)\big) \\ &\leq \tfrac{1}{c}\,\big\|z(0)\big\|_{\mathbf{L}^1}\big\|z^\sharp(0)\big\|_{\mathbf{L}^1}. \end{aligned}$$

Using functionals of the form (179), one can control the source terms

$$\mathscr{O}(1)\cdot\sum_{j\neq k}\Big[|v^j v^k| + |v^j_x v^k| + |v^j w^k| + |v^j_x w^k| + |w^j w^k|\Big]$$

accounting for interaction of waves of different families.

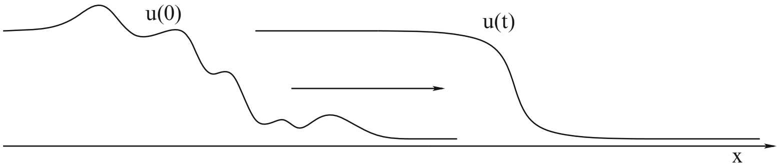


Fig. 61 As $t \to +\infty$, the solution to a scalar viscous conservation law is expected to approach a traveling wave profile

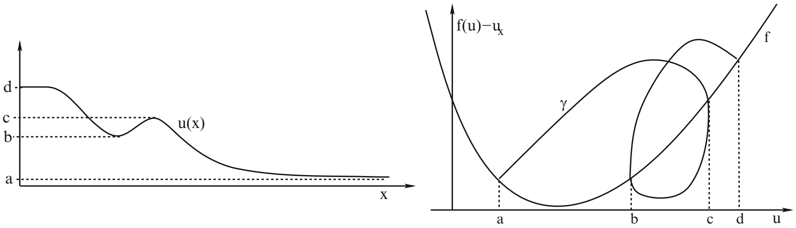


Fig. 62 *Left*: the graph of a function $u = u(x)$. *Right*: the corresponding curve $x \mapsto \gamma(x) = \Big(u(x),\ f(u(x)) - u_x(x)\Big)$

2. Lyapunov functionals for a scalar viscous conservation law
Consider a scalar conservation law with viscosity:

$$u_t + f(u)_x \;=\; u_{xx}. \tag{181}$$

We seek functionals that decrease in time, along every solution of (181). As $t \to +\infty$, we expect that the solution will approach a viscous traveling wave profile. One could thus look for a Lyapunov functional describing how far u is from a viscous traveling wave profile (Fig. 61).

For this purpose, it is convenient to adopt a variable transformation. Given a scalar function $u = u(x)$, consider the curve (Fig. 62)

$$\gamma \doteq \begin{pmatrix} u \\ f(u) - u_x \end{pmatrix} \;=\; \begin{pmatrix} \text{conserved quantity} \\ \text{flux} \end{pmatrix} \tag{182}$$

Observe that $u(\cdot)$ is a traveling wave profile if and only if the corresponding curve γ is a segment. Indeed

$$-\frac{u_t}{u_x} \;=\; \frac{f(u)_x - u_{xx}}{u_x} \;=\; \text{constant} \;=\; [\text{wave speed}]$$

if and only if

$$\frac{d}{du}\Big[f(u) - u_x\Big] \;=\; \Big[f(u) - u_x\Big]_x \cdot \frac{1}{u_x} \;=\; \text{constant}.$$

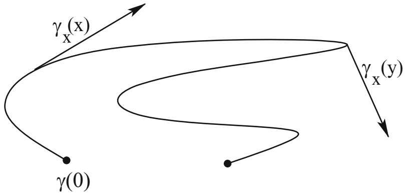


Fig. 63 Defining the area functional

If now $u = u(t, x)$ provides a solution to the viscous conservation law (181), the corresponding curve γ in (182) evolves according to the vector equation

$$\gamma_t + f'(u)\gamma_x = \gamma_{xx}. \tag{183}$$

Recalling that

$$\gamma \doteq \begin{pmatrix} u \\ f(u) - u_x \end{pmatrix}, \qquad \gamma_x = \begin{pmatrix} v \\ w \end{pmatrix} \doteq \begin{pmatrix} u_x \\ -u_t \end{pmatrix}. \tag{184}$$

we find two functionals associated with (183). One is

$$\textbf{Curve Length:} \qquad L(\gamma) \doteq \int |\gamma_x|\, dx = \int \sqrt{v^2 + w^2}\, dx. \tag{185}$$

Indeed, a direct computation yields

$$\frac{d}{dt} L\big(\gamma(t)\big) = -\int \frac{|v| \Big[(w/v)_x \Big]^2}{\big(1 + (w/v)^2\big)^{3/2}}\, dx.$$

Using functionals of this type, one controls the source terms

$$\mathcal{O}(1) \cdot \left| v^j \left(\frac{w^j}{v^j} \right)_x \right|^2 \qquad \text{(change in wave speed, quadratic).}$$

The second functional is (see Fig. 63)

$$\textbf{Area functional:} \qquad Q(\gamma) \doteq \frac{1}{2} \iint_{x<y} \Big| \gamma_x(x) \wedge \gamma_x(y) \Big|\, dx\, dy \tag{186}$$

If γ evolves in the direction of curvature, then Q controls the area swept by the curve: $|dA| \leq -dQ$. This can best be understood thinking of polygonal approximations (Fig. 64). If γ is a polygonal with sides $\mathbf{v}_j$, the double integral in (186) is computed by a finite sum:

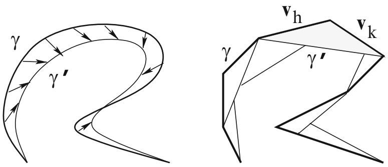


Fig. 64 The decrease in the area functional bounds the area swept by the curve in its motion

$$Q(\gamma) \;=\; \frac{1}{2}\sum_{i<j}\big|\mathbf{v}_i \wedge \mathbf{v}_j\big|. \tag{187}$$

If we now replace two consecutive edges $\mathbf{v}_h, \mathbf{v}_k$ by a single segment, the area of the corresponding triangle is

$$|dA| \;=\; \frac{1}{2}\big|\mathbf{v}_h \wedge \mathbf{v}_k\big| \leq -dQ$$

Indeed, the term $\frac{1}{2}\big|\mathbf{v}_h \wedge \mathbf{v}_k\big|$ is now missing from the sum in (187), while the sum of all other terms remains the same, or decreases.

Recalling (183)–(184), we now compute

$$-\frac{dQ}{dt} \;\geq\; \left|\frac{dA}{dt}\right| \;=\; \int |\gamma_t \wedge \gamma_x|\,dx \;=\; \int |\gamma_{xx} \wedge \gamma_x|\,dx \;=\; \int |v_x w - v w_x|\,dx.$$

As a consequence, the integral over time of the right hand side can be estimated by

$$\int_0^\infty \int |v_x w - v w_x|\,dx\,dt \;\leq\; \int_0^\infty \left|\frac{d}{dt} Q\big(\gamma(t)\big)\right| dt \;\leq\; Q\big(\gamma(0)\big)$$

Using functionals of this type, one can control the source terms

$$\mathscr{O}(1)\cdot |v_x^j w^j - v^j w_x^j| \qquad\qquad \text{(change in wave speed, linear).}$$

8.4 Continuous Dependence on the Initial Data

The techniques described in the previous section provide uniform estimates on the total variation of a solution u to the system (163). Similar techniques can also be used to estimate the size of first order perturbations.

Indeed, let u be solution of (163) and assume that, for each $\varepsilon > 0$, the function

$$u^\varepsilon(t,x) \;=\; u(t,x) + \varepsilon\, z(t,x) + o(\varepsilon)$$

is also a solution, with $o(\varepsilon)$ denoting an infinitesimal of higher order w.r.t. ε. Inserting the above expansion in (163) and collecting terms of order ε, one finds that the function z must satisfy the the linearized variational equation

$$z_t + \big[DA(u)\cdot z\big]u_x + A(u)z_x \;=\; z_{xx}. \tag{188}$$

Assuming that the total variation of u remains small, one can prove the estimate

$$\big\|z(t,\cdot)\big\|_{\mathbf{L}^1} \;\le\; L\,\big\|z(0,\cdot)\big\|_{\mathbf{L}^1} \qquad \text{for all } \; t\ge 0\,, \tag{189}$$

for a uniform constant L. The above estimate is valid for every solution u of (163) having small total variation and every $\mathbf{L}^1$ solution of the corresponding system (188).

Relying on (189), a standard homotopy argument yields the Lipschitz continuity of the flow of (163) w.r.t. the initial data, uniformly in time. Indeed, let any two solutions u, v of (163) be given (Fig. 46). We can connect them by a smooth path of solutions u^θ, whose initial data satisfy

$$u^\theta(0,x) \;\doteq\; \theta u(0,x) + (1-\theta)v(0,x) \qquad\qquad \theta\in[0,1].$$

The distance $\big\|u(t,\cdot)-v(t,\cdot)\big\|_{\mathbf{L}^1}$ at any later time $t>0$ is clearly bounded by the length of the path $\theta\mapsto u^\theta(t)$. In turn, this can be computed by integrating the norm of a tangent vector. Calling $z^\theta \doteq du^\theta/d\theta$, each vector z^θ is a solution of the corresponding (188), with u replaced by u^θ. Using (190) we thus obtain

$$\begin{aligned}\big\|u(t,\cdot)-v(t,\cdot)\big\|_{\mathbf{L}^1} \;&\le\; \int_0^1 \left\|\frac{d}{d\theta}u^\theta(t)\right\|_{\mathbf{L}^1} d\theta \;=\; \int_0^1 \big\|z^\theta(t)\big\|_{\mathbf{L}^1}\,d\theta\\ &\le\; L\int_0^1 \big\|z^\theta(0)\big\|_{\mathbf{L}^1}\,d\theta \;=\; L\,\big\|u(0,\cdot)-v(0,\cdot)\big\|_{\mathbf{L}^1}.\end{aligned} \tag{190}$$

8.5 *The Semigroup of Vanishing Viscosity Limit Solutions*

The estimates on the total variation and on the continuous dependence on the initial data, obtained in the previous sections were valid for solutions of the system (163) with unit viscosity matrix. By the simple rescaling of coordinates $t\mapsto\varepsilon t$, $x\mapsto\varepsilon x$, all of the above estimates remain valid for solutions u^ε of the system $(155)_\varepsilon$. In this way one obtains the a priori bounds (157) and (158).

As soon as the global BV bounds are established, by a compactness argument one obtains the existence of a strong limit $u^{\varepsilon_m} \to u$ in $\mathbf{L}^1_{loc}$, for some sequence $\varepsilon_m \to 0$. In the conservative case where $A = Df$, by Lemma 1 in Sect. 2 this limit $u = u(t, x)$ provides a weak solution to the Cauchy problem (162).

At this stage, it only remains to prove that the limit is unique, i.e. it does not depend on the choice of the sequence $\varepsilon_m \to 0$. For a system in conservative form, and with the standard assumption (H) that each field is either genuinely nonlinear or linearly degenerate, we can apply Theorem 7 in Sect. 7, and conclude that the limit of vanishing viscosity approximations is unique and coincides with the limit of Glimm and of front tracking approximations.

To handle the general non-conservative case, some additional work is required. Relying on the analysis in [6], one first considers Riemann initial data and shows that in this special case the vanishing viscosity solution is unique and can be accurately described. In a second step, one proves that any weak solution obtained as limit vanishing viscosity approximations is also a "viscosity solution", i.e. it satisfies the local integral estimates (E1)–(E2) in Sect. 7.2, where $U^\sharp$ is now the unique solution of a Riemann problem obtained as limit of viscous approximations [6]. By an argument introduced in [10], a Lipschitz semigroup is completely determined as soon as one specifies its local behavior for piecewise constant initial data. Characterizing its trajectories as "viscosity solutions" one thus establishes the uniqueness of the semigroup of vanishing viscosity limits.

9 Extensions and Open Problems

With the papers [7,20,34], the well-posedness of the Cauchy problem for hyperbolic conservation laws in one space dimension has been essentially settled, within the class of solutions with small total variation. Extensions of these well-posedness results to the initial-boundary value problem and to balance laws with source terms can be found in [30] and in [1], respectively.

A major remaining open problem concerns the solutions with large total variation. Results in this direction can be found in [23] and [35]. As proved by M. Lewicka [41], for a large class of hyperbolic systems the solutions are unique and depend continuously on the initial data, as long as their total variation remains bounded. The key question is whether the total variation can blow up in finite time, if the initial data is sufficiently large. An example constructed by K. Jenssen [38] shows that this can indeed happen, for some strictly hyperbolic system. One should remark, however, that the 3×3 system considered in [38] does not come from any realistic physical model. In particular, it does not admit any strictly convex entropy. One may thus conjecture that the presence of a strictly convex entropy restricts the possibility of a finite time blow up. More specifically, it is an important open problem to understand whether finite blow up in the total variation norm can occur for solutions to the Euler equations of gas dynamics.

We remark that, since hyperbolic conservation laws are a class of nonlinear evolution equations, one might expect to observe some rich dynamics: periodic orbits, bifurcation, chaotic behavior, etc... However, the present theory does not include any of this. The reason is that, as long as one considers only solutions with small total variation, the dynamics is mostly trivial. As proved by T.P. Liu [47], letting time $t \to +\infty$, every solution with small total variation converges asymptotically to the solution of a Riemann problem. It is only for large BV solutions that some interesting dynamics will likely be observed—provided that some global existence theorem can be established.

In connection with vanishing viscosity approximations, uniform BV bounds for systems of balance laws with dissipative sources were established in [24]. Viscous approximations to the initial-boundary value problem, with suitable boundary conditions, have been studied by Ancona and Bianchini [2].

Up to now, all results on a priori BV bounds, stability and convergence of viscous approximations have dealt with "artificial viscosity", assuming that the diffusion coefficient is independent of the state u. A more realistic model would be

$$u_t + f(u)_x = (B(u)u_x)_x\,, \tag{191}$$

where B is a positive definite viscosity matrix, possibly depending on the state u. It remains an outstanding open problem to establish similar results in connection with the more general system (191).

Appendix

We collect here some results of mathematical analysis, which were used in previous sections.

9.1 Compactness Theorems

Let Ω be an open subset of $I\!R^m$. We denote by $\mathbf{L}^1_{loc}(\Omega;\, I\!R^n)$ the space of locally integrable functions on Ω. This is the space of all functions $u : \Omega \mapsto I\!R^n$ whose restriction to every compact subset $K \subset \Omega$ is integrable. The space $\mathbf{L}^1_{loc}$ is not a normed space. However, it is a Fréchet space: for every compact $K \subset \Omega$, the mapping

$$u \mapsto \int_K |u(x)|\, dx$$

is a seminorm on $\mathbf{L}^1_{loc}$.

Next, consider a (possibly unbounded) interval $J \subseteq I\!R$ and a map $u : J \mapsto I\!R^n$. The *total variation* of u is defined as

$$\text{Tot.Var.}\{u\} \doteq \sup \left\{ \sum_{j=1}^{N} \left| u(x_j) - u(x_{j-1}) \right| \right\}, \tag{1}$$

where the supremum is taken over all $N \geq 1$ and all $(N+1)$-tuples of points $x_j \in J$ such that $x_0 < x_1 < \cdots < x_N$. If the right hand side of (1) is bounded, we say that u has bounded variation, and write $u \in BV$.

Lemma A.1 (properties of functions with bounded variation). *Let $u :]a,b[\mapsto I\!R^n$ have bounded variation. Then, for every $x \in]a,b[$, the left and right limits*

$$u(x-) \doteq \lim_{y \to x-} u(y), \qquad\qquad u(x+) \doteq \lim_{y \to x+} u(y)$$

are well defined. Moreover, u has at most countably many points of discontinuity.

By the above lemma, if u has bounded variation, we can redefine the value of u at each point of jump by setting $u(x) \doteq u(x+)$. In particular, if we are only interested in the $\mathbf{L}^1$-equivalence class of a BV function u, by possibly changing the values of u at countably many points we can assume that u is right continuous.

We state below a version of Helly's compactness theorem, which provides the basic tool in the proof of existence of weak solutions. For a proof, see [11].

Theorem A.1 (Compactness for a family of BV functions). *Consider a sequence of functions $u_\nu : [0,\infty[\times I\!R \mapsto I\!R^n$ with the following properties.*

$$\textit{Tot.Var.}\big\{u_\nu(t,\cdot)\big\} \leq C, \qquad \big|u_\nu(t,x)\big| \leq M \qquad \textit{for all } t,x, \tag{2}$$

$$\int_{-\infty}^{\infty} \big|u_\nu(t,x) - u_\nu(s,x)\big|\,dx \leq L|t-s| \qquad \textit{for all } t,s \geq 0, \tag{3}$$

for some constants C, M, L. Then there exists a subsequence u_μ which converges to some function u in $\boldsymbol{L}^1_{\mathrm{loc}}\big([0,\infty) \times I\!R;\ I\!R^n\big)$. This limit function satisfies

$$\int_{-\infty}^{\infty} \big|u(t,x) - u(s,x)\big|\,dx \leq L|t-s| \qquad \textit{for all } t,s \geq 0. \tag{4}$$

The point values of the limit function u can be uniquely determined by requiring that

$$u(t,x) = u(t,x+) \doteq \lim_{y \to x+} u(t,y) \qquad \textit{for all } t,x. \tag{5}$$

In this case, one has

$$\textit{Tot.Var.}\big\{u(t,\cdot)\big\} \leq C, \qquad \big|u(t,x)\big| \leq M \qquad \textit{for all } t,x. \tag{6}$$

9.2 An Elementary Error Estimate

Let $\mathscr{D}$ be a closed subset of a Banach space E and consider a Lipschitz continuous semigroup $S : \mathscr{D} \times [0,\infty[\mapsto \mathscr{D}$. More precisely, assume that

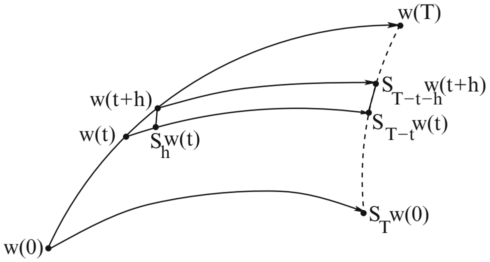


Fig. 65 Comparing the approximate solution w with the trajectory of the semigroup having the same initial data

(i) $S_0 u = u, \qquad S_s S_t u = S_{s+t} u.$
(ii) $\|S_t u - S_s v\| \ \leq \ L \cdot \|u - v\| + L' \cdot |t - s|.$

Given a Lipschitz continuous map $w : [0, T] \mapsto \mathscr{D}$, the following theorem estimates the difference between w and the trajectory of the semigroup S starting at $w(0)$. For the proof we again refer to [11].

Theorem A.2 (Error estimate for a Lipschitz flow). *Let $S : \mathscr{D} \times [0, \infty[\mapsto \mathscr{D}$ be a continuous flow satisfying the properties (i)–(ii). For every Lipschitz continuous map $w : [0, T] \mapsto \mathscr{D}$ one then has the estimate*

$$\left\| w(T) - S_{T} w(0) \right\| \ \leq L \ \int_0^T \left\{ \liminf_{h \to 0+} \ \frac{\left\| w(t+h) - S_h w(t) \right\|}{h} \right\} \ dt. \tag{7}$$

Remark 9. The integrand in (7) can be regarded as the instantaneous error rate for w at time t. Since the flow is uniformly Lipschitz continuous, during the time interval $[t, T]$ this error is amplified at most by a factor L (see Fig. 65).

9.3 The Center Manifold Theorem

Let A be an $n \times n$ matrix and consider the Cauchy problem for a linear system of O.D.E's with constant coefficients

$$\dot{x} = Ax\,, \qquad\qquad x(0) = \bar{x}. \tag{8}$$

The explicit solution can be written as

$$x(t) = e^{tA} \bar{x}\,, \qquad\qquad e^{tA} \doteq \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}.$$

We say that a subspace $V \subset I\!R^n$ is **invariant** for the flow of (8) if $x \in V$ implies $e^{At} x \in V$ for all $t \in I\!R$. A natural way to decompose the space $I\!R^n$ as the sum of three invariant subspaces is now described. Consider the eigenvalues of A, i.e. the

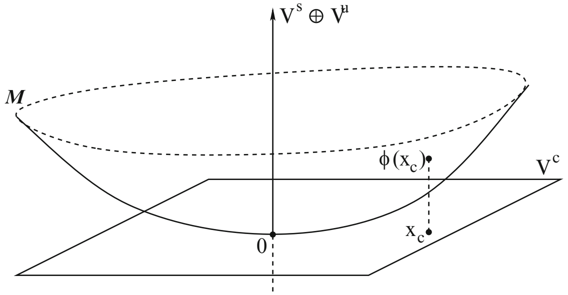


Fig. 66 The center subspace V^c and the center manifold $\mathscr{M}$, tangent to V^c at the origin

zeroes of the polynomial $p(\zeta) \doteq \det(\zeta I - A)$. These are finitely many points in the complex plane.

The space $I\!R^n$ can then be decomposed as the sum of a stable, an unstable and a center subspace, respectively spanned by the (generalized) eigenvectors corresponding to eigenvalues with negative, positive and zero real part. We thus have

$$I\!R^n = V^s \oplus V^u \oplus V^c$$

with continuous projections

$$\pi_s : I\!R^n \mapsto V^s\,, \qquad \pi_u : I\!R^n \mapsto V^u\,, \qquad \pi_c : I\!R^n \mapsto V^c\,,$$

$$x \;=\; \pi_s x + \pi_c x + \pi_u x.$$

These projections commute with A and hence with the exponential e^{At} as well:

$$\pi_s e^{At} = e^{At}\pi_s\,, \qquad \pi_u e^{At} = e^{At}\pi_u\,, \qquad \pi_c e^{At} = e^{At}\pi_c.$$

In particular, these subspaces are invariant for the flow of (8).

Next, consider the nonlinear system

$$\dot{x} = f(x). \tag{9}$$

Assume that $f(0) = 0$ and $Df(0) = A$, so that (8) provides a first order Taylor approximation to (9). According to the center manifold theorem, the nonlinear system (9) admits an invariant manifold $\mathscr{M}$, which at the origin is tangent to the center subspace V^c, as shown in Fig. 66. In the following theorem, the solution of (9) with initial data $x(0) = x_0$ will be denoted by $t \mapsto x(t, x_0)$. For a proof we refer to [13].

Theorem A.3 (Existence and properties of center manifold). *Let $f : I\!R^n \mapsto I\!R^n$ be a vector field in $\mathscr{C}^{k+1}$ (here $k \geq 1$), with $f(0) = 0$. Consider the matrix $A = Df(0)$, and let V^s, V^u, V^c be the corresponding stable, unstable, and center subspaces. Then there exists $\delta > 0$ and a local center manifold $\mathscr{M}$ with the following properties.*

(i) There exists a $\mathscr{C}^k$ function $\phi : V^c \mapsto I\!R^n$ with $\pi_c\,\phi(x_c) = x_c$ such that

$$\mathscr{M} \;=\; \Big\{\phi(x_c)\,;\quad x_c \in V^c\,,\quad |x_c| < \delta\Big\}.$$

(ii) The manifold $\mathscr{M}$ is locally invariant for the flow of (9), i.e. $x_0 \in \mathscr{M}$ implies $x(t, x_0) \in \mathscr{M}$, for all t sufficiently close to zero.

(iii) $\mathscr{M}$ is tangent to V^c at the origin.

(iv) Every globally bounded orbit remaining in a suitably small neighborhood of the origin is entirely contained inside $\mathscr{M}$.

(v) Given any trajectory such that $x(t) \to 0$ as $t \to +\infty$, there exists $\eta > 0$ and a trajectory $t \mapsto y(t) \in \mathscr{M}$ on the center manifold such that

$$e^{\eta t}\,\big|x(t) - y(t)\big| \to 0 \qquad\qquad as \quad t \to +\infty.$$

References

1. D. Amadori, L. Gosse, G. Guerra, Global BV entropy solutions and uniqueness for hyperbolic systems of balance laws. Arch. Ration. Mech. Anal. **162**, 327–366 (2002)
2. F. Ancona, S. Bianchini, Vanishing viscosity solutions of hyperbolic systems of conservation laws with boundary, in *"WASCOM 2005"–13th Conference on Waves and Stability in Continuous Media* (World Scientific, Hackensack, 2006), pp. 13–21
3. F. Ancona, A. Marson, Existence theory by front tracking for general nonlinear hyperbolic systems. Arch. Ration. Mech. Anal. **185**, 287–340 (2007)
4. F. Ancona, A. Marson, A locally quadratic Glimm functional and sharp convergence rate of the Glimm scheme for nonlinear hyperbolic systems. Arch. Ration. Mech. Anal. **196**, 455–487 (2010)
5. P. Baiti, H.K. Jenssen, On the front tracking algorithm. J. Math. Anal. Appl. **217**, 395–404 (1998)
6. S. Bianchini, On the Riemann problem for non-conservative hyperbolic systems. Arch. Ration. Mech. Anal. **166**, 1–26 (2003)
7. S. Bianchini, A. Bressan, Vanishing viscosity solutions to nonlinear hyperbolic systems. Ann. Math. **161**, 223–342 (2005)
8. A. Bressan, Contractive metrics for nonlinear hyperbolic systems. Indiana Univ. J. Math. **37**, 409–421 (1988)
9. A. Bressan, Global solutions of systems of conservation laws by wave-front tracking. J. Math. Anal. Appl. **170**, 414–432 (1992)
10. A. Bressan, The unique limit of the Glimm scheme. Arch. Ration. Mech. Anal. **130**, 205–230 (1995)
11. A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem* (Oxford University Press, Oxford, 2000)
12. A. Bressan, BV solutions to systems of conservation laws by vanishing viscosity, C.I.M.E. course in Cetraro, 2003, in *Springer Lecture Notes in Mathematics*, vol. 1911, ed. by P. Marcati (Springer-Verlag, Berlin, 2007), pp. 1–78
13. A. Bressan, A tutorial on the Center Manifold Theorem, C.I.M.E. course in Cetraro, 2003, in *Springer Lecture Notes in Mathematics*, vol. 1911, ed. by P. Marcati (Springer-Verlag, Berlin, 2007), pp. 327–344

14. A. Bressan, R.M. Colombo, The semigroup generated by 2×2 conservation laws. Arch. Ration. Mech. Anal. **133**, 1–75 (1995)
15. A. Bressan, P. Goatin, Oleinik type estimates and uniqueness for $n \times n$ conservation laws. J. Differ. Equat. **156**, 26–49 (1999)
16. A. Bressan, P. LeFloch, Uniqueness of weak solutions to hyperbolic systems of conservation laws. Arch. Ration. Mech. Anal. **140**, 301–317 (1997)
17. A. Bressan, M. Lewicka, A uniqueness condition for hyperbolic systems of conservation laws. Discrete. Cont. Dyn. Syst. **6**, 673–682 (2000)
18. A. Bressan, A. Marson, Error bounds for a deterministic version of the Glimm scheme. Arch. Ration. Mech. Anal. **142**, 155–176 (1998)
19. A. Bressan, T. Yang, On the convergence rate of vanishing viscosity approximations. Comm. Pure Appl. Math. **57**, 1075–1109 (2004)
20. A. Bressan, T.P. Liu, T. Yang, L^1 stability estimates for $n \times n$ conservation laws. Arch. Ration. Mech. Anal. **149**, 1–22 (1999)
21. A. Bressan, G. Crasta, B. Piccoli, Well posedness of the Cauchy problem for $n \times n$ systems of conservation laws. Am. Math. Soc. Mem. **694** (2000)
22. A. Bressan, K. Jenssen, P. Baiti, An instability of the Godunov scheme. Comm. Pure Appl. Math. **59**, 1604–1638 (2006)
23. C. Cheverry, Systèmes de lois de conservation et stabilité BV [Systems of conservation laws and BV stability]. Mém. Soc. Math. Fr. **75** (1998) (in French)
24. C. Christoforou, Hyperbolic systems of balance laws via vanishing viscosity. J. Differ. Equat. **221**, 470–541 (2006)
25. M.G. Crandall, The semigroup approach to first order quasilinear equations in several space variables. Isr. J. Math. **12** 108–132, (1972)
26. C. Dafermos, Polygonal approximations of solutions of the initial value problem for a conservation law. J. Math. Anal. Appl. **38**, 33–41 (1972)
27. C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics* (Springer, Berlin, 1999)
28. R.J. DiPerna, Global existence of solutions to nonlinear hyperbolic systems of conservation laws. J. Differ. Equat. **20**, 187–212 (1976)
29. R. DiPerna, Convergence of approximate solutions to conservation laws. Arch. Ration. Mech. Anal. **82**, 27–70 (1983)
30. C. Donadello, A. Marson, Stability of front tracking solutions to the initial and boundary value problem for systems of conservation laws. Nonlinear Differ. Equat. Appl. **14**, 569-592 (2007)
31. L.C. Evans, *Partial Differential Equations* (American Mathematical Society, Providence, 1998)
32. L.C. Evans, R.F. Gariepy, *Measure Theory and Fine Properties of Functions* (CRC Press, Boca Raton, Fl, 1992)
33. M. Garavello, B. Piccoli, in *Traffic Flow on Networks. Conservation Laws Models* (AIMS Series on Applied Mathematics, Springfield, 2006)
34. J. Glimm, Solutions in the large for nonlinear hyperbolic systems of equations. Comm. Pure Appl. Math. **18**, 697–715 (1965)
35. J. Glimm, P. Lax, Decay of solutions of systems of nonlinear hyperbolic conservation laws. Am. Math. Soc. Mem. **101** (1970)
36. J. Goodman, Z. Xin, Viscous limits for piecewise smooth solutions to systems of conservation laws. Arch. Ration. Mech. Anal. **121**, 235–265 (1992)
37. H. Holden, N.H. Risebro, *Front Tracking for Hyperbolic Systems of Conservation Laws* (Springer, New York, 2002)
38. H.K. Jenssen, Blowup for systems of conservation laws. SIAM J. Math. Anal. **31**, 894–908 (2000)
39. S. Kruzhkov, First-order quasilinear equations with several space variables. Mat. Sb. **123**, 228–255 (1970). English translation in Math. USSR Sb. **10**, 217–273 (1970)
40. P.D. Lax, Hyperbolic systems of conservation laws II. Comm. Pure Appl. Math. **10**, 537–566 (1957)
41. M. Lewicka, Well-posedness for hyperbolic systems of conservation laws with large BV data. Arch. Ration. Mech. Anal. **173**, 415–445 (2004)

42. T.-T. Li, *Global Classical Solutions for Quasilinear Hyperbolic Systems* (Wiley, Chichester, 1994)
43. M. Lighthill, G. Whitham, On kinematic waves, II. A theory of traffic flow on long crowded roads. Proc. R. Soc. Lond. A **229**, 317–345 (1955)
44. T.P. Liu, The Riemann problem for general systems of conservation laws. J. Differ. Equat. **18**, 218–234 (1975)
45. T.P. Liu, The entropy condition and the admissibility of shocks. J. Math. Anal. Appl. **53**, 78–88 (1976)
46. T.P. Liu, The deterministic version of the Glimm scheme. Comm. Math. Phys. **57**, 135–148 (1977)
47. T.P. Liu, Linear and nonlinear large-time behavior of solutions of general systems of hyperbolic conservation laws. Comm. Pure Appl. Math. **30**, 767–796 (1977)
48. T.-P. Liu, T. Yang, L^1 stability of conservation laws with coinciding Hugoniot and characteristic curves. Indiana Univ. Math. J. **48**, 237–247 (1999)
49. T.-P. Liu, T. Yang, L^1 stability of weak solutions for 2×2 systems of hyperbolic conservation laws. J. Am. Math. Soc. **12**, 729–774 (1999)
50. Y. Lu, *Hyperbolic Conservation Laws and the Compensated Compactness Method* (Chapman & Hall/CRC, Boca Raton, 2003)
51. O. Oleinik, Discontinuous solutions of nonlinear differential equations. Am. Math. Soc. Transl. **26**, 95–172 (1963)
52. B. Riemann, Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. Göttingen Abh. Math. Cl. **8**, 43–65 (1860)
53. F. Rousset, Viscous approximation of strong shocks of systems of conservation laws. SIAM J. Math. Anal. **35**, 492–519 (2003)
54. S. Schochet, Sufficient conditions for local existence via Glimm's scheme for large BV data. J. Differ. Equat. **89**, 317–354 (1991)
55. S. Schochet, The essence of Glimm's scheme, in *Nonlinear Evolutionary Partial Differential Equations*, ed. by X. Ding, T.P. Liu (American Mathematical Society/International Press, Providence, RI, 1997), pp. 355–362
56. D. Serre, *Systems of Conservation Laws*, vols. 1–2 (Cambridge University Press, Cambridge, 1999)
57. J. Smoller, *Shock Waves and Reaction-Diffusion Equations* (Springer, New York, 1983)
58. S.H. Yu, Zero-dissipation limit of solutions with shocks for systems of hyperbolic conservation laws. Arch. Ration. Mech. Anal. **146**, 275–370 (1999)
59. A.I. Volpert, The spaces BV and quasilinear equations. Math. USSR Sbornik **2**, 225–267 (1967)

Derivation of Non-local Macroscopic Traffic Equations and Consistent Traffic Pressures from Microscopic Car-Following Models*

Dirk Helbing

Abstract This contribution compares several different approaches allowing one to derive macroscopic traffic equation directly from microscopic car-following models. While it is shown that some conventional approaches lead to theoretical problems, it is proposed to use an approach reminding of smoothed particle hydrodynamics to avoid gradient expansions. The derivation circumvents approximations and, therefore, demonstrates the large range of validity of macroscopic traffic equations, without the need of averaging over many vehicles. It also gives an expression for the "traffic pressure", which generalizes previously used formulas. Furthermore, the method avoids theoretical inconsistencies of macroscopic traffic models, which have been criticized in the past by Daganzo and others.

1 Introduction

In order to describe the dynamics of traffic flows, a large number of mathematical models has been developed. The analysis of the spatio-temporal features and statistics of traffic patterns has often been done with methods from non-linear dynamics and statistical physics. An overview of modeling approaches and methods is, for example, given in [5, 9, 19, 20]. Among these are cellular automata, "microscopic" car-following models, "mesoscopic" gas-kinetic, and macroscopic traffic models.

*First published in: The European Physical Journal B **69**(4), 539–548, DOI: 10.1140/epjb/e2009-00192-5 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ).

D. Helbing (✉)
ETH Zurich, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland
e-mail: dhelbing@ethz.ch

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_3,

Cellular automata can often be interpreted as discretized versions of car-following models, while gas-kinetic models have frequently been used to derive macroscopic from microscopic models. Such derivations were driven by the desire to improve phenomenological specifications of macroscopic traffic models [15, 16, 22], which were criticized to have unrealistic properties [6]. However, the derivation of gas-kinetic models from car-following models usually simplifies the interactions among vehicles by a collisional approach assuming immediate braking maneuvers. Moreover, the derivation of macroscopic traffic models from gas-kinetic ones terminates an infinite and poorly converging series expansion, which replaces dynamical equations for higher moments of the velocity distribution by simplified equilibrium relationships [12].

Although this leads to macroscopic equations which work well in most theoretical and practical aspects [26], the implications of the approximations are hardly known. Moreover, the approach seems to require an averaging over at least 100 vehicles for each speed class and spatial location. While this constitutes no problem for gases with 10^{23} particles within a small volume, for traffic flows this would require an averaging over spatial intervals much greater than the scale on which traffic flow changes. Hence, it is not well understood, whether or why macroscopic traffic equations can be used at all.

In this paper, we will therefore focus on attempts to derive macroscopic traffic equations directly from microscopic ones. Doing so, we will compare three different approaches: First, we study the gradient expansion approach in Sect. 2. Second, we turn to the linear interpolation approach in Sect. 3. Third, we discuss an approach reminding of smoothed particle hydrodynamics in Sect. 4 and compare the results with macroscopic traffic models such as the Payne model, the Aw–Rascle model, and a non-local traffic model. In the conclusions of Sect. 5, we summarize and discuss our results, in particular with regard to the mathematical form of the traffic pressure and the theoretical consistency of macroscopic traffic models.

2 The Gradient Expansion Approach

Already in the 1970s, Payne [23, 24] used a gradient expansion approach to derive a macroscopic velocity equation complementing the continuity equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}\left[\rho(x,t)V(x,t)\right] = 0. \tag{1}$$

It relates the vehicle density $\rho(x,t)$ at location x and time t with the average velocity $V(x,t)$ or the vehicle flow

$$Q(x,t) = \rho(x,t)V(x,t), \tag{2}$$

respectively, and describes the conservation of the number of vehicles [28].

Payne derived his model from Newell's car-following model [21]

$$v_i(t+\tau) = v_o\big(d_i(t)\big)\,, \tag{3}$$

which assumes that the speed $v_i(t)$ of vehicle i at time t will be adjusted with a delay of τ to some optimal speed v_o, which depends on the distance $d_i(t) = x_{i-1}(t) - x_i(t)$ between the location of the leading vehicle $x_{i-1}(t)$ and the location $x_i(t)$ of the following car.

Payne identified microscopic and macroscopic velocities as follows:

$$\begin{aligned} v_i(t+\tau) &= V(x + V\,\tau, t+\tau) \\ &\approx V(x,t) + V\,\tau\,\frac{\partial V(x,t)}{\partial x} + \tau\,\frac{\partial V(x,t)}{\partial t}. \end{aligned} \tag{4}$$

Then, Taylor approximations (gradient expansions) were used in several places. For example, Payne substituted the inverse of the distance d_i to the leading vehicle by the density ρ at the place $x + d_i(t)/2$ in the middle between the leading and the following vehicle. In this way, he obtained

$$\begin{aligned} \frac{1}{d_i(t)} &= \rho\left(x + \frac{d_i(t)}{2}, t\right) = \rho\left(x + \frac{1}{2\rho}, t\right) \\ &\approx \rho(x,t) + \frac{1}{2\rho}\frac{\partial\rho(x,t)}{\partial x}\,. \end{aligned} \tag{5}$$

When defining the so-called equilibrium velocity $V_e(\rho)$ through

$$V_e(\rho) = v_o\left(\frac{1}{\rho}\right) \qquad \text{or} \qquad V_e\left(\frac{1}{d_i}\right) = v_o(d_i)\,, \tag{6}$$

a first order Taylor approximation and (5) imply

$$\begin{aligned} v_o\big(d_i(t)\big) &= V_e\left(\frac{1}{d_i(t)}\right) \\ &\approx V_e\big(\rho(x,t)\big) + \frac{1}{2\rho(x,t)}\,\frac{dV_e(\rho)}{d\rho}\,\frac{\partial\rho(x,t)}{\partial x}\,. \end{aligned} \tag{7}$$

Starting from the previous equations, one finally arrives at Payne's macroscopic velocity equation

$$\frac{\partial V}{\partial t} + V\frac{\partial V}{\partial x} = \frac{1}{\tau}\left[V_e(\rho) - \frac{D(\rho)}{\rho}\frac{\partial\rho}{\partial x} - V(x,t)\right], \tag{8}$$

where we have introduced the density-dependent diffusion

$$D(\rho) = -\frac{1}{2}\frac{dV_{\rm e}(\rho)}{\partial\rho} = \frac{1}{2}\left|\frac{dV_{\rm e}(\rho)}{d\rho}\right| \geq 0\,. \tag{9}$$

The single terms of (8) have the following interpretation: The term $V\partial V/\partial x$ is called the *transport term* and describes a motion of the velocity profile with the vehicles. The term $-[D(\rho)/(\rho\,\Delta t)]\partial\rho/\partial x$ is called *anticipation term*, as it reflects the reaction of drivers to the traffic situation in front of them. The *relaxation term* $[V_{\rm e}(\rho) - V]/\Delta t$ delineates the adaptation of the average velocity $V(x,t)$ to the density-dependent *equilibrium velocity* $V_{\rm e}(\rho)$ with a delay τ.

Other authors have applied similar gradient expansions to the optimal velocity model defined by

$$\frac{dv_i(t)}{dt} = \frac{1}{\tau}\Big[v_{\rm o}\big(d_i(t)\big) - v_i(t)\Big] \tag{10}$$

with $dd_i/dt = v_{i-1}(t) - v_i(t)$, see e.g. [4,17]. Equation (10) results from the Newell model (3) by a first-order Taylor approximation $v_i(t+\tau) \approx v_i(t) + \tau\, dv_i/dt$. Regarding the derivation of macroscopic traffic equations from the optimal velocity model, it is also worth reading [4, 17].

One weakness of the gradient expansion approach is that its validity implicitly requires small gradients. However, it is well-known that many microscopic and macroscopic traffic equations give rise to emergent traffic jams, which are related with *steep* gradients. That calls for the consideration of higher-order terms and leads to macroscopic traffic equations that are not anymore simple and well tractable (even numerically). Let us, therefore, study other approaches to determine macroscopic from microscopic equations.

3 The Linear Interpolation Approach

The optimal velocity model may be also written in the form

$$\frac{dv_i}{dt} = a_i(t) = \frac{v^0 - v_i(t)}{\tau} + f\big(d_i(t)\big)\,, \tag{11}$$

where $a_i(t)$ denotes the acceleration, v^0 the "desired velocity" or "free speed", and

$$f(d_i) = \frac{v_{\rm o}(d_i) - v^0}{\tau} \leq 0 \tag{12}$$

the repulsive interaction among the leading vehicle $i-1$ and its follower i.

In [14], it has been suggested to establish a micro-macro link between microscopic and macroscopic traffic variables by the definitions

$$\rho(x,t) = \frac{\frac{1}{x_i(t) - x_{i+1}(t)}\big[x_{i-1}(t) - x\big]}{x_{i-1}(t) - x_i(t)} + \frac{\frac{1}{x_{i-1}(t) - x_i(t)}\big[x - x_i(t)\big]}{x_{i-1}(t) - x_i(t)}, \tag{13}$$

$$V(x,t) = \frac{v_i(t)\big[x_{i-1}(t) - x\big] + v_{i-1}(t)\big[x - x_i(t)\big]}{x_{i-1}(t) - x_i(t)}, \tag{14}$$

$$A(x,t) = \frac{a_i(t)\big[x_{i-1}(t) - x\big] + a_{i-1}(t)\big[x - x_i(t)\big]}{x_{i-1}(t) - x_i(t)}. \tag{15}$$

These definitions assume that the macroscopic variables in the vehicle locations $x = x_i(t)$ would be given by the microscopic ones, while in locations x *between* two vehicles, they would be defined by linear interpolation.

Let us consider the consequences of such an approach. For this, we determine the partial derivative of

$$G(x,t) = \frac{g_i(t)\big[x_{i-1}(t) - x\big] + g_{i-1}(t)\big[x - x_i(t)\big]}{x_{i-1}(t) - x_i(t)} \tag{16}$$

with respect to x, which gives

$$\frac{\partial G(x,t)}{\partial x} = \frac{-g_i(t) + g_{i-1}(t)}{x_{i-1}(t) - x_i(t)} \tag{17}$$

for any specification of $g_i(t)$, for example, $g_i(t) = v_i(t)$. The partial derivative with respect to time is

$$\begin{aligned}\frac{\partial G(x,t)}{\partial t} = {} & \frac{\frac{dg_i(t)}{dt}\big[x_{i-1}(t) - x\big] + g_i(t)\frac{dx_{i-1}(t)}{dt}}{x_{i-1}(t) - x_i(t)} \\ & + \frac{\frac{dg_{i-1}(t)}{dt}\big[x - x_i(t)\big] - g_{i-1}(t)\frac{dx_i(t)}{dt}}{x_{i-1}(t) - x_i(t)} \\ & - \frac{\left(\frac{dx_{i-1}(t)}{dt} - \frac{dx_i(t)}{dt}\right) g_i(t)\big[x_{i-1}(t) - x\big]}{\big[x_{i-1}(t) - x_i(t)\big]^2} \\ & - \frac{\left(\frac{dx_{i-1}(t)}{dt} - \frac{dx_i(t)}{dt}\right) g_{i-1}(t)\big[x - x_i(t)\big]}{\big[x_{i-1}(t) - x_i(t)\big]^2}.\end{aligned} \tag{18}$$

For $g_i(t) = v_i(t) = dx_i/dt$ and with $dv_i/dt = a_i(t)$, this formula simplifies to the following expression:

$$
\begin{aligned}
\frac{\partial V(x,t)}{\partial t} &= \frac{a_i(t)\big[x_{i-1}(t) - x\big] + v_i(t)v_{i-1}(t)}{x_{i-1}(t) - x_i(t)} \\
&\quad + \frac{a_{i-1}(t)\big[x - x_i(t)\big] - v_{i-1}(t)v_i(t)}{x_{i-1}(t) - x_i(t)} \\
&\quad - \frac{v_{i-1}(t) - v_i(t)}{x_{i-1}(t) - x_i(t)} \\
&\quad \times \frac{v_i(t)\big[x_{i-1}(t) - x\big] + v_{i-1}(t)\big[x - x_i(t)\big]}{x_{i-1}(t) - x_i(t)} \\
&= A(x,t) - \frac{\partial V(x,t)}{\partial x} V(x,t). \qquad (19)
\end{aligned}
$$

As a consequence, we find the exact relationship

$$
\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} = A(x,t). \qquad (20)
$$

This would be fully compatible with Payne's macroscopic traffic equation (8), if

$$
A(x,t) = \frac{1}{\tau}\Big[V_{\rm e}(\rho) - V(x,t)\Big] - \frac{D(\rho)}{\tau\rho(x,t)}\frac{\partial \rho}{\partial x}. \qquad (21)
$$

However, the expression for $g_i(t) = 1/[x_{i-1}(t) - x_i(t)]$ does not simplify in a way that would finally lead to the continuity equation (1). Therefore, a micro-macro link based on the linear interpolation (16) of the microscopic variables $g_i(t)$ does not exactly imply the conservation of the number of vehicles, i.e. it is theoretically not consistent. Nevertheless, it works surprisingly well in practise [14].

In the next section, we will see that the interpolation approach fails because it does not reflect the non-locality of the correct macroscopic traffic equations, see (39) or (47). The dependence on gradients makes the model too isotropic, while vehicles should only respond to the traffic situation ahead of them, but not behind them. This problem is usually taken care of by hyperbolic schemes such as the Godunov scheme, as used for example in [2]. This scheme naturally discretizes the velocity in a downwind way, which avoids the isotropy problem of Payne's model and similar ones [6].

To avoid this problem, [18] suggests a hybrid Lagrangian approach. This is based on a transformation into Lagrangian coordinates, i.e. a moving coordinate system. As a result, the continuity equation (1) becomes linear. For piecewise linear ρ and V, the result can then be transformed back into Eulerian coordinates, i.e. into the stationary frame of reference. In the following, we will present an alternative method that yields macroscopic traffic equations from microscopic ones directly, without the need of transformation into Lagrangian coordinates.

4 An Approach Reminding of Smooth Particle Hydrodynamics

4.1 Derivation of the Continuity Equation

In this section, we will start with the derivation of the continuity equation from the equation of motion $dx_i/dt = v_i$, using a "trick" that I learned from Isaac Goldhirsch. For this, we represent the location $x_i(t)$ of an element i in space by a delta function $\delta(x - x_i(t))$, which may be treated here like a very narrow Gaussian distribution. Moreover, we introduce a *symmetrical* smoothing function

$$s(x' - x) = s(|x' - x|) = s(x - x'), \tag{22}$$

for example, a Gaussian distribution with a finite variance or a differentiable approximation of a triangular function or a rectangular one. The smoothing function shall be normalized by demanding

$$\int\limits_{-\infty}^{\infty} dx'\, s(x' - x) = 1 \tag{23}$$

for any value of x. With this, we define the local density

$$\rho(x,t) = \int\limits_{-\infty}^{\infty} dx'\, s(x' - x) \sum_i \delta\big(x' - x_i(t)\big) \tag{24}$$

$$= \sum_i s(x_i(t) - x)\,. \tag{25}$$

Herein, we sum up over all particles i. Note that the replacement of the conventional formula $\sum_i \delta(x_i(t) - x)$ for the vehicle density by the formula $\sum_i s(x_i(t) - x)$ reminds of a substitution of point-like particles by "fuzzy" particles, which is the idea behind smoothed particle hydrodynamics. Nevertheless it should be remembered that we have formally related the smoothing function $s(x' - x)$ to locations x' in the stationary frame of reference, and not to the moving vehicles themselves.

Now, we define the average velocity $V(x,t)$ as usual via a weighted average with the weight function $\delta(x' - x_i(t))s(x' - x)$:

$$V(x,t) = \frac{\int\limits_{-\infty}^{\infty} dx' \sum\limits_i v_i(t)\delta\big(x' - x_i(t)\big)s(x'-x)}{\int\limits_{-\infty}^{\infty} dx' \sum\limits_i \delta\big(x - x_i(t)\big)s(x'-x)}$$

$$= \frac{\int\limits_{-\infty}^{\infty} dx' \sum\limits_i v_i(t)\delta\big(x' - x_i(t)\big)s(x'-x)}{\rho(x,t)}$$

$$= \frac{\sum\limits_i v_i(t)s(x_i(t)-x)}{\sum\limits_i s(x_i(t)-x)}$$

$$= \frac{\sum\limits_i v_i(t)s(x_i(t)-x)}{\rho(x,t)} . \tag{26}$$

This implies the well-known fluid-dynamic flow relationship

$$Q(x,t) = \rho(x,t)V(x,t). \tag{27}$$

Differentiation of (24) with respect to time and application of the chain rule gives

$$\frac{\partial\rho(x,t)}{\partial t}$$

$$= \int\limits_{-\infty}^{\infty} dx' \sum_i \left(-\frac{dx_i}{dt}\right) \cdot \left[\frac{\partial}{\partial x'}\delta\Big(x' - x_i(t)\Big)\right] s(x'-x)$$

$$= \int\limits_{-\infty}^{\infty} dx' \sum_i v_i(t)\delta\Big(x' - x_i(t)\Big)\left[\frac{\partial}{\partial x'}s(x'-x)\right] , \tag{28}$$

where we have applied partial integration to obtain the last results. That is, we have used the theorem

$$\int\limits_{-\infty}^{\infty} dx' \left[\frac{\partial}{\partial x'}u(x')\right] v(x')$$

$$= \Big[u(x)v(x)\Big]_{-\infty}^{\infty} - \int\limits_{-\infty}^{\infty} u(x')\left[\frac{\partial}{\partial x'}v(x')\right], \tag{29}$$

considering the vanishing of the first term after the equality sign due to the vanishing of $u(x)v(x)$ at the boundaries. Taking into account the symmetry of the smoothing function $s(x'-x)$, we may replace $\partial s(x'-x)/\partial x'$ by $-\partial s(x'-x)/\partial x$, which finally yields (1) as follows:

$$\begin{aligned}\frac{\partial\rho(x,t)}{\partial t} &= -\frac{\partial}{\partial x}\int\limits_{-\infty}^{\infty} dx' \sum_i v_i(t)\delta\Big(x' - x_i(t)\Big)s(x'-x) \\ &= -\frac{\partial}{\partial x}\Big[\rho(x,t)V(x,t)\Big] . \end{aligned} \tag{30}$$

To obtain this desired result, we have finally applied the definition (26) of the average velocity $V(x,t)$. As a consequence of this, the validity of the continuity equation does not require an averaging over large numbers of entities, i.e. macroscopic volumes to average over. This makes the equation absolutely fundamental and explains its large range of validity.

4.2 Derivation of the Macroscopic Velocity Equation

In order to derive the equation for the average velocity, we start by deriving the formula

$$\rho(x,t)V(x,t) = \sum_i v_i(t)s\big(x_i(t) - x\big) \tag{31}$$

for the vehicle flow with respect to time. This gives

$$\begin{aligned}\frac{\partial}{\partial t}\big[\rho(x,t)V(x,t)\big] &= \sum_i \frac{dv_i(t)}{dt}s\big(x_i(t)-x\big) \\ &\quad + \sum_i v_i(t)\frac{\partial}{\partial x_i}\Big[s\big(x_i(t)-x\big)\Big]\frac{dx_i(t)}{dt} \\ &= \sum_i a_i(t)s\big(x_i(t)-x\big) \\ &\quad - \frac{\partial}{\partial x}\sum_i [v_i(t)]^2 s\big(x_i(t)-x\big) . \end{aligned} \tag{32}$$

Introducing $\delta v_i(x,t) = v_i(t) - V(x,t)$ and defining the velocity variance

$$\theta(x,t) = \frac{\int\limits_{-\infty}^{\infty} dx' \ \sum_i [v_i(t) - V(x,t)]^2 \delta\big(x' - x_i(t)\big) s(x' - x)}{\int\limits_{-\infty}^{\infty} dx' \ \sum_i \delta\big(x' - x_i(t)\big) s(x' - x)}$$
$$= \frac{\sum_i [v_i(t) - V(x,t)]^2 s(x_i(t) - x)}{\sum_i s(x_i(t) - x)}$$
$$= \frac{\sum_i [\delta v_i(x,t)]^2 s(x_i(t) - x)}{\rho(x,t)} \qquad (33)$$

similarly to the average velocity (26), we can make the decomposition

$$\sum_i [v_i(t)]^2 s\big(x_i(t) - x\big)$$
$$= \sum_i [V(x,t) + \delta v_i(x,t)]^2 s\big(x_i(t) - x\big)$$
$$= \sum_i \Big\{[V(x,t)]^2 + 2V(x,t)\delta v_i(x,t)$$
$$+ [\delta v_i(x,t)]^2\Big\} s\big(x_i(t) - x\big)$$
$$= \rho(x,t)[V(x,t)]^2 + 2\rho(x,t)V(x,t)\big[V(x,t) - V(x,t)\big]$$
$$+ \rho(x,t)\theta(x,t)\,, \qquad (34)$$

where we have considered

$$\sum_i \delta v_i(x,t) s\big(x_i(t) - x\big)$$
$$= \sum_i \Big[v_i(t) - V(x,t)\Big] s\big(x_i(t) - x\big)$$
$$= Q(x,t) - \rho(x,t)V(x,t) = 0\,, \qquad (35)$$

see (26) and (25). Altogether, we get

$$\frac{\partial}{\partial t}\big[\rho(x,t)V(x,t)\big] = -\frac{\partial}{\partial x}\Big\{\rho(x,t)\big[V(x,t)^2 + \theta(x,t)\big]\Big\}$$
$$+ \sum_i a_i(t) s\big(x_i(t) - x\big)\,. \qquad (36)$$

Now, we carry out the partial differentiation applying the product rule of Calculus. Taking into account

$$\rho(x,t)\frac{\partial V(x,t)}{\partial t} = -V(x,t)\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial}{\partial t}\big[\rho(x,t)V(x,t)\big] \tag{37}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial x}\Big\{\big[\rho(x,t)V(x,t)\big]V(x,t)\Big\} \\
&\quad = \rho(x,t)V(x,t)\frac{\partial V}{\partial x} \\
&\qquad + V(x,t)\frac{\partial}{\partial x}\Big[\rho(x,t)V(x,t)\Big],
\end{aligned} \tag{38}$$

with (36) we obtain the following:

$$\begin{aligned}
&\rho(x,t)\frac{\partial V(x,t)}{\partial t} \\
&\quad = -V(x,t)\frac{\partial \rho(x,t)}{\partial t} - V(x,t)\frac{\partial}{\partial x}\big[\rho(x,t)V(x,t)\big] \\
&\qquad -\rho(x,t)V(x,t)\frac{\partial V(x,t)}{\partial x} - \frac{\partial}{\partial x}\big[\rho(x,t)\theta(x,t)\big] \\
&\qquad + \sum_i a_i(t)s\big(x_i(t)-x\big)\,.
\end{aligned} \tag{39}$$

Inserting the continuity equation (30) for $\partial\rho/\partial t$ and dividing the above equation by $\rho(x,t)$ finally yields the velocity equation

$$\begin{aligned}
&\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} \\
&\quad = -\frac{1}{\rho(x,t)}\frac{\partial}{\partial x}\big[\rho(x,t)\theta(x,t)\big] \\
&\qquad + \frac{1}{\rho(x,t)}\sum_i a_i(t)s\big(x_i(t)-x\big)\,.
\end{aligned} \tag{40}$$

Inserting (11) for $a_i(t)$, we find

$$\begin{aligned}
&\sum_i a_i(t)s\big(x_i(t)-x\big) \\
&\quad = \sum_i \left[\frac{v^0 - v_i}{\tau} + \sum_i f\big(d_i(t)\big)\right] s\big(x_i(t)-x\big) \\
&\quad = \frac{v^0 - V(x,t)}{\tau} + \sum_i f\big(d_i(t)\big)s\big(x_i(t)-x\big)\,.
\end{aligned} \tag{41}$$

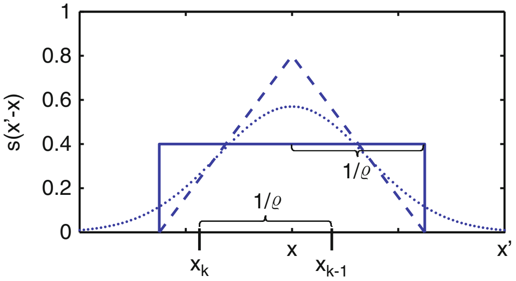


Fig. 1 Illustration of *rectangular* (—), *triangular* (– –), and Gaussian (···) smoothing functions $s(x' - x)$. x_k and x_{k-1} are the locations of the two closest vehicles k and $k - 1$ with respect to a reference location x. Their distance $1/\varrho = x_{k-1} - x_k$ determines the size $2/\varrho$ of the smoothing range chosen in the calculations of the main text

For further simplification, let us now specify the smoothing function by the rectangular function

$$s(x_i - x) = \frac{\varrho}{2} \cdot \begin{cases} 1 \text{ if } |x_i - x| \le 1/\varrho \\ 0 \text{ otherwise,} \end{cases} \tag{42}$$

with a large enough smoothing window of length $\Delta x = 2/\varrho$ (see Fig. 1). Then, the number of vehicles i within the smoothing interval $[x - 1/\varrho, x + 1/\varrho]$ is expected to be $\rho\,\Delta x = 2\rho/\varrho$, where ρ represents the average vehicle density in this interval. Therefore,

$$\rho(x,t) = \sum_i s\big(x_i(t) - x\big) = \frac{2\rho}{\varrho}\frac{\varrho}{2} = \rho, \tag{43}$$

which shows the consistency of this approach.

If the smoothing parameter ϱ is specified via the inverse vehicle distance

$$\varrho = \varrho_k = \frac{1}{d_k} = \frac{1}{x_{k-1} - x_k} = \rho(x,t) \quad \text{for} \quad x_k < x \le x_{k-1}, \tag{44}$$

the smoothing window of length $\Delta x = 2/\varrho$ will usually contain only two vehicles $k - 1$ and k with $x_k < x \le x_{k-1}$ (see Fig. 1). With this, the sum over i reduces to two terms with $i = k$ and $i = k - 1$ only. This finally yields

$$\begin{aligned} \rho(x,t)V(x,t) &= \sum_i v_i(t) s\big(x_i(t) - x\big) \\ &= v_k(t) s\big(x_k(t) - x\big) + v_{k-1}(t) s\big(x_{k-1}(t) - x\big) \\ &= \frac{\varrho}{2}\big[v_{k-1}(t) + v_k(t)\big] \\ &= \rho(x,t)\frac{v_{k-1}(t) + v_k(t)}{2} \end{aligned} \tag{45}$$

and, considering (44),

$$\begin{aligned}\sum_i s(x_i(t)-x)f\big(d_i(t)\big) &= \frac{\varrho}{2}f(d_k)+\frac{\varrho}{2}f(d_{k-1})\\ &= \frac{\varrho}{2}f\left(\frac{1}{\varrho_k}\right)+\frac{\varrho}{2}f\left(\frac{1}{\varrho_{k-1}}\right)\\ &= \frac{\rho(x,t)}{2}f\left(\frac{1}{\rho(x,t)}\right)\\ &\quad+\frac{\rho(x,t)}{2}f\left(\frac{1}{\rho(x+1/\rho,t)}\right). \qquad (46)\end{aligned}$$

In summary, the macroscopic velocity equation related to the optimal velocity model corresponds to[1]

$$\begin{aligned}&\frac{\partial V(x,t)}{\partial t}+V(x,t)\frac{\partial V(x,t)}{\partial x}\\ &= -\frac{1}{\rho(x,t)}\frac{\partial}{\partial x}\big[\rho(x,t)\theta(x,t)\big]+\frac{v^0-V(x,t)}{\tau}\\ &\quad+\frac{1}{2}f\left(\frac{1}{\rho(x,t)}\right)+\frac{1}{2}f\left(\frac{1}{\rho(x+1/\rho,t)}\right). \qquad (47)\end{aligned}$$

Note that the last line of this equation contains more terms, if more than two vehicles are located in the spatial interval between $x-1/\varrho$ and $x+1/\varrho$, as it can happen due to density variations. Since this does not affect a numerical implementation of the macroscopic equations (40) and (41), we do not need to be concerned about this. Equation (43) anyway remains unchanged.

At the cost of less straight-forward analytical evaluation, it is also possible to use other than rectangular smoothing functions (see Fig. 1). A triangular function, for example, puts less weight on the boundaries of the smoothing window, so it will make little difference whether there are two or three cars in the smoothing range. Using the specification

$$s(x_i-x)=\max\big[\varrho(1-\varrho|x_i-x|),0\big] \qquad (48)$$

and considering

[1] If another smoothing function is applied, the last term of (47) is replaced by a similar weighted mean value, as (41) reveals, but the essence stays the same. That is, the way of looking at the microscopic equations (i.e. the way of defining the density and velocity moments) potentially has some influence on the dynamics, but it is expected to be small.

$$|x_{k-1} - x| + |x - x_k| = (x_{k-1} - x) + (x - x_k) = x_{k-1} - x_k = \frac{1}{\varrho} \tag{49}$$

shows that a triangular specification leads to the same consistent density measurement:

$$\begin{aligned}\rho(x,t) &= s\big(x_{k-1}(t) - x\big) + s\big(x_k(t) - x\big) \\ &= 2\varrho - \varrho^2(x_{k-1} - x_k) = \varrho. \end{aligned} \tag{50}$$

4.3 Discussion of the Non-locality

The crucial point of (47) is its *non-locality*. The dependence on $x + 1/\rho(x,t)$ reflects the anticipatory behavior of drivers, who react to the traffic situation *ahead* of them. From the point of view of traffic simulation, the non-locality does not constitute a problem. Non-local traffic models such as the gas-kinetic based traffic model summarized in Appendix 5 can be even more efficient numerically than second-order models with diffusion terms, that would result from a gradient expansion.

In fact, the reason for the numerical inefficiency of explicit solvers for partial differential equations is the *diffusion instability*, which must be avoided by small time discretizations [13]. As pointed out by Daganzo [6], a diffusion term also implies theoretical inconsistencies such as the possible occurrence of negative velocities at upstream jam fronts. Therefore, it should be underlined that numerical inefficiencies and theoretical inconsistencies can be avoided by working with the non-local velocity equation rather than with the gradient expansion of it, which will be looked at in the next section.

4.4 Comparison with Other Macroscopic Traffic Models

4.4.1 The Payne Model

Despite the before-mentioned problems, we will now carry out a Taylor expansion of the non-local terms in (47), exclusively for the sake of comparison with other traffic models. A first-order approximation gives

$$\begin{aligned}& f\left(\frac{1}{\rho(x+1/\rho,t)}\right) \\ &\approx f\left(\frac{1}{\rho(x,t) + \frac{\partial\rho(x,t)}{\partial x}\frac{1}{\rho(x,t)}}\right)\end{aligned}$$

$$\approx f\left(\frac{1}{\rho(x,t)}\left(1-\frac{\partial\rho(x,t)}{\partial x}\frac{1}{\rho(x,t)^2}\right)\right)$$

$$\approx f\left(\frac{1}{\rho(x,t)}\right)+\frac{df(d)}{dd}\cdot\left(-\frac{\partial\rho(x,t)}{\partial x}\frac{1}{\rho(x,t)^3}\right), \tag{51}$$

where we have applied the geometric series expansion $1/(1-z) \approx 1+z+\dots$ Note that the relation $\rho = 1/d$ and

$$V_\mathrm{e}(\rho) = V_\mathrm{e}\left(\frac{1}{d}\right) = v_\mathrm{o}(d) = v^0 + \tau f(d) = v^0 + \tau f\left(\frac{1}{\rho}\right) \tag{52}$$

imply

$$\frac{df(d)}{dd} = \left(\frac{d}{d\rho}\frac{V_\mathrm{e}(\rho)-v^0}{\tau}\right)\frac{d\rho}{dd} = \frac{1}{\tau}\frac{dV_\mathrm{e}(\rho)}{d\rho}\cdot\left(-\frac{1}{d^2}\right)$$

$$= -\frac{\rho^2}{\tau}\frac{dV_\mathrm{e}(\rho)}{d\rho}. \tag{53}$$

Therefore, using (46), we finally obtain:

$$\sum_i s(x_i(t)-x)f(t) \approx \rho(x,t)f\left(\frac{1}{\rho(x,t)}\right)+\frac{1}{2\tau}\frac{dV_\mathrm{e}(\rho)}{d\rho}\frac{\partial\rho(x,t)}{\partial x}. \tag{54}$$

Considering $V_\mathrm{e}(\rho) = v^0 + \tau f(\rho)$ and defining the “traffic pressure” as

$$P(x,t) = \rho(x,t)\theta(x,t) + \frac{v^0 - V_\mathrm{e}(\rho)}{2\tau}, \tag{55}$$

the corresponding macroscopic velocity equation becomes

$$\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x}$$

$$= -\frac{1}{\rho(x,t)}\frac{\partial P(x,t)}{\partial x} + \frac{V_\mathrm{e}(\rho)-V(x,t)}{\tau}. \tag{56}$$

If the velocity variance θ is zero, this model corresponds exactly to Payne’s macroscopic traffic model with the pressure term [23, 24]

$$P(\rho) = \frac{V^0 - V_\mathrm{e}(\rho)}{2\tau}. \tag{57}$$

As a check of consistency between the Payne model and the optimal velocity model, one may perform an instability analysis of both models. Such an analysis is carried out in [10] and demonstrates indeed that the instability conditions and the characteristic velocities are compatible, as expected.

4.4.2 The Macroscopic Traffic Model by Aw and Rascle

Note that Daganzo has seriously criticized macroscopic traffic equations of the type (56) [1]. For example, he studied the case of a vehicle queue of maximum density $\rho = \rho_{\text{jam}}$ and speed $V = V_{\text{e}}(\rho_{\text{jam}}) = 0$, the end of which was assumed to be at some location $x = x_0$. In this situation, (56) predicts $V = 0$ and $dV/dt = \partial V/\partial t + V\partial V/\partial x < 0$ for the last vehicle in the queue, i.e. the occurrence of negative velocities, if pressure relations such as $P = \rho\theta_0 - \eta_0 \partial V/\partial x$ with non-negative parameters θ_0 and η_0 are assumed [15].

In order to overcome Daganzo's criticism, Aw and Rascle have proposed the macroscopic velocity equation

$$\frac{\partial}{\partial t}[V + p(\rho)] + V\frac{\partial}{\partial x}[V + p(\rho)] = 0 \tag{58}$$

with $p(\rho) = \rho^\gamma$ [1]. Let us study, how this model relates to the previous macroscopic models. For this purpose, let us apply the chain rule of Calculus to obtain

$$\begin{aligned} &\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} \\ &\quad = -\frac{dp(\rho)}{d\rho}\frac{\partial \rho(x,t)}{\partial t} - V(x,t)\frac{dp(\rho)}{d\rho}\frac{\partial \rho(x,t)}{\partial x}\,. \end{aligned} \tag{59}$$

Inserting the continuity equation (30) for $\partial\rho/\partial t$ on the right-hand side, we get

$$\begin{aligned} &\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} \\ &\quad = \frac{dp(\rho)}{d\rho}\frac{\partial}{\partial x}\big[\rho(x,t)V(x,t)\big] - V(x,t)\frac{dp(\rho)}{d\rho}\frac{\partial \rho(x,t)}{\partial x} \\ &\quad = \rho(x,t)\frac{dp(\rho)}{d\rho}\frac{\partial V(x,t)}{\partial x}. \end{aligned} \tag{60}$$

This model can be rigorously derived from particular car-following models [2]. By comparison with the macroscopic velocity equation (56) we see that the model by Aw and Rascle does not have a relaxation term $[V_{\text{e}}(\rho) - V(x,t)]/\tau$, which would correspond to the limit $\tau \to \infty$. Moreover, we find

$$-\frac{1}{\rho(x,t)}\frac{\partial P(x,t)}{\partial x} = \rho(x,t)\frac{dp(\rho)}{d\rho}\frac{\partial V(x,t)}{\partial x}\,. \tag{61}$$

Therefore, the traffic pressure according to the model of Aw and Rascle is a function of the velocity gradient rather than the density gradient, in contrast to Payne's pressure term (57). Consequently, Aw's and Rascle's pressure term must result in a different way than Payne's one, i.e. from a different kind of car-following model [2]. In order to illustrate this, let us now discuss a generalization of the optimal velocity model and its macroscopic counterpart.

4.4.3 Non-local Macroscopic Traffic Models

It is well-known [11] that the optimal velocity model may produce accidents, if the initial condition, the optimal velocity function $v_{\mathrm{o}}(d)$, and the parameter τ are not carefully chosen. In order to have both, the emergence of traffic jams and the avoidance of accidents, we need to assume that the repulsive interaction force among vehicles does not only depend on the vehicle distance $d_i(t) = x_{i-1}(t) - x_i(t)$, but also on the vehicle velocity $v_i(t)$ (to reflect the dependence of the safe distance on the vehicle speed) or on the relative velocity

$$\Delta v_i(t) = v_i(t) - v_{i-1}(t) = -\frac{dd_i}{dt} \, . \tag{62}$$

The corresponding generalization of the acceleration equation (11) reads

$$\frac{dv_i}{dt} = a_i(t) = \frac{v^0 - v_i(t)}{\tau} + f\Big(d_i(t), v_i(t), \Delta v_i(t)\Big) \, . \tag{63}$$

This also changes the associated macroscopic traffic equation. Namely, (47) has to be replaced by

$$\begin{aligned}
&\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} \\
&\quad = -\frac{1}{\rho(x,t)}\frac{\partial}{\partial x}\big[\rho(x,t)\theta(x,t)\big] + \frac{v^0 - V(x,t)}{\tau} \\
&\qquad + \frac{1}{2} f\left(\frac{1}{\rho(x,t)}, V(x,t), \Delta V(x,t)\right) \\
&\qquad + \frac{1}{2} f\left(\frac{1}{\rho(x+1/\rho,t)}, V(x+1/\rho,t), \Delta V(x+1/\rho,t)\right) \, .
\end{aligned} \tag{64}$$

For the sake of comparison with other macroscopic traffic models and linear stability analyses, let us perform a Taylor approximation of this. First, we may write

$$
\begin{aligned}
& f\left(\frac{1}{\rho(x+1/\rho,t)}, \Delta V(x+1/\rho,t), V(x+1/\rho,t)\right) \\
& \approx f\left(\frac{1}{\rho(x,t)}, \Delta V(x,t), V(x,t)\right) \\
& +\frac{\partial f}{\partial d}\frac{dd}{d\rho}\Big[\rho(x+1/\rho,t)-\rho(x,t)\Big] \\
& +\frac{\partial f}{\partial v}\Big[V(x+1/\rho,t)-V(x,t)\Big] \\
& +\frac{\partial f}{\partial \Delta v}\Big[\Delta V(x+1/\rho,t)-\Delta V(x,t)\Big] . \qquad (65)
\end{aligned}
$$

Then, we may insert $dd/d\rho = -1/\rho^2$,

$$
\rho(x+1/\rho,t)-\rho(x,t) \approx \frac{\partial\rho}{\partial x}\,\frac{1}{\rho}, \qquad (66)
$$

and

$$
V(x+1/\rho,t)-V(x,t) \approx \frac{\partial V}{\partial x}\,\frac{1}{\rho} . \qquad (67)
$$

Furthermore, considering $\Delta v_i(t) = -dd_i/dt$, $\rho(x,t) = 1/d_i(t)$, and the continuity equation $d\rho/dt = \partial\rho/\partial t + V\partial\rho/\partial x = -\rho\,\partial V/\partial x$, we get

$$
\begin{aligned}
\Delta V(x,t) &= -\frac{d}{dt}\left(\frac{1}{\rho(x,t)}\right) = \frac{1}{\rho(x,t)^2}\,\frac{d\rho(x,t)}{dt} \\
&= -\frac{1}{\rho(x,t)}\,\frac{\partial V(x,t)}{\partial x} \\
&\approx V(x,t) - V(x+1/\rho,t) \qquad (68)
\end{aligned}
$$

and

$$
\begin{aligned}
& \Delta V(x+1/\rho,t) - \Delta V(x,t) \\
& \approx \frac{\partial \Delta V}{\partial x}\,\frac{1}{\rho} \approx -\frac{1}{\rho}\frac{\partial}{\partial x}\left(\frac{\partial V}{\partial x}\frac{1}{\rho}\right) \\
& = \frac{1}{\rho^3}\,\frac{\partial\rho}{\partial x}\frac{\partial V}{\partial x} - \frac{1}{\rho^2}\,\frac{\partial^2 V}{\partial x^2} \approx -\frac{1}{\rho^2}\,\frac{\partial^2 V}{\partial x^2} , \qquad (69)
\end{aligned}
$$

since a linearization drops products of gradient terms such as $(\partial\rho/\partial x)(\partial V/\partial x)$ (which are assumed to be smaller than the linear terms). Altogether, with $dd/d\rho = -1/\rho^2$ we can write

$$
\begin{aligned}
& f\left(\frac{1}{\rho(x+1/\rho,t)}, V(x+1/\rho,t), \Delta V(x+1/\rho,t)\right) \\
& \approx f\left(\frac{1}{\rho(x,t)}, \Delta V(x,t), V(x,t)\right) - \frac{1}{\rho^3}\frac{\partial f}{\partial d}\frac{\partial \rho}{\partial x} \\
& + \frac{1}{\rho}\frac{\partial f}{\partial v}\frac{\partial V}{\partial x} - \frac{1}{\rho^2}\frac{\partial f}{\partial \Delta v}\frac{\partial^2 V}{\partial x^2}\,. \qquad (70)
\end{aligned}
$$

With the definition

$$
V_{\mathrm{o}}(\rho, V, \Delta V) = v^0 + \tau f\left(\frac{1}{\rho}, V, \Delta V\right), \qquad (71)
$$

we may finally write

$$
\begin{aligned}
\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} &= -\frac{1}{\rho}\frac{\partial}{\partial x}\big[\rho(x,t)\theta(x,t)\big] \\
&\quad + \frac{V_{\mathrm{o}}(\rho, V, \Delta V) - V(x,t)}{\tau} - \frac{1}{2\rho^3}\frac{\partial f}{\partial d}\frac{\partial \rho}{\partial x} \\
&\quad + \frac{1}{2\rho}\frac{\partial f}{\partial v}\frac{\partial V}{\partial x} - \frac{1}{2\rho^2}\frac{\partial f}{\partial \Delta v}\frac{\partial^2 V}{\partial x^2}\,. \qquad (72)
\end{aligned}
$$

Furthermore, let us assume that the variance can be approximated as a function of the density and the average velocity:

$$
\theta(x,t) = \theta_{\mathrm{e}}\big(\rho(x,t), V(x,t)\big)\,. \qquad (73)
$$

With the definitions

$$
\frac{\partial P_1}{\partial \rho} = \theta_{\mathrm{e}}(\rho, V) + \rho\,\frac{\partial \theta_{\mathrm{e}}(\rho, V)}{\partial \rho} + \frac{1}{2\rho^2}\,\frac{\partial f(1/\rho, V, \Delta V)}{\partial d}, \qquad (74)
$$

$$
\frac{\partial P_2}{\partial V} = \rho\,\frac{\partial \theta_{\mathrm{e}}(\rho, V)}{\partial V} - \frac{1}{2}\,\frac{\partial f(1/\rho, V, \Delta V)}{\partial v}\,, \qquad (75)
$$

$$
\eta = -\frac{1}{2\rho^2}\,\frac{\partial f(1/\rho, V, \Delta V)}{\partial \Delta v} \qquad (76)
$$

(where η should be greater than zero), we may also write the linearized macroscopic traffic equations as

$$\begin{aligned}&\frac{\partial V(x,t)}{\partial t}+V(x,t)\frac{\partial V(x,t)}{\partial x}\\&\quad=-\frac{1}{\rho}\frac{\partial P_1}{\partial\rho}\frac{\partial\rho}{\partial x}-\frac{1}{\rho}\frac{\partial P_2}{\partial V}\frac{\partial V}{\partial x}+\eta\frac{\partial^2 V}{\partial x^2}\\&\quad+\frac{V_{\mathrm{o}}(\rho,\Delta V,V)-V(x,t)}{\tau}\,.\end{aligned}\qquad(77)$$

The term $\eta\partial^2 V/\partial x^2$ can be interpreted as viscosity term and has a smoothing effect. Further viscosity (and diffusion) terms may be derived by second-order Taylor expansions. *It is interesting to note that the pressure term containing P_2 looks similar to (61). Therefore, it is possible to derive Aw's and Rascle's model from a suitably specified microscopic traffic model [2].*

5 Summary, Discussion, and Conclusions

In this paper, we have discussed several approaches to derive macroscopic traffic equations from microscopic car-following models. It has been pointed out that a Taylor approximation may be used only for linear stability analyses, as the gradients would otherwise often be too large for the approximation to work. Further undesirable consequences of a gradient expansion are the possible occurrence of negative velocities, diffusion instabilities, and inefficient numerical solution methods.

The linear interpolation approach often works well in practise [14], but it is theoretically inconsistent as it violates the continuity equation which is required for the conservation of the vehicle number. In contrast, the approach reminding of smoothed particle hydrodynamics was suited in all respects. It led to a non-local macroscopic traffic model, which partially reminds of the non-local gas-kinetic based traffic model [26] (see Appendix). In order to reach a realistic traffic dynamics (in particular accident avoidance if a vehicle with speed v^0 approaches a standing car), one needs to take into account that the repulsive vehicle interactions not only depend on the vehicle distance, but also on the relative velocity and the vehicle velocity. This leads to a specification of the traffic pressure which contains variance-dependent terms, additional terms proportional to $\partial\rho/\partial x$ as in Payne's model, and further terms proportional to $\partial V/\partial x$ as in Aw's and Rascle's model. While the variance-dependent term describes dispersion effects, Payne's, Aw's and Rascle's terms reflect effects of vehicle interactions. Furthermore note that, in case of multi-lane traffic, the additional inter-lane variance

$$\Theta(x,t)=\frac{1}{L}\sum_{l=1}^{L}\frac{\rho_l(x,t)}{\rho(x,t)}[V_l(x,t)-V(x,t)]^2,\qquad(78)$$

must be added to the inner-lane variance $\theta(x,t)$, where $\rho_l(x,t)$ is the density and $V_l(x,t)$ the average velocity in lane l at location x and time t [9,25].

Let us finally discuss whether the above "smoothed particle hydrodynamics approach" may lead to inconstencies such as extremely high densities. An unrealistic car-following model may, in fact, imply a theoretically inconsistent macroscopic traffic model, but a plausible microscopic model should generate a plausible macroscopic one: Specifically, the preservation of the order of vehicles requires a car-following model that does not produce accidents. Examples for this are the intelligent driver model (IDM) [27] or the Gipps model [7]. Furthermore, if the car-following model implies that vehicles keep a minimum distance of d_{min}, as the IDM does, this will translate into a maximum density $\rho_{\text{jam}} = 1/d_{\text{min}}$ in the equivalent macroscopic traffic model. This can be seen from (43) with $\rho = 1/d_{\text{min}}$. Therefore, in order to obtain a realistic macroscopic traffic model, one needs to make a suitable specification of the repulsive interaction force f. Generally, it is advised to work with speed-dependent interaction forces. An example for a microscopically derived macroscopic traffic model that takes into account the finite space requirements of vehicles is the non-local gas-kinetic-based traffic model (see Appendix).

Acknowledgements The author would like to thank for the inspiring discussions with the participants of the Workshop on "Multiscale Problems and Models in Traffic Flow" organized by Michel Rascle and Christian Schmeiser at the Wolfgang Pauli Institute in Vienna from May 5–9, 2008, with partial support by the CNRS.

Appendix: The Non-local, Gas-Kinetic Based Traffic Model

For comparison, let us shortly recall the form of the non-local gas-kinetic based traffic model (GKT model). This has been derived via a collision approximation [26] and can be written in the form of (56) with $P(x,t) = \rho(x,t)\theta(x,t)$, but $V_{\text{e}}(\rho)$ must be replaced by a non-local expression

$$V_{\text{g}}(\rho, V, \theta, \rho_+, V_+, \theta_+) = v^0 \underbrace{-\tau[1 - p(\rho_+)]\chi(\rho_+)\rho_+ B(\Delta)}_{\text{repulsive interaction term}} . \tag{79}$$

Here, the index "+" indicates evaluation at the advanced "interaction point" $x + s_0 + TV$, where s_0 represents the minimum vehicle distance and TV the velocity-dependent safety distance. The related non-locality has some effects that other macroscopic models generate by their pressure and viscosity terms. The dependence of the non-local repulsive interaction on the effective dimensionless velocity difference

$$\Delta = \frac{V - V_+}{\sqrt{\theta - 2r\sqrt{\theta\theta_+} + \theta_+}} \tag{80}$$

takes into account effects of the velocity variances θ, θ_+, and velocity correlations r among successive cars [25]. Furthermore, the "Boltzmann factor"

$$B(\varDelta) = \left(\theta - 2r\sqrt{\theta\theta_+} + \theta_+\right)\left[\varDelta N(\varDelta) + \left(1 + \varDelta^2\right)E(\varDelta)\right] \tag{81}$$

in the braking term is monotonically increasing with $\varDelta V$. It contains the normal distribution

$$N(\varDelta) = \frac{\mathrm{e}^{-\varDelta^2/2}}{\sqrt{2\pi}} \tag{82}$$

and the Gaussian error function

$$E(\varDelta) = \int\limits_{-\infty}^{\varDelta} dz\, N(z). \tag{83}$$

To close the system of equations, the velocity correlation r is specified as a function of the density in accordance with empirical observations. Moreover, for a description of the presently known properties of traffic flows it seems sufficient to set

$$\theta = A(\rho)V^2\,. \tag{84}$$

This guarantees that the velocity variance will vanish whenever the average velocity goes to zero, but it will be positive otherwise. It should be noted that the variance prefactor A is higher in congested traffic than in free traffic [26]. The *"effective cross section"* is, finally, specified via

$$[1 - p(\rho)]\chi(\rho) = \frac{v^0 \rho T^2}{\tau A(\rho_{\text{jam}})(1 - \rho/\rho_{\text{jam}})^2}\,, \tag{85}$$

where T is the safe time headway and ρ_{jam} the maximum vehicle density. This formula makes also sense in the low-density limit $\rho \to 0$, where $\chi \to 1$ and $p \to 1$.

A linear stability analysis of the non-local traffic model can be done via a gradient expansion. It results in equations of the kind (77) and further viscosity and diffusion terms [8].

References

1. A. Aw, M. Rascle, Resurrection of "second order" models of traffic flow. SIAM J. Appl. Math. **60**(3), 916–938 (2000)
2. A. Aw, A. Klar, T. Materne, M. Rascle, Derivation of continuum traffic flow models from microscopic follow-the-leader models. SIAM J. Appl. Math. **63**, 259–278 (2002)
3. M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, Dynamical model of traffic congestion and numerical simulation. Phys. Rev. E **51**, 1035–1042 (1995)

4. P. Berg, A. Mason, A. Woods, Continuum approach to car-following models. Phys. Rev. E **61**, 1056–1066 (2000)
5. D. Chowdhury, L. Santen, A. Schadschneider, Statistical physics of vehicular traffic and some related systems. Phys. Rep. **329**, 199 (2000)
6. C.F. Daganzo, Requiem for second-order fluid approximations of traffic flow. Transp. Res. B **29**, 277–286 (1995)
7. P.G. Gipps, A behavioral car-following model for computer simulation. Transp. Res. B **15**, 105–111 (1981)
8. D. Helbing, Derivation and empirical validation of a refined traffic flow model. Phys. A **233**, 253–282 (1996). See also http://arxiv.org/abs/cond-mat/9805136
9. D. Helbing, Traffic and related self-driven many-particle systems. Rev. Mod. Phys. **73**, 1067–1141 (2001)
10. D. Helbing, A. Johansson, On the controversity around Daganzo's requiem for and Aw-Rascle's resurrection of second-order traffic flow models. Eur. Phys. J. B **69**, 549–562 (2009)
11. D. Helbing, M. Schreckenberg, Cellular automata simulating experimental properties of traffic flows. Phys. Rev. E **59**, R2505–R2508 (1999)
12. D. Helbing, M. Treiber, Enskog equations for traffic flow evaluated up to Navier-Stokes order. Granul. Matter **1**, 21–31 (1998)
13. D. Helbing, M. Treiber, Numerical simulation of macroscopic traffic equations. Comput. Sci. Eng. **1**(5), 89–98 (1999)
14. D. Helbing, A. Hennecke, V. Shvetsov, M. Treiber, Micro- and macrosimulation of freeway traffic. Math. Comput. Model. **35**, 517–547 (2002)
15. B.S. Kerner, P. Konhäuser, Cluster effect in initially homogeneous traffic flow. Phys. Rev. E **48**, 2335–2338 (1993)
16. R.D. Kühne, M.B. Rödiger, Macroscopic simulation model for freeway traffic with jams and stop-start waves, in *Proceedings of the 1991 Winter Simulation Conference*, ed. by B.L. Nelson, W.D. Kelton, G.M. Clark (Society for Computer Simulation International, Phoenix, 1991), pp. 762–770
17. H.K. Lee, H.-W. Lee, D. Kim, Macroscopic traffic models from microscopic car-following models. Phys. Rev. E **64**, 056126 (2001)
18. S. Moutari, M. Rascle, A hybrid Lagrangian model based on the Aw-Rascle traffic flow model. SIAM J. Appl. Math. **68**, 413–436 (2007)
19. T. Nagatani, The physics of traffic jams. Rep. Prog. Phys. **65**, 1331–1386 (2002)
20. K. Nagel, Multi-Agent Transportation Simulations, see http://www2.tu-berlin.de/fb10/ISS/FG4/archive/sim-archive/publications/book/
21. G.F. Newell, Nonlinear effects in the dynamics of car following. Oper. Res. **9**, 209–229 (1961)
22. M. Papageorgiou, *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control* (Springer, Heidelberg, 1983)
23. H.J. Payne, Models of freeway traffic and control, in *Mathematical Models of Public Systems*, vol. 1, ed. by G.A. Bekey (Simulation Council, La Jolla, 1971), pp. 51–61
24. H.J. Payne, A critical review of a macroscopic freeway model, in *Research Directions in Computer Control of Urban Traffic Systems*, ed. by W.S. Levine, E. Lieberman, J.J. Fearnsides (American Society of Civil Engineers, New York, 1979), pp. 251–265
25. V. Shvetsov, D. Helbing, Macroscopic dynamics of multi-lane traffic. Phys. Rev. E **59**, 6328–6339 (1999)
26. M. Treiber, A. Hennecke, D. Helbing, Derivation, properties, and simulation of a gas-kinetic-based, non-local traffic model. Phys. Rev. E **59**, 239–253 (1999)
27. M. Treiber, A. Hennecke, D. Helbing, Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**, 1805–1824 (2000)
28. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)

On the Controversy Around Daganzo's Requiem for and Aw–Rascle's Resurrection of Second-Order Traffic Flow Models*

Dirk Helbing and Anders Johansson

Abstract Daganzo's criticisms of second-order fluid approximations of traffic flow [C. Daganzo, Transp. Res. B **29**, 277–286 (1995)] and Aw and Rascle's proposal how to overcome them [A. Aw and M. Rascle, SIAM J. Appl. Math. 60, 916–938 (2000)] have stimulated an intensive scientific activity in the field of traffic modeling. Here, we will revisit their arguments and the interpretations behind them. We will start by analyzing the linear stability of traffic models, which is a widely established approach to study the ability of traffic models to describe emergent traffic jams. Besides deriving a collection of useful formulas for stability analyses, the main attention is put on the characteristic speeds, which are related to the group velocities of the linearized model equations. Most macroscopic traffic models with a dynamic velocity equation appear to predict *two* characteristic speeds, one of which is *faster* than the average velocity. This has been claimed to constitute a theoretical inconsistency. We will carefully discuss arguments for and against this view. In particular, we will shed some new light on the problem by comparing Payne's macroscopic traffic model with the Aw–Rascle model and macroscopic with microscopic traffic models.

*First published in: The European Physical Journal B **69**(4), 549–562, DOI: 10.1140/epjb/e2009-00182-7 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ).

D. Helbing (✉) · A. Johansson
ETH Zurich, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland
e-mail: dhelbing@ethz.ch; andersj@ethz.ch

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_4,
© Springer-Verlag Berlin Heidelberg 2013

1 Introduction

Understanding traffic congestion has puzzled not only traffic engineers, but also a large number of physicists [1–4]. Scientists have been particularly interested in emergent traffic jams, which are related to instabilities in the traffic flow. Such instabilities have been found in empirical data [5], but also in recent experiments [6].

The theoretical analysis is usually done by computer simulation or by linear stability analysis. Both techniques have been used since the early days of traffic engineering [7] and traffic physics [8, 9]. Here, we will perform the analysis for macroscopic *and* microscopic models *in parallel*, as there should be a correspondence between the properties of both kinds of models. In contrast to previous publications, the analysis of macroscopic traffic equations is done for a model that considers a dependence of the optimal velocity function and the traffic pressure on the average velocity, not only the density. Such a dependence results for models which represent vehicle interactions realistically, taking into account a velocity-dependent safety distance [10]. This is, for example, important to avoid accidents, and it changes the instability conditions significantly (see Sect. 3).

Besides determining the stability threshold, a particular focus will be put on the calculation of the group velocities of the partial differential equations underlying the macroscopic traffic model (see Sect. 3.2). For clarity, the definition of the group velocities will be compared with those of phase velocities and of characteristic speeds. All three definitions describe propagation processes of waves. It will be shown, that they lead to identical results under certain circumstances, but not necessarily so.

Furthermore, we will derive conditions under which one of the group velocities is greater than the average velocity. In Sect. 2, we will shortly summarize the main points of the controversial discussion that this observation has triggered. We will also address Daganzo's other criticisms of second-order fluid approximations of traffic flow [11]. After the formal analysis in Sect. 3, Sect. 4 will be dedicated to a careful discussion of the results. In particular, we will analyze different conceivable reasons for characteristic speeds faster than the vehicle speeds: (1) artifacts due to approximations underlying second-order macroscopic traffic models, (2) indirect long-range *forward* interactions with *followers* on a circular road, (3) the definition of the propagation speed of perturbations, (4) the variability of vehicle velocities, (5) the interpretation of characteristic speeds. Since characteristic speeds are primarily perceived as a problem of second-order macroscopic traffic models, in Sect. 5 we will compare them with the group velocities predicted by microscopic traffic models. Finally, we will summarize our results in Sect. 6.

2 Summary of the Controversy Regarding Second-Order Traffic Flow Models

In the area of macroscopic traffic flow modeling, it is common to formulate equations for the vehicle density $\rho(x,t)$ as a function of space x and time t and for the average velocity $V(x,t)$. The most well-known model, sometimes called the LWR model, was proposed by Lighthill, Whitham, and Richards [12,13]. It is based on the continuity equation

$$\frac{\partial \rho(x,t)}{\partial t} + V(x,t)\frac{\partial \rho(x,t)}{\partial x} = -\rho(x,t)\frac{\partial V(x,t)}{\partial x} \tag{1}$$

for the density and a speed-density relationship

$$V(x,t) = V_{\mathrm{e}}\big(\rho(x,t)\big) \tag{2}$$

or, alternatively, a "fundamental diagram" $Q(x,t) = Q_{\mathrm{e}}(\rho(x,t))$ for the vehicle flow $Q(x,t) = \rho(x,t)V(x,t)$. Obviously, the LWR model is based on a (hyperbolic) partial differential equation of first order. A detailed analysis is given in [12, 14]. It is well-known, that it describes the generation of shock waves characterized by discontinuous density changes.

Therefore, in his famous "Requiem for Second-Order Fluid Approximations of Traffic Flow" [11], Carlos Daganzo correctly notes on page 285 that, "Besides a coarse representation of shocks, other deficiencies of the LWR theory include its failure to describe *platoon diffusion* properly … and its inability to explain the instability of heavy traffic, which exhibits oscillatory phenomena on the order of minutes." However, he also criticizes theoretical inconsistencies of alternative models, which, at that time, were mainly second-order models containing diffusion, pressure, or viscosity terms. The Payne–Whitham model [15, 16, 28], for example, has a dynamic velocity equation of the form

$$\begin{aligned}\frac{\partial V(x,t)}{\partial t} &+ V(x,t)\frac{\partial V(x,t)}{\partial x} \\ &= -\frac{\nu}{\rho(x,t)}\frac{\partial \rho(x,t)}{\partial x} + \frac{1}{\tau}\Big[V_{\mathrm{e}}\big(\rho(x,t)\big) - V(x,t)\Big]\end{aligned} \tag{3}$$

with

$$\nu = -\frac{1}{2\tau}\frac{dV_{\mathrm{e}}(\rho)}{d\rho} = \frac{1}{2\tau}\left|\frac{dV_{\mathrm{e}}(\rho)}{d\rho}\right| \geq 0\,. \tag{4}$$

Here, the term containing ν is called *anticipation term*, while the last term is known as *relaxation term*. $V_{\mathrm{e}}(\rho)$ denotes the equilibrium velocity and τ the relaxation time.

Some of the second-order models, including the Payne–Whitham model [15,16], can be derived from car-following models by certain approximations. This involves gradient expansions of non-local, forwardly directed (i.e. anisotropic) vehicle interactions [10]. Such approximations are problematic, since they lead to terms containing spatial derivatives, which imply undesired backward interaction effects as well. The related theoretical inconsistencies were elaborated by Daganzo. In the following, we will summarize his critique by quotes from [11] (page numbers in square brackets):

1. **Lack of anisotropy:** "A fluid particle responds to stimuli from the front and from behind, but a car is an anisotropic particle that mostly responds to frontal stimuli" [p. 279].
2. **Insufficient description of jam fronts:** "The width of a traffic shock only encompasses a few vehicles", while second-order models involving viscosity terms would typically imply extended jam fronts [p. 279]. Daganzo argues that "the smoothness of the shock is inherently unreasonable" [p. 282], because "spacings and density must change abruptly whenever the road behind is empty" [p. 282]. Based on the analysis of concrete examples, Daganzo further finds that "the cars at the end of the queue move back and the behavior spreads to the remaining vehicles in the queue . . . from the back to the front!" [p. 283]. Further on, new arrivals of vehicles would "compress a queue from behind" [p. 283].
3. **Insufficient representation of acceleration processes and driver characteristics:** According to the "relaxation" mechanism for the velocity distribution assumed in the gas-kinetic traffic model by Prigogine et al. [17], the "desired speed distribution is a property of the road and not the drivers, as noted by Paveri-Fontana (1975)" [p. 280]. However, "Unlike molecules, vehicles have personalities (e.g., aggressive and timid) that remain unchanged by motion" [p. 279], and models should make sure "that interactions do not change the 'personality' (agressive/timid) of any car" [p. 280]. Therefore, "a slow car should be virtually unaffected by its interaction with faster cars passing it (or queueing behind it) . . ." [p. 280].

A further criticism concerns the propagation speeds of perturbations in the traffic flow, predicted by second-order traffic models, which will be addressed after we have replied to the above, well-taken points:

1. The lack of anisotropy is a consequence of gradient expansions and can be avoided by non-local macroscopic traffic models [10], such as the gas-kinetic-based traffic model (GKT model) [18, 19].
2. Non-local traffic models can represent sharp shock fronts well, as has been demonstrated for the GKT model [20]. They are also capable of avoiding negative vehicle velocities, if properly specified [20]. For example, the speed variance θ appearing in some macroscopic traffic models, in particular in the "pressure term" (see below) must vanish, whenever the average velocity V vanishes. This can be reached by a relationship of the form $\theta(\rho, V) = \alpha(\rho)V^2$ with a suitable, density-dependent function $\alpha(\rho) \geq 0$ [18, 19].
3. The personality of drivers can be represented by multi-class traffic models [19, 21, 22]. Moreover, the unrealistic acceleration-behavior implied by Prigogine's gas-kinetic traffic model [17] has been overcome by the gas-kinetic model by

Paveri-Fontana [23] and its generalizations to different driver-vehicle classes [19, 21]. In these models, it is *not* the velocity *distribution* which relaxes to a *desired* velocity distribution (which would imply discontinuous velocity jumps at a certain rate). Rather they describe a continuous adaptation of individual vehicle velocities to their desired speeds.

Let us now turn to the discussion of the "characteristic speeds". Characteristic speeds relate to the eigenvalues of hyperbolic partial differential equations. They determine the solutions for given initial and boundary conditions, in particular which locations influence the solution at other locations at a given time [24, 25] (see Appendix 1). The characteristic speeds are also important for the stability of numerical solution schemes for partial differential equations [26].

What implications does this have for macroscopic traffic models based on systems of hyperbolic partial differential equations with source terms? In his "Requiem for second-order fluid approximations of traffic flow" [11], Daganzo argues that "high-order models always exhibit one characteristic speed greater than the macroscopic fluid velocity. ... This is highly undesirable because it means that the future conditions of a traffic element are, in part, determined by what is happening ... BEHIND IT! ... it is a manifestation of the erroneous cause and effect relationship between current and future variables that is at the heart of all high-order models" [p. 281].

Is this violation of causality a result of crude approximations underlying second-order macroscopic traffic models? Or could the assumption of circular boundary conditions explain an influence from behind, even in the case where vehicle interactions are exclusively directed to the front? Or is the faster characteristic speed related to vehicle interactions at all? Until today, the problem of characteristic speeds is puzzling, and it has stimulated many scientists to develop and investigate improved macroscopic traffic models [27–36]. Here, we restrict our discussion to the most prominent example: In their "Resurrection of 'second order' models of traffic flow" [27], Aw and Rascle propose a new model with two characteristic speeds, one of which is smaller than and the other one equal to V, where V denotes the macroscopic vehicle speed. Details are discussed in Sect. 4.1. While, without any doubt, such an approach is interesting and worth pursuing, we will address the question, whether it is *necessary* to overcome the problem pointed out by Daganzo. This issue must be analyzed very carefully in order to exclude misunderstandings and to avoid jumping to a conclusion. To provide a complete chain of arguments, the main text of this paper is supplemented by several appendices.

3 Linear Instability of Macroscopic Traffic Models

Let us start our analysis with the continuity equation (1) for the vehicle density $\rho(x,t)$ and a macroscopic equation for the average velocity $V(x,t)$ of the type derived at the end of Sect. 4.4.3 of [10]: Assuming repulsive vehicle interactions that depend on the vehicle distance and vehicle speed, but (for simplicity) not on the relative velocity, it reads

$$\begin{aligned}\frac{\partial V(x,t)}{\partial t} &+ V(x,t)\frac{\partial V(x,t)}{\partial x} \\ &= -\frac{1}{\rho}\frac{\partial P_1(\rho,V)}{\partial \rho}\frac{\partial \rho(x,t)}{\partial x} - \frac{1}{\rho}\frac{\partial P_2(\rho,V)}{\partial V}\frac{\partial V(x,t)}{\partial x} \\ &\quad + \frac{V_o(\rho,V) - V(x,t)}{\tau}\,. \end{aligned} \tag{5}$$

Herein, P_1 and P_2 are contributions to the "traffic pressure", and $V_o(\rho, V)$ is the "optimal velocity" function.

Our stability analysis starts with an initial state of uniform vehicle density ρ_e. The related stationary and homogeneous (i.e. time- and location-independent) solution is obtained by setting the partial derivatives $\partial/\partial t$ and $\partial/\partial x$ to zero. In this way, (5) yields the implicit equation

$$V_e(\rho_e) = V_o\big(\rho_e, V_e(\rho_e)\big) \tag{6}$$

for the equilibrium speed $V_e(\rho_e)$. With this, we can define the deviations

$$\delta\rho(x,t) = \rho(x,t) - \rho_e \qquad \text{and} \qquad \delta V(x,t) = V(x,t) - V_e\,. \tag{7}$$

Inserting $\rho(x,t) = \rho_e + \delta\rho(x,t)$ and $V(x,t) = V_e + \delta V(x,t)$ into the continuity equation, performing Taylor approximations, where necessary, and dropping all non-linear terms because of the assumption of small deviations $\delta\rho(x,t)/\rho_e \ll 1$ and $\delta V(x,t)/V_e \ll 1$, we end up with the following linearized equation:

$$\frac{\partial\,\delta\rho(x,t)}{\partial t} + V_e(\rho_e)\frac{\partial\,\delta\rho(x,t)}{\partial x} = -\rho_e\frac{\partial\,\delta V(x,t)}{\partial x}\,. \tag{8}$$

Analogously, the linerarized dynamical equation for the average velocity becomes

$$\begin{aligned}\frac{\partial\,\delta V(x,t)}{\partial t} &+ V_e\frac{\partial\,\delta V(x,t)}{\partial x} \\ &= -\frac{1}{\rho_e}\left[\frac{\partial P_1(\rho_e,V_e)}{\partial\rho}\frac{\partial\,\delta\rho(x,t)}{\partial x} + \frac{\partial P_2(\rho_e,V_e)}{\partial V}\frac{\partial\,\delta V(x,t)}{\partial x}\right] \\ &\quad + \frac{1}{\tau}\left[\frac{\partial V_o(\rho_e,V_e)}{\partial\rho}\,\delta\rho(x,t)\right. \\ &\quad \left. + \frac{\partial V_o(\rho_e,V_e)}{\partial V}\,\delta V(x,t) - \delta V(x,t)\right]. \end{aligned} \tag{9}$$

The terms on the right-hand side in the first square bracket may be considered to describe dispersion and interaction effects contributing to the "traffic pressure", while the terms in the second square bracket result from the so-called relaxation term, i.e. the adaptation of the average velocity $V(x,t)$ to some "optimal velocity" $V_o(\rho, V)$ with a relaxation time τ.

As is shown in Appendix 2, a linear stability analysis of (8) and (9) leads to the characteristic polynomial

$$(\tilde{\lambda})^2 + \tilde{\lambda}\left[\frac{\mathrm{i}\kappa}{\rho_e}\frac{\partial P_2}{\partial V} + \frac{1}{\tau}\left(1 - \frac{\partial V_o}{\partial V}\right)\right] \\ + \mathrm{i}\kappa\rho_e\left(-\frac{\mathrm{i}\kappa}{\rho_e}\frac{\partial P_1}{\partial \rho} + \frac{1}{\tau}\frac{\partial V_o}{\partial \rho}\right) = 0\,. \tag{10}$$

It has the two solutions (eigenvalues)

$$\tilde{\lambda}_\pm(\rho_e,\kappa) = \lambda_\pm(\rho_e,\kappa) - \mathrm{i}\tilde{\omega}_\pm(\rho_e,\kappa) \\ = -\frac{1}{2\hat{\tau}} - \frac{\mathrm{i}\kappa}{2\rho_e}\frac{\partial P_2}{\partial V} \pm \sqrt{\Re \pm \mathrm{i}|\Im|} \tag{11}$$

with

$$\frac{1}{\hat{\tau}(\rho_e,\kappa)} = \frac{1}{\tau}\left(1 - \frac{\partial V_o}{\partial V}\right) \geq 0\,, \tag{12}$$

$$\Re(\rho_e,\kappa) = \frac{1}{4\hat{\tau}^2} - \kappa^2\frac{\partial P_1}{\partial \rho} - \frac{\kappa^2}{4\rho_e^2}\left(\frac{\partial P_2}{\partial V}\right)^2\,, \tag{13}$$

$$\pm|\Im(\rho_e,\kappa)| = -\frac{\kappa\rho_e}{\tau}\frac{dV_o}{d\rho} + \frac{\kappa}{2\rho_e\hat{\tau}}\frac{\partial P_2}{\partial V}\,. \tag{14}$$

Here, we have used the abbreviations

$$\tilde{\lambda} = \lambda - \mathrm{i}\tilde{\omega} \qquad \text{and} \qquad \tilde{\omega} = \omega - \kappa V_e(\rho_e)\,. \tag{15}$$

As the square root contains a complex number, it is difficult to see the sign of the real value λ of $\tilde{\lambda}$. However, we may apply the formula

$$\sqrt{\Re \pm \mathrm{i}|\Im|} = \sqrt{\frac{1}{2}\left(\sqrt{\Re^2 + \Im^2} + \Re\right)} \\ \pm \mathrm{i}\sqrt{\frac{1}{2}\left(\sqrt{\Re^2 + \Im^2} - \Re\right)}\,, \tag{16}$$

which is derived in Appendix 3. From this and (11), we get the following relationship for the real part of the eigenvalues $\tilde{\lambda}_\pm(\rho_e,\kappa)$:

$$\lambda_\pm(\rho_e,\kappa) = \mathrm{Re}\big(\tilde{\lambda}_\pm(\rho_e,\kappa)\big) = -\frac{1}{2\hat{\tau}} \pm \sqrt{\frac{1}{2}\left(\sqrt{\Re^2 + \Im^2} + \Re\right)}. \tag{17}$$

The expression for the imaginary part gives

$$
\begin{aligned}
-\tilde{\omega}_{\pm}(\rho_{\mathrm{e}},\kappa) &= \mathrm{Im}\big(\tilde{\lambda}_{\pm}(\rho_{\mathrm{e}},\kappa)\big) \\
&= -\frac{\kappa}{2\rho_{\mathrm{e}}}\frac{\partial P_2}{\partial V} \pm \sqrt{\frac{1}{2}\left(\sqrt{\Re^2+\Im^2}-\Re\right)}\,.
\end{aligned} \tag{18}
$$

3.1 *Derivation of the Instability Condition*

A transition from stable to unstable behavior, i.e. the change from negative to positive values of $\lambda_{\pm}(\rho_{\mathrm{e}},\kappa)$ occurs only for the eigenvalue $\tilde{\lambda}_{+}(\rho_{\mathrm{e}},\kappa)$, namely under the condition

$$
\lambda_{+}(\rho_{\mathrm{e}},\kappa) = -\frac{1}{2\hat{\tau}} + \sqrt{\frac{1}{2}\left(\sqrt{\Re^2+\Im^2}+\Re\right)} = 0\,. \tag{19}
$$

This implies

$$
\left(\frac{1}{4\hat{\tau}^2} - \frac{\Re}{2}\right)^2 = \frac{1}{4}(\Re^2+\Im^2) \tag{20}
$$

and, therefore,

$$
\frac{1}{16\hat{\tau}^4} = \frac{\Re}{4\hat{\tau}^2} + \frac{\Im^2}{4}\,. \tag{21}
$$

Inserting the above definitions of $\Im$ and $\Re$, we eventually find

$$
\begin{aligned}
&\frac{\kappa^2}{4\hat{\tau}^2}\left[\frac{\partial P_1}{\partial \rho} + \frac{1}{4\rho_{\mathrm{e}}^2}\left(\frac{\partial P_2}{\partial V}\right)^2\right] \\
&\quad = \frac{1}{4}\left(-\frac{\kappa\rho_{\mathrm{e}}}{\tau}\frac{\partial V_{\mathrm{o}}}{\partial \rho} + \frac{\kappa}{2\rho_{\mathrm{e}}\hat{\tau}}\frac{\partial P_2}{\partial V}\right)^2\,.
\end{aligned} \tag{22}
$$

From this and definition (12), we can derive the following condition for the instability threshold:

$$
\frac{1}{\hat{\tau}}\sqrt{\frac{\partial P_1}{\partial \rho} + \frac{1}{4\rho_{\mathrm{e}}^2}\left(\frac{\partial P_2}{\partial V}\right)^2} = -\frac{\rho_{\mathrm{e}}}{\tau}\frac{\partial V_{\mathrm{o}}}{\partial \rho} + \frac{1}{2\rho_{\mathrm{e}}\hat{\tau}}\frac{\partial P_2}{\partial V}\,. \tag{23}
$$

Assuming the relationships $\partial V_{\mathrm{o}}(\rho)/\partial\rho \le 0$, $\partial V_{\mathrm{o}}/\partial V \le 0$, and $\partial P_2/\partial V \le 0$, the condition for $\mathrm{Re}(\tilde{\lambda}_{+}) > 0$ becomes

$$
\begin{aligned}
\rho_{\mathrm{e}}\left|\frac{\partial V_{\mathrm{o}}}{\partial \rho}\right| > &\left[\sqrt{\frac{\partial P_1}{\partial \rho} + \frac{1}{4\rho_{\mathrm{e}}^2}\left(\frac{\partial P_2}{\partial V}\right)^2} + \frac{1}{2\rho_{\mathrm{e}}}\left|\frac{\partial P_2}{\partial V}\right|\right] \\
&\times\left(1 + \left|\frac{\partial V_{\mathrm{o}}}{\partial V}\right|\right)\,.
\end{aligned} \tag{24}
$$

We notice that this instability condition is *not* fulfilled, if the average velocity $V_{\mathrm{o}}(\rho, V)$ changes little with the density ρ, which is typically the case for small densities and, in many models, also for large ones. However, $\lambda_{+}(\rho_{\mathrm{e}}, \kappa)$ may be greater than zero at medium densities, where $|dV_{\mathrm{e}}/d\rho|$ is large according to empirical observations. The related instability mechanism is based on a reduction of the average velocity with increasing density. Due to the continuity equation, this tends to cause a further compression (but the "traffic pressure" terms P_1 and P_2 partially counteract this re-inforcement mechanism).

As a consequence of the inequality (24), we can state that the speed-dependence of the traffic pressure term P_2 and the optimal velocity V_{o} tends to make traffic flow more stable with respect to perturbations. The speed-dependence also resolves problems related to the fact that $\partial P_1/\partial\rho$ may become negative in a certain density range. This would imply a negative discriminant of the square root, if the negative contribution $\partial P_1/\partial\rho < 0$ was not compensated for by $(\partial P_2/\partial V)^2/(4\rho_{\mathrm{e}}^2)$ [10]. The case $\partial P_1/\partial\rho < 0$ could also cause negative accelerations and speeds, particularly at the end of congestion areas, which would not be realistic [11]. Again, the second pressure contribution P_2 can resolve the problem, if properly chosen.

3.2 Characteristic Speeds, Phase, and Group Velocities

When neglecting the relaxation term (i.e. in the limit $\tau \to \infty$), the so-called characteristics may be imagined as (parametrized) space-time lines, along which the solution of a macroscopic traffic model based on partial differential equations does not change in time. In Appendix 1, we derive the characteristics of the linearlized equations (8) and (9). In the following, we will compare the characteristic speeds $C_j(\rho_{\mathrm{e}}) = V_{\mathrm{e}}(\rho_{\mathrm{e}}) + c_j(\rho_{\mathrm{e}})$ given by (66) with the phase velocities $V_{\mathrm{e}}(\rho_{\mathrm{e}}) + \tilde{\omega}_{\pm}(\rho_{\mathrm{e}}, \kappa)/\kappa$ and the group velocities $V_{\mathrm{e}}(\rho_{\mathrm{e}}) + \partial\tilde{\omega}_{\pm}(\rho_{\mathrm{e}}, \kappa)/\partial\kappa$ resulting from the above linear instability analysis. While the phase velocity describes the propagation of a single wave mode, the group velocity describes the propagation of a wave packet composed of waves with different wave numbers κ (see Appendix 4 for details). The group velocity is usually considered to represent the speed of information propagation.[1] Due to dispersion effects, we may have $\partial\tilde{\omega}_{\pm}(\rho_{\mathrm{e}}, \kappa)/\partial\kappa \neq \tilde{\omega}_{\pm}(\rho_{\mathrm{e}}, \kappa)/\kappa$.

Let us first study the situation in the limit $\tau \to \infty$ of arbitrarily slow adaptation to changed traffic conditions. Considering the definitions (12) to (14), we find $1/\hat{\tau}(\kappa) = 0$, $|\Im(\rho_{\mathrm{e}}, \kappa)| = 0$, and

$$\Re(\rho_{\mathrm{e}}, \kappa) = -\kappa^2 \frac{\partial P_1}{\partial\rho} - \frac{\kappa^2}{4\rho_{\mathrm{e}}^2}\left(\frac{\partial P_2}{\partial V}\right)^2 . \tag{25}$$

[1] A typical example is the modulation of electromagnetic waves used to transfer information via radio.

For $\Re \le 0$, we have $\sqrt{\Re^2 + \Im^2} = |\Re| = -\Re$ and, due to (17) and (18), we obtain

$$\lambda_\pm = 0 \qquad \text{and} \qquad \tilde{\omega}_\pm = -\frac{\kappa}{2\rho_\mathrm{e}} \left| \frac{\partial P_2}{\partial V} \right| \mp \sqrt{|\Re(\rho_\mathrm{e}, \kappa)|} \tag{26}$$

in the limit $\tau \to \infty$. This implies

$$\begin{aligned} \frac{\partial \tilde{\omega}_\pm(\rho_\mathrm{e}, \kappa)}{\partial \kappa} &= \frac{\tilde{\omega}_\pm(\rho_\mathrm{e}, \kappa)}{\kappa} \\ &= -\frac{1}{2\rho_\mathrm{e}} \left| \frac{\partial P_2}{\partial V} \right| \mp \sqrt{\frac{\partial P_1}{\partial \rho} + \frac{1}{4\rho_\mathrm{e}^2} \left(\frac{\partial P_2}{\partial V} \right)^2}. \end{aligned} \tag{27}$$

Therefore, group and phase velocity in the limit $\tau \to \infty$ are the same. A comparison with (66) shows that they also agree with the characteristic speeds. This is expected, because of $\lambda_\pm = 0$, which means that the wave amplitudes do not grow or decay—they just propagate along the characteristics.

For *finite* values of τ, which are typical for *real* traffic flows, the phase and group velocities may be different, and they also do not need to agree with the characteristic speeds, as we will see below: The group velocities, i.e. the propagation speeds of small perturbations, are given by

$$\begin{aligned} C_l(\rho_\mathrm{e}, \kappa) = \frac{\partial \omega_l(\rho_\mathrm{e}, \kappa)}{\partial \kappa} &= V_\mathrm{e}(\rho_\mathrm{e}) + \frac{\partial \tilde{\omega}_l(\rho_\mathrm{e}, \kappa)}{\partial \kappa} \\ &= V_\mathrm{e}(\rho_\mathrm{e}) + c_l(\rho_\mathrm{e}, \kappa), \end{aligned} \tag{28}$$

as derived in Appendix 4. Obviously, there are two group velocities $C_\pm = V_\mathrm{e} + c_\pm$, which can be determined by differentiation of the expression for $\tilde{\omega}_\pm(\rho_\mathrm{e}, \kappa)$ given in (18):

$$c_\pm(\rho_\mathrm{e}, \kappa) = +\frac{1}{2\rho_\mathrm{e}} \frac{\partial P_2}{\partial V} \mp \frac{\partial}{\partial \kappa} \sqrt{\frac{1}{2} \left(\sqrt{\Re^2 + \Im^2} - \Re \right)}. \tag{29}$$

Considering $\partial P_2 / \partial V \le 0$ and

$$\begin{aligned} \frac{1}{2} \left(\sqrt{\Re^2 + \Im^2} - \Re \right) &= \frac{1}{2} \left(\sqrt{\Re^2 + \Im^2} + \Re \right) - \Re \\ &= \left(\lambda_\pm + \frac{1}{2\hat{\tau}} \right)^2 - \Re, \end{aligned} \tag{30}$$

which is implied by (17) and (18), we may also write

$$c_\pm(\rho_\mathrm{e}, \kappa) = -\frac{1}{2\rho_\mathrm{e}} \left| \frac{\partial P_2}{\partial V} \right| \mp \frac{\partial}{\partial \kappa} \sqrt{\left(\lambda_\pm + \frac{1}{2\hat{\tau}} \right)^2 - \Re}. \tag{31}$$

Taking into account (13), this is generally not the same as $\tilde{\omega}_\pm(\rho_\mathrm{e}, \kappa)/\kappa$, i.e. the phase velocities differ. Interestingly enough, however, at the stability threshold given by $\lambda_+ = 0$, we find

$$
\begin{aligned}
c_+(\rho_\mathrm{e},\kappa) &= -\frac{1}{2\rho_\mathrm{e}}\left|\frac{\partial P_2}{\partial V}\right| - \frac{\partial}{\partial\kappa}\sqrt{\frac{1}{4\hat{\tau}^2}-\Re} \\
&= -\frac{1}{2\rho_\mathrm{e}}\left|\frac{\partial P_2}{\partial V}\right| - \sqrt{\frac{\partial P_1}{\partial\rho}+\frac{1}{4\rho_\mathrm{e}^2}\left(\frac{\partial P_2}{\partial V}\right)^2}\,. \qquad (32)
\end{aligned}
$$

At the stability threshold we furthermore have $\lambda_- = -1/\hat{\tau}$. Inserting this into (31) reveals

$$
\begin{aligned}
c_-(\rho_\mathrm{e},\kappa) &= -\frac{1}{2\rho_\mathrm{e}}\left|\frac{\partial P_2}{\partial V}\right| + \frac{\partial}{\partial\kappa}\sqrt{\frac{1}{4\hat{\tau}^2}-\Re} \\
&= -\frac{1}{2\rho_\mathrm{e}}\left|\frac{\partial P_2}{\partial V}\right| + \sqrt{\frac{\partial P_1}{\partial\rho}+\frac{1}{4\rho_\mathrm{e}^2}\left(\frac{\partial P_2}{\partial V}\right)^2}\,. \qquad (33)
\end{aligned}
$$

The same expressions are found for the phase velocities. A comparison with (66) shows that they also agree with the characteristic speeds. Note that c_+ is *smaller* than zero. However, we have $c_- \le 0$ (corresponding to characteristic speeds *slower* than the average vehicle velocity or equal to it) *only* if

$$\sqrt{\frac{\partial P_1}{\partial\rho}+\frac{1}{4\rho_\mathrm{e}^2}\left(\frac{\partial P_2}{\partial V}\right)^2} \le \frac{1}{2\rho_\mathrm{e}}\left|\frac{\partial P_2}{\partial V}\right| \qquad (34)$$

or

$$0 \le -\frac{\partial P_1}{\partial\rho} \le \frac{1}{4\rho_\mathrm{e}^2}\left(\frac{\partial P_2}{\partial V}\right)^2\,. \qquad (35)$$

4 Discussion

For the discussion of our results regarding the characteristic speeds, let us study two particular models first, the Payne model [15, 16] and the Aw–Rascle model [27].

4.1 Characteristic Speeds in the Aw–Rascle Model

The model proposed by Aw and Rascle [27] corresponds to (1) and (5) with $\tau \to \infty$,

$$\frac{\partial P_1(\rho,V)}{\partial\rho} = 0 \qquad \text{and} \qquad \frac{\partial P_2(\rho,V)}{\partial V} = -\gamma\rho(x,t)^{\gamma+1} \le 0\,, \qquad (36)$$

see [10]. γ is a positive constant.

This implies $1/\hat{\tau} = 0$, $\Re(\kappa) = -\kappa^2(\partial P_2/\partial V)^2/(4\rho_e^2) < 0$ and $|\Im(\kappa)| = 0$. Therefore, (29) implies

$$\begin{aligned} c_\pm(\rho_e,\kappa) &= -\frac{1}{2\rho_e}\left|\frac{\partial P_2}{\partial V}\right| \mp \frac{\partial}{\partial\kappa}\sqrt{\frac{1}{2}\left(|\Re| - \Re\right)} \\ &= -\frac{1}{2\rho_e}\left|\frac{\partial P_2}{\partial V}\right| \mp \frac{1}{2\rho_e}\left|\frac{\partial P_2}{\partial V}\right| . \end{aligned} \tag{37}$$

This leads to $c_+ = -\gamma\rho(x,t)^\gamma$ and $c_- = 0$, corresponding to the characteristic speeds $V - \gamma\rho(x,t)^\gamma$ and V, in agreement with Aw's and Rascle's calculations [27]. That is, their model does not have a characteristic speed faster than the average vehicle speed, which elegantly avoids the problem raised by Daganzo [11].

However, is it really necessary to *exclude* the existence of a characteristic speed faster than the vehicle speeds? In order to address this problem, we will now study Payne's macroscopic traffic model, which has received most of the criticism. We do this primarily for the sake of illustration, while we are well aware of the weaknesses of this model (like the possibility of backward moving vehicles at upstream jam fronts for certain initial conditions). Therefore, the authors of this paper generally prefer the use of *non-local* macroscopic traffic models [10], but this is not the issue to be discussed, here.

4.2 Payne's Traffic Model

Payne's macroscopic traffic model [15, 16] has a solely density-dependent optimal velocity

$$V_o(\rho, V) = V_e(\rho) \tag{38}$$

and the pressure gradients

$$\frac{\partial P_1(\rho,V)}{\partial\rho} = \frac{1}{2\tau}\left|\frac{dV_e(\rho)}{d\rho}\right| \ge 0\,, \qquad \frac{\partial P_2(\rho,V)}{\partial V} = 0\,. \tag{39}$$

This simplifies the instability condition (24) considerably, and we get

$$\rho_e\left|\frac{dV_e(\rho_e)}{d\rho}\right| > \frac{1}{2\rho_e\tau}\,. \tag{40}$$

Traffic flow becomes unstable, if the equilibrium velocity $V_e(\rho)$ decreases too rapidly with an increase in the density ρ, and greater relaxation times τ tend to imply larger instability regimes. For the characteristic speeds at the instability threshold, with $\rho_e|dV_e/d\rho| = 1/(2\rho_e\tau)$ we find

$$c_\pm(\rho_e) = \mp\sqrt{\frac{\partial P_1}{\partial\rho}} = \mp\sqrt{\frac{1}{2\tau}\left|\frac{dV_e(\rho_e)}{d\rho}\right|} = \mp\rho_e\left|\frac{dV_e(\rho_e)}{d\rho}\right|\,. \tag{41}$$

Clearly, $c_-(\rho)$ is non-negative, i.e. the related characteristic speed $V_e(\rho) + c_-(\rho)$ tends to be larger than the average vehicle speed $V_e(\rho)$. Nevertheless, by demanding $V_e(\rho) + c_-(\rho) \le V^0$, e.g. by assuming a linear speed-density function

$$V_e(\rho) = V^0 \left(1 - \frac{\rho}{\rho_{\text{jam}}}\right) , \tag{42}$$

one could still reach that the characteristic speed $V_e(\rho) + c_-(\rho)$ lies within the variability of the vehicle speeds. In fact, we have $c_\pm = 0$ whenever the vehicle speed cannot vary, namely at density zero and at maximum density, where $\rho_e|dV_e(\rho_e)/d\rho| = 0$. However, do we *need* to impose such conditions on the characteristic speed and the speed-density relationship? This shall be addressed in the following and in Sect. 5.

In connection with this question, it is interesting to note that, according to (33) and (41), the group velocity c_+ corresponding to the solution with the *unstable* eigenvalue λ_+ is *negative* with respect to the average velocity V_e. In contrast, propagation at the positive speed c_- with respect to the average velocity V_e is related with an *eigenmode* that *decays quickly, basically at the rate at which the vehicle speeds adjust.* Therefore, the forwardly propagating mode cannot emerge by itself. It could only be produced by a particular specification of the initial condition, enforcing a finite amplitude of the forwardly moving mode. We will come back to this in Sect. 5.

It is noteworthy that already Whitham performed a thorough analysis of the speeds characterizing the traffic dynamics in what is known as the Payne model today (see [14], Chaps. 3 and 10). He showed that the linearized partial differential equations (8) and (9), when specified in accordance with (38) and (39), can be cast into the equation

$$\begin{aligned}
&\frac{\partial \delta\rho(x,t)}{\partial t} + \left(V_e(\rho) + \rho \frac{dV_e(\rho)}{d\rho}\right) \frac{\partial \delta\rho(x,t)}{\partial x} \\
&\quad = -\tau \left(\frac{\partial}{\partial t} + \left[V_e(\rho) + c_+(\rho)\right] \frac{\partial}{\partial x}\right) \\
&\qquad \times \left(\frac{\partial}{\partial t} + \left[V_e(\rho) + c_-(\rho)\right] \frac{\partial}{\partial x}\right) \delta\rho(x,t) \,.
\end{aligned} \tag{43}$$

Whitham was perfectly aware of the fact that the characteristic speed $V_e(\rho) + c_-(\rho)$ was faster than the average vehicle velocity $V_e(\rho)$, but not at all worried about this. His perception was that all three velocities were meaningful, and that the kinematic speed $V_e(\rho) + \rho\, dV_e/d\rho$ would dominate in the limit of small values of τ (which implies stable vehicle flows). However, the open problem is still, how a characteristic speed $V_e(\rho) + c_-(\rho) > V_e(\rho)$ can be interpreted, without violating causality.

4.3 Characteristic Speeds Vs. Vehicle Speeds

In physical systems, it is not necessarily surprising to find characteristic speeds faster than the average speed. Let us illustrate this for the example of sound propagation. In one spatial dimension, this is described by the continuity equation (1) in combination with the one-dimensional velocity equation

$$\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} = -\frac{1}{\rho}\frac{\partial \mathscr{P}(\rho)}{\partial x}\,. \tag{44}$$

These so-called Euler equations [37] can be considered to model frictionless fluid or gas flows in one dimension. Compared to the velocity equation (5), we have dropped the relaxation term $[V_\mathrm{e}(\rho) - V]/\tau$. Therefore, we do not have an equilibrium velocity-density relation $V_\mathrm{e}(\rho)$, now.

In order to determine the solution of the above equations, one can derive linearized equations for the case of small deviations $\delta\rho(x,t) = \rho(x,t) - \rho_\mathrm{e}$ and $\delta V(x,t) = V(x,t) - V_\mathrm{e}$ from the stationary and homogeneous solution $\rho(x,t) = \rho_\mathrm{e}$ and $V(x,t) = V_\mathrm{e} = 0$. The quantity ρ_e corresponds to the average density of the fluid or gas.

Inserting (7) into (1) and (44) and neglecting non-linear terms in the small deviations $\delta\rho$, δV results in

$$\frac{\partial \delta\rho(x,t)}{\partial t} + V_\mathrm{e}\frac{\partial \delta\rho(x,t)}{\partial x} = -\rho_\mathrm{e}\frac{\partial \delta V(x,t)}{\partial x} \tag{45}$$

and

$$\frac{\partial \delta V(x,t)}{\partial t} + V_\mathrm{e}\frac{\partial \delta V(x,t)}{\partial x} = -\frac{1}{\rho_\mathrm{e}}\frac{d\mathscr{P}(\rho_\mathrm{e})}{d\rho}\frac{\partial \delta\rho(x,t)}{\partial x}\,. \tag{46}$$

Considering $V_\mathrm{e} = 0$, deriving (45) with respect to t, and (46) with respect to x yields

$$\frac{\partial^2 \delta\rho(x,t)}{\partial t^2} + \rho_\mathrm{e}\frac{\partial^2 \delta V(x,t)}{\partial t\,\partial x} = 0 \tag{47}$$

and

$$\frac{\partial^2 \delta V(x,t)}{\partial x\,\partial t} = -\frac{1}{\rho_\mathrm{e}}\frac{d\mathscr{P}(\rho_\mathrm{e})}{d\rho}\frac{\partial^2 \delta\rho(x,t)}{\partial x^2}\,. \tag{48}$$

Inserting (48) into (47) finally gives the so-called wave equation

$$\frac{\partial^2 \delta\rho(x,t)}{\partial t^2} - \hat{c}^2\frac{\partial^2 \delta\rho(x,t)}{\partial x^2} = 0\,, \tag{49}$$

which is well-known from one-dimensional sound propagation. The constant

$$\hat{c} = \sqrt{\frac{d\mathscr{P}(\rho_\mathrm{e})}{d\rho}}\,, \tag{50}$$

corresponds to the speed of sound. In order to determine the spatio-temporal solution of (49), we rewrite this equation, inspired by the relationship $(a^2 - b^2) = (a + b)(a - b)$:

$$\left(\frac{\partial}{\partial t} + \hat{c}\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} - \hat{c}\frac{\partial}{\partial x}\right)\delta\rho(x,t) = 0\,. \tag{51}$$

According to this equation, perturbations propagate backward and forward at the speed $\pm\hat{c}$, although the average speed is $V = 0$. However, for gases we may assume an approximate pressure law of the form $\mathscr{P} = \rho\theta_0$ [37], where θ_0 is the velocity variance of gas molecules. Hence, the speed of sound is given by $\hat{c} = \sqrt{\theta_0}$, i.e. by the standard deviation of velocities. As a consequence, the speed of sound *can* actually be propagated by the mobility of gas molecules.

In a similar way, we can understand characteristic speeds faster than the average vehicle speed in the macroscopic model of Phillips [38] or Kühne [8], Kerner and Konhäuser [39], and Lee et al. [40]. Their pressure functions are also given by the formula "density times velocity variance". Therefore, the faster characteristic speed of *these* macroscopic traffic models is expected to lie within the range of individual vehicle speeds.[2]

As we have seen above, the situation is generally different for Payne's model. However, it is illustrative to note that $V_{\rm o}(\rho) + c_+(\rho)$ may become *negative*, even when all vehicles move *forward*. That is, it is *possible* to have characteristic speeds *outside* of the range of vehicle speeds: According to (41) and (15), the slower characteristic speed at the instability threshold is

$$\begin{aligned} V_{\rm e}(\rho) + c_+(\rho) &= V_{\rm e}(\rho) - \rho\left|\frac{dV_{\rm e}(\rho)}{d\rho}\right| \\ &= V_{\rm e}(\rho) + \rho\frac{dV_{\rm e}(\rho)}{d\rho} = \frac{dQ_{\rm e}(\rho)}{d\rho}\,. \end{aligned} \tag{52}$$

Since $Q_{\rm e}(\rho) = \rho V_{\rm e}(\rho)$ represents the "fundamental diagram", $dQ_{\rm e}(\rho)/d\rho$ describes the negative speed of kinematic waves in the congested regime [14]. This does not constitute *any* theoretical inconsistency, even if $V_{\rm e}(\rho_{\rm e}) + c_+(\rho) < 0$. In fact, we all know situations involving negative group velocities from dissolving congestion fronts, e.g. when a traffic light turns green: There, the negative propagation speed just results from the fact that the congestion front moves backward, whenever vehicles leave a congested area with some delay. Hence, the negative characteristic speed does not describe the speed of cars. It reflects the propagation of *gaps* rather than vehicles.

Therefore, could we have a similar mechanism that generates characteristic speeds *faster* than the vehicle speeds? If vehicles would react to their leaders

[2] Note that the existence of perturbations in the traffic flow always implies a variation of the vehicle speeds.

with a *negative* delay, this would in fact be the case, but it would violate causality. Therefore, all possible explanations for characteristic speeds faster than the vehicle speeds considered so far have failed to resolve the problem. However, the problem may still be a result of the approximations underlying second-order macroscopic traffic models. As we have indicated before, the gradient expansion required to derive them implies some degree of backward interactions. Therefore, it is conceivable that following vehicles would cause their leaders to accelerate, even beyond their desired speed V^0.

If this would be the explanation of a characteristic speed faster than the average speed V or free speed V^0, we should not observe it in microscopic traffic models with forward interactions only. Therefore, we will now determine the characteristic speeds of the optimal velocity model [3]. This car-following is chosen, because the Payne model can be considered as a macroscopic approximation of it (see [10] and references therein). Besides, we will compare the instability conditions of both models.

5 Linear Instability and Characteristic Speeds of the Optimal Velocity Model

We have seen that macroscopic traffic models behave unstable with respect to small perturbations in a certain density range, where the average velocity changes too rapidly with the density. The same is true for many car-following models. As an example, we will shortly discuss the dynamic behavior of the optimal velocity model. While its stability has been already studied in the past [3], we will focus here on the characteristic speeds, in order to show that characteristic speeds greater than the average velocity are not an artifact of macroscopic traffic models.

According to the optimal velocity model, the change of the speed $v_i(t)$ of vehicle i is given by

$$\frac{dv_i}{dt} = \frac{v_{\rm o}\big(d_i(t)\big) - v_i(t)}{\tau} \tag{53}$$

and the temporal change of the distance $d_i(t) = x_{i-1}(t) - x_i(t)$ to the leading vehicle $i-1$ is determined by

$$\frac{dd_i}{dt} = v_{i-1}(t) - v_i(t)\,. \tag{54}$$

In the above equations, the distance-dependent function $v_{\rm o}(d_i)$ is called the optimal velocity function and τ is again the relaxation time for adjustments of the speed.

Appendix 5 sketches the linear stability analysis of the optimal velocity model. In the following, we will focus on the analysis of the group velocity $c_\pm$ with respect to the average velocity $v_{\rm o}(d_{\rm e})$, i.e. the velocity at which perturbations are expected to propagate. Relative to the average motion of vehicles with speed $v_{\rm e}(d_{\rm e})$, the characteristic speeds are

$$\begin{aligned} c_{\pm}(d_{\rm e}, k) &= \frac{\partial \tilde{\omega}_{\pm}(d_{\rm e}, \kappa)}{\partial \kappa} = \frac{L}{2\pi} \frac{\partial \tilde{\omega}_{\pm}(d_{\rm e}, k)}{\partial k} \\ &= \mp \frac{L}{2\pi} \frac{\partial}{\partial k} \sqrt{\frac{1}{2}\left(\sqrt{\Re^2 + \Im^2} - \Re\right)} . \end{aligned} \tag{55}$$

This can be derived analogously to (29), using (16) and $\kappa = 2\pi k/L$. According to (31) and due to the series expansion $\cos(x) \approx 1 - x^2/2$, at the instability threshold with $\lambda_+ = 0$ and $d v_{\rm o}(d_{\rm e})/dd = 1/(2\tau)$, we obtain with (105)

$$\begin{aligned} c_{\pm}(d_{\rm e}, k) &= \mp \frac{L}{2\pi} \frac{\partial}{\partial k} \sqrt{\left(\frac{1}{2\tau}\right)^2 - \Re} \\ &= \mp \frac{L}{2\pi} \frac{\partial}{\partial k} \sqrt{\frac{1}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \big[1 - \cos(2\pi k/N)\big]} \\ &\approx \mp \frac{L}{2\pi} \frac{\partial}{\partial k} \sqrt{\frac{1}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \frac{1}{2} \left(\frac{2\pi k}{N}\right)^2} \\ &= \mp \frac{L}{N} \sqrt{\frac{1}{2\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd}} = \mp d_{\rm e} \sqrt{\frac{1}{2\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd}} \qquad (56) \\ &= \mp d_{\rm e} \sqrt{\left(\frac{d v_{\rm o}(d_{\rm e})}{dd}\right)^2} = \mp d_{\rm e} \frac{d v_{\rm o}(d_{\rm e})}{dd} . \qquad (57) \end{aligned}$$

It is remarkable that the group velocity of the optimal velocity model can again exceed the average vehicle velocity $v_{\rm o}(d_{\rm e})$, namely by an amount $c_{-}(d_{\rm e}) = d_{\rm e}\, d v_{\rm o}(d_{\rm e})/d_{\rm e} > 0$. Moreover, it can be shown that the instability thresholds and the related characteristic speeds are the same as for the Payne model (see Appendix 6). This confirms that the Payne model may be viewed as macroscopic approximation of the optimal velocity model (see [10] and references therein). In view of these results, it is hard to argue that a characteristic speed faster than the vehicle speeds constitutes primarily a theoretical inconsistency of certain kinds of *macroscopic* traffic models. Quite unexpectedly, it also occurs for microscopic traffic models that, according to computer simulations, behave reasonably well.

Therefore, the approximations underlying the Payne model cannot be the problem for the existence of a characteristic speed faster than the vehicle speeds. However, it is interesting to note that the larger group velocity $v_{\rm o}(d_{\rm e}) + c_{-}(d_{\rm e})$ is related to a *negative* real part λ_{-} of the eigenvalue $\tilde{\lambda}_{-}$. According to (29), the fast characteristic speed $V_{\rm e}(\rho_{\rm e}) + c_{-}(\rho_{\rm e})$ of macroscopic second-order models is related to a negative eigenvalue $\lambda_{-}(\rho_{\rm e})$ as well, see (17). Therefore, *the related eigenmode decays quickly, and it will be hard to observe in reality. In particular, the faster propagating mode may not emerge by itself.* A closer analysis shows that both,

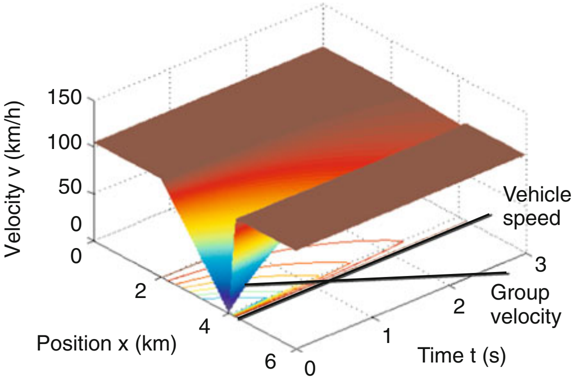


Fig. 1 Simulation result of the optimal velocity model with $v_o(d) = v^0\{\tanh[(d-l)/s_0 - 1.2] + \tanh(1.2)\}/2$, $v^0 = 115$ km/h, $s_0 = 50$ m, and $l = 4$ m. We have chosen a particular initial condition, where all vehicles started with a distance $d_e = 200$ m to their respective leader, but some vehicles i had a speed $v_i(0) < v_o(d_e)$ in the beginning. As a consequence, these vehicles adjusted their speeds to the optimal velocity. The relevant point here is that followers reach the optimal velocity (or certain fractions of it) earlier than their respective leaders. That is, for the particular initial condition chosen here, the perturbation in the speeds propagates *faster* than the vehicle speeds. This effect, however, does not violate causality, as the earlier acceleration of upstream cars is not triggered by interactions with followers—it just results from the relaxation term. Therefore, the perturbation disappears on a time scale that is determined by the relaxation time $\tau = 1$ s, as predicted by the real part of the eigenvalue $\tilde{\lambda}_-$, see (104). The relaxation takes longer for larger values of τ. In the limit $\tau \to \infty$, the perturbation does not decay anymore, but according to (104), we then have $c_\pm \to 0$. Therefore, despite its fast speed, the perturbation did not overtake the first car upstream of the initial perturbation in our simulations, when the parameters were chosen in a way that avoided accidents. This confirms the validity of the causality principle

for the optimal velocity model and the Payne model, λ_- *is of the order* $-1/\tau$, *i.e. related to the relaxation time* τ *of vehicles.* We will see that this observation is highly relevant for understanding perturbations that move faster than the vehicles do.

After all, does the fast characteristic speed *really* constitute a theoretical inconsistency? Not so, if we can find initial conditions, for which a following car accelerates or decelerates earlier than the leading car does, although the leader does *not* react to the follower. In fact, such initial condition can be constructed: Fig. 1 shows the result of a computer simulation with N vehicles on a circular road of length L. We assume that all vehicles have the distance $d = d_e = L/N$ initially. Moreover, all vehicles, with the exception of ten subsequent vehicles, are assumed to have the initial speed $v_o(d_e)$. Furthermore, the speed of the last of the ten vehicles is set to 0 (or v^0), the speed of the first one to $v_o(d_e)$. The speeds of the vehicles in between are determined by linear interpolation. For this scenario, it is quite natural that the *last* of the 10 vehicles accelerates (or decelerates) *first*, since it experiences the largest deviation of its actual velocity $v_i(0)$ from the optimal velocity $v_o(d_e)$. However, *as this earlier acceleration (or deceleration) is not interaction-induced, it does not violate causality.* The large characteristic speed in macroscopic traffic models can be understood in a similar way.

6 Summary, Conclusions, and Outlook

In this paper, we have started with a discussion of Daganzo's sharp criticism of second-order macroscopic traffic flow models [11]. We have argued that most of the deficiencies identified by Daganzo were fully justified, but could be overcome in the course of time by improved macroscopic traffic models, particularly by nonlocal multi-class models. However, the issue of characteristic speeds faster than the average vehicle speed was still an open, controversial problem, as it seems to violate causality. In order to study it, we have performed a linear instability analysis of a generalized macroscopic traffic model, which took into account speed-dependencies of the optimal velocity and the traffic pressure terms. Such speed-dependencies occur, for example, in Aw's and Rascle's model [27]. They result when realistic vehicle interactions are considered, and when the possibility of accidents and negative vehicle speeds shall be avoided [10, 41]. Requirements for reasonable models seem to be

$$\frac{\partial V_o(\rho, V)}{\partial \rho} \le 0\,, \qquad \frac{\partial V_o(\rho, V)}{\partial V} \le 0\,, \qquad \frac{\partial P_2(\rho, V)}{\partial V} \le 0\,, \tag{58}$$

and

$$\frac{\partial P_1(\rho, V)}{\partial \rho} + \frac{1}{4\rho^2}\left(\frac{\partial P_2(\rho, V)}{\partial V}\right)^2 > 0\,. \tag{59}$$

These conditions are, for example, fulfilled by the gas-kinetic-based traffic model (GKT model), see [43].

Our main attention was dedicated to the characteristic speeds (or group velocities) rather than the instability thresholds. In the following, we summarize the main results:

1. While the characteristic speeds may generally differ from the group and the phase velocities, in the limit $\tau \to \infty$ of a vanishing source (relaxation) term, they are all the same. Therefore, using a different definition of propagation speeds does not resolve the problem of characteristic speeds faster than the (average or maximum) vehicle speed.
2. Velocity-dependent pressure terms tend to reduce the characteristic speeds, see (31). This is best illustrated by Aw's and Rascle's model, where the fast characteristic agrees with the average vehicle speed.
3. Most macroscopic traffic models have a characteristic speed *faster* than the average velocity, but it may still be within the variability of the vehicle speeds, see (42) and Sect. 4.3.
4. In some models like the Payne model, the characteristic speeds can move slower than the slowest vehicle and faster than the fastest vehicle. The first case is related to delayed acceleration maneuvers at jam fronts and related to gap propagation during jam dissolution, but the second case remained a mystery for a long time.
5. The *faster* characteristic speed is related with a *negative* real part of the eigenvalue. This causes a quick decay of the corresponding eigenmode, basically

at the rate, at which the vehicle speed is adjusted. Therefore, this eigenmode will not emerge by itself (see Sect. 3.2).

6. If the faster characteristic speed were a result of interactions with following vehicles in a circular road geometry (where *following* vehicles influence the *downstream* flow as well), the fast eigenmode should decay with the length L of the circular road, not with the relaxation time τ. Therefore, periodic boundary conditions cannot be responsible for a characteristic speed faster than the vehicle speeds. This has also been verified with simulations.[3]
7. A characteristic speed faster than the vehicle speeds cannot be explained as a result of the approximations underlying macroscopic second-order models, as it is also found for microscopic car-following models, in which vehicle interactions are forwardly directed and velocities are restricted to a range between zero and some maximum speed. For the macroscopic Payne model and the optimal velocity model, we have shown a correspondence not only of the instability thresholds, but also of formulas for the group velocities (see Appendix 6).
8. Assuming particular initial conditions, characteristic speeds faster than the average vehicle speed could be *demonstrated* to exist in computer simulations, where followers accelerate (or decelerate) *before* their leaders do (see Fig. 1). *As these acceleration (or deceleration) processes are induced by artificial initial perturbations rather than by vehicle interactions, this does* ***not*** *imply a violation of causality.*

Given these findings, we conclude that characteristic speeds faster than the average speed of vehicles do not constitute a theoretical inconsistency of traffic models and do not need to be "healed" by particularly constructed traffic models.[4] From our point of view, the problem is that characteristic speeds are hard to imagine. In fact, there is no direct correspondence to particle or vehicle velocities (see Sects. 4.3 and Appendix 4). The group velocity is nothing more than a matter of phase relations between oscillations of successive vehicles in an eigenmode, and the interpretation as speed of information transmission is sometimes misleading.

Acknowledgements

Author contributions: DH performed the analytical calculations and proposed the initial conditions for the simulation presented in Fig. 1. AJ generated the computational results and prepared the figure.

Acknowledgment: DH would like to thank for the inspiring discussions with the participants of the Workshop on "Multiscale Problems and Models in Traffic Flow" organized by Michel Rascle and Christian Schmeiser at the Wolfgang Pauli Institute in Vienna from May 5–9, 2008, with partial support by the CNRS.

[3] Simulations for open boundary conditions basically yield the same results as for periodic boundary conditions, given the system (in terms of the road length L) is sufficiently large.

[4] Of course, this does not speak against models of the Aw–Rascle type.

Appendix 1 Hyperbolic Sets of Partial Differential Equations and Characteristic Speeds

Let us rewrite (8) and (9) in the form of a system of linear partial differential equations. With

$$\begin{aligned} S(\delta\rho, \delta V) = \frac{1}{\tau}\Bigg[& \frac{\partial V_{\mathrm{o}}(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial \rho}\,\delta\rho(x,t) \\ & + \frac{\partial V_{\mathrm{o}}(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial V}\,\delta V(x,t) - \delta V(x,t)\Bigg] \end{aligned} \tag{60}$$

we obtain

$$\frac{\partial}{\partial t}\begin{pmatrix} \delta\rho(x,t) \\ \delta V(x,t) \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} \delta\rho(x,t) \\ \delta V(x,t) \end{pmatrix} = \begin{pmatrix} 0 \\ S \end{pmatrix} \tag{61}$$

with

$$\underline{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} V_{\mathrm{e}}(\rho_{\mathrm{e}}) & \rho_{\mathrm{e}} \\ \frac{1}{\rho_{\mathrm{e}}}\frac{\partial P_1(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial \rho} & V_{\mathrm{e}}(\rho_{\mathrm{e}}) + \frac{1}{\rho_{\mathrm{e}}}\frac{\partial P_2(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial V} \end{pmatrix}. \tag{62}$$

As will be shown below, the solution of this system of partial differential equations is given by the *initial* condition $\delta\rho(x,0)$ and $\delta V(x,0)$. The solution procedure consists basically of two steps: On the one hand, we must determine the so-called characteristics, and on the other hand, we must solve a set of ordinary differential equations to find the solutions along them (see [42] and footnote 3): With $\mathbf{u}(x,t) = \big(\delta\rho(x,t), \delta V(x,t)\big)'$ and $\mathbf{S} = (0, S)'$ (where the prime indicates a transposed, i.e. a column vector), we can rewrite (61) as

$$\frac{\partial \mathbf{u}(x,t)}{\partial t} + \underline{A}\,\frac{\partial \mathbf{u}(x,t)}{\partial x} = \mathbf{S} = \underline{B}\,\mathbf{u}(x,t)\,. \tag{63}$$

The source term can be rewritten as $\mathbf{S} = \underline{B}\,\mathbf{u}(x,t)$ with

$$\underline{B} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \frac{1}{\tau}\frac{\partial V_{\mathrm{o}}(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial \rho} & \frac{1}{\tau}\left(\frac{\partial V_{\mathrm{o}}(\rho_{\mathrm{e}}, V_{\mathrm{e}})}{\partial V} - 1\right) \end{pmatrix}. \tag{64}$$

Now, let C_j denote the eigenvalues of the matrix $\underline{A}$. The values of $C_j = V_{\mathrm{e}}(\rho_{\mathrm{e}}) + c_j$ satisfying $\det(\underline{A} - C_j\,\underline{1}) = 0$ are given by the characteristic polynomial

$$c_j^2 - \frac{c_j}{\rho_{\mathrm{e}}}\frac{\partial P_2}{\partial V} - \frac{\partial P_1}{\partial \rho} = 0\,, \tag{65}$$

which results in

$$c_j = \frac{1}{2\rho_e}\frac{\partial P_2}{\partial V} \pm \sqrt{\frac{1}{4\rho_e^2}\left(\frac{\partial P_2}{\partial V}\right)^2 + \frac{\partial P_1}{\partial \rho}}\,. \tag{66}$$

Furthermore, let $\mathbf{z}_j$ be the eigenvectors related with the eigenvalues $C_j = V_e + c_j$, i.e.

$$\underline{A}\,\mathbf{z}_j = C_j \mathbf{z}_j\,. \tag{67}$$

Finally, let $\underline{R} = (R_{ij})$ be the matrix containing the eigenvectors $\mathbf{z}_j$ as their jth column, and $\mathbf{y}(x,t) = \underline{R}^{-1}\mathbf{u}(x,t)$ or $\mathbf{u}(x,t) = \underline{R}\,\mathbf{y}(x,t)$. Then, inserting this into (63) and multiplying the result with the inverse matrix $\underline{R}^{-1}$ of $\underline{R}$ yields

$$\frac{\partial y_j(x,t)}{\partial t} + C_j\frac{\partial y_j(x,t)}{\partial x} = (\underline{R}^{-1}\mathbf{S})_j = (\underline{R}^{-1}\underline{B}\,\underline{R}\,\mathbf{y})_j\,. \tag{68}$$

For $\mathbf{S} = 0$ (corresponding to the limiting case $\tau \to \infty$), we have

$$y_j(x,t) = y_j(x - C_j t, 0)\,, \tag{69}$$

which means that the solution does not change in time along the characteristics $x_j(t) = C_j t$. The quantities C_j are called the characteristic speeds.[5] If $\mathbf{u}(x,0)$ is the initial condition, the solution of the set of partial differential equations is

$$u_i(x,t) = \sum_j R_{ij}\,y_j(x - C_j t, 0) \tag{70}$$

with $\mathbf{y}(x,0) = \underline{R}^{-1}\mathbf{u}(x,0)$.[6] Therefore, the spatio-temporal solution $\mathbf{u}(x,t)$ is fully determined by the initial condition. In other words, the future state of the system is given by its previous state, and the principle of causality should be valid.

[5] The idea behind the characteristics is to introduce a parameterization $t(s_1,s_2)$, $x(s_1,s_2)$, which is defined by $\partial t/\partial s_j = 1$ and $\partial x/\partial s_j = C_j$. Then, one can rewrite (68) as $\frac{\partial y_j}{\partial s_j} = \frac{\partial y_j(x,t)}{\partial t}\frac{\partial t}{\partial s_j} + \frac{\partial y_j(x,t)}{\partial x}\frac{\partial x}{\partial s_j} = (\underline{R}^{-1}\underline{B}\,\underline{R}\,\mathbf{y})_j$. In the generalized coordinates s_1 and s_2, the partial differential equations in x and t we were starting with, turn into *ordinary* differential equations. These are much easier to solve.

[6] Note that formulas (69) and (70) only apply to the limiting case $\tau \to \infty$, where the relaxation term of the macroscopic traffic model vanishes.

Appendix 2 Stability Analysis for Macroscopic Traffic Models

In order to understand the dynamics of traffic flows, it is important to find out whether and under what conditions variations in the traffic flow can grow and eventually cause traffic congestion. For this, it is useful to make the solution ansatz

$$\begin{aligned}\delta\rho(x,t) &= \delta\rho_0\, \exp\big(\mathrm{i}\kappa x + (\lambda - \mathrm{i}\omega)t\big) = \delta\rho_0\, \mathrm{e}^{\lambda t}\, \mathrm{e}^{\mathrm{i}(\kappa x - \omega t)}\,,\\ \delta V(x,t) &= \delta V_0\, \exp\big(\mathrm{i}\kappa x + (\lambda - \mathrm{i}\omega)t\big) = \delta V_0\, \mathrm{e}^{\lambda t}\, \mathrm{e}^{\mathrm{i}(\kappa x - \omega t)}\,.\end{aligned} \tag{71}$$

Because of $\exp(\mathrm{i}\kappa x) = \cos(\kappa x) + \mathrm{i}\sin(\kappa x)$ (see Appendix 3), ansatz (71) assumes that the perturbation of the stationary and homogeneous traffic situation can be represented as a periodic function with the wave number κ and wavelength $2\pi/\kappa$. The wave frequency of (71) is ω, while $\delta\rho_0\, \exp(\lambda t)$ and $\delta V_0\, \exp(\lambda t)$ are the amplitudes at time t. That is, if the "growth rate" λ is greater than zero, even small perturbations will eventually grow, which can give rise to "phantom traffic jams". For $\lambda < 0$, however, the initial perturbation will be damped out and the stationary and homogeneous solutions will be re-established, i.e. it is stable with respect to small perturbations.

Below we will see that, for each specification of κ and the average density ρ_e, there exist two solutions $l \in \{+,-\}$ with the frequencies $\omega_l(\kappa)$ and the growth rates $\lambda_l(\kappa)$. All the corresponding specifications of ansatz (71) are solutions of the linearized partial differential equations. The same applies to their superpositions. The general solution for an *arbitrary* initial perturbation is of the form

$$\begin{aligned}\delta\rho(x,t) &= \sum_{l\in\{+,-\}} \int d\kappa\, \delta\rho_0^l(\kappa) \exp\big(\mathrm{i}\kappa x + \big[\lambda_l(\kappa) - \mathrm{i}\omega_l(\kappa)\big]t\big)\,,\\ \delta V(x,t) &= \sum_{l\in\{+,-\}} \int d\kappa\, \delta V_0^l(\kappa) \exp\big(\mathrm{i}\kappa x + \big[\lambda_l(\kappa) - \mathrm{i}\omega_l(\kappa)\big]t\big)\,.\end{aligned} \tag{72}$$

In order to find the possible κ-dependent wave numbers ω and growth rates λ, we insert ansatz (71) into the linearized macroscopic traffic equations (8) and (9) and use the relationship $\mathrm{i}^2 = -1$. The result can represented as an eigenvalue problem:

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} \delta\rho_0 \\ \delta V_0 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{73}$$

where

$$M_{11} = -\tilde{\lambda}\,, \tag{74}$$

$$M_{12} = -\mathrm{i}\kappa\rho_{\mathrm{e}}\,, \tag{75}$$

$$M_{21} = -\frac{\mathrm{i}\kappa}{\rho_{\mathrm{e}}}\frac{\partial P_1}{\partial\rho} + \frac{1}{\tau}\frac{\partial V_{\mathrm{o}}}{d\rho}\,, \tag{76}$$

$$M_{22} = -\tilde{\lambda} - \frac{\mathrm{i}\kappa}{\rho_{\mathrm{e}}}\frac{\partial P_2}{\partial V} + \frac{1}{\tau}\frac{\partial V_{\mathrm{o}}}{\partial V} - \frac{1}{\tau} \tag{77}$$

and

$$\tilde{\lambda} = \lambda - \mathrm{i}\tilde{\omega} \qquad \text{with} \qquad \tilde{\omega} = \omega - \kappa V_{\mathrm{e}}(\rho_{\mathrm{e}})\,. \tag{78}$$

Equation (73) is fulfilled only for certain values of $\tilde{\lambda}(\kappa)$, the so-called "eigenvalues". These depend on the average density ρ_{e} and solve the *characteristic polynomial* of second order in $\tilde{\lambda}$, which is obtained by determining the determinant

$$\det(\underline{M}) = M_{11}M_{22} - M_{21}M_{12} \tag{79}$$

of the matrix $\underline{M}$ and requiring that it becomes zero. The corresponding characteristic polynomial is given by (10).

Appendix 3 Derivation of Formula (19)

Remember that a complex number

$$z = \Re + \mathrm{i}\Im = r\mathrm{e}^{\mathrm{i}\varphi} = r\cos(\varphi) + \mathrm{i}r\sin(\varphi) \tag{80}$$

can be represented in two-dimensional space with coordinates $\Re = \mathrm{Re}(z) = r\cos(\varphi)$ and $\Im = \mathrm{Im}(z) = r\sin(\varphi)$, respectively, called the real part and the imaginary part. The absolute value is given as

$$r = \sqrt{\Re^2 + \Im^2} = \sqrt{(\Re + \mathrm{i}\Im)(\Re - \mathrm{i}\Im)} = \sqrt{z\bar{z}} = |z|\,, \tag{81}$$

where $\bar{z} = \Re - \mathrm{i}\Im = r\mathrm{e}^{-\mathrm{i}\varphi}$ is the conjugate complex number. The angle φ is determined by

$$\tan(\varphi) = \frac{\sin(\varphi)}{\cos(\varphi)} = \frac{\Im}{\Re} = \frac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\,, \tag{82}$$

and the exponential functions is defined as for real numbers by the infinite series expansion

$$\exp(z) = \mathrm{e}^z = \sum_{l=0}^{\infty}\frac{z^l}{l!}\,, \tag{83}$$

where $l! = l \cdot (l-1) \dots 2 \cdot 1$. Therefore, the relationships for exponential functions apply also to the case of complex numbers, i.e. the product of two complex numbers $z_1 = \Re_1 + \mathrm{i}\Im_1 = r_1 \mathrm{e}^{\mathrm{i}\varphi_1}$ and $z_2 = \Re_2 + \mathrm{i}\Im_2 = r_2 \mathrm{e}^{\mathrm{i}\varphi_2}$ is given by

$$\begin{aligned} z_1 z_2 &= \big(\Re_1\Re_2 - \Im_1\Im_2\big) + \mathrm{i}\big(\Re_1\Im_2 + \Im_1\Re_2\big) \\ &= r_1 \mathrm{e}^{\mathrm{i}\varphi_1} r_2 \mathrm{e}^{\mathrm{i}\varphi_2} = r_1 r_2 \mathrm{e}^{\mathrm{i}(\varphi_1+\varphi_2)} \\ &= r_1 r_2 \cos(\varphi_1 + \varphi_2) + \mathrm{i} r_1 r_2 \sin(\varphi_1 + \varphi_2) \,. \end{aligned} \tag{84}$$

As the real and imaginary part are linearly independent of each other, this implies $\Re_1\Re_2 - \Im_1\Im_2 = r_1 r_2 \cos(\varphi_1 + \varphi_2)$ and $\Re_1\Im_2 + \Im_1\Re_2 = r_1 r_2 \sin(\varphi_1 + \varphi_2)$. The inverse of a complex number is given by

$$\frac{1}{z} = \frac{1}{r\mathrm{e}^{\mathrm{i}\varphi}} = \frac{\mathrm{e}^{-\mathrm{i}\varphi}}{r} \,. \tag{85}$$

The imaginary unit i has the property $\mathrm{i}^2 = -1$ and may, therefore, be written as $\mathrm{i} = \sqrt{-1} = \mathrm{e}^{\mathrm{i}\pi/2}$.

The square of complex numbers

$$z = r\mathrm{e}^{\pm\mathrm{i}\varphi} = r\big[\cos(\varphi) \pm \mathrm{i}\sin(\varphi)\big], \tag{86}$$

can, on the one hand, be written as

$$z^2 = r^2\Big[\cos^2(\varphi) \pm 2\mathrm{i}\cos(\varphi)\sin(\varphi) - \sin^2(\varphi)\Big] \,. \tag{87}$$

On the other hand, using the well-known law $\mathrm{e}^{x_1} \cdot \mathrm{e}^{x_2} = \mathrm{e}^{x_1+x_2}$ for the exponential function, we find the alternative representation

$$z^2 = r^2\big(\mathrm{e}^{\pm\mathrm{i}\varphi}\big)^2 = r^2\mathrm{e}^{\pm\mathrm{i}2\varphi} = r^2\big[\cos(2\varphi) \pm \mathrm{i}\sin(2\varphi)\big] \,. \tag{88}$$

Comparing the real parts and using the trigonometric relationship $\sin^2(x) + \cos^2(x) = 1$, we find

$$\cos(2\varphi) = 1 - 2\sin^2(\varphi) = 1 - 2\big[1 - \cos^2(\varphi)\big] = 2\cos^2(\varphi) - 1 \,, \tag{89}$$

from which we can derive the trigonometric formulas

$$\sin^2(\varphi/2) = \frac{1}{2}\big[1 - \cos(\varphi)\big] \tag{90}$$

and

$$\cos^2(\varphi/2) = \frac{1}{2}\big[1 + \cos(\varphi)\big] \,. \tag{91}$$

Therefore, the square root of a complex number is given by

$$\begin{aligned}\sqrt{z} &= \sqrt{r}\mathrm{e}^{\pm \mathrm{i}\varphi/2} = \sqrt{r}\big[\cos(\varphi/2) \pm \mathrm{i}\sin(\varphi/2)\big] \\ &= \sqrt{\frac{1}{2}\big[r + r\cos(\varphi)\big]} \pm \mathrm{i}\sqrt{\frac{1}{2}\big[r - r\cos(\varphi)\big]}. \end{aligned} \tag{92}$$

Considering $\Re = r\cos(\varphi)$, $\Im = r\sin(\varphi)$, and $\Re^2 + \Im^2 = r^2$, we end up with the desired equation

$$\sqrt{\Re \pm \mathrm{i}|\Im|} = \sqrt{\frac{1}{2}\Big(\sqrt{\Re^2 + \Im^2} + \Re\Big)} \pm \mathrm{i}\sqrt{\frac{1}{2}\Big(\sqrt{\Re^2 + \Im^2} - \Re\Big)}. \tag{93}$$

Appendix 4 Meaning of the Group Velocity

Let us start with the representation (72) of the general solution of the linearized system of equations, focussing (for simplicity) on the case $\lambda_l(\kappa) = 0$ and assuming a "Gaussian wave packet" with

$$\delta\rho_0^l(\kappa) = \frac{\mathrm{e}^{-(\kappa-\kappa_0)^2/(2\theta)}}{\sqrt{2\pi\theta}}\,. \tag{94}$$

Via the linear Taylor approximation $\omega_l(\kappa) = \omega_l(\kappa_0) + C_l\,\Delta\kappa$ with $C_l = d\omega_l(\kappa_0)/d\kappa$ and $\Delta\kappa = (\kappa - \kappa_0)$, from (72) we get

$$\begin{aligned}&\delta\rho(x,t) \\ &\quad = \sum_{l\in\{+,-\}} \int\limits_{-\infty}^{\infty} d\kappa\, \frac{\mathrm{e}^{-(\kappa-\kappa_0)^2/(2\theta)}}{\sqrt{2\pi\theta}} \mathrm{e}^{\mathrm{i}[\kappa x - \omega_l(\kappa)t]} \\ &\quad = \sum_{l\in\{+,-\}} \mathrm{e}^{\mathrm{i}[\kappa_0 x - \omega_l(\kappa_0)t]} \int\limits_{-\infty}^{\infty} d\Delta\kappa\, \frac{\mathrm{e}^{-(\Delta\kappa)^2/(2\theta)}}{\sqrt{2\pi\theta}} \mathrm{e}^{\mathrm{i}[\Delta\kappa x - C_l t]} \\ &\quad = \sum_{l\in\{+,-\}} \mathrm{e}^{\mathrm{i}[\kappa_0 x - \omega_l(\kappa_0)t]} \underbrace{\int\limits_{-\infty}^{\infty} d\Delta\kappa\, \frac{\mathrm{e}^{-[\Delta\kappa - \mathrm{i}\theta(x - C_l t)]^2/(2\theta)}}{\sqrt{2\pi\theta}}}_{=1} \\ &\qquad \times \mathrm{e}^{-\theta(x - C_l t)^2/2} \\ &\quad = \sum_{l\in\{+,-\}} \mathrm{e}^{\mathrm{i}[\kappa_0 x - \omega_l(\kappa_0)t]} \mathrm{e}^{-\theta(x - C_l t)^2/2}\,. \end{aligned} \tag{95}$$

While the single waves of frequency $\omega_l(\kappa)$ move with the "phase velocity" $x/t = \omega_l(\kappa)/\kappa$, it turns out that their superposition behaves like a wave with frequency $\omega_l(\kappa_0)$ and speed $x/t = \omega_l(\kappa_0)/\kappa_0$. However, the wave packet or, more exactly speaking, its amplitude $\mathrm{e}^{-\theta(x-C_l t)^2/2}$ is moving with the group velocity $x/t = C_l = d\omega_l(\kappa)/d\kappa$. Note that the case $C_l > \omega_l(\kappa_0)/\kappa_0$, in which the group velocity is greater than the phase velocity (wave velocity), is possible. It is called "anomalous dispersion".

Appendix 5 Linear Stability Analysis of the Optimal Velocity Model

For a linear stability analysis of the optimal velocity model, we imagine the situation of N vehicles i distributed over a circular road of length L. This allows us to assume periodic boundary conditions. The stationary solution for this case is given by $dv_i/dt = 0$ and $dd_i/dt = 0$, which implies

$$\begin{aligned} d_i(t) &= d_\mathrm{e} = L/N = \text{const.} \\ v_{i-1}(t) &= v_i(t) = v_\mathrm{o}(d_\mathrm{e}) = \text{const.} \end{aligned} \tag{96}$$

We are now interested how the deviations from this solution, i.e. the variables

$$\begin{aligned} \delta d_i(t) &= d_i(t) - d_\mathrm{e}\,, \\ \delta v_i(t) &= v_i(t) - v_\mathrm{o}(d_\mathrm{e})\,, \end{aligned} \tag{97}$$

develop in time, assuming that the initial deviations are small, i.e. $\delta d_i(0) \ll d_\mathrm{e}$ and $\delta v_i(0) \ll v_\mathrm{e}(d_\mathrm{e})$. For this, we linearize the model equations (53) and (54) around the stationary and homogeneous solution. This results in

$$\begin{aligned} \frac{d\delta v_i(t)}{dt} &= \frac{1}{\tau}\left(\frac{dv_\mathrm{o}(d_\mathrm{e})}{dd}\delta d_i(t) - \delta v_i(t)\right), \\ \frac{d\delta d_i(t)}{dt} &= \delta v_{i-1}(t) - \delta v_i(t)\,. \end{aligned} \tag{98}$$

For the analysis of stability, we use the solution ansatz

$$\begin{aligned} \delta v_j(t) &= \delta v_0\,\mathrm{e}^{\mathrm{i}2\pi jk/N+\tilde{\lambda}t} = \delta v_0\,\mathrm{e}^{\mathrm{i}j\kappa L/N+\tilde{\lambda}t}\,, \\ \delta d_j(t) &= \delta d_0\,\mathrm{e}^{\mathrm{i}2\pi jk/N+\tilde{\lambda}t} = \delta d_0\,\mathrm{e}^{\mathrm{i}j\kappa L/N+\tilde{\lambda}t}\,, \end{aligned} \tag{99}$$

where $\kappa = 2\pi k/L$ is the so-called wave number, which is inversely proportional to the wave length $2\pi/\kappa = L/k$. Note that, due to the assumed periodic boundary conditions, possible wavelength are fractions L/k of the length L or the circular road. The shortest wave length is given by the average vehicle distance $d_{\rm e} = L/N$, i.e. $k \in \{1, 2, \dots, N\}$. Summing up the functions (99) over these values of k results in the *Fourier representation* of $\delta v_j(t)$ and $\delta d_j(t)$:

$$\delta v_j(t) = \sum_{k=1}^{N} \delta v_k {\rm e}^{{\rm i}2\pi jk/N + \tilde{\lambda} t} \, ,$$

$$\delta d_j(t) = \sum_{k=1}^{N} \delta d_k {\rm e}^{{\rm i}2\pi jk/N + \tilde{\lambda} t} \, . \tag{100}$$

The parameters δv_k and δd_k are determined by the initial conditions of all vehicles j. $\tilde{\lambda} = \lambda - {\rm i}\tilde{\omega}$ are the so-called *eigenvalues*, whose real part λ describes an exponential growth (if $\lambda > 0$) or decay (if $\lambda < 0$), and whose imaginary part $\tilde{\omega}$ reflects oscillation frequencies. δd_0 and δv_0 denote oscillation amplitudes. Inserting this into (98) and dividing by ${\rm e}^{{\rm i}2\pi jk/N + \tilde{\lambda} t}$, we finally obtain

$$\tilde{\lambda} \delta v_0 = \frac{1}{\tau} \left(\frac{d v_{\rm o}(d_{\rm e})}{dd} \delta d_0 - \delta v_0 \right) \, , \tag{101}$$

$$\tilde{\lambda} \delta d_0 = \delta v_0 {\rm e}^{-{\rm i}2\pi k/N} - \delta v_0 = \delta v_0 \Big({\rm e}^{-{\rm i}2\pi k/N} - 1 \Big) \, . \tag{102}$$

Multiplying (101) with $\tilde{\lambda}$ and inserting (102) for $\tilde{\lambda}\, \delta d_0$ in the square brackets gives, after division by δv_0, the *characteristic polynomial* in the eigenvalues $\tilde{\lambda}$, namely

$$\tilde{\lambda}^2 + \frac{1}{\tau} \tilde{\lambda} - \frac{1}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \Big({\rm e}^{-{\rm i}2\pi k/N} - 1 \Big) = 0 \, . \tag{103}$$

The solutions $\tilde{\lambda}(d_{\rm e}, k)$ of this polynomial are the eigenvalues. They read

$$\tilde{\lambda}_\pm(d_{\rm e}, k) = -\frac{1}{2\tau} \pm \sqrt{\frac{1}{4\tau^2} + \frac{1}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \Big({\rm e}^{-{\rm i}2\pi k/N} - 1 \Big)} \, . \tag{104}$$

Again, the square root contains a complex number, which makes it difficult to see the sign of the real value $\lambda_\pm$ of $\tilde{\lambda}_\pm$. However, considering ${\rm e}^{\pm {\rm i}\varphi} = \cos(\varphi) \pm {\rm i} \sin(\varphi)$ and defining the real part

$$\Re = \frac{1}{4\tau^2} - \frac{1}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \big[1 - \cos(2\pi k/N) \big] \tag{105}$$

of the expression under the root and its imaginary part

$$\Im = -\frac{\sin(2\pi k/N)}{\tau} \frac{d v_{\rm o}(d_{\rm e})}{dd} \, , \tag{106}$$

we can again apply the useful formula (16). From this we can conclude that $\lambda = \mathrm{Re}(\tilde{\lambda}) = 0$ if

$$\frac{1}{16\tau^4} = \frac{\Re}{4\tau^2} + \frac{\Im^2}{4}\,, \tag{107}$$

see (21). Inserting (105) and (106), we find

$$\frac{\sin^2(2\pi k/N)}{4\tau^2}\left(\frac{dv_{\mathrm{o}}(d)}{dd}\right)^2 = \frac{1}{4\tau^3}\frac{dv_{\mathrm{o}}(d)}{dd}\big[1-\cos(2\pi k/N)\big]\,, \tag{108}$$

which finally results in the condition

$$\frac{dv_{\mathrm{o}}(d_{\mathrm{e}})}{dd} = \frac{1-\cos(2\pi k/N)}{\tau\sin^2(2\pi k/N)} \overset{k\to 0}{=} \frac{1}{2\tau}\,. \tag{109}$$

The limit $2\pi k/N \to 0$ follows from $\cos(\varphi) \approx 1-\varphi^2/2$ and $\sin(\varphi) \approx \varphi$ in the limit of small wave numbers $\kappa = 2\pi k/L$, i.e. large wave lengths $2\pi/\kappa = L/k$.

It can be demonstrated by numerical analyses that

$$\frac{dv_{\mathrm{o}}(d_{\mathrm{e}})}{dd} > \frac{1}{2\tau} \tag{110}$$

constitutes the instability condition of the optimal velocity model (53) [3]. In other words, if the velocity changes too strongly with the distance, small variations of the vehicle distance or speed will grow and finally cause emergent waves, i.e. the formation of one or several traffic jams. Since the origin of such a breakdown can be infinitesimally small, these traffic jams seem to have no origin. In such situations, one speaks of "phantom traffic jams". A closer analysis for realistic speed-distance relationships $v_{\mathrm{o}}(d)$ shows that traffic tends to be unstable at medium densities $\rho = 1/d$, while it tends to be stable at small and large densities (where the speed does not change much with a variation in the distance). Only a sufficient reduction in the adaptation time τ can avoid an instability of traffic flow, while large delays in the velocity adjustment lead to growing perturbations of traffic flow.

Appendix 6 Correspondence of the Optimal Velocity Model with the Macroscopic Payne Model

As the Payne model has been claimed to be a macroscopic approximation of the optimal velocity model (see [10] and citations therein), it is interesting to compare the instability conditions and characteristic speeds of both models. Therefore, let us make the identifications

$$\rho = \frac{1}{d}\,, \quad V_{\mathrm{e}}(\rho) = v_{\mathrm{o}}\left(\frac{1}{\rho}\right)\,. \tag{111}$$

Then, with the chain rule and the quotient rule of Calculus we can derive

$$\begin{aligned}\left|\frac{dV_\mathrm{e}(\rho)}{d\rho}\right| &= -\frac{dV_\mathrm{e}(\rho)}{d\rho} = -\frac{dv_\mathrm{o}(1/\rho)}{d\rho} = -\frac{dv_\mathrm{o}(d)}{dd}\frac{dd}{d\rho} \\ &= \frac{dv_\mathrm{o}(d)}{dd}\cdot\frac{1}{\rho^2}\,. \end{aligned} \tag{112}$$

Inserting this into (40) gives

$$\rho_\mathrm{e}\left|\frac{dV_\mathrm{e}}{d\rho}\right| = \frac{1}{\rho_\mathrm{e}}\frac{dv_\mathrm{o}(d)}{dd} > \frac{1}{2\rho_\mathrm{e}\tau} \tag{113}$$

or

$$\frac{dv_\mathrm{o}(d_\mathrm{e})}{dd} > \frac{1}{2\tau} \qquad \text{and} \qquad \rho_\mathrm{e}\left|\frac{dV_\mathrm{e}(\rho_\mathrm{e})}{d\rho}\right| = d_\mathrm{e}\frac{dv_\mathrm{o}(d_\mathrm{e})}{dd}\,, \tag{114}$$

where $d_\mathrm{e} = 1/\rho_\mathrm{e}$. This shows the agreement of the instability conditions (40) and (110) and of the characteristic speeds (41) and (57) at the instability threshold.

References

1. D. Chowdhury, L. Santen, A. Schadschneider, Statistical physics of vehicular traffic and some related systems. Phys. Rep. **329**, 199 (2000)
2. D. Helbing, Traffic and related self-driven many-particle systems. Rev. Mod. Phys. **73**, 1067–1141 (2001)
3. T. Nagatani, The physics of traffic jams. Rep. Prog. Phys. **65**, 1331–1386 (2002)
4. K. Nagel, Multi-Agent Transportation Simulations, see http://www2.tu-berlin.de/fb10/ISS/FG4/archive/sim-archive/publications/book/
5. M. Schönhof, D. Helbing, Empirical features of congested traffic state and their implications for traffic modeling. Transp. Sci. **41**, 135–166 (2007)
6. Y. Sugiyama, M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S.-i. Tadaki, S. Yukawa, Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam. New J. Phys. **10**, 033001 (2008)
7. R. Herman, E.W. Montroll, R.B. Potts, R.W. Rothery, Traffic dynamics: Analysis of stability in car following. Oper. Res. **7**, 86–106 (1959)
8. R.D. Kühne, M.B. Rödiger, Macroscopic simulation model for freeway traffic with jams and stop-start waves, in *Proceedings of the 1991 Winter Simulation Conference*, ed. by B.L. Nelson, W.D. Kelton, G.M. Clark (Society for Computer Simulation International, Phoenix, 1991), pp. 762–770
9. M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, Dynamical model of traffic congestion and numerical simulation. Phys. Rev. E **51**, 1035–1042 (1995)
10. D. Helbing, Derivation of non-local macroscopic traffic equations and consistent traffic pressures from microscopic car-following models. Eur. Phys. J. B **69**, 539–548 (2009)
11. C.F. Daganzo, Requiem for second-order fluid approximations of traffic flow. Transp. Res. B **29**, 277–286 (1995)
12. M. Lighthill, G. Whitham, On kinematic waves: II. A theory of traffic on long crowded roads. Proc. R. Soc. Lond. A **229**, 317–345 (1955)
13. P.I. Richards, Shock waves on the highway. Oper. Res. **4**, 42–51 (1956)

14. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)
15. H.J. Payne, Models of freeway traffic and control, in *Mathematical Models of Public Systems*, vol. 1, ed. by G.A. Bekey (Simulation Council, La Jolla, 1971), pp. 51–61
16. H.J. Payne, A critical review of a macroscopic freeway model, in *Research Directions in Computer Control of Urban Traffic Systems*, ed. by W.S. Levine, E. Lieberman, J.J. Fearnsides (American Society of Civil Engineers, New York, 1979), pp. 251–265
17. I. Prigogine, R. Herman, *Kinetic Theory of Vehicular Traffic* (Elsevier, New York, 1971)
18. M. Treiber, A. Hennecke, D. Helbing, Derivation, properties, and simulation of a gas-kinetic-based, non-local traffic model. Phys. Rev. E **59**, 239–253 (1999)
19. V. Shvetsov, D. Helbing, Macroscopic dynamics of multi-lane traffic. Phys. Rev. E **59**, 6328–6339 (1999)
20. D. Helbing, M. Treiber, Numerical simulation of macroscopic traffic equations. Comput. Sci. Eng. **1**(5), 89–99 (1999)
21. C.K.J. Wagner, Verkehrsflußmodelle unter Berücksichtigung eines internen Freiheitsgrades. Ph.D. Thesis, TU Munich, 1997
22. S.P. Hoogendoorn, P.H.L. Bovy, Continuum modeling of multiclass traffic flow. Transp. Res. B **34**(2), 123–146 (2000)
23. S.L. Paveri-Fontana, On Boltzmann-like treatments for traffic flow. A critical review of the basic model and an alternative proposal for dilute traffic analysis. Transp. Res. **9**, 225–235 (1975)
24. L.C. Evans, *Partial Differential Equations* (American Mathematical Society, Providence, 1998)
25. S.F. Farlow, *Partial Differential Equations for Scientists and Engineers* (Dover, New York, 1993)
26. R.J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhäuser, Basel, 1992)
27. A. Aw, M. Rascle, Resurrection of “second order” models of traffic flow. SIAM J. Appl. Math. **60**(3), 916–938 (2000)
28. A. Klar, R. Wegener, Kinetic derivation of macroscopic anticipation models for vehicular traffic. SIAM J. Appl. Math. **60**(5), 1749–1766 (2000)
29. J.M. Greenberg, Extensions and amplifications of a traffic model of Aw and Rascle. SIAM J. Appl. Math. **62**(3), 729–745 (2001)
30. H.M. Zhang, A non-equilibrium traffic model devoid of gas-like behavior. Transp. Res. B **36**, 275–290 (2002)
31. P. Goatin, The Aw-Rascle vehicular traffic flow model with phase transitions. Math. Comput. Model. **44**, 287–303 (2006)
32. M. Garavello, B. Piccoli, Traffic flow on a road network using the Aw-Rascle model. Comm. Partial Differ. Equat. **31**, 243–275 (2006)
33. F. Siebel, W. Mauser, Synchronized flow and wide moving jams from balanced vehicular traffic. Phys. Rev. E **73**, 066108 (2006)
34. Z.-H. Ou, S.-Q. Dai, P. Zhang, L.-Y. Dong, Nonlinear analysis in the Aw-Rascle anticipation model of traffic flow. SIAM J. Appl. Math. **67**(3), 605–618 (2007)
35. J.-P. Lebacque, S. Mammar, H. Haj-Salem, The Aw-Rascle and Zhang’s model: Vacuum problems, existence and regularity of the solutions of the Riemann problem. Transp. Res. B **41**, 710–721 (2007)
36. F. Berthelin, P. Degond, M. Delitala, M. Rascle, A model for the formation and evolution of traffic jams. Arch. Ration. Mech. Anal. **187**, 185–220 (2008)
37. J. Keizer, *Statistical Thermodynamics of Nonequilibrium Processes* (Springer, New York, 1987)
38. W.F. Phillips, A kinetic model for traffic flow with continuum implications. Transp. Plan. Tech. **5**, 131–138 (1979)
39. B.S. Kerner, P. Konhäuser, Cluster effect in initially homogeneous traffic flow. Phys. Rev. E **48**(4), R2335–R2338 (1993)
40. H.Y. Lee, H.-W. Lee, D. Kim, Origin of synchronized traffic flow on highways and its dynamic phase transitions. Phys. Rev. Lett. **81**, 1130–1133 (1998)

41. M. Treiber, A. Hennecke, D. Helbing, Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**, 1805–1824 (2000)
42. A. Hood, Characteristics, in *Encyklopedia of Nonlinear Science*, ed. by A. Scott (Routledge, New York, 2005)
43. D. Helbing, Derivation and empirical validation of a refined traffic flow model. Phys. A **233**, 253–282 (1996). See also http://arxiv.org/abs/cond-mat/9805136

Theoretical vs. Empirical Classification and Prediction of Congested Traffic States*

Dirk Helbing, Martin Treiber, Arne Kesting, and Martin Schönhof

Abstract Starting from the instability diagram of a traffic flow model, we derive conditions for the occurrence of congested traffic states, their appearance, their spreading in space and time, and the related increase in travel times. We discuss the terminology of traffic phases and give empirical evidence for the existence of a phase diagram of traffic states. In contrast to previously presented phase diagrams, it is shown that "widening synchronized patterns" are possible, if the maximum flow is located inside of a metastable density regime. Moreover, for various kinds of traffic models with different instability diagrams it is discussed, how the related phase diagrams are expected to approximately look like. Apart from this, it is pointed out that combinations of on- and off-ramps create different patterns than a single, isolated on-ramp.

1 Introduction

While traffic science makes a clear distinction between free and congested traffic, the empirical analysis of spatiotemporal congestion patterns has recently revealed an unexpected complexity of traffic states. Early contributions in traffic physics

*First published in: The European Physical Journal B **69**(4), 583–598, DOI: 10.1140/epjb/e2009-00140-5 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ).

D. Helbing (✉)
ETH Zurich, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland
e-mail: dhelbing@ethz.ch

M. Treiber · A. Kesting · M. Schönhof
Institute for Transport & Economics, Dresden University of Technology, Andreas-Schubert-Str. 23, 01062 Dresden, Germany
e-mail: martin.treiber@vwi.tu-dresden.de; kesting@vwi.tu-dresden.de; martin@vwi.tu-dresden.de

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_5,
© Springer-Verlag Berlin Heidelberg 2013

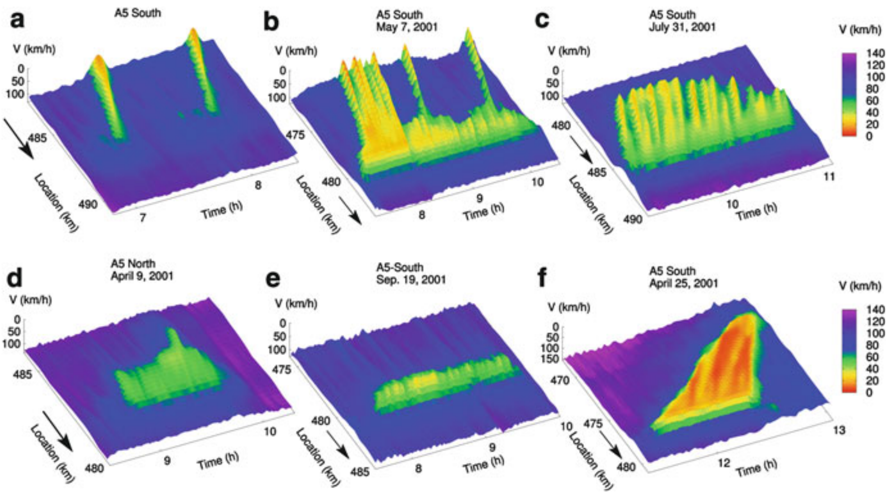


Fig. 1 Examples of elementary patterns of congested traffic measured on the German freeway A5 close to Frankfurt. For better illustration of the traffic patterns, speeds are displayed *upside down*. The driving direction is indicated by *arrows*. *Top row*: (**a**) Moving clusters (MC), (**b**) stop-and-go waves (SGW), (**c**) oscillating congested traffic (OCT). *Bottom row*: (**d**) Widening synchronized pattern (WSP), (**e**) pinned localized cluster (PLC), and (**f**) homogeneous congested traffic (HCT). The spatiotemporal velocity fields have been reconstructed from 1-min data of double-loop detector cross sections using the "adaptive smoothing method" protect[50]

focussed on the study of so-called "phantom traffic jams" [56], i.e. traffic jams resulting from minor perturbations in the traffic flow rather than from accidents, building sites, or other bottlenecks. This subject has recently been revived due to new technologies facilitating experimental traffic research [48]. Related theoretical and numerical stability analyses were—and still are—often carried out for setups with periodic boundary conditions. This is, of course, quite artificial, as compared to real traffic situations. Therefore, in response to empirical findings [27], physicists have pointed out that the occurrence of congested traffic on real freeways normally results from a *combination* of three ingredients [13, 31]:

1. A high traffic volume (defined as the freeway flow plus the actual on-ramp flow, see below).
2. A spatial inhomogeneity of the freeway (such as a ramp, gradient, or change in the number of usable lanes).
3. A temporary perturbation of the traffic flow (e.g. due to lane changes [30] or long-lasting overtaking maneuvers of trucks [12, 44]).

The challenge of traffic modeling, however, goes considerably beyond this. It would be favorable, if the traffic dynamics could be understood on the basis of elementary traffic patterns [44] such as the ones depicted in Fig. 1, and if complex traffic patterns (see, e.g., Fig. 2) could be understood as combinations of them, considering interaction effects.

It was proposed that the occurrence of elementary congested traffic states could be classified and predicted by a phase diagram [13, 53]. Furthermore, it was

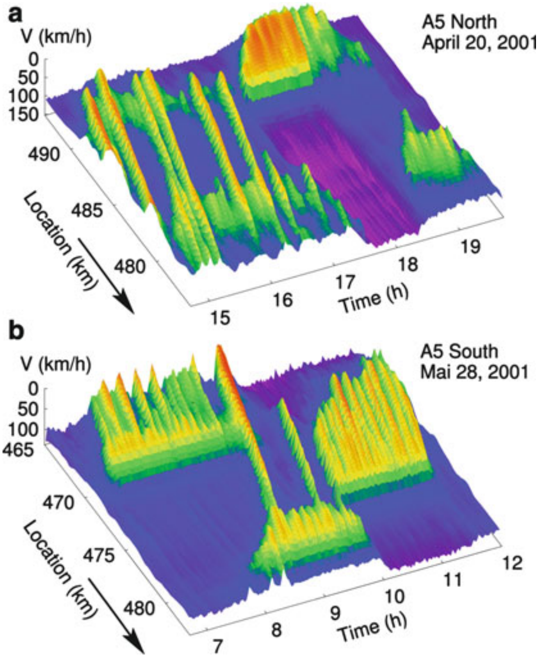


Fig. 2 Two examples of complex traffic states measured on the German freeway A5 close to Frankfurt. *Top*: On the A5 North, an accident occurs at $x = 487.5$ km at the time $t = 17{:}13$ h, which causes a HCT pattern that turns into an OCT pattern as the upstream traffic flow goes down. The capacity drop related to the congestion pattern reduces the downstream flow and leads to a dissolution of the previous SGW pattern over there around $t = 18{:}00$ h. *Bottom*: On the freeway A5 South, the stop-and-go waves induced by a bottleneck at $x = 480$ km replace the OCT at the bottleneck near $x = 470$ km. At time $t = 9{:}50$ h, the waves induce an accident at $x = 478.33$ km, which triggers a new OCT pattern further upstream. The related capacity drop, in turn, causes the previous OCT state at $x \approx 480$ km to dissolve

suggested that this phase diagram can be derived from the instability diagram of traffic flow and the outflow from congested traffic. This idea has been taken up in many other publications, also as a means of studying, visualizing, and classifying properties of traffic models [3, 15, 35, 36]. However, it has been claimed that the phase diagram approach would be insufficient [20]. While some of the criticism is due to misunderstandings, as will be shown in Sect. 7.1, the classical phase diagrams lack, in fact, the possibility of "widening synchronized patterns" (WSP) proposed by Kerner and Klenov [23], see Fig. 1d.

In this paper, we will start in Sect. 2 with a discussion of the somewhat controversial notion of "traffic phases" and the clarification that we use it to distinguish congestion patterns with a qualitatively different spatiotemporal appearance. In Sect. 3 we will show that existing models can produce all the empirically observed patterns of Fig. 1, when simulated in an open system with a bottleneck. We will then present a derivation and explanation of the idealized, schematic phase diagram of traffic states in Sect. 4. In contrast to previous publications, we will assume that the critical density ρ_{c2}, at which traffic becomes linearly unstable, is greater than the density $\rho_{\max}$, where the maximum flow is reached (see Appendix 1 for details). As a consequence, we will find that "widening synchronized pattern" *do* exist within the phase diagram approach, even for models with a fundamental diagram. While this analysis is carried out for single, isolated bottlenecks, Sect. 5 will introduce how to generalize it to the case of multi-ramp setups. In Sect. 6, we will discuss other possible types of phase diagrams, depending on the stability properties of the considered model. Afterwards, in Sect. 7, we will present recent empirical data supporting our theoretical phase diagram. Sections 7.1 and 8 will finally try to overcome some misunderstandings regarding the phase diagram concept and summarize our findings.

2 On the Definition of Traffic Phases

Before we present the phase diagram of traffic states, it must be emphasized that some confusion arises from the different use of the term "(traffic) phase". In thermodynamics, a "phase" corresponds to an equilibrium state in a region of the parameter space of thermodynamic variables (such as pressure and temperature), in which the appropriate free energy is *analytic*, i.e., all first and higher-order derivatives with respect to the thermodynamic variables exist. One speaks of a *first-order* phase transition, if a first derivative, or "order parameter", is discontinuous, and of a "second-order" or "continuous" phase transition, if the first derivatives are continuous but a second derivative (the "susceptibility") diverges. What consequences does this have for defining "traffic phases"?

Although traffic flow is a self-driven nonequilibrium system, it has been shown [1] that much of the equilibrium concepts can be transferred to driven or self-driven non-equilibrium systems by appropriately redefining them. Furthermore, concepts of classical thermodynamics have been successfully applied to nonequilibrium physical and nonphysical systems, yielding quantitatively correct results. This includes, for example, the application of the fluctuation-dissipation theorem [33] (originally referring to equilibrium phenomena) to vehicular traffic [51].

In contrast to classical thermodynamics, *nonequilibrium* phase transitions are *possible* in one-dimensional systems [7]. However, according to the definition of phase transitions, one needs to make sure that details of the boundary conditions or finite-size effects do not play a role for the characteristic properties of the phase. Furthermore, one must define suitable order parameters or susceptibilities. While the first propositions have been already made a decade ago [38], there is no general agreement regarding the quantity that should be chosen for the order parameter. Candidates include the density, the fraction of vehicles in the congested state [38], the average velocity or flow, or the variance of density, velocity, or flow. Whenever one observes a discontinuous or hysteretic transition in a large enough system, there is no need to define an order parameter, as this already implies a first-order phase transition. For *continuous*, symmetry-braking phase transitions, the deviation from the more symmetric state (e.g. the amplitude of density variations as compared to the homogeneous state) seems to be an appropriate order parameter.

To summarize the above points, it appears that thermodynamic phases *can*, in fact, be defined for traffic flow. In connection with transitions between different traffic states at bottlenecks, we particularly mention the notion of *boundary-induced phase transitions* [32,43,46]. Here, the boundary conditions have been mainly used as a means to control the average density in the open system under consideration.

In publications on traffic, a "phase" is often interpreted as "traffic pattern" or "traffic state with a typical spatio-temporal appearance". Such states depend on the respective boundary conditions. In this way, models with several phases can produce a multitude of spatiotemporal patterns. It should become clear from these considerations that the various proposed "phase diagrams" do *not* relate to *thermodynamic* phases, but classify spatio-temporal states, as is common in systems theory. In these non-thermodynamic phase diagrams, the "phase space" is spanned by certain control parameters, e.g. by suitably parameterized boundary conditions, by inhomogeneities (bottleneck strengths), or by model parameters [54]. For example, the phase diagrams discussed in [13, 20] and this paper contain the axes "main inflow" (i.e., an upstream boundary condition) and "on-ramp flow" (characterizing the bottleneck strength).

In any case, *empirical* observations of the traffic dynamics relate to the spatiotemporal traffic patterns, and not to the thermodynamic phases. Therefore, the quality of a traffic model should be assessed by asking whether it can produce all observed kinds of spatio-temporal traffic patterns, including the conditions for their appearance.

3 Congested Traffic States

When simulating traffic flow with the "microscopic" intelligent driver model (IDM) [53], the optimal velocity model (OVM) [2], the non-local, gas-kinetic-based traffic model (GKT) [52], or the "macroscopic" Kerner–Konhäuser model [24] (with the parameter set chosen by Lee et al. [34]), we find free traffic flow and different kinds of congestion patterns, when the ramp flow Q_{on} and the upstream arrival flow Q_{up} on the freeway are varied. The diversity of traffic patterns is

1. Due to the possibility of having either *locally constraint* or *spatially extended* congestion.[1]
2. Due to the possibility of having stable, unstable or free traffic flows.

[1] Note that traffic patterns which appear to be localized, but continue to grow in size, belong to the spatially extended category of traffic states. Therefore, "widening moving clusters" (WMC) are classified as extended congested traffic, while the similarly looking "moving localized clusters" (MLC) are not. According to Fig. 6, however, the phases of both states are located next to each other, so one could summarize both phases by one area representing "moving clusters" (MC).

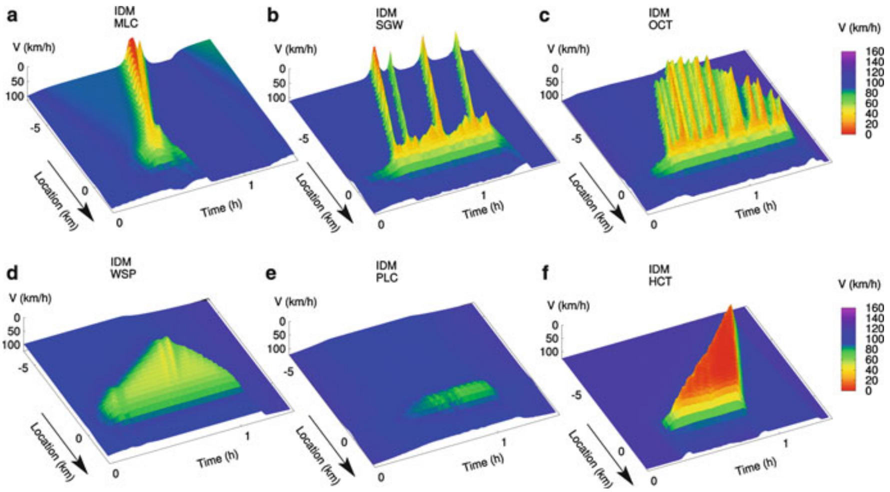


Fig. 3 Simulation of traffic on a freeway with an on-ramp at location $x = 0$ km using the intelligent driver model (IDM) [53] with parameters corresponding to an instability diagram as illustrated in Fig. 4d. The macroscopic velocity field was extracted from the simulated trajectories by placing virtual detectors every 500 m and determining the velocity with the same method [50] that has been used for the data. Depending on the respective traffic flows on the ramp and on the freeway, different kinds of congested traffic states emerge: a moving cluster (MC), a pinned localized cluster (PLC), ("triggered") stop-and-go waves (SGW), oscillating congested traffic (OCT), or homogeneous congested traffic (HCT). During the first minutes of the simulation, the flows on the freeway and the on-ramp were increased from low values to their final values. Since the assumed flows fall into a metastable traffic regime, the actual breakdown was initiated by additional perturbations of the ramp flow

Typical representatives of congested traffic patterns obtained by computer simulations with the intelligent driver model [53] are shown in Fig. 3. Notice that all empirical patterns displayed in Fig. 1 can be reproduced.

One can distinguish the different traffic states (i.e. congestion patterns) by analyzing the temporal *and* spatial dependence of the average velocity $V(x,t)$: If $V(x,t)$ stays above a certain threshold V_{crit}, where x is varied within a homogeneous freeway section upstream of a bottleneck, we call the traffic state *free traffic* (FT), otherwise congested traffic.[2] If these speeds fall below V_{crit} only over a short freeway subsection, and the length of this section is approximately stable or stabilizes over time, we talk about *localized clusters* (LC), otherwise of *spatially extended congestion* states (see also footnote 1).

According to our simulations, there are two forms of localized clusters: *Pinned localized clusters* (PLC) stay at a fixed location over a longer period of time, while *moving localized clusters* (MLC) propagate upstream with the characteristic

[2] A typical threshold for German freeways would be $V_{\text{crit}} \approx 80$ km/h.

speed c_0. These states have to be contrasted with extended congested traffic[3]: *Stop-and-go waves* (SGW) may be interpreted as a sequence of several moving localized clusters. Alternatively, they may be viewed as special case of *oscillating congested traffic* (OCT), but with free traffic flows of about $Q_{\text{out}} \gtrsim$ 1,800 vehicles/h/lane between the upstream propagating jams. Generally, however, OCT is just characterized by oscillating speeds in the congested range, i.e. unstable traffic flows. If the speeds are congested over a spatially extended area, but not oscillating,[4] we call this *homogeneous congested traffic* (HCT). It is typically related with low vehicle velocities.

In summary, besides free traffic, the above mentioned and some other traffic models predict five different, spatio-temporal patterns of congested traffic states at a simple on-ramp bottleneck: PLC, MLC, SGW, OCT, and HCT. Similar traffic states have been identified for flow-conserving bottlenecks in car-following models [8,14], and for on-ramps and other types of bottlenecks in macroscopic models [13,34].

In contrast to this past work, we have also simulated an additional traffic pattern (Fig. 3d). This pattern has a similarity to the *widening synchronized pattern* (WSP) proposed by Kerner in the framework of his three-phase traffic theory [22]. In the following section, we show how this pattern may be understood in the framework of models with a fundamental diagram.

4 Derivation and Explanation of the Phase Diagram of Traffic States

It turns out that the possible traffic patterns in open systems with bottlenecks are mainly determined by the instability diagram (see Fig. 4), no matter if the model is macroscopic or microscopic. This seems to apply at least for traffic models with a fundamental diagram, which we will focus on in the following sections. Due to the close relationship with the instability diagram, the preconditions for the possible occurrence of the different traffic states can be illustrated by a *phase diagram*. Figures 5 and 6 show two examples. Each area of a phase diagram represents the combinations of upstream freeway flows Q_{up} and bottleneck strengths ΔQ, for which a certain kind of traffic state can exist.

It is obvious that an on-ramp flow $Q_{\text{on}}(t)$, for example, causes a bottleneck, as it consumes some of the capacity of the freeway. $Q_{\text{on}}(t)$ represents the flow actually entering the freeway via the on-ramp, i.e. the flow *leaving* the on-ramp and not the flow *entering* the on-ramp.[5] We assume that $Q_{\text{on}}(t)$ is known through a

[3] Which includes "widening moving clusters" (see Fig. 1a and footnote 1).

[4] When averaging over spatial and temporal intervals that sufficiently eliminate effects of heterogeneity and pedal control in real vehicle traffic.

[5] When the freeway is busy, it may happen that these two flows are different and that a queue of vehicles forms on the on-ramp. Of course, it is an interesting question to determine how the

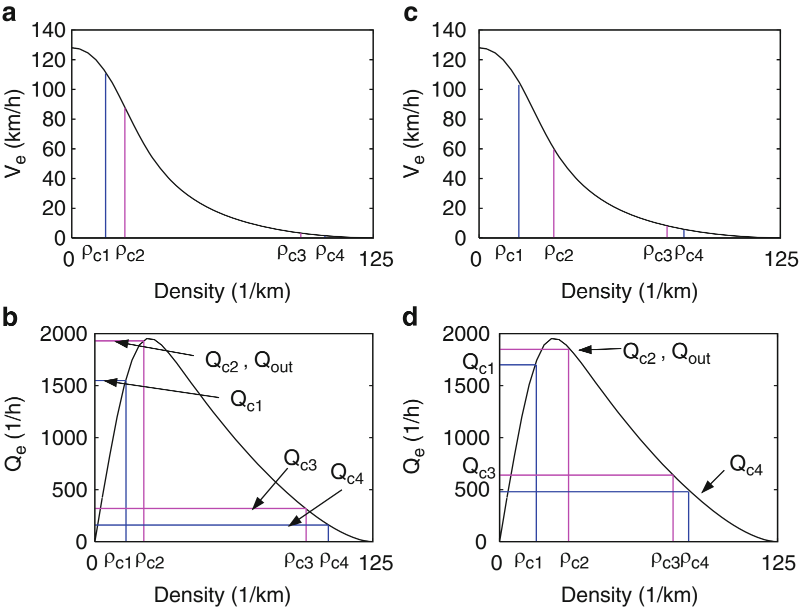


Fig. 4 Illustration of stable, linearly unstable, and metastable density regimes within velocity-density diagrams $V_\mathrm{e}(\rho)$ (*top*) and the flow-density diagrams $Q_\mathrm{e}(\rho)$ (*bottom*). Traffic is stable for $\rho < \rho_{c1}$ and $\rho > \rho_{c4}$ and linearly unstable for $\rho_{c2} < \rho < \rho_{c3}$. These two regimes are separated by a low-density and a high-density region of metastable traffic given by the intervals $\rho_{c1} < \rho < \rho_{c2}$ and $\rho_{c3} < \rho < \rho_{c4}$, respectively. In the metastable regimes, perturbations in the traffic flow grow, if their size is larger than a certain critical amplitude [10], otherwise they fade away. The critical amplitude is largest towards the boundaries $\rho_{\mathrm{c}1}$ and $\rho_{\mathrm{c}4}$ of unconditionally stable traffic flow, while it goes to zero towards the boundaries $\rho_{\mathrm{c}2}$ and $\rho_{\mathrm{c}3}$ of linearly unstable traffic. Note that the metastable and unstable regimes may vanish for certain traffic models or parameter specifications. The possible types of congested traffic patterns depend on the existence of the different stability regimes and on the relative position of their boundaries with respect to the density $\rho_{\max}$ at capacity $Q_{\max}$ (maximum flow). The *left figures* show the situation for $\rho_{c2} < \rho_{\max}$, the *right figures* the situation for $\rho_{c2} > \rho_{\max}$

suitable measurement. Having clarified this, we define the *bottleneck strength* due to an on-ramp by the entering ramp flow, divided by the number I_fr of freeway lanes:

$$\Delta Q(t) = \Delta Q_\mathrm{on}(t) = \frac{Q_\mathrm{on}(t)}{I_\mathrm{fr}} . \tag{1}$$

This is done so, because the average flow ΔQ added to each freeway lane by the on-ramp flow corresponds to the capacity that is not available anymore for the traffic flow Q_up coming from the upstream freeway section. As a consequence, congestion

entering ramp flow depends on the freeway flow, but this is not the focus of attention here, as this formula is not required for the following considerations.

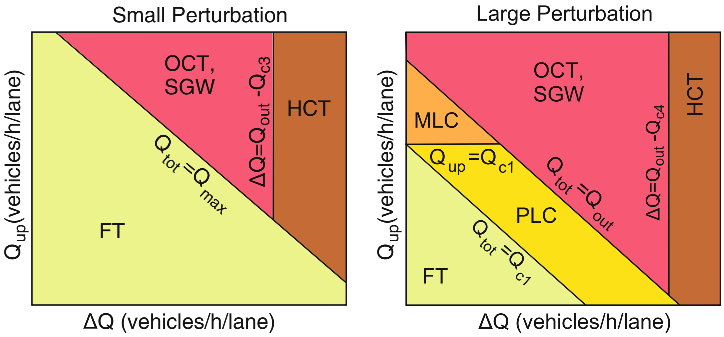


Fig. 5 Schematic (idealized) phase diagrams for the expected traffic patterns as a function of the upstream freeway flow $Q_{\rm up}$ and the ramp flow ΔQ, as studied in [13, 53]. The *left figure* is for negligible, the *right figure* for large perturbations. The situation for medium-sized perturbations can lie anywhere between these two extremes. For example, in the area marked as PLC, one may find free traffic or pinned localized clusters, or in some of the area attributed to HCT, one may find SGW or OCT states. The assumed instability diagram underlying the above schematic phase diagrams is depicted in Fig. 4a, b. With $\rho_{\rm c1} < \rho_{\rm c2} < \rho_{\rm max} < \rho_{\rm c3} < \rho_{\rm c4} < \rho_{\rm jam}$, it assumes no degeneration of the critical densities $\rho_{{\rm c}k}$ and a stable flow at high densities. Note that, for illustrative reasons, we have set aside the exact correspondence of the flow values $Q_{{\rm c}k}$

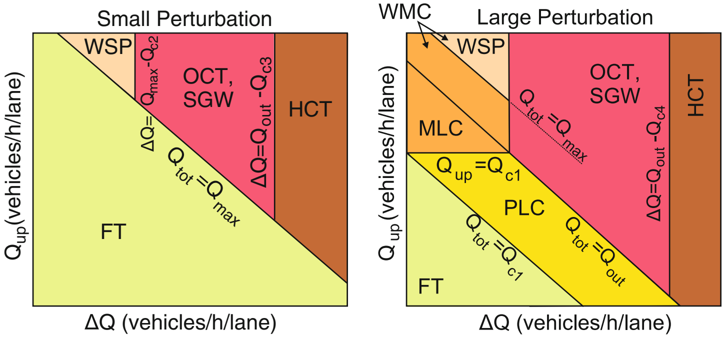


Fig. 6 Schematic phase diagram as in Fig. 5, but for the instability diagram represented by Fig. 4c, d. In contrast to Fig. 5, traffic flow at capacity is metastable ($\rho_{c1} < \rho_{\rm max} < \rho_{c2}$), which leads to a greater variety of traffic states in the *upper left corner* of the phase diagram. In particularly, we find "widening synchronized patterns" (WSP). "OCT, SGW" means that one expects to find oscillating congested traffic or stop-and-go waves, but not necessarily both. Together with "widening moving clusters" (WMC, see footnote 1) they form the area of extended oscillatory congestion. However, the WMC and MLC phases may also be summarized by one area representing "moving clusters" (MC)

may form upstream of the ramp. In the following, we will have to determine the density inside the forming congestion pattern and where in the instability diagram it is located. It will turn out that, given certain values of $Q_{\rm up}$ and ΔQ, the different regions of the phase diagram can be related with the respectively observed or simulated spatiotemporal patterns. We distinguish free traffic and different kinds of localized congested traffic as well as different kinds of extended congested traffic.

When contrasting our classification of traffic states with Kerner's one [20], we find the following comparison helpful:

1. According to our understanding, what we call "extended congested traffic" may be associated with Kerner's "synchronized flow". In particular, the area where Kerner's phase diagrams predict a "general pattern" matches well with the area, where we expect OCT and HCT states.
2. "Moving clusters"[6] may be associated with "wide moving jams" and/or "moving synchronized patterns" (MSP).
3. "Stop-and-go waves" appear to be related with multiple "wide moving jams" generated by the "pinch effect".
4. "Pinned localized clusters" may related to Kerner's "localized synchronized pattern" (LSP).
5. Kerner's "widening synchronized pattern" (WSP) and "dissolving general pattern" (DGP) did not have a correspondence with results of our own computer simulations so far. These states are predicted to appear for high freeway flows and low bottleneck strengths. In the following subsections, we report that, quite unexpectedly, similar results are found for certain traffic models having a fundamental diagram.

The phase diagram can not only be determined *numerically*. It turns out that the borderlines between different areas (the so-called phase boundaries) can also be *theoretically* understood, based on the flows

$$Q_{ck} = Q_e(\rho_{ck}) \tag{2}$$

at the instability thresholds ρ_{ck} ($k = 1, \ldots, 4$), the maximum flow capacity $Q_{\max}$ under free flow conditions, and the dynamic flow capacity, i.e. the characteristic outflow Q_{out} from congested traffic [25] (see Fig. 4). $Q_e(\rho)$ represents the equilibrium flow-density relationship, which is also called the "fundamental diagram".

The exact shape and location of the separation lines between different kinds of traffic states depend on the traffic model and its parameter values.[7] Furthermore, the characteristic outflow Q_{out} typically depends on the type and strength of the bottleneck.[8] For the sake of simplicity of our discussion, however, we will assume constant outflows Q_{out} in the following.

The meaning of the different critical density thresholds ρ_{ck} and flow thresholds $Q_{ck} = Q_e(\rho_{ck})$, respectively, is described in the caption of Fig. 4. Note that the

[6] That is, "moving localized clusters" (MLC) and "widening moving clusters" (WMC), see footnote 1 and Sect. 4.2.

[7] Since the model parameters characterize the prevailing driving style as well as external conditions such as weather conditions and speed limits, the separation lines ("phase boundaries") and even the existence of certain traffic patterns are subject to these factors, see Sect. 7.

[8] For example, in most models, the outflow Q_{out} downstream of an on-ramp bottleneck decreases with the bottleneck strength and increases with the length of the on-ramp [53, 55].

density ρ_{c2} may be smaller or larger than the density $\rho_{\max}$ at capacity, where the maximum flow $Q_{\max}$ is reached. Previous computer simulations and phase diagrams mostly assumed parameters where traffic at capacity is linearly unstable ($\rho_{c2} < \rho_{\max} < \rho_{c3}$), which is depicted in Figs. 4a, b. However, in some traffic models such as the IDM [53], the stability thresholds can be controlled in a flexible way by varying their model parameters (see Appendix 2). In the following, we will focus on the case where traffic at capacity is metastable ($\rho_{c2} > \rho_{\max} > \rho_{c1}$), cf. Fig. 4c, d.[9] As will be shown in the next subsection, this appears to offer an alternative explanation of the "widening synchronized pattern" (WSP) introduced in [22], see Fig. 3d. Simpler cases will be addressed in Sect. 6 below.

4.1 Transition to Congested Traffic for Small Bottlenecks

In the following, we restrict our considerations to situations with one bottleneck only, namely a single on-ramp. Combinations of off- and on-ramps are not covered by this section. They will be treated later on (see Sect. 5).

For matters of illustration, we assume a typical rush hour scenario, in which the total traffic volume

$$Q_{\text{tot}}(t) = Q_{\text{up}}(t) + \Delta Q(t)\,, \tag{3}$$

i.e. the sum of the flow $Q_{\text{up}}(t)$ sufficiently upstream of the ramp bottleneck and the on-ramp flow $\Delta Q(t)$ per freeway lane, is increasing with time t. As long as traffic flows freely, the flow downstream of the bottleneck corresponds to the total flow $Q_{\text{tot}}(t)$, while the upstream flow is $Q_{\text{up}}(t)$.

When the total flow $Q_{\text{tot}}(t)$ exceeds the critical density ρ_{c1}, it enters the metastable density regime. That is, large enough perturbations may potentially grow and cause a breakdown of the traffic flow. However, often the perturbations remain comparatively small, and the total traffic volume rises so quickly that it eventually exceeds the maximum freeway capacity

$$Q_{\text{tot}} = Q_{\text{up}} + \Delta Q > Q_{\max} = \max_{\rho} Q_{\text{e}}(\rho) = Q_{\text{e}}(\rho_{\max})\,. \tag{4}$$

This is reflected in the left phase diagram in Fig. 6 by the diagonal line separating the states "FT" and "WSP". (Note that $\rho_{\max}$ represents the density, for which the maximum free traffic flow occurs, not the jam density ρ_{jam}.)

When the total traffic volume Q_{tot} exceeds the maximum capacity $Q_{\max}$, a platoon of vehicles will form upstream of the bottleneck. Since, in this section, we assume *metastable* traffic at maximum capacity $Q_{\max}$ (see Fig. 7 in [12]), this will not instantaneously lead to a traffic breakdown with an associated capacity

[9] The IDM parameters for plots (a) and (b) are given by $v_0 = 128$ km/h, $T = 1$ s, $s_0 = 2$ m, $s_1 = 10$ m, $a = 0.8$ m/s^2, and $b = 1.3$ m/s^2. To generate plots (c) and (d), the acceleration parameter was increased to $a = 1.3$ m/s^2, while the other parameters were left unchanged.

drop. Thus, the flow downstream of the bottleneck remains limited to $Q_{\max}$ (at least temporarily). As the on-ramp flow takes away an amount ΔQ of the maximum capacity $Q_{\max}$, the (maximum) flow upstream of the bottleneck is given by

$$Q_{\text{bot}} = Q_{\max} - \Delta Q \,. \tag{5}$$

When the actual upstream flow Q_{up} exceeds this value, a mild form of congestion will result upstream of the ramp. The density of the forming vehicle platoon is predicted to be

$$\rho_{\text{bot}} = \rho_{\text{cg}}(Q_{\text{bot}}) = \rho_{\text{cg}}\big(Q_{\max} - \Delta Q\big) > \rho_{\max} \,, \tag{6}$$

where $\rho_{\text{cg}}(Q)$ is the density corresponding to a stationary and homogeneous *congested* flow of value Q (i.e. it is the inverse function of the "congested branch" of the fundamental diagram).

According to the equation for the propagation speed of shockwaves (see [57]), the upstream front of the forming vehicle platoon is expected to propagate upstream at the speed

$$C_1(t) = \frac{Q_{\text{up}} - Q_{\text{bot}}}{\rho_{\text{fr}}(Q_{\text{up}}) - \rho_{\text{cg}}(Q_{\text{bot}})} \,, \tag{7}$$

where $\rho_{\text{fr}}(Q)$ is the density of stationary and homogeneous traffic at a given flow Q (i.e. the inverse function of the "free branch" of the fundamental diagram.)

Note that this high-flow situation can persist for a significant time period only, if the flow Q_{bot} in the platoon is stable or metastable. This is the case if one of the following applies:

(a) The traffic flow is unconditionally stable for all densities such as in the Lighthill–Whitham model [37]. This will be discussed in Sect. 6 below.
(b) Traffic flow at capacity is metastable and the bottleneck is sufficiently weak. This gives rise to the *widening synchronized pattern* (WSP), as will be discussed in the rest of this subsection.

By WSP, we mean a semi-congested extending traffic state without large-scale oscillations or significant velocity drops below, say, 30–40 km/h [20]. Putting aside stochastic accelerations or heterogeneous driver-vehicle populations, this corresponds to (meta-)stable vehicle platoons at densities greater than, but close to the density $\rho_{\max}$ at capacity. This can occur when ρ_{bot} lies in the metastable density range, i.e. $\rho_{\max} < \rho_{\text{bot}} < \rho_{\text{c2}}$, corresponding to $Q_{\max} > Q_{\text{bot}} = Q_{\max} - \Delta Q > Q_{\text{c2}}$ or

$$\Delta Q < Q_{\max} - Q_{\text{c2}}. \tag{8}$$

In Fig. 6, this condition belongs to the area left of the vertical line separating the WSP and OCT states. If the bottleneck strength ΔQ becomes greater than $Q_{\max} - Q_{\text{c2}}$, or if $\rho_{\text{cg}}(Q_{\text{bot}})$ lies in the metastable regime and perturbations in the traffic flow are large enough, traffic flow becomes unstable and breaks down. After the related capacity drop by the amount

$$\Delta Q_{\mathrm{drop}} = Q_{\max} - Q_{\mathrm{out}}\,, \tag{9}$$

the new, "dynamic" capacity will be given by the outflow Q_{out} from congested traffic [53]. Obviously, the capacity drop causes the formation of more serious congestion.[10] This is illustrated in the right phase diagram of Fig. 6 by the offset between the diagonal lines separating free traffic from WSP and the other extended congested states (OCT, SGW, and HCT). In the following, we will focus on the traffic states *after* the breakdown of freeway capacity from $Q_{\max}$ to Q_{out} has taken place.

4.2 Conditions for Different Kinds of Congested Traffic After the Breakdown of Traffic Flow

For the sake of simplicity, we will assume the case

$$Q_{\mathrm{c4}} < Q_{\mathrm{c3}} < Q_{\mathrm{c1}} \le Q_{\mathrm{out}} \le Q_{\mathrm{c2}} < Q_{\max}\,, \tag{10}$$

which seems to be appropriate for real traffic (particularly in Germany). However, depending on the choice of model parameters, other cases are possible. The conclusions may be different, then, but the line of argumentation is the same. In the following, we will again assume $\rho_{\mathrm{c2}} \ge \rho_{\max}$, so that the maximum flow $Q_{\max}$ is metastable. Therefore, it can persist for some time, until the maximum flow state is destabilized by perturbations or too high traffic volumes $Q_{\mathrm{tot}}(t)$, which eventually cause a breakdown of the traffic flow. (For $\rho_{\mathrm{c2}} < \rho_{\max}$, the capacity drop happens automatically, whenever $Q_{\mathrm{tot}}(t) > Q_{\max}$.)

After the breakdown of traffic flow, the traffic situation downstream is given by the outflow Q_{out} from (seriously) congested traffic. As the actually entering ramp flow requires the capacity ΔQ per lane, the flow upstream of the bottleneck is limited to

$$Q_{\mathrm{cong}} = Q_{\mathrm{out}} - \Delta Q\,. \tag{11}$$

In analogy to (7), the upstream front of this congested flow is expected to propagate with the velocity

$$C_2(t) = \frac{Q_{\mathrm{up}} - Q_{\mathrm{cong}}}{\rho_{\mathrm{fr}}(Q_{\mathrm{up}}) - \rho_{\mathrm{cg}}(Q_{\mathrm{cong}})}\,, \tag{12}$$

as the upstream freeway flow Q_{up} is assumed to be free. The downstream end of the congested flow Q_{cong} remains located at the bottleneck [9].

Figure 7 shows that the propagation of the upstream front according to (12) agrees remarkably well with empirical observations, not only for homogeneous

[10] And the condition $Q_{\mathrm{up}} + \Delta Q < Q_{\mathrm{out}}$ for the gradual *dissolution* of the resulting congestion pattern is harder to fulfil than the condition $Q_{\mathrm{up}} + \Delta Q < Q_{\max}$ implied by (4).

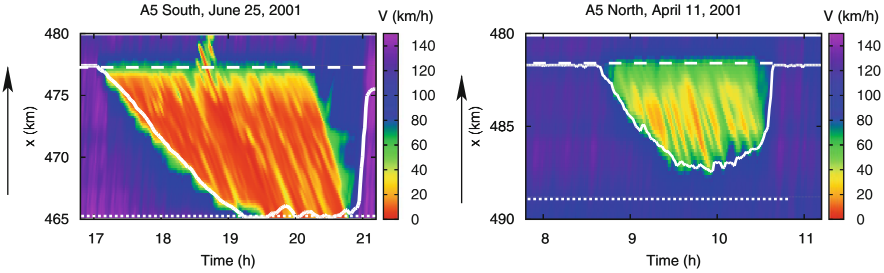


Fig. 7 Examples of congestion patterns on the German freeway A5 close to Frankfurt, for which data have been provided to the authors between kilometers 465 and 492. The figures analyze the propagation of the upstream front of a region of congested traffic (white solid line) according to (12), for empirical HCT (*left*) and OCT (*right*). In both plots, the driving direction is upwards (as indicated by the *arrows*). The upstream flow Q_{up} was determined from a detector cross section whose location is indicated by a *dotted white line*, while the bottleneck flow was determined from detectors of a nearby cross section (*dashed white line*). When determining the flows, the time delay caused by the finite propagation velocities $dQ_{\text{e}}(\rho)/d\rho$ from the detectors to the upstream front was taken care of. The congestion patterns were chosen such that there were no ramps at or between the two detector cross sections. Otherwise, the determination of Q_{up} and Q_{cong} would have been more complicated. The free and congested densities were calculated with a simple, triangular fundamental diagram. Therefore, $\rho_{\text{fr}}(Q) = Q/V_0$ and $\rho_{\text{cg}}(Q) = \rho_{\text{jam}}(1 - QT)$, where the following parameters were chosen: $V_0 = 120$ km/h, $\rho_{\text{jam}} = 100$ vehicle/km/lane, and $T = 2$ s

congested flow but also for the OCT pattern. Since the location of the congestion front is given by *integration* of (12) over time, oscillations of the input quantities of this equation are automatically averaged out.

The resulting congestion pattern depends on the stability properties of the vehicle density

$$\rho_{\text{cong}} = \rho_{\text{cg}}(Q_{\text{cong}}) = \rho_{\text{cg}}\big(Q_{\text{out}} - \Delta Q\big) \tag{13}$$

in the congested area, where the outflow Q_{out} from seriously congested traffic represents the effective freeway capacity under congested conditions and ΔQ the capacity taken away by the bottleneck. In view of this stability dependence, let us now discuss the meaning of the critical densities $\rho_{\text{c}k}$ or associated flows $Q_{\text{c}k} = Q_{\text{e}}(\rho_{\text{c}k})$, respectively, for the phase diagram.

If $\rho_{\text{c2}} < \rho_{\text{cong}} < \rho_{\text{c3}}$, we expect unstable, oscillatory traffic flow (OCT or SGW). For $\rho_{\text{c3}} \le \rho_{\text{cong}} < \rho_{\text{c4}}$, the congested flow is metastable, i.e. it depends on the perturbation amplitude: One may either have oscillatory patterns (for large enough perturbations) or homogeneous ones (for small perturbations). Moreover, for $\rho_{\text{cong}} \ge \rho_{\text{c4}}$ (given that the critical density ρ_{c4} is smaller than ρ_{jam}), we expect homogeneous, i.e. non-oscillatory traffic flows.

Expressing this in terms of flows rather than densities, one would expect the following: Oscillatory congestion patterns (OCT or SGW) should be possible for $Q_{\text{c2}} > Q_{\text{cong}} = Q_{\text{out}} - \Delta Q > Q_{\text{c4}}$, i.e. in the range

$$Q_{\text{out}} - Q_{\text{c2}} < \Delta Q < Q_{\text{out}} - Q_{\text{c4}}\,, \tag{14}$$

where we have considered $Q_{\text{c2}} \ge Q_{\text{out}}$.

The assumption that the densities between ρ_{out} with $Q_{\text{e}}(\rho_{\text{out}}) = Q_{\text{out}}$ and Q_{max} are *metastable*, as we assume here, has interesting implications: A linear *instability* would cause a single moving cluster to trigger further local clusters and, thereby, so-called "triggered stop-and-go waves" (TSG or SGW) [13]. A metastability, in contrast, can suppress the triggering of additional moving clusters, which allows the persistence of a *single* moving cluster, if the bottleneck strength ΔQ is small. As, for $Q_{\text{tot}} > Q_{\text{out}}$, the related flow conditions fall into the area of extended congested traffic, the spatial extension of such a cluster will grow. Therefore, one may use the term "widening moving cluster" (WMC).

Furthermore, according to our computer simulations, the capacity downstream of a widening moving cluster may eventually revert from Q_{out} to Q_{max}. This happens in the area, where "widening synchronized patterns" (WSP) can appear.[11] Therefore, rather than by (14), the bottleneck strengths characterizing OCT or SGW states are actually given by

$$Q_{\text{max}} - Q_{\text{c2}} < \Delta Q < Q_{\text{out}} - Q_{\text{c4}}\,, \tag{15}$$

where the lower boundary corresponds to the boundary of the WSP state, see (8). We point out that a capacity reversion despite congestion is a special feature of traffic models with $\rho_{\text{c2}} > \rho_{\text{max}}$.

Homogeneous congested traffic (the definition of which does not cover the homogeneous WSP state) is expected to be possible for $Q_{\text{cong}} = Q_{\text{out}} - \Delta Q < Q_{\text{c3}}$, i.e. (meta-) stable flows at high densities. This corresponds to

$$\Delta Q > Q_{\text{out}} - Q_{\text{c3}}\,. \tag{16}$$

The occurrence of *extended* congested traffic like HCT and OCT requires an additional condition: The total flow must exceed the freeway capacity Q_{out} during serious congestion,[12] i.e. we must have

$$Q_{\text{tot}} = Q_{\text{up}} + \Delta Q > Q_{\text{out}}\,. \tag{17}$$

Localized congestion patterns, in contrast, require $Q_{\text{tot}} \leq Q_{\text{out}}$ and can be triggered for $Q_{\text{tot}} > Q_{\text{c1}}$, which implies

$$Q_{\text{c1}} < Q_{\text{tot}} = Q_{\text{up}} + \Delta Q \leq Q_{\text{out}}\,. \tag{18}$$

[11] In this connection, it is interesting to remember Kerner's "dissolving general pattern" (DGP), which is predicted under similar flow conditions.

[12] One may also analyze the situation with the shock wave equation: Spatially expanding congested traffic results, if the speed of the downstream shock front of the congested area (which is usually zero) minus the speed of the upstream shock front (which is usually negative) gives a positive value.

We can distinguish at least two cases: On the one hand, if

$$Q_{\mathrm{c1}} < Q_{\mathrm{up}} < Q_{\max}\,, \tag{19}$$

the flow upstream of the congested area is *metastable,* which allows jams (and large enough perturbations) to propagate upstream. In this case, we speak of *moving localized clusters* (MLC). Their propagation speed $c_0 = -15 \pm 5\,\mathrm{km/h}$ is given by the slope of the jam line [17].

On the other hand, if

$$Q_{\mathrm{up}} \le Q_{\mathrm{c1}} \tag{20}$$

or $\rho_{\mathrm{fr}}(Q_{\mathrm{up}}) < \rho_{\mathrm{c1}}$, traffic flow upstream of the bottleneck is *stable*. Under such conditions, perturbations and, in particular, localized congestion patterns cannot propagate upstream, and they stay at the location of the bottleneck. In this case, one speaks of *pinned localized clusters* (PLC).[13]

We underline that the actual outflow $\tilde{Q}_{\mathrm{out}}$ from localized clusters corresponds, of course, to their inflow $Q_{\mathrm{up}} + \Delta Q$ (otherwise they would grow or shrink in space). Therefore, the actual outflow $\tilde{Q}_{\mathrm{out}}$ of localized congestion patterns can be smaller than Q_{out}, i.e. smaller than the outflow of *serious* congestion.

5 Combinations of On- and Off-Ramps

We see that the instability diagram implies a large variety of congestion patterns already in the simple simulation scenario of a homogeneous freeway with a single ramp. The possible congestion patterns are even richer in cases of complex freeway setups. All combinations of the previously discussed, "elementary" traffic patterns are possible. Furthermore, we expect particular patterns due to interactions among patterns through spillover effects. For illustration, let us focus here on the combination of an on-ramp with an off-ramp further upstream. This freeway design is illustrated in Fig. 8 and often built to reduce the magnitude of traffic breakdowns, since it is favorable when vehicles leave the freeway before new ones enter. Nevertheless, the on-ramp and the off-ramp bottleneck can get coupled, namely when congestion upstream of the on-ramp reaches the location of the off-ramp.

[13] Since pinned localized clusters rarely constitute a *maximum* perturbation, they can also occur at higher densities and flows, as long as $Q_{\mathrm{up}} < Q_{\mathrm{c2}}$. Therefore, MLC and PLC states can *coexist* in the range $Q_{\mathrm{c1}} < Q_{\mathrm{up}} < Q_{\mathrm{c2}}$. For most traffic models and bottleneck types, congestion patterns with $Q_{\mathrm{tot}} \approx Q_{\mathrm{c1}}$ do *not* exist, since *localized* congestion patterns do not correspond to *maximum* perturbations. The actual lower boundary $\tilde{Q}_{\mathrm{c1}}$ for the overall traffic volume Q_{tot} that generates congestion is somewhat higher than Q_{c1}, but usually lower than Q_{c2}. Considering the metastability of traffic flow in this range and the decay of the critical perturbation amplitude from ρ_{c1} to ρ_{c2} [10], this behavior is expected. However, for some models and parameters, one may even have $\tilde{Q}_{\mathrm{c1}} > Q_{\mathrm{out}}$. In such cases, PLC states would not be possible under *any* circumstances.

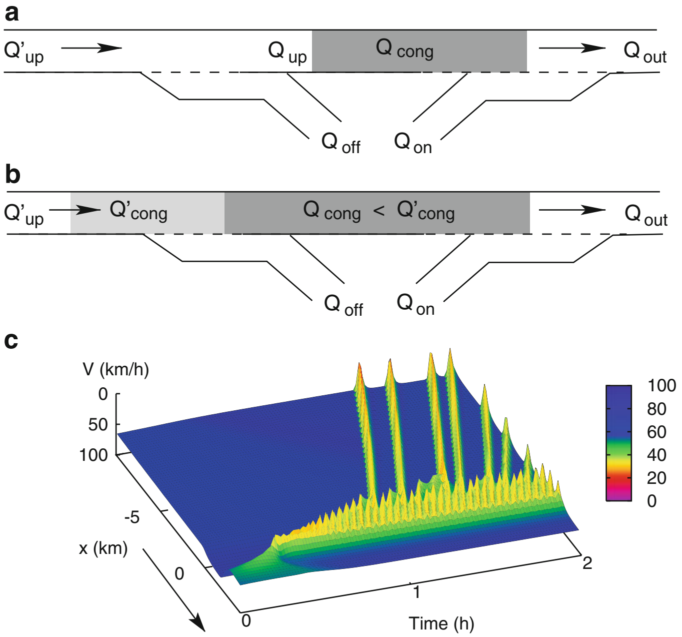


Fig. 8 Combination of an on-ramp bottleneck with an upstream off-ramp. (**a**) When the flow $Q_{\rm up} = Q'_{\rm up} - \Delta Q_{\rm off}$ upstream of the on-ramp exceeds $Q_{\rm cong}$, which is defined as the outflow $Q_{\rm out}$ from congested traffic minus the on-ramp flow $\Delta Q_{\rm on} = Q_{\rm on}/I_{\rm fr}$, congested traffic upstream of the on-ramp (*dark grey area*) is expected to grow. (**b**) As soon as the congested area extends up to the location of the off-ramp, the off-ramp bottleneck is activated. Its effective outflow $Q'_{\rm out}$ is given by the congested flow $Q_{\rm cong}$ upstream of the on-ramp, while congested flow $Q'_{\rm cong}$ upstream of the off-ramp is higher by the amount $\Delta Q_{\rm off} = Q_{\rm off}/I_{\rm fr}$ of the off-ramp flow. (**c**) Spatiotemporal velocity field resulting from a computer simulation with the gas-kinetic-based traffic model (GKT) [52], which allows to treat ramps easily. The *arrow* indicates the driving direction. One can clearly see pronounced stop-and-go waves emanating from an area of oscillating congested traffic

What would a bottleneck analysis analogous to the one in Sect. 4 predict for this setup? In order to discuss this, let us again denote the outflow capacity downstream of the on-ramp by $Q_{\rm out}$, its bottleneck strength equivalent to the on-ramp flow $Q_{\rm rmp} = Q_{\rm on}$ by $\Delta Q_{\rm on} = Q_{\rm rmp}/I_{\rm fr} \geq 0$, the upstream flow by $Q_{\rm up}$, and the average congested flow resulting immediately upstream of the on-ramp by $Q_{\rm cong}$. In contrast, we will denote the same quantities relating to the area of the off-ramp by primes ($'$), but we will introduce the abbreviation $-\Delta Q_{\rm off} = Q'_{\rm rmp}/I_{\rm fr} \leq 0$ for the effect of the off-ramp flow $Q'_{\rm rmp} \leq 0$.

According to Fig. 8, we observe the following dynamics: First, traffic breaks down at the strongest bottleneck, which is the on-ramp. If $Q_{\rm up} > Q_{\rm out} - \Delta Q_{\rm on}$, congested flow of size $Q_{\rm cong} = Q_{\rm out} - \Delta Q_{\rm on}$ expands, and eventually reaches the location of the off-ramp, see Fig. 8a. Afterwards, the freeway capacity downstream of the off-ramp suddenly drops from $Q'_{\rm out} = Q_{\rm out}$ to the congested flow

$$Q'_{\text{out}} = Q_{\text{cong}} = Q_{\text{out}} - \Delta Q_{\text{on}} \tag{21}$$

due to a spillover effect. This abrupt change in the bottleneck capacity restricts the capacity for the flow *upstream* of the off-ramp to

$$Q'_{\text{cong}} = Q_{\text{cong}} + \Delta Q_{\text{off}} \geq Q_{\text{cong}} \,. \tag{22}$$

This higher flow capacity implies either free flow or milder congestion upstream of the off-ramp. If Q'_{cong} is smaller than the previous outflow capacity Q_{out}, we have a bottleneck along the off-ramp, and its effective strength ΔQ is given by the *difference* of these values:

$$\Delta Q = Q_{\text{out}} - (Q_{\text{cong}} + \Delta Q_{\text{off}}) = \Delta Q_{\text{on}} - \Delta Q_{\text{off}} \,. \tag{23}$$

That is, the bottleneck strength is defined as the amount of outflow from congested traffic which cannot be served by the off-ramp and the downstream freeway flow. For $Q_{\text{cong}} + \Delta Q_{\text{off}} \geq Q_{\text{out}}$, no bottleneck occurs, which corresponds to a bottleneck strength $\Delta Q = 0$. This finally results in the expression

$$\Delta Q = \max(\Delta Q_{\text{on}} - \Delta Q_{\text{off}}, 0) \leq \Delta Q_{\text{on}} \tag{24}$$

[9]. Whenever $\Delta Q_{\text{off}} > \Delta Q_{\text{on}}$, there is no effective bottleneck upstream of the off-ramp, i.e. the off-ramp bottleneck is de-activated. For $\Delta Q = \Delta Q_{\text{on}} - \Delta Q_{\text{off}} > 0$, however, the resulting congested flow upstream of the off-ramp becomes

$$Q'_{\text{cong}} = Q'_{\text{out}} + \Delta Q_{\text{off}} = Q_{\text{out}} - \Delta Q_{\text{on}} + \Delta Q_{\text{off}} = Q_{\text{out}} - \Delta Q \,. \tag{25}$$

In conclusion, if congested traffic upstream of an on-ramp reaches an upstream off-ramp, the off-ramp becomes a bottleneck of strength ΔQ, which is given by the difference between the on-ramp and the off-ramp flows (or zero, if this difference would be negative).

Since $\Delta Q \leq \Delta Q_{\text{on}}$ according to (24) and $Q'_{\text{cong}} \geq Q_{\text{cong}}$ according to (22), the congestion upstream of the off-ramp tends to be "milder" than the congestion upstream of the on-ramp. The resulting traffic pattern is often characterized by homogeneous or oscillating congested traffic between the off-ramp and the on-ramp, and by stop-and-go waves upstream of the off-ramp, i.e. it has typically the appearance of a "pinch effect" [26] (see Fig. 8c). For this reason, Kerner also calls the "pinch effect" a "general pattern" [20].[14]

[14] Oscillatory congestion patterns upstream of off-ramps are further promoted by a behavioral feedback, since drivers may decide to leave the freeway in response to downstream traffic congestion.

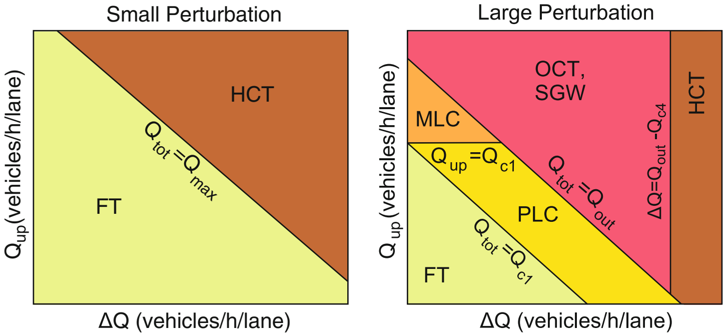


Fig. 9 Schematic phase diagram for traffic flow *without* an extended linearly unstable density regime ($\rho_{c2} = \rho_{c3}$), when the traffic flow at capacity (at the density $\rho_{\max}$ corresponding to the maximum flow) is assumed to be metastable ($\rho_{c1} < \rho_{\max} < \rho_{c4}$)

6 Other Phase Diagrams and Universality Classes of Models

The phase diagram approach can also be used for a classification of traffic models. By today, there are hundreds of traffic models, and many models have a similar goodness of fit, when parameters are calibrated to empirical data [4,5,11,29,41,42]. It is, therefore, difficult, if not impossible, to determine "the best" traffic model. However, one can classify models according to topologically equivalent phase diagrams. Usually, there would be several models in the same universality class, producing qualitatively the same set of traffic patterns under roughly similar conditions. Among the models belonging to the same universality class, one could basically select *any* model. According to the above, the differences in the goodness of fit are usually not dramatic. Models with many model parameters may even suffer from insignificant parameters or parameters, which are hard to calibrate, at the cost of predictive power. Therefore, it is most reasonable to choose the *simplest* representative of a universality class which, however, should fulfil minimum requirements regarding theoretical consistency.

Before we enter the comparison with empirical data, let us discuss a number of phase diagrams expected for certain kinds of traffic models. Particular specifications of the optimal velocity model, for example, are linearly unstable for one density $\rho_{c2} = \rho_{c3}$ only, but show unstable behavior in an extended density regime for sufficiently large perturbations (i.e. extended metastable regimes) [10]. The schematic phase diagram expected in this case is shown in Fig. 9. Some other traffic models have linearly unstable and metastable regimes, but do not show a restabilisation at very high densities, i.e. $\rho_{c4} = \rho_{\text{jam}}$ (see Fig. 10), and sometimes one even has $\rho_{c3} = \rho_{\text{jam}}$ [28,47] (see Fig. 11). In the latter case, homogeneous congested traffic does not exist. In models such as the IDM, the restabilisation depends on the chosen parameter values [55], see also Appendix 2.

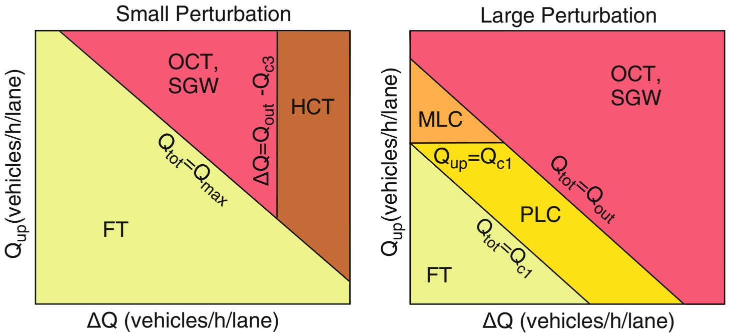


Fig. 10 Schematic phase diagram for the case of an incomplete destabilization at high densities, $\rho_{c3} < \rho_{c4} = \rho_{\text{jam}}$, when traffic at capacity is assumed to be linearly unstable ($\rho_{c1} < \rho_{c2} < \rho_{\max}$)

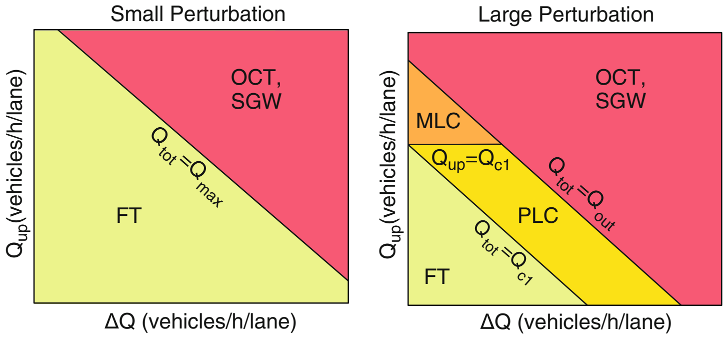


Fig. 11 Schematic phase diagram for traffic flow exhibiting both, metastable and linearly unstable density regimes, with unstable flow at capacity ($\rho_{c1} < \rho_{c2} < \rho_{\max}$), but no restabilisation for very high densities ($\rho_{c3} = \rho_{c4} = \rho_{\text{jam}}$)

In most of the currently studied traffic models, one has either *both*, linearly unstable *and* metastable density ranges, or unconditionally stable traffic. In principle, however, models with linearly unstable but no metastable regimes are conceivable. For example, they may be established by taking a conventional model and introducing a dependence on the square of the velocity gradient (in macroscopic models) or the velocity difference (in microscopic models).

A linearly unstable model without metastable ranges would correspond to $\rho_{c2} = \rho_{c1}$ and $\rho_{c4} = \rho_{c3}$. For such models, we do not expect any multi-stability (see Fig. 12), and localized congested traffic would only be possible under special conditions [36, 44] (e.g. on freeway sections between off- and on-ramps). If, in addition, there is no restabilisation (i.e. $\rho_{c3} = \rho_{\text{jam}}$), only free traffic and oscillating congested traffic should exist. This seems to reflect the situation for the classical Nagel–Schreckenberg model [39], although the situation is somewhat unclear, since this model is stochastic and an exact distinction between free and congested states is difficult in this model.

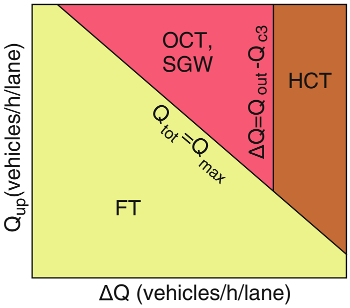


Fig. 12 Schematic phase diagram, if there are only stable and linearly unstable, but no metastable density regimes ($\rho_{c1} = \rho_{c2}$, $\rho_{c3} = \rho_{c4}$). Furthermore, traffic at capacity is assumed to be unstable ($\rho_{c1} < \rho_{\max} < \rho_{c4}$)

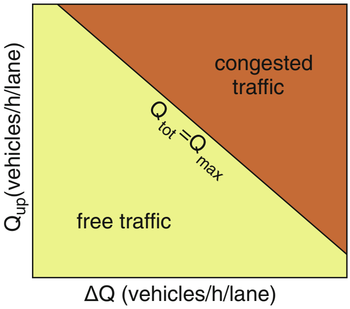


Fig. 13 Schematic phase diagram, if traffic is unconditionally stable ($\rho_{c1} = \rho_{c2} = \rho_{c3} = \rho_{c4}$). The most prominent example is the Lighthill–Whitham model [37], but many other models (including the gas-kinetic-based traffic model (GKT) [52] and the IDM) can be parameterized to reproduce this case

Finally, we would like to discuss the fluid-dynamic model by Lighthill and Whitham [37], which does not display *any* instabilities [8] and, consequently, has only homogeneous patterns, namely free traffic for $Q_{\text{tot}} \le Q_{\max}$ and (homogeneous) extended congested traffic for $Q_{\text{tot}} > Q_{\max}$ (which corresponds to a vehicle platoon behind the bottleneck). This is illustrated in Fig. 13. The two phases can also be distinguished locally, if temporal correlations are considered: While perturbations in free traffic travel in forward direction, in the congested regime they travel backward.

We underline again that, by changing model parameters (corresponding to different driving styles), the resulting instability and phase diagrams of many traffic models change as well. For example, the IDM can be parameterized to generate most of the stability diagrams discussed in this contribution. Since different parameter values correspond to different driving styles or prevailing velocities, this may explain differences between empirical observations in different countries. For example, oscillating congested traffic seems to occur less frequent in the United States [6, 58].

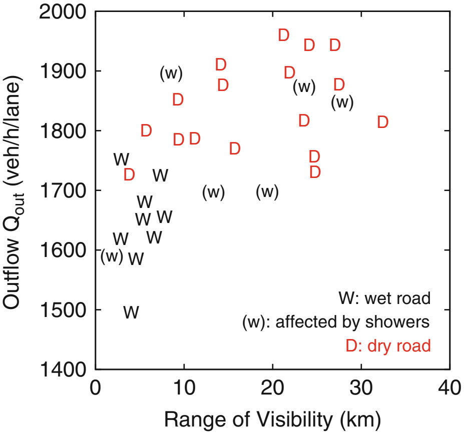


Fig. 14 The outflows Q_{out} of congestion patterns correlate significantly with the weather-dependent range of visibility (meteorological optical range $3/\sigma$). This was determined by measuring the extinction coefficient σ, using multiple laser reflections. "W" denotes traffic breakdowns when the road surface was wet, while "D" stands for a dry road surface. Note that there are three cases of congestion marked with a lower-case "w", which were related with a short and light shower only, so that the outflow values stayed comparatively high. For details see [45]

Finally, note that somewhat different phase diagrams result for models that are characterized by a complex vehicle dynamics and no existence of a fundamental diagram [21, 28]. Nevertheless, similarities can be discovered (see Sect. 4).

7 Empirical Phase Diagram

The remaining challenge in this paper is to find the universality class that fits the stylized facts of traffic dynamics well. Here, we will primarily demand that it fits the empirical phase diagram, i.e. reproduces all elementary congestion patterns observed, and not more. We have evaluated empirical data from the German freeway A5 close to Frankfurt. Due to the weather-dependence of the outflows Q_{out} (see Fig. 14), it is important to scale all flows by the respective measurements of Q_{out}. This naturally collapses the area of localized congested traffic states to a line. As Fig. 15 shows, the phase diagram after scaling the flows is very well compatible with the theoretical phase diagrams of Figs. 5 and 6. Since the determination of the empirical phase diagram did not focus on the detection of "widening synchronized patterns", it does not allow us to clearly distinguish between the two phase diagrams, i.e. to decide whether $\rho_{c2} > \rho_{\max}$ or $\rho_{c2} < \rho_{\max}$. However, the empirical WSP displayed in Fig. 1 suggests that Fig. 6 corresponding to $\rho_{c2} > \rho_{\max}$ would be

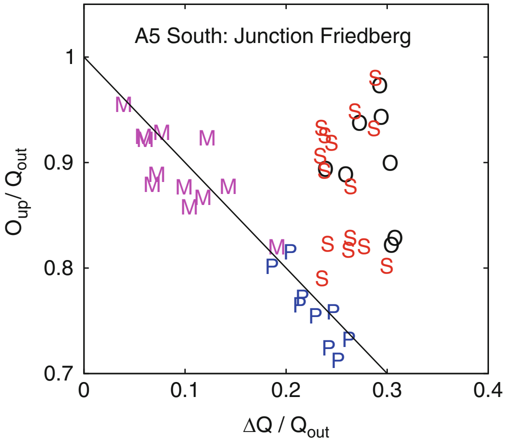


Fig. 15 Empirical phase diagram, where the flows have been scaled by the respective outflows Q_{out} (after [44]). The data represent the congested traffic states observed on the German freeway A5 at Junction Friedberg in direction South (M = moving localized cluster, S = stop-and-go waves, O = oscillating congested traffic, P = pinned localized cluster). It can be clearly seen that the non-extended traffic states are scattered around the line $Q_{\text{tot}}/Q_{\text{out}} = (Q_{\text{up}} + \Delta Q)/Q_{\text{out}} = 1$, as expected, while the extended traffic states are above this line. Moreover, pinned localized clusters, moving localized clusters, and stop-and-go waves/oscillating congested traffic are well separated from each other. Homogeneous congested traffic, but not other traffic states were observed for $\Delta Q/Q_{\text{out}} \gtrsim 0.5$ (see [44])

the right choice. Another piece of evidence for this is the metastability of vehicle platoons forming behind overtaking trucks (see [12]).[15]

7.1 Reply to Criticisms of Phase Diagrams for Traffic Models with a Fundamental Diagram

In the following, we will face the criticism of the phase diagram approach by Kerner [16, 20]:

1. Models containing a fundamental diagram could not explain the wide scattering of flow-density data observed for “synchronized” congested traffic flow. This is definitely wrong, as a wide scattering is excellently reproduced by considering the wide distribution of vehicle gaps, partially due to different vehicle classes such as cars and trucks [40, 49]. Note that, for a good reproduction of empirical

[15] The existence of “widening moving clusters”, see Sect. 4.2 and Fig. 1a, supports this view as well.

measurements, it is important to apply the same measurement procedure to empirical and simulated data, in particular the data aggregation over a finite time period.

2. As the ramp flow or the overall traffic volume Q_{tot} is increasing, the phase diagram approach would predict the transitions free traffic → moving or pinned localized cluster → stop-and-go traffic/oscillating congested traffic → homogeneous congested traffic. However, this would be wrong because (i) homogeneous congested traffic would not exist [18, 19], and (ii) according to the "pinch effect", wide moving jams (i.e. moving localized clusters) should occur *after* the occurrence of "synchronized flow" (i.e. extended congested traffic) [17].

   We reply to (i) that it would be easy to build a car-following model with a fundamental diagram that produces no HCT states,[16] but according to empirical data, homogeneous congested traffic *does* exist (see Fig. 1f), but it occurs very rarely and only for extremely large bottleneck strengths exceeding $\Delta Q \approx 0.5\, Q_{\text{out}}$ [44]. As freeways are dimensioned such that bottlenecks of this size are avoided, HCT occurs primarily when freeway lanes are closed after a serious accident. In other words, when excluding cases of accidents from the data set, HCT states will normally not be found.

   Moreover, addressing point (ii), Kerner is wrong in claiming that our theoretical phase diagram would necessarily *require* moving localized clusters to occur *before* the transition to stop-and-go waves or oscillating congested traffic. This misunderstanding might have occurred by ignoring the dependence of the resulting traffic state on the perturbation size. The OCT pattern of Fig. 3c clearly shows that a *direct* transition from free traffic flow to oscillating congested traffic is possible in cases of small perturbations. The same applies to a fast increase in the traffic volume $Q_{\text{tot}}(t)$, which is typical during rush hours.

3. The variability of the empirical outflow Q_{out} would not be realistically accounted for by traffic models with a fundamental diagram. This variability, however, does not require an explanation based on complex vehicle dynamics. To a large extent, it can be understood by variations in the weather conditions (see Fig. 14) and in the flow conditions on the freeway lanes in the neighborhood of ramps [44], which is particularly affected by a largely varying truck fraction [49].

In summary, the phase diagram approach for traffic models with a fundamental diagram has been criticized with invalid arguments.

7.2 *On the Validity of Traffic Models*

In the past decades, researchers have proposed a large number of traffic models and it seems that there often exist several different explanations for the same observation(s) [55]. As a consequence, it is conceivable that there are models

[16] An example would be the IDM with the parameter choice $s_1 = 0$.

which are macroscopically correct (in terms of reproducing the observed congestion patterns discussed above), but microscopically wrong. In order to judge the validity of competing traffic models, we consider it necessary to compare models in a *quantitative* way, based on empirical data. This should include

- A definition of suitable performance measures (such as the deviation between simulated and measured travel times or velocity profiles).
- The implementation and parameter calibration of the competing models with *typical* empirical data sets.
- The comparison of the performance of the competing models for different *test* data sets of representative traffic situations.

Based on data sets of car-following experiments, such analyses have, for example, been performed with a number of follow-the-leader models [4,5,11,29,41,42], with good results in particular for the intelligent driver model (IDM) [4, 5, 29]. If there is no statistically significant difference in the performance of two models (based on an analysis of variance), preference should be given to the simpler one, according to Einstein's principle that a model should be always as simple as possible, but not simpler.

We would like to point out that over-fitting of a model must be avoided. This may easily happen for models with many parameters. Fitting such models to data will, of course, tend to yield smaller errors than fitting models with a few parameters only. Therefore, one needs to make a significance analysis of parameters that *adjusts* for the number of parameters, as it is commonly done in statistical analyses. Reproducing a certain *calibration* data set well does not necessarily mean that an independent *test* data set will be well reproduced. While the *descriptive* capability of models with many parameters is often high, models with fewer parameters may have a higher *predictive* capability, as their parameters are often easier to calibrate.

This point is particularly important, since it is known that traffic flows fluctuate considerably, especially in the congested regime. So, one may pose the question whether these fluctuations are meaningful dynamical features of traffic flows or just noise. To some extent, this depends on the question to be addressed by the model, i.e. how fine-grained predictions the model shall be able to make. There are certainly systematic sources of fluctuations, such as lane-changes, in particular by vehicles entering or leaving the freeway via ramps, different types of vehicles, and different driver behaviors [29, 42]. Such issues would most naturally be addressed by multi-lane models considering lane changes and heterogeneous driver-vehicle units [30]. Details like this may, in particular, influence the outflow Q_{out} of congested traffic flow (see Fig. 17 in [44]). When trying to understand the empirically observed variability of the outflow, however, one also needs to take the variability of the *weather conditions* and the visibility into account (see Fig. 14). In order to show that car-following models with a fundamental diagram are inferior to other traffic models in terms of reproducing microscopic features of traffic flows (even when multi-lane multi-class features are considered), one would have to show with standard statistical procedures that these other models can explain a larger share of the empirically observed variance, and that the difference in the explanation power is significant, even when the number of model parameters is considered. To the knowledge of the authors, however, such a statistical analysis has not been presented so far.

8 Summary, Conclusions, and Outlook

After a careful discussion of the term "traffic phase", we have extended the phase diagram concept to traffic models with a fundamental diagram that are not only capable of reproducing congestion patterns such as localized clusters, stop-and-go waves, oscillatory congested traffic, or homogeneous congested traffic, but also "widening synchronized patterns" (WSP) and "widening moving clusters" (WMC). The discovery of these states for the case, where the maximum traffic flow lies in the metastable density regime, was quite unexpected. It offers an alternative and—from our point of view—simpler interpretation of some of Kerner's empirical findings. A particular advantage of starting from models with a fundamental diagram is the possibility of *analytically* deriving the schematic phase diagram of traffic states from the instability diagram, which makes the approach *predictive*.

Furthermore, we have discussed how the phase diagram approach can be used to classify models into universality classes. Models within one universality class are essentially equivalent, and one may choose any, preferably the simplest representative satisfying minimum requirements regarding theoretical consistency. The universality class should be chosen in agreement with empirical data. These were well represented by the schematic phase diagram in Fig. 6. Furthermore, we have demonstrated that one needs to implement the full details of a freeway design, in particularly all on- and off-ramps, as these details matter for the resulting congestion patterns. Multi-ramp designs lead to congestion patterns composed of several elementary congestion patterns, but spillover effects must be considered. In this way, a simple explanation of the "pinch effect" [17] and the so-called "general pattern" [20] results. We have also replied to misunderstandings of the phase diagram concept.

In conclusion, the phase diagram approach is a simple and natural approach, which can explain empirical findings well, in particular the dependence of traffic patterns on the flow conditions. Note that the phase diagram approach is a metatheory rather than a model. It can be theoretically derived from the instability diagram of traffic flows and the self-organized outflow from seriously congested traffic. This is not a triviality and, apart from this, the phase diagram approach is more powerful than the instability diagram itself: It does not only allow predictions regarding the possible appearance of traffic patterns and possible transitions between them. It also allows to predict whether it is an extended or localized traffic pattern, or whether a localized cluster moves or not. Furthermore, it facilitates the prediction of the spreading dynamics of congestion in space, as reflected by (7) and (12). This additionally requires formula (11), which determines how the bottleneck strength ΔQ determines the effective flow capacity Q_{cong} of the upstream freeway section.

Acknowledgements DH and MT are grateful for the inspiring discussions with the participants of the Workshop on "Multiscale Problems and Models in Traffic Flow" organized by Michel Rascle and Christian Schmeiser at the Wolfgang Pauli Institute in Vienna from May 5–9, 2008, with partial support by the CNRS. Furthermore, the authors would like to thank for financial support by the Volkswagen AG within the BMBF research initiative INVENT and the *Hessisches Landesamt für Straßen und Verkehrswesen* for providing the freeway data.

Appendix 1 Modeling of Source and Sink Terms (In- and Outflows)

In this appendix, we will focus on the case of a freeway section with a single bottleneck such as an isolated on-ramp. Scenarios with several bottlenecks are discussed in Sect. 5.

In order to derive the appropriate form of source and sink terms due to on- or off-ramps, we start from the continuity equation, which reflects the conservation of the number of vehicles. If $\rho_*(x,t)$ represents the one-dimensional density of vehicles at time t and a location x along the freeway, and if $Q_*(x,t)$ represents the vehicle flow measured at a cross section of the freeway, the continuity equation can be written as follows:

$$\frac{\partial \rho_*(x,t)}{\partial t} + \frac{\partial Q_*(x,t)}{\partial x} = 0\,. \tag{26}$$

Now, assume that $I(x)$ is the number of freeway lanes at location x. We are interested in the density $\rho(x,t) = \rho_*(x,t)/I(x)$ and traffic flow $Q(x,t) = Q_*(x,t)/I(x)$ per freeway lane. Inserting this into the continuity equation (26) and carrying out partial differentiation, applying the product rule of Calculus, we get

$$\begin{aligned}\frac{\partial}{\partial t}\big[I(x)\rho(x,t)\big] &= I(x)\frac{\partial \rho(x,t)}{\partial t} \\ &= -\frac{\partial}{\partial x}\Big[I(x)Q(x,t)\Big] \\ &= -Q(x,t)\frac{dI(x)}{dx} - I(x)\frac{\partial Q(x,t)}{\partial x}\,. \end{aligned} \tag{27}$$

Rearranging the different terms, we find

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(x,t)}{\partial x} = -\frac{Q(x,t)}{I(x)}\frac{dI(x)}{dx}\,. \tag{28}$$

The first term of this equation looks exactly like the continuity equation for the density $\rho_*(x,t)$ over the whole cross section at x. The term on the right-hand side of the equality sign describes an increase of the density $\rho(x,t)$ per lane, whenever the number of freeway lanes is reduced ($\partial I(x)/\partial x < 0$) and all vehicles have to squeeze into the remaining lanes. In contrast, the density per lane $\rho(x,t)$ goes down, if the width of the road increases ($\partial I(x)/\partial x > 0$).

It is natural to treat on- and off-ramps in a similar way by the continuity equation

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(x,t)}{\partial x} = \nu_+(x,t) - \nu_-(x,t) \tag{29}$$

with source terms $\nu_+(x,t)$ and sink terms $-\nu_-(x,t)$. For example, if a one-lane on-ramp flow $Q_{\text{on}}(t)$ is entering the freeway uniformly over an effectively used ramp length of L_{eff}, we have $dI(x)/dx = 1/L_{\text{eff}}$, which together with (27) and (29) implies

$$\nu_+(x,t) = \begin{cases} \dfrac{Q_{\text{on}}(t)}{I_{\text{fr}} L_{\text{eff}}} & \text{for } x_{\text{rmp}} - \frac{L_{\text{eff}}}{2} < x < x_{\text{rmp}} + \frac{L_{\text{eff}}}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

$I_{\text{fr}} = I(x_{\text{rmp}} \pm L_{\text{eff}}/2)$ denotes the number of freeway lanes *upstream* and *downstream* of the ramp, which is assumed to be the same, here. The sink term due to off-ramp flows $Q_{\text{off}}(t) \geq 0$ has the form

$$\nu_-(x,t) = \begin{cases} \dfrac{Q_{\text{off}}(t)}{I_{\text{fr}} L_{\text{eff}}} & \text{for } x_{\text{rmp}} - \frac{L_{\text{eff}}}{2} < x < x_{\text{rmp}} + \frac{L_{\text{eff}}}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

Appendix 2 Parameter Dependence of the Instability Thresholds in the Intelligent Driver Model

The acceleration function $a_{\text{IDM}}(s, v, \Delta v)$ of the intelligent driver model (IDM) [53] depends on the gap s to the leading vehicle, the velocity v, and the velocity difference Δv (positive, when approaching). It is given by

$$a_{\text{IDM}}(s, v, \Delta v) = a \left[1 - \left(\frac{v}{v_0} \right)^4 - \left(\frac{s^*(v, \Delta v)}{s} \right)^2 \right], \tag{32}$$

where

$$s^*(v, \Delta v) = s_0 + s_1 \sqrt{\frac{v}{v_0}} + T v + \frac{v\, \Delta v}{2\sqrt{ab}}. \tag{33}$$

For identical driver-vehicle units, there exists a one-parameter class of homogeneous and stationary solutions defining the "microscopic" fundamental diagram $v_e(s)$ *via* $a_{\text{IDM}}(s, v_e(s), 0) = 0$. From a standard linear analysis around this solutions it follows that the IDM is linearly stable if the condition

$$\frac{\partial a_{\rm IDM}}{\partial s} \le \frac{\partial a_{\rm IDM}}{\partial v}\left(\frac{\partial a_{\rm IDM}}{\partial \Delta v} + \frac{1}{2}\frac{\partial a_{\rm IDM}}{\partial v}\right) \tag{34}$$

is fulfilled. With the micro-macro relation

$$s = \frac{1}{\rho} - l_{\rm veh}, \quad \text{where } l_{\rm veh} = 6\text{ m}, \tag{35}$$

this defines the stability boundaries ρ_{c2} and ρ_{c3} as a function of the model parameters v_0 (desired velocity), T (desired time headway), a (desired acceleration), b (desired deceleration), s_0 (minimum gap), and s_1 (gap parameter; if nonzero, the fundamental diagram has an inflection point). The overall stability can be controlled most effectively by the acceleration a. Setting the other parameters to the values used in Fig. 4 [$v_0 = 128$ km/h, $T = 1$ s, $s_0 = 2$ m, $s_1 = 10$ m, and $b = 1.3$ m/s^2, we obtain

- Unconditional linear stability for $a \geq 1.68$ m/s^2.
- Linear instability in the density range $\rho_{c2} \le \rho \le \rho_{c3}$ for 0.95 m/s$^2 \le a \le 1.68$ m/s^2, where $\rho_{c2} > \rho_{\max}$ and $\rho_{c3} < \rho_{\rm jam}$. In this situation, corresponding to Fig. 4c, d, the instability range lies completely on the "congested" side of the fundamental diagram.
- Finally, for $a \le 0.95$ m/s^2, the linear instability also extends to the "free branch" of the fundamental diagram ($\rho_{c2} < \rho_{\max}$), corresponding to Fig. 4a, b.

The upper instability threshold ρ_{c3} can be controlled nearly independently from the lower instability threshold ρ_{c2} by the gap parameters s_0 and s_1. Generally, ρ_{c3} increases with decreasing values of s_1. In particular, if $s_1 = 0$, one obtains the analytical result $\rho_{c3} = (l_{\rm veh} + s_0)^{-1}$ for any $a < s_0/T^2$, and unconditional linear stability for $a > s_0/T^2$. As can be seen from the last expression, the instability generally becomes more pronounced for decreasing values of the time headway parameter T, which is plausible.

The additional influence of the parameter b according to computer simulations is plausible as well: With decreasing values of b, the sensitivity with respect to velocity differences increases, and the instability tends to decrease. Further simulations suggest that the IDM has metastable density areas only when linearly unstable densities exist. Metastability at densities above the linear instability range additionally requires $s_1 > 0$.

References

1. P.F. Arndt, Phys. Rev. Lett. **84**, 814 (2000)
2. M. Bando, K. Hasebe, K. Nakanishi, A. Nakayama, Phys. Rev. E **58**, 5429 (1998)
3. R. Barlovic, T. Huisinga, A. Schadschneider, M. Schreckenberg, Phys. Rev. E **66**(4), 046113 (2002)
4. E. Brockfeld, R.D. Kühne, A. Skabardonis, P. Wagner, Transp. Res. Rec. **1852**, 124 (2003)

5. E. Brockfeld, R.D. Kühne, P. Wagner, Transp. Res. Rec. **1876**, 62 (2004)
6. M.J. Cassidy, R.L. Bertini, Transp. Res. B Methodol. **33**, 25 (1999)
7. M. Evans, Y. Kafri, H. Koduvely, D. Mukamel, Phys. Rev. Lett. **80**, 425 (1998)
8. D. Helbing, Rev. Mod. Phys. **73**, 1067 (2001)
9. D. Helbing, J. Phys. A **36**(46), L593 (2003)
10. D. Helbing, M. Moussaid, Eur. Phys. J. B **69**, 571–581 (2009)
11. D. Helbing, B. Tilch, Phys. Rev. E **58**, 133 (1998)
12. D. Helbing, B. Tilch, Eur. Phys. J. B **68**, 577–586 (2009)
13. D. Helbing, A. Hennecke, M. Treiber, Phys. Rev. Lett. **82**, 4360 (1999)
14. D. Helbing, A. Hennecke, V. Shvetsov, M. Treiber, Math. Comput. Model. **35**, 517 (2002)
15. R. Jiang, Q. Wu, B. Wang, Phys. Rev. E **66**(3), 36104 (2002)
16. B.S. Kerner, in *Proceedings of the Third International Symposium on Highway Capacity*, vol. 2, ed. by R. Rysgaard (Road Directorate, Denmark, 1998), pp. 621–641
17. B. Kerner, Phys. Rev. Lett. **81**(17), 3797 (1998)
18. B.S. Kerner, Transp. Res. Rec. **1710**, 136 (2000)
19. B. Kerner, Phys. Rev. E **65**, 046138 (2002)
20. B.S. Kerner, *The Physics of Traffic* (Springer, Heidelberg, 2004)
21. B.S. Kerner, Phys. A **399**, 379 (2004)
22. B. Kerner, S. Klenov, J. Phys. A **35**, L31 (2002)
23. B. Kerner, S. Klenov, J. Phys. A Math. Gen. **35**, L31 (2002)
24. B. Kerner, P. Konhäuser, Phys. Rev. E **48**(4), 2335 (1993)
25. B. Kerner, H. Rehborn, Phys. Rev. E **53**, R1297 (1996)
26. B. Kerner, H. Rehborn, Phys. Rev. E **53**, R4275 (1996)
27. B. Kerner, H. Rehborn, Phys. Rev. Lett. **79**, 4030 (1997)
28. B. Kerner, S. Klenov, D. Wolf, J. Phys. A **35**(47), 9971 (2002)
29. A. Kesting, M. Treiber, Transp. Res. Rec. **2088**, 148 (2008)
30. A. Kesting, M. Treiber, D. Helbing, Transp. Res. Rec. **1999**, 86 (2007)
31. A. Kesting, M. Treiber, M. Schönhof, D. Helbing, Transp. Res. C Emerg. Tech. **16**(6), 668 (2008)
32. J. Krug, Phys. Rev. Lett. **67**, 1882 (1991)
33. R. Kubo, Rep. Prog. Phys. **29**, 255 (1966)
34. H. Lee, H. Lee, D. Kim, Phys. Rev. Lett. **81**, 1130 (1998)
35. H.Y. Lee, H.W. Lee, D. Kim, Phys. Rev. E **59**, 5101 (1999)
36. H. Lee, H. Lee, D. Kim, Phys. Rev. E **62**, 4737 (2000)
37. M. Lighthill, G. Whitham, Proc. R. Soc. Lond. A **229**, 317 (1955)
38. S. Lübeck, M. Schreckenberg, K.D. Usadel, Phys. Rev. E **57**, 1171 (1998)
39. K. Nagel, M. Schreckenberg, J. Phys. I Fr. **2**, 2221 (1992)
40. K. Nishinari, M. Treiber, D. Helbing, Phys. Rev. E **68**, 067101 (2003)
41. S. Ossen, S.P. Hoogendoorn, Transp. Res. Rec. **1934**, 13 (2005)
42. S. Ossen, S.P. Hoogendoorn, B.G. Gorte, Transp. Res. Rec. **1965**, 121 (2007)
43. V. Popkov, L. Santen, A. Schadschneider, G.M. Schütz, J. Phys. A Math. Gen. **34**, L45 (2001)
44. M. Schönhof, D. Helbing, Transp. Sci. **41**, 1 (2007)
45. M. Schönhof, Ph.D. thesis, Technische Universität Dresden (Unpublished)
46. G. Schütz, E. Domany, J. Stat. Phys. **72**, 277 (1993)
47. F. Siebel, W. Mauser, Phys. Rev. E **73**(6), 66108 (2006)
48. Y. Sugiyama, M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S. Tadaki, S. Yukawa, New J. Phys. **10**, 033001 (2008)
49. M. Treiber, D. Helbing, J. Phys. A **32**(1), L17 (1999)
50. M. Treiber, D. Helbing, Coop. Transp. Dyn. **1**, 3.1 (2002). Internet Journal, see http://www.TrafficForum.org/journal
51. M. Treiber, D. Helbing, Eur. Phys. J. B **68**, 607–618 (2009)
52. M. Treiber, A. Hennecke, D. Helbing, Phys. Rev. E **59**, 239 (1999)
53. M. Treiber, A. Hennecke, D. Helbing, Phys. Rev. E **62**, 1805 (2000)
54. M. Treiber, A. Kesting, D. Helbing, Phys. A **360**, 71 (2006)

55. M. Treiber, A. Kesting, D. Helbing, Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts. Transport. Res. B Methodological **44**(8–9), 983–1000 (2010)
56. J. Treiterer, J. Myers, in *Proceedings of the 6th International Symposium on Transportation and Traffic Theory*, ed. by D. Buckley (Elsevier, New York, 1974), p. 13
57. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley-Interscience, New York, 1974)
58. B. Zielke, R. Bertini, M. Treiber, *Empirical Measurement of Freeway Oscillation Characteristics: An International Comparison*. Transportation Research Board Annual Meeting, Paper #08-0300, 2008

Self-Organized Network Flows*

Dirk Helbing, Jan Siegmeier, and Stefan Lämmer

Abstract A model for traffic flow in street networks or material flows in supply networks is presented, that takes into account the conservation of cars or materials and other significant features of traffic flows such as jam formation, spillovers, and load-dependent transportation times. Furthermore, conflicts or coordination problems of intersecting or merging flows are considered as well. Making assumptions regarding the permeability of the intersection as a function of the conflicting flows and the queue lengths, we find self-organized oscillations in the flows similar to the operation of traffic lights.

1 Introduction

Material flows are found in many places of the world. This concerns, for example, traffic flows in urban areas or flows of commodities in logistic systems. There is also some similarity with material flows in production or biological systems, from cells over bodies upto ecological food chains. Many of these material flows are not of diffusive nature or going on in continuous space. They are often directed and organized in networks. In comparison with data flows in information networks,

*First published in: Networks and Heterogeneous Media **2**(2), 193Ú–210 (2007), ©aimSciences.org, reproduction with kind permission.

D. Helbing (✉)
ETH Zurich, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland
e-mail: dhelbing@ethz.ch

J. Siegmeier · S. Lämmer
Institute for Transport and Economics, Dresden University of Technology, Andreas-Schubert-Str. 23, 01062 Dresden, Germany
e-mail: siegmeier@vwi.tu-dresden.de; stefan.laemmer@tu-dresden.de

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_6,
© Springer-Verlag Berlin Heidelberg 2013

however, there are conservation laws, which can be used to set up equations for material flows in networks. It turns out, however, that this is not a trivial task. While there is already a controversial discussion about the correct equations representing traffic flows along road sections [6, 11, 25, 31], their combination in often complex and irregular networks poses further challenges. In particular, there have been several publications on the treatment of the boundary conditions at nodes (connections) of several network links (i.e. road sections) [4, 7, 9, 10, 16, 21–24, 26]. In particular, the modelling of merging and intersecting flows is not unique, as there are many possible forms of organization, including the use of traffic lights. Then, however, the question comes up how these traffic lights should be operated, coordinated, and optimized. In order to address these questions, in Sect. 2 we formulate a simple model for network flows, which contains the main ingredients of material or traffic flows. Section 3 will then discuss the treatment of diverges, merges, and intersections. Equations for the interaction-dependent permeability at merging zones and intersections will be formulated in Sect. 4. We will see that, under certain conditions, they lead to spontaneous oscillations, which have features similar to the operation of traffic lights. Finally, Sect. 5 summarizes and concludes this paper.

2 Flows in Networks

The following section will start with a summary of the equations derived for traffic flows in networks in a previous paper. These equations are based on the following assumptions:

- The road network can be decomposed into road sections of homogeneous capacity (links) and nodes describing their connections.
- The traffic dynamics along the links is sufficiently well described by the Lighthill–Whitham model, i.e. the continuity equation for vehicle conservation and a flow-density relationship ("fundamental diagram"). This assumes adiabatic speed adjustments, i.e. that acceleration and deceleration times can be neglected.
- The parameters of vehicles such as the maximum speed V_i^0 and the safe time headway T are assumed to be identical in the same road section, and who enters a road section first exits first (FIFO principle). That is, overtaking is assumed to be negligible.
- The fundamental diagram can be well approximated by a triangular shape, with an increasing slope V_i^0 at low densities and a decreasing slope c in the congested regime. This implies two constant characteristic speeds: While V_i^0 corresponds to the free speed or speed limit on road section i,

$$-c = -\frac{1}{\rho_{\max} T} \tag{1}$$

the dissolution speed of the downstream front of a traffic jam and the velocity of upstream propagation of perturbations in congested traffic. While $\rho_{\max}$ denotes the maximum vehicle density in vehicle queues, $T \approx 1.8$s is the safe time gap between two successive vehicles.

- The vehicle density in traffic jams is basically constant.

These assumptions may be compensated for by suitable corrections [11], but already the model below displays a rich spectrum of spatio-temporal behaviors and contains the main elements of traffic dynamics we are interested in here.

2.1 Flow Conservation Laws

In the following, we will introduce our equations for traffic flows in networks very shortly, as a detailed justification and derivation has been given elsewhere [12, 13, 16]. These equations are also meaningful for pipeline networks [3] (if complemented by equations for momentum conservation), logistic systems [20], or supply networks [15]. Our notation is illustrated in Fig. 1.

Compared to [16], we will use a simplified notation, here.[1] The *arrival flow* $A_j(t)$ denotes the actual inflow of vehicles into the upstream end of road section j, while $O_j(t)$ is the actual *departure flow*, i.e. the flow of vehicles leaving road section j at its downstream end. The quantity

$$\widehat{Q}_j = \left(T + \frac{1}{V_j^0 \rho_{\max}} \right)^{-1} = \frac{\rho_{\max}}{1/c + 1/V_j^0} \tag{2}$$

represents the maximum in- or outflow of road section j. All the above quantities refer to flows *per lane*. I_j is the number of lanes and L_j the length of road section j. $l_j(t) \le L_j$ is the length of the congested area on link j (measured from the downstream end), and ΔN_j is the number of stopped or delayed vehicles, see (16) and (14). With these definitions, we can formulate constraints for the *actual* arrival and departure flows, which are given by the *potential arrival flows* $\widehat{A}_j(t)$ and the *potential departure flows* $\widehat{O}_i(t)$, respectively.

The actual arrival flow $A_j(t)$ is limited by the maximum inflow $\widehat{Q}_j$, if road section j is not fully congested ($l_j(t) < L_j$). Otherwise (if $l_j = L_j$) it is limited by the actual departure flow $O_j(t - L_j/c)$ a time period L_j/c before, as it requires this time period until the downstream flow value has propagated upto the upstream end of the road section by forward movement of vehicles under congested traffic conditions. This implies

[1] The arrival flow $A_j(t)$ has previously been denoted by $Q_j^{\text{arr}}(t)$, the potential arrival flow $\widehat{A}_j(t)$ by $Q_j^{\text{arr,pot}}(t)$, the departure flow $O_j(t)$ by $Q_j^{\text{dep}}(t)$ and the potential departure flow $\widehat{O}_j(t)$ by $Q_j^{\text{dep,pot}}(t)$.

$$0 \leq A_j(t) \leq \widehat{A}_j(t) := \begin{cases} \widehat{Q}_j & \text{if } l_j(t) < L_j \\ O_j(t - L_j/c) & \text{if } l_j(t) = L_j \,. \end{cases} \tag{3}$$

Moreover, the potential departure flow $\widehat{O}_i(t)$ of road section i is given by its permeability $\gamma_i(t)$ times the maximum outflow $\widehat{Q}_i$ from this road section, if vehicles are queued up ($\Delta N_i > 0$) and waiting to leave. Otherwise (if $\Delta N_i = 0$) the outflow is limited by the permeability times the arrival flow A_i a time period L_i / V_i^0 before, as this is the time period that entering vehicles need to reach the end of road section i when moving freely at the speed V_i^0. This gives the additional relationship

$$0 \leq O_i(t) \leq \widehat{O}_i(t) := \gamma_i(t) \begin{cases} A_i(t - L_i / V_i^0) & \text{if } \Delta N_i(t) = 0 \\ \widehat{Q}_i & \text{if } \Delta N_i(t) > 0 \,. \end{cases} \tag{4}$$

The permeability $\gamma_i(t)$ for traffic flows at the downstream end of section i can assume values between 0 and 1. In case of a traffic light, $\gamma_i(t) = 1$ corresponds to a green light for road section i, while $\gamma_i(t) = 0$ corresponds to a red or amber light.

Alternatively and shorter than (3) and (4) one can write

$$\widehat{A}_j(t) = \max\left[\widehat{Q}_j \Theta(l_j(t) < L_j), O_j(t - L_j/c)\right] \tag{5}$$

and

$$\widehat{O}_i(t) = \gamma_i(t) \max\left[\widehat{Q}_i \Theta(\Delta N_i > 0), A_i(t - L_i / V_i^0)\right], \tag{6}$$

where the Heaviside function Θ is 1, if the argument (inequality) has the logical value "true", otherwise it is 0. Note that the above treatment of the traffic flow in a road section requires the specification of the boundary conditions only, as we have integrated up Lighthill's and Whitham's partial differential equation over the length of the road section. The dynamics in the inner part of the section can be easily reconstructed from the boundary conditions thanks to the constant characteristic speeds. However, a certain point of the road section may be determined either from the upstream boundary (in the case of free traffic) or by the downstream boundary (if lying in the congested area, i.e. behind the upstream congestion front). Therefore, we have a switching between the influence of the upstream and the downstream boundary conditions, which makes the dynamics both, complicated and interesting. This switching results from the maximum functions above and implies also that material flows in networks are described by hybrid equations. Although the dynamics is determined by linear ordinary differential equations in all regimes, the switching between the regimes can imply a complex dynamics and even deterministic chaos [30].

Complementary to the above equations, we have now to specify the constraints for the nodes, i.e. the connection, merging, diverging or intersection points of the homogeneous road sections. Let the ingoing links be denoted by the index i and the outgoing ones by j. To distinguish quantities more easily when we insert concrete values $1, 2, \ldots$ for i and j, we mark quantities of outgoing links additionally by a prime ($'$).

Due to the condition of flow conservation, the arrival flow into a road section j with I'_j lanes must agree with the sum of the fractions α_{ij} of all outflows $I_i O_i(t)$ turning into road section i. Additionally, the arrival flows are limited, i.e. we have

$$I'_j A'_j(t) = \sum_i I_i O_i(t) \alpha_{ij} \le I'_j \widehat{A}'_j(t) \tag{7}$$

for all j. Of course, the turning fractions $\alpha_{ij} \ge 0$ are normalized due to flow conservation:

$$\sum_j \alpha_{ij}(t) = 1\,. \tag{8}$$

In cases of no merging flows, (7) simplifies to

$$I'_j A'_j(t) = I_i O_i(t) \alpha_{ij} \le I'_j \widehat{A}'_j(t) \tag{9}$$

for all j. At the same time, $0 \le O_i(t) \le \widehat{O}_i(t)$ must be fulfilled for all i. Together, this implies

$$O_i(t) \le \min \left[\widehat{O}_i(t), \min_j \left(\frac{I'_j \widehat{A}'_j}{I_i \alpha_{ij}} \right) \right] \tag{10}$$

for all i.

The advantage of the above model is that it contains the most important elements of the traffic dynamics in networks. This includes the transition from free to congested traffic flows due to lack of capacity, the propagation speeds of vehicles and congested traffic, spillover effects (i.e. obstructions when entering fully congested road sections) and, implicitly, load-dependent travel times as well.

2.2 Two Views on Traffic Jams

Let us study the traffic dynamics on the road sections in more detail. Traffic jams can be handled in two different ways: First by determining the number of cars that are delayed compared to free traffic or, second, by determining fronts and ends of traffic jams. The former method is more simple, but it cannot deal correctly with spill-over effects, when the end of a traffic jam reaches the end of a road section. Therefore, the first method is sufficient only in situations where the spatial capacity of road sections is never exceeded.

2.2.1 Method 1: Number of Delayed Vehicles

The first method just determines the difference between the number $N_i^{\text{in}}(t)$ of vehicles that would reach the end of road section i upto time t and the number $N_i^{\text{out}}(t)$ of vehicles that actually leave the road section upto this time. $N_i^{\text{in}}(t)$ just corresponds to the number of vehicles which have entered the road section upto time $t - L_i / V_i^0$, as L_i / V_i^0 is the free travel time. This implies

$$N_i^{\text{in}}(t) = \int_0^t dt' \; A_i(t' - L_i / V_i^0) \, , \tag{11}$$

while the number of vehicles that have actually left the road section upto time t is

$$N_i^{\text{out}}(t) = \int_0^t dt' \; O_i(t') \, . \tag{12}$$

Hence, the number $\Delta N_i(t)$ of delayed vehicles is given by

$$\Delta N_i(t) = \int_0^t dt' \; [A_i(t' - L_i / V_i^0) - O_i(t')] \geq 0 \, . \tag{13}$$

Alternatively, one can use the following differential equation for the temporal change in the number of delayed vehicles:

$$\frac{d\Delta N_i}{dt} = A_i(t - L_i / V_i^0) - O_i(t) \, . \tag{14}$$

In contrast, the number of *all* vehicles on road section i (independently of whether they are delayed or not) changes in time according to

$$\frac{dN_i}{dt} = A_i(t) - O_i(t) \, . \tag{15}$$

2.2.2 Method 2: Jam Formation and Resolution

In our simple macroscopic traffic model, the formation and resolution of traffic jams is described by the shock wave equations, where we have the two characteristic speeds V_i^0 (the free speed) and c (the jam resolution speed). According to the theory of shock waves [27,35], the upstream end of a traffic jam, which is located at a place $l_i(t) \geq 0$ upstream of the end of road section i, is moving at the speed

$$\frac{dl_i}{dt} = -\frac{A_i\big(t - [L_i - l_i(t)]/V_i^0\big) - O_i\big(t - l_i(t)/c\big)}{\rho_1(t) - \rho_2(t)} \tag{16}$$

with the (free) density

$$\rho_1(t) = A_i\big(t - [L_i - l_i(t)]/V_i^0\big)/V_i^0 \tag{17}$$

immediately before the upstream shock front and the (congested) density

$$\rho_2(t) = [1 - TO_i\big(t - l_i(t)/c\big)]\rho_{\max} \tag{18}$$

immediately downstream of it. This is, because free traffic is upstream of the shock front, and congested traffic downstream of it (for details see (1.6) and (1.4) in [16]). In contrast, the downstream front of a traffic jam is moving at the speed

$$-\frac{0 - O_i\big(t - l_i(t)/c\big)}{\rho_{\max} - O_i\big(t - l_i(t)/c\big)/V_i^0} = \frac{O_i\big(t - l_i(t)/c\big)}{\rho_{\max} - O_i\big(t - l_i(t)/c\big)/V_i^0}\,, \tag{19}$$

since congested traffic with zero flow is upstream of the shock front and free traffic flow occurs downstream of it.

2.2.3 Comparison of the Two Methods

Let us discuss a simple example to make the differences of both descriptions clearer. For this, we assume that, at time $t = 0$, traffic flow on the overall road section i is free, i.e. any traffic jam has resolved and there are no delayed vehicles. The flow shall be stopped by a red traffic light for a time period t_0. At time $t = t_0$, the traffic light shall turn green, and the formed traffic jam shall resolve. For the arrival flow, we simply assume a constant value A_i, and the road section shall be long enough to take up the forming traffic jam. Moreover, the departure flow shall be O_i. Then, according to method 1, the number of delayed vehicles at time t_0 is

$$\Delta N_i(t_0) = A_i t_0\,, \tag{20}$$

and it is reduced according to

$$\Delta N_i(t) = A_i t_0 - (O_i - A_i)(t - t_0)\,. \tag{21}$$

Therefore, any delays are resolved after a time period

$$t - t_0 = \frac{A_i t_0}{O_i - A_i} = \frac{\Delta N_i(t_0)}{O_i - A_i}\,, \tag{22}$$

i.e. at time

$$t_2 = t_0 \frac{O_i}{O_i - A_i}\,. \tag{23}$$

Afterwards, $\Delta N_i(t) = 0$.

In contrast, the end of the traffic jam grows with the speed

$$\frac{dl_i}{dt} = -\frac{A_i - 0}{A_i/V_i^0 - (1-0)\rho_{\max}} = \frac{1}{\rho_{\max}/A_i - 1/V_i^0} =: C_i \,. \tag{24}$$

Therefore, we have $l_i(t_0) = C_i t_0$. Surprisingly, this is greater than $\Delta N_i(t_0)/\rho_{\max}$, i.e. the expected length of the traffic jam based on the number of delayed vehicles. The reason is that the delay of a vehicle joining the traffic jam at location $x_i = L_i - l_i$ is noticed at the downstream end of the road section only after a time period l_i/V_i^0.

The resolution of the traffic jam starts from the downstream end with the speed

$$\frac{0 - \widehat{Q}_i}{\rho_{\max} - \widehat{Q}_i/V_i^0} = \frac{-1}{\rho_{\max}/\widehat{Q}_i - 1/V_i^0} = -c \,, \tag{25}$$

if the outflow is free (i.e. $O_i = \widehat{Q}_i$), otherwise with the speed

$$\frac{0 - O_i}{\rho_{\max} - (\rho_{\max} - O_i/c)} = -c \,, \tag{26}$$

since congested traffic with zero flow and maximum density is upstream of the shock front.

Obviously, the jam resolution has reached the further growing, upstream jam front when $C_i t = c(t - t_0)$. Therefore, the jam of density $\rho_{\max}$ has disappeared after a time period $t - t_0 = C_i t_0/(c - C_i)$, i.e. at time

$$t_1 = c t_0/(c - C_i)\,. \tag{27}$$

Surprisingly, it can be shown that $t_1 < t_2$, i.e. the traffic jam resolves before the number of delayed vehicles reaches a value of zero. In fact, it still takes the time $C_i t_1/\widehat{V}_i^0$ until the last delayed vehicle has left the road section, where

$$\widehat{V}_i^0 = \frac{A_i - O_i}{A_i/V_i^0 - (\rho_{\max} - O_i/c)} \tag{28}$$

is the shock front between free upstream traffic flow and the congested outflow O_i, which usually differs from the speed $V_i = O_i/[(1 - TQ_i)\rho_{\max}]$ of outflowing vehicles. For $O_i = \widehat{Q}_i$, we have $\widehat{V}_i^0 = V_i^0$ because of $1/c = \rho_{\max}/\widehat{Q}_i - 1/V_i^0$.

Undelayed traffic starts when this shock front reaches the end of the road section, i.e. at time

$$t_2 = t_1\left(1 + \frac{C_i}{\widehat{V}_i^0}\right) = \frac{t_0}{1 - C_i/c}\left(1 + \frac{C_i(A_i/V_i^0 - \rho_{\max}) + C_i O_i/c}{A_i - O_i}\right). \tag{29}$$

Inserting $C_i(A_i/V_i^0 - \rho_{\max}) = -A_i$ eventually gives $t_2 = t_0 O_i/O_i - A_i)$. This agrees perfectly with the above result for the first method (based on vehicle delays rather than traffic jams).

In conclusion, both methods of dealing with traffic jams are consistent, and delayed vehicles occur as soon as traffic jam formation begins. However, according to method 1, a queued vehicle at position $x_i = L_i - l_i$ is counted as delayed only after an extra time period l_i/V_i^0, but it is counted as undelayed after the same extra time period. This is because method 1 counts on the basis of vehicle arrivals at the downstream end of road section i.

As it is much simpler to use the method 1 based on determining the number of delayed vehicles than using method 2 based on determining the movement of shock fronts, we will use method 1 in the following. More specifically, in (3) we will replace $l_j(t) < L_j$ by $\Delta N_j(t) < N_j^{\max} := L_j\rho_{\max}$ and $l_j(t) = L_j$ by $\Delta N_j(t) = N_j^{\max}$. This corresponds to a situation in which the vehicles would not queue up along the road section, but at the downstream end of the road section, like in a wide parking lot or on top of each other. As long as road section j is not fully congested, this difference does not matter significantly. If it is fully congested, the dynamics will potentially be different, defining a modified model of material network flows. However both, the original and the modified model fulfill the conservation equation and show spillover effects.

2.2.4 Calculation of Cumulative and Maximum Individual Waiting Times

In [12], we have derived a delay differential equation to determine the travel time $T_i(t)$ of a vehicle entering road section i at time t (see also [1,2,5]):

$$\frac{dT_i(t)}{dt} = \frac{A_i(t)}{O_i(t + T_i(t))} - 1 \,. \tag{30}$$

According to this, the travel time $T_i(t)$ increases with time, when the arrival rate A_i at the time t of entry exceeds the departure rate O_i at the leaving time $t+T_i(t)$, while it decreases when it is lower. It is remarkable that this formula does not explicitly depend on the velocities on the road section, but only on the arrival and departure rates.

Another method to determine the travel times is to integrate up over the number of vehicles arriving in road section i,

$$N_i^A(t) = \int\limits_0^t dt'\ A_i(t') = N_i^{\text{in}}(t + L_i/V_i^0)\,, \tag{31}$$

and over the number of vehicles leaving it,

$$N_i^O(t) = \int\limits_0^t dt'\ O_i(t') = N_i^{\text{out}}(t)\,, \tag{32}$$

starting at time $t = 0$ when there are no vehicles in the road. If $T_i'(t)$ denotes the time at which $N_i^O(t + T_i'(t)) = N_i^A(t)$, then $T_i'(t)$ is the travel time of a vehicle entering road section i at time t and

$$T_i(t) = T_i'(t) - L_i / V_i^0 \tag{33}$$

is its waiting time.

Another interesting quantity is the cumulative waiting time $T_i^{\mathrm{c}}(t)$, which is determined by integrating up over the number ΔN_i of all delayed vehicles. We obtain

$$\begin{aligned} T_i^{\mathrm{c}}(t) &= \int_0^t dt'\ \Delta N_i(t') = \int_0^t dt'\ [N_i^{\mathrm{in}}(t' - L_i/V_i^0) - N_i^{\mathrm{out}}(t')] \\ &= \int_0^t dt' \int_0^{t'} dt''\ [A_i(t'' - L_i/V_i^0) - O_i(t'')] \end{aligned} \tag{34}$$

and the differential equation

$$\frac{dT_i^{\mathrm{c}}(t)}{dt} = \Delta N_i(t) = \int_0^t dt'\ [A_i(t' - L_i/V_i^0) - O_i(t')] \,. \tag{35}$$

For a constant arrival flow A_i and a red traffic light from $t = 0$ to $t = t_0$ (i.e. $O_i(t) = 0$), we find

$$T_i^{\mathrm{c}} = \frac{A_i t^2}{2} \,. \tag{36}$$

In this time period, a number of $N_i(t) = A_i t$ vehicles accumulates, which gives an average waiting time of

$$\frac{T_i^{\mathrm{c}}(t_0)}{\Delta N(t_0)} = \frac{t_0}{2} \tag{37}$$

at the end of the red light. The first vehicle has to wait twice as long, namely, a time period t_0.

3 Treatment of Merging, Diverging and Intersection Points

While the last section has given general formulas that must be fulfilled at nodes connecting two or more different links, in the following we will give some concrete examples, how to deal with standard elements of street networks. For previous treatments of traffic flows at intersections see, for example, [7, 9, 24, 26].

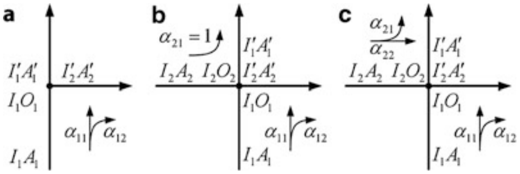


Fig. 1 Schematic illustration of the (**a**) diverging, (**b**) merging, and (**c**) intersecting flows discussed in this paper

3.1 Diverging Flows: One Inflow and Several Outflows

In the case of one road section i diverging into several road sections j (see Fig. 1a), (10) and (5) to (7) imply

$$O_i(t) \le \min\left\{\gamma_i(t)\max\left[\widehat{Q}_i\Theta(\Delta N_i > 0), A_i\left(t - \frac{L_i}{V_i^0}\right)\right],\right.$$
$$\left.\min_j\left[\frac{I'_j}{I_i\alpha_{ij}}\max\left(\widehat{Q}_j\Theta(l_j < L_j), O_j(t - L_j/c)\right)\right]\right\} \tag{38}$$

for all i. If we assume that downstream road sections are never completely congested, this simplifies to

$$O_i(t) = \min\left\{Q_i, \gamma_i\max\left[\widehat{Q}_i\Theta(\Delta N_i > 0), A_i\left(t - L_i/V_i^0\right)\right]\right\} \tag{39}$$

with

$$Q_i = \min_j\left(\frac{I'_j\widehat{Q}_j}{I_i\alpha_{ij}}\right). \tag{40}$$

Otherwise

$$Q_i(t) = \min_j\left[\max\left(\frac{I'_j\widehat{Q}_j}{I_i\alpha_{ij}}\Theta(l_j < L_j), \frac{I'_jO_j(t - \frac{L_j}{c})}{I_i\alpha_{ij}}\right)\right]. \tag{41}$$

3.2 Merging Flows: Two Inflows and One Outflow

We assume a flow $I_1O_1(t)$ that splits into two flows $I_1O_1(t)\alpha_{11}$ (going straight) and $I_1O_1(t)\alpha_{12}$ (turning right), but a right-turning flow $I_2O_2(t)$ merging with flow $I_1O_1(t)\alpha_{11}$, as in turn-right-on-red setups (see Fig. 1b). For this situation, we have the equations

$$I'_1A'_1(t) = I_1O_1(t)\alpha_{11} + I_2O_2(t) \le I'_1\widehat{A}'_1(t)\,, \tag{42}$$

$$I'_2A'_2(t) = I_1O_1(t)\alpha_{12} \le I'_2\widehat{A}'_2(t)\,. \tag{43}$$

One can derive

$$0 \leq O_1 = \min\left[\widehat{O}_1(t), \frac{I_1'\widehat{A}_1'(t) - I_2 O_2(t)}{I_1\alpha_{11}}, \frac{I_2'\widehat{A}_2'(t)}{I_1\alpha_{12}}\right] \tag{44}$$

and

$$0 \leq O_2 = \min\left[\widehat{O}_2(t), \frac{I_1'\widehat{A}_1'(t) - I_1 O_1(t)\alpha_{11}}{I_2}\right]. \tag{45}$$

Let us set

$$O_1 = \min\left[\widehat{O}_1, \frac{I_1'\widehat{A}_1'(t)}{I_1\alpha_{11}}, \frac{I_2'\widehat{A}_2'(t)}{I_1\alpha_{12}}\right] \tag{46}$$

and

$$O_2(O_1) = \min\left[\widehat{O}_2(t), \frac{I_1'\widehat{A}_1'(t) - I_1 O_1\alpha_{11}}{I_2}\right]. \tag{47}$$

Then, it can be shown that $O_2(t) \geq 0$ and $O_1(t) \leq [I_1'\widehat{A}_1'(t) - I_2 O_2(t)]/(I_1\alpha_{11})$, as demanded. If $O_1(t)$ is chosen a value ΔO_1 smaller than specified in (46), but O_2 is still set to the maximum related value $O_2(O_1 - \Delta O_1)$ according to (47), the overall flow

$$F = I_1 O_1 + I_2 O_2 \tag{48}$$

is reduced as long as $\alpha_{11} < 1$, since this goes along with additional turning flows (while the number of lanes does not matter!). Therefore, it is optimal to give priority to the outflow $O_1(t)$ according to (46) and to add as much outflow $O_2(t)$ as capacity allows. This requires suitable flow control measures, otherwise the optimum value of the overall flow F could not be reached. In fact, the merging flow would "steel" some of the capacity reserved for the "main" flow ($i = 1$), which would reduce the possible outflow $O_1(t)$ and potentially cause a breakdown of free traffic flow, as it is known from on-ramp areas of freeways [32] .

3.3 A Side Road Merging with a Main Road

Compared to the last section, the situation simplifies, if we have just a side road or secondary turning flow merging with a the flow of a main road without any turning flow away from the main road. In this case, we have $\alpha_{11} = 1$ and $\alpha_{12} = 0$, which leaves us with the relationships

$$O_1 = \min\left[\widehat{O}_1, \frac{I_1'\widehat{A}_1'(t)}{I_1}\right] \tag{49}$$

and

$$O_2(O_1) = \min\left[\widehat{O}_2(t), \frac{I_1'\widehat{A}_1'(t) - I_1 O_1}{I_2}\right]. \tag{50}$$

according to (46) and (47).

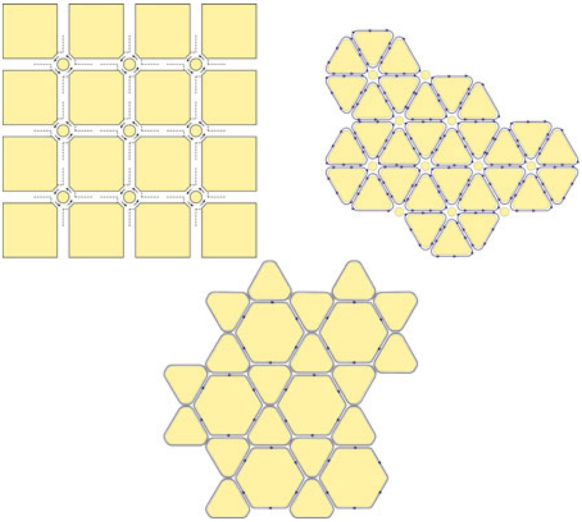


Fig. 2 Three examples for intersection-free designs of urban road networks

3.4 Intersection-Free Designs of Road Networks

With the formulas for the treatment of merges and diverges in the previous sections, it is already possible to simulate intersection-free designs of urban road networks, which do not need any traffic light control. The most well-known design of intersection-free nodes are roundabouts (see the upper left illustration in Fig. 2). It is, however, also possible to construct other intersection-free designs based on subsequent merges and diverges of flows with different destinations. Two examples are presented in Fig. 2b, c.

Although intersection-free designs require the driver to take small detours, such a road network will normally save travel time and fuel, given that the traffic volume is not too low. This is because intersections then need to be signalized in order to be safe and efficient.[2] Traffic signals, however, imply that vehicles will often be stopped for considerable time intervals. This causes significant delays, at least for vehicles not being served by a green wave. Intersection-free designs, in contrast, do not necessarily require vehicles to stop. Therefore, the average speeds are expected to be higher and the travel times lower than for road networks with intersections. This has significant implications for urban transport planning, if intersections cannot be avoided by bridges or tunnels.

[2] Of course, a first-come-first-serve or right-before-left rule will be sufficient at small traffic volumes.

3.5 Two Inflows and Two Outflows

The treatment of intersecting flows is more complicated than the treatment of merges and diverges. Moreover, the resulting flows are only uniquely defined, if additional rules are introduced such as the optimization of the overall flow. Let us here treat the case of an intersection with two inflows and two outflows (see Fig. 1c). Equation (5) implies the inequalities

$$\begin{aligned} 0 \le I'_1 A'_1(t) &= I_1 O_1(t)\alpha_{11} + I_2 O_2(t)\alpha_{21} \le I'_1 \widehat{A}'_1(t)\,, \\ 0 \le I'_2 A'_2(t) &= I_1 O_1(t)\alpha_{12} + I_2 O_2(t)\alpha_{22} \le I'_2 \widehat{A}'_2(t) \end{aligned} \tag{51}$$

with the constraints

$$\begin{aligned} 0 \le O_1(t) &\le \widehat{O}_1(t)\,, \\ 0 \le O_2(t) &\le \widehat{O}_2(t)\,, \end{aligned} \tag{52}$$

so that $I'_j A'_j(t) \ge 0$ is automatically fulfilled. The constraints (52) define an rectangular area of possible O_i-values in the O_1-O_2 plane, where the size of the rectangle varies due to the time-dependence of $\widehat{O}_i(t)$. The inequalities (51) can be rewritten as

$$O_2(t) \le \frac{I'_1 \widehat{A}_1(t) - I_1 O_1(t)\alpha_{11}}{I_2 \alpha_{21}} =: a_1 - b_1 O_1(t)\,, \tag{53}$$

and

$$O_2(t) \le \frac{I'_2 \widehat{A}_2(t) - I_1 O_1(t)\alpha_{12}}{I_2 \alpha_{22}} =: a_2 - b_2 O_1(t)\,. \tag{54}$$

They potentially cut away parts of this rectangle, and the remaining part defines the convex set of feasible points (O_1, O_2) at time t. We are interested to identify the "optimal" solution (O_1^*, O_2^*), which maximizes the overall flow

$$\sum_j I'_j A'_j(t) = \sum_i I_i O_i(t)\,. \tag{55}$$

As this defines a linear optimization problem, the optimal solution corresponds to one of the corners of the convex set of feasible points, namely the one which is touched first by the line

$$O_2 = \frac{Z - I_1 O_1}{I_2}\,, \tag{56}$$

when we reduce Z from high to low values.

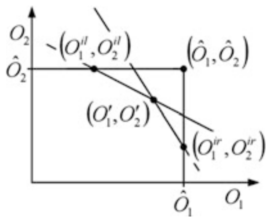


Fig. 3 Illustration of the possible optimal solutions for two intersecting flows (see text for details)

Let us, therefore, determine all possible corners of the convex set and the conditions, under which they correspond to the optimal solution. We will distinguish the following cases:

(a) *None* of the boundary lines (53) and (54) corresponding to the equality signs cuts the rectangle defined by $0 \le O_1(t) \le \widehat{O}_1(t)$ and $0 \le O_2(t) \le \widehat{O}_2(t)$ in more than 1 point. This case applies, if $a_1 - b_1\widehat{O}_1 \ge \widehat{O}_2$ and $a_2 - b_2\widehat{O}_1 \ge \widehat{O}_2$, as $a_i \ge 0$ and $b_i \ge 0$ implies that both lines are falling or at least not increasing. Since the line (56) reflecting the goal function is falling as well, the optimal point is

$$(O_1^*, O_2^*) = (\widehat{O}_1, \widehat{O}_2)\,, \tag{57}$$

i.e. the outer corner of the rectangle corresponding to the potential or maximum possible departure flows (see Fig. 3).

(b) Only *one* of the two boundary lines/border lines, $O_2(t) = a_1 - b_1 O_1(t)$ or $O_2(t) = a_2 - b_2 O_1(t)$, cuts the rectangle in *more* than one point. Let us assume, this holds for line i, i.e. $a_i - b_i\widehat{O}_1 < \widehat{O}_2$. Then, the left cutting point

$$(O_1^{i1}, O_2^{i1}) = \begin{cases} \Big((a_i - \widehat{O}_2)/b_i, \widehat{O}_2\Big) & \text{if } a_i > \widehat{O}_2\,, \\ (0, a_i) & \text{otherwise} \end{cases} \tag{58}$$

is the optimal point if $I_1/I_2 < b_i$, i.e. if the slope I_1/I_2 of the goal function (56) is smaller than the one of the cutting border line. Otherwise, if $I_1/I_2 > b_i$, the optimal point is given by the right cutting point

$$(O_1^{i\mathrm{r}}, O_2^{i\mathrm{r}}) = \begin{cases} (\widehat{O}_1, a_i - b_i\widehat{O}_1) & \text{if } a_i > b_i\widehat{O}_1\,, \\ (a_i/b_i, 0) & \text{otherwise} \end{cases} \tag{59}$$

(see Fig. 3).

(c) If *both* border lines cut through the rectangle, but one of them lies above the other line, then only the lower line determines the optimal solution, which can be obtained as in case (b). Case (c) occurs if $a_2 - b_2 O_1^{11} > a_1 - b_1 O_1^{11}$ and $a_2 - b_2 O_1^{1\mathrm{r}} > a_1 - b_1 O_1^{1\mathrm{r}}$ (line 1 is the lower one) or if $a_2 - b_2 O_1^{11} < a_1 - b_1 O_1^{11}$ and $a_2 - b_2 O_1^{1\mathrm{r}} < a_1 - b_1 O_1^{1\mathrm{r}}$ (line 2 is the lower one).

(d) The boundary lines cut each other in the inner part of the rectangle. This occurs if $a_1 - b_1\widehat{O}_1 < \widehat{O}_2$ and $a_2 - b_2\widehat{O}_1 < \widehat{O}_2$. Then, the left-most cutting point

(O_1^{i1}, O_2^{i1}) is the optimal solution, if the slope I_1/I_2 of the goal function is smaller than the smallest slope of the two boundary lines, while it is the lower right cutting point $(O_1^{i\mathrm{r}}, O_2^{i\mathrm{r}})$, if I_1/I_2 is greater than the steepest slope of the two boundary lines, otherwise, the cutting point of the two boundary lines,

$$(O_1', O_2') = \left(\frac{a_2 - a_1}{b_2 - b_1}, \frac{a_1 b_2 - b_1 a_2}{b_2 - b_1} \right) \tag{60}$$

is the optimal point (see Fig. 3). Mathematically speaking, we have

$$(O_1^*, O_2^*) = \begin{cases} (O_1^{11}, O_2^{11}) & \text{if } I_1/I_2 < b_1 < b_2, \\ (O_1', O_2') & \text{if } b_1 < I_1/I_2 < b_2, \\ (O_1^{2\mathrm{r}}, O_2^{2\mathrm{r}}) & \text{if } b_1 < b_2 < I_1/I_2, \\ (O_1^{2\mathrm{l}}, O_2^{2\mathrm{l}}) & \text{if } I_1/I_2 < b_2 < b_1, \\ (O_1', O_2') & \text{if } b_2 < I_1/I_2 < b_1, \\ (O_1^{1\mathrm{r}}, O_2^{1\mathrm{r}}) & \text{if } b_2 < b_1 < I_1/I_2, \end{cases} \tag{61}$$

It is astonishing that the simple problem of two intersecting traffic flows has so many different optimal solutions, which sensitively depend on the parameter values. This can reach from situations where both outgoing road sections experience the maximum possible outflows upto situations, where the outflow in the system-optimal point becomes zero for one of the road sections. A transition from one optimal solution to another one could easily be triggered by changes in the turning fractions α_{ij} entering the parameters a_i and b_i, for example due to time-dependent turning fractions $\alpha_{ij}(t)$.

3.6 Inefficiencies Due to Coordination Problems

An interesting question is how to actually establish the flows corresponding to the system optima that were determined in the previous sections on merging and intersecting flows. Of course, zero flows can be enforced by a red traffic light, while maximum possible flows can be established by a node design giving the right of way to one road (the "main" road). However, it is not so easy to support an optimum point corresponding to mixed flows, such as (O_1', O_2'). That would need quite tricky intersection designs or the implementation of an intelligent transportation system ensuring optimal gap usage, e.g. based on intervehicle communication. Only in special cases, the task could be performed by a suitable traffic light control.

In normal merging or intersection situations, there will always be coordination problems [19] when entering or crossing another flow, if the traffic volumes reach a certain level. This will cause inefficiencies in the usage of available road capacity, i.e. mixed flows will not be able to use the full capacity. Such effects can be modelled by specifying the corresponding permeabilities $\gamma_i(t)$ as a function of the merging flows, particularly the main flow or crossing flow. The deviation of $\gamma_i(t)$ from 1 will then be a measure for the inefficiency. A particularly simple, phenomenological specification would be

$$\gamma_2(t) = \frac{1}{1 + a\mathrm{e}^{b(O_1 - O_2)}} , \tag{62}$$

where the own outflow O_2 supports a high permeability and the intersecting outflow O_1 suppresses it. However, rather than using such a phenomenological approach, the permeability could also be calculated analytically, based on a model of gap statistics, since large enough vehicle gaps are needed to join or cross a flow. Such kinds of calculations have been carried out in [18, 28, 33, 34].

4 Towards a Self-Organized Traffic Light Control

In [16], it has been pointed out that, for not too small arrival flows, an oscillatory service at intersections reaches higher intersection capacities and potentially shorter waiting times than a first-in-first-out service of arriving vehicles. This is due to the fact that the outflow of queued vehicles is more efficient than waiting for the arrival of other freely flowing vehicles, which have larger time gaps. For similar reasons, pedestrians are passing a bottleneck in an oscillatory way [14], and also two intersecting flows tend to organize themselves in an oscillatory way [17, 18].

Therefore, using traffic lights at intersections is natural and useful, if operated in the right way. However, how to switch the traffic lights optimally? While this is a solvable problem for single traffic lights, the optimal coordination of many traffic lights [29] is a really hard (actually NP hard) problem [8]. Rather than solving a combinatorial optimization problem, here, we want to suggest a novel approach, which needs further elaboration in the future. The idea is to let the network flows self-organize themselves, based on suitable equations for the permeabilities $\gamma_i(t)$ as a function of the outflows $O_i(t)$ and the numbers $\Delta N_i(t)$ of delayed vehicles.

Here, we will study the specification

$$\gamma_1(t) = \frac{1}{1 + a\mathrm{e}^{b(O_2 - O_1) - cD}} \tag{63}$$

and

$$\gamma_2(t) = \frac{1}{1 + a\mathrm{e}^{b(O_1 - O_2) + cD}} , \tag{64}$$

which generalizes formula (62). While the relative queue length

$$D(t) = \Delta N_1(t) - \Delta N_2(t) \tag{65}$$

quantifies the pressure to increase the permeability γ_1 for road Sect. 1, the outflow $O_2(t)$ from the road Sect. 2 resists this tendency, while the flow $O_1(t)$ on road Sect. 1 supports the permeability. The increasing pressure eventually changes the resistance

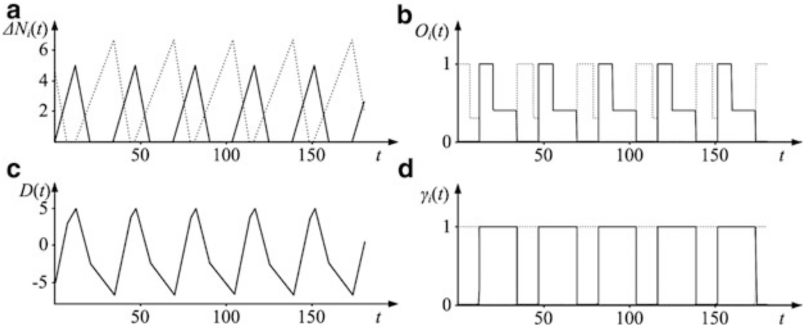


Fig. 4 Illustration of the dynamics of self-organized oscillations in the permeabilities and the resulting flows for a single intersection with constant inflows (see text for details). Note that the road section with the higher inflow (arrival rate) is served longer, and its queues are shorter (see *solid lines*)

threshold and the service priority. An analogous situation applies to the permeability γ_2 for road Sect. 2, where the pressure corresponds to $-D$, which is again the difference in queue length. a, b, and c are non-negative parameters. a may be set to 1, while c must be large enough to establish a sharp switching. Here, we have assumed $c = 100$. The parameter b allows to influence the switching frequency f, which is approximately proportional to b. We have adjusted the frequency f to the cycle time

$$T^{\text{cyc}} = \frac{2\tau}{1 - (A_1 + A_2)/\widehat{Q}}, \tag{66}$$

which results if the switching (setup) time ("yellow traffic light") is $\tau = 5$s and a green light is terminated immediately after a queue has dissolved after lifting the red light.[3] The corresponding parameter value is

$$b = \frac{500}{\widehat{Q} - (A_1 + A_2)}. \tag{67}$$

Figure 4 shows a simulation result for $A_1/\widehat{Q} = 0.3$ and $A_2/\widehat{Q} = 0.4$. The properties of the corresponding specification of the permeabilities $\gamma_i(t)$ are as follows:

- $\gamma_i(t)$ is non-negative and does not exceed the value 1.
- For the sum of permeabilities and $a \geq 1$, we have

[3] If ΔT_1 and ΔT_2 denote the green time periods for the intersecting flows 1 and 2, respectively, the corresponding red time periods for a periodic signal control are ΔT_2 and ΔT_1, to which the switching setup time of duration τ must be added. From formula (23) and with $O_i = \widehat{Q}$ we obtain $\Delta T_1 = (\Delta T_2 + \tau)\widehat{Q}/(\widehat{Q} - A_1)$ and $\Delta T_2 = (\Delta T_1 + \tau)\widehat{Q}/(\widehat{Q} - A_2)$. Using the definition $T^{\text{cyc}} = \Delta T_1 + \tau + \Delta T_2 + \tau$ for the cycle time, we finally arrive at (66).

$$\gamma_1 + \gamma_2 = \frac{2 + a(\mathrm{e}^{E} + \mathrm{e}^{-E})}{1 + a^2 + a(\mathrm{e}^{E} + \mathrm{e}^{-E})} \le 1\,, \tag{68}$$

where we have introduced the abbreviation

$$E = b(O_1 - O_2) + c(\varDelta N_1 - \varDelta N_2)\,. \tag{69}$$

The sum is close to 1 for large absolute values of E, while for $E \approx 0$ the overall permeability $\gamma_1 + \gamma_2$ is small.

- For large enough values of ab and for $c, A_1, A_2 > 0$, the equations for the permeability do not have a stable stationary solution. This can be concluded from

$$\frac{dE}{dt} = b\left(\frac{dO_1}{dt} - \frac{O_2}{dt}\right) + c\left(\frac{d\varDelta N_1}{dt} - \frac{d\varDelta N_2}{dt}\right) \tag{70}$$

together with

$$\frac{d\varDelta N_i}{dt} = A_i - O_i(t) \tag{71}$$

and

$$O_i(t) = \gamma_i(t) \max[\widehat{Q}\,\Theta(\varDelta N_i > 0), A_i]\,, \tag{72}$$

see (14) and (6). As $dD/dt = d\varDelta N_1/dt - d\varDelta N_2/dt$ varies around zero, the same applies to $D(t)$, which leads to oscillations of the permeabilities $\gamma_i(t)$.

- With the specification (67) of parameter b, the cycle time is approximately proportional to the overall inflow $(A_1 + A_2)$.
- The road section with the higher flow gets a longer green time period (see Fig. 4).

If the above self-organized traffic flows shall be transfered to a new principle of traffic light control, phases with $\gamma_i(t) \approx 1$ could be interpreted as green phases and phases with $\gamma_i(t) \approx 0$ as red phases. Inefficient, intermediate switching time periods for certain choices of parameter values could be translated into periods of a yellow traffic light.

5 Summary and Outlook

We have presented a simple model for conserved flows in networks. Although our specification has been illustrated for traffic flows in urban areas, similar models are useful for logistic and production system or even transport in biological cells or bodies. Our model considers propagation speeds of entities and congestion fronts, spill-over effects, and load-dependent transportation times.

We have also formulated constraints for network nodes. These constraints contain several minimum and maximum functions, which implies a multitude of possible cases even for relatively simple intersections. It turns out that the arrival and departure flows of diverges have uniquely defined values, while merges or intersections have a set of feasible solutions. This means, the actual result may sensitively depend on the intersection design. For mathematical reasons, we have determined flow-optimizing solutions for two merging and two intersecting flows. However, it is questionable whether these solutions can be established in reality without the implementation of intelligent transport systems facilitating optimal gap usage: In many situations, coordination problems between vehicles in merging or intersection areas cause inefficiencies, which reduce their permeability.

In fact, at not too small traffic volumes, it is better to have an oscillation between minimum and maximum permeability values. Therefore, we have been looking for a mechanism producing emergent oscillations between high and low values. According to our proposed specification (which is certainly only one of many possible ones), the transition between high and low permeability was triggered, when the difference between the queue lengths of two traffic flows competing for the intersection capacity exceeded a certain value. The resulting oscillatory service could be used to *define* traffic phases. One potential advantage of such an approach would be that the corresponding traffic light control would be based on the self-organized dynamics of the system. Further work in this direction seems very promising.

Acknowledgements The authors are grateful for partial financial support by the German Research Foundation (research projects He 2789/5-1, 8-1) and by the "Cooperative Center for Communication Networks Data Analysis", a NAP project sponsored by the Hungarian National Office of Research and Technology under grant No. KCKHA005.

References

1. V. Astarita, Flow propagation description in dynamic network loading models, in *Proceedings of the IV International Conference on Applications of Advanced Technologies in Transportation Engineering (AATT)*, ed. by Y.J. Stephanedes, F. Filippi (Capri, Italy, 1995), pp. 599–603
2. V. Astarita, Node and link models for traffic simulation. Math. Comput. Model. **35**, 643–656 (2002)
3. M.K. Banda, M. Herty, A. Klar, Gas flow in pipeline networks. Netw. Heterogeneous Media **1**, 41–56 (2006)
4. G. Bretti, R. Natalini, B. Piccoli, Numerical approximations of a traffic flow model on networks. Netw. Heterogeneous Media **1**, 57–84 (2006)
5. M. Carey, Y.E. Ge, M. McCartney, A whole-link travel-time model with desirable properties. Transp. Sci. **37**, 83–96 (2003)
6. C.F. Daganzo, Requiem for second-order fluid approximations of traffic flow. Transp. Res. B **29**, 277–286 (1995)
7. C.F. Daganzo, The cell transmission model, Part II: Network traffic. Transp. Res. B **29**, 79–93 (1995)
8. B. De Schutter, Optimizing acyclic traffic signal switching sequences through an extended linear complementarity problem formulation. Eur. J. Oper. Res. **139**, 400–415 (2002)
9. M. Garavello, B. Piccoli, *Traffic Flow on Networks* (American Institute of Mathematical Sciences, Springfield, 2006)

10. S. Goettlich, M. Herty, A. Klar, Modelling and optimization of supply chains on complex networks. Comm. Math. Sci. **4**, 315–330 (2006)
11. D. Helbing, Traffic and related self-driven many-particle systems. Rev. Mod. Phys. **73**, 1067–1141 (2001)
12. D. Helbing, A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks. J. Phys. Math. Gen. **36**, L593–L598 (2003)
13. D. Helbing, Production, supply, and traffic systems: A unified description, in *Traffic and Granular Flow '03*, ed. by S.P. Hoogendoorn, S. Luding, P.H.L. Bovy, M. Schreckenberg, D.E. Wolf (Springer, Berlin, 2005), pp. 173–188
14. D. Helbing, P. Molnár, Social force model for pedestrian dynamics. Phys. Rev. E **51**, 4282–4286 (1995)
15. D. Helbing, S. Lämmer, T. Seidel, P. Seba, T. Platkowski, Physics, stability and dynamics of supply networks. Phys. Rev. E **70**, 066116 (2004)
16. D. Helbing, S. Lämmer, J.-P. Lebacque, Self-organized control of irregular or perturbed network traffic, in *Optimal Control and Dynamic Games*, ed. by C. Deissenberg, R.F. Hartl (Springer, Dordrecht, 2005), pp. 239–274
17. D. Helbing, L. Buzna, A. Johansson, T. Werner, Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transp. Sci. **39**, 1–24 (2005)
18. D. Helbing, R. Jiang, M. Treiber, Analytical investigation of oscillations in intersecting flows of pedestrian and vehicle traffic. Phys. Rev. E **72**, 046130 (2005)
19. D. Helbing, A. Johansson, J. Mathiesen, M.H. Jensen, A. Hansen, Analytical approach to continuous and intermittent bottleneck flows. Phys. Rev. Lett. **97**, 168001 (2006)
20. D. Helbing, T. Seidel, S. Lämmer, K. Peters, Self-organization principles in supply networks and production systems. in *Econophysics and Sociophysics*, ed. by B.K. Chakrabarti, A. Chakraborti, A. Chatterjee (Wiley, New York, 2006)
21. M. Herty, A. Klar, Modeling, simulation, and optimization of traffic flow networks. SIAM Appl. Math. **64**, 565–582 (2003)
22. M. Herty, A. Klar, Simplified dynamics and optimization of large scale traffic flow networks. Math. Mod. Meth. Appl. Sci. **14**, 579–601 (2004)
23. M. Herty, S. Moutari, M. Rascle, Optimization criteria for modelling intersections of vehicular traffic flow. NHM **1**, 275–294 (2006)
24. M. Hilliges, W. Weidlich, A phenomenological model for dynamic traffic flow in networks. Transp. Res. B **29**, 407–431 (1995)
25. B. Kerner, *The Physics of Traffic* (Springer, Berlin, 2004)
26. J.-P. Lebacque, M.M. Khoshyaran, First-order macroscopic traffic flow models: Intersection modeling, network modeling, in *16th International Symposium on Transportation and Traffic Theory*, ed. by H.S. Mahmasani (Elsevier, Amsterdam, 2005), pp. 365–386
27. M.J. Lighthill, G.B. Whitham, On kinematic waves: II. A theory of traffic on long crowded roads. Proc. R. Soc. Lond. A **229**, 317–345 (1955)
28. A.J. Mayne, Some further results in the theory of pedestrians and road traffic. Biometrika **41**, 375–389 (1954)
29. C.H. Papadimitriou, J.N. Tsitsiklis, The complexity of optimal queuing network control. Math. Oper. Res. **24**, 293–305 (1999)
30. K. Peters, U. Parlitz, Hybrid systems forming strange billiards. Int. J. Bifurcat. Chaos **13**, 2575–2588 (2003)
31. M. Schönhof, D. Helbing, Empirical features of congested traffic states and their implications for traffic modelling. Transp. Sci. **41**, 135–166 (2007)
32. M. Treiber, A. Hennecke, D. Helbing, Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**, 1805–1824 (2000)
33. R.J. Troutbeck, Average delay at an unsignalized intersection with two major each having a dichotomized headway distribution. Transp. Sci. **20**, 272–286 (1986)
34. R.J. Troutbeck, W. Brilon, Unsignalized intersection theory, in *Traffic Flow Theory: A State-of-the-Art Report*, ed. by N. Gartner, H. Mahmassani, C.H. Messer, H. Lieu, R. Cunard, A.K. Rathi (Transportation Research Board, Washington, 1997), pp. 8.1–8.47
35. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)

Operation Regimes and Slower-is-Faster-Effect in the Control of Traffic Intersections*

Dirk Helbing and Amin Mazloumian

Abstract The efficiency of traffic flows in urban areas is known to crucially depend on signal operation. Here, elements of signal control are discussed, based on the minimization of overall travel times or vehicle queues. Interestingly, we find different operation regimes, some of which involve a "slower-is-faster effect", where a delayed switching reduces the average travel times. These operation regimes characterize different ways of organizing traffic flows in urban road networks. Besides the optimize-one-phase approach, we discuss the procedure and advantages of optimizing multiple phases as well. To improve the service of vehicle platoons and support the self-organization of "green waves", it is proposed to consider the price of stopping newly arriving vehicles.

1 Introduction

The study of urban traffic flows has attracted the interest of physicists for quite a while (see, e.g., [3, 9, 36, 45]). This includes the issue of traffic light control and the resulting dynamics of vehicle flows [5,7,12,15,37,41,49]. Theoretical investigations in this direction have primarily focussed on single intersections and grid-like street networks, e.g. adaptive control [1, 13, 14] of a single traffic light or coordination of traffic lights in Manhattan-like road networks with unidirectional roads and periodic boundary conditions. Some of the fascination for traffic light control is due to the

*First published in: The European Physical Journal B **70**(2), 257–274, DOI: 10.1140/epjb/e2009-00213-5 (2009), © EDP Sciences, Societa Italiana di Fisica, Springer-Verlag 2009, reproduction with kind permission of The European Physical Journal (EPJ).

D. Helbing (✉) · A. Mazloumian
ETH Zurich, UNO D11, Universitätstr. 41, 8092 Zurich, Switzerland
e-mail: dhelbing@ethz.ch; mseyyed@ethz.ch

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_7,
© Springer-Verlag Berlin Heidelberg 2013

relationship with the synchronization of oscillators [10, 29, 34] and other concepts of self-organization [16, 17, 24, 25, 28, 33, 38, 44].

The efficiency of traffic light control is essential to avoid or at least delay the collapse of traffic flows in traffic networks, particularly in urban areas. It is also crucial for attempts to reduce the fuel consumption and CO_2 emissions of vehicles. Both, delay times and acceleration maneuvers (i.e. the number of stops faced by vehicles)[1] cause additional fuel consumption and additional CO_2 emissions [32]. Within the USA alone, the cost of congestion per year is estimated to be 63.1 billion US$, related with 3.7 billion hours of delays and 8.7 billion liters of “wasted” fuel [43]. Climate change and political goals to reduce CO_2 emissions force us to rethink the design and operation of traffic systems, which contributes about one third to the energy consumption of industrialized countries. On freeways, traffic flows may eventually be improved by automated, locally coordinated driving, based on new sensor technologies and intervehicle communication [31, 42].

But what are options for urban areas? There, traffic lights are used to resolve conflicts of intersecting traffic streams. In this way, they avoid accidents and improve the throughput at moderate or high traffic volumes. For a discussion of the related traffic engineering literature, including the discussion of traffic light coordination and adaptive signal control, see [25, 33] and references therein. In the following, we will focus our attention on some surprising aspects of traffic flow optimization.

1.1 Paradoxical Behavior of Transport Systems

Besides Braess’ paradox (which is related to selfish routing) [4, 40, 48], the slower-is-faster effect is another counter-intuitive effect that seems to occur in many transport networks. It has been found for pedestrian crowds, where a rush of people may delay evacuation [22].

Slower-is-faster effects have fascinated scientists for a long time. Smeed [46], for example, discussed “some circumstances in which vehicles will reach their destinations earlier by starting later”, but Ben-Akiva and de Palma [2] showed that this effect disappeared under realistic assumptions. Moreover, it is known from queuing theory that idle time can decrease the work in process (i.e. basically the queue length) in cyclically operated production systems under certain circumstances, particularly when the variance in the setup times is large [8]. These circumstances, however, do not seem to be very relevant for traffic light control. Nevertheless, there are many examples of slower-is-faster effects in traffic, production, and logistic systems, and it has been suggested that the phenomenon is widespread in networked systems with conflicting flows that are competing for prioritization [21, 27].

[1] For formulas to estimate these quantities as a function of the utilization of the service capacity of roads see [20].

While there are numerical algorithms to exploit this effect systematically to improve the performance of these systems [21], there have been only a few analytical studies of the slower-is-faster effect [23,26,47]. Therefore, we will put a particular focus on the study of conditions leading to this counterintuitive, but practically relevant effect.

Our paper is structured as follows: While Sect. 2 specifies the traffic system investigated in this paper, Sect. 3 discusses the throughput of intersections. Section 4 continues with the problem of minimizing travel times, while Sect. 4.5 discusses the minimization of queue lengths. The challenge in these sections is to come up with a concept that still leads to reasonably simple formulas, allowing one to study the behavior of the proposed signal control analytically. A successful approach in this respect is the "optimize-one-phase approach", which seems justified by the short intervals, over which traffic flows can be anticipated reliably. Among the operation regimes resulting from the optimization process are also some with extended green times, corresponding to a "slower-is-faster effect" (see Sect. 4.4). A further improvement of signal operation is reached by applying multi-phase optimization, when flow constraints are taken into account. As Sect. 5 shows, this approach leads to a variety of plausible operation regimes. A summary and discussion is presented in Sect. 6. Complementary, Appendix 1 will discuss the "price" of stopping vehicles, which is an interesting concept to support moving vehicle platoons (and, thereby, the self-organization of "green waves"). For a more sophisticated, but analytically less accessible approach to the self-organization of coordinated traffic lights and vehicle streams in road networks see [21,33,35].

2 Specification of the Traffic System Under Consideration

In this paper, we will first focus on the study of a single traffic intersection with uniform arrival flows, before we discuss later how to extend our control concept in various ways. Furthermore, for simplicity we will concentrate on the study of a traffic light control with *two* green phases only, which is generalized in Appendix 2. As the traffic organization in parts of Barcelona shows, a two-phase control is *sufficient*, in principle, to reach all points in the road network: Just assume unidirectional flows in all streets with alternating directions. Then, in each phase, traffic either flows straight ahead and/or turns (right or left, depending on the driving direction in the crossing road). Hence, two intersecting unidirectional roads imply two possible traffic phases, which alternate (see Fig. 1).

While the optimization approach discussed in the following can be also applied to time-dependent arrival flows A_1 and A_2 per lane, when numerical solution methods are applied, for the sake of analytical tractability and closed formulas we will focus here on the case of constant flows over the short time periods involved in our optimization. I_j will represent the number of lanes of road section j, and it will be assumed that vehicles passing a green light can freely enter the respective downstream road section. Analogously to [18,20,28], the departure flows

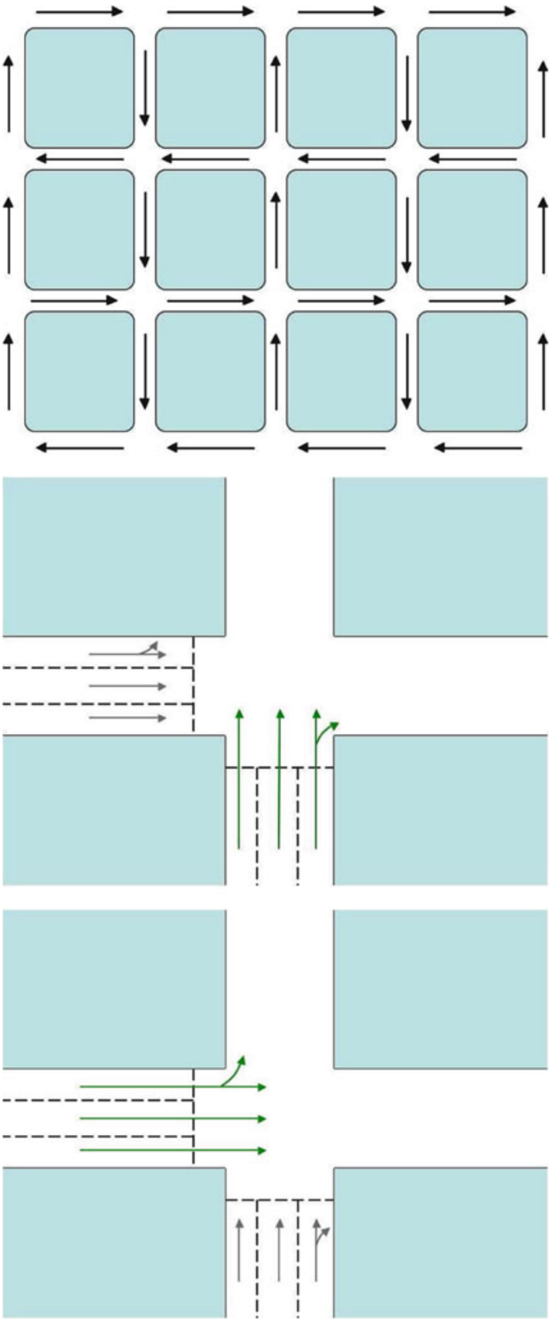


Fig. 1 *Top*: Schematic illustration of the unidirectional street layout in the center of Barcelona. *Center* and *bottom*: illustration of the two traffic phases, during which vehicles can move straight ahead or turn (either *right* or *left*, depending on the direction of the crossing road)

$\gamma_j(t)O_j(t)$ (as long as the traffic flows are not obstructed by the downstream traffic conditions) are given by the possible outflows $O_j(t)$ (which vary with time t), multiplied with the permeabilities $\gamma_j(t)$. The latter reflect the states of the traffic lights. During amber and red time periods, the permeabilities $\gamma_j(t)$ are zero, as there is no outflow, while $\gamma_j(t) = 1$ during green phases. Note that the departure flows $\gamma_j(t)O_j(t)$ may split up into a straight and a turning flow after the traffic light, but for our further considerations, this is not relevant. The possible outflows $O_j(t)$ are determined by the equation

$$O_j(t) = \begin{cases} \widehat{Q}_j & \text{if} \Delta N_j(t) > 0, \\ A_j(t - \mathscr{T}_j^0) & \text{otherwise.} \end{cases} \tag{1}$$

Herein, $\widehat{Q}_j$ is the service rate per lane during the green phase as long as there is a finite number $\Delta N_j(t) > 0$ of delayed vehicles behind the traffic light (i.e. $\widehat{Q}_j$ corresponds to the characteristic outflow from congested traffic). $A_j(t)$ represents the time-dependent arrival rate of vehicles per lane and $A_j(t - \mathscr{T}_j^0)$ the rate of vehicles arriving at the traffic light under free flow conditions, where $\mathscr{T}_j^0$ denotes the free travel time needed to pass road section j. In case of constant arrival rates A_j, the dependence on the time point t and the delay by the free travel time $\mathscr{T}_j^0$ can be dropped.

In the following, we will use some additional variables and parameters: T_j shall denote the minimum green time, after which the vehicle queue in road section j is fully dissolved (i.e. after which $\Delta N_j = 0$ and $O_j = A_j$). In contrast, ΔT_j will stand for the actual green time period. Consequently,

$$\Delta t_j = \Delta T_j - T_j \tag{2}$$

(if greater than zero) represents the excess green time, during which we have a free vehicle flow with $\gamma_j(t)O_j(t) = A_j$. τ_j shall be the setup time *before* the green phase ΔT_j for road section j. For illustrative reasons, it is also called the "amber time (period)", although it is usually somewhat longer than that. The sum

$$T_{\text{cyc}} = \tau_1 + \Delta T_1 + \tau_2 + \Delta T_2 \tag{3}$$

is normally called the cycle time. Note, however, that we do not need to assume *periodic* operation. Within the framework of our model assumptions, we may consider *stepwise* constant flows. That is, the arrival flows may vary from one cycle (or even one green time period) to the next. Under such conditions, each green phase is adjusted to the changing traffic situation.

Finally note that we do not consider pedestrian flows in this paper. In order to take them into account, one would have to consider additional traffic phases for the service of pedestrians. Alternatively, one could select the setup times τ_j for vehicles so large that they cover the amber time for vehicles plus a sufficient time for pedestrians to cross the road.

3 Consideration of Traffic Flows

The art of traffic control is to manipulate the permeabilities $\gamma_j(t)$ in a way that optimizes a given goal function. In fact, when the traffic volume is high enough, an oscillatory service corresponding to the operation of a traffic light can increase the effective intersection capacity as compared to the application of a first-come-first-serve rule for arriving vehicles [25, 28]: While the red and amber lights (corresponding to $\gamma_j(t) = 0$) cause vehicles to queue up and wait, this implies a high flow rate and an efficient service of vehicles when the traffic light turns green s(i.e. $\gamma_j(t) = 1$).

One natural concept of traffic flow optimization would be to maximize the average overall throughput. This is measured by the function

$$G_{\mathrm{t}}(t) = \frac{1}{t} \sum_j \int_0^t dt' \, \gamma_j(t') O_j(t') \, . \tag{4}$$

Due to (1), $G_{\mathrm{t}}(t)$ depends not only on the outflows $O_j(t)$, but also on the inflows $A_j(t)$ to the system. This makes $G_{\mathrm{t}}(t)$ basically dependent on the time-dependent origin-destination matrices of vehicle flows.

The numbers of vehicles accumulating during the red and amber time periods are

$$I_1 \Delta N_1^{\max} = I_1 A_1 (\tau_2 + \Delta T_2 + \tau_1) \tag{5}$$

and

$$I_2 \Delta N_2^{\max} = I_2 A_2 (\tau_1 + \Delta T_1 + \tau_2) \, , \tag{6}$$

where $\Delta N_j^{\max}$ represents the maximum number of delayed vehicles per lane in road section j, if the vehicle queue in it has been fully cleared before. I_j is the number of lanes. As the service rate of queued vehicles during the green time ΔT_j is $\widehat{Q}_j$, and A_j is the arrival rate of additional vehicles at the end of the queue, the minimum green time required to dissolve the queue is given by

$$T_j = \frac{\Delta N_j^{\max}}{\widehat{Q}_j - A_j} \, . \tag{7}$$

From (5) to (7) we obtain

$$T_1 = \frac{A_1}{\widehat{Q}_1 - A_1} \left(\tau_2 + \Delta T_2 + \tau_1 \right) . \tag{8}$$

Assuming $\Delta T_j = T_j$ (i.e. no excess green times) and inserting (7) yields

$$T_1 = \frac{A_1}{\widehat{Q}_1 - A_1}\left(\tau_2 + \frac{A_2(\tau_1 + T_1 + \tau_2)}{\widehat{Q}_2 - A_2} + \tau_1\right) \tag{9}$$

for the clearing time T_1, or

$$T_1 = (\tau_1 + \tau_2)\frac{\frac{A_1}{\widehat{Q}_1 - A_1}\left(1 + \frac{A_2}{\widehat{Q}_2 - A_2}\right)}{1 - \frac{A_1 A_2}{(\widehat{Q}_1 - A_1)(\widehat{Q}_2 - A_2)}} \,. \tag{10}$$

With the analogous formula for T_2 we can determine the related cycle time, if the traffic light turns red immediately when all queued vehicles have been served. After a few intermediate mathematical steps, we finally get

$$T^{\text{cyc}} = \tau_1 + T_1 + \tau_2 + T_2 = \frac{\tau_1 + \tau_2}{1 - A_1/\widehat{Q}_1 - A_2/\widehat{Q}_2} \,. \tag{11}$$

Moreover, one can show [20]

$$T_j = \frac{A_j}{\widehat{Q}_j} T^{\text{cyc}} \,. \tag{12}$$

We can see that the cycle time and the clearing times T_j diverge in the limit

$$\frac{A_1}{\widehat{Q}_1} + \frac{A_2}{\widehat{Q}_2} \to 1 \,. \tag{13}$$

If this expression (11) becomes negative, the vehicle queues in one or both ingoing road sections are growing larger and larger in time, as the intersection does not have enough capacity to serve both arrival flows. See [20] for a discussion of this case.

Note that (11) determines the *smallest* cycle time that allows to serve all queued vehicles within the green time periods. Let us study now the effect of *extending* the green time periods ΔT_j beyond T_j: The average throughput of the intersection is given by the overall flow of vehicles during one cycle time $T_{\text{cyc}} = \tau_1 + \Delta T_1 + \tau_2 + \Delta T_2$. During that time period, a total number $(I_1 A_1 + I_2 A_2)T_{\text{cyc}}$ of vehicles is arriving in the two considered road sections. If all arriving vehicles are served during the cycle time T_{cyc}, the average throughput is

$$G_{\text{t}} = \frac{(I_1 A_1 + I_2 A_2)T_{\text{cyc}}}{T^{\text{cyc}}} = I_1 A_1 + I_2 A_2 \,. \tag{14}$$

Therefore, in the case where we do not have an accumulation of vehicles over time, which requires sufficient green times ($\Delta T_j > T_j$) and a sufficient resulting service capacity

$$\frac{I_1 \widehat{Q}_1 \, \Delta T_1 + I_2 \widehat{Q}_2 \, \Delta T_2}{T_{\text{cyc}}} \ge I_1 A_1 + I_2 A_2 \,, \tag{15}$$

the throughput is determined by the sum $I_1 A_1 + I_2 A_2$ of the overall arrival flows. Consequently, excess green times $\Delta t_j = \Delta T_j - T_j > 0$ do not lead to smaller or larger intersection throughputs. But under what conditions should a green phase be extended, if at all? This shall be addressed in the next sections.

4 Travel-Time-Oriented Signal Operation

Rather than on a consideration of the flow, we will now focus on the *cumulative waiting time*

$$F(t) = \sum_j I_j \int_0^t dt' \int_0^{t'} dt''[A_j - \gamma_j(t'')O_j(t'')] \tag{16}$$

and minimize its average growth over a time period t to be defined later. This corresponds to a minimization of the function

$$\begin{aligned} G(t) &= \frac{1}{t} \sum_j I_j \int_0^t dt' \Delta N_j(t') \\ &= \frac{1}{t} \sum_j I_j \int_0^t dt' \int_0^{t'} dt''[A_j - \gamma_j(t'')O_j(t'')] \,, \end{aligned} \tag{17}$$

which quantifies the time average of the overall delay time. The term on the right-hand side describes the increase of the overall waiting time proportionally to the number ΔN_j of delayed cars, which is given by the integral over the difference between the arrival and departure flows [20, 28].

Note that the formula (17) makes an implicit simplification by assuming that delays occur only in the vehicle queues behind traffic lights, while no delays accumulate under uncongested flow conditions. This assumes a triangular flow-density diagram, which, however, seems to be sufficiently justified for urban traffic flows [18, 28]. Moreover, while approaching a vehicle queue, it usually does not matter, when vehicles travel more slowly than the speed limit allows: If they would travel faster, they would be queued earlier, i.e. the delay would stay the same. In other words, most of the time it is irrelevant, whether vehicles lose their time in the vehicle queue or by decelerating before.

4.1 The Optimize-One-Phase Approach

When minimizing the goal function $G(t)$, it is essential upto what time t we extend the integral. In principle, it is possible to integrate over a full cycle or even many cycles of traffic operation, but the resulting formulas do not provide an intuitive understanding anymore. We will, therefore, focus on the optimization of a single phase, with full amber time periods τ_j in the beginning and τ_{j+1} at the end. This turns out to result in explicit and plausible formulas, while some other approaches we have tried, did not result in well interpretable results. Besides this practical aspect, when analytical results shall be obtained, the specification $t = \tau_1 + \Delta T_j + \tau_2$ chosen in the following makes sense: It "charges" the switching-related inefficiencies to the road that "wants" to be served. The switching of a traffic light should lead to a temporary *in*crease in traffic performance. After completion of each green phase, the travel time optimization is repeated, so that one can compose the traffic light schedule as a sequence of optimized single phases (see Appendix 2 for details).

In Sect. 5, we will show that a multi-phase optimization yields better results, but requires a higher degree of sophistication. The treatment of situations with varying or pulsed traffic flows is even more difficult and can usually be solved only numerically. This issue is addressed in [33].

In our calculations, we will assume that the green time for road Sect. 2 lasted for a time period ΔT_2 and ended at time $t = 0$. That is, we have now to determine the optimal duration ΔT_1 of the green phase for road Sect. 1 after an intermediate amber time period τ_1. For this, we minimize the function

$$G_1(\tau_1 + \Delta T_1 + \tau_2) = \frac{F_1(\tau_1 + \Delta T_1 + \tau_2)}{\tau_1 + \Delta T_1 + \tau_2}\,, \tag{18}$$

where the subscript "1" of G and F refers to road Sect. 1, for which the green phase is determined. Assuming a step-wise constant outflow with $\gamma_j O_j = \widehat{Q}_j$, if $\Delta N_j > 0$, but $\gamma_j O_j = A_j$, if $\Delta N_j = 0$, and $\gamma_j O_j = 0$, if $\gamma_j = 0$, the integral over t'' results in a stepwise linear function, and the function $F_1(t)$ is characterized by quadradic dependencies. We will distinguish two cases: (a) The green time is potentially terminated *before* all queued vehicles have been served (i.e. $\Delta T_i \le T_i$), or (b) it is potentially *extended* (i.e. $\Delta T_i \ge T_i$). Let us start with the first case.

(a) **No excess green time** ($\Delta T_1 \le T_1$): In this case, $A_2(t'') - \gamma_2(t'')O_2(t'') = A_2$ for $0 \le t'' \le \tau_1 + \Delta T_1 + \tau_2$, i.e. over the period ΔT_1 of the green time for road Sect. 1 and the amber time periods τ_j and τ_{j+1} before and after it. In addition,

$$A_1 - \gamma_1(t'')O_1(t'') = \begin{cases} A_1 & \text{if } 0 \le t'' < \tau_1, \\ A_1 - \widehat{Q}_1 & \text{if } \tau_1 \le t'' < \tau_1 + \Delta T_1, \\ A_1 & \text{otherwise.} \end{cases} \tag{19}$$

Using the abbreviation

$$\Delta N_1^{\max} = \Delta N_1(\tau_1) = \Delta N_1(0) + A_1\tau_1 \,, \tag{20}$$

we get

$$\begin{aligned}
F_1^{\mathrm{a}}(\tau_1 + \Delta T_1 + \tau_2) = I_1\bigg\{ & \Delta N_1(0)\tau_1 + A_1\frac{{\tau_1}^2}{2} \\
& + \Delta N_1^{\max}\Delta T_1 - (\widehat{Q}_1 - A_1)\frac{{\Delta T_1}^2}{2} \\
& + [\Delta N_1^{\max} - (\widehat{Q}_1 - A_1)\Delta T_1]\tau_2 + A_1\frac{{\tau_2}^2}{2}\bigg\} \\
& + I_2\bigg[\Delta N_2(0)(\tau_1 + \Delta T_1 + \tau_2) + A_2\frac{(\tau_1 + \Delta T_1 + \tau_2)^2}{2}\bigg] \\
= I_1\bigg[& \Delta N_1(0)(\tau_1 + \Delta T_1 + \tau_2) + \frac{A_1}{2}(\tau_1 + \Delta T_1 + \tau_2)^2 \\
& - \frac{\widehat{Q}_1}{2}\Delta T_1(\Delta T_1 + 2\tau_2)\bigg] \\
& + I_2\bigg[\Delta N_2(0)(\tau_1 + \Delta T_1 + \tau_2) + A_2\frac{(\tau_1 + \Delta T_1 + \tau_2)^2}{2}\bigg],
\end{aligned} \tag{21}$$

where the superscript "a" refers to case (a). Dividing the above function by $(\tau_1 + \Delta T_1 + \tau_2)$ and making the plausible assumption $\tau_1 = \tau_2$ of equal amber time periods for simplicity, we gain

$$\begin{aligned}
G_1^{\mathrm{a}}(\tau_1 + \Delta T_1 + \tau_2) = I_1\bigg[& \Delta N_1(0) + \widehat{Q}_1\tau_2 \\
& - (\widehat{Q}_1 - A_1)\frac{\tau_1 + \Delta T_1 + \tau_2}{2}\bigg] \\
& + I_2\bigg[\Delta N_2(0) + A_2\frac{\tau_1 + \Delta T_1 + \tau_2}{2}\bigg].
\end{aligned} \tag{22}$$

If $I_1(\widehat{Q}_1 - A_1) < I_2A_2$, i.e. when the number of queued vehicles in road section2 grows faster than it can be reduced in road Sect. 1, the minimum of this function is reached for $\Delta T_1 = 0$, corresponding to a situation where it is not favorable to turn green for section $j = 1$. For

$$I_1(\widehat{Q}_1 - A_1) > I_2A_2 \,, \tag{23}$$

the value of G_1^{a} goes down with growing values of ΔT_1, and the minimum is reached for a value $\Delta T_1 \geq T_1$.

(b) **Potential green time extension** ($\Delta T_1 \geq T_1$): Let us assume that we (possibly) have an excess green time, i.e. $\Delta t_i = \Delta T_i - T_i \geq 0$. In this case,

$$A_1 - \gamma_1(t'')O_1(t'') = \begin{cases} A_1 & \text{if } 0 \leq t'' < \tau_1, \\ A_1 - \widehat{Q}_1 & \text{if} \tau_1 \leq t'' < \tau_1 + T_1, \\ A_1 & \text{if } t'' \geq \tau_1 + \Delta T_1, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Considering that now, $\Delta N_1(t') = 0$ for $\tau_1 + T_1 \leq t' < \tau_1 + \Delta T_1$, and introducing the *clearing time*

$$T_1 = \frac{\Delta N_1^{\max}}{\widehat{Q}_1 - A_1} = \frac{\Delta N_1(0) + A_1\tau_1}{\widehat{Q}_1 - A_1}\,, \tag{25}$$

we obtain

$$\begin{aligned} F_1^{\mathrm{b}}(\tau_1 + \Delta T_1 + \tau_2) &= I_1\Big[\Delta N_1(0)\tau_1 + A_1\frac{{\tau_1}^2}{2} + \Delta N_1^{\max}T_1 \\ &\quad -(\widehat{Q}_1 - A_1)\frac{{T_1}^2}{2} + A_1\frac{{\tau_2}^2}{2}\Big] \\ &\quad + I_2\left[\Delta N_2(0)(\tau_1 + \Delta T_1 + \tau_2) + A_2\frac{(\tau_1 + \Delta T_1 + \tau_2)^2}{2}\right] \\ &= I_1\left[\Delta N_1^{\max}\tau_1 + \frac{A_1}{2}({\tau_2}^2 - {\tau_1}^2) + \frac{(\Delta N_1^{\max})^2}{2(\widehat{Q}_1 - A_1)}\right] \\ &\quad + I_2\left[\Delta N_2(0)(\tau_1 + \Delta T_1 + \tau_2) + A_2\frac{(\tau_1 + \Delta T_1 + \tau_2)^2}{2}\right] \end{aligned} \tag{26}$$

Assuming again $\tau_1 = \tau_2$ for simplicity, introducing the abbreviation

$$E_1 = \Delta N_1^{\max}\tau_1 + \frac{(\Delta N_1^{\max})^2}{2(\widehat{Q}_1 - A_1)}\,, \tag{27}$$

and dividing (26) by $(\tau_1 + \Delta T_1 + \tau_2)$ yields

$$\begin{aligned} G_1^{\mathrm{b}}(\tau_1 + \Delta T_1 + \tau_2) &= \frac{I_1 E_1}{\tau_1 + \Delta T_1 + \tau_2} \\ &\quad + I_2\left[\Delta N_2(0) + A_2\frac{\tau_1 + \Delta T_1 + \tau_2}{2}\right]. \end{aligned} \tag{28}$$

This expression shall be minimized under the constraint $\Delta T_1 \geq T_1$. In order to determine the minimum, we set the derivative with respect to ΔT_1 to zero and get

$$0 = \frac{dG_1^{\mathrm{b}}(\tau_1 + \Delta T_1 + \tau_2)}{d\,\Delta T_1} = -\frac{I_1 E_1}{(\tau_1 + \Delta T_1 + \tau_2)^2} + \frac{I_2 A_2}{2}\,. \tag{29}$$

The minimum is located at

$$(\tau_1 + \Delta T_1 + \tau_2)^2 = \frac{2 I_1 E_1}{I_2 A_2}\,, \tag{30}$$

if $\Delta T_1 \geq T_1$. Considering (25), $\Delta T_1 \geq T_1$ implies

$$(\tau_1 + \Delta T_1 + \tau_2)^2 \geq \left(\tau_1 + \frac{\Delta N_1^{\max}}{\widehat{Q}_1 - A_1} + \tau_2 \right)^2 . \tag{31}$$

With (30) this leads to the condition

$$\begin{aligned} &\frac{(\Delta N_1^{\max})^2}{\widehat{Q}_1 - A_1} \left(\frac{I_1}{I_2 A_2} - \frac{1}{\widehat{Q}_1 - A_1} \right) \\ &+ 2\Delta N_1^{\max} \left(\frac{I_1 \tau_1}{I_2 A_2} - \frac{\tau_1 + \tau_2}{\widehat{Q}_1 - A_1} \right) \geq (\tau_1 + \tau_2)^2 . \end{aligned} \tag{32}$$

If inequality (32) is not fulfilled, we must have $\Delta T_1 < T_1$.

For completeness, we note that

$$G_1^{\mathrm{a}}(\tau_1 + T_1 + \tau_2) = G_1^{\mathrm{b}}(\tau_1 + T_1 + \tau_2)\,, \tag{33}$$

i.e. the goal function G_1 is continuous in $\Delta T_1 = T_1$, while it must not be smooth. Moreover, $\Delta N_1(0) = A_1(\tau_2 + \Delta T_2)$ and $\Delta N_2(0) = 0$, if the vehicle queues have been fully cleared before the traffic light is switched. The case where the queue is not fully dissolved is treated in [20].

4.2 Transformation to Dimensionless Variables and Parameters

For an analysis of the system behavior, it is useful to transform variables and parameters to dimensionless units. Such dimensionless units are, for example, the capacity utilizations

$$u_i = \frac{A_i}{\widehat{Q}_i} \tag{34}$$

of the road sections i and the relative size

$$\kappa = \frac{I_1 A_1}{I_2 A_2} = \frac{I_1 u_1 \widehat{Q}_1}{I_2 u_2 \widehat{Q}_2} = \frac{u_1}{u_2} K \tag{35}$$

of the arrival flows, where

$$K = \frac{I_1 \widehat{Q}_1}{I_2 \widehat{Q}_2} \,. \tag{36}$$

Furthermore, we may scale the green times ΔT_i by the sum of amber time periods $\tau_1 + \tau_2$, which defines the dimensionless green times

$$\sigma_i = \frac{\Delta T_i}{\tau_1 + \tau_2} \tag{37}$$

and the dimensionless clearing times

$$\hat{\sigma}_j = \frac{T_j}{\tau_1 + \tau_2} = \frac{\Delta N_1^{\max}}{(1-u_1)\widehat{Q}_i(\tau_1 + \tau_2)} \,. \tag{38}$$

In order to express the previous relationships exclusively by these quantities, we must consider that a number $A_1(\tau_2 + \Delta T_2 + \tau_1)$ of vehicles per lane accumulates during the time period $(\tau_2 + \Delta T_2 + \tau_1)$, in which the vehicle flow on road Sect. 1 is not served. With (20) this implies

$$\Delta N_1^{\max} = \Delta N_1(0) + A_1 \tau_1 = A_1(\tau_2 + \Delta T_2 + \tau_1) \,, \tag{39}$$

if the vehicle queue in road Sect. 1 has been fully cleared during the previous green time. Then, we have

$$\frac{\Delta N_1^{\max}}{\tau_1 + \tau_2} = A_1(1 + \sigma_2) \,, \tag{40}$$

and from (27) and (22) we get

$$\frac{2E_1}{(\tau_1 + \tau_2)^2} = A_1(1+\sigma_2)\frac{2\tau_1}{\tau_1 + \tau_2} + \frac{(A_1)^2(1+\sigma_2)^2}{\widehat{Q}_1 - A_1} \,. \tag{41}$$

With $A_1 = u_1 \widehat{Q}_1$ and $\tau_1 = \tau_2$, (30) belonging to the case of extended green time for road Sect. 1 can be written as

$$(1+\sigma_1)^2 = [1 + \tilde{\sigma}_1(\sigma_2)]^2 = \kappa\left[(1+\sigma_2) + \frac{u_1}{1-u_1}(1+\sigma_2)^2\right] . \tag{42}$$

The solution of this equation defines the relationship $\tilde{\sigma}_1(\sigma_2)$ for the optimal scaled green time period σ_1 as a function of σ_2, if the green time for road Sect. 1 is extended. Moreover, in dimensionless variables, the condition (32) for green time extension becomes

$$\frac{u_1}{1-u_1}(1+\sigma_2)^2\left(\kappa-\frac{u_1}{1-u_1}\right)+(1+\sigma_2)\left(\kappa-\frac{2u_1}{1-u_1}\right)\geq 1 \tag{43}$$

or

$$\left[\frac{u_1(1+\sigma_2)^2}{1-u_1}+(1+\sigma_2)\right]\left(\kappa-\frac{u_1}{1-u_1}\right)\geq\frac{1+u_1\sigma_2}{1-u_1}\,. \tag{44}$$

However, we can check for green time extension also in a different way, since the extension condition $\Delta T_1 > T_1$ can be written as $\tilde{\sigma}_1 > \hat{\sigma}_1$. Using (25), (38) and (39), the dimensionless green time σ_1 for the case of no green time extension may be presented as

$$\sigma_1=\hat{\sigma}_1(\sigma_2)=\frac{A_1(1+\sigma_2)}{\widehat{Q}_1-A_1}=\frac{u_1(1+\sigma_2)}{1-u_1} \tag{45}$$

or

$$1+\sigma_1=\frac{1+u_1\sigma_2}{1-u_1}\,. \tag{46}$$

Moreover, from $\sigma_1=\hat{\sigma}_1(\sigma_2)$ follows

$$\frac{\sigma_1}{1+\sigma_1+\sigma_2}=\frac{u_1(1+\sigma_2)}{(1-u_1)\Big[1+\frac{u_1}{1-u_1}(1+\sigma_2)+\sigma_2\Big]}=u_1\,. \tag{47}$$

That is, in the case where road Sect. 1 is completely cleared, but there is no green time extension, the green time fraction

$$\frac{\Delta T_1}{T_{\text{cyc}}}=\frac{\sigma_1}{1+\sigma_1+\sigma_2} \tag{48}$$

agrees with the utilization u_1. Moreover, one can show

$$\frac{\partial}{\partial\sigma_1}\left(\frac{\sigma_1}{1+\sigma_1+\sigma_2}\right)=\frac{1+\sigma_2}{(1+\sigma_1+\sigma_2)^2}>0\,. \tag{49}$$

Therefore, $\sigma_1 > \hat{\sigma}_1$ implies a green time fraction greater than u_1, and we have excess green time for road Sect. 1, if

$$\frac{\tilde{\sigma}_1(\sigma_2)}{1+\tilde{\sigma}_1(\sigma_2)+\sigma_2}>u_1\,. \tag{50}$$

An analogous condition must be fulfilled, if excess green times on road Sect. 2 shall be optimal. It reads

$$\frac{\tilde{\sigma}_2(\sigma_1)}{1+\sigma_1+\tilde{\sigma}_2(\sigma_1)} > u_2\,, \tag{51}$$

where

$$[1+\tilde{\sigma}_2(\sigma_1)]^2 = \frac{1}{\kappa}\left[(1+\sigma_1) + \frac{u_2}{1-u_2}(1+\sigma_1)^2\right], \tag{52}$$

which has been gained by interchanging indices 1 and 2 and replacing κ by $1/\kappa$ in (42).

4.3 Control Strategies and Slower-is-Faster Effect

Based on the results of Sect. 4.1 and the scaled formulas of Sect. 4.2, we can now formulate control strategies for a single traffic light within the optimize-one-phase approach:

(i) **Terminate the green light for road Sect. 1 immediately**, corresponding to $\sigma_1 = 0$, if condition (23) is violated, i.e. if

$$1 - u_1 \le \frac{u_2}{K} \tag{53}$$

is fulfilled. To obtain the dimensionless form of this inequality, we have considered $A_j = u_j I_j \widehat{Q}_j$ and (36). In case (i), travel time optimization for one phase advises against turning green for road Sect. 1. Of course, in reality, drivers cannot be stopped forever. Either, one would have to give them a short green phase after a maximum tolerable time period, or at least one would have to allow vehicles to turn on red, i.e. to merge the crossing flow, whenever there is a large enough gap between two successive vehicles. Alternatively, one may apply an optimize-multiple-phases approach, see Sect. 5. It implies a service of side roads even when the intersection capacity is insufficient to satisfy all inflows completely.

(ii) **Terminate the green phase for road Sect. 1, when the vehicle queue is completely resolved**, if conditions (53) and (44) are violated. In this case, the scaled green time σ_1 is given by (45).

(iii) **Extend the green times for road Sect. 1** in accordance with formula (42), if the condition (44) is fulfilled. The recommended delay in the switching time constitutes a *slower-is-faster effect*. In this situation, it takes some additional time to accumulate enough vehicles on road Sect. 2 to guarantee an efficient service in view of the inefficiencies caused by the setup times τ_j.

In Fig. 2, operation regime (i) is indicated in white and operation regime (iii) in red, while operation regime (ii) is shown in green, if road Sect. 2 is served, otherwise in orange.

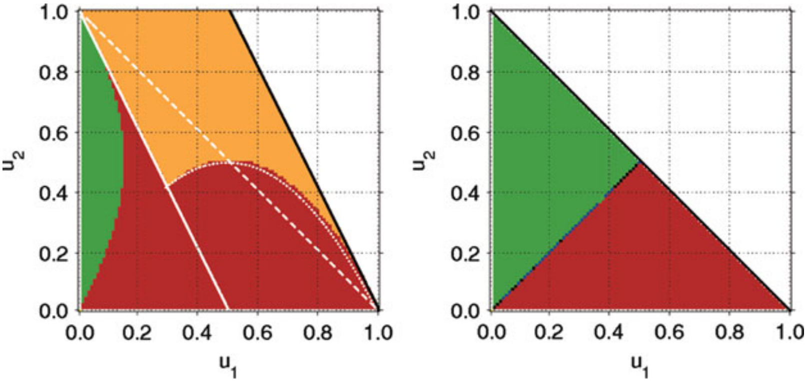


Fig. 2 Operation regimes of (periodic) signal control for $K = 1$ (*left*) and $K = 3$ (*right*) as a function of the utilizations u_j of both roads j according to the one-phase travel time optimization approach. For each combination of u_1 and u_2, the operation regime has been determined after convergence of the signal control procedure described in Appendix 2. The *separating lines* are in good agreement with our analytical calculations. For example, the *solid falling lines* are given by (54), while the *dotted parabolic line* in the *right illustration* corresponds to $u_2 = Ku_1(1 - u_1)$ and results by equalizing (42) with the square of (46), assuming $\sigma_2 = 0$ (i.e. no service of road section 2). The different operation regimes are characterized as follows: In the *green triangular* or *parabolic area* to the left of both illustrations, where the utilization u_1 of road Sect. 1 is sufficiently small, the service of road Sect. 2 is extended. In the adjacent *red area* below the *white area* (*left*) or the *solid line* (*right*), road Sect. 2 is just cleared, while above the *separating line* $u_2 = 1-Ku_1$, road Sect. 2 is not served at all. Road Sect. 1, in contrast, gets just enough *green* time to clear the vehicle queue in the *green area* (and the *orange area* towards the *top* of the *right illustration*), while it gets extended *green* time in the *red area* towards the *bottom*, where the utilization u_2 of road Sect. 2 is sufficiently small. In the *white area* given by $u_2 > K(1 - u_1)$, road Sect. 1 gets no *green* time anymore. Between the *dashed* and the *solid white lines*, road Sect. 2 is not served, although there would be enough capacity to satisfy the vehicle flows in both roads. Improved operation regimes are presented in Fig. 4

4.4 Operation Regimes for Periodic Operation

In the previous section, we have determined the optimal green time period σ_1 for road Sect. 1, assuming that the last green time period σ_2 for road Sect. 2 and $N_1(0)$ were given. Of course, σ_1 will then determine σ_2, etc. If the utilizations u_j are constant and not too high, the sequence of green phases converges towards a *periodic* signal operation (see Fig. 3). It will be studied in the following. While the formulas for the determination of σ_1 were derived in Sect. 4.2, the corresponding formulas for σ_2 can be obtained by interchanging the indices 1 and 2 and replacing κ by $1/\kappa$ in all formulas. In principle, there could be the following cases, if we restrict ourselves to reasonable solutions with $\sigma_j \geq 0$:

(0) According to travel time minimization, one or both road sections should not be served, if (53) is fulfilled for one or both of the road sections. This case occurs if

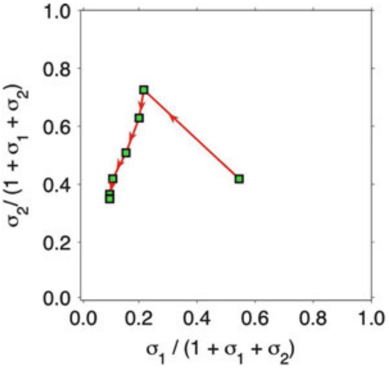


Fig. 3 *Green* time fraction $\sigma_2/(1+\sigma_1+\sigma_2)$ for road Sect. 2 vs. *green* time fraction $\sigma_1/(1+\sigma_1+\sigma_2)$ for road Sect. 1, if we apply the signal control algorithm described in Appendix 2 to a randomly chosen initial queue length $\Delta N_1(0)$ in road Sect. 1 and $K = 2$ (i.e. road Sect. 1 has two times as many lanes as road Sect. 2). One can clearly see that the *green* time fractions quickly converge towards values that do not change anymore over time. The solution corresponds to periodic signal operation

$$1 - u_1 - \frac{u_2}{K} \le 0 \quad \text{or} \quad 1 - Ku_1 - u_2 \le 0 \tag{54}$$

(see the area above the white solid line in the right illustration of Fig. 2). According to this, service should focus on the main flow, while crossing flows should be suppressed, thereby enforcing a re-routing of traffic streams when this would be favorable to minimize travel times. Of course, in such situations vehicles should still be allowed to turn on red and to merge the crossing flow, when vehicle gaps are large enough.

(1) Both green time periods are terminated as soon as the respective vehicle queues are fully dissolved. In this case, we should have the relationships $\sigma_1 = \hat{\sigma}_1(\sigma_2)$ and $\sigma_2 = \hat{\sigma}_2(\sigma_1)$, where $\hat{\sigma}_j$ is defined in (45). After a few steps, the condition $\sigma_1 = \hat{\sigma}_1(\hat{\sigma}_2(\sigma_1))$ implies

$$\sigma_j = \hat{\sigma}_j = \frac{u_j}{1 - u_1 - u_2} \tag{55}$$

and

$$\frac{\sigma_j}{1 + \sigma_1 + \sigma_2} = u_j \, . \tag{56}$$

According to (56), the green time fraction of each road section in case (1) should be proportional to the respective utilization u_j of the flow capacity.

(2) Road Sect. 2 gets an excess green time, while the green phase of road Sect. 1 ends after the dissolution of the vehicle queue (see green area in Fig. 2). In this case we should have $\sigma_1 = \hat{\sigma}_1(\tilde{\sigma}_2(\sigma_1))$, where $(1+\tilde{\sigma}_2)$ is defined by formula (52). This gives

$$\sigma_1 = \frac{u_1}{1-u_1}\sqrt{\frac{1}{\kappa}\Big((1+\sigma_1) + \frac{u_2}{1-u_2}(1+\sigma_1)^2\Big)}\,, \tag{57}$$

which eventually leads to a quadratic equation for σ_1, namely

$$\begin{aligned}&\big[u_1{u_2}^2 - K(1-u_1)^2(1-u_2)\big]{\sigma_1}^2 \\ &+u_1u_2(1+u_2)\sigma_1 + u_1u_2 = 0\,.\end{aligned} \tag{58}$$

To determine σ_2, we can either use the relationship $\sigma_2 = \tilde{\sigma}_2(\sigma_1)$ or invert the formula $\sigma_1 = \hat{\sigma}_1(\sigma_2)$. Doing the latter, (45) gives

$$\sigma_2 = \frac{1-u_1}{u_1}\sigma_1 - 1\,. \tag{59}$$

According to (47) and (51), the occurrence of case (2) requires that the resulting solution satisfies

$$\frac{\sigma_1}{1+\sigma_1+\sigma_2} = u_1 \quad \text{and} \quad \frac{\sigma_2}{1+\sigma_1+\sigma_2} > u_2\,. \tag{60}$$

(3) Road Sect. 1 gets an excess green time, while the green phase of road Sect. 2 ends after the dissolution of the vehicle queue (see red area in Fig. 2). The formulas for this case are obtained from the ones of case (2) by interchanging the indices 1 and 2 and replacing κ by $1/\kappa$.

(4) Both road sections get excess green time periods. This case would correspond to $\sigma_1 = \tilde{\sigma}_1(\tilde{\sigma}_2(\sigma_1))$, and the solutions should fulfil

$$\frac{\sigma_1}{1+\sigma_1+\sigma_2} > u_1 \quad \text{and} \quad \frac{\sigma_2}{1+\sigma_1+\sigma_2} > u_2\,. \tag{61}$$

According to numerical results (see Fig. 2), cases (0), (2), and (3) do all exist, while the conditions for cases (1) and (4) are not fulfilled. Note, however, that small vehicle flows should better be treated as discrete or pulsed rather than continuous flows, in order to reflect the arrival of single vehicles (see [19] for their possible treatment within a continuous flow framework). In other words, for rare vehicle arrivals, we either have $u_1 > 0$ and $u_2 = 0$, or we have $u_2 > 0$ and $u_1 = 0$. Hence, the case of small utilizations u_j will effectively imply green time extensions for both road sections due to the discreteness of the flow, and it allows single vehicles to pass the traffic light without previously stopping at the red light.

Summarizing the above, one-phase optimization provides extra green times for road sections, as long as both of them are fully served. While in one road section, this slower-is-faster effect allows some vehicles to pass the traffic light without stopping, in the other road section it causes the formation of a longer vehicle queue, which supports an efficient service of a substantial number of vehicles after the traffic light turns green. In this connection, it is useful to remember that switching is costly due to the amber times, which are “lost” service times.

4.5 Minimization of Vehicle Queues

We have seen that travel time minimization implies the possibility of case (0), where one of the road sections (the side road) in not being served. This case should not occur as long as the intersection capacity is not fully used. According to (13) and (34), the intersection capacity is sufficient, if

$$u_1 + u_2 \leq 1 \tag{62}$$

As the inequalities (54) and (62) do not agree, conditions may occur, where the vehicle queue in one road section (a side road) continuously increases, even though the intersection capacity would allow to serve both flows (see the orange and red areas above the dashed white line in Fig. 2). This can result in an "unstable" signal control scheme, which causes undesired spillover effects and calls for a suitable stabilization strategy [33]. As we will see in the following, this problem can be overcome by minimizing vehicle queues rather than travel times.

Conditions (54) and (62) agree, if $K = 1$, particularly when $I_1 = I_2$ and $\widehat{Q}_1 = \widehat{Q}_2$. Therefore, let us assume this case in the following, corresponding to

$$\kappa = \frac{I_1 A_1}{I_2 A_2} = \frac{I_1 u_1 \widehat{Q}_1}{I_2 u_2 \widehat{Q}_2} = \frac{u_1}{u_2} \,. \tag{63}$$

$\widehat{Q}_1 = \widehat{Q}_2$ holds, when the street sections downstream of the intersection do not impose a bottleneck. Furthermore, $I_1 = I_2 = 1$ corresponds to a minimization of the average *queue length* rather than the average delay time. Such a minimization of the queue length makes a lot of sense and means that the optimization is made from the perspective of the traffic network rather than from the perspective of the driver. This minimizes spillover effects and, at the same time, keeps travel times low.

4.6 Complexity of Traffic Light Control

It is interesting that already a single intersection with constant arrival flows shows a large variety of operation regimes. In order to get an idea of the complexity of optimal traffic light control in general, let us ask about the dimension of the phase space. For such an analysis, it is common to transform all parameters to dimensionless form, as above. In this way, all formulas are expressed in terms of relative flows such as

$$\kappa = \frac{I_1 A_1}{I_2 A_2} \,, \quad u_1 = \frac{A_1}{\widehat{Q}_1} \,, \quad u_2 = \frac{A_2}{\widehat{Q}_2} \,. \tag{64}$$

Parameters like

$$\frac{I_2(\widehat{Q}_2 - A_2)}{I_1 A_1} \quad \text{and} \quad \frac{I_1(\widehat{Q}_1 - A_1)}{I_2 A_2} \tag{65}$$

can be expressed through the previous set of parameters. A single intersection with two phases only is characterized by the two parameters u_1 and u_2, if queue minimization is performed, and one additional parameter κ, if travel time is minimized. Therefore, the optimal operation of n intersections depends on 2^n (or even 3^n) parameters. In view of this, it is obvious that the optimal coordination of traffic lights in an urban road network constitutes a hard computational problem [39].

The consideration of non-uniform arrival flows further complicates matters. If the traffic flows are not constant, but characterized by vehicle platoons, the *phase* of traffic light control can be significant for intersection capacity [20]. Therefore, the mutual coordination of neighboring traffic lights has a significant impact [20]. This issue is, for example, addressed in [25, 33].

5 Optimize-Multiple-Phases Approach

Under certain circumstances, it may be reasonable to *interrupt* the service of a vehicle queue to clear the way for a large flow of newly arriving vehicles in the other road section. Such an interruption may be interpreted as another slower-is-faster effect, occurring in situations where the interruption-induced delay of vehicles in one road section is overcompensated for by the avoidance of delay times in the other road section. Such effects involving several green phases can clearly not be studied within the optimization of a single phase. One would rather need an approach that optimizes two or more phases simultaneously.

In the optimize-two-phases approach, it appears logical to optimize the goal function

$$G_{12}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2) = \frac{F_{12}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2)}{\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2}\,, \tag{66}$$

which considers the waiting times in the successive green phase ΔT_2 as well. The average delay time $G_{12}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2)$ is minimized by variation of *both* green time periods, ΔT_1 and ΔT_2. The optimal green times are characterized by vanishing partial derivatives $\partial G_{12}/\partial \Delta T_j$. Therefore, we must find those values ΔT_1 and ΔT_2 which fulfil

$$\frac{\partial G_{12}}{\partial \Delta T_j} = \frac{\dfrac{\partial F_{12}}{\partial \Delta T_j}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2) - F_{12}}{(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2)^2} = 0\,. \tag{67}$$

This implies the balancing principle

$$\frac{\partial F_{12}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2)}{\partial \Delta T_1} = \frac{\partial F_{12}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2)}{\partial \Delta T_2} \tag{68}$$

which is known from other optimization problems as well, e.g. in economics [11]. Condition (68) allows one to express the green time ΔT_2 as a function of the green time ΔT_1. Both values can then be fixed by finding minima of $G_{\rm w}(\tau_1 + \Delta T_1 + \tau_2 + \Delta T_2(\Delta T_1))$. When this optimization procedure is applied after completion of each phase, it is expected to be adaptive to changing traffic conditions. However, a weakness of the above approach is its neglecting of the flows in the optimization procedure. Therefore, the resulting intersection throughput may be poor, and flows would not necessarily be served, when the intersection capacity would allow for this. Therefore, we will now modify the multiple-phase optimization in a suitable way, focussing on the two-phase case.

5.1 Combined Flow-and-Delay Time Optimization

The new element of the following approach is the introduction of flow constraints into the formulation of the delay time minimization. For this, let us start with the formula for the average delay time $\mathscr{T}_j^{\rm av}$ in road section j derived in [20]. It reads

$$\mathscr{T}_j^{\rm av} = \frac{(1-f_j)^2}{(1-u_j)} \frac{T_{\rm cyc}}{2} \tag{69}$$

with

$$T_{\rm cyc} = \tau_1 + \Delta T_1 + \tau_2 + \Delta T_2 = (\tau_1 + \tau_2)(1 + \sigma_1 + \sigma_2) \tag{70}$$

and

$$1 - f_j = \frac{T_{\rm cyc} - \Delta T_j}{T_{\rm cyc}} = \frac{(1 + \sigma_1 + \sigma_2) - \sigma_j}{1 + \sigma_1 + \sigma_2} \, . \tag{71}$$

As the number of vehicles arriving on road section j during the time period $T_{\rm cyc}$ is given by $I_j A_j T_{\rm cyc} = I_j u_j \widehat{Q}_j T_{\rm cyc}$, the scaled overall delay time of vehicles over the two green phases $\Delta T_1, \Delta T_2$ and amber time periods τ_1, τ_2 covered by the cycle time $T_{\rm cyc}(\Delta T_1, \Delta T_2)$ is given by

$$\begin{aligned} G &= \frac{\sum\limits_{j=1}^{2} \mathscr{T}_j^{\rm av} I_j u_j \widehat{Q}_j T_{\rm cyc}}{T_{\rm cyc}} \\ &= \sum_{j=1}^{2} \frac{[(1 + \sigma_1 + \sigma_2) - \sigma_j]^2}{2(1-u_j)(1 + \sigma_1 + \sigma_2)} I_j u_j \widehat{Q}_j (\tau_1 + \tau_2) \, . \end{aligned} \tag{72}$$

Let us now set $\theta_j = \theta_j(\sigma_1, \sigma_2) = 0$, if $\sigma_j \le \hat{\sigma}_j$ (corresponding to $\sigma_j/(1 + \sigma_1 + \sigma_2) \le u_j$), and $\theta_j = 1$ otherwise. The dimensionless clearing time

$$\hat{\sigma}_j = \frac{u_j}{(1-u_j)}(1+\sigma_1+\sigma_2-\sigma_j) \tag{73}$$

was defined in (45). With this, we will minimize the scaled overall delay time (72) in the spirit of the optimize-two-cycles approach, but under the constraint that the average outflow

$$\begin{aligned}\overline{O} &= \sum_{j=1}^{2} \frac{I_j \widehat{Q}_j \{\Delta T_j(1-\theta_j) + [T_j + u_j(\Delta T_j - T_j)]\theta_j\}}{T_{\text{cyc}}} \\ &= \sum_{j=1}^{2} \frac{I_j \widehat{Q}_j \{\sigma_j(1-\theta_j) + [(1-u_j)\hat{\sigma}_j + u_j\sigma_j]\theta_j\}}{1+\sigma_1+\sigma_2}\end{aligned} \tag{74}$$

reaches the maximum throughput

$$\widehat{O}(u_1,u_2) = \min\Big(G_{\text{t}}(u_1,u_2), O_{\max}(u_1,u_2)\Big). \tag{75}$$

The maximum throughput corresponds to the overall flow $G_{\text{t}}(u_1,u_2) = I_1A_1 + I_2A_2 = u_1I_1\widehat{Q}_1 + u_2I_2\widehat{Q}_2$, as long as the capacity constraint (62) is fulfilled. Otherwise, if the sum of arrival flows exceeds the intersection capacity, the maximum throughput is given by[2]

$$\begin{aligned}O_{\max}(u_1,u_2) &= \max_{\substack{x_j \le u_j \\ x_1+x_2=1}} \big(x_1I_1\widehat{Q}_1 + x_2I_2\widehat{Q}_2\big) \\ &= \max_{1-u_2\le x_1\le u_1} I_2\widehat{Q}_2\big[Kx_1 + (1-x_1)\big] \\ &= \begin{cases} I_2\widehat{Q}_2\big[(K-1)u_1+1\big] & \text{if } K \ge 1 \\ I_2\widehat{Q}_2\big[1-(1-K)(1-u_2)\big] & \text{if } K < 1. \end{cases}\end{aligned} \tag{76}$$

Demanding the flow constraint

$$\overline{O}\Big(\sigma_1(u_1,u_2),\sigma_2(u_1,u_2)\Big) = \widehat{O}(u_1,u_2) \tag{77}$$

and considering (73), we can derive

$$\widehat{O} = \sum_j I_j\widehat{Q}_j\left[\frac{\sigma_j(1-\theta_j)}{1+\sigma_1+\sigma_2} + u_j\theta_j\right]. \tag{78}$$

[2] If the cycle time T_{cyc} is limited to a certain maximum value $T_{\text{cyc}}^{\max}$, one must replace the constraint $x_1 + x_2 \le 1$ by $x_1 + x_2 \le 1 - (\tau_1+\tau_2)/T_{\text{cyc}}^{\max}$ and $1-u_2$ by $1-u_2-(\tau_1+\tau_2)/T_{\text{cyc}}^{\max}$.

This implies a linear relationship between σ_1 and σ_2. If the denominator is non-zero, we have:

$$\sigma_1(\sigma_2) = \frac{\theta_1 u_1 I_1 \widehat{Q}_1 + \theta_2 u_2 I_2 \widehat{Q}_2 - \widehat{O}}{\widehat{O} - (1-\theta_1) I_1 \widehat{Q}_1 - \theta_1 u_1 I_1 \widehat{Q}_1 - \theta_2 u_2 I_2 \widehat{Q}_2} + \frac{(1-\theta_2) I_2 \widehat{Q}_2 + \theta_1 u_1 I_1 \widehat{Q}_1 + \theta_2 u_2 I_2 \widehat{Q}_2 - \widehat{O}}{\widehat{O} - (1-\theta_1) I_1 \widehat{Q}_1 - \theta_1 u_1 I_1 \widehat{Q}_1 - \theta_2 u_2 I_2 \widehat{Q}_2} \sigma_2 . \tag{79}$$

By demanding the flow constraint, we can guarantee that all arriving vehicles are served as long as the intersection capacity is sufficient, while we will otherwise use the maximum possible intersection capacity. As a consequence, operation regime (0) of the one-phase optimization, which neglected the service of at least one road section, cannot occur within this framework. Instead, it is replaced by an operation regime, in which the vehicle queue in one road section is fully cleared, while the vehicle queue in the other road section is served in part.[3] Of course, this will happen only, if the intersection capacity is insufficient to serve both flows completely (i.e. in the case $1-u_1-u_2 < 0$). If $K > 1$ (i.e. the main flow is on road Sect. 1), we have

$$\frac{\sigma_1}{1+\sigma_1+\sigma_2} = u_1 \quad \text{and} \quad \frac{\sigma_2}{1+\sigma_1+\sigma_2} = (1-u_1) \, . \tag{80}$$

If $K < 1$, the indices 1 and 2 must be interchanged.

Operation regime (1) is still defined as in Sect. 4.4 and characterized by

$$\sigma_j = \frac{u_j}{1-u_1-u_2} \, , \qquad \frac{\sigma_j}{1+\sigma_1+\sigma_2} = u_j \, . \tag{81}$$

In contrast to the one-phase optimization approach, this “normal case” of signal operation occurs in a large parameter area of the two-phase optimization approach (see blue area in Fig. 4). It implies that both green times are long enough to dissolve the vehicle queues, but not longer.

The case, where both green phases are extended, is again no optimal solution. We will, therefore, finally focus on case (2), where the vehicle queue in road Sect. 1 is just cleared ($\theta_1 = 0$), while road Sect. 2 gets an excess green time ($\theta_2 = 1$).

[3] In this case, we do not expect a periodic signal control anymore, as the growing vehicle queue in one of the road sections, see [20], has to be considered in the signal optimization procedure. Our formulas for one-phase optimization can handle this case due to the dependence on $\Delta N_j(0)$. In the two-phase optimization procedure, we would have to add $\sum_j I_j \, \Delta N_j(0)$ to formula (72), where $\Delta N_j(0) = A_j T^k_{\text{cyc}} - \widehat{Q}_j \, \Delta T^k_j$ denotes the number of vehicles that was not served during the kth cycle $T^k_{\text{cyc}} = \tau_1 + \Delta T^k_1 + \tau_2 + \Delta T^k_2$. This gives an additional term $\sum_j u_j I_j \widehat{Q}_j (\tau_1 + \tau_2) \sum_k (1 + \sigma^k_1 + \sigma^k_2 - \sigma^k_j / u_j)$ in (72).

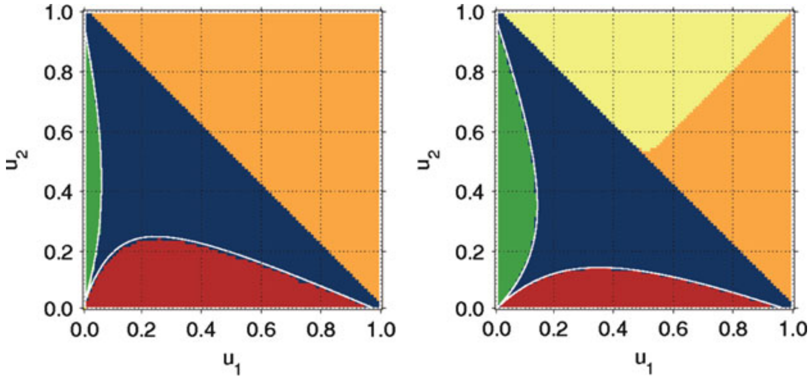


Fig. 4 Operation regimes of periodic signal control as a function of the utilizations u_j of both road sections according to the two-phase optimization approach, assuming $K = 1$, corresponding to equal roads (*left*), and $K = 3$, corresponding to a three-lane road 1 and a one-lane road 2 (*right*). For most combinations of utilizations (if u_1 is not too different from u_2), the *green* phases are terminated as soon as the corresponding road sections are cleared (see the *blue area* below the *falling diagonal line*). However, extended *green* times for road Sect. 1 result (see the *red area* along the u_1 axis), if the utilization of road Sect. 2 is small. In contrast, if the utilization of road Sect. 1 is small, extended *green* times should be given to road Sect. 2 (see the *green area* along the u_2 axis). The *white separating lines* between these areas correspond to (94), (95) fit the numerical results well. Above the line $u_2 = 1 - u_1$, the intersection capacity is insufficient to serve the vehicle flows in *both* road sections. In this area, the two-phase optimization gives solutions where road Sect. 1 is fully cleared, but road Sect. 2 is served in part (*orange area* towards the *right*), or vice versa (*yellow area* towards the *top* in the *left figure*)

With $\widehat{O} = G_t = u_1 I_1 \widehat{Q}_1 + u_2 I_2 \widehat{Q}_2$, (79) yields the simple constraint

$$\sigma_1(\sigma_2) = \frac{u_1}{1 - u_1}(1 + \sigma_2) \, , \tag{82}$$

which corresponds to (45). It implies

$$\frac{d\sigma_1}{d\sigma_2} = \frac{u_1}{1 - u_1} \, , \qquad 1 + \sigma_1 = \frac{1 + u_1 \sigma_2}{1 - u_1} \, , \tag{83}$$

and

$$1 + \sigma_1 + \sigma_2 = \frac{1 + \sigma_2}{1 - u_1} \, , \qquad \frac{\sigma_1}{1 + \sigma_1 + \sigma_2} = u_1 \, . \tag{84}$$

We will now determine the minimum of the goal function G by setting the derivative $\partial G / \partial \sigma_1$ to zero, considering

$$\frac{d\hat{\sigma}_2(\sigma_1)}{d\sigma_1} = \frac{u_2}{1 - u_2} \, . \tag{85}$$

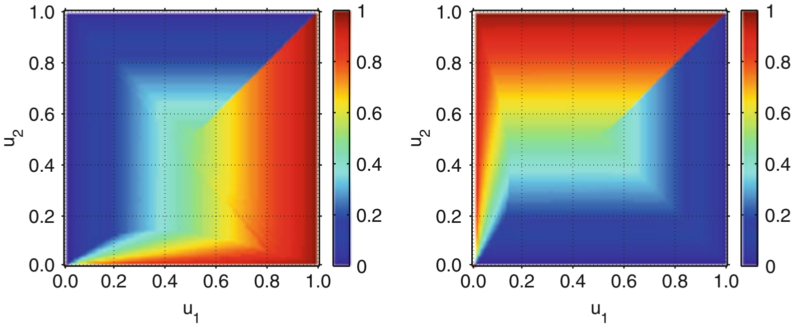


Fig. 5 Optimal *green* time fractions $\Delta T_j/T_{\rm cyc} = \sigma_j/(1+\sigma_1+\sigma_2)$ for road section $j = 1$ (*left*) and road section $j = 2$ (*right*) as a function of the utilizations u_j of both roads j, assuming periodic signal operation according to the two-phase optimization approach with $K = 1$. For combinations (u_1, u_2) with several solutions (with extended *green* time and without), we display the solution which minimizes the goal function (72). The results are qualitatively similar to the ones belonging to the one-phase optimization approach displayed in Fig. 7, but we find periodic solutions above the capacity line $u_2 = 1 - u_1$, where one road section (the one with the greater utilization) is fully cleared, while the other one is served in part

Multiplying the result with $2(1-u_1)^3(1-u_2)(1+\sigma_1+\sigma_2)^2/(I_2\widehat{Q}_2)$, we find the following relationship:

$$\begin{aligned}&2u_1u_2(1+\sigma_2)(1+u_1\sigma_2) + 2Ku_1(1-u_1)(1-u_2)(1+\sigma_2)^2\\ &\quad = u_2(1+u_1\sigma_2)^2 + Ku_1(1-u_1)(1-u_2)(1+\sigma_2)^2\,, \end{aligned} \tag{86}$$

which finally leads to

$$\begin{aligned}(1+\sigma_2)^2 &= \frac{u_2(1-u_1)^2}{u_1{}^2u_2 + Ku_1(1-u_1)(1-u_2)}\\ &= \frac{(1-u_1)^2}{u_1{}^2 + \kappa(1-u_1)(1-u_2)}\,. \end{aligned} \tag{87}$$

According to (60), for an extended green time on road Sect. 2, the condition

$$\frac{\sigma_2}{1+\sigma_1+\sigma_2} > u_2 \tag{88}$$

must again be fulfilled. If the solution $\sigma_2(u_1, u_2)$ of (87) satisfies this requirement, it can be inserted into (82) to determine the scaled green time period $\sigma_1(u_1, u_2)$ as a function of the capacity utilizations u_1 and u_2 within the framework of the optimize-two-phases approach. The corresponding results are displayed in Figs. 4–6. A generalization to signal controls with more than two phases is straight forward.

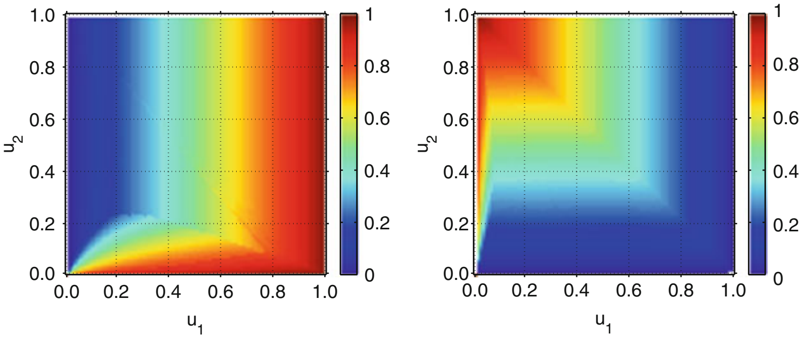


Fig. 6 Same as Fig. 5, but for $K = 3$, corresponding to a three-lane road Sect. 1 (arterial road) and a one-lane road Sect. 2 (crossing side road)

Finally, let us calculate the separating line between case (1) and case (2). Inserting (82) into (72), we can express the goal function G as a function H of a single variable σ_2:

$$H(\sigma_2) = G(\hat{\sigma}_1(\sigma_2), \sigma_2)\,. \tag{89}$$

As (82) holds for both cases, an exact clearing of road Sect. 2 or an excess green time for it, the functional dependence of goal function (89) on σ_2 must be the same for both cases. Now, on the one hand, we may apply (81) for the case without excess green time, which yields

$$1 + \sigma_2 = \frac{1 - u_1}{1 - u_1 - u_2} \quad \text{and} \quad (1 + \sigma_2)^2 = \frac{(1 - u_1)^2}{(1 - u_1 - u_2)^2}\,. \tag{90}$$

On the other hand, in the case of excess green time, we may use (87). The goal function must be the same along the separating line between both cases, which requires

$$\frac{(1 - u_1)^2}{(1 - u_1 - u_2)^2} = \frac{(1 - u_1)^2}{{u_1}^2 + \kappa(1 - u_1)(1 - u_2)}\,. \tag{91}$$

This implies

$$\kappa(1 - u_1)(1 - u_2) = (1 - u_1 - u_2)^2 - {u_1}^2 \tag{92}$$

or

$$\kappa(1 - u_1)(1 - u_2) = (1 - 2u_1 - u_2)(1 - u_2)\,. \tag{93}$$

The finally resulting equation for the separating line between the regimes with and without excess green time is given by

$$\frac{1}{\kappa} = \frac{u_2}{K u_1} = \frac{1 - u_1}{1 - 2u_1 - u_2}\,. \tag{94}$$

As Fig. 4 shows, this analytical result fits the result of our numerical optimization very well. The separating line between case (1) and case (3) is derived analogously. It may also be obtained by interchanging the subscripts 1 and 2 and substituting κ by $1/\kappa$, yielding

$$\kappa = K\frac{u_1}{u_2} = \frac{1 - u_2}{1 - 2u_2 - u_1} \, . \tag{95}$$

6 Summary, Discussion, and Outlook

We have studied the control of traffic flows at a single intersection. Such studies have been performed before, but we have focussed here on some particular features:

- For the sake of a better understanding, we were interested in deriving analytical formulas, even though this required some simplifications.
- A one-phase minimization of the overall travel times in all road sections tended to give excess green times to the main flow, i.e. to the road section with the larger number of lanes or, if the number of lanes is the same ($K = 1$), to the road section with the larger utilization (see Fig. 2). The excess green time can lead to situations where one of the vehicle flows is not served, although there would be enough service capacity for all flows.
- A minimization of vehicle queues rather than travel times simplifies the relationships through the special settings $\widehat{Q}_j = \widehat{Q}$ and $I_j = 1$, resulting in $K = 1$. Moreover, these settings guarantee that the case of no service only occurs, if the intersection capacity is exceeded.
- An optimize-multiple-phases approach considering flow constraints gives the best results among the optimization methods considered. It makes sure that both roads are served even when the intersection capacity is exceeded.
- For all considered optimization approaches, we have derived different operation regimes of traffic signals control: One of them is characterized by ending a green time period upon service of the last vehicle in the queue, which implies that all vehicles are stopped once by a traffic signal. However, we have also found conditions under which it is advised to delay switching for one of the road sections ("slower-is-faster effect"), which allows some vehicles to pass the signal without stopping.
- Compared to the one-phase optimization, a two-phase optimization tends to give much less excess green times, in particular if the utilizations of the road sections are comparable. We hypothesize that this is an effect of the short-sightedness of the one-phase optimization: It does not take into account future delay times caused by current excess green times. This hypothesis is confirmed by Fig. 7 (which is to be contrasted with the left illustration in Fig. 2). It specifies the green time durations according to (42) and (45) of the one-phase optimization, but selects the solution that minimizes the average delay time (72) over two phases.

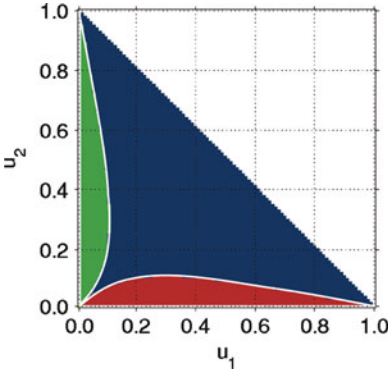


Fig. 7 Operation regimes of periodic signal control as a function of the utilizations u_j of both roads, if one specifies the clearing times and excess *green* times according to (45) and (42) of the one-phase optimization, but selects the solution that minimizes the overall delay time (72) over two successive phases. For most combinations of utilizations (if u_1 is not too different from u_2), the *green* phases are terminated as soon as the corresponding road sections are cleared (see the *blue area* below the *falling diagonal line*). However, extended *green* times for road Sect. 1 result (see the *red area* along the u_1 axis), if the utilization of road Sect. 2 is small. In contrast, if the utilization of road Sect. 1 is small, extended *green* times should be given to road Sect. 2 (see *green area* along the u_2 axis) [6]. Above the line $u_2 = 1 - u_1$, the intersection capacity is insufficient to serve the vehicle flows in *both* road sections

- Although the multi-phase optimization approach provides extended green times in a considerably smaller area of the parameter space spanned by the utilizations u_j, the slower-is-faster effect still persists when signal settings are optimized over a full cycle time (as we effectively did with the periodic two-phase optimization approach). The slower-is-faster effect basically occurs when the utilization of a road section is so small that it requires some extra time to collect enough vehicles for an efficient service during the green phase, considering the efficiency losses by switching traffic lights during the amber phases.
- In complementary appendices, we discuss traffic controls with more than two phases and an exponentially weighted goal function for short-term traffic optimization. Furthermore, we propose how to take into account the effect of stopping newly arriving vehicles and how to assess its impact as compared to queues of waiting vehicles. As stopping vehicles causes additional delay times, it becomes often favorable to implement excess green times (i.e. to apply the slower-is-faster effect").

In summary, our approach successfully delivers analytical insights into various operation regimes of traffic signal control, including the occurring slower-is-faster effects. Moreover, as the two-phase optimization approach takes care of side roads and minor flows, it has similar effects as the stabilization rule that was introduced in [33] to compensate for unstable service strategies. This stabilization rule tries to avoid spillover effects via an earlier green time by the next traffic light downstream.

Note that spillover effects imply growing delay times even in road sections which have a green light. Therefore, if the utilization is greater than the intersection capacity, travel time minimization may additionally demand to interrupt the green times of the next traffic lights upstream (in favor of a road section that could be successfully left by vehicles when a green light would be given to them). This effectively requires to generalize the traffic light control principle discussed before towards a consideration of the traffic conditions in upstream *and* downstream road sections. Such a control is considerably more complicated and will be addressed in future publications, based on formulas and principles developed in [20, 21].

6.1 Self-Organized Traffic Light Control

Our restriction to analytical calculations implied certain simplifications such as the assumption of two traffic phases, the assumption of constant arrival flows, and no obstructions of the outflow. However, these restrictions can be easily overcome by straight-forward generalizations (see Appendices). The assumption of constant arrival flows, for example, is not needed. Assuming a short-term prediction based on upstream flow measurements [35], the expected delay times or queue lengths can be determined via the integral (17). The optimal solution must then be numerically determined, which poses no particular problems. Although the behavior may become somewhat more complicated and the boundaries of the operation regimes may be shifted, we expect that the above mentioned signal operation modes and the control parameters $u_1 = A_1/\widehat{Q}_1$, $u_2 = A_2/\widehat{Q}_2$, and $\kappa = I_1 A_1/(I_2 A_2)$ still remain relevant.

In the following, we show that the optimize-one-phase approach works surprisingly well, when it is applied to signal-controlled networks with their typical, pulsed vehicle flows. Rather than performing strict travel time optimization, however, we use a simplified approach that determines exponential averages $A'_j(t)$ of the arrival flows $A_j(t)$ according to

$$A'_j(t) = \alpha_j A_j(t) + (1 - \alpha_j) A'_j(t-1) , \qquad (96)$$

and inserts these values into the formulas for the control strategies that were derived for constant arrival flows. The averaging parameters α_j are specified such that the average vehicle speed over 30 min is maximized.

Figure 8 shows simulation results for a Barcelona kind of road network (see Fig. 1) with 72 links, the lengths of which are uniformly distributed between 100 and 200 m. For simplicity, the turning fractions have been set to 1/2 for all intersections, the setup times τ_j to $\tau = 5$ s. Traffic flows were simulated in accordance with the section-based traffic model [18, 28]. The parameters determining the assumed triangular flow-density relationship on the road sections are the safe time headway $T = 1.8$ s, the maximum density $\rho_{\max} = 140$ vehicles/km, and the speed limit V_j^0, which is either set to 50 or to 70 km/h. As one can see, the average speed for the

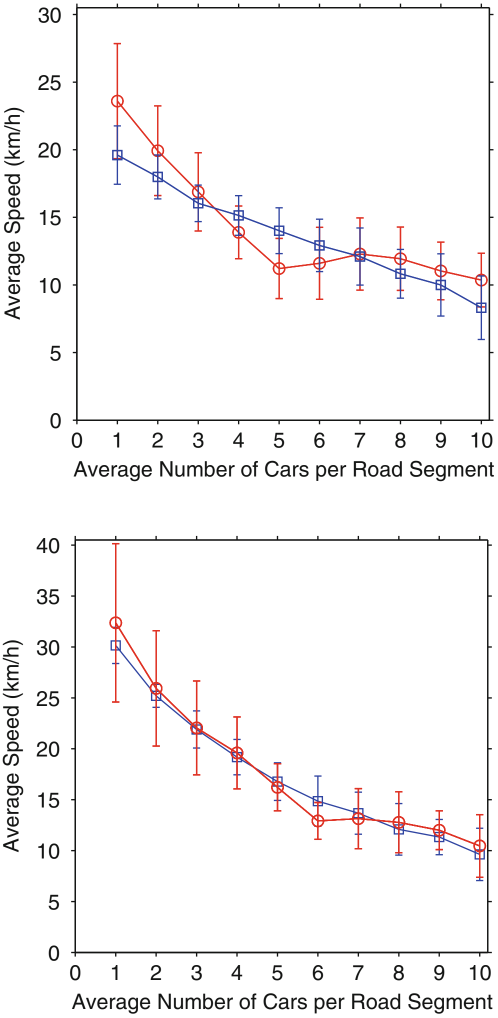


Fig. 8 Comparison of the average velocity resulting for an optimized fixed cycle time control (*red circles*) with a self-organized control based on the optimize-one-phase approach (*blue squares*) for a speed limit $V_j^0 = 50\,\text{km/h}$ (*top*) and $V_j^0 = 70\,\text{km/h}$ (*bottom*). Both control approaches perform similarly well. *Error bars* represent standard deviations. Details of the simulation scenarios are given in the main text

self-organized traffic light control performs similarly well as a fixed cycle strategy, where the cycle time is adjusted to the traffic volume. Specifically, the green times ΔT_j are linearly increased from 15 s for an average number of 1 car per road section up to 60 s for an average number of 10 cars per road segment. The offsets of the green phases are optimized by means of Particle Swarm Optimization (PSO) [30]. This serves to minimize the stopping of moving vehicle platoons.

A more detailed, numerical comparison of fixed cycle control schemes with self-organized traffic light controls for urban road networks will be presented in forth-coming publications. Note that, in [33], a somewhat more sophisticated self-control principle has been studied, which involves a short-term anticipation based on measurements of the arrival flows A_j. This self-control performs particularly well in cases of heterogeneous road networks and stochastically varying arrival flows, and it can create coordinated flow patterns similar to "green waves" (where vehicle platoons are not stopped at every traffic light).

Acknowledgements *Author contributions:* DH set up the model, performed the analytical calculations, and wrote the manuscript. AM produced the numerical results and figures, and derived the separating lines given by (94) and (95).
Acknowledgements: The authors would like to thank for partial support by the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01 08-2, the VW Foundation Project I/82 697 on "Complex Self-Organizing Networks of Interacting Machines: Principles of Design, Control, and Functional Optimization", the Daimler-Benz Foundation Project 25-01.1/07 on "BioLogistics", and the "Cooperative Center for Communication Networks Data Analysis", a NAP project sponsored by the Hungarian National Office of Research and Technology under Grant No. KCKHA005.

Appendix 1 Considering the Price of Stopping Vehicles

The previous considerations have only taken into account delays by vehicles in a vehicle queue. However, it would also make sense to consider the price of stopping vehicles. In particular, it must be possible that a large flow of moving vehicles in one road section is prioritized to a short queue of standing vehicles in the other road section. But how can we assess the relative disadvantage of stopping newly arriving vehicles as compared to stopping the service of a vehicle queue at the intersection? If the arrival flow is not large enough, it would certainly be better to continue serving the standing vehicle queue in the other road until it is fully dissolved.

We pursue the following approach: While the flow model used before implicitly assumes instantaneous vehicle accelerations and decelerations, we will now consider that, in reality, a finite vehicle acceleration a causes additional delays of $V_j^0/(2a)$, where V_j^0 denotes the free speed or speed limit. Furthermore, the reaction time T_r must be taken into account as well. This leads to an additional delay of

$$T'_j = T_\mathrm{r} + \frac{V_j^0}{2a} \tag{97}$$

for each vehicle that leaves a queue. T_r is of the order of the safe time gap T. Note that delays $V_j^0/(2b)$ due to a finite deceleration b do not additionally contribute to the delay times, as it does not matter whether delayed vehicles spend their time decelerating or stopped.[4]

Furthermore, we must determine the rate at which such additional delays are produced. This is given by the rate at which freely moving vehicles join the end of a traffic jam, i.e. by

[4] The finite deceleration only matters slightly, when the exact moment must be determined when a road section becomes fully congested.

$$\rho_{\text{jam}}|C_j| = \frac{\rho_{\text{jam}}}{\rho_{\text{jam}}/A_j - 1/V_j^0} \geq A_j\,, \tag{98}$$

where ρ_{jam} denotes the density of vehicles per lane in a standing queue. The propagation speed

$$C_j = \frac{A_j - 0}{A_j/V_j^0 - \rho_{\text{jam}}} \tag{99}$$

of the upstream front of the queue corresponds to the propagation speed of shock fronts, see [18,28,50]. Depending on the values of C_j (or A_j) and T_j', newly arriving vehicles can have an impact $T_j' C_j \rho_{\text{jam}}$ equivalent to about $\Delta N_j = 10$ queued vehicles.

Summarizing the above considerations, we suggest to replace the goal function $G_1(t)$ by the generalized formula

$$\widehat{G}_1(t) = \frac{1}{t}\sum_j I_j \int\limits_0^t dt' \Big[\Delta N_j(t') + T_j'|C_j|\rho_{\text{jam}}\Theta(\Delta N_j > 0)\Big]\,, \tag{100}$$

where $\Theta(\Delta N_j > 0) = 1$, if $\Delta N_j > 0$, and $\Theta(\Delta N_j > 0) = 0$ otherwise. In case (a) with $\Delta T_i \leq T_i$, we find

$$\begin{aligned}\widehat{F}_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2) = {} & F_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2) \\ & + I_1 T_1'|C_1|\rho_{\text{jam}}(\tau_1 + \Delta T_1 + \tau_2) \\ & + I_2 T_2'|C_2|\rho_{\text{jam}}(\tau_1 + \Delta T_1 + \tau_2)\,.\end{aligned} \tag{101}$$

This implies

$$\begin{aligned}\widehat{G}_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2) = {} & G_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2) \\ & + I_1 T_1'|C_1|\rho_{\text{jam}} + I_2 T_2'|C_2|\rho_{\text{jam}}\end{aligned} \tag{102}$$

with $G_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2)$ according to (22). Therefore, the partial derivative of $\widehat{G}_1^{\text{a}}(\tau_1 + \Delta T_i + \tau_2)$ with respect to ΔT_1 remains unchanged, and we find the same optimal green time period $\Delta T_1 = 0$ or $\Delta T_1 \geq T_1$. However, in case (b) with $\Delta T_1 \geq T_1$, we obtain

$$\begin{aligned}\widehat{F}_1^{\text{b}}(\tau_1 + \Delta T_i + \tau_2) = {} & F_1^{\text{b}}(\tau_1 + \Delta T_i + \tau_2) \\ & + I_1 T_1'|C_1|\rho_{\text{jam}}(\tau_1 + T_1 + \tau_2) \\ & + I_2 T_2'|C_2|\rho_{\text{jam}}(\tau_1 + \Delta T_1 + \tau_2)\,,\end{aligned} \tag{103}$$

which implies

$$\begin{aligned}\widehat{G}_1^{\mathrm{b}}(\tau_1 + \Delta T_i + \tau_2) = G_1^{\mathrm{b}}(\tau_1 + \Delta T_i + \tau_2) \\ + I_1 T_1' |C_1| \rho_{\mathrm{jam}} + I_2 T_2' C_2 \rho_{\mathrm{jam}} \\ - I_1 T_1' |C_1| \rho_{\mathrm{jam}} \frac{\Delta T_1 - T_1}{\tau_1 + \Delta T_1 + \tau_2}\end{aligned} \tag{104}$$

with $G_1^{\mathrm{b}}(\tau_1 + \Delta T_i + \tau_2)$ according to (28). In cases where an excess green time is favorable, the corresponding formula for the green time duration becomes

$$(\tau_1 + \Delta T_1 + \tau_2)^2 = \frac{2I_1}{I_2 A_2} [E_1 + T_1' |C_1| \rho_{\mathrm{jam}} (\tau_1 + T_1 + \tau_2)] , \tag{105}$$

i.e. the optimal green times tend to be longer. In order to support excess green times, the condition $(\tau_1 + \Delta T_1 + \tau_2)^2 \geq (\tau_1 + T_1 + \tau_2)^2$ must again be fulfilled, which requires

$$\begin{aligned}\frac{(\Delta N_1^{\max})^2}{\widehat{Q}_1 - A_1} \left(\frac{I_1}{I_2 A_2} - \frac{1}{\widehat{Q}_1 - A_1} \right) + 2 \Delta N_1^{\max} \left(\frac{I_1 \tau_1}{I_2 A_2} - \frac{\tau_1 + \tau_2}{\widehat{Q}_1 - A_1} \right) \\ \geq (\tau_1 + \tau_2)^2 - \frac{2 I_1 T_1' |C_1| \rho_{\max}}{I_2 A_2} \left(\tau_1 + \frac{\Delta N_1^{\max}}{\widehat{Q}_1 - A_1} + \tau_2 \right) .\end{aligned} \tag{106}$$

Comparing this with formula (32), we can see that the threshold for the implementation of excess green times $\Delta T_j > T_j$ is reduced. Therefore, excess green times will be implemented more frequently, as this reduces the number of stopped vehicles.

Appendix 2 More Than Two Traffic Phases

The above formulas for the optimize-one-phase approach can be easily generalized to multiple traffic phases of more complicated intersections as in the case of Barcelona's center (see Fig. 1). For

$$\Delta T_i \leq T_i = \frac{\Delta N_i^{\max}}{\widehat{Q}_i - A_i} \tag{107}$$

with

$$\Delta N_i^{\max} = \Delta N_i(0) + A_i \tau_i , \tag{108}$$

for example, we can derive from (102)

$$\begin{aligned}\widehat{G}_i^{\mathrm{a}}(\tau_i + \Delta T_i + \tau_{i+1}) = I_i \Bigg[& \Delta N_i(0) + \widehat{Q}_i \tau_{i+1} \\ & -(\widehat{Q}_i - A_i)\frac{\tau_i + \Delta T_i + \tau_{i+1}}{2}\Bigg] \\ & + \sum_{j(\neq i)} I_j \left[\Delta N_j(0) + A_j \frac{\tau_i + \Delta T_i + \tau_{i+1}}{2}\right] \\ & + \sum_j I_j T_j' |C_j| \rho_{\mathrm{jam}} \, . \end{aligned} \tag{109}$$

In contrast, for $\Delta T_i \geq T_i$ and with

$$E_i = \Delta N_i^{\max} \tau_i + \frac{(\Delta N_i^{\max})^2}{2(\widehat{Q}_i - A_i)} \, , \tag{110}$$

from (104) and (28) we obtain

$$\begin{aligned}\widehat{G}_i^{\mathrm{b}}(\tau_i + \Delta T_i + \tau_{i+1}) = & \frac{I_i E_i}{\tau_i + \Delta T_i + \tau_{i+1}} \\ & + \sum_{j(\neq i)} I_j \left[\Delta N_j(0) + A_j \frac{\tau_i + \Delta T_i + \tau_{i+1}}{2}\right] \\ & + \sum_j I_j T_j' |C_j| \rho_{\mathrm{jam}} - I_i T_i' |C_i| \rho_{\mathrm{jam}} \frac{\Delta T_i - T_i}{\tau_i + \Delta T_i + \tau_{i+1}} \, . \end{aligned} \tag{111}$$

The minimum of this function is reached for

$$(\tau_i + \Delta T_i + \tau_{i+1})^2 = \frac{I_i E_i + I_i T_i' |C_i| \rho_{\mathrm{jam}} (\tau_i + T_i + \tau_{i+1})}{\sum_{j(\neq i)} I_j A_j / 2} \, . \tag{112}$$

The occurrence of excess green time requires $(\tau_i + \Delta T_i + \tau_{i+1})^2 \geq (\tau_i + T_i + \tau_{i+1})^2$, i.e.

$$\begin{aligned} & \frac{(\Delta N_i^{\max})^2}{\widehat{Q}_i - A_i} \left(\frac{I_i}{\sum_{j(\neq i)} I_j A_j} - \frac{1}{\widehat{Q}_i - A_i} \right) + 2 \Delta N_i^{\max} \left(\frac{I_i \tau_1}{\sum_{j(\neq i)} I_j A_j} - \frac{\tau_i + \tau_{i+1}}{\widehat{Q}_i - A_i} \right) \\ & \geq (\tau_i + \tau_{i+1})^2 - \frac{2 I_i T_i' |C_i| \rho_{\mathrm{jam}}}{\sum_{j(\neq i)} I_j A_j} \left(\tau_i + \frac{\Delta N_i^{\max}}{\widehat{Q}_i - A_i} + \tau_{i+1} \right) . \end{aligned} \tag{113}$$

It can be seen that the existence of more traffic phases is unfavorable for providing excess green times. For their existence, a small number of phases is preferable.

Procedure of Traffic Signal Control

Based on the above formulas, the next green phase i is determined as follows:

1. Set the time t to zero, after the last green phase i' has been completed.
2. Apply the required service time (amber time) of duration $\tau_{i'+1}$ and set $\tau_j = \tau_{i'+1}$ for all road sections j. Then, calculate $\Delta N_j^{\max}$ and E_j for all j with formulas (108) and (110).
3. During the service time, determine the green times ΔT_j and T_j with and without green time extension, for each road section j with formulas (112) and (107).
4. If $\Delta T_j > T_j$ and $\widehat{G}_j^{\rm b}(\tau_j + \Delta T_j + \tau_{j+1}) < \widehat{G}_j^{\rm a}(\tau_j + T_j + \tau_{j+1})$, see (111) and (109), consider the implementation of the extended green time ΔT_j and set $\widehat{G}_j = \widehat{G}_j^{\rm b}(\tau_j + \Delta T_j + \tau_{j+1})$. Otherwise consider the implementation of the clearing time T_j and set $\widehat{G}_j = \widehat{G}_j^{\rm a}(\tau_j + T_j + \tau_{j+1})$, but if $\widehat{G}_j^{\rm a}(\tau_j + \tau_{j+1}) < \widehat{G}_j$, set $\Delta T_j = 0$ and $\widehat{G}_j = \widehat{G}_j^{\rm a}(\tau_j + \tau_{j+1})$.
5. Among all road sections j' different from the previously selected one i', choose that one i for service, for which the expected average travel time $\widehat{G}_i$ is smallest (i.e. $\widehat{G}_i = \min_{j(\neq i')} \widehat{G}_j$). Implement the selected green phase ΔT_i.
6. Update the length of the vehicle queue in road section i according to

$$\Delta N_i(\tau_i + \Delta T_i) = 0 \tag{114}$$

   and the queue lengths in all other road sections $j \neq i$ according to

$$\Delta N_j(\tau_i + \Delta T_i) = \Delta N_j(0) + A_j(\tau_i + \Delta T_i)\,. \tag{115}$$

   If road section was not served ($\Delta T_i = 0$), update the vehicle queues in *all* road sections j (including i) according to (115).
7. At the end of the corresponding green time duration ΔT_i, set $i' = i$ and continue with step 1.

The optimize-multiple-phases approach can be generalized in a similar way. Then, among all solutions satisfying preset flow constraints, that multi-phase solution is chosen, which minimizes the goal function and does not start with a service of the previously served road section. In order to flexibly adjust to varying traffic conditions, one may repeat the optimization after completion of one phase rather than after completion of all the phases considered in the multi-phase optimization.

Appendix 3 Limited Forecast Time Horizon

While traffic light optimization is an NP-hard problem [39], we have simplified it here considerably by restricting ourselves to local optimization and to limited time horizons. Both simplifications may imply a potentially reduced traffic performance in the urban street network, but this loss of performance is small if traffic lights adjust to arriving vehicle platoons [33]. The reliable look-ahead times are anyway very limited for fundamental reasons (see the Appendix in [33]). Therefore, one can restrict traffic light optimization to time periods $1/\lambda$, over which the traffic forecast can be done with sufficient accuracy. When traffic lights are switched frequently, the value of $1/\lambda$ of the forecast time horizon will go down.

Note that an optimization based on unreliable long-term forecasts will yield bad results. Therefore, it is not only *justified*, but also *successful* to replace the optimization of one or several full cycles by the optimization of, say, two phases. Alternatively, one may minimize the exponentially weighted travel times, i.e. minimize the function

$$\widetilde{G} = \sum_j \lambda I_j \int\limits_0^\infty dt \; \mathrm{e}^{-\lambda t} \Big[\Delta N_j(t) + T'_j |C_j| \rho_{\mathrm{jam}} \Theta(\Delta N_j > 0) \Big] \qquad (116)$$

by variation of the duration and sequence of green phases. While this approach is less suited for an analytical optimization, it reminds of formulations of discounted functions in economics [11]. Goal function (116) can be optimized *numerically*, limiting the evaluation of the integral to the range $t < 3/\gamma$.

References

1. R. Barlovic, T. Huisinga, A. Schadschneider, M. Schreckenberg, Adaptive traffic light control in the ChSch model for city traffic, in *Traffic and Granular Flow'03*, ed. by P.H.L. Bovy, S.P. Hoogendoorn, M. Schreckenberg, D.E. Wolf (Springer, Berlin, 2004)
2. M. Ben-Akiva, A. de Palma, Some circumstances in which vehicles will reach their destinations earlier by starting later: Revisited. Transp. Sci. **20**, 52–55 (1986)
3. O. Biham, A.A. Middleton, D. Levine, Self-organization and a dynamical transition in traffic-flow models. Phys. Rev. A **46**, R6124–R6127 (1992)
4. D. Braess, Über ein Paradoxon der Verkehrsplanung. Unternehmensforsch. **12**, 258–268 (1968)
5. E. Brockfeld, R. Barlovic, A. Schadschneider, M. Schreckenberg, Optimizing traffic lights in a cellular automaton model for city traffic. Phys. Rev. E **64**, 056132 (2001)
6. C. Chase, P.J. Ramadge, On real-time scheduling policies for flexible manufacturing systems. IEEE Trans. Autom. Contr. **37**(4), 491–496 (1992)
7. D. Chowdhury, A. Schadschneider, Self-organization of traffic jams in cities: Effects of stochastic dynamics and signal periods. Phys. Rev. E **59**, R1311–R1314 (1999)
8. R.B. Cooper, S.-C. Niu, M.M. Srinivasan, When does forced idle time improve performance in polling models? Manag. Sci. **44**, 1079–1086 (2000)
9. J. Esser, M. Schreckenberg, Microscopic simulation of urban traffic based on cellular automata. Int. J. Mod. Phys. B **8**, 1025–1036 (1997)
10. B. Faieta, B.A. Huberman, Firefly: a synchronization strategy for urban traffic control. Internal Report No. SSL-42, Xerox PARC, Palo Alto, 1993
11. G. Feichtinger, R.F. Hartl, *Optimale Kontrolle ökonomischer Prozesse [Optimal Control of Economic Processes]* (de Gruyter, Berlin, 1986)
12. M. Fouladvand, M. Nematollahi, Optimization of green-times at an isolated urban crossroads. Eur. Phys. J. B **22**, 395–401 (2001)

13. M.E. Fouladvand, M.R. Shaebani, Z. Sadjadi, Simulation of intelligent controlling of traffic flow at a small city network. J. Phys. Soc. Jpn. **73**, 3209 (2004)
14. M.E. Fouladvand, Z. Sadjadi, M.R. Shaebani, Optimized traffic flow at a single intersection: traffic responsive signalization. J. Phys. A **37**, 561–576 (2004)
15. M. Garavello, B. Piccoli, *Traffic Flow on Networks* (American Institute of Mathematical Sciences, Springfield, 2006); Y. Chitour, B. Piccoli, Traffic circles and timing of traffic lights for cars flow. Discrete Cont. Dyn. Syst. B **5**, 599–630 (2005)
16. C. Gershenson, Self-Organizing traffic lights. Complex Syst. **16**, 29–53 (2005)
17. S.-B. Cools, C. Gershenson, B. D'Hooghe, Self-organizing traffic lights: A realistic simulation, in *Advances in Applied Self-Organizing Systems*, ed. by M. Prokopenko (Springer, New York, 2007)
18. D. Helbing, A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks. J. Phys. Math. Gen. **36**, L593–L598 (2003)
19. D. Helbing, Production, supply, and traffic systems: A unified description. in *Traffic and Granular Flow '03*, ed. by S.P. Hoogendoorn, S. Luding, P.H.L. Bovy, M. Schreckenberg, D.E. Wolf (Springer, Berlin, 2005), pp. 173–188
20. D. Helbing, Derivation of a fundamental diagram for urban traffic flow. Eur. Phys. J. B **70**, 229–241 (2009)
21. D. Helbing, S. Lämmer, Verfahren zur Koordination konkurrierender Prozesse oder zur Steuerung des Transports von mobilen Einheiten innerhalb eines Netzwerkes [Method for coordination of concurrent processes for control of the transport of mobile units within a network]. Patent WO/2006/122528 (2006)
22. D. Helbing, I. Farkas, T. Vicsek, Simulating dynamical features of escape panic. Nature **407**, 487–490 (2000)
23. D. Helbing, R. Jiang, M. Treiber, Analytical investigation of oscillations in intersecting flows of pedestrian and vehicle traffic. Phys. Rev. E **72**, 046130 (2005)
24. D. Helbing, R. Jiang, M. Treiber, Analytical investigation of oscillations in intersecting flows of pedestrian and vehicle traffic. Phys. Rev. E **72**, 046130 (2005); R. Jiang, D. Helbing, P.K. Shukla, Q.-S. Wu, Inefficient emergent oscillations in intersecting driven many-particle flows. Phys. A **368**, 567–574 (2006)
25. D. Helbing, S. Lämmer, J.-P. Lebacque, Self-organized control of irregular or perturbed network traffic, in *Optimal Control and Dynamic Games*, ed. by C. Deissenberg, R.F. Hartl (Springer, Dordrecht, 2005), pp. 239–274
26. D. Helbing, A. Johansson, J. Mathiesen, M.H. Jensen, A. Hansen, Analytical approach to continuous and intermittent bottleneck flows. Phys. Rev. Lett. **97**, 168001 (2006)
27. D. Helbing, T. Seidel, S. Lämmer, K. Peters (2006) Self-organization principles in supply networks and production systems, in *Econophysics and Sociophysics—Trends and Perspectives*, ed. by B.K. Chakrabarti, A. Chakraborti, A. Chatterjee (Wiley, Weinheim, 2006), pp. 535–558
28. D. Helbing, J. Siegmeier, S. Lämmer, Self-organized network flows. Netw. Heterogeneous Media **2**, 193–210 (2007)
29. D.-w. Huang, W.-n. Huang, Traffic signal synchronization. Phys. Rev. E **67**, 056124 (2003)
30. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, Perth, 27 November 1995 to 01 December 1995, pp. 1942–1948
31. A. Kesting, M. Treiber, M. Schönhof, D. Helbing, Extending adaptive cruise control to adaptive driving strategies. Transp. Res. Rec. **2000**, 16–24 (2007)
32. S. Lämmer, Reglerentwurf zur dezentralen Online-Steuerung von Lichtsignalanlagen in Straßennetzwerken [Controller design for a decentralized control of traffic lights in urban road networks]. Ph.D. Thesis, Dresden University of Technology, 2007
33. S. Lämmer, D. Helbing, Self-control of traffic lights and vehicle flows in urban road networks. J. Stat. Mech. (JSTAT), **2008**, P04019 (2008)
34. S. Lämmer, H. Kori, K. Peters, D. Helbing Decentralised control of material or traffic flows in networks using phase-synchronisation. Phys. A **363**, 39–47 (2006)

35. S. Lämmer, R. Donner, D. Helbing, Anticipative control of switched queueing systems. Eur. Phys. J. B **63**, 341–348 (2008)
36. T. Nagatani, Jamming transition in the traffic-flow model with two-level crossings. Phys. Rev. E **48**, 3290–3294 (1993)
37. T. Nagatani, Control of vehicular traffic through a sequence of traffic lights positioned with disordered interval. Phys. Stat. Mech. Appl. **368**, 560–566 (2006)
38. T. Nakatsuji, S. Seki, T. Kaku, Development of a self-organizing traffic control system using neural network models. Transp. Res. Rec. **1324**, 137–145 (1995)
39. C.H. Papadimitriou, J.N. Tsitsiklis, The complexity of optimal queuing network control. Math. Oper. Res. **24**, 293–305 (1999)
40. T. Roughgarden, *Selfish Routing and the Price of Anarchy* (MIT, Cambridge, 2005)
41. M. Sasaki, T. Nagatani, Transition and saturation of traffic flow controlled by traffic lights. Phys. Stat. Mech. Appl. **325** 531–546 (2003)
42. M. Schönhof, M. Treiber, A. Kesting, D. Helbing, Autonomous detection and anticipation of jam fronts from messages propagated by intervehicle communication. Transp. Res. Rec. **1999**, 3–12 (2007)
43. D. Schrank, T. Lomax, *The 2005 Urban Mobility Report* (Texas Transportation Institute, College Station Texas, 2005)
44. K. Sekiyama, J. Nakanishi, I. Takagawa, T. Higashi, T. Fukuda, Self-organizing control of urban traffic signal network. IEEE Int. Conf. Syst. Man. Cybern. **4**, 2481–2486 (2001)
45. P.M. Simon, K. Nagel, Simplified cellular automaton model for city traffic. Phys. Rev. E **58**, 1286–1295 (1998)
46. R.J. Smeed, Some circumstances in which vehicles will reach their destinations earlier by starting later. Transp. Sci. **1**, 308–317 (1967)
47. H.-U. Stark, C.J. Tessone, F. Schweitzer, Slower-is-faster: Fostering consensus formation by heterogeneous inertia. Adv. Complex Syst. **11**(4), 551–563 (2008)
48. R. Steinberg, R.E. Stone, The prevalence of paradoxes in transportation equilibrium problems. Transp. Sci. **22**, 231–241 (1988)
49. B.A. Toledo, V. Munoz, J. Rogan, C. Tenreiro, J.A. Valdivia, Modeling traffic through a sequence of traffic lights. Phys. Rev. E **70**, 016107 (2004)
50. G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)

Modeling and Optimization of Scalar Flows on Networks

Simone Göttlich and Axel Klar

Abstract Detailed models based on partial differential equations characterizing the dynamics on single arcs of a network (roads, production lines, etc.) are considered. These models are able to describe the dynamical behavior in a network accurately. On the other hand, for large scale networks often strongly simplified dynamics or even static descriptions of the flow have been widely used for traffic flow or supply chain management due to computational reasons. In this paper, a unified presentation highlighting connections between the above approaches are given and furthermore, a hierarchy of dynamical models is developed including models based on partial differential equations and nonlinear algebraic equations or even combinatorial models based on linear equations. Special focus is on optimal control problems and optimization techniques where combinatorial and continuous optimization approaches are discussed and compared.

1 Introduction

Modeling and simulation of flows on networks, like traffic flow on highways or supply chain models have been investigated intensively during the last decade.

On the one hand detailed models based on partial differential equations describing the dynamics on single arcs of the network have been constantly developed and improved. These models are able to describe the dynamical behavior accurately including special features of the dynamics like jam/shock propagation or stop-and-go waves in traffic flow. To describe flows on networks these detailed models have also been used. However, the number of arcs which can be treated by such an approach is restricted, in particular, if optimization problems have to

S. Göttlich (✉) · A. Klar
University of Mannheim, School of Business Informatics and Mathematics, A5, 6, 68131 Mannheim, Germany
e-mail: goettlich@uni-mannheim.de; klar@itwm.fhg.de

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_8,

be solved. On the other hand large scale networks with strongly simplified dynamics or even static descriptions of the flow have been widely investigated for traffic flow or supply chain management. In particular, optimal control problems for traffic flow on networks arising from traffic or supply chain management are a major focus of research in this field.

The purpose of this paper is to give a unified presentation highlighting connections between the above approaches. A hierarchy of simplified dynamical models is developed for different fields of applications starting with the "correct" dynamical description based on partial differential equations and ending with nonlinear algebraic equations or even combinatorial models based on linear equations. These models give a reasonably accurate description of the dynamics and, at the same time, are solvable for large scale networks.

Special focus is on optimal control problems and optimization techniques. Various combinatorial and continuous optimization techniques, like adjoint calculus and solution methods for mixed–integer problems, are discussed. Using strongly simplified models large scale networks can be optimized with combinatorial approaches in real-time. However, including more complex (in particular nonlinear) dynamics reduces the advantage of the combinatorial algorithms compared to continuous optimization procedures.

These topics are illustrated for two different fields: traffic networks and supply chains.

Concerning traffic flow, modelling and simulation has been investigated intensively during the last years, see, for example, [5, 11, 19–22, 38, 43, 47, 48, 52] and many others. Macroscopic models using partial differential equations have been developed for cumulative (averaging over all lanes) or multi-lane descriptions of traffic on unidirectional roads. For the present investigations we are interested in traffic flow models for road networks using scalar models based on partial differential equations. That means we use a cumulative description of traffic on each road not distinguishing between single lanes and a scalar macroscopic traffic model. First several possibilities to define suitable conditions at the junctions are presented to obtain well posed network solutions. Then simplified models are derived. Finally this hierarchy of models is used to optimize the distribution of traffic through the network.

For the simulation of PDE-based supply chain models different approaches have been introduced during the last years; see for example [1, 3, 4, 13, 14, 17, 18]. These models describe the evolution of flows, in particular the flow of goods, in supply networks. However, in many applications the simulation and prediction of long time behavior is only one important item. A further important aspect in supply chain decision making are optimization problems, for example maximizing output of a production process or minimizing used buffers. These optimization problems can be formulated on a continuous level with constraints consisting of partial or ordinary differential equations. Then, naturally an adjoint calculus is used for the efficient computation of optimal control parameters. Alternatively, we will see that models based on mixed–integer programming (MIP) can be used to find the optimal load balance on the interconnections between different entities.

The paper is organized along the following lines: In Sect. 2 traffic flow networks are considered. Section 2.1 considers scalar PDE's for traffic flow and the associated network models. In particular, the coupling conditions are considered in detail. In Sect. 2.2 and also in Sect. 2.3 a hierarchy of simplified network models is developed. Finally, Sect. 2.3 looks at optimization and control procedures for these models. In Sect. 3, we focus on supply chain models. We start with the introduction of a network model governed by scalar partial differential equations in Sect. 3.1. Similarities and differences to traffic flow models are addressed. Optimization techniques will be discussed in Sect. 3.2. The Lagrange formalism (i.e. adjoint equations) is compared to a mixed–integer approach.

Parts of this work have been taken from the articles [15, 24, 25] (traffic flow models) as well as [16, 17, 36] (supply chain models).

2 Traffic Flow Networks

This section is organized in the following way: in Sect. 2.1 traffic network models based on partial differential equations are discussed. This includes, in particular, the dynamic models and the statement and comparison of different coupling conditions at junctions. Section 2.2 describes several simplifications of the models developed in the previous subsection. Finally, Sect. 2.3 is concerned with optimization procedures for the different models of the model hierarchy developed before. Additionally, linearized models are considered and related to mixed-integer problems. Finally, the optimization routines are compared numerically.

2.1 Network Models Based on Scalar Partial Differential Equations

First several definitions concerning the network are given. Moreover, the equations used to describe the flow on the network are specified.

Definition 2.1 (Network definition).[1,2] A traffic flow network is a finite, connected directed graph, where, in addition, we may attach a finite number of directed curves extending to infinity. The roads are numbered by i and the set of all roads is $I = (1, 2, \dots, K)$. The junctions are numbered by j and the set of junctions is $J = (J_1, J_2, \dots, J_M)$. Each road is modelled by the interval $[a_i, b_i]$, where a_i and b_i can be infinity.

[1] Follow Holden and Risebro in [32].

[2] See Fig. 1.

Continuous traffic models have been introduced by several people. A classical model is due to Aw/Rascle [5], see also [6,19]. The cumulative form of the model is

$$\partial_t \rho + \partial_x(\rho u) = 0 \tag{1}$$
$$\partial_t(\rho u) + \partial_x(\rho u^2) + c(\rho)\partial_x \rho = \frac{1}{T(\rho)}\rho[U(\rho) - u]$$

where ρ describes the density on the whole road, and u the mean velocity. $f(\rho) = \rho U(\rho)$ is the so called fundamental diagram and T is a relaxation time. $c = c(\rho)$ describes the anticipation of the drivers. As long as T is small, the above model is approximated by the well-known Lighthill–Whitham equations [45]:

$$\partial_t \rho + \partial_x f(\rho) = 0 \tag{2}$$

with $f(\rho) = \rho U(\rho)$.

Remark 2.2. Here we use the Lighthill–Whitham model (2). We note that a traffic network definition based on the full (1) can be done as well, see [26].

From now on, we assume that the Lighthill–Whitham equation holds on the network away from intersections, i.e.,

$$\partial_t \rho_i(x,t) + \partial_x f(\rho_i(x,t)) = 0 \quad \forall i \in I, x \in (a_i, b_i), t \geq 0 \tag{3}$$
$$\rho_i(x,0) = \rho_{i,0}(x) \quad \forall x \in (a_i, b_i)$$

where the density on road i is denoted by ρ_i. The maximal density is for the following assumed to be $\rho_{max} = 1$, if not otherwise stated. We consider entropic solutions on each of the single roads. A weak solution of the network problem has been defined in [32] as a solution in the sense of distributions with test functions which are smooth at the intersections: Let $\phi_i = \phi_i(x,t), i = 1,\ldots,K$ be smooth test functions with compact support in $[a_i, b_i] \times \mathbf{R}^+$. Suppose that the test functions are smooth across the junctions, i.e., for example for an ingoing road i and an outgoing road j of the same junction they fulfill

$$\phi_i(b_i,t) = \phi_j(a_j,t) \text{ and } \partial_x \phi_i(b_i,t) = \partial_x \phi_j(a_j,t), t \geq 0.$$

$\{\rho_i\}_{i=1,\ldots,K}$ is called a weak solution if it satisfies

$$\sum_{i=1}^{K} \left(\int_0^\infty \int_{a_i}^{b_i} [\rho_i \partial_t \phi_i + f(\rho_i)\partial_x \phi_i]\, dxdt + \int_{a_i}^{b_i} \rho_{i,0}(x)\phi_i(x,0)dx \right) = 0.$$

One is looking now for a well defined problem on the whole network. To do so one has to give a formulation of a well defined problem at each junction. Consider a junction in the network with n incoming roads labeled by $i = 1,\ldots,n$ and m

outgoing roads labeled by $i = n+1, \dots, n+m$. One obtains (the Rankine–Hugoniot condition) that the following condition is satisfied for the above weak solutions:

$$\sum_{i=1}^{n} f(\rho_i(b_i, t)) = \sum_{i=n+1}^{n+m} f(\rho_i(a_i, t)) \quad \forall t. \tag{4}$$

We define, following again [32],

Definition 2.3 (Weak solutions of a Riemann problem at a junction). By the definition of a weak solution of the Riemann problem for a junction $J_j \in J$, we mean a weak solution of the initial value problem (3) for the network consisting of the single junction with n incoming and m outgoing roads all extending to infinity. The initial data are given by

$$\rho_{i,0}(x) = \tilde{\rho}_i, \ \forall x \in [a_i, b_i], i = 1, \dots n + m, \tag{5}$$

where $\tilde{\rho}_i$ are constants.

Consider now the Riemann initial data (5) with $\tilde{\rho}_i = \rho_{i,0}(x = b_i)$ for incoming roads and $\tilde{\rho}_i = \rho_{i,0}(x = a_i)$ for outgoing roads for a single junction. Assuming a unique solution for the problem at the junction, we denote the solution at the junction, i.e., at $x = b_i$ for incoming and at $x = a_i$ for outgoing roads, by

$$(\overline{\rho}_1, \dots, \overline{\rho}_{n+m}).$$

If suitable restrictions on the $\overline{\rho}_i$ are imposed, see below, it turns out that these states are independent of time, see [10, 32].

Once the $\overline{\rho}_i$ are determined the original problem at the junction is solved in the usual way by solving a Riemann problem with initial data $\rho_{i,0}(x), x < b_i$ and $\overline{\rho}_i, x = b_i$ for incoming roads and the corresponding procedure for outgoing roads.

However, the above coupling condition (4) is obviously not sufficient to obtain a unique weak solution $\overline{\rho}_1, \dots, \overline{\rho}_{n+m}$ at the junction. The main question in defining traffic flow on a network is the definition of suitable additional coupling conditions at the junctions to obtain unique values for $(\overline{\rho}_1, \dots, \overline{\rho}_{n+m})$, i.e., a unique solution of the Riemann problem at each junction. The difference between the models considered in this survey is the way extra conditions are imposed at the junctions.

For the following we assume as in [32] that

$$f(0) = f(1) = 0 \ \text{ and } \ \exists \sigma \in (0, 1) : f'(\sigma) = 0 \text{ and } (\rho - \sigma) f'(\rho) < 0 \quad \forall \rho \neq \sigma \tag{6}$$

This condition is fulfilled by any reasonable model for the fundamental diagram, see for example [39].

We will now give a brief summary of two traffic network models, the Holden/Risebro and Coclite/Piccoli models. For further details we refer to [32] and [10].

2.1.1 The Approach of Holden and Risebro

We consider as before a junction with n incoming roads and m outgoing roads labeled by $i = 1, \ldots, n + m$. Given the constant initial values $\rho_{i,0}$, we need to determine a unique solution $\overline{\rho_i}$ satisfying the coupling condition. After determining the state $\overline{\rho_i}$ the solution to the junction is the solution of the Riemann problem on each road i as described above. Thus the solution consists of waves emerging from the junction. The following restrictions are necessary to obtain reasonable speed directions for the emerging waves, i.e., negative speeds on incoming and positive on outgoing roads.

$$\begin{array}{lll}
\overline{\rho}_i \in [\sigma, 1] & \rho_{i,0} \geq \sigma & i = 1, \ldots, n \\
\overline{\rho}_i \in \{\rho_{i,0}\} \cup [\tau(\rho_{i,0}), 1] & \rho_{i,0} \leq \sigma & i = 1, \ldots, n \\
\overline{\rho}_i \in [0, \sigma] & \rho_{i,0} \leq \sigma & i = n+1, \ldots, n+m \\
\overline{\rho}_i \in [0, \tau(\rho_{i,0})] \cup \{\rho_{i,0}\} & \rho_{i,0} \geq \sigma & i = n+1, \ldots, n+m
\end{array} \tag{7}$$

where for each $\rho \neq \sigma, \rho \in [0, 1]$ the value $\tau(\rho)$ is the unique number $\tau(\rho) \neq \rho$, s.t. $f(\rho) = f(\tau(\rho))$. Thus $\rho < \sigma \Rightarrow \tau(\rho) > \sigma$ and vice versa. The coupling condition (4) reads

$$\sum_{i=1}^{n} f(\overline{\rho}_i) = \sum_{i=n+1}^{n+m} f(\overline{\rho}_i). \tag{8}$$

To obtain a unique weak solution Holden and Risebro introduced an additional "entropy condition" at the junction: Let g be a differentiable, strictly concave function of a single variable and define E as follows:

$$E(\overline{\rho}_1, \ldots, \overline{\rho}_{n+m}) = \sum_{i=1}^{n+m} g(\gamma_i) \tag{9}$$

where $\gamma_i = \gamma(\overline{\rho}_i) = f(\overline{\rho}_i)/f(\sigma)$. Then one can prove,[3] that the solution of the following problem yields a unique solution in the sense of Definition 2.3:

$$\text{Maximize } E \text{ subject to (8) and (7)} \tag{10}$$

Remark 2.4. The Holden/Risebro approach describes a general framework for defining well-posed traffic network problems. In particular, they show that the main task is to find a condition to replace the "entropy condition" at the junction. To obtain applicable conditions one has to be more specific in the choice of suitable

[3] Theorem 1.1 in [32].

conditions at the junctions. This has been done by Coclite and Piccoli in the model described below.

2.1.2 The Approach of Coclite/Piccoli

Consider again a junction with n incoming roads and m outgoing roads. Coclite and Piccoli use the same basic idea, i.e., find $\overline{\rho}_i$ to given $\rho_{i,0}$ and solve a Riemann problem on each road. Therefore, similar restrictions of the possible values of $\overline{\rho}_i$ are necessary:

$$\begin{array}{llll} \overline{\rho}_i & \in [\sigma, 1] & \rho_{i,0} \geq \sigma & i = 1, \dots, n \\ \overline{\rho}_i & \in \{\rho_{i,0}\} \cup (\tau(\rho_{i,0}), 1] & \rho_{i,0} \leq \sigma & i = 1, \dots, n \\ \overline{\rho}_i & \in [0, \sigma] & \rho_{i,0} \leq \sigma & i = n+1, \dots, n+m \\ \overline{\rho}_i & \in [0, \tau(\rho_{i,0})) \cup \{\rho_{i,0}\} & \rho_{i,0} \geq \sigma & i = n+1, \dots, n+m. \end{array} \tag{11}$$

Instead of the function g, a matrix $A = (\alpha_{ki}) \in \mathbf{R}^{m \times n}$ is introduced in this model. A describes the percentages of drivers who want to drive from road i to k and is assumed to fulfill the following assumptions

$$\alpha_{ki} \neq \alpha_{ki'}, \forall i \neq i' \text{ and } 0 < \alpha_{kn} < 1 \text{ and } \sum_{k=n+1}^{n+m} \alpha_{ki} = 1 \quad \forall i. \tag{12}$$

Then the following additional coupling condition is assumed to be satisfied for the junction $j \in J$

$$f(\overline{\rho}_k) = \sum_{i=1}^{n} \alpha_{ki} f(\overline{\rho}_i) \quad \forall k = n+1, \dots, n+m. \tag{13}$$

Furthermore, Coclite/Piccoli introduce a function E, which measures the flux of the incoming roads

$$E(\overline{\rho}_1, \dots, \overline{\rho}_n) = \sum_{i=1}^{n} f(\overline{\rho}_i). \tag{14}$$

Then it is proved [4] that the following problem has a unique solution in the sense of definition (2.3)

$$\text{Maximize } E \text{ subject to (13) and (11)} \tag{15}$$

[4] Theorem 3.1 in [10].

Remark 2.5. As mentioned in [10] for example a junction with two incoming ($n = 2$) and one outgoing ($m = 1$) road is not covered by this model. In this case condition (12) is not fulfilled anymore. The result is not unique, as simple examples show. A solution has been suggested in [10], see below.

The Coclite/Piccoli Model for Two Ingoing and One Outgoing Road

We consider a junction with two incoming roads $n=2$ and one outgoing road $m=1$. The initial densities on roads i are given by $\rho_{1,0}, \rho_{2,0}, \rho_{3,0}$. The corresponding fluxes are denoted as $\gamma_{i,0} = f(\rho_{i,0})$. Denote the maximum of the flux by $f(\sigma)$. We denote the sets of valid resulting fluxes γ_i by Ω_i. For the incoming roads $i = 1, 2$ this is

$$\rho_{i,0} \leq \sigma \Rightarrow \Omega_i = [0, \gamma_{i,0}] \tag{16}$$

and

$$\rho_{i,0} \geq \sigma \Rightarrow \Omega_i = [0, f(\sigma)] \tag{17}$$

For the outgoing road $i = 3$

$$\rho_{i,0} \leq \sigma \Rightarrow \Omega_i = [0, f(\sigma)] \tag{18}$$

and

$$\rho_{i,0} \geq \sigma \Rightarrow \Omega_i = [0, \gamma_{i,0}]. \tag{19}$$

It is easy to see, that if we define c_i such that

$$\Omega_i = [0, c_i] \tag{20}$$

and if $c_1 + c_2 > c_3$, the solution of the Coclite/Piccoli conditions is not unique: Assume $c_1 < c_2$, then the family (γ_1, γ_2), where $\gamma_1 \in [c_3 - c_2, c_1]$ and $\gamma_2 = c_3 - \gamma_1$, maximizes $E(\gamma_1, \gamma_2)$ and are admissible solutions in the sense of Coclite/Piccoli.

The above problem can be solved for example by the following simple procedure suggested in [10]: Obviously the problem only arises, if the possible flux on the incoming roads is larger than the maximal flux, the outgoing road can handle. In this dense traffic situation one may assume that the cars of the incoming roads move to the outgoing road in an alternating way resulting in an equal flux on the two incoming lanes. Thus we get the following conditions for the fluxes at the junction:

(1) $\gamma_1 + \gamma_2 \in \Omega_3$, $\gamma_1 \in \Omega_1, \gamma_2 \in \Omega_2$.
(2) If $c_1 + c_2 \leq c_3$: Maximize E w.r.t. (1).
(3) If $c_1 + c_2 > c_3$: Maximizes E w.r.t. (1) and $\gamma_1 = \gamma_2$.

The corresponding densities $\overline{\rho}_i$ are then found as usual in the Coclite/Piccoli model: Using condition (11) they are uniquely determined, iff the γ_i are uniquely determined. To see that the above set of conditions yields indeed unique $\gamma_i, i = 1, \dots, 3$ for arbitrary $\gamma_{i,0}$ we have to consider the case $c_1 + c_2 > c_3$. Then we have to look for γ_1, γ_2 such that

$$\begin{aligned} &\max \gamma_1 + \gamma_2 \quad \text{w.r.t.} \\ &\quad \gamma_1 = \gamma_2 \\ &\quad 0 \leq \gamma_1 \leq c_1, 0 \leq \gamma_2 \leq c_2, \gamma_1 + \gamma_2 \leq c_3. \end{aligned}$$

Obviously, the unique solution is $\gamma_1 = \min\{c_1, c_2, c_3/2\}, \gamma_2 = \gamma_1, \gamma_3 = \gamma_1 + \gamma_2$.

Remark 2.6. The above conditions for example for two outgoing and one ingoing road are called FIFO (first in, first out) in the traffic engineering literature, see for example [12].

One Ingoing and Two Outgoing Roads, FIFO Versus NONFIFO Models

We consider a junction with one incoming road $n = 1$ and two outgoing roads $m = 2$. We use the same notation as in the previous section, i.e. we define $\gamma_{i,0}$ and the sets Ω_i as above depending on whether incoming or outgoing roads are considered. In this case the matrix A in the notation of Coclite/Piccoli is $A = (\alpha_{2,1}, \alpha_{3,1})$, where $\alpha_{2,1} + \alpha_{3,1} = 1$. The conditions of Coclite and Piccoli are

(1) $\gamma_1 \in \Omega_1, \alpha_{j,1}\gamma_1 \in \Omega_j$ for $j = 2, 3$.
(2) Maximize γ_1 w.r.t. (1).
(3) $\gamma_j = \alpha_{j,1}\gamma_1,\ j = 2, 3$.

Using as in the previous appendix $\Omega_i = [0, c_i], i = 1, 2, 3$ we obtain

$$\gamma_1 = \min\{c_1, c_2/\alpha_{2,1}, c_3/\alpha_{3,1}\}. \tag{21}$$

This is exactly, what is known as the FIFO (first in, first out) model of a junction [12].

A typical situation is the following: suppose one of the outgoing roads is completely filled, i.e. $\rho_{j,0} = \rho_{\max}$. Then the resulting flux at the junction will be $\gamma_{i,0} = 0, i = 1, 2, 3$. No car can pass the junction.

For highways one observes, that this is not always an appropriate description. Drivers wishing to go for the full road will not necessarily block the whole junction. Models treating such a situation are so-called NON-FIFO models, [44]. Such a model can be included in the above context by setting

(1) $\gamma_j \in \Omega_j$ and $\gamma_j/\alpha_{j,1} \in \Omega_1$ for $j = 2, 3$.
(2) Maximize γ_j w.r.t. (1) for $j = 2, 3$.
(3) $\gamma_1 = \sum_{j=2}^{3} \gamma_j$.

We obtain as before that $\gamma_i \in \Omega_i$ for $i = 1, 2, 3$. Moreover

$$\begin{aligned}\gamma_j &= \min\{\alpha_{j,1}c_1, c_j\},\ j = 2, 3\\ \gamma_1 &= \min\{\alpha_{2,1}c_1, c_2\} + \min\{\alpha_{3,1}c_1, c_3\}.\end{aligned}$$

We observe that the flux generated by the NON-FIFO conditions is greater or equal to the flux of the FIFO conditions.

2.1.3 A Smooth Approximations of Junctions by Multi-Lane Models

Traffic flow on unidirectional roads at a motorway junction is in reality described by a number of roads, which merge or disperse. A suitable accurate model for these intersections are multi-lane equations. Examples can be found in [37] or [23]. For the Aw/Rascle approach the full multi-lane equations with N roads labeled by α and with $\rho_\alpha(x,t)$ as density and $u_\alpha(x,t)$ as velocity on lane α read using the Kronecker symbol $\delta_{\alpha,\beta}$:

$$\begin{aligned}\partial_t\rho_\alpha + \partial_x(\rho_\alpha u_\alpha) &= \Big(\frac{1}{T^L_{\alpha-1}}\rho_{\alpha-1} - \frac{1}{T^R_\alpha}\rho_\alpha\Big)(1-\delta_{\alpha,1})\\ &\quad+\Big(\frac{1}{T^R_{\alpha+1}}\rho_{\alpha+1} - \frac{1}{T^L_\alpha}\rho_\alpha\Big)(1-\delta_{\alpha,N})\\ \partial_t(\rho_\alpha u_\alpha) + \partial_x(\rho_\alpha u_\alpha^2) + c(\rho_\alpha)\partial_x u_\alpha &= \Big(\frac{1}{T^L_{\alpha-1}}u_{\alpha-1}\rho_{\alpha-1} - \frac{1}{T^R_\alpha}u_\alpha\rho_\alpha\Big)(1-\delta_{\alpha,1})\\ &\quad+\Big(\frac{1}{T^R_{\alpha+1}}u_{\alpha+1}\rho_{\alpha+1} - \frac{1}{T^L_\alpha}u_\alpha\rho_\alpha\Big)(1-\delta_{\alpha,N})\\ &\quad+\frac{1}{T(\rho_\alpha)}\rho_\alpha[U(\rho_\alpha) - u_\alpha],\end{aligned}$$

where T^L_α and T^R_α are lane changing rates from lane α to the left or right. They depend on $\rho_{\alpha+1}$ and $\rho_{\alpha-1}$. For a detailed derivation of these lane changing rates and of T and c from microscopic/kinetic models we refer the reader to [37]. Similar to the above considerations the following scalar multi-lane model is a simplification of these equations:

$$\begin{aligned}\partial_t\rho_\alpha + \partial_x(\rho_\alpha U(\rho_\alpha)) &= 1cm\Big(\frac{1}{T^L_{\alpha-1}}\rho_{\alpha-1} - \frac{1}{T^R_\alpha}\rho_\alpha\Big)(1-\delta_{\alpha,1})\\ &\quad+\Big(\frac{1}{T^R_{\alpha+1}}\rho_{\alpha+1} - \frac{1}{T^L_\alpha}\rho_\alpha\Big)(1-\delta_{\alpha,N}).\end{aligned}$$

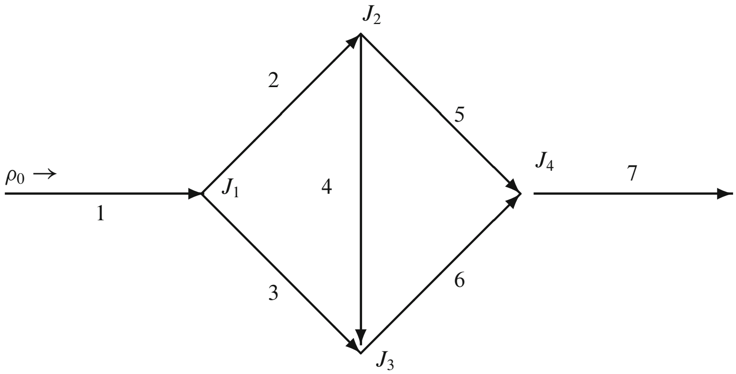


Fig. 1 Prototype of a network

In general, on the whole network, traffic is accurately described by these multi-lane equations. The basic idea of the approach is to use the following simplifications: On the roads of the network a cumulative model, treating the whole road by one equation not resolving the single lanes, is assumed to be sufficient as in the models described in the previous sections.

Near the junctions we introduce a zooming, i.e., a more detailed description of the situation. This is done by introducing—as usual in boundary layer theory, see, for example, [9]—a new enlarged spatial coordinate. Then, the asymptotic values of the resulting problem are used to determine the desired new states $\overline{\rho}_i$ at the junctions. A more detailed description of the procedure is given in Sect. 2.1.3.
To simplify the following discussion we focus on the prototype network of Fig. 1 with only two different types of junctions. As usual the number of incoming roads is denoted by n and the outgoing by m. The flux f of the above models is given by $f(\rho) = \rho U(\rho)$.

Modelling in the Case n=1 and m=2

We consider the case of two outgoing and one ingoing road. Similar to the model by Coclite/Piccoli it is assumed that the drivers have a tendency to take one of the outgoing roads, but in the multi-lane model the final decision depends also on the local traffic situation at the junction.

Example 1. The junction is modeled by two lane traffic, where the first lane splits into two roads, see Fig. 2. In the area (x_0, x_1) vehicles are changing only from the first to the second lane, i.e., we have $N = 2$, $T_2^R = 0$, T_1^L is chosen as a linearly increasing function for $x \in (x_0, x_1)$. The dependence of $1/T_1^L$ on ρ_2 is

$$1/T_1^L = \omega(1 - \rho_2)/T^*(x) \tag{22}$$

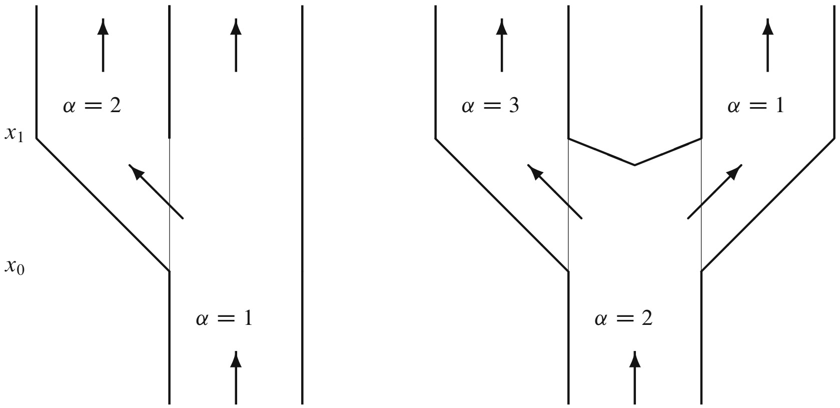


Fig. 2 Schematic diagram for examples 1 (*left*) and 2 (*right*)

with $1/T^*$ as before. Here the tendency of the drivers to take one of the two lanes is modeled by the variable $\omega \in [0, 1]$.

Thus, the equations are for $\alpha = 1, 2$ and $\forall t > 0$

$$\begin{array}{ll} x < x_0 & \partial_t \rho_1 + \partial_x(\rho_1 U(\rho_1)) = 0 \\ x \in (x_0, x_1) & \partial_t \rho_\alpha + \partial_x(\rho_\alpha U(\rho_\alpha)) = (-1)^\alpha \omega (1 - \rho_2)\rho_1 / T^*(x) \\ & \rho_2 U(\rho_2)(x_0, t) = 0 \\ x > x_1 & \partial_t \rho_\alpha + \partial_x(\rho_\alpha U(\rho_\alpha)) = 0 \end{array} \tag{23}$$

with the initial conditions

$$\begin{array}{llll} x < x_1 & \rho_1(x, 0) = \rho_{1,0}, & x > x_1 & \rho_1(x, 0) = \rho_{2,0}, \\ \mathrm{f} < x_1 & \rho_2(x, 0) = 0, & x > x_1 & \rho_2(x, 0) = \rho_{3,0}. \end{array} \tag{24}$$

Example 2. A more symmetric three-lane approach is the following, see Fig. 2: From x_0 to x_1 we consider three roads $i = 1, 2, 3$, where the left and right lanes are emerging from the middle one. The vehicles change from the middle lane to the other two lanes for $x \in (x_0, x_1)$. At x_1 the middle lane is closed. There are two nonzero interaction rates $1/T_2^L, 1/T_2^R$. They differ, according to different desired distributions of the incoming traffic to the two outgoing lanes. Thus the thresholds are with $1/T^*$ linear increasing in $x \in (x_0, x_1)$:

$$1/T_2^R = \omega_1 (1 - \rho_1)/T^*(x) \tag{25}$$

$$1/T_2^L = \omega_3 (1 - \rho_3)/T^*(x), \tag{26}$$

where $\omega_1 + \omega_3 = 1, \omega_i \in (0, 1)$ controls the distribution of the vehicles. This yields the following set of equations

$$\begin{array}{ll} x < x_0 & \partial_t \rho_2 + \partial_x(\rho_2 U(\rho_2)) = 0 \\ x \in (x_0, x_1) & \partial_t \rho_\alpha + \partial_x(\rho_\alpha U(\rho_\alpha)) = \omega_\alpha(1-\rho_\alpha)\rho_2/T^*(x), \text{ for } \alpha = 1,3 \\ & \partial_t \rho_2 + \partial_x(\rho_2 U(\rho_2)) = -[\omega_1(1-\rho_1) + \omega_3(1-\rho_3)]\rho_2/T^*(x) \\ x > x_1 & \partial_t \rho_\alpha + \partial_x(\rho_\alpha U(\rho_\alpha)) = 0, \ \alpha = 1,3 \\ x = x_0 & \rho_1 U(\rho_1) = \rho_3 U(\rho_3) = 0 \\ x = x_1 & \rho_2 U(\rho_2) = 0 \end{array} \tag{27}$$

with the initial conditions

$$\begin{array}{llll} x < x_1 & \rho_1(x,0) = 0, & x > x_1 & \rho_1(x,0) = \rho_{2,0}, \\ x < x_1 & \rho_2(x,0) = \rho_{1,0}, & x > x_1 & \rho_3(x,0) = \rho_{3,0}, \\ x < x_1 & \rho_3(x,0) = 0. & & \end{array} \tag{28}$$

Remark 2.7. In both models the final choice of the drivers which road to take depends on the actual density of the respective road. The drivers decide according to the local traffic situation. This is different to the treatment in the model of Coclite/Piccoli.

Remark 2.8. In a similar way a junction with two incoming and one outgoing road is treated, see [24].

Derivation of the Coupling Conditions

We consider the above problems and assume that the region of the junction is of order $\varepsilon << 1$, i.e., $[x_0, x_1]$ is rescaled like $[\varepsilon x_0, \varepsilon x_1]$. The lane changing rates have to be rescaled accordingly with $1/\varepsilon$:

$$\frac{1}{T_\alpha^{L,R}} \rightarrow \frac{1}{\varepsilon T_\alpha^{L,R}}.$$

As mentioned before we introduce near the junctions a zooming, i.e., a more detailed description of the situation. This is achieved by introducing a new enlarged spatial coordinate

$$x_J = \frac{x}{\varepsilon}. \tag{29}$$

Introducing x_J into the above problems and introducing a new time scale $t_j = \frac{t}{\varepsilon}$ leads to the same problems as before with time and spatial ranges extended to infinity. Thus, we are looking now for the asymptotic values, i.e., the solutions at $x = \pm\infty$ and $t = \infty$ of the problems defined in the above subsection. Numerically, we determine the solutions at $x = x_1^+$ and $x = x_0^-$, respectively. These values are then used as definition for the final values $\overline{\rho}_i$ at the junctions for the in- and outgoing lanes. More exactly, the desired asymptotic states $\overline{\rho}_i$ are given by

$$n = 1, m = 2 \text{ (Ex. 1)} : \overline{\rho_1} = \rho_2(x = x_0^-), \overline{\rho_2} = \rho_1(x_1^+), \overline{\rho_3} = \rho_2(x_1^+)$$
$$n = 1, m = 2 \text{ (Ex. 2)} : \overline{\rho_1} = \rho_2(x = x_0^-), \overline{\rho_2} = \rho_1(x_1^+), \overline{\rho_3} = \rho_3(x_1^+)$$

where $\rho_1(x)$, $\rho_2(x)$ and $\rho_3(x)$ are the solutions of the different problems described in the last subsections at time $t = \infty$. After calculating these values the solution in the sense of Definition 2.3 is calculated as solution to the problems defined by the initial values $\rho_{i,0}$ and $\overline{\rho}_i$ for each road i as in the previous approaches by Holden/Risebro and Coclite/Piccoli.

Remark 2.9. One observes easily that the coupling condition (4) is fulfilled.

2.1.4 Comparison of the Models

Here we show a numerical comparison of the scalar multi-lane model with the Coclite/Piccoli model.

We compare both models for the junction with $n = 1, m = 2$. For comparison we use the simple flux-function $f(x) = 4x(1-x)$, i.e., $U(\rho) = 4(1-\rho)$.

We consider the Coclite/Piccoli model with an equal distribution of vehicles to the two outgoing lanes:

$$A = (1/2, 1/2)^T$$

and determine the resulting states $\bar{\rho}_1, \bar{\rho}_2, \bar{\rho}_3$ for different initial states $\rho_{1,0}, \rho_{2,0}, \rho_{3,0}$.

For the scalar multi-lane model we use Example 2 and assume an equal tendency of the drivers to choose one of the two lanes:

$$\omega_1 = \omega_3.$$

With this choice of parameters a simulation of the full system is done using a second order method for hyperbolic equations with relaxation term, see [34]. The asymptotic states $\bar{\rho}_1, \bar{\rho}_2, \bar{\rho}_3$ are determined using the same initial states as before. For the simulation the multi-lane model was calculated until the residuum $(res = f(\rho_2(x_0)) - f(\rho_1(x_1)) - f(\rho_3(x_1)))$ was less than a small constant h and the time derivative was less than h.

The results are compared in Figs. 3 and 4. The figures show the differences between the final states $\bar{\rho}_1, \bar{\rho}_2, \bar{\rho}_3$ of both models.

In Fig. 3 the above differences in final values $\bar{\rho}_1, \bar{\rho}_2, \bar{\rho}_3$ are shown for initial values $\rho_{3,0} = 0$ and 10×10 different initial states $\rho_{1,0}$ and $\rho_{2,0}$. Since the differences are very small, we use in three of the four pictures in Fig. 3 a logarithmic scale. The last picture shows the absolute difference in normal scale between both models.

One observes that the two models give nearly coinciding results if the outgoing roads were initially empty. In case the roads are initially not empty larger differences in the behaviour of the models can be seen. For example in Fig. 4 the case $\rho_{3,0} = 7/9$

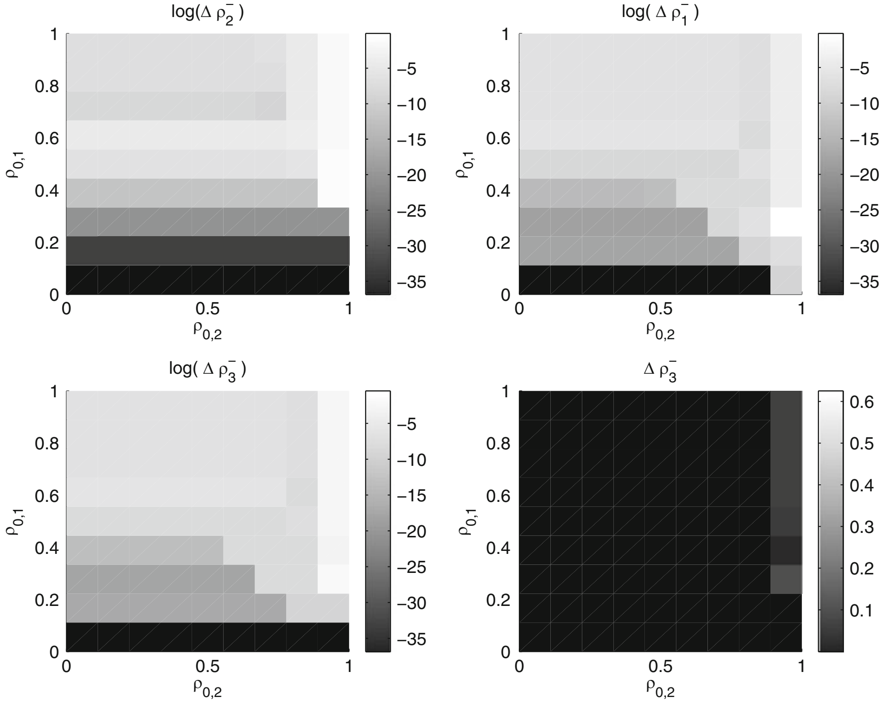


Fig. 3 Comparison of multi-lane and Coclite/Piccoli models for $\rho_{3,0} = 0$

is shown. The same 10×10 initial states $\rho_{1,0}$ and $\rho_{2,0}$ on road 1 and 2 are considered. One observes that the differences of the final state $\overline{\rho}_3$ on the outgoing road are in this case of order 1 for large values of $\rho_{1,0}$ and $\rho_{2,0}$.

To conclude a simple analytically solvable case with $\rho_{1,0} = \frac{1}{9}, \rho_{2,0} = \frac{8}{9}, \rho_{3,0} = 1$ is considered: the resulting stationary values of the multi-lane model are $\overline{\rho}_1 = \frac{1}{9}$, $\overline{\rho}_2 = \frac{8}{9}, \overline{\rho}_3 = 1$, since $f(\overline{\rho}_1) = f(\overline{\rho}_2)$. Road 3 is already full such that no vehicle can change to this lane. The results of the Coclite/Piccoli model are in contrast: $\overline{\rho}_1 = 1, \overline{\rho}_2 = 0, \overline{\rho}_3 = 1$, since by the definition of the admissible sets, it is $\overline{\rho}_3 \in \{1\}$ and thus $0 = f(\overline{\rho}_3) = \frac{1}{2} f(\overline{\rho}_1)$. Therefore we must have $f(\overline{\rho}_i) = 0$ for all i. Obviously, the different behaviour is due to the fact, that the choice of the drivers which way to take is fixed in the Coclite/Piccoli model, whereas in the multi-lane model the choice of the drivers depends on the local traffic situation.

Comments

- With the multi-lane approach one obtains results which are qualitatively comparable with the existing models. However, the quantitative comparison also shows that there are several situations where this model gives different results: we

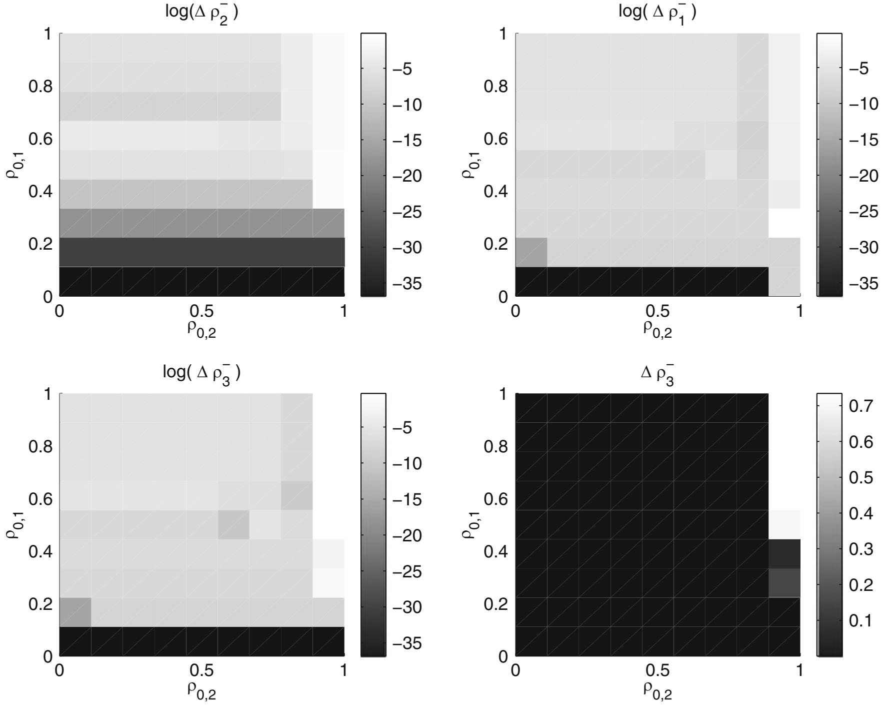


Fig. 4 Comparison of multi-lane and Coclite/Piccoli models for $\rho_{3,0} = 7/9$

observe differences between the Coclite/Piccoli approach and the multi-lane model for dense traffic situations. This is partially due to the fact that the wishes of the drivers are fixed in the Coclite/Piccoli model, whereas they depend on the local traffic situation in the multi-lane model.

- To circumvent the high computation times of the multilane model, the necessary data (the asymptotic values) have to be stored before the computation of a large network in lookup tables.

2.2 Simplified Dynamics on the Network

In the following the roads are labeled by $j = 1, \dots, J$ and junctions are labeled by $i = 1, \dots, I$. Variables without index indicate the whole vector of all indexed quantities. We assume for sake of simplicity to have a network with only one incoming and one outgoing road. This assumption is not strict, all results can be easily obtained in the case of several ingoing and outgoing roads.

We allow only junctions with a total of three roads. Thus having two different situations, i.e. a junction where two roads merge to one or a junction, where a road disperses in two others. The coupling conditions from the last section imply for a dispersing junction i, that the matrix $A = A^{(i)}$ is given by

$$A^{(i)} = \begin{bmatrix} \alpha_i \\ 1 - \alpha_i \end{bmatrix} \in \mathbf{R}^{2\times 1}. \tag{30}$$

The value $0 < \alpha_i < 1$ distributes the flux from the ingoing to the outgoing roads.

We derive from the PDE model two different simplifications of traffic flow on a network. The first one is based on a simple finite volume spatial discretization of the PDE model. The second one is based on Front-Tracking.

2.2.1 An ODE-Type Model

We derive an approximate model by considering the evolution of density averages on the road. Consider the equation $(\rho_j)_t + f_j(\rho_j)_x = 0$ for $x \in [a,b], t \in [0,T]$. Integrating over $[a,m]$ and $[m,b], a < m = a + \frac{b-a}{2} < b$ yields

$$\partial_t \rho_j^{(a)}(t) = -\frac{2}{L}\Big(f_j(\rho_j(m,t)) - f_j(\rho_j(a,t))\Big) \tag{31}$$

$$\partial_t \rho_j^{(b)}(t) = \frac{2}{L}\Big(f_j(\rho_j(m,t)) - f_j(\rho_j(b,t))\Big) \tag{32}$$

where $L = b - a$ and $\rho_j^{(a)}(t) = \frac{2}{L}\int_a^m \rho_j(x,t)dx$ and $\rho_j^{(b)}(t) = \frac{2}{L}\int_m^b \rho_j(x,t)dx$. We use the following approximation for $\rho_j(m,t)$:

$$\rho_j(m,t) = \frac{1}{2}\big(\rho_j^{(a)}(t) + \rho_j^{(b)}(t)\big). \tag{33}$$

Thus, one obtains in a straightforward way a coupled system of two ordinary differential equations. Initial conditions are $\rho_j^{(a)}(0) = \frac{2}{L}\int_a^m \rho_{j,0}(x)dx$ and $\rho_j^{(b)}(0) = \frac{2}{L}\int_m^b \rho_{j,0}(x)dx$.

The above ODE model is then closed using for $\rho_j(a,t)$ and $\rho_j(b,t)$ the coupling conditions. We set

$$\partial_t \rho_j^{(a)}(t) = -\frac{2}{L}\Big(f_j(\rho_j(m,t)) - f_j(\bar{\rho}_j^a(t))\Big) \tag{34}$$

$$\partial_t \rho_j^{(b)}(t) = \frac{2}{L}\Big(f_j(\rho_j(m,t)) - f_j(\bar{\rho}_j^b(t))\Big) \tag{35}$$

with the following approximation of the coupling conditions:

$$\bar{\rho}_j^a(t) = F_a^j(\rho_j^{(a)}(t), \rho_k^{(a/b)}(t), \rho_l^{(a/b)}(t)) \tag{36}$$

$$\bar{\rho}_j^b(t) = F_b^j(\rho_j^{(b)}(t), \rho_m^{(a/b)}(t), \rho_n^{(a/b)}(t)) \tag{37}$$

where a or b are chosen for in- and outgoing roads at the junctions respectively. In the above formulas we assume that road j connects two junctions. At $x = a$ we have a junction with the roads j, k, l and the function F_a for the coupling. Similarly at the junction at $x = b$ we have the roads j, m, n and the function F_b.

Finally, we discretize in time with stepwidth τ. We call this time the "update time" of the ODE model. The full model for a road j with fixed time τ reads

$$\begin{aligned}
\rho_j^{(a)}(t+\tau) - \rho_j^{(a)}(t) &= -\frac{2\tau}{L}\Big(f_j(\frac{\rho_j^{(a)}(t) + \rho_j^{(b)}(t)}{2}) - f_j(\bar{\rho}_j^a(t))\Big) \\
\rho_j^{(b)}(t+\tau) - \rho_j^{(b)}(t) &= \frac{2\tau}{L}\Big(f_j(\frac{\rho_j^{(a)}(t) + \rho_j^{(b)}(t)}{2}) - f_j(\bar{\rho}_j^b(t))\Big) \\
\bar{\rho}_j^a(t) &= F_a^j(\rho_j^{(a)}(t), \rho_k^{(a/b)}(t), \rho_l^{(a/b)}(t)) \\
\bar{\rho}_j^b(t) &= F_b^j(\rho_j^{(b)}(t), \rho_m^{(a/b)}(t), \rho_n^{(a/b)}(t))
\end{aligned}$$

We have the following remark on the choice of the update time τ.

Remark 2.10. The above approach can be seen as Finite-difference-discretization of the partial differential equation with only three points in space. Considering the CFL condition for the scheme we have a restriction to the update time in the ODE model, i.e. to be more precise we have

$$\tau \leq \frac{L}{2\max\limits_{x,j}\{f_j'(x)\}} \tag{38}$$

Remark 2.11. The model presented in this section is based on a coarse discretization of the PDE model. Thus, in particular, features like the speed of propagation of disturbances are not any more the same as in the original model.

2.2.2 An Algebraic Approach

In this section the traffic flow model is reduced to a system of algebraic equations. This is achieved by considering a simplified situation concerning the inflow into the network and tracking single waves through the network. In contrast to the static network models often used by traffic engineers the present approach still contains simplified dynamics—being at the same time not much more complicated and expensive from a computational point of view. For the following we assume that no backwards going shock wave appears, this means optimizing in a way, s.t. no traffic jam occurs. We start with an initially empty network and refer to the end of the section for the case of partially filled networks. Moreover, we restrict for the moment to constant inflow ρ_0.

We assign two values $\rho_{j,0} \in \mathbf{R}$ and $t_j \in \mathbf{R}^+$ to each road j in the network. The value $\rho_{j,0}$ is an approximation of the density $\rho_j(x,t)$ while t_j denotes the arrival time of a wave at road j, see below.

We use the following bounds for $\rho_{j,0}$ and t_j.

$$0 \leq \rho_{j,0} \leq \sigma, \ 0 \leq t_j \leq T. \tag{39}$$

The assumption on t_j is obvious, but the assumption on $\rho_{j,0}$ is critical. It ensures, that the direction of the flow through the network is one-way only. For the ingoing road j_0 to the network we set $\rho_{j_0,0} = \rho_0$. Then we set the values $\rho_{j,0}$ such that they satisfy the coupling conditions at the junctions. This determines the values $\rho_{j,0}$ and the conditions simplify under the restriction (39) and formula (30). We can express them as algebraic equations:

For a node $i \in \{1, \ldots, I\}$ with the roads k, l, m where k is the incoming and l, m are the outgoing roads we have

$$\rho_{l,0} = f_l^{-1}(\alpha^{(i)} f_k(\rho_{k,0})), \rho_{m,0} = f_m^{-1}((1-\alpha^{(i)}) f_k(\rho_{k,0})), \tag{40}$$

when $0 < \alpha < 1$ distributes in direction of road l. For a junction with two incoming k, l and one outgoing road m we obtain

$$\rho_{m,0} = f_m^{-1}(f_l(\rho_{l,0}) + f_k(\rho_{k,0})). \tag{41}$$

The solution to (40), resp. (41) is unique, iff it exists. If the capacity of the outgoing roads of any junction is not sufficient there is no solution $\rho_{j,0}$ subject to (39) and (40), resp. (41). This reflects the occurrence of a backwards going shock wave.

Remark 2.12. As an example note, that for a flux functions of the type $f_k(x) = 4x(1-x/M_k)$ the conditions read $2\rho_{l,0} = M_l - \sqrt{M_l^2 - \alpha^{(i)} M_l f_k(\rho_{k,0})}$ and similar for $\rho_{m,0}$. For the other junction we obtain $2\rho_{m,0} = M_m - \sqrt{M_m^2 - M_m \chi}$, $\chi = f_l(\rho_{l,0}) + f_k(\rho_{k,0})$.

The arrival times t_j defined by (42) and (43) describe an approximation of the time when the step function arrive at road j. Starting with a Riemann problem

$$\partial_t \rho_j + \partial_x f_j(\rho_j) = 0, \quad \rho_j(x,0) = \begin{bmatrix} \rho_{j,0} & x \leq a \\ 0 & x > a \end{bmatrix}$$

with concave fluxfunction f_j, a rarefaction wave is the correct solution. We simplify this by approximating the wave by a discontinuity. We restrict ourselves to track only one shock on each road. The speed of the wave is approximated with the Rankine-Hugoniot speed $s_j = \frac{f_j(\rho_{j,0})}{\rho_{j,0}}$. The arrival times are approximated as follows: For the ingoing road j_0 we set $t_{j_0} = 0$. In the case of a junction, where one road j disperse in two others k, l we set

$$t_k = t_l = t_j + \frac{b-a}{s_j} \tag{42}$$

where $s_j = \frac{f_j(\rho_{j,0})}{\rho_{j,0}}$.

In the case of a junction with two incoming roads k, l and one outgoing road j the situation is more complicated. We set

$$t_j = (t_l + \frac{b-a}{s_l})\frac{\rho_{l,0}}{\rho_{l,0}+\rho_{k,0}} + (t_k + \frac{b-a}{s_k})\frac{\rho_{k,0}}{\rho_{l,0}+\rho_{k,0}}. \tag{43}$$

This choice is motivated by the following calculations: Let $t^{(1)} < t^{(2)}$ denote the time, when the shocks from road k and l reach the beginning of road j. Thus we have the values $\rho_{k,0}, \rho_{l,0}$ given. We again assume to have one shock on road j instead of rarefaction waves. The travelling speeds are given by $s_1 = \frac{f(\rho_{k,0})}{\rho_{k,0}}$ and $s_2 = \frac{f_j(\rho_{j,0})-f_j(\rho_{k,0})}{\rho_{j,0}-\rho_{k,0}}$. The values $\rho_{j,0}$ are determined by the coupling condition, i.e. $f(\rho_{j,0}) = f(\rho_{k,0}) + f(\rho_{l,0})$. Then we have

$$\begin{aligned}\int_0^T \int_a^b \rho_j(x,t)dxdt &= (T-t^{(1)})(b-a)\rho_{1,0} - \frac{\rho_{k,0}}{2s_1}(b-a)^2 \\ &+ (T-t^{(2)})(b-a)(\rho_{j,0}-\rho_{k,0}) - \frac{\rho_{j,0}-\rho_{k,0}}{2s_2}(b-a)^2\end{aligned}$$

The idea is to approximate the above integral by $(T-t_j)(b-a)\rho_{j,0} - \frac{\rho_{j,0}}{2s_j}(b-a)^2$. If f is linear, then the correct choice for t_j would be:

$$t_j = t^{(1)}\frac{\rho_{1,0}}{\rho_{1,0}+\rho_{2,0}} + t^{(2)}\frac{\rho_{2,0}}{\rho_{1,0}+\rho_{2,0}} \tag{44}$$

This is used as well as an approximation in the nonlinear case. Finally, to obtain formula (43) we use (44) together with $t^{(1)} = (t_l + \frac{b-a}{s_l})$ and $t^{(2)} = (t_k + \frac{b-a}{s_k})$.

Thus we have defined a purely algebraic model for traffic flow on road networks without backwards going shock waves.

Remark 2.13. The treatment of a partially filled network is also possible. Assume we have the initial densities $p_{j,0}$ on the road j given, where all values are consistent with the conditions at the junctions. $p_{j,0}$ is constant for the whole road and such that $p_{j,0} < \sigma_j$. We start, as before, with an inflow $\rho_{1,0} < \sigma$ as above. Then similar considerations as the above, yield the following expression for an integral on j with $(L := b-a)$

$$\int_0^T \int_a^b \rho(x,t)dxdt = Lt_j p_{j,0} + L\rho_{j,0}(T-t_j) - \frac{\rho_{j,0}-p_{j,0}}{2s_j}L^2 \tag{45}$$

where now s_j is given by

$$s_j = \frac{f_j(\rho_{j,0}) - f_j(p_{j,0})}{\rho_{j,0} - p_{j,0}} \tag{46}$$

Using this definition of s_j one can approximate the arrival times of the ingoing wave t_j in the same way as before with s_j as in (46).

Remark 2.14. Nonconstant initial data, for example piecewise constant initial data, can be treated in the same way. For each wave the arrival times have to be tracked and we have to assume that the waves do not interact.

Remark 2.15. The above model contains simplifications at several points. However, still the dynamical behaviour is similar to that of the PDE model in the simple case considered here. If the ingoing flow structure is becoming more complicated, in particularly time dependent as discussed in Remark 2.14 the present procedure will be similar to a Front-Tracking approximation of the PDE.

2.3 *Optimization*

In this section we consider the optimal distribution of traffic flow in a given network. The notation is as in Sect. 2.2. As mentioned above the network consists of junctions with three roads with either two joining or two dispersing roads. The junctions with two dispersing roads are numbered with $i = 1, \dots, M$. Junction i is assumed to have an ingoing road j and outgoing roads k, l. Using again the notation of Sect. 2.2, i.e. formula (30) we have M matrices $A^{(i)}, i = 1, \dots, M$ given. Each matrix $A^{(i)}$ is described by $\alpha_i \in \mathbf{R}$ giving the flux distribution. These are the control parameters to optimize the flow in the network. Hence, we have a total of M real-valued controls $\alpha = (\alpha_1, \dots, \alpha_M)$. The set of all possible controls is given by $S = [0, 1]^M$. These parameters are given by recommendations to the drivers to take one of the outgoing roads.

2.3.1 Cost Functionals

A Cost Functional Measuring the Outflow

A first functional could be obtained by measuring the flow on the outgoing road. For a given time period $[0, T]$ we want to distribute the traffic in the network, such that the maximal possible outflow is achieved for a given inflow $\rho_0 = \rho_0(t)$. we define the cost functional J_1 as

$$J_1(\alpha; T, \rho_0) = \int_0^T f(\rho_{out}(x_0, t))dt. \tag{47}$$

The minimization problem to be solved is

$$\min_{\alpha \in S}(-J_1)$$

where S is the set of controls.

A Cost Functional Measuring the Time Cars Remain in the Network

Another functional is given by considering the time and space averaged densities. Minimizing this functional means finding the fastest way through the network, since by the fundamental diagram a low density is connected to high velocities. One obtains the functional

$$J_2(\alpha; T, \rho_0) = \sum_{j=1}^{J} \int_0^T \int_a^b \rho_j(x, t)dxdt. \tag{48}$$

The problem to be solved is $\min_{\alpha \in S} J_2,\ S$.

Relations Between the Functionals

A relation between the two functionals (47) and (48) is given by

Lemma 2.16. *$J_2 = -c_1 J_1 + c_2$ with $c_1 = c_1(\alpha) \geq 0$ and $c_2 \geq 0$, c_2 depending only on the inflow to the network.*

Density and flux are related via the hyperbolic conservation law, i.e. we have on each road j

$$\partial_t \rho_j + \partial_x f_j(\rho_j) = 0 \quad \forall x \in (a, b) t \in [0, T] \tag{49}$$

Assume an empty network at time $t = 0$ and $\rho_1(a, t) = \rho_0(t)$, where road 1 is the ingoing road to the network. Using the conservation law we have

$$\int_0^t \partial_t \rho_j(x, s)ds = \rho_j(x, t) - \rho_j(x, 0) = -\int_0^t \partial_x f_j(\rho_j(x, s))ds.$$

We get using the initial conditions

$$\begin{aligned} J_2 &= \sum_{j=1}^{J} \int_0^T \int_a^b \rho_j(x,t)dxdt = -\sum_{j=1}^{J} \int_0^T \int_a^b \int_0^t \partial_x f_j(\rho_j(x,s))dsdxdt \\ &= \int_0^T \int_0^t \sum_{j=1}^{J} -f_j(\rho_j(b,s)) + f_j(\rho_j(a,s))dsdt \end{aligned} \tag{50}$$

For a simplification we consider a network with only one ingoing and one outgoing road. Then one of the two cases hold for all interior roads j of the network.

1. $f_j(\rho_j(a,t)) = f_k(\rho_k(b,t)) + f_l(\rho_l(b,t))$ for a junction, where road k and l merge to j.
2. $f_l(\rho_l(a,t)) + f_k(\rho_k(a,t)) = f_j(\rho_j(b,t))$ for a junction, where road j disperse to k and l.

Let us denote the incoming road as road 1 and the outgoing as road n_0. Then we have $\forall s$

$$\sum_{j=2}^{n_0} f_j(\rho_j(a,s)) = \sum_{j=1}^{n_0-1} f_j(\rho_j(b,s)). \tag{51}$$

The index $j = 1$ appears in the second sum, since two roads $j_1, j_2 > 1$ are connected to the incoming road. Using the above equality (51) in (50):

$$J_2 = \int_0^T \int_0^t -f_{n_0}(\rho_{n_0}(b,s)) + f_1(\rho_1(a,s))dsdt \tag{52}$$

Thus, minimizing J_2 means maximizing a functional, which only depends on the out- and inflow of the network. To obtain the lemma we use the following equality for $\xi \in (0,T)$ and g continuous and positive.

$$\int_0^T \int_0^t g(s)dsdt = \int_0^T Tg(t) - tg(t)dt = (T-\xi)\int_0^T g(t)dt$$

Since $b_{n_0} > a_{n_0}$ is arbitrary we finally derive

$$J_2 = -c_1 \int_0^T f_{n_0}(\rho_{n_0}(b,t))dt + c_2 = -c_1 J_1 + c_2, \tag{53}$$

where $c_1 = T - \xi$, $\xi = \xi(\alpha)$ depending on the control parameters and

$$c_2 = \int_0^T \int_0^t f_1(\rho_1(a,s))dsdt$$

depending only on the inflow.

Functional for the ODE Model

The functional J_2 reads in the context of the ode model for a total of J roads and $t \in [0, T]$

$$J_2(\alpha; T; \rho_0) = \int_0^T \sum_{j=1}^{J} \frac{b-a}{2} \left(\rho_j^{(a)}(t) + \rho_j^{(b)}(t) \right) dt \tag{54}$$

where $\rho_j^{(a)}(t) = \frac{2}{b-a} \int_a^m \rho_j(x,t)dx$ and $\rho_j^{(b)} = \frac{2}{b-a} \int_m^b \rho_j(x,t)dx$ and m denotes the midpoint of the interval $[a, b]$.

Functional for the Algebraic Model

We derive an expression for the functional J_2 in the context of the algebraic model. In Sect. 2.2 we assumed, that we have only one wave of height $\rho_{j,0}$ on each road j. This step arrives at road j at time t_j and travels with speed s_j on road j. Hence we have the following approximation of the functional.

$$\int_0^T \int_a^b \rho_j(x,t)dxdt = (T - t_j)(b-a)\rho_{j,0} - \frac{\rho_{j,0}}{2s_j}(b-a)^2 \tag{55}$$

Summing up over all roads yields

$$J_2(\alpha; T, \rho_0) = \sum_{j=1}^{J} (T - t_j) L \rho_{j,0} - \frac{\rho_{j,0}}{2s_j} L^2 \tag{56}$$

where $s_j = f_j(\rho_{j,0})/\rho_{j,0}$, $L = b - a$ and $\rho_{j,0}, t_j$ are given by the expressions in Sect. 2.2. One proceeds similarly for partially filled networks.

2.3.2 The Optimization Problem for the PDE Network

On each edge j, the traffic dynamics are described by the Lighthill Whitham equations. For the following considerations we restrict ourselves to the Coclite/Piccoli model of junctions. Here, we restrict ourselves for simplicity to networks with only two types of junctions with a total of three incident roads, see Fig. 5. We use the following notations.

$$M_j = \max f_j(\rho),\ \sigma_j = \text{argmax} f_j(\rho). \tag{57}$$

We consider the case of a single junction v and constant initial data $\bar{\rho}_j$. We denote the given values by $\tilde{p}_j \in \mathbf{R}^+$ and the initial value by $\bar{\rho}_j \in \mathbf{R}^+$. δ_v^- denote the set of incoming roads and δ_v^+ the set of outgoing roads. Then, we consider the problem

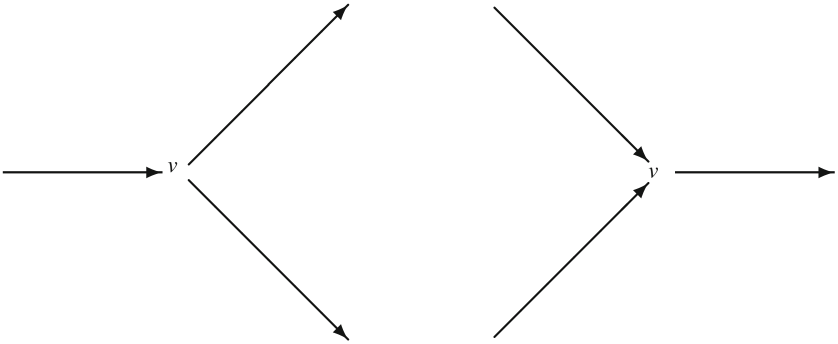


Fig. 5 Considered types for a junction v. The used notation is $\delta_v^- = \{j_0\}$, $\delta_v^+ = \{j_1, j_2\}$ (*left*) and $\delta_v^- = \{j_1, j_2\}$, $\delta_v^+ = \{j_3\}$ (*right*), respectively

$$\begin{aligned} &\partial_t \rho_j + \partial_x f_j(\rho_j) = 0 \\ &j \in \delta_v^+ : \rho_j(x,0) = \begin{bmatrix} \bar{\rho}_j & x > a_j \\ \tilde{p}_j & x \leq a_j \end{bmatrix} \text{ resp.} \\ &j \in \delta_v^- : \rho_j(x,0) = \begin{bmatrix} \tilde{p}_j & x \geq b_j \\ \bar{\rho}_j & x \leq b_j \end{bmatrix} \end{aligned} \tag{58}$$

A solution ρ_j of (58) satisfies also (3). Since the conservation of flux holds, there are certain restriction on $\tilde{p}_j$. They can be expressed explicitly by (60) using the following definition of the function $\rho \mapsto \tau(\rho)$:

$$\text{for given } \rho \text{ define } \tau = \tau_j(\rho) \text{ to be } \tau \neq \rho,\ f_j(\tau) = f_j(\rho). \tag{59}$$

The restrictions are

$$\begin{aligned} &j \in \delta_v^- : \quad \tilde{p}_j \in \begin{bmatrix} \{\bar{\rho}_j\} \cup]\tau_j(\bar{\rho}_j), \rho_{j,\max}] & \text{if } \bar{\rho}_j < \sigma_j \\ [\sigma, \rho_{j,\max}] & \text{if } \bar{\rho}_j \geq \sigma_j \end{bmatrix} \\ &j \in \delta_v^+ : \quad \tilde{p}_j \in \begin{bmatrix} [0, \sigma_j) & \text{if } \bar{\rho}_j < \sigma_j \\ \{\bar{\rho}_j\} \cup [0, \tau_j(\bar{\rho}_j)[& \text{if } \bar{\rho}_j \geq \sigma_j \end{bmatrix}. \end{aligned} \tag{60}$$

Depending to which intervall $\tilde{p}_j$ belongs, the wave generated by the Riemann problem (58) is either a shock wave or a rarefaction wave. As additional constraints we use

Case 1: Consider a single junction v where road j_0 disperse in two roads j_1 and j_2. A value $\alpha_v \in \mathbf{R}$ with $0 < \alpha_v < 1$ specifying the percentage of drivers coming from road j_0 and driving to j_1 is introduced. Then

$$\begin{aligned} f_{j_1}(\rho_{j_1}(a_{j_1}+,\cdot)) &= \alpha_v f_{j_0}(\rho_{j_0}(b_{j_0}-,\cdot)) \\ f_{j_2}(\rho_{j_2}(a_{j_2}+,\cdot)) &= (1-\alpha_v) f_{j_0}(\rho_{j_0}(b_{j_0}-,\cdot)). \end{aligned} \tag{61}$$

Unique values $\tilde{p}_j$, $j = j_0, j_1, j_2$, are be found by solving the maximization problem

$$\max f_{j_0}(\tilde{p}_j)\text{s.t.}(60), (61), (58) \tag{62}$$

Expression (62) does not allow an explicit representation of the boundary conditions. If we neglect the possibility of shock waves, especially backwards going shock waves on the incoming street j_0, the situation is much simpler. Therefore we assume the following

$$\bar{\rho}_j, \rho_j(x,t) \leq \sigma_j, \ \forall j. \tag{63}$$

Since we omit shock waves on j_0 we obtain instead of the maximization problem (62) an explicit formula for calculating $\tilde{p}_j$:

$$\tilde{p}_{j_0} = \bar{\rho}_{j_0}, \quad f_{j_1}(\tilde{p}_{j_1}) = \alpha_v f_{j_0}(\tilde{p}_{j_0}), \quad f_{j_2}(\tilde{p}_{j_2}) = (1-\alpha_v) f_{j_0}(\tilde{p}_{j_0}).$$

Equation (64) is well-defined due to (63) and yields unique values $\tilde{p}_{j_1}, \tilde{p}_{j_2}$.

Case 2: Consider a single junction v where roads j_1 and j_2 merge to j_3. Flux conservation through the junction implies

$$f_{j_3}(\rho_{j_3}(a_{j_3}+,\cdot)) = f_{j_1}(\rho_{j_1}(b_{j_1}-,\cdot)) + f_{j_2}(\rho_{j_2}(b_{j_2}-,\cdot)). \tag{64}$$

The unique values $\tilde{p}_j$, $j = j_i, i = 1, 2, 3$, are chosen as before. Define maximal possible fluxes by

$$j \in \delta_v^- = \{j_1, j_2\}: \quad \gamma_j = \left(\begin{matrix} f_j(\bar{\rho}_j) & \text{if } \bar{\rho}_j < \sigma_j \\ M_j & \text{if } \bar{\rho}_j \geq \sigma_j \end{matrix} \right]$$

$$j \in \delta_v^+ = \{j_3\}: \quad \gamma_j = \left[\begin{matrix} M_j & \text{if } \bar{\rho}_j < \sigma_j \\ f_j(\bar{\rho}_j) & \text{if } \bar{\rho}_j \geq \sigma_j \end{matrix} \right]$$

and solve the maximization problem:

$$\begin{aligned} &\text{If } \gamma_{j_1} + \gamma_{j_2} > \gamma_{j_3} \max \sum_{j \in \delta_v^-} f_j(\tilde{p}_j) \text{ s.t. } (64), (60) \text{ and } f_{j_1}(\tilde{p}_{j_1}) = f_{j_2}(\tilde{p}_{j_2}) \\ &\text{If } \gamma_{j_1} + \gamma_{j_2} \leq \gamma_{j_3} \max \sum_{j \in \delta_v^-} f_j(\tilde{p}_j) \text{ s.t. } (64), (60). \end{aligned} \tag{65}$$

Again we obtain an explicit representation of the boundary conditions when assuming (63).

$$\begin{aligned} \tilde{p}_j &= \bar{\rho}_j \;\; j = j_1 \text{ and } j = j_2 \\ f_{j_3}(\tilde{p}_{j_3}) &= \sum_{j \in \delta_v^-} f_j(\tilde{p}_j) \end{aligned} \tag{66}$$

Thus, in this simple case of no backwards travelling wave, the coupling conditions at the junctions are essentially given by the drivers wishes for a diverging junction. For a converging junction they are given by the equality of fluxes together with the requirement that the fluxes from the two ingoing roads are equal in the dense case.

Now, optimal control problems can be investigated. Typically, the average time spent by the drivers in the network is minimized. This means we consider the objective function

$$J(\alpha_1, \dots, \alpha_{|V|}) = \int_0^T \sum_{j=1}^{|E|} \int_{a_j}^{b_j} \rho_j(x,t)dxdt. \tag{67}$$

This function has to be minimized with respect to the control variables α_v. We solve the problem:

$$\min_{0<\alpha_1,\dots,\alpha_{|V|}<1} J(\alpha_1, \dots, \alpha_{|V|}) \tag{68}$$

$$\text{subject to: } \rho_j \text{ is solution of (3) with coupling conditions}$$

$$\text{at the junctions given by (62) and (65).}$$

A solution to this problem yields an optimal distribution of a traffic flow in a network including all dynamics, like jam propagation etc.

Alternatively we can optimize the above function in the case of no backwards going shock wave. This implies replacing conditions (62) and (65) by (64) and (66). However, even in this case optimization of networks with a large number of roads in reasonable time is beyond any computational possibility.

The Optimization Problem for the Simplified Nonlinear Model

In this section we consider the simplified model from Sect. 2.2. We start with an initially empty network and refer to the end of the section for the case of partially filled networks. Moreover, for simplicity, we restrict to constant inflow $\rho_{j,0}$ applied as boundary condition at the incoming road to the network. For the geometry of the network we use the same assumptions as in the previous section, i.e. we assume to have only junctions connecting at most three roads, like in Fig. 5.

The assumption of no backwards going shock waves is imposed as in (63), i.e.

$$\rho_j(x,t) \leq \sigma_j \;\; \forall j.$$

We assign two values $p_j \in \mathbf{R}$ and $t_j \in \mathbf{R}^+$ to each road j of the network. The value p_j is an approximation of the density $\rho_j(x,t)$ while t_j denotes the arrival time of a wave at road j. The following bounds are obvious.

$$0 \leq p_j \leq \sigma_j, \; 0 \leq t_j \leq T. \tag{69}$$

Due to (63) we can express the coupling conditions (64) and (66) in the form (64). We translate them in terms of p_j and obtain:

Case 1:

$$\begin{aligned} &\delta_v^- = \{j_0\}, \delta_v^+ = \{j_1, j_2\} \\ &p_{j_1} = f_{j_1}^{-1}(\alpha_v f_{j_0}(p_{j_0})), \quad p_{j_2} = f_{j_2}^{-1}((1-\alpha_v) f_{j_0}(p_{j_0})). \end{aligned}$$

Case 2:

$$\begin{aligned} &\delta_v^- = \{j_1, j_2\}, \delta_v^+ = \{j_3\} \\ &p_{j_3} = f_{j_3}^{-1}(f_{j_1}(p_{j_1}) + f_{j_2}^{-1}(f_j(p_{j_2}))) \end{aligned}$$

For the ingoing roads to the network we set $p_j = \rho_{j,0}$. In Case 1 the parameters $0 < \alpha_v < 1$ distribute traffic at junction v in the direction of road i. Hence, p_j is determined solely by fulfilling the coupling conditions at the junctions. The arrival times of the waves t_j are defined as in Sect. 2.2 using

$$s_j = \frac{f_j(p_j)}{p_j}$$

by the formula

$$t_{j_1} = t_{j_2} = t_{j_0} + \frac{b-a}{s_{j_0}}. \tag{70}$$

In the case of a junction with two incoming roads j_1, j_2 and one outgoing road j_3 the situation is more complicated. We set as in Sect. 2.2

$$t_{j_3} = (t_{j_1} + \frac{b-a}{s_{j_1}})\frac{p_{j_1}}{p_{j_1}+p_{j_2}} + (t_{j_2} + \frac{b-a}{s_{j_2}})\frac{p_{j_2}}{p_{j_1}+p_{j_2}}. \tag{71}$$

Finally, the full simplified nonlinear model reads with $L_j = b_j - a_j$ and $s_j = f_j(p_j)/p_j$:

$$\left.\begin{array}{r}\text{For junctions of merging type:}\\ t_k = (t_i + \frac{L_i}{s_i})\frac{p_i}{p_i+p_j} + (t_j + \frac{L_j}{s_j})\frac{p_j}{p_i+p_j}\\ p_k = f_k^{-1}(f_i(p_i) + f_j(p_j)).\\ \text{For junctions of dispersing type:}\\ t_i = t_j = t_k + \frac{L_k}{s_k}\\ p_i = f_i^{-1}(\alpha_v f_k(p_k)), \quad p_j = f_j^{-1}((1-\alpha_v) f_k(p_k)).\\ \text{For the road entering the network:}\\ p_j = \rho_0, t_j = 0.\end{array}\right\} \tag{72}$$

The objective function reads

$$J(\alpha; T, \rho_0) = \sum_{j\in E}(T - t_j)L_j p_j - \frac{p_j}{2s_j}L_j^2. \tag{73}$$

Herein T is a fixed time and ρ_0 is the inflow to the network. It turns out that also for this simplification the minimization problem min J subject to the constraints above still needs large computation times for very large networks due to the nonlinearities in the coupling conditions for p_j and t_j. For numerical results we refer to the subsequent sections.

The Optimization Problem for Linearized Models

In this section the previously introduced model is further simplified obtaining a linear model accessible to discrete optimization techniques. The basic idea is the reformulation of the above model in terms of the flux $q_j := p_j u^e(p_j)$. We introduce the notation

$$\tau_j(q_j) := \frac{1}{u^e(f_j^{-1}(q_j))} \tag{74}$$

and obtain $p_j = q_j \tau_j(q_j)$.

The coupling conditions at the junctions read

$$\left.\begin{array}{r}\text{For junctions of merging type:}\\ q_k = q_i + q_j\\ \text{For junctions of dispersing type:}\\ q_i + q_j = q_k.\\ \text{For all roads}\\ M_j \geq q_j \geq 0.\end{array}\right\} \tag{75}$$

Note that the control variable α_v does not appear in the above formulation. Therefore the values of q_i, q_j are not solely defined by q_k. The function J is given in terms of q_j by

$$J(q_j; T, \rho_0) = \sum_{j \in E} \Big(TL_j - t_j L_j - \frac{\tau_j(q_j) L_j^2}{2} \Big) \tau(q_j) q_j. \tag{76}$$

Then the complete model and the optimization problem reads

$$\left.\begin{aligned} \min_{q_j \ j \in E} J(q_j; T, \rho_0) \\ \text{where for junctions of dispersing type:} \\ t_i = t_j = t_k + L_k \tau_k(q_k) \\ q_i + q_j = q_k \\ \text{where for junctions of merging type:} \\ t_k = (t_i + L_i \tau_i(q_i)) \tfrac{q_i \tau_i(q_i)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)} + (t_j + L_j \tau_j(q_j)) \tfrac{q_j \tau_j(q_j)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)} \\ q_k = q_i + q_j \\ \text{where for roads ingoing to the network:} \\ q_j = f_0(\rho_0), t_j = 0 \\ \text{where for all roads:} \\ M_j \geq q_j \geq 0 \end{aligned}\right\} \tag{77}$$

This model is still equivalent to the nonlinear model described above. We derive different (linear!) models from this formulation and refer to the subsequent sections for numerical results.

Linear Models with Dynamics

The coupling conditions at the junctions are linear in q_j but nonlinear in t_j. We use different possibilities to linearize the coupling t_j. In the numerical tests it turns out that the crucial point is the proper discretization of the weight w appearing in the case of merging junctions, i. e.

$$w_i(q_i, q_j) := \frac{q_i \tau_i(q_i)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)}, \quad w_j(q_i, q_j) := \frac{q_j \tau_j(q_j)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)}$$

We propose two different approaches and compare the results numerically in the next section.

(**A**) We approximate

$$w_i, w_j \sim \tilde{w} = \frac{1}{2}$$

and calculate the first order Taylor expansion $\tilde{\tau}_j(q) = \tau_j(0) + q\tau_j'(0)$ as an approximation for $\tau_j(q)$. That means we linearize the model globally around 0. Neglecting higher order terms, we obtain the following linear equations:

$$\begin{aligned}&\text{Dispersing junctions}\\ &t_i = t_j = t_k + L_k \tilde{\tau}_k(q_k)\\ &\text{Merging junctions}\\ &t_k = \big(t_i + L_i \tilde{\tau}_i(q_i)\big)\cdot \tilde{w} + \big(t_j + L_j \tilde{\tau}_j(q_j)\big)\cdot \tilde{w}.\end{aligned} \tag{78}$$

(**B**) Instead of linearizing the functions globally, we discretize the problem using piecewise linear approximations. The junctions of merging type are now approximated by piecewise linear functions on triangles, a more refined approximation as in case A. For each junction k of the merging type consider

$$a_k(q_i, q_j) := L_i \tau_i(q_i) \frac{q_i \tau_i(q_i)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)} + L_j \tau_j(q_j) \frac{q_j \tau_j(q_j)}{q_i \tau_i(q_i) + q_j \tau_j(q_j)} \tag{79}$$

As an example note that for $f(\rho) = 4\rho(1 - \rho/M_i)$ and $M_i = M_j = 1$ the contour lines of a_k are given in Fig. 6.

We introduce $N_i \cdot N_j$ discretization points (ξ_v^k, η_w^k) with $0 = \xi_1^k < \xi_2^k < \ldots < \xi_{N_i-1}^k < \xi_{N_i}^k = M_i$ and $0 = \eta_1^k < \eta_2^k < \ldots < \eta_{N_j-1}^k < \eta_{N_j}^k = M_j$. Denote Δ a partition of the grid of discretization points into triangles and introduce a binary variable $y_{(p_1,p_2,p_3)}^k \in \{0, 1\}$ for each triangle $(p_1, p_2, p_3) \in \Delta$. The identification of the proper triangle corresponding to the incoming fluxes q_i, q_j is done by the next equations. Exactly one triangle has to be selected:

$$\sum_{(p_1,p_2,p_3)\in\Delta} y_{(p_1,p_2,p_3)}^k = 1. \tag{80}$$

Once one triangle is selected, the values of q_i, q_j can be encoded as convex combination of its corners. For this, introduce a continuous variable $\lambda_{v,w}^k \geq 0$ for each discretization point (ξ_v^k, η_w^k), which are coupled to q_i and q_j as follows:

$$q_i = \sum_{v=1}^{N_i} \sum_{w=1}^{N_j} \xi_v^k \cdot \lambda_{v,w}, \quad q_j = \sum_{v=1}^{N_i} \sum_{w=1}^{N_j} \eta_w^k \cdot \lambda_{v,w}. \tag{81}$$

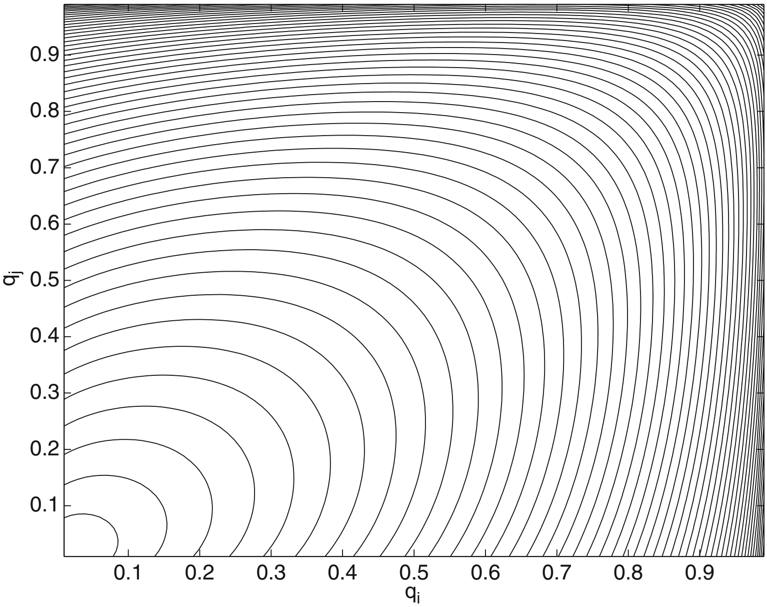


Fig. 6 Contour lines of the nonlinear weight function $a_k(q_i, q_j)$ for $q_i, q_j \in [0, 1]$

The convex combination condition is

$$\sum_{v=1}^{N_i} \sum_{w=1}^{N_j} \lambda_{v,w} = 1. \tag{82}$$

Only those three values $\lambda_{p_1}, \lambda_{p_2}, \lambda_{p_3}$ may be non-zero that correspond to the selected triangle by (80):

$$y^k_{(p_1,p_2,p_3)} \leq \lambda_{p_1} + \lambda_{p_2} + \lambda_{p_3}, \quad \forall\ (p_1, p_2, p_3) \in \Delta. \tag{83}$$

To introduce $\tilde{a}_k$ as a piecewise linear approximation of $a_k(q_i, q_j)$, we add the following equation to the model:

$$\tilde{a}_k = \sum_{v=1}^{N_i} \sum_{w=1}^{N_j} a_k(\xi_v, \eta_w) \cdot \lambda_{v,w} \tag{84}$$

The junctions of dispersing type are approximated as in case A, whereas for the junctions of merging type, we use a blending of $\tilde{a}$ as above and $\tilde{w}$ as in case A:

$$\begin{aligned}&\text{Dispersing junctions}\\ &t_i = t_j = t_k + L_k \tilde{\tau}_k(q_k)\\ &\text{Merging junctions}\\ &t_k = (t_i + t_j)\cdot \tilde{w} + \tilde{a}_k.\end{aligned} \tag{85}$$

For any linearization A or B we linearize the objective function (76) as follows. For every $j \in E$ we introduce D_q variables $0 \le y_i^j \le \frac{M_j}{D_q}$ and let the flux be represented by

$$q_j = \sum_{i=1}^{D_q} y_i^j. \tag{86}$$

Functional J is approximated by

$$\tilde{J}(q_j; T, \rho_0) := \sum_{j\in E} z_j, \tag{87}$$

where we introduce for every edge $j \in E$ and every $k = 1, \ldots, D_t$ the inequality

$$\begin{aligned}&\sum_{i=1}^{D_q}\left(G\left(\frac{(i+1)\cdot M_j}{D}, T\cdot 2^{k-D_t}\right) - G\left(\frac{i\cdot M_j}{D}, T\cdot 2^{k-D_t}\right)\right)\cdot \frac{D}{M_j}\cdot y_i^j\\ &\le z_j + M\cdot(1-u_{jk}),\end{aligned} \tag{88}$$

where M is a sufficiently big value and G is defined by

$$G(\xi,\zeta) := \left(T - \zeta - \frac{\tau(\xi)L_j}{2}\right)\cdot L_j\tau(\xi)\xi, \tag{89}$$

and we assume that $G(\cdot,\zeta)$ is convex for every $\zeta \in [0, T]$. Moreover, u_{jk} is a binary variable for every $j \in E$ and $k = 1, \ldots, D_t$, where $u_{jk} = 1$ if $t_j \le T\cdot 2^{k-D_t}$. Thus we add the following inequalities to the model:

$$t_j \ge T\cdot 2^{k-D_t}(1-u_{jk}), \tag{90}$$

for all $j \in E$ and $k = 1, \ldots, D_t$. Summarizing we obtain a linear mixed-integer model with dynamics given by

$$\left.\begin{aligned}
\min_{z_j, y_i^j, u_{jk}, \lambda^k_{v,w}, q_j, t_j} \quad & \tilde{J} \\
\text{where for junctions of dispersing type:} & \\
t_i = t_j = t_k + L_k \tilde{\tau}_k(q_k) & \\
q_i + q_j = q_k & \\
\text{where for junctions of merging type:} & \\
\text{case A: } t_k = \big(t_i + L_i \tilde{\tau}_i(q_i)\big) \cdot \tilde{w} + \big(t_j + L_j \tilde{\tau}_j(q_j)\big) \cdot \tilde{w} & \\
\text{case B: } t_k = (t_i + t_j) \cdot \tilde{w} + \tilde{a}_k & \\
q_k = q_i + q_j & \\
\text{where for roads ingoing to the network:} & \\
q_j = f_0(\rho_0), t_j = 0 & \\
\text{where for all roads:} & \\
M_j \geq q_j \geq 0 & \\
\text{and where } u_{jk}, z_j, \lambda^k_{v,w} \text{ are coupled as introduced.} &
\end{aligned}\right\} \tag{91}$$

Remark 2.17. In the above modelling we set the discretization points as $T2^{k-D_t}$ for $k = 1, \ldots, D_t$. This produces a log-scale distribution of discretization points in $[0, T]$. Other distributions are also possible. For example, if we identically distribute we obtain

$$t_j \geq T \frac{k-1}{D_t - 1}(1 - u_{jk}) \tag{92}$$

instead of (90). The proper choice depends on the size of the network geometry and the scaling of T.

Linear Model Without Dynamics

We assume $t_j = 0$ which models a static traffic flow network. We obtain a linear function $\tilde{J}$ from (76) by a piecewise linear approximation of J. For this, we introduce D variables $0 \leq y_i^j \leq \frac{M_j}{D}$ for every edge $j \in E$. Then the flux q_j is represented by

$$q_j = \sum_{i=1}^{D} y_i^j. \tag{93}$$

Now J is approximated by

$$\tilde{J}(q_j; T, \rho_0) := \sum_{j \in E} \sum_{i=1}^{D} \left(G\left(\frac{(i+1) \cdot M_j}{D}\right) - G\left(\frac{i \cdot M_j}{D}\right)\right) \cdot \frac{D}{M_j} \cdot y_i^j, \tag{94}$$

where G is defined by

$$G(\xi) := \left(TL_j - \frac{\tau(\xi)L_j^2}{2} \right) \cdot \tau(\xi) \cdot \xi. \tag{95}$$

Again, we assume $G(\cdot)$ to be convex. Summarizing, we have the following model

$$\left.\begin{aligned} \min_{y_i^j, q_j} \tilde{J} \\ \text{where for roads connected to a junction } v\text{:} \\ \sum_{j \in \delta_v^+} q_j = \sum_{j \in \delta_v^-} q_j \\ \text{where for roads ingoing to the network:} \\ q_j = f_0(\rho_0) \\ \text{where for all roads:} \\ M_j \geq q_j \geq 0 \\ \text{and } y_i^j \text{ satisfies (93)} \end{aligned}\right\} \tag{96}$$

Remark 2.18. **Min cost flow model**

We again assume $t_j = 0$, i.e. the static network case. Instead of a piecewise linear approximation of our objective function (76) we additionally assume a simplified dynamic: If the function

$$u_j^e(\rho) = c_j$$

is constant for all j, then by definition

$$\tau(q_j) = \frac{1}{c_j}.$$

The function (76) reads

$$\bar{J}(q_j; T, \rho_0) = \sum_{j \in E} \omega_j q_j, \tag{97}$$

where ω_j are constants given by

$$\omega_j = T\frac{L_j}{c_j} - \frac{L_j^2}{2c_j^2}.$$

Together with the linear coupling conditions and the lower bounds for q_j we obtain the classical min cost flow problem:

$$\left.\begin{aligned}
&\min_{q_j} \sum_{j \in E} \omega_j q_j \\
&\text{where for roads connected to a junction } v\text{:} \\
&\sum_{j \in \delta_v^+} q_j = \sum_{j \in \delta_v^-} q_j \\
&\text{where for roads ingoing to the network:} \\
&q_j = f_0(\rho_0) \\
&\text{and where for all roads:} \\
&M_j \geq q_j \geq 0
\end{aligned}\right\} \tag{98}$$

Unfortunately the assumption $u_j^e = \text{constant}_j$ is not a realistic approximation of a typical fundamental diagram. For reasonable fundamental diagrams we refer to [39]. At least we have to assume $u_j^e(x)$ is linear.

Remark 2.19. We note that for the linear models there is a strong connection to the traffic flow models proposed by Möhring et al. see, for example [40, 41]. Especially the occurrence of the so-called transit-times shows the close relation between the models. However, the cost function for the linear problem differs due to the derivation starting with partial differential equations. The starting point of the models introduced in [41] are the "transit times" τ_e which are assumed to be known functions. They describe the time needed by a flow to pass the arc e. In our formulation the "transit times" are the functions derived by (74), i.e.

$$q \to \tau_j(q) L_j.$$

The case of constant transit times is named "static flow problems" in [41]. In our introduced model this reflects the situation $u_j^e(\rho)$ constant. As pointed out by Möhring, et. al. this can not be a realistic assumption. Therefore, they introduced "static traffic flows with congestion" by assuming a dependency of τ_e on q. In our model this approach is reflected by the introduced linear model without dynamics. However, we see by our derivation that congestion in form of backwards going shock waves are not covered by those models, c.f. numerical results below.

2.3.3 Numerical Comparison of Approaches

In this subsection we compare the computing times for different models and networks.

Testcase for Comparing Network Models

For the purpose of comparing the models we introduce a network with two controls and seven roads as in Fig. 7. As an example we use the smooth and concave family of flux functions

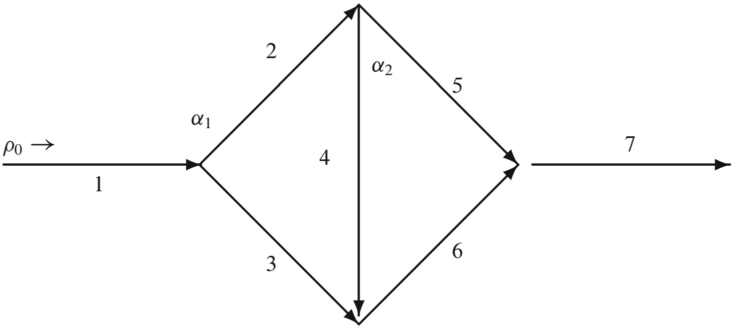


Fig. 7 Example of a network

$$f_j(\rho) = \rho u_j^e(\rho) = 4\rho(1 - \rho/M_j). \tag{99}$$

The function $\tau_j(\rho)$ is then given by,

$$\tau_j(q) = \frac{M_j}{2\left(M_j + \sqrt{M_j^2 - M_j q}\right)}, \quad 0 \leq q \leq M_j. \tag{100}$$

If not stated otherwise we assume

$$T = 5 \text{ and } L_j := b_j - a_j = 1 \ \ \forall j = 1, \ldots, 7. \tag{101}$$

We define q_0 to be the known inflow given at $x = a_1$.

Comparison of the Values of the Objective Function

We compare the derived models on the sample network. We compute the objective function of the corresponding model for all admissible choices of the control variables α_1 and α_2. In the context of the linear models this implies to compute the objective for all choices $q_1, \ldots, q_7$ satisfying the constraints. As described in Sect. 2.3.2 the fluxes q_j and the controls are related. For example we obtain for the first roads of our sample network

$$q_1 = q_0, \ q_2 = \alpha_1 q_1, \ q_3 = (1 - \alpha_1) q_1.$$

In all subsequent plots we draw contour lines of the objective function against α_1 and α_2. We choose different maximal fluxes M_j on the roads to obtain different test cases.

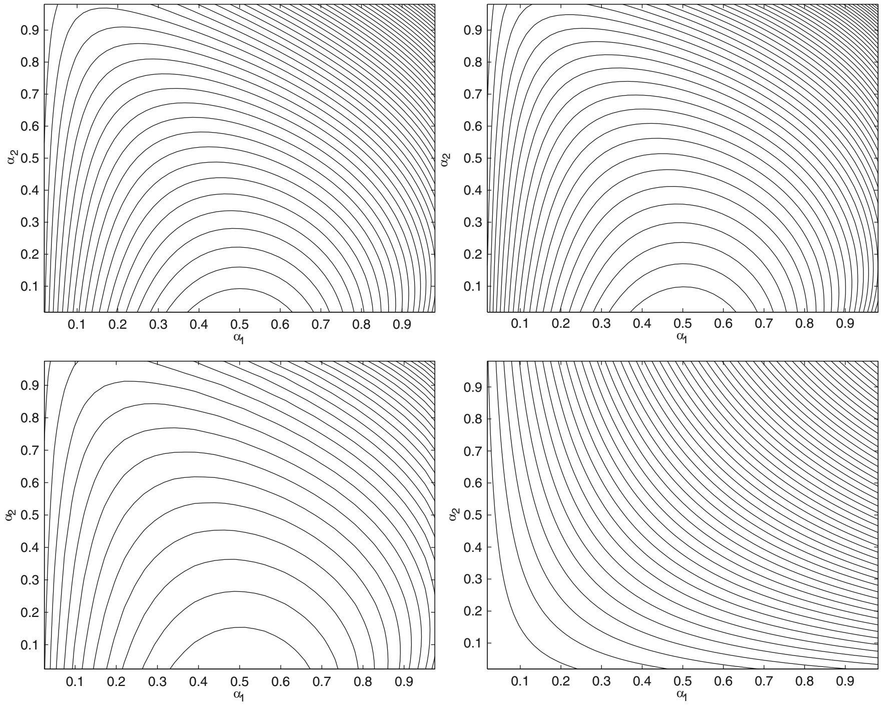


Fig. 8 Test Case 1: Contour lines of the functions for partial differential equation, simplified nonlinear, linear without dynamics and min cost flow model (*top left to bottom right*)

Test Case 1: Free Flow

We set $M_j = 1$ for all roads and $q_0 = 96\% M_1$. We compute the objective function (67) by a trapezoid rule. The underlying partial differential equations is solved by a first-order Godnuov scheme with $N = 100$ discretization points for each road j. The objective function (73) is computed by the formulas given in Sect. 2.3.2. For the linear models we computed the function with $D = 1{,}000$ variables for each edge j. Note, that the function $\xi \to G(\xi)$ is at least monotone for the choice (99). For comparison we include a plot of the function for the min cost flow problem (97) where we set $u_j^e(\rho) = 2$ for this calculation. The results are given in Fig. 8. The minimizer of all problems is $(\alpha_1, \alpha_2) = (\frac{1}{2}, 0)$. In case of the min cost flow problem we loose the uniqueness of the minimizer. Furthermore, the qualitative behaviour differs significantly from the other models.

Test Case 2: Backwards Going Shock Waves

When deriving the simplified models we neglected backwards going shock waves. This was an essential part of the simplification of the dynamics. We compare

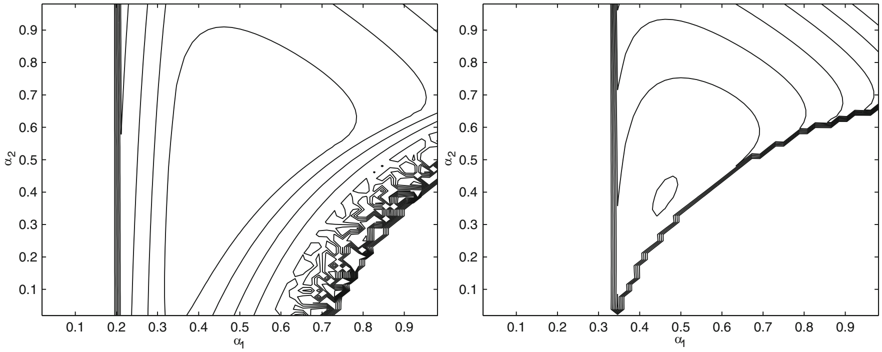


Fig. 9 Test Case 2: Contour lines for the functions for pde and simplified nonlinear model (*left to right*)

the simplified nonlinear model (72) with the model based on partial differential equations (68) in a case with backwards going shock waves. We set $M_1 = M_2 = M_4 = M_6 = 2,\quad M_3 = 1,\quad M_5 = 0.5$ and $q_0 = 75\% M_1$. We used the same discretization as previously and compare the contour lines of (67) and (73) in Fig. 9. We observe that in the pde case the domain of admissible controls is larger than in the case of the simplified nonlinear model. This effect is due to backwards going shock waves which occur on some roads in the pde model. Controls generating these waves are not admissible in the simplified nonlinear model. In our special case the region for the optimal control coincides. We skip results on the linear model since they approximate the algebraic model.

Test Case 3: Linearization of the Dynamics

In this case we consider the influence of the linearizations. We use the following setting $M_j = 2,\ q_0 = 96\% M_1$ and $L_1 = L_7 = L_5 = 2, L_4 = L_6 = 1, L_2 = 2.5, L_3 = 15$. We compare the qualitative behaviour of the objective function for the simplified nonlinear model with the linear models with dynamics given in Sect. 2.3.2. We compare the different discretization, Cases A and B. The results are given in Fig. 10. We used $D_q = D_t = 100$ variables for the discretization of the flux and the time on each road for any linearized model. We calculate Cases B with $N_i = N_j = 5$ and $N_i = N_j = 25$, respectively, discretization points for each junction of the merging type.

Optimization on the Sample Network

We consider the optimization problems introduced and compare computing times on the sample network.

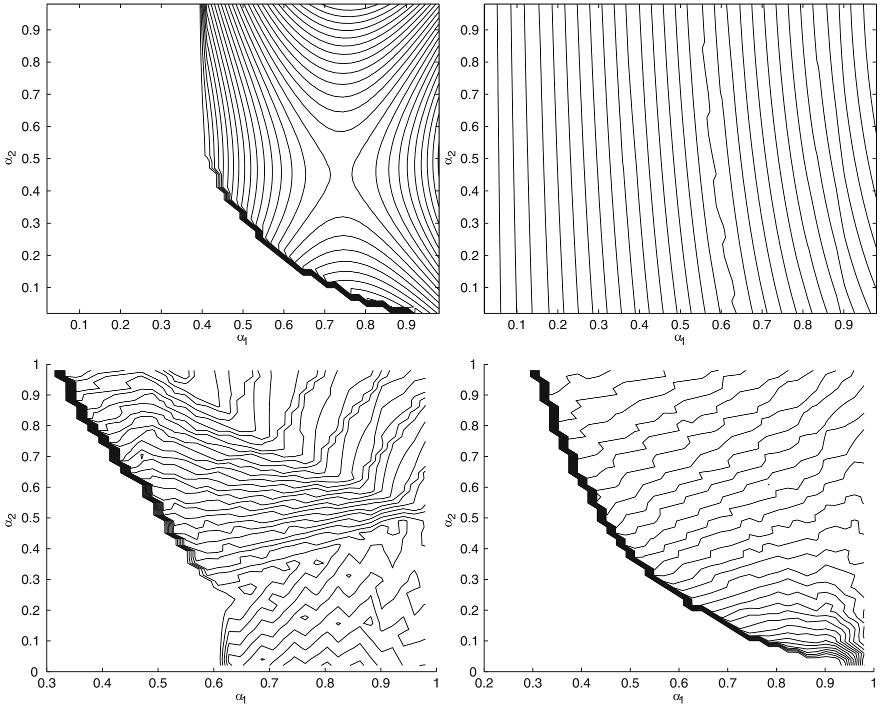


Fig. 10 Test Case 3: Contour lines for the functions for simplified nonlinear (73) and different linear models with dynamics (87). Simplified nonlinear (*upper left*), Case A (*upper right*), Case B with $N_i = N_j = 5$, resp. $N_i = N_j = 25$ (*lower row left to right*)

As in the previous section we solved the partial differential equations model (68) with a Godunov scheme with N discretization points. The objective function is discretized using the trapezoid rule. For standard nonlinear optimization routines we need at least the gradient of the objective function. We compute an approximation by finite differences. Other approaches (using adjoint calculus) are investigated in [27, 28]. In case of the simplified nonlinear model (72) the gradient can be calculated analytically.

For all nonlinear optimization problems the L-BFGS-B optimizer of Byrd, Lu, Nocedal and Zhu [7,8,53] is used. This method is a gradient projection method with a limited memory BFGS approximation of the Hessian and is capable to consider bound constraints. The default settings are $m = 17$, $factr = 1.d + 5$, $pgtol = 1.d - 8$ and $isbmin = 1$.

The linear model without dynamics (96) is a pure linear programming problem. We solved it using ILOG CPLEX 8.1 [33]. As a default strategy, we set the network simplex method to solve the linear programs. For our test-cases, this method outperforms other solution techniques, such as primal or dual simplex. In case of the linear model with dynamics (91) we have a mixed-integer problem.

Model and Scheme	Parameters	CPU time
Godunov scheme for pde model	N=100	135.65 s
Godunov scheme for pde model	N=50	45.17 s
Simplified nonlinear model		0.05 s
Linear Model with dynamics (B)	$D_q = D_t = 100, N_i \cdot N_j = 25$	0.02 s
Linear Model without dynamics	$D_q = 100$	0.01 s

Fig. 11 CPU times for sample network and different models

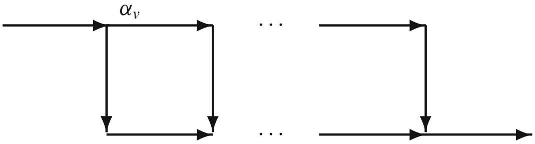


Fig. 12 General layout of a large scale network

Among the currently most successful methods for solving these problems are linear programming based branch-and-bound algorithms, where the underlying linear programming relaxations are possibly strengthened by cutting planes. Fortunately, todays state-of-the-art commercial MIP-solvers (such as CPLEX [33]) are able to handle mixed-integer programs even for our large size problem instances.

For the setting of Test Case 1 we have the following result on the computational times (CPU times), see Fig. 11. The parameters (D_q, D_t) describes the discretization of the nonlinear function. The parameter $N_i \cdot N_j$ describes the total number of discretization points for the function $a_k(\cdot,\cdot)$ at the merging junctions. Therefore, the only models reasonable to test on large scale networks are the simplified nonlinear and the linear models.

Large Scale Network Optimization

The network considered next is shown in Fig. 12.

There, every node in the top row is controlable via a separate control α_v. There are only one source and one sink. The prescribed inflow is again $q_0 = 96\% M_1$ and all streets have the same maximal flux, $M_j = 1.0$ Then, the optimal controls are $\alpha_1 = 0.5$ and $\alpha_v = 1.0, \ \forall v \neq 1$. The results are given in Fig. 13. The number of discretization points for the flux q per road is denoted by D_q and for the time by D_t. The number of discretization points for each function a_k, c.f. (79), in model B is denoted by $N_i N_j$. Note that all nodes in the bottom row are of the merging type. To improve the performance of CPLEX we increased the optimality gap from 0.001% (default setting) to 10%. We present results for other optimality gaps, too.

Model	# Roads	D_q	D_t	$N_i N_j$	Gap	CPU time
Simplified nonlinear model	240	n.a.	n.a.	n.a.	n.a.	6 s
Linear with dynamics (B)		10	10	25	1%	11 m
		10	10	25	10%	3.8 m
		10	10	9	0.1%	2.6 m
		10	10	9	10%	57 s
Linear with dynamics (A)		100	10	n.a.	0.1%	33.08 s
		10	10	n.a.	0.1%	4.78 s
Linear without dynamics		100	n.a.	n.a.	0.1%	<0.01 s
Simplified nonlinear model	1’500	n.a.	n.a.	n.a.	n.a.	57 m
Linear with dynamics (B)		10	10	25	10%	4.7 h
		10	10	9	10%	26 m
		5	5	9	10%	5 m
Linear with dynamics (A)		100	10	n.a.	0.1%	180.01 m
		10	10	n.a.	0.1%	13.69 m
Linear without dynamics		1000	n.a.	n.a.	0.1%	24.98 s
		100	n.a.	n.a.	0.1%	12.75 s
		5	n.a.	n.a.	0.1%	1.8 s
Simplified nonlinear model	15’000	n.a.	n.a.	n.a.	n.a.	>4d
Linear with dynamics (B)		5	5	9	10%	6.2 h
Linear without dynamics		100	n.a.	n.a.	n.a.	22.79 m
		10	n.a.	n.a.	n.a.	7.33 m
Linear without dynamics	150’000	10	n.a.	n.a.	n.a.	16.77 h

Fig. 13 CPU times for large scale networks. n.a. is short for not available since those quantities do not appear in the corresponding models.

2.4 Summary

- A hierachy a traffic network models ranging from PDE models to simple combinatorial models of min cost flow type has been developed.
- A variety of different network topologies has been investigated. Combinatorial and continuous optimization approaches using these models have been implemented and compared.
- The investigation shows the advantages and disadvantages of the different models and optimization procedures. In particular, for very large networks discrete optimization procedures are superior in terms of computation time.
- However, the simplified models developed here do not contain more complicated dynamic situations like backwards going shocks, i.e., traffic jams. To include such situations one has to use the original PDE model or to derive more sophisticated models from the PDE network.
- One could combine the models described here in a coupling strategy for very large networks. The main part of the network can be simulated using simple linear models. More complicated dynamic models may be used in regions where the detailed dynamic behaviour is important.

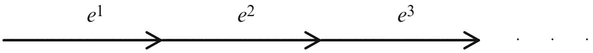


Fig. 14 A serial supply chain

3 Modeling Supply Networks

Continuous models are used to describe many physical problems as for example traffic flow on road networks (cf. Sect. 2), gas transportation through pipelines, telecommunications networks, drinking water systems and many more. In the context of supply chains, we introduce continuous models which are computationally feasible and accurately describe the dynamic behavior of the system. The presented approaches are reasonable in the situation of a large number of parts and can be derived formally from particle simulations (so-called Discrete Event Simulations), see [3]. Therein, a detailed time recursion is derived and used to deduce a model consisting of continuous equations. This model will serve as a basis for extensions to more general supply chains. We present the final continuous network model for supply chains, as stated in [17, 18].

This chapter is organized as follows. Section 3.1 introduces basic terminology and the way how to model supply chains with scalar partial differential equations. We develop coupling conditions and discuss their necessity. In Sect. 3.2 we introduce the continuous optimal control problem that is to be investigated. We present the optimality system for the discrete as well as the continuous supply chain model and discuss the discretization leading to a mixed-integer model. Numerical experiments contain results of the adjoint-based method and the mixed-integer problem.

3.1 Network Models Based on Scalar Conservation Laws

To derive a first network model we consider a particular simple situation: a serial supply chain consisting of M suppliers.
Here, each supplier ships all its goods directly to the next supplier as in Fig. 14. Basically, every supplier is now characterized by the parameters length L, processing time T and maximal processing rate μ. Here, the processing time T is the time which is needed to finish a single production stage. It is assumed that each supplier is available at all times and there will be no unexpected shut-downs. Since suppliers may have different processing rates, it may happen that goods have to wait until the next operations can be performed. Therefore, buffers are installed between the suppliers. In our approach the buffers have unlimited capacity. We treat the problem of limited capacities later. The total amount of goods in the system is denoted by N.
To start introducing continuous supply chain models, we briefly summarize the work of Armbruster, Degond and Ringhofer [3]. Therein, they propose a continuous model (a conservation law for the part density) for interacting processors in a serial

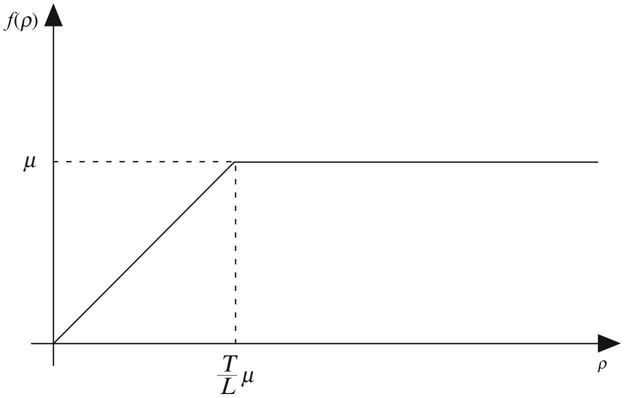


Fig. 15 The fundamental diagram of (105)

production line. This model provides the basis for the derivation of a general supply network model.

$$\partial_t \rho(x,t) + \partial_x \min\left\{\frac{L}{T}\rho(x,t),\ \mu(x)\right\} = 0, \tag{102}$$

$$\rho(x,0) = \rho_0(x), \tag{103}$$

with inflow conditions

$$f(0,t) = f^{in}(t). \tag{104}$$

Due to different processing rates μ in the system, one obtains δ-distributions on the level of the solution for ρ. These δ-distributions (bottlenecks) are natural, modeling the queues in the system, but do not allow for a simple theoretical treatment of the equation.

Note that the flux function is given by

$$f : \mathbf{R}_0^+ \to [0,\mu], \quad f(\rho) = \min\left\{\mu, \frac{L}{T}\rho\right\}. \tag{105}$$

with positive constants μ, T, L. Clearly, f is Lipschitz with constant $L_f = \frac{L}{T}$ (Fig. 15).

For some notational convenience we denote $v := L/T$ in the following.

We now define weak solutions for the general Cauchy problem (102) in the sense of Kruzkov [42]:

Definition 3.1 (Weak solution). A locally bounded and measurable function $\rho(x,t)$ on $\mathbf{R} \times \mathbf{R}_0^+$ is called an admissible weak solution to (102), if for any non-decreasing function $h(\rho)$ and any smooth non-negative function ϕ with compact support in $\mathbf{R} \times \mathbf{R}_0^+$,

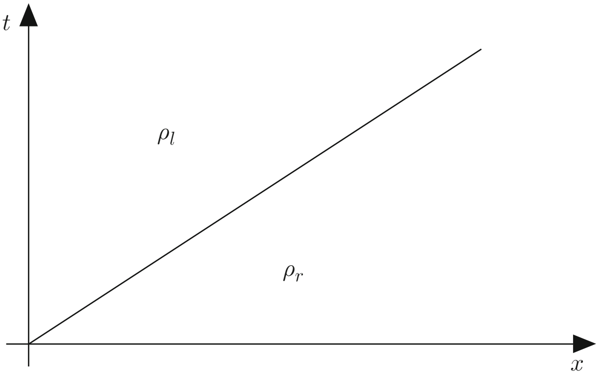


Fig. 16 The solution to the Riemann problem if $\rho_l < \rho_r$

$$\int_0^\infty \int_{-\infty}^\infty (I(\rho)\phi_t + F(\rho)\phi_x)\,dx\,dt + \int_{-\infty}^\infty I(\rho_0)\phi(x,0)\,dx \geq 0 \tag{106}$$

where $I(\rho) = \int^\rho h(\xi)\,d\xi$ and $F(\rho) = \int^\rho h(\xi) f'(\xi)\,d\xi$.

We consider the Riemann problem for (102) with initial data

$$\rho_0(x,0) = \begin{cases} \rho_l, \ x \leq 0 \\ \rho_r, \ x \geq 0 \end{cases} \tag{107}$$

where $\rho_l, \rho_r \in \mathbf{R}_0^+$.

For the flux function (105) with $v = 1$, the solution of Riemann problems for (102) and (105) is either given by (108) or by (109) (Fig. 16):

If $\rho_l < \rho_r$, then the weak admissible solution to (102), (105) and (107) is given by

$$\rho(x,t) = \begin{cases} \rho_l & -\infty < \frac{x}{t} \leq \frac{f(\rho_r)-f(\rho_l)}{\rho_r-\rho_l} \\ \rho_r & \frac{f(\rho_r)-f(\rho_l)}{\rho_r-\rho_l} < \frac{x}{t} < \infty \end{cases} \tag{108}$$

If instead $\rho_r < \rho_l$, we distinguish three cases. If $\rho_l \leq \mu$ or if $\rho_r \geq \mu$ the solution is given by (108). In the remaining case $\rho_r < \mu < \rho_l$ we obtain the solution (Fig. 17)

$$\rho(x,t) = \begin{cases} \rho_l & -\infty < \frac{x}{t} \leq \frac{f(\rho_l)-\mu}{\rho_l-\mu} \\ \mu & \frac{f(\rho_l)-\mu}{\rho_l-\mu} < \frac{x}{t} \leq \frac{\mu-f(\rho_r)}{\mu-\rho_r} \\ \rho_r & \frac{\mu-f(\rho_r)}{\mu-\rho_r} < \frac{x}{t} < \infty \end{cases} \tag{109}$$

Note that the Rankine–Hugoniot condition is either $\frac{f(\rho_l)-\mu}{\rho_l-\mu} = 0$ or $\frac{\mu-f(\rho_r)}{\mu-\rho_r} = 1$.

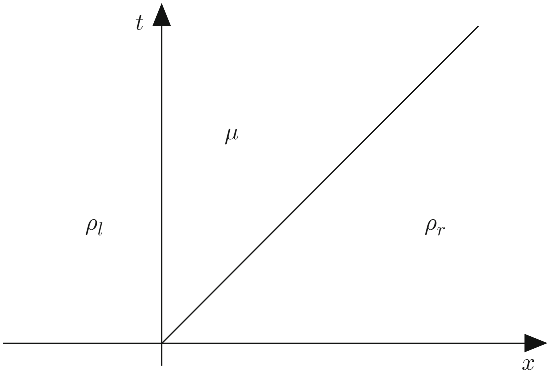


Fig. 17 The solution to the Riemann problem if $\rho_r < \mu < \rho_l$

Summarizing, this scalar PDE-model describes the production process within a serial supply chain. However, different processing rates may lead to δ-distributions in the density which do not allow for a simple theoretical treatment. To avoid these bottlenecks and to extend the model to more complex supply chain geometries we present an alternative modeling approach.

3.1.1 General Network Topologies

We present an extension of the model mentioned above. By adding extra equations for the queues we obtain new equations for the density avoiding the above δ-distributions. This leads to a system of conservation laws coupled to ordinary differential equations. In this framework, situations with real networks having multiple incoming and outgoing arcs for each vertex are easily included. One further advantage is the easy accessibility to existence theory of the network problem, see [29].

We give a definition of a supply network and explain the different scenarios.

Definition 3.2 (Supply network definition). A supply network is a directed graph $(\mathscr{V}, \mathscr{A})$ consisting of a set of arcs $\mathscr{A}$ and a set of vertices $\mathscr{V}$. Each supplier is mapped on to one arc. The length of the supplier corresponding to arc $e \in \mathscr{A}$ is given by the interval $L^e = [a^e, b^e]$.

The maximal processing rate μ^e and the processing velocity $v^e := L^e / T^e$ of each supplier are constant parameters on each arc. According to the assumption that each supplier possesses a buffer, we locate the buffer at the vertex v in front of the supplier. For a fixed vertex v, the set of ingoing arcs is denoted by δ_v^- and the set of outgoing arcs by δ_v^+. In the case of more than one outgoing arc, we introduce distribution rates $A^{v,e}(t)$, $v \in \mathscr{V}_d$, where $\mathscr{V}_d \subset \mathscr{V}$ denotes the set of dispersing junctions, cf. Fig. 18. Those rates describe the distribution of incoming parts among

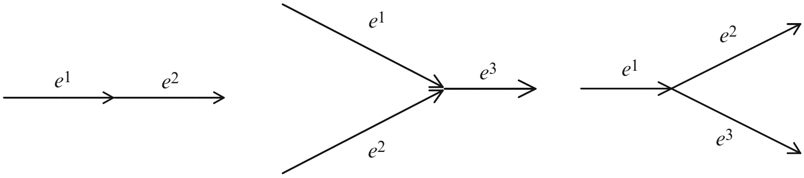


Fig. 18 Possible types of intersections in the network labeled as ordinary (**a**), merging (**b**) and dispersing (**c**) nodes (*from left to right*)

the outgoing suppliers and are later subject to optimization. The functions $A^{v,e}$ are required to satisfy $0 \leq A^{v,e}(t) \leq 1$ and $\sum_{e\in\delta_v^+} A^{v,e}(t) = 1$ for all times $t > 0$. Next, we introduce the continuous supply network model which consists of a coupled system of differential equations. These kind of equations arise whenever the relationship between changing quantities (modeled by functions) and their rates of change (expressed as derivatives) is known. This relationship can be derived from particle simulations, see [3].

The continuous model describes the evolution of the density of goods $\rho^e(x,t)$ at x in time t inside each supplier e and the time evolution of the buffer $q^e(t)$ belonging to supplier e. On each arc e, the density $\rho^e(x,t)$ is transported with velocity v^e, if the flux of goods is less than the maximal processing rate, i.e., ρ^e satisfies the transport equation

$$\partial_t \rho^e + \partial_x f^e(\rho^e) = 0, \tag{110}$$

where the relation between flux and density is given by

$$f^e(\rho^e) = v^e \rho^e.$$

Whenever a supplier is connected to another supplier of possibly different processing rate μ^e, we introduce a buffering zone for the incoming but not yet processed goods. To describe the buffering we introduce the time–dependent function $q^e(t)$ describing the load of the buffer at time t.

$$\partial_t q^e(t) = A^{v,e}(t) \sum_{\bar{e}\in\delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}(x_v^{\bar{e}},t)) - f^e(\rho^e(x_v^e,t)) \tag{111}$$

The dynamics of the buffering is governed by the difference of all incoming and outgoing goods at the connection point x_v: If the queue is empty, the outgoing flux is either a percentage of the sum of all incoming fluxes given by $A^{v,e}(t)$ or the maximal processing rate μ^e. In the first case the buffer remains empty, in the second case the buffer increases. Last, if the buffer is full, the buffer is always reduced with a capacity determined by the distribution rates $A^{v,e}$ and the capacities of the connected arcs.

$$f^e(\rho^e(x_v^e,t)) = \begin{cases} \min\{A^{v,e}(t)\Big(\sum_{\bar{e}\in\delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}(x_v^{\bar{e}},t))\Big), \mu^e\}; & q^e(t) = 0 \\ \mu^e; & q^e(t) > 0 \end{cases} \tag{112}$$

Finally, we obtain the continuous supply chain model for the evolution of $(\rho^e, q^e)_{e\in\mathscr{A}}$ on the network $(\mathscr{V}, \mathscr{A})$ by (110)–(112). The implementation of this model is done by applying standard numerical schemes, see e.g. [17]. So far, the proposed continuous model is based on only a few assumptions and thus extendible to more complicated settings.

Remark 3.3. Note that applying a left-sided Upwind scheme to (102), a straightforward computation shows that (102) and (110)–(112) are equivalent.

3.2 Optimization Problems

A fundamental question arising in the context of managing supply chains is the *optimal* design. Depending on the actual application several aspects are of importance: inventory costs, storage limitations, distribution of goods or external supply and demand.

In this subsection, we concentrate on the optimal routing of goods through the network in order to achievemaximal output atminimumcost and further constraints. The formulation of an optimization problem is based on the continuous model presented before. In summary, (113)–(117) constitute a constrained optimal control problem where the constraints are given by linear transport and ordinary differential equations. The controls are the distribution rates $A^{v,e}$ and the dependent states are the vectors $\boldsymbol{\rho}^e := (\rho^e)_{e\in\mathscr{A}}$ and $\mathbf{A}^v := (A^{v,e})_{e\in\delta_v^+}$.

$$\min_{A^{v,e}(t), v\in\mathscr{V}_d} \sum_{e\in\mathscr{A}} \int_0^T \int_{a^e}^{b^e} f^e(\rho^e(x,t))\,dx\,dt + \int_0^T q^e(t)\,dt \tag{113}$$

$$\text{subject to } e \in \mathscr{A},\ v \in \mathscr{V},\ t \in (0,T),\ x \in [a^e, b^e] \tag{114}$$

$$\partial_t \rho^e(x,t) + \partial_x f^e(\rho^e(x,t)) = 0 \tag{115}$$

$$\partial_t q^e(t) = A^{v,e}(t) \sum_{\bar{e}\in\delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}(x_v^{\bar{e}}, t)) - f^e(\rho^e(x_v^e, t)) \tag{116}$$

$$f^e(\rho^e(x_v^e,t)) = \begin{cases} \min\{A^{v,e}(t)\Big(\sum_{\bar{e}\in\delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}(x_v^{\bar{e}},t))\Big), \mu^e\}; & q^e(t) = 0 \\ \mu^e; & q^e(t) > 0 \end{cases} \tag{117}$$

To solve this PDE–ODE restricted optimization problem, two ways are of interest: adjoint equations or mixed–integer programming. The former has been successfully applied in different areas. Among the variety of literature we only mention some examples, like optimal control of fluid flows [30], optimal semiconductor design [31] or general initial value control of hyperbolic equations [50, 51]. For optimal control in the context of networks we refer to [27] for an adjoint calculus in the context of traffic flow networks. To compute the optimal control, the continuous optimality system is discretized and usually solved by a descent type method, see [35, 46, 49]. This approach can therefore be seen as *optimize-then-discretize*. Alternatively, we can proceed by first discretizing the constraints and cost functional and then optimize the finite-dimensional optimization problem; this strategy is known as *discretize-then-optimize*. For our supply network model the discretization can be chosen, such that the optimization problem is in fact a mixed-integer programming problem, see Sect. 3.2.2. This is mainly due to the fact that the governing dynamics in the supply network are linear in the state (but not in the control) variables. Moreover, further extensions to the mixed-integer problem have been investigated, e.g., finite size buffers, inflow profile optimization or processor shut-down due to maintenance.

In Sect. 3.2.1, we derive the discrete as well as the continuous optimality system and show that the former is a valid discretization of the discretized continuous optimality system, i.e., both approaches *discretize-then-optimize* and *optimize-then-discretize* lead to the same continuous optimal control if the discretization width tends to zero. Furthermore, we investigate the numerical properties of the two approaches by comparing computing times for the solution to the mixed-integer model with a steepest descent method based on the adjoint equations.

3.2.1 Adjoint Equations

We are concerned with the numerical solution to the previous optimal control problem. For further investigation we apply the following modifications and simplifications: First, in order to avoid the discontinuous dependence on the queue-length in (117), we make use of the reformulation presented in [2]. There, (117) has been replaced

$$f^e(\rho^e(x_v^e,t)) = \min\{\mu^e, \frac{q^e(t)}{\varepsilon}\} \quad \text{with } \varepsilon \ll 1. \tag{118}$$

See [2] for further remarks. Since adjoint calculus requires the constraints to be differentiable, we replace the function $y \to \min(y/\varepsilon, \mu^e)$ in (118) by any smooth approximation $\psi^{e,\delta}(y)$ for the computations following. To be more precise, we assume there are families of smooth functions $\{\psi^{e,\delta}\}$ such that

$$\lim_{\delta\to 0} \psi^{e,\delta}(y) = \min(y/\varepsilon, \mu^e)\ \forall y, \forall e. \tag{119}$$

For notational convenience we drop the superindex δ in the following, since the calculations remain true for all $\delta > 0$. Third, we simplify the notation by introducing functions $h^e(\boldsymbol{\rho}, \mathbf{A}^{v,e})$ for each edge e (resp. $\tilde{e}$) and fixed $v \in \mathscr{V}$ such that $e \in \delta_v^+$ (resp. $\tilde{e} \in \delta_v^+$). We define

$$h^e(\boldsymbol{\rho}^e, \mathbf{A}^v) = A^{v,e}(t) \sum_{\bar{e} \in \delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}), \forall e \in \delta_v^+ \backslash \{\tilde{e}\}, \tag{120}$$

$$h^{\tilde{e}}(\boldsymbol{\rho}^e, \mathbf{A}^v) = \left(1 - \sum_{e \neq \tilde{e}} A^{v,e}(t)\right) \sum_{\bar{e} \in \delta_v^-} f^{\bar{e}}(\rho^{\bar{e}}). \tag{121}$$

Note that with this definition the assumption $\sum_{e \in \delta_v^+} A^{v,e} = 1$ can be omitted. For example, for an intersection with $\delta_v^- = \{1\}$ and $\delta_v^+ = \{2, 3\}$ we have the more explicit form

$$h^2(\boldsymbol{\rho}^e, \mathbf{A}^v) = A^{v,2}(t) f^1(\rho^1), \; h^2(\boldsymbol{\rho}^e, \mathbf{A}^v) = (1 - A^{v,2}(t)) f^1(\rho^1). \tag{122}$$

Finally, we summarize the previous modifications and restate the optimal control problem for all $e \in \mathscr{A}$, $v \in \mathscr{V}$, $t \in (0, T)$, $x \in [a^e, b^e]$:

$$\min_{A^{v,e}(t), v \in \mathscr{V}_d} \sum_{e \in \mathscr{A}} \int_0^T \int_{a^e}^{b^e} v^e \rho^e(x,t)\, dx\, dt + \int_0^T q^e(t)\, dt \tag{123}$$

subject to

$$\partial_t \rho^e(x,t) + v^e \partial_x \rho^e(x,t) = 0, \; \rho^e(x,0) = 0, v^e \rho^e(a,t) = \psi^e(q^e) \tag{124}$$

$$\partial_t q^e(t) = h^e(\boldsymbol{\rho}^e, \mathbf{A}^v) - \psi^e(q^e), \; q^e(0) = 0. \tag{125}$$

As technical detail we need to introduce boundary data for those arcs $e \in \mathscr{A}$ which are incoming to the network, i.e., such that $\delta_v^- = \emptyset$. Here, we assume inflow data $\rho_0(t)$ to be given and set $\rho^e(a,t) = \rho_0(t)$ for all $v \in \mathscr{V}$ and $e \in \delta_v^+$ and $\delta_v^- = \emptyset$. From now on we neglect this technical point.

Derivation of Optimality Systems for the Optimal Control Problem

We first derive a discrete optimality system and solve the latter directly by nonlinear optimization methods. This approach is known as *first discretize then optimize*. Formally, one can also derive the continuous optimality system and discretize the latter. This method is referred to as *first optimize then discretize*. We present the corresponding results in the following and comment on the relation between the two approaches *first discretize then optimize* and *first optimize then discretize*.

Optimality System of the Discrete Optimal Control Problem

First, we consider the discrete optimality system. A coarse grid discretization in space of (124) is obtained by just a two-point Upwind discretization and (125) is discretized using the explicit Euler method. Each arc has length L^e and we introduce a step size Δt such that the CFL condition for each arc and the stiffness restriction of the ordinary differential equation are met. The time steps t_j are numbered by $j = 0, \dots, T$. We use the following abbreviations for all e, j:

$$\rho_j^{e,b} := \rho^e(b^e, t_j),\ \rho_j^{e,a} := \rho^e(a^e, t_j),\ q_j^e := q^e(t_j), A_j^{v,e} := A^{v,e}(t_j) \tag{126}$$

$$h_j^e := h^e(\boldsymbol{\rho}^e(x, t_j), \mathbf{A}^v(t_j)). \tag{127}$$

Due to the boundary condition $v^e\rho^e(a,t) = \psi^e(q^e(t))$ we replace the discrete variable $\rho_j^{e,a}$ by $\psi^e(q_j^e)$ and therefore, $\rho_j^{e,a}$ does not appear explicitly in the discrete optimal control problem below. For the initial data we have

$$\rho_0^{e,b} = \rho_0^{e,a} = q_0^e = 0,\ \forall e. \tag{128}$$

Finally, the discretization of problem (123)–(125) reads for $j \geq 1, e \in \mathscr{A}, v \in \mathscr{V}$:

$$\min_{\mathbf{A}^v,\, v \in \mathscr{V}_d} \sum_{e \in \mathscr{A}} \sum_{j=1}^{T-1} \Delta t \left(\frac{L^e}{2} (\psi^e(q_j^e) + v^e \rho_j^{e,b}) + q_j^e \right) \tag{129}$$

subject to

$$\rho_{j+1}^{e,b} = \rho_j^{e,b} + \frac{\Delta t}{L^e} (\psi(q_j^e) - v^e \rho_j^{e,b}) \tag{130}$$

$$q_{j+1}^e = q_j^e + \Delta t (h_j^e - \psi^e(q_j^e)) \tag{131}$$

For deriving the discrete optimality system we state the precise definition of h^e in the case of the following intersections, see Fig. 18. In case A h^2 is independent of $\mathbf{A}^v$ and we have $h^2(\boldsymbol{\rho}^e, \mathbf{A}^v) = v^1\rho^1(b,t)$. Similarly, in case B we obtain $h^3(\boldsymbol{\rho}^e, \mathbf{A}^v) = v^1\rho^1(b,t) + v^2\rho^2(b,t)$. Finally, as already stated, we have in the controlled case C : $h^2(\boldsymbol{\rho}^e, \mathbf{A}^v) = A^{v,2}(t)v^1\rho^1(b,t),\ h^3(\boldsymbol{\rho}^e, \mathbf{A}^v) = (1 - A^{v,2}(t))v^1\rho^1(b,t)$.

Now it is straightforward to derive the discrete optimality system for (129)–(131). We denote the Lagrange multipliers for the discretized partial differential equation by λ_j^e and for the discretized ordinary differential equation by p_j^e. The discrete Lagrangian is given by

$$L(\boldsymbol{\rho}_j^e, \mathbf{q}_j^e, \mathbf{A}_j^v, \boldsymbol{\lambda}_j^e, \boldsymbol{p}_j^e) = \sum_{e \in \mathscr{A}} \sum_{j=1}^{T-1} \Delta t \left(\frac{L^e}{2} (\psi^e(q_j^e) + v^e \rho_j^{e,b}) + q_j^e \right) \tag{132}$$

$$- \sum_{e \in \mathscr{A}} \sum_{j=1}^{T} \Delta t L^e \lambda_j^e \left(\frac{\rho_{j+1}^{e,b} - \rho_j^{e,b}}{\Delta t} - \frac{\psi(q_j^e) - v^e \rho_j^{e,b}}{L^e} \right) \tag{133}$$

$$- \sum_{e \in \mathscr{A}} \sum_{j=1}^{T} \Delta t\ p_j^e \left(\frac{q_{j+1}^e - q_j^e}{\Delta t} - (h_j^e - \psi(q_j^e)) \right), \tag{134}$$

if we set $\lambda^e_T = p^e_T = 0$. Assuming sufficient constraint qualifications the first–order optimality system is given by (130) and (131) and the following additional equations for $j \le T, e \in \mathscr{A}$ and $v \in \mathscr{V}$:

$$\lambda^e_{j-1} = \Delta t \, \frac{v^e}{2} + \lambda^e_j - \frac{\Delta t}{L^e} \left(\phi^e_j - v^e \lambda^e_j \right), \tag{135}$$

$$\phi^e_j := \sum_{\bar{e} \in \delta^+_v \text{ s.t. } e \in \delta^-_v} p^{\bar{e}}_j \frac{\partial}{\partial \rho^e} h^{\bar{e}}_j, \tag{136}$$

$$p^e_{j-1} = \Delta t (1 + \frac{L^e}{2} (\psi^e)'(q^e_j)) + p^e_j - \Delta t \left(p^e_j - \lambda^e_j \right) (\psi^e)'(q^e_j), \tag{137}$$

$$0 = \sum_{e \in \delta^+_v} p^e_j \, \frac{\partial}{\partial A^{v,\bar{e}}} h^e_j \tag{138}$$

The summation in the definition of the function ϕ^e is understood in the following way: For a fixed intersection $v \in \mathscr{V}$ such that $e \in \delta^-_v$ we sum over all $\bar{e} \in \delta^+_v$. Hence, the function ϕ^e depends on the type of intersection and for clarity we state its explicit form for the cases $A - C$ introduced above: In case A we have $\phi^2_j = 0$ and $\phi^1_j = p^1_j v^1$. In case B we obtain $\phi^1_j = p^3_j v^3$ and $\phi^2_j = p^3_j v^3$. Finally, for the interesting case C we find $e = 1$ which implies $\phi^1_j = A^{v,2} p^2_j v^2 + (1 - A^{v,2}) p^3_j v^3$. Furthermore, we obtain with the previous definitions for $\bar{e} \neq \tilde{e}$:

$$\sum_{e \in \delta^+_v} p^e_j \, \partial_{A^{v,\bar{e}}} h^e_j = \left(p^{\bar{e}}_j - p^{\tilde{e}}_j \right) \sum_{e \in \delta^-_v} v^e \rho^e_j . \tag{139}$$

Summarizing, the optimality system to (129)–(131) is given by (130),(131), and (135)–(138). Changing the objective functions only affects the first term on the right hand side in formulas (135) and (137).

Optimality System of the Continuous Optimal Control Problem

We turn our attention to the continuous optimality system for (123)–(125); we will show that the optimality system (130), (131) and (135)–(138) derived above is a valid discretization of the former. For the derivation of the continuous optimality system to (123)–(125) the Lagrangian reads

$$L(\boldsymbol{\rho}^e, \mathbf{A}^v, \mathbf{q}^e, \Lambda^e, \mathbf{P}^e) = \sum_{e \in \mathscr{A}} \int_0^T \int_{a^e}^{b^e} v^e \rho^e \, dx dt + \int_0^T q^e \, dt \tag{140}$$

$$-\sum_{e\in\mathscr{A}}\int_0^T\int_{a^e}^{b^e} \Lambda^e\partial_t\rho^e + \Lambda^e v^e\partial_x\rho^e\,dx\,dt \tag{141}$$

$$-\sum_{e\in\mathscr{A}}\int_0^T P^e\left(\partial_t q^e - h^e(\boldsymbol{\rho}^e,\mathbf{A}^v) + \psi^e(q^e)\right)dt \tag{142}$$

In this setup the adjoint variables are denoted as $\Lambda^e(x,t)$ and $P^e(x,t)$; we use capital letters to highlight their difference from the previously introduced quantities λ_j^e and p_j^e. The relation between these variables is discussed below. We formally obtain the continuous optimality system for all $t, x \in [a^e, b^e], e \in \mathscr{A}$ as

$$\partial_t\rho^e + v^e\partial_x\rho^e = 0,\ \rho^e(x,0) = 0,\ v^e\rho^e(a,t) = \psi^e(q^e), \tag{143}$$

$$\partial_t q^e = h^e(\boldsymbol{\rho}^e,\mathbf{A}^v) - \psi^e(q^e),\ q^e(0) = 0, \tag{144}$$

$$-\partial_t\Lambda^e - v^e\partial_x\Lambda^e = v^e,\ \Lambda^e(x,T) = 0, \tag{145}$$

$$v^e\Lambda^e(b,t) = \sum_{\bar{e}\in\delta_v^+\text{ s.t. }e\in\delta_v^-} P^{\bar{e}}(t)\,\frac{\partial}{\partial\rho^{\bar{e}}}h^{\bar{e}}(\boldsymbol{\rho}^e,\mathbf{A}^v), \tag{146}$$

$$-\partial_t P^e = 1 - (P^e - \Lambda^e(a,t))\,(\psi^e)'(q^e),\ P^e(T) = 0, \tag{147}$$

$$\sum_{e\in\delta_v^+} P^e\,\frac{\partial}{\partial A^{v,\bar{e}}}h^e(\boldsymbol{\rho}^e,\mathbf{A}^v) = 0. \tag{148}$$

Recall that in the limit case $\delta = 0$, we have by definition $\psi^e(y) \to \min\{y/\varepsilon, \mu^e\}$. Therefore, we obtain

$$(\psi^e)'(q^e) \to \frac{1}{\varepsilon}H(\mu^e - q^e/\varepsilon),\ \delta \to 0, \tag{149}$$

where $H(x)$ is the Heaviside function. Hence, in the limit the dynamics of the adjoint queue P^e is governed by a discontinuous right–hand side.

Finally, we show that in fact (135)–(138), (130), (131) is a suitable discretization of (143)–(148). We proceed by reformulating the discrete optimal control problem in the introduced variables defined by

$$\Lambda_j^{e,a} := \lambda_j^e - \frac{L^e}{2},\ \ P_j^e := p_j^e, \tag{150}$$

Then, (135)–(138),(130), (131) read

$$\frac{\rho_{j+1}^{e,b}-\rho_j^{e,b}}{\Delta t} = -\frac{v^e}{L^e}(\rho_j^{e,b}-\rho_j^{e,a}),\ \rho_0^e = 0,\ v^e\rho_j^{e,a} = \psi^e(q_j^e), \tag{151}$$

$$\frac{q_{j+1}^e-q_j^e}{\Delta t} = h_j^e - \psi^e(q_j^e),\ q_0^e = 0 \tag{152}$$

$$\frac{\Lambda_{j-1}^{e,a}-\Lambda_j^{e,a}}{\Delta t} = v^e - \frac{v^e}{L^e}\left(\Lambda_j^{e,b}-\Lambda_j^{e,a}\right),\ \Lambda_T^{e,a} = 0, \tag{153}$$

$$v^e\Lambda_j^{e,b} = \sum_{\bar{e}\in\delta_v^+ \text{ s.t. } e\in\delta_v^-} p_j^{\bar{e}}\frac{\partial}{\partial\rho^e}h_j^{\bar{e}}, \tag{154}$$

$$\frac{P_{j-1}^e-P_j^e}{\Delta t} = 1-\left(P_j^e-\Lambda_j^{e,a}\right)(\psi^e)'(q_j^e), \tag{155}$$

$$0 = \sum_{e\in\delta_v^+} P_j^e\ \frac{\partial}{\partial A^{v,\bar{e}}}h_j^e \tag{156}$$

Obviously, (151)–(156) is an Upwind and explicit Euler discretization of (143)–(148). Note that the discrete Lagrangian multiplier λ_j^e and the discretized Lagrange multiplier $\Lambda_j^{e,a}$ satisfy

$$\Lambda_j^{e,a} = \lambda_j^e + O(L^e) \tag{157}$$

and L^e is in fact the discretization stepwidth in space. Therefore, if we formally let $L^e, \Delta t \to 0$ for $L^e/\Delta t$ fixed, we see that $\lambda^e \to \Lambda^e$ and furthermore, the discrete Lagrangian tends to the continuous Lagrangian.

3.2.2 Mixed-Integer Model

We focus on the issue of reformulating the optimal control problem (123)–(125). As an alternative to the adjoint equations, the optimal control problem can been expressed as an mixed-integer model. This is possible, if one introduces a coarse grid discretization of (123)–(125).

This is possible, since (115) does not allow for complex dynamics like backwards travelling shock waves. Hence, we propose a two-point Upwind discretization of each arc e. Finally, a reformulation of (118) using binary variables yields the mixed–integer model for supply chains. The details are as follows: For each fixed arc $e \in \mathscr{A}$ we introduce two variables for the flux at the boundary and a single variable for the queue for each time t of a timegrid $t = 1, \dots, N_T$

$$\mathrm{f}_t^e := f^e(\rho^e(a^e,t)), \quad \mathrm{g}_t^e = f^e(\rho^e(b^e,t)), \quad \mathrm{q}_t^e := q^e(t) \quad \forall e, t. \tag{158}$$

A two-point Upwind discretization in space and time of (115) is given by

$$\mathrm{g}_{t+1}^e = \mathrm{g}_t^e + \frac{\Delta t}{L^e} v^e \left(\mathrm{f}_t^e - \mathrm{g}_t^e \right), \quad \forall e, t, \tag{159}$$

where we use the same time discretization Δt for all arcs e. Condition (118) is reformulated by introducing binary variables $\zeta_t^e \in \{0, 1\}$ for $e \in \mathscr{A}, t = 1, \ldots, N_T$ and given by

$$\mu^e \zeta_t^e \leq \mathrm{f}_t^e \leq \mu^e, \tag{160}$$

$$\frac{\mathrm{q}_t^e}{\varepsilon} - M \zeta_t^e \leq \mathrm{f}_t^e \leq \frac{\mathrm{q}_t^e}{\varepsilon}, \tag{161}$$

$$\mu^e \zeta_t^e \leq \frac{\mathrm{q}_t^e}{\varepsilon} \leq \mu^e (1 - \zeta_t^e) + M \zeta_t^e, \tag{162}$$

where M is a sufficiently large constant. To be more precise, M may be set to $\frac{T}{\varepsilon} \max_{e \in \mathscr{A}} \mu^e$.

Next, we need to reformulate the coupling conditions (116). We introduce variables h_t^e for the total inflow to arc e at $x = a^e$ and require the following equalities for each vertex $v \in \mathscr{V}$:

$$\sum_{e \in \delta_v^+} \mathrm{h}_t^e = \sum_{e \in \delta_v^-} \mathrm{g}_t^e \ \forall v, t \tag{163}$$

$$\mathrm{q}_{t+1}^e = \mathrm{q}_t^e + \Delta t \left(\mathrm{h}_t^e - \mathrm{f}_t^e \right), \forall e, t. \tag{164}$$

Note that we use an explicit time discretization of the ordinary differential equation. This is mainly due to the fact that an implicit discretization would introduce an additional coupling between different arcs on the network. On the contrary, the explicit discretization introduces only a local coupling between the arcs connected at a fixed vertex $v \in \mathscr{V}$. From condition (118), we observe that the ordinary differential equation is stiff, whenever $0 < q^e(t) \leq \varepsilon \mu^e$. A suitable discretization of the ordinary and partial differential equation should satisfy a stiffness condition and the CFL condition as well. Hence, we choose Δt as

$$\Delta t = \min\{\varepsilon; L^e / v^e : e \in \mathscr{A}\}. \tag{165}$$

in the case of the coarsest discretization. A natural choice for ε is $\varepsilon = \Delta x / v^e$, since, as already mentioned above, $q^e(t)/\varepsilon$ represents a relaxed flux. More detailed, we know that a flux can be rewritten as the product of the part density and the processing velocity, $f^e = v^e \rho^e$. Due to the fact that the density at the first discretization point $x = a^e$ of the processor is the same as $q^e / \Delta x$, the parameter ε is determined.

This leads to the condition

$$\Delta t = \min\{L^e / v^e : e \in \mathscr{A}\}. \tag{166}$$

Moreover, we have the following box constraints for all $e \in \mathscr{A}$, $\forall t = 1, \ldots, N_T$

$$0 \leq \mathrm{f}_t^e \leq \mu^e, \quad 0 \leq \mathrm{g}_t^e \leq \mu^e, \quad 0 \leq \mathrm{q}_t^e. \tag{167}$$

Finally, we assign initial data to $\mathrm{f}_1^e, \mathrm{g}_1^e$ and q_1^e. For a discretization of the cost functional we use a trapezoid rule in space and a rectangle rule in time and obtain:

$$\sum_{e,t} \Delta t \, \frac{L^e}{2} \Big(\mathscr{F}(\mathrm{f}_t^e / v^e, \mathrm{q}_t^e) + \mathscr{F}(\mathrm{g}_t^e / v^e, \mathrm{q}_t^e) \Big). \tag{168}$$

Summarizing, the mixed–integer model derived by discretization of the network formulation of the supply chain dynamics is given by

$$\begin{gathered} \min_{A_t^{v,e}, v \in \mathscr{V}_d} \quad (168) \\ \text{subject to } (159) - (164), (167). \end{gathered} \tag{169}$$

A few remarks are added. First, it is a matter of simple calculations to recover the entries of the distribution vectors $A_t^{v,e}$ from the values of h_t^e. Second, other objective functionals can be envisioned and in the case of a nonlinearity in (113), we might have to introduce additional binary variables to obtain a mixed–integer approximation. This is standard and can be found for example in [15]. Third, if we use an implicit discretization of the ordinary differential equation (116),

$$\mathrm{q}_{t+1}^e = \mathrm{q}_t^e + \Delta t \left(\mathrm{h}_{t+1}^e - \mathrm{f}_{t+1}^e \right), \tag{170}$$

we end up with no restriction on the time step. From the continuous point of view such an approach is not favorable due to the additional introduced strong coupling between all arcs in the network. We conclude the modeling with the following remark: In the particular case of a supply chain consisting of a sequence of processors and vertices of degree at most two, there is no possibility to distribute parts. In this case, the mixed–integer model coincides with the two-point Upwind discretization of the partial differential equation and both yield the same dynamics. The mixed–integer problem reduces to a feasibility problem in this case.

Model Extensions

In real-world examples the introduced model is too simple to give realistic results. Hence, we propose a few extensions to the mixed–integer model (169) on an arbitrary network.

1. **Finite size buffers**
   Usually, in the design of production lines, it is mandatory to limit the size of the buffering queues q_t^e. This condition can be implemented in the mixed–integer context by adding box constraints as follows:

$$\mathrm{q}_t^e \leq \text{ const }, \ \forall e, t. \tag{171}$$

Similarly, we can add the constraints $q^e(t) \leq$ const to the continuous problem (123)–(125) and obtain a state constrained optimal control problem.

2. **Optimal inflow profile**
Under the assumption of finite sizes in the buffering queues, the question arises to find the maximum possible inflow to the network, such that the buffering capacities of the queues are not exceeded.
This can be modeled by replacing the cost functional (168) or (123), respectively, by the following objective function

$$\max \sum_{e \in \mathscr{A}', t} \mathrm{f}_t^e, \tag{172}$$

where $\mathscr{A}' \subset \mathscr{A}$ is the set of all inflow arcs of the network.

3. **Processor Shut-Down due to maintenance**
Maintenance of processors can also be included in the mixed–integer model: Assume that processor $\tilde{e}$ has to be switched off for maintenance for N consecutive time intervals. Further assume that this period can be chosen freely during the whole simulation time $t = 1, \dots, N_T$. Then, we supplement the mixed–integer model with the following condition

$$\mathrm{h}_{t+l}^{\tilde{e}} \leq \max\{\mu^e : e \in \mathscr{A}\}|\mathscr{A}| \cdot (1 - \phi_t^{\tilde{e}}), \forall t, \forall l = 0, \dots, N-1, \tag{173}$$

$$\sum_{t=1}^{N_T} \phi_t^{\tilde{e}} = 1, \tag{174}$$

where for each processor $e \in \mathscr{A}$ and every time t we introduce a binary variable $\phi_t^e \in \{0, 1\}$ that indicates whether process e is shut-down at time t. If $\phi_{t_0}^{\tilde{e}} = 1$, then the maintenance interval starts at time t_0, and in the time interval $t_0, t_0 + N$, the processor $\tilde{e}$ is not available.

3.3 Numerical Results

As we have seen, there are two numerical approaches for solving the optimal control problem. On the one hand, we use a steepest descent method for a suitable cost functional.We consecutively solve the equations of state (130) and (131) for a given initial control $\mathbf{A}_0^v \equiv 0$ and the adjoint equations (135)–(137) which in turn are needed to evaluate the gradient (139). Using the Armijo–Goldstein rule for the choice of the stepsizes we update the controls $\mathbf{A}_0^v$ and iterate the described procedure.

On the other hand, we reformulate the original problem as a mixed–integer programming (MIP) model that is usually solved by a Branch-and-Bound algorithm.

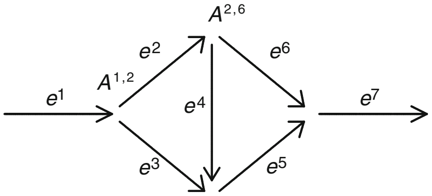


Fig. 19 Sample network with controls $A_j^{1,2}$ and $A_j^{2,6}$

The essential difference to (131) is to rewrite the nonlinearity in (118) by introducing binary variables ζ_t^e. This leads finally to a mixed–integer problem and not just a linear programming (LP) model. For solving the mixed-integer problem the standard optimization software solver ILOG CPLEX [33] is used.

3.3.1 Gradient Computations

At first we compare the gradient of the cost functional obtained by finite differences to the gradient obtained by the adjoint equations for a suitable network. We use the network depicted in Fig. 19 for this test since it has only two variable controls $A_j^{1,2}$ and $A_j^{2,6}$ at time j (recall that $A_j^{1,3} = 1 - A_j^{1,2}$ and $A_j^{2,5} = 1 - A_j^{2,6}$ due to the coupling conditions). We discretize the control–space $[0,1] \times [0,1]$ using 16 points in both the $A_j^{1,2}$ and $A_j^{2,6}$ component.

We set the time–horizon $T = 4$, use $NT = 200$ time–intervals and set $\varepsilon = 1$. We use a one–sided forward difference scheme to compare the gradient at time–interval j, $j = 1, \ldots, NT$:

$$\partial_{A_j^{v,e}} J(\mathbf{A}^v) := \frac{J(\mathbf{A}^v + \delta) - J(\mathbf{A}^v)}{\delta} \tag{175}$$

where $\delta = 0.001$. For the cost–functional we chose the nonlinear function

$$J(\mathbf{A}^v) := \left(\sum_{e \in \mathscr{A}} \sum_j \Delta t \left(\frac{L^e}{2} (\psi(q_j^e) + v^e \rho_j^{e,b}) + q_j^e \right) \right)^2 . \tag{176}$$

Further, we set $L^3 = L^6 = 10$ and $L^e = 1$ for $e \in \mathscr{A} \backslash \{e^3, e^6\}$. The processing rates are $\mu^e = 1,\ \forall e$. This implies that the lowest functional value should be attained for $A_j^{1,2} = 1$ and $A_j^{2,6} = 0$ for all j as confirmed in Fig. 20. The inflow–profile on e^1 is chosen as

$$f^{in}(t) = \begin{cases} 0.852 & t \le 2 \\ 0 & t > 2 \end{cases}$$

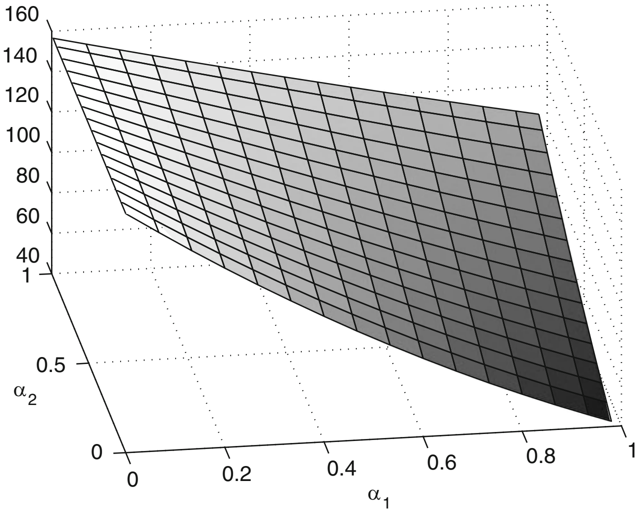


Fig. 20 Plot of the cost functional (176) corresponding to Fig. 19

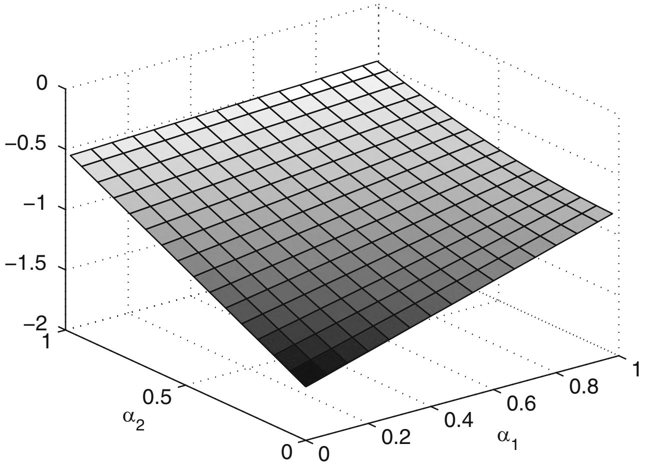


Fig. 21 First component of the gradient computed by the adjoint scheme at $j = 50$

With this inflow profile the gradient w.r.t. $A_{50}^{1,2}$ is nonzero and is depicted in Fig. 21. The controls $A_j^{1,2}$ with $j > 2$ can be chosen arbitrarily since the inflow is zero and hence the gradient w.r.t. these controls needs to vanish. However, since in this particular setup queue 2 is nonempty at time $j = 200$, the gradient w.r.t. $A_{200}^{2,6}$ does not vanish, cf. Fig. 22. The relative error in the second component is of order $1e^{-8}$ and can be found in Fig. 23.

Finally, we mention that we have conducted extensive test with different objective–functionals and varying the parameters $\varepsilon \in [0.01, 1]$, $\delta \in \{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$

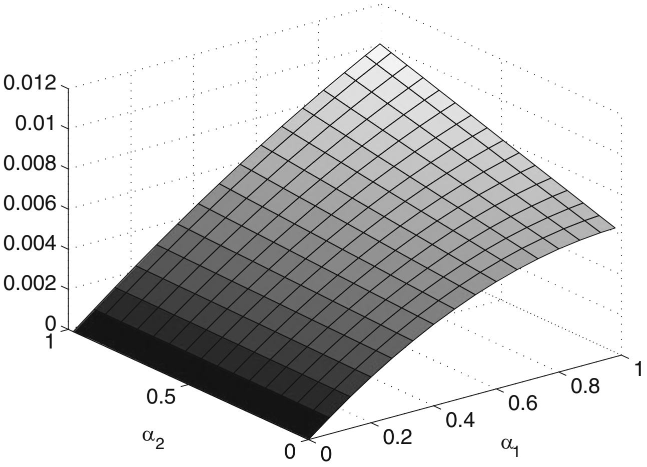


Fig. 22 Second component of the gradient computed by the adjoint scheme at $j = 200$

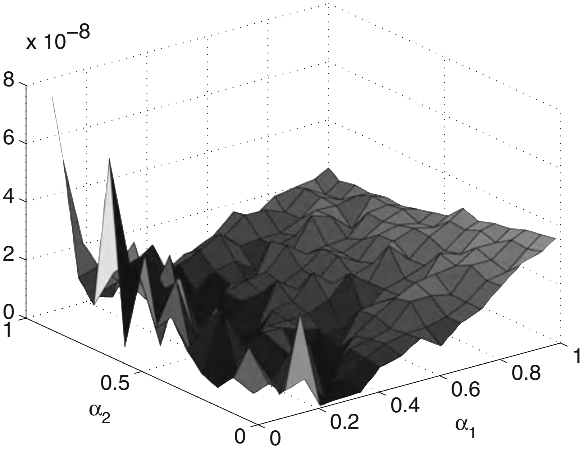


Fig. 23 Relative error in partial w.r.t. $A^{2,6}_{200}$ at $j = 200$

and $NT \in [20, 400]$; we never encountered a relative error in the gradient larger than $1e^{-6}$.

Quality of Solutions of Discrete Adjoint Calculus Compared with the Mixed–Integer Model

As a next step, we compare results computed by the adjoint approach and the mixed–integer programming (MIP) model. We show that this kind of discretization induces

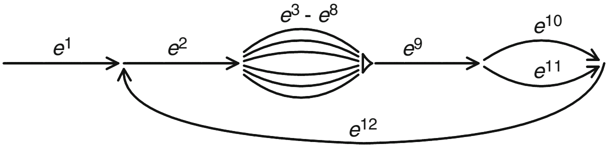


Fig. 24 Sample network

Table 1 Processing rates μ^e

e	1	2	3	4	5	6	7	8	9	10	11	12
μ^e	100	8	10	0.5	0.5	10	0.5	2	20	3.5	2.5	8

same results for the cost functional as the discrete adjoint approach by focussing on the optimal control problem of routing of goods through a network.

In the following, we consider the network in Fig. 24. It consists of 11 processors and queues and we have the six free controls $A^{2,3}(t)$, $A^{2,4}(t)$, $A^{2,5}(t)$, $A^{2,6}(t)$, $A^{2,7}(t)$ and $A^{9,10}(t)$. The artificial arc 1 is used to prescribe an inflow profile which is given by

$$f(t) = \begin{cases} 0.5 & 0 \leq t \leq \frac{T}{4} \\ 0.1 & \frac{T}{4} < t \leq \frac{T}{2} \\ 0.3 & \frac{T}{2} < t \leq \frac{3}{4}T \\ 0 & \frac{3}{4}T < t \leq T \end{cases} \tag{177}$$

Our goal is to maximize the output of processor 12 on a given time–interval $[0, T]$. We use an equidistant time–discretization with NT time–intervals and choose the following reduced cost functional

$$J(\mathbf{A}^v) = \sum_{j=2}^{NT+1} -\frac{v^{12}\rho_j^{12,b}}{j}. \tag{178}$$

In the example below we define $T = 200$, $NT = 400$, $\varepsilon = 1$ and set $L^e = v^e = 1$ for all edges except for $e = 2$; here we use $L^2 = 1$ and $v^2 = 2$. The corresponding processing rates are given in Table 1.

In the following we present results for the optimal routing problem by pointing out similarities and differences between the adjoint and discrete approach. The computation of the adjoint approach takes $37.781s$ using 32 iterations and for the MIP $16.60s$ using 16 iterations. In the adjoint approach, we terminate the iteration if the relative error of two consecutive iterates is less than $tol := 1e^{-6}$—consistent with the default accuracy in ILOG CPLEX [33]. Both approaches yield an optimal functional value of $J^*(\mathbf{A}^v) = -6.49$.

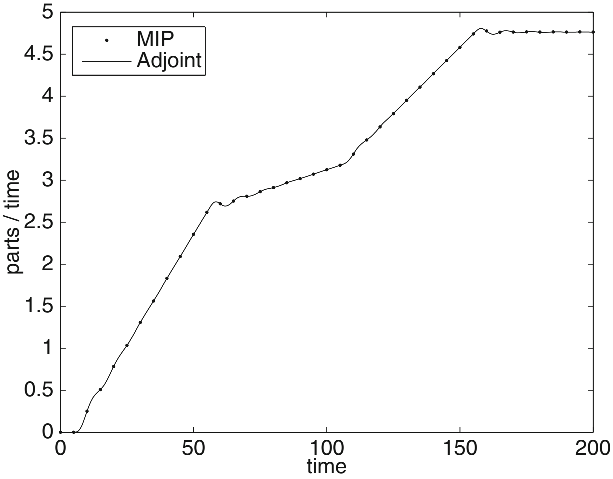


Fig. 25 Optimal output for processor 12

In Fig. 25 we plot the optimal outflow profile computed by the two approaches. We observe that for this particular example the curves coincide. However, the computed optimal controls and time evolution of the queues differ considerably. In Figs. 26 and 27 we plot the optimal control feeding parts into queue 10. Furthermore, we present the evolution over time for the queue 10 in Figs. 28 and 29 for the MIP and the adjoint approach, respectively, and the maximum queue–length in Figs. 28 and 29.

Although the optimal functional values coincide we see that we do not have a unique minimizer to our optimal control problem.

Computational Times

The numerical results conclude with a comparison of computational times of the adjoint–based approach and the mixed–integer formulation. Our computations are performed on the network given in Fig. 24 with default parameters $v^e = L^e = 1, e = 2, \dots, 12$, $\varepsilon = 1$ and time horizon $T = 200$. To obtain a stable discretization both models have to satisfy the following restriction:

$$\Delta t \leq \min\{\varepsilon; \frac{L^e}{v^e} : e \in \mathscr{A}\}. \tag{179}$$

Resulting from (179) the parameter NT describes the number of time intervals. We increase NT by varying the ratio of L^1/v^1. The MIP is solved using the interior point method implemented in ILOG CPLEX [33].

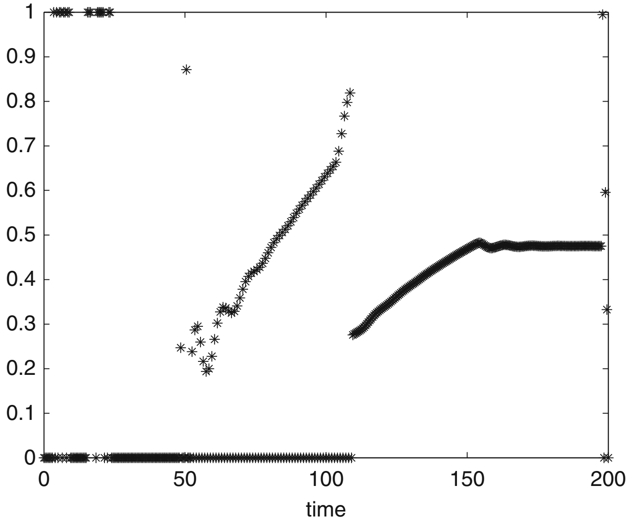


Fig. 26 Plot of the distribution rate $A_j^{9,10}$ computed by the MIP

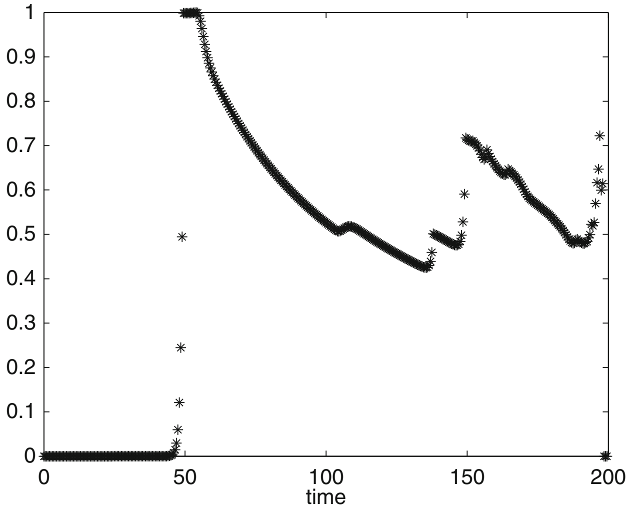


Fig. 27 Plot of the distribution rate $A_j^{9,10}$ computed by the adjoint approach

As Table 2 indicates the MIP is superior if one wants to use up to approximately 600 time-steps (corresponding to $\Delta t \in [0.\overline{3}, 1]$). As NT increases the adjoint approach becomes more attractive. For values of $\Delta t < 0.\overline{3}$ it computes an optimal solution faster than the MIP. At present the MIP fails to compute a solution for $\Delta t \leq 0.05$ since the system becomes too large and the preprocessing procedure produces infeasible solutions.

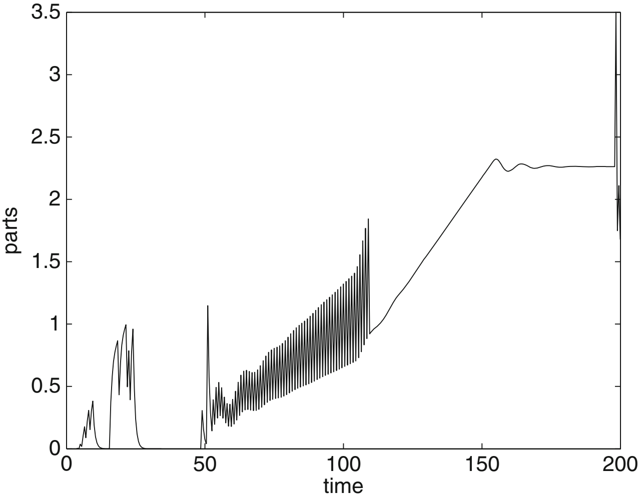


Fig. 28 Optimal queue length q_j^{10} computed by the MIP

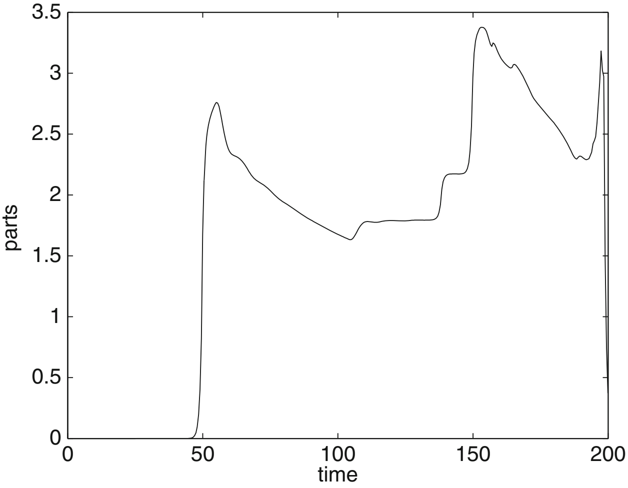


Fig. 29 Optimal queue length q_j^{10} computed by the adjoint approach

Table 2 CPU times in seconds for sample network Fig. 24

NT	Adjoint	MIP
200	7.31	5.52
400	26.10	17.06
800	45.10	68.09
2,000	124.58	592.61

3.4 Summary

- Based on the supply chain model of Armbruster et al. [3] we have presented a reformulation consisting of queues and processors modeled as a coupled system of partial and ordinary equations. The new model allows for an existence theory directly for the density ρ since δ-distributions are avoided, see [29].
- Different optimization techniques to supply networks have been applied. We have derived a continuous and a discrete optimality system and shown that the latter can be interpreted as an upwind and explicit Euler discretization of the former. On the other hand, we have deduced a simplified model by using a straightforward two point discretization of the equations on each arc. For the optimization of the supply chain model the resulting equations are interpreted as a mixed–integer problem.
- The results of the adjoint equations are compared to the ones obtained by a mixed-integer formulation. For the testcases under consideration, the optimal solutions (i.e., the optimal values of the objective function) of the adjoint approach and the MIP formulation introduced coincide. However, the optimal controls differ qualitatively since they are not unique. The usage of the adjoint method as presented here is limited. With the MIP more complex and praxis-relevant questions can be modeled and solved quite easily; processor shutdown due to maintenance or storage limitations are just two examples.

Acknowledgements We wish to thank all our collaborators and co-authors, in particular Michael Herty, Armin Fügenschuh and Alexander Martin. Parts of this work have been taken from the articles [15, 24, 25] as well as [16, 17, 36].

References

1. D. Armbruster, D. Marthaler, C. Ringhofer, Kinetic and fluid model hierarchies for supply chains. SIAM J. Multiscale Model. **2**, 43–61 (2004)
2. D. Armbruster, C. de Beer, M. Freitag, T. Jagalski, C. Ringhofer, Autonomous control of production networks using a pheromone approach. Phys. A **363**, 104–114 (2006)
3. D. Armbruster, P. Degond, C. Ringhofer, A model for the dynamics of large queuing networks and supply chains.. SIAM J. Appl. Math. **66**, 896–920 (2006)
4. D. Armbruster, P. Degond, C. Ringhofer, Kinetic and fluid models for supply chains supporting policy attributes. Bull. Inst. Math. Acad. Sinica **2**, 433–460 (2007)
5. A. Aw, M. Rascle, Resurrection of second order models of traffic flow. SIAM J. Appl. Math. **60**, 916–938 (2000)
6. A. Aw, A. Klar, T. Materne, M. Rascle, Derivation of continuum flow traffic models from microscopic follow the leader models. SIAM J. Appl. Math. **63**, 259–278 (2002)
7. R. Byrd, J. Nocedal, R. Schnabel, Representations of quasi-newton matrices and their use in limited memory methods. Math. Program. **63**, 129–156 (1994)
8. R. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**, 1190–1208 (1995)
9. C. Cercignani, *The Boltzmann Equation and its Applications* (Springer, New York, 1988)

10. G. Coclite, M. Garavello, B. Piccoli, Traffic flow on a road network. SIAM J. Math. Anal. **36**, 1862–1886 (2005)
11. R. Colombo, Hyperbolic phase transitions in traffic flow. SIAM J. Appl. Math. **63**, 708–721 (2002)
12. C. Daganzo, A continuum theory of traffic dynamics for freeways with special lanes. Transp. Res. B **31**, 83–102 (1997)
13. C.F. Daganzo, A theory of supply chains. in *Lecture Notes in Economics and Mathematical Systems*, vol. 526 (Springer, Berlin, 2003), p. viii, 123
14. C. D'Apice, R. Manzo, A fluid dynamic model for supply chains. Netw. Heterogenous Media **1**, 379–389 (2006)
15. A. Fügenschuh, M. Herty, A. Klar, A. Martin, Combinatorial and continuous models for the optimization of traffic flows on networks. SIOPT **16**, 1155–1176 (2006)
16. A. Fügenschuh, S. Göttlich, M. Herty, A. Klar, A. Martin, A discrete optimization approach to large scale supply networks based on partial differential equations. SIAM J. Sci. Comput. **30**, 1490–1507 (2008)
17. S. Göttlich, M. Herty, A. Klar, Network models for supply chains. Comm. Math. Sci. **3**, 545–559 (2005)
18. S. Göttlich, M. Herty, A. Klar, Modelling and optimization of supply chains on complex networks. Comm. Math. Sci. **4**, 315–330 (2006)
19. J. Greenberg, Extension and amplification of the Aw-Rascle model. SIAM J. Appl. Math. **62**, 729–745 (2001)
20. J. Greenberg, A. Klar, M. Rascle, Congestion on multilane highways. SIAM J. Appl. Math. **63**, 818–833 (2003)
21. M. Günther, A. Klar, T. Materne, R. Wegener, Multivalued fundamental diagrams and stop and go waves for continuum traffic flow equations. SIAM J. Appl. Math. **64**, 468–483 (2003)
22. D. Helbing, Improved fluid dynamic model for vehicular traffic. Phys. Rev. E **51**, 3164 (1995)
23. D. Helbing, A. Greiner, Modeling and simulation of multi-lane traffic flow. Phys. Rev. E **55**, 5498–5507 (1975)
24. M. Herty, A. Klar, Modelling, simulation and optimization of traffic flow networks. SIAM J. Sci. Comput. **25**, 1066–1087 (2003)
25. M. Herty, A. Klar, Simplified dynamics and optimization of large scale traffic networks. Math. Models Meth. Appl. Sci. **14**, 1–23 (2004)
26. M. Herty, M. Rascle, Coupling conditions for a class of second order models for traffic flow. SIAM J. Math. Anal. **38**, 595–616 (2006)
27. M. Herty, M. Gugat, A. Klar, G. Leugering, Optimal control for traffic flow networks. J. Optim. Theor. Appl. **126**, 589–615 (2005)
28. M. Herty, K. Kirchner, A. Klar, Instantaneous control for traffic flow. Math. Methods Appl. Sci. **30**, 153–169 (2006)
29. M. Herty, A. Klar, B. Piccoli, Existence of solutions for supply chain models based on partial differential equations.. SIAM J. Math. Anal. **39**, 160–173 (2007)
30. M. Hinze, K. Kunisch, Second order methods for optimal control of time-dependent fluid flow. SIAM J. Contr. Optim. **40**, 925–946 (2001)
31. M. Hinze, R. Pinnau, An optimal control approach to semiconductor design. M3AS **12**, 89–107 (2002)
32. H. Holden, N. Risebro, A mathematical model of traffic flow on a network of unidirectional road. SIAM J. Math. Anal. **4**, 999–1017 (1995)
33. IBM ILOG CPLEX. IBM Deutschland Gmbh, 71137 Ehningen. Information available at http://www-01.ibm.com/software/integration/optimization
34. S. Jin, Z. Xin, The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Comm. Pure Appl. Math. **48**, 235–276 (1995)
35. C. Kelley, *Iterative Methods for Linear and Nonlinear Equations* (SIAM, Philadelphia, 1995)
36. C. Kirchner, M. Herty, S. Göttlich, A. Klar, Optimal control for continuous supply network models. Netw. Heterogenous Media **1**, 675–688 (2006)

37. A. Klar, R. Wegener, A hierachy of models for multilane vehicular traffic I: Modeling. SIAM J. Appl. Math. **59**, 983–1001 (1998)
38. A. Klar, R. Wegener, Kinetic derivation of macroscopic anticipation models for vehicular traffic. SIAM J. Appl. Math. **60**, 1749–1766 (2000)
39. A. Klar, R. Kuehne, R. Wegener, Mathematical models for vehicular traffic. Surv. Math. Ind. **6**, 215 (1996)
40. E. Köhler, M. Skutella, Flows over time with load-dependent transit times, in *SODA'02 Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 174–183 (2002)
41. E. Köhler, M. Skutella, R.H. Möhring, Traffic networks and flows over time, in Book Algorithmics of Large and Complex Networks, Springer-Verlag Berlin, Heidelberg, 166–196 (2009)
42. S.N. Kruzkov, First order quasi linear equations in several independent variables. Math. USSR Sbornik **10**, 217 (1970)
43. R. Kühne, Macroscopic freeway model for dense traffic, in *9th International Symposium on Transportation and Traffic Theory*, ed. by N. Vollmuller (VNU Science Press, Utrecht, 1984), pp. 21–42
44. J. Lebacque, M. Khoshyaran, First order macroscopic traffic flow models for networks in the context of dynamic assignment, *Transportation Planning*, ed. by M. Patriksson, K.A.P.M. Labbe, 119–140 (2002)
45. M. Lighthill, J. Whitham, On kinematic waves. Proc. R. Soc. Edinb. **A229**, 281–345 (1983)
46. S.G. Nash, A. Sofer, *Linear and Nonlinear Programming* (The McGraw-Hill Companies, New York/St. Louis/San Francisco, 1996)
47. P. Nelson, A kinetic model of vehicular traffic and its associated bimodal equilibrium solutions. Transp. Theor. Stat. Phys. **24**, 383–408 (1995)
48. H. Payne, FREFLO: A macroscopic simulation model of freeway traffic. Transp. Res. Rec. **722**, 68–75 (1979)
49. P. Spellucci, *Numerische Verfahren der nichtlinearen Optimierung* (Birkhäuser, Basel, 1993)
50. S. Ulbrich, A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms. SIAM J. Contr. Optim. **41**, 740 (2002)
51. S. Ulbrich, Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws. Syst. Contr. Lett. **3**, 309 (2003)
52. G. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974)
53. C. Zhu, R. Byrd, J. Lu, J. Nocedal, L-bfgs-b: Fortran subroutines for large scale bound constrained optimization. Technical Report, NAM-11, EECS Department, Northwestern University, 1994

Control and Stabilization of Waves on 1-d Networks

Enrique Zuazua

Abstract We present some recent results on control and stabilization of waves on 1-d networks.

The fine time-evolution of solutions of wave equations on networks and, consequently, their control theoretical properties, depend in a subtle manner on the topology of the network under consideration and also on the number theoretical properties of the lengths of the strings entering in it. Therefore, the overall picture is quite complex.

In this paper we summarize some of the existing results on the problem of controllability that, by classical duality arguments in control theory, can be reduced to that of observability of the adjoint uncontrolled system. The problem of observability refers to that of recovering the total energy of solutions by means of measurements made on some internal or external nodes of the network. They lead, by duality, to controllability results guaranteeing that L^2-controls located on those nodes may drive sufficiently smooth solutions to equilibrium at a final time. Most of our results in this context, obtained in collaboration with R. Dáger, refer to the problem of controlling the network from one single external node. It is, to some extent, the most complex situation since, obviously, increasing the number of controllers enhances the controllability properties of the system. Our methods of proof combine sidewise energy estimates (that in the particular case under consideration can be derived by simply applying the classical d'Alembert's formula), Fourier series representations, non-harmonic Fourier analysis, and number theoretical tools.

E. Zuazua (✉)
BCAM – Basque Center for Applied Mathematics, Alameda Mazarredo, 14, E-48009 Bilbao, Basque Country, Spain

Ikerbasque, Basque Foundation for Science, Alameda Urquijo 36-5, Plaza Bizkaia, 48011, Bilbao, Basque Country, Spain
e-mail: zuazua@bcamath.org

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics 2062, DOI 10.1007/978-3-642-32160-3_9,

These control results belong to the class of the so-called open-loop control systems.

We then discuss the problem of closed-loop control or stabilization by feedback. We present a recent result, obtained in collaboration with J. Valein, showing that the observability results previously derived, regardless of the method of proof employed, can also be recast a posteriori in the context of stabilization, so to derive explicit decay rates (as $t \to \infty$) for the energy of smooth solutions. The decay rate depends in a very sensitive manner on the topology of the network and the number theoretical properties of the lengths of the strings entering in it.

In the end of the article we also present some challenging open problems.

1 Introduction and Main Results

This article is devoted to the presentation of some results on wave propagation phenomena in multi-link or multi-body structures constituted by a planar network of linear vibrating strings and undergoing vertical displacements.

There exists a rich mathematical literature on multi-body mechanical systems constituted by coupled flexible or elastic elements as strings, beams, membranes or plates since their practical relevance is huge. In most cases they are systems of Partial Differential Equations (PDE) on networks or graphs. The interested reader is referred to the books [10, 37] for an introduction to the theory of Partial Differential Equations on networks which is an active subject since the early 1980s [32, 38]. In [23] wide information may be found on modeling and control issues. We also refer to [24] for a systematic analysis of the application of domain decomposition techniques for networks. But elasticity and flexible structures are not the only motivation for dealing with wave equations on graphs or networks. This topic is also closely related to many other applications such as water supply and irrigation, in which case the relevant models are often the Saint Venant equations, a first order hyperbolic system (see [15, 16]).

The model we address in these notes is, to some extent, the simplest one in this context but, as we shall see, it is complex enough to present a rich variety of new qualitative properties. Indeed, the interaction between the different components of the multi-link structure generates new dynamics that can not be predicted by simply analyzing the dynamics of each component separately. Doing that requires taking into account various ingredients as the topology of the graph of the network, the lengths of the strings entering in it, the boundary conditions on the external nodes, the joint conditions, etc.

The goal of these notes is to present some by now well-known results that illustrate this complex dynamics, indicating the needed analytical tools and pointing towards some open problems and directions of research. We mainly focus on the control theoretical problems of observation, control and stabilization. These issues are intrinsically interesting but, in fact, constitute a way of analyzing and describing the fine propagation properties of waves in these media. We mainly focus on the case where controllers, observers and dampers are located in one single external node of the network. This is somehow the most degenerate situation, in which, control theoretical properties are harder to be fulfilled. The methods and ideas we develop for addressing this case can then easily be adapted to deal with other problems in which, in particular, several controllers are located in different nodes (internal or external ones) of the network.

We follow closely our previous book on the subject [7], devoted mainly to the problem of controllability and our more recent article on the stabilization [41], incorporating some new results and material.

As we mentioned before, we consider the scalar 1-d wave equation on a finite planar network of strings. Deformations are assumed to be perpendicular to the reference plane. The main advantage of considering this model, as compared to other more complex equations or systems along the graph, is that, while waves propagate within one of the strings, one can have a complete and explicit representation through the classical d'Alembert formula. This allows to easily follow the propagation of the energy along each individual string. But, the overall dynamics turns out to be rather complex, due to the interaction of the various strings at the joints. Indeed, when waves reach a node or junction point, part of the energy bounces back and part of it is transmitted to the other strings with the same common node. This occurs whenever some wave reaches a node or the external boundary (in which, in the case of conservative boundary conditions, the whole energy bounces back).

Then, the overall picture necessarily depends on a number of ingredients:

- The topology of the graph.
- The lengths of the various strings constituting the graph.
- The boundary conditions imposed at the extremes of the graph.
- The joint conditions.

In these notes we consider the simplest model involving the so-called Kirchhoff type joint conditions. Other joint conditions can also be considered so that the model under consideration is well-posed. That is for instance the case when imposing dynamical point-mass equations on the joints. But, in that case, the dynamics is even more complex since the phenomena we address here have to be complemented with the possibility that waves have a different degree of regularity on the various strings involved in the network, a fact that was observed in [17] in the simplest case of two vibrating strings connected by a point mass and later extended to the multi-dimensional case in [21, 27].

Thus, the results we present here are not exhaustive, by any means. However, most of the ideas and methods we develop here can be adapted and extended to more sophisticated and realistic wave models in networks.

As we mentioned above, one of the issues we address is that of *observability*. It concerns, roughly speaking, the issue of determining whether one can estimate the total energy of vibrations by partial measurements made, for instance, in one or several interior or external nodes of the network.[1] It is therefore intimately related to the way the energy of solutions is distributed along the various components of the multi-structure, as time evolves. This problem is relevant, not only because it is a way of analyzing deeply the nature of vibrations, but because it is also of immediate application in the context of inverse and control problems.

We also present the consequences of the observability properties in what concerns *controllability* issues. In this context, we are interested in driving the solutions to a given final state by means of the action of one or several controllers located in some of the internal nodes and/or the extremes of the network.[2] The problems of observability and controllability are dual one to each other and, therefore, the observability inequalities have immediate consequences in the controllability setting.

It is however important to underline that one of the difficulties related to dealing with networks and not the standard wave equation in an open domain of the Euclidean space or a smooth manifold is that, even if observability holds, the observed norm is weaker than the energy of the system, in analogy of the well-known behavior for the 1-d wave equation with point-wise interior observations.[3]

As we shall see, for instance, when the network is a tree, observing/controlling in all but one external vertices suffices to get full observation or control in the natural energy spaces (see [25]). This case is similar to that of the wave equation in a bounded domain with a control on a sufficient large subset of the boundary, fulfilling the so called *Geometric Control Condition (GCC)* by Bardos, Lebeau and Rauch [4]. But the problem becomes immediately much more complex when the control misses two external vertices. Then, diophantine approximation issues enter, as it happens for the internal point-wise control of the 1-d wave equation [13]. The situation is even more complex when the graph contains closed circuits. Then there may exist eigen-vibrations of the network that remain concentrated and trapped in that circuit, without being propagated to the rest of the network. In those cases, obviously, it is impossible to achieve the observation and/or control property if the observer or controller is not located on the circuit where the solution is trapped. But whether a circuit may support a localized eigen-vibration depends also strongly on the number theoretical properties of the lengths of the strings composing the circuit. This is an issue that is not completely well understood.

Our main result for general networks asserts that the problem of observability or controllability for a sufficiently large time (twice the total length of the network) is equivalent to the property that all eigen-vibrations to be observable. The later is, obviously, a necessary condition for observability and controllability. Our result

[1] As mentioned above, most of this article is devoted to the case in which the observation is only done on one external node of the network.

[2] Once more, we shall focus in the case in which one single controller acts on one of the external nodes of the network.

[3] We refer to [48, 50] for relatively complete and updated surveys on the state of the art of the observability and controllability of PDE's.

shows that it is also sufficient for observability/controllability to take place in spaces that can be described in Fourier series in terms of a summability condition of the Fourier coefficients with suitable weights. This is done using a corollary due to Haraux and Jaffard [18] of the celebrated Beurling–Malliavin's Theorem. However, characterizing the rate of decay of these weights for high frequencies (or, in other words, the spaces in which observability/controllability holds) in terms of the topological and geometrical properties of the graph is an open problem.

The overall picture is quite complex, and still not complete. We shall summarize the known results in this topic in Sect. 3.

In what concerns the problem of stabilization, recently, a black-box strategy has been developed in [41] allowing to automatically transfer the known observability/controllability results into stabilization ones. This provides a new way of getting stabilization results and complements the existing literature on the subject (some of the main references are collected in the bibliography at the end of the paper). Roughly speaking, whenever the wave process in the network is observable/controllable by some internal or exterior nodes, then the system can also be stabilized by feedback laws acting on the same nodes. But, of course, there is also a price to pay for the fact that the observation/control properties only hold in weaker spaces. In the context of stabilization, this amounts to get slow decay rates for smooth solutions (say, in the domain of the generator of the semigroup) and not exponential ones. The decay rate, roughly speaking, is polynomial when there is a loss of a finite number of derivatives in the observation/control process, but it may be even slower, say, logarithmic, when an infinite number of derivatives is lost in the observation/control process. Once again, the precise weak norm in which observability and/or controllability holds, depends on diophantine properties of the mutual lengths of the strings of the network.

The same issues arise for all other models like beams, Schrödinger or heat equations. The theory of observation and control of these models in open domains of $\mathbb{R}^n$ is by now quite well developed (we refer to the survey articles [47, 48, 50] for an updated account of the developments in this field). However, very little is known in the context of PDE's on networks. However, as pointed out in [7], one can transform the results obtained in the context of the wave equation in networks into results on the control of these systems in the same networks. In [7] this was proved to be true using the classical strategy by D.L. Russell [39] that was the first one to observe that the control to zero of the heat equation can be derived as a consequence of the exact controllability of the wave equation in domains of the Euclidean space. Recently, this issue has been further developed and clarified by L. Miller by the so-called transmutation method (see [35]), using the Kannai transform. We shall not develop this issue here but, for these models, as expected, due to the infinite speed of propagation, the observability inequalities hold in an arbitrarily small time [7]. It is however important to underline that, so far, the direct analysis of the control/observation properties of the Schrödinger and heat equations on networks has not been addressed.

As we have already mentioned, this article collects the existing results on simple 1-d models on networks. Much remains to be done in this field. At the end of this article we include a list of open problems and possible subjects of future research.

For those who will address these topics for the first time, we refer to [34] for an introduction to some of the most elementary tools on the controllability of PDE's and to the survey articles [47, 48, 50], for a description of the state of the art in this field.

This article is organized as follows. Section 2 is devoted to present the model under consideration: the wave equation on a 1-d network of strings. In Sect. 3 we make a brief presentation of known results on the observability and controllability of this model. In Sect. 4 we present known results on the problem of stabilization. In Sect. 5 we present and discuss some possible further developments of the methods and results in the paper and some open problems and future directions of research.

2 The Wave Equation on a Network

Let us first recall some definitions and notations about 1-d networks used in the paper. We refer to [7, 33, 36, 42] for more details.

A 1-d network $\mathscr{N}$ is a connected set of $\mathbb{R}^n$, $n \geq 1$, defined by

$$\mathscr{N} = \bigcup_{j=1}^{M} e_j$$

where e_j is a curve that we identify with the interval $(0, l_j)$, $l_j > 0$, and such that for $k \neq j$, $\overline{e_j} \cap \overline{e_k}$ is either empty or a common end called a vertex or a node (here $\overline{e_j}$ stands for the closure of e_j).

For a function $u : \mathscr{N} \longrightarrow \mathbb{R}$, we set $u^j = u_{|e_j}$ the restriction of u to the edge e_j.

We denote by $\mathscr{E} = \{e_j \,;\, 1 \leq j \leq M\}$ the set of edges of $\mathscr{N}$, by $\mathscr{V}$ the set of external nodes of $\mathscr{N}$, and by N the number of these external nodes. For a fixed vertex v, let

$$\mathscr{E}_v = \{j \in \{1, \ldots, M\} \,;\, v \in \overline{e_j}\}$$

be the set of edges having v as vertex. If card $(\mathscr{E}_v) = 1$, v is an exterior node, while if card $(\mathscr{E}_v) \geq 2$, v is an interior one. We denote by $\mathscr{V}_{ext}$ the set of exterior nodes and by $\mathscr{V}_{int}$ the set of interior ones. For $v \in \mathscr{V}_{ext}$, the single element of $\mathscr{E}_v$ is denoted by j_v.

Now we consider a planar network of elastic strings that undergo small perpendicular vibrations. At rest, the network coincides with a planar graph G contained in that plane.

Let us suppose that the function $u^j = u^j(t, x) : \mathbb{R} \times [0, \ell_j] \to \mathbb{R}$ describes the transversal displacement in time t of the string that coincides at rest with the edge

$\mathbf{e}_j$. Then, for every $t \in \mathbb{R}$, the functions u^j, $j = 1, ..., M$, define a function $\bar{u}(t)$ on G with components $u^j : \mathbb{R} \times [0, \ell_j] \to \mathbb{R}$ given by $u^j(t, x) = u^j(t, x_j(x))$.

As a model of the motion of the network, we assume that the displacements u^j satisfy the following non-homogeneous system

$$\begin{cases} u^j_{tt} - u^j_{xx} = 0 & \text{in } \mathbb{R} \times [0, \ell_j], \quad j = 1, ..., M, \\ u^{j_{\mathbf{v}_1}}(t, \mathbf{v}_1) = h(t) & t \in \mathbb{R}, \\ u^{j_{\mathbf{v}_i}}(t, \mathbf{v}_i) = 0 & t \in \mathbb{R}, \quad i = 2, ..., N, \\ u^j(t, \mathbf{v}) = u^k(t, \mathbf{v}) & t \in \mathbb{R}, \quad \mathbf{v} \in \mathscr{V}_{int}, \ j, k \in \mathscr{E}_{\mathbf{v}}, \\ \sum_{j \in I_{\mathbf{v}}} \partial_n u^i(t, \mathbf{v}) = 0 & t \in \mathbb{R}, \quad \mathbf{v} \in \mathscr{V}_{int}, \\ u^j(0, x) = u^j_0(x), \ \ u^j_t(0, x) = u^j_1(x) & x \in [0, \ell_j], \quad j = 1, ..., M. \end{cases} \tag{1}$$

The first equation in this system represents the classical 1-d wave equation on the network. Within each of the M strings of the network the d'Alembert equation is fulfilled. The second and third equalities reflect the condition that over the exterior node $\mathbf{v}_1$ a control $h = h(t)$ acts to regulate its displacement, while the remaining $N-1$ exterior nodes, are fixed. The fourth and fifth relations constitute the Kirchhoff joint conditions, expressing the continuity of the network and the balance of forces at the interior nodes. Finally, the last equation imposes the initial deformation and velocity of the strings (i.e., at time $t = 0$). The pair $(\bar{u}_0, \bar{u}_1)$ is called *initial state* of the network.

Here and in the sequel $\partial_n u^j(t, \mathbf{v})$ denotes the exterior normal derivative of u^j at the node $\mathbf{v}$.

Thus, (1) corresponds to a network with one controlled exterior node. Similar problems can be formulated when the controller acts on an interior node or when several controllers act simultaneously, either on interior or exterior nodes. We refer to [7] for a discussion of some of these problems.

For a proper functional analysis of this system, it is convenient to introduce the following Hilbert spaces:

$$V = \{\bar{u} \in \prod_{i=1}^{M} H^1(0, \ell_i) : u^i(\mathbf{v}) = u^j(\mathbf{v}) \text{ if } \mathbf{v} \in \mathscr{V}_{int} \text{ and } u^i(\mathbf{v}) = 0 \text{ if } \mathbf{v} \in \mathscr{V}_{ext}\},$$

$$H = \prod_{i=1}^{M} L^2(0, \ell_i),$$

endowed with the Hilbert structures

$$< \bar{u}, \bar{w} >_V := \sum_{i=1}^{M} < u^i, w^i >_{H^1(0,\ell_i)} = \sum_{i=1}^{M} \int_0^{\ell_i} u^i_x w^i_x dx,$$

$$< \bar{u}, \bar{w} >_H := \sum_{i=1}^{M} < u^i, w^i >_{L^2(0,\ell_i)} = \sum_{i=1}^{M} \int_0^{\ell_i} u^i w^i dx,$$

respectively. Besides, we will denote by $U = L^2(0,T)$, the space of controls. We also denote by W the product eergy space $W = V \times H$.

Since the imbedding $V \subset H$ is dense and compact, when H is identified with its dual H' by means of the Riesz–Fréchet isomorphism, we can define the operator $-\Delta_G : V \to V'$ by

$$\langle -\Delta_G \bar{u}, \bar{v} \rangle_{V' \times V} = \langle \bar{u}, \bar{v} \rangle_V.$$

The operator $-\Delta_G$ is an isometry from V to V'. The notation $-\Delta_G$ is justified by the fact that, for smooth functions $\bar{u} \in V$, the operator $-\Delta_G$ coincides with the Laplace operator.

The spectrum of the operator $-\Delta_G$ is formed by an increasing positive sequence $(\mu_n)_{n\in\mathbb{N}}$ of eigenvalues. The corresponding eigenfunctions $(\bar{\theta}_n)_{n\in\mathbb{N}}$ may be chosen to form an orthonormal basis of H.

The spaces V and H may be characterized as

$$V = \left\{ \bar{u} = \sum_{n\in\mathbb{N}} u_n \bar{\theta}_n : \quad ||\bar{u}||_V^2 := \sum_{n\in\mathbb{N}} \mu_n u_n^2 < \infty \right\},$$

$$H = \left\{ \bar{u} = \sum_{n\in\mathbb{N}} u_n \bar{\theta}_n : \quad ||\bar{u}||_H^2 := \sum_{n\in\mathbb{N}} u_n^2 < \infty \right\},$$

and the norms of V and H are equivalent to $||.||_V$ and $||.||_H$, respectively. The spaces V and H are Hilbert spaces with respect to the scalar products that generate the corresponding norms.

System (1) can be shown to be well-posed in an appropriate functional setting by means of the standard *transposition method* (see [30]).

To implement the method of transposition we need to consider the adjoint system[4]:

$$\begin{cases} \phi^j_{tt} - \phi^j_{xx} = 0 & \text{in } \mathbb{R} \times [0, \ell_j], \quad j = 1, ..., M, \\ \phi^{j_{\mathbf{v}_j}}(t, \mathbf{v}_j) = 0 & t \in \mathbb{R}, \quad j = 1, ..., N, \\ \phi^j(t, \mathbf{v}) = \phi^k(t, \mathbf{v}) & t \in \mathbb{R}, \quad \mathbf{v} \in \mathscr{V}_{int}, \ j, k \in \mathscr{E}_{\mathbf{v}}, \\ \sum_{j\in I_{\mathbf{v}}} \partial_n \phi^j(t, \mathbf{v}) = 0 & t \in \mathbb{R}, \quad \mathbf{v} \in \mathscr{V}_{int}, \\ \phi^j(0, x) = \phi^j_0(x), \quad \phi^j_t(0, x) = \phi^j_1(x) & x \in [0, \ell_j], \quad j = 1, ..., M. \end{cases} \tag{2}$$

The solution of the adjoint system (2) with initial data

$$\bar{\phi}_0 = \sum_{n\in\mathbb{N}} \phi_{0,n} \bar{\theta}_n, \qquad \bar{\phi}_1 = \sum_{n\in\mathbb{N}} \phi_{1,n} \bar{\theta}_n, \tag{3}$$

[4] More rigorously, for the adjoint system, the initial data should be given at time $t = T$, but the system under consideration being time-reversible, we may consider equally that the initial data are given at $t = 0$.

can be written in Fourier series as follows:

$$\bar{\phi}(t,x) := \sum_{n\in\mathbb{N}} (\phi_{0,n} \cos\sqrt{\mu_n}t + \frac{\phi_{1,n}}{\sqrt{\mu_n}} \sin\sqrt{\mu_n}t)\bar{\theta}_n(x). \tag{4}$$

When $(\bar{\phi}_0, \bar{\phi}_1) \in V \times H$, by standard variational or semigroup methods it can be shown that the solution $\bar{\phi}$ satisfies

$$\bar{\phi} \in C([0,T];V) \bigcap C^1([0,T];H), \tag{5}$$

for all $T > 0$.

For a classical smooth solution $\bar{u}$ of (1), the energy is defined as the sum of the energies of its components, that is,

$$\mathbf{E}_{\bar{u}}(t) := \sum_{j=1}^{M} \mathbf{E}_{u^j}(t) \quad \text{with} \quad \mathbf{E}_{u^j}(t) := \frac{1}{2}\int_0^{\ell_j} \left(\left| u_t^j(t,x) \right|^2 + \left| u_x^j(t,x) \right|^2 \right) dx.$$

This energy satisfies

$$\frac{d}{dt}\mathbf{E}_{\bar{u}}(t) = \sum_{j=1}^{M} u_t^j(t,\mathbf{v}_j)\partial_n u^j(t,\mathbf{v}_j). \tag{6}$$

In particular, for the adjoint system (2), the energy is conserved for all t:

$$\mathbf{E}_{\bar{\phi}}(t) = \mathbf{E}_{\bar{\phi}}(0),$$

for every $t \in \mathbb{R}$. Besides, if the initial data are as in (3) then

$$\mathbf{E}_{\bar{\phi}} = \frac{1}{2}\sum_{n\in\mathbb{N}} (\mu_n \phi_{0,n}^2 + \phi_{1,n}^2) = \frac{1}{2}(||\bar{\phi}_0||_V^2 + ||\bar{\phi}_1||_H^2). \tag{7}$$

For every $s \in \mathbb{R}$ we also consider the Hilbert spaces

$$V^s := \left\{ \bar{u} = \sum_{n\in\mathbb{N}} u_n \bar{\theta}_n : \quad \|\bar{u}\|_s^2 := \sum_{n\in\mathbb{N}} \mu_n^s |u_n|^2 < \infty \right\}, \tag{8}$$

$$h^s := \left\{ (u_n) : \quad \|(u_n)\|_s^2 := \sum_{n\in\mathbb{N}} \mu_n^s |u_n|^2 < \infty \right\}, \tag{9}$$

endowed with the norms $\|\cdot\|_s$, where (u_n) denotes a sequence of real numbers u_n. The canonical isomorphism $\sum_{n\in\mathbb{N}} u_n \bar{\theta}_n \to (u_n)$ is an isometry between V^s and h^s.

Let us observe that V^s is the domain of $(-\Delta_G)^{\frac{s}{2}}$ considered as an unbounded operator from H to H. Besides, $V = V^1$ and $H = V^0$.

Further, we introduce the Hilbert spaces

$$\mathscr{W}^s := V^s \times V^{s-1},$$

endowed with the natural product structures. We then have

$$\mathscr{W}^1 = V \times H, \qquad \mathscr{W}^0 = H \times V'.$$

For initial state $(\bar{\phi}_0, \bar{\phi}_1) \in \mathscr{W}^s$ the solution of the homogeneous problem (2) may be defined by (5) and

$$\bar{\phi} \in C(\mathbb{R}; V^s) \bigcap C^1(\mathbb{R}; V^{s-1}).$$

Furthermore, the solutions of the adjoint system, for all $T > 0$ finite and every exterior node $\mathbf{v} \in \mathscr{V}_{ext}$ satisfy the following *hidden regularity* inequality

$$\int_0^T |\partial_n \phi^j(t, \mathbf{v})|^2 dt \leq C \mathbf{E}_{\bar{\phi}}. \tag{10}$$

The inequality (10) may be proved using d'Alembert formula for the representation of the solutions of the wave equation in each string of the network, or multiplier techniques (see [25]).

If we multiply the first equation in (2) by u^j and integrate over $[0, t] \times [0, \ell_i]$ it holds, after integration by parts,

$$\int_0^t h \partial_n \phi^1(\tau, \mathbf{v}_1) d\tau = \sum_{j=1}^{M} \int_0^{\ell_j} \Big(u^j(t, x) \phi_t^j(t, x) - u_t^j(t, x) \phi^j(t, x) \Big) dx \,|_0^t.$$

We consider this identity as the definition of weak solution $\bar{u}$ of (1) in the sense of distributions. Given $h \in L^2(0, T)$, as a consequence of (10), this solution is well-defined, unique and, by (10), has the property

$$\bar{u} \in C([0, T] : H) \bigcap C^1([0, T]; V'), \tag{11}$$

together with the estimate

$$||\bar{u}||_{L^\infty(0,T:H)} + ||\bar{u}_t||_{L^\infty(0,T:V')} \leq C[||(\bar{u}_0, \bar{u}_1)||_{H \times V'} + ||h||_{L^2(0,T)}]. \tag{12}$$

The control problem in time T consists in determining for which initial states it is possible to choose the control $h \in L^2(0, T)$, such that the system reaches the equilibrium position at time T. Depending on how strict we are on requiring the state to reach equilibrium, several notions or degrees of controllability may be distinguished.

More precisely, given $T > 0$, we say that the initial state $(\bar{u}_0, \bar{u}_1) \in H \times V'$, is **exactly controllable** (or simply controllable) **in time** T, if there exists a function $h \in L^2(0, T)$, such that the solution of (1) with initial state $(\bar{u}_0, \bar{u}_1)$ satisfies

$$\bar{u}|_{t=T} = \bar{u}_t|_{t=T} = \bar{0}.$$

The system is said to be **approximately controllable in time** T when for every $\varepsilon > 0$ there exists a control h such that the corresponding solutions $\bar{u}^\varepsilon$ verifies

$$\left\| \left(\bar{u}^\varepsilon|_T, \bar{u}^\varepsilon_t|_T\right) \right\|_{H \times V'} < \varepsilon.$$

Here we shall mainly focus on the problem of controllability and present the existing results guaranteeing that the system is controllable within a class of initial data that one might identify.

Using the definition of solutions of the state equation by means of transposition the control property can be characterized in the following manner:

Proposition 2.1. *The initial state* $(\bar{u}_0, \bar{u}_1) \in H \times V'$ *is controllable in time* T *with control* $h \in U$ *if, and only if, for every* $(\bar{\phi}_0, \bar{\phi}_1) \in V \times H$ *the following equality holds*

$$-\langle \bar{u}_0, \bar{\phi}_1 \rangle_H + \langle \bar{u}_1, \bar{\phi}_0 \rangle_{V' \times V} = \int_0^T h(t) \partial_n \phi^{j_{\mathbf{v}_1}}(t, \mathbf{v_1}) dt, \tag{13}$$

where $\bar{\phi}$ *is the solution of system (2) with initial state* $(\bar{\phi}_0, \bar{\phi}_1)$.

The relation (13) suggests a minimization algorithm for the construction of the control h. If we look for the control in the form $h = -\partial_n \bar{\psi}(\mathbf{v}_1, t)$, where $\bar{\psi}$ is a solution of the homogeneous system (2), then the equality (13) is the Euler equation $I'(\bar{\psi}_0, \bar{\psi}_1) = 0$ corresponding to the quadratic functional $I : V \times H \to \mathbb{R}$ defined by

$$I(\bar{\phi}_0, \bar{\phi}_1) = \frac{1}{2} \int_0^T |\partial_n \phi^i(t, \mathbf{v}_1)|^2 dt + \langle \bar{u}_0, \bar{\phi}_1 \rangle - \langle \bar{u}_1, \bar{\phi}_0 \rangle.$$

Therefore, if $(\bar{\psi}_0, \bar{\psi}_1)$ is a minimizer of I, the relation (13) will be verified. The functional I is continuous and convex. So, in order to guarantee the controllability of an initial state $(\bar{u}_0, \bar{u}_1) \in H \times V'$ it is sufficient that I be coercive. This is the central idea of the Hilbert Uniqueness Method (HUM) introduced by J.-L. Lions in [29].

The coercivity of the functional is equivalent to the following observability inequality:

$$||(\bar{\phi}_0, \bar{\phi}_1)||_*^2 \leq C \int_0^T |\partial_n \phi^i(t, \mathbf{v}_1)|^2 dt, \tag{14}$$

for all solutions of the adjoint system, where $||\cdot||_*$ stands for a norm to be identified.

Once the norm $||\cdot||_*$ has been identified and the observability inequality (14) proved, the controllability property can be guaranteed to hold for all initial data $(\bar{u}_0, \bar{u}_1)$ in $(\mathscr{W}^*)'$, the dual of $\mathscr{W}^*$, the Hilbert space obtained as the closure of $\mathscr{W}^1$ with respect to the norm $||\cdot||_*$.

The main issue is then the obtention of inequalities of the form (14), giving quantitative informations about the norm $||\cdot||_*$, so that the spaces in which observability and controllability hold, $\mathscr{W}^*$ and $(\mathscr{W}^*)'$, respectively, might be identified. These issues depend in a very sensitive manner on the topological and number theoretical properties of the network under consideration.

As we shall see later, the problem of stabilization can also be solved once the observability inequality (14) is well understood. Indeed, we shall present a black-box strategy recently developed in collaboration with J. Valein [41], allowing to get observability inequalities (and, consequently, decay properties) for the solutions of wave equations in networks with dissipative boundary conditions, as a consequence of their conservative counterparts.

3 Main Results on Observability and Controllability

3.1 *Summary of Known Results*

The state of the art in what concerns the observability/controllability problem is more or less the one presented in [7], where the following three cases, in increasing complexity, were discussed. We summarize here the known main results.

- **The star.** In the star-like network a finite number of strings are connected on a single point by one of their extremes. This is a particular case of a tree-like network that we shall discuss below (Fig. 1).

  If the observation/control acts on all but one external vertices of the star, one gets observability/controllability in the optimal energy spaces. In other words, we get (14) with $\mathscr{W}^* = \mathscr{W}^1$.

  To the contrary, in the opposite case in which the observation/control is only applied in one external vertex, as we are doing here, then the space of observation and/or control can be described in Fourier series by means of suitable weights depending on the lengths of the strings entering in the star. These weights depend on the ratios of the lengths of the strings and, in particular, on their irrationality properties. In case when some of the non-controlled strings are mutually rational, some of these weights vanish and then the observability/controllability properties fail to hold. To the contrary, when they are all mutually irrational, all these weights are strictly positive but their lower envelop tends to zero for high frequencies so that there is always some loss in the spaces in which the observability/controllability problems are solvable. How important this loss is depends on the diophantine approximation properties of the quotients of the lengths. In particular, the weights may degenerate

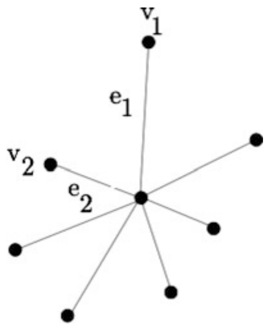


Fig. 1 A star-shaped network

exponentially when some of the quotient of the lengths is a Liouville number. But, regardless of the diophantine properties and the nature of the spaces in which the observability/controllability properties hold, the time needed for observation turns out to be twice the sum of all the lengths of the strings of the networks.

It is interesting to analyze the relation of this result with the so-called Geometric Control Condition (GCC) introduced by Bardos, Lebeau and Rauch [4] in the context of the boundary observation and/or control of the wave equation in bounded domains of $\mathbb{R}^n$. The GCC requires that all the rays of Geometric Optics enter the observation region in a finite, uniform time, which turns out to be the minimal one for observation/control. In the case of the star shaped network this would correspond to the maximum of sum of the lengths of any pair of two strings.

But this time is insufficient for the control from only one end-point. As we mentioned, above, indeed, the time needed is twice the sum of all the lengths of all the strings of the star-shaped network. This control time is closer to the one gets when one string is controlled at an interior point or two strings are controlled by a single control on a common vertex. In that case the minimal control time is $2(\ell_1 + \ell_2)$ and not $2\max(\ell_1, \ell_2)$, ℓ_1, ℓ_2 being the lengths of the two strings. The wave equation is a second order problem and therefore, even in 1-d, for a point-wise observation mechanism to be efficient we need to measure not only the position, but also the space derivative. This implies that a necessary condition for observation/control is that all waves pass twice through the observation point. This is guaranteed when the time of control is larger than $2(\ell_1 + \ell_2)$. But, in fact, passing twice by the observation point is not sufficient either. The irrationality of the ratio ℓ_1/ℓ_2 is needed to guarantee that, when passing through the observation point the second time, the solution is not exactly at the configuration as in the first crossing, which, of course, would make the second observation to be insufficient too. Finally, even when ℓ_1/ℓ_2 is irrational, we cannot get a uniform bound of the energy of the solution but rather a weaker measurement in a weaker norm. The nature of this norm, which is represented in Fourier series by means of some weights depending on ℓ_1/ℓ_2, depends very strongly on the irrationality class to

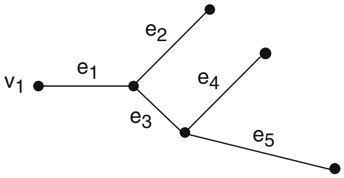


Fig. 2 A tree-shaped network

which the number ℓ_1/ℓ_2 belongs. In fact, even in the most favorable case, i.e., when ℓ_1/ℓ_2 is an algebraic number of degree two, one looses one derivative with respect to the expected energy norm.

We refer to [7] for an in depth discussion of the problem of simultaneous control of a finite number of strings and its connections with the problem of the control of star networks.

- **The tree.** The tree-like network is a generalization of the star-like one. As we said above, it is well known that, when all but one external nodes of the network are observed on a tree-like configuration, the whole energy of solutions may be observed (see [25]). This can be easily seen by sidewise energy estimates for the solutions of the wave equation. In this case the observation inequality holds in the sharp energy space in a time which is twice the length of the longest path joining the points of the network with some of the observed ends, which is much smaller than twice the total length of the network, which was the time needed for the observation from a single end in the case of stars mentioned above. This smaller observability time is the one that coincides with the one given by the GCC in the case of waves in domains of the Euclidean space (Fig. 2).

  In the opposite case in which the observation is made at one single extreme of the tree-like network, the observation time turns out to be, again, twice the sum of the lengths of the strings forming the network.

  But for the observability inequality to be true in the case of the tree one needs a condition extending the one that, in the case of stars, requires the strings to have mutually irrational lengths. In [7] it was observed that this condition can be recast in spectral terms: two strings have mutually irrational lengths if and only if their Dirichlet spectra have empty intersection.

  The latter condition turns out to be the appropriate one to be extended to general trees: the wave equation on a tree is observable from one end if and only if the spectra of all pairs of subtrees of the tree that match on an interior node are disjoint.

  This allows showing, in particular, that, generically within the class of trees (i. e. for almost all tress), this property is satisfied and then, the wave process is observable/controllable from one single node. But the space in which the observability/controllability holds depends in a subtle manner on the distance between the various spectra of the corresponding subtrees and how it vanishes asymptotically at high frequencies.

Note however that the identification of the precise norm $||\cdot||_*$ in which the observability inequality (14) holds is a delicate issue.

- **General networks.** The characterization we have given of controllable stars and trees is hard to be extended to general graphs. Indeed, in the general case, we lack of a natural ordering on the graph to analyze the propagation of waves and, for instance, when the graph contains cycles, the condition of empty intersection of subgraphs is hard to extend. Actually, as we mentioned above, the presence of closed circuits may trap the waves thus making impossible the controllability/observability properties to hold from an external node.

  Thus, in the analysis of general graphs, we proceed in a different way by applying a consequence of the celebrated Beurling–Malliavin's Theorem on the completeness of families of real exponentials obtained by Haraux and Jaffard in [18] when analyzing the control of plates. Using the min-max principle, one can show that the spectral density of a general graph is the same as that of a single string whose length, L, is the sum of the lengths of all the strings entering in the network. Then, when the time is greater than twice the total length, as a consequence of Beurling–Malliavin's Theorem, we deduce that there exist some Fourier weights so that the observation property holds in the corresponding weighted norm if and only if all the eigenfunctions of the network are observable.

  So far we do not know of any necessary and sufficient condition guaranteeing that all the eigenfunctions are observable in the general case. However, this condition, in the particular case of stars and trees discussed above turns out to be sharp and equivalent to the ones we have identified in each particular case: (a) the condition that lengths of the strings are mutually irrational in the case of stars or (b) that the spectra of all pairs of subtrees with a common end-point to be mutually disjoint in the more general case of trees.

3.2 The Weighted Observability Inequality

In the previous section we have described the main existing results on the observability of graphs distinguishing three different cases, in increasing complexity: the star, the tree and general graphs. In each case, under suitable assumptions, we obtain the observability inequality (14) for a suitable norm $||\cdot||_*$. This norm can be characterized in terms of the Fourier coefficients by suitable weights. This subsection is devoted to explain this fact, which plays a critical role in the control and stabilization results one can get out of this analysis, and that will be discussed in the next section.

Recall that if we suppose that $(\bar{\phi}^{(0)}, \bar{\phi}^{(1)}) \in \mathscr{W}^1$, then problem (2) admits a unique solution

$$\bar{\phi} \in C(\mathbb{R};\, V) \cap C^1(\mathbb{R};\, H).$$

The observability inequalities we have described can be rewritten in terms of the Fourier expansion (4) as follows:

$$\sum_{n\geq 1} c_n^2(\mu_n \phi_{0,n}^2 + \phi_{1,n}^2) \leq C \int_0^T \left| \frac{\partial \phi_1}{\partial x}(\mathbf{v}_1, t) \right|^2 dt. \tag{15}$$

This holds in the situations described above, under the corresponding assumptions on the network, for T large enough (twice the sum of the lengths of all the strings entering in the network, $T > 2L$) and for a suitable observability constant $C > 0$ and weights $\{c_n\}_{n\geq 1}$. The norm $|| \cdot ||_*$ arising in the observability inequality is therefore as follows:

$$||(\bar{\phi}_0, \bar{\phi}_1)||_* = \left[\sum_{n\geq 1} c_n^2(\mu_n \phi_{0,n}^2 + \phi_{1,n}^2) \right]^{1/2}. \tag{16}$$

Obviously, the nature of this norm depends on how fast the weights $\{c_n\}_{n\geq 1}$ tend to zero as $n \to \infty$.

Recall however that, in each case, extra assumptions are needed to ensure that the weights c_n^2 are strictly positive for every $n \in \mathbb{N}^*$.

One of the most interesting open problems in this context is to give sharp sufficient conditions on the network so that these weights have a given asymptotic lower bound as $n \to \infty$. At this respect, the case of the star network is the simplest one: it then suffices to impose diophantine conditions on the quotients of the lengths of the strings entering in the network to get those lower bounds.

4 Stabilization

4.1 Problem Formulation

So far we have considered an open-loop control problem. In this section we discuss the closed-loop counterpart in which the goal is to find suitable feedback mechanisms ensuring the decay as $t \to \infty$ of solutions.

Recall that, in the control and observation problems above we have distinguished one vertex $\mathbf{v}_1$ among all the exterior ones $\mathcal{V}_{ext}$: the one in which the control or the observation is being applied. The rest of the nodes in which the homogeneous Dirichlet boundary condition holds for the control problem is denoted by $\mathcal{V}_{\mathcal{D}}$. In this way, we distinguish the conservative exterior nodes, $\mathcal{V}_{\mathcal{D}}$, in which we impose Dirichlet homogeneous boundary conditions, and the one in which the damping term is effective, $\mathbf{v}_1$. To simplify the notation, we will assume that $\mathbf{v}_1$ is located at the end 0 of the edge e_1.

The system under consideration then reads as follows:

$$\begin{cases} \frac{\partial^2 y^j}{\partial t^2} - \frac{\partial^2 y^j}{\partial x^2} = 0 & 0 < x < l_j,\ t > 0,\ \forall j \in \{1, ..., M\}, \\ y^j(\mathbf{v}, t) = y^l(\mathbf{v}, t) & \forall j,\ l \in \mathscr{E}_{\mathbf{v}},\ \mathbf{v} \in \mathscr{V}_{int}, t > 0, \\ \sum\limits_{j \in \mathscr{E}_{\mathbf{v}}} \frac{\partial y^j}{\partial n_j}(\mathbf{v}, t) = 0 & \forall \mathbf{v} \in \mathscr{V}_{int}, t > 0, \\ y^{j_{\mathbf{v}}}(\mathbf{v}, t) = 0 & \forall \mathbf{v} \in \mathscr{V}_{\mathscr{D}}, t > 0, \\ \frac{\partial y^1}{\partial x}(0, t) = \frac{\partial y^1}{\partial t}(0, t) & \forall t > 0, \\ \bar{y}(0) = \bar{y}_0,\ \frac{\partial \bar{y}}{\partial t}(0) = \bar{y}_1, & \end{cases} \tag{17}$$

where $\partial y^j / \partial n_j(\mathbf{v}, .)$ stands for the outward normal (space) derivative of y^j at the vertex $\mathbf{v}$. Similarly the normal derivative at the vertex $\mathbf{v}_1 = 0$ where the dissipative boundary condition is imposed is denoted by $-\partial y^1(0, t)/\partial x$, y^1 being the deformation of the first edge with extreme $\mathbf{v}_1 = 0$. The deformation of the network at that point is given by $y^1(0, t)$. As usual, we denote by $\overline{y}$ the vector $\overline{y} = (y^j)_{j=1,...,M}$.

The above system has been considered in a number of articles where the decay rate of solutions has been investigated in some specific examples and, recently, an unified treatment has been given in [41]. We briefly present here the main ideas and results.

In order to study system (17) we need a proper functional setting which is slightly different to the one considered until now because of the damped boundary condition on one of the nodes. To be more precise, the space V above has to be replaced by:

$$V_{\mathscr{D}} = \{\bar{y} \in \prod_{j=1}^{M} H^1(0, \ell_j) : y^j(\mathbf{v}) = y^k(\mathbf{v}) \text{ if}$$
$$\mathbf{v} \in \mathscr{V}_{int}, \forall j,\ k \in \mathscr{E}_{\mathbf{v}} \text{ and } y^{j_{\mathbf{v}}}(\mathbf{v}) = 0 \text{ if } \mathbf{v} \in \mathscr{V}_{\mathscr{D}}\}.$$

The only difference between the space V above and the new one $V_{\mathscr{D}}$ is that, in the later, we do not impose the homogeneous Dirichlet boundary condition on $\mathbf{v}_1 = 0$.

It is easy to see by semigroup methods that this dissipative system is well posed in in the Hilbert space

$$\mathscr{W}_{\mathscr{D}} := V_{\mathscr{D}} \times H,$$

equipped with the canonical norm.

Then, for an initial datum in $\mathscr{W}_{\mathscr{D}} := V_{\mathscr{D}} \times H$, there exists a unique solution such that

$$\overline{y} \in C([0, \infty); V_{\mathscr{D}}) \bigcap C^1([0, \infty); H). \tag{18}$$

Moreover, the solutions remain in $D(\mathscr{A}_{\mathscr{D}})$, the domain of the operator $\mathscr{A}_{\mathscr{D}}$, for all $t > 0$ whenever the initial data belong to $D(\mathscr{A}_{\mathscr{D}})$:

$$D(\mathscr{A}_{\mathscr{D}}) := \{(y, z) \in (V_{\mathscr{D}} \cap \prod_{j=1}^{M} H^2(0, l_j)) \times V_{\mathscr{D}} \ : \ \frac{\partial y^1}{\partial x}(0)$$

$$= z^1(0) \, ; \ \sum_{j \in \mathscr{E}_v} \frac{\partial y^j}{\partial n_j}(\mathbf{v}) = 0, \ \forall \mathbf{v} \in \mathscr{V}_{int}\}.$$

For this dissipative system the energy satisfies the energy dissipation law

$$\frac{d}{dt}\mathbf{E}_{\bar{y}}(t) = -\left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 \leq 0, \tag{19}$$

and therefore it is decreasing.

Integrating the expression (19) between 0 and T, we obtain

$$\int_0^T \left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 dt = \mathbf{E}_{\bar{y}}(0) - \mathbf{E}_{\bar{y}}(T) \leq \mathbf{E}_{\bar{y}}(0).$$

This estimate implies that $\frac{\partial y^1}{\partial t}(0, \cdot)$ belongs to $L^2(0, T)$ for finite energy solutions.

The main goal of this section is to show how the results of previous sections on observability/controllability can be used to derive energy decay rates as $t \to \infty$ for smooth solutions in $D(\mathscr{A}_{\mathscr{D}})$. Obviously, the better the observability/controllability results, faster decay rates will be obtained.

Note, however, that in the context of observability/controllability we have considered only Dirichlet boundary conditions while in here we are imposing a dissipative boundary condition on one node. Thus, we need to reduce the problem of getting decay results for the damped systems into the one of observability inequalities for the conservative one with Dirichlet boundary conditions on all the exterior nodes. To do this we proceed in two steps:

- We first reduce it to the case of conservative Dirichlet–Neumann boundary conditions.
- To later reduce it to the case of purely Dirichlet conditions.

As we shall see, overall, this reduction argument allows obtaining an observability inequality for $\bar{y}$ out of the known ones for the solutions of the Dirichlet problem. The obtained observability inequality reads

$$\mathbf{E}_{\bar{y}}^{-}(0) \leq C \int_0^T \left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 dt, \tag{20}$$

for an energy $\mathbf{E}_{\bar{y}}^{-}(0)$ that we shall make precise below but that, definitely, will be weaker than the energy norm.

To obtain explicit decay rates out of this weak observability inequality we use an interpolation inequality which is a variant of the one from Bégout and Soria [5] and which is a generalization of Hölder's inequality. For this to be done we need to assume more regularity of the initial data.

To be more precise we shall consider initial data $(\bar{y}_0, \bar{y}_1) \in X_s := [D(\mathscr{A}_{\mathscr{D}}), \mathscr{W}_{\mathscr{D}}]_{1-s}$ for $0 < s < 1/2$ and deduce an interpolation inequality of the form

$$1 \leq \Phi_s \left(\frac{\mathbf{E}^-_{\bar{y}}(0)}{C\mathbf{E}_{\bar{y}}(0)} \right) \frac{\|(\bar{y}_0, \bar{y}_1)\|^2_{X_s}}{C'\mathbf{E}_{\bar{y}}(0)}, \tag{21}$$

where Φ_s is an increasing function which depends on s and on the weak energy $\mathbf{E}^-_{\bar{y}}$ under consideration.

The previous interpolation inequality implies

$$\mathbf{E}^-_{\bar{y}}(0) \geq C\mathbf{E}_{\bar{y}}(0)\Phi_s^{-1} \left(\frac{\mathbf{E}_{\bar{y}}(0)}{C' \|((\bar{y}_0, \bar{y}_1)\|^2_{X_s}} \right).$$

With (19) and (20), we obtain

$$\mathbf{E}_{\bar{y}}(0) - \mathbf{E}_{\bar{y}}(T) \geq C\mathbf{E}_{\bar{y}}(0)\Phi_s^{-1} \left(\frac{\mathbf{E}_{\bar{y}}(0)}{C' \|(\bar{y}_0, \bar{y}_1)\|^2_{X_s}} \right),$$

which implies, by the semigroup property (see Ammari and Tucsnak [1])

$$\forall t > 0, \mathbf{E}_{\bar{y}}(t) \leq C\Phi_s \left(\frac{1}{t+1} \right) \|(\bar{y}_0, \bar{y}_1)\|^2_{X_s}. \tag{22}$$

Obviously, the decay rate in (22) depends on the behaviour of the function Φ_s near 0. Thus, in order to determine the explicit decay rate, we need to have a sharp description of the function Φ_s, which depends on s and on the energies $\mathbf{E}_{\bar{y}}$ and $\mathbf{E}^-_{\bar{y}}$ and thus on the nature of the weak energy $\mathbf{E}^-_{\bar{y}}$ in an essential way and this depends on the topology of the network and the number theoretical properties of the lengths of the strings entering in it.

This approach allows getting in a systematic way decay rates for the energy of smooth solutions of the damped system as a consequence of the observability properties of the undamped one.

The key ingredients of the proof that remain to be developed are the following:

- To get the weak observability inequality (20) out of the previous results on the observability of the Dirichlet problem.
- To derive the interpolation inequality (21) with a precise estimate on the behavior of Φ_s.

4.2 Observability for the Damped System

This subsection is devoted to explain how the weak observability inequality (20) can be proved as a consequence of the results of previous sections on the Dirichlet problem on the same network.

Let us explain how the observability results of the purely Dirichlet case discussed in the previous sections can be applied directly to get an inequality of the form (20) for the solutions of (17).

For that, we decompose $\bar{y}$, the solution of (17), as the sum of $\bar{\phi}$, solution of (2), and a reminder term $\bar{\epsilon}$:

$$\bar{y} = \bar{\phi} + \bar{\epsilon}.$$

Recall that $\bar{\phi}$ is a solution of (2) with appropriate initial data $(\bar{y}_0 - y_0^1(0)\bar{\gamma},\ \bar{y}^1)$, where $\bar{\gamma}$ is a given smooth function such that $\gamma^1(0) = 1$ and vanishing on all other external nodes.

Applying (15) to the solution $\bar{\phi}$ of (2), we obtain the following weighted observability estimate (note that $y^1(0,0) = y_0^1(0)$)

$$||(\bar{y}_0 - y_0^1(0)\bar{\gamma},\ \bar{y}_1)||_*^2 \le C_T \int_0^T \left(\frac{\partial \phi^1}{\partial x}(0,\ t) \right)^2 dt, \tag{23}$$

where the weak norm $||\cdot||_*$ is defined with weights $(c_n^2)_n$ that tend to zero as $n \to \infty$, depending on the network, as described in the previous sections. It is however important to underline that (23) holds under the same assumptions on the network needed for observability to hold for the Dirichlet problem (2) and provided $T > 2L$.

The reminder term ϵ is the solution of the following non-homogeneous Dirichlet problem:

$$\begin{cases} \frac{\partial^2 \epsilon^j}{\partial t^2} - \frac{\partial^2 \epsilon^j}{\partial x^2} = 0 & \forall x \in (0,\ l_j),\ t > 0,\ \forall j \in \{1, ..., M\}, \\ \epsilon^j(v,\ t) = \epsilon^l(\mathbf{v},\ t) & \forall j,\ l \in \mathscr{E}_{\mathbf{v}},\ \mathbf{v} \in \mathscr{V}_{int}, t > 0, \\ \sum_{j \in \mathscr{E}_v} \frac{\partial \epsilon^j}{\partial n_j}(\mathbf{v},\ t) = 0 & \forall \mathbf{v} \in \mathscr{V}_{int}, t > 0, \\ \epsilon_{j_{\mathbf{v}}}(\mathbf{v}, t) = 0 & \forall \mathbf{v} \in \mathscr{V}_{\mathscr{D}}, t > 0, \\ \epsilon(0,\ t) = y(0,\ t) & t > 0, \\ \bar{\epsilon}(0) = y_0^1(0)\bar{\gamma},\ \frac{\partial \bar{\epsilon}}{\partial t}(0) = \bar{0}. \end{cases} \tag{24}$$

Note that ϵ satisfies a non-homogeneous Dirichlet boundary condition at $x = 0$. Actually it coincides with the initial value of the solution y^1 of (17) at that point. We know that the solution $\bar{y}$ of the dissipative problem, because of the energy dissipation law, is such that $\partial y^1(0,\ \cdot)/\partial t \in L^2(0,\ T)$, so that the non-homogeneous Dirichlet boundary condition belongs to $H^1(0,\ T)$.

Proceeding in this manner, the following result was proved in [41]:

Theorem 4.1. **([41])**. *Assume that the network is such that the weighted observability inequality (23) is satisfied for $T > 2L$ for the conservative system (2) with Dirichlet boundary conditions at all the exterior nodes. Define the weak energy $\mathbf{E}_{\bar{y}}^{-}(0)$ by*

$$\mathbf{E}_{\bar{y}}^{-}(0) := \frac{1}{2}\left[||(\bar{y}_0 - y_0^1(0)\bar{\gamma},\ \bar{y}_1)||_*^2 + y_0^1(0)^2\right]. \tag{25}$$

Then for all $T > 2L$, there exists $C_T > 0$ such that all solution $\bar{y}$ of (17) satisfies the weak observability inequality (20).

Note that $\left(\mathbf{E}_{\bar{y}}^{-}(0)\right)^{\frac{1}{2}}$ as above defines a norm in the space of initial data $(\bar{y}_0, \bar{y}_1) \in \mathscr{W}_{\mathscr{D}}$. Indeed, when $\mathbf{E}_{\bar{y}}^{-}(0)$ vanishes, $y_0^1(0) = 0$. Thus $(\bar{y}_0, \bar{y}_1) \in \mathscr{W}$ and then $\mathbf{E}_{\bar{y}}^{-}(0) = \mathbf{E}_{\bar{y}}(0)$, and, by assumption, $\left(\mathbf{E}_{\bar{y}}(0)\right)^{\frac{1}{2}}$ defines a norm in $\mathscr{W}$.

Let us now present a sketch of the proof of Theorem 4.1.

Note that, standard results on the hidden regularity of the wave equation guarantee that, for all $T > 0$ there exists $C_T > 0$ such that the solutions $\bar{y}$ of (17) and $\bar{\epsilon}$ of (24) satisfy the following estimate

$$\int_0^T \left(\frac{\partial \epsilon^1}{\partial x}(0, t)\right)^2 dt \le C_T \int_0^T \left[\left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 + \left(y^1(0, t)\right)^2\right] dt. \tag{26}$$

Despite of the fact that we are here working with the wave equation on a network, this result is of local nature and therefore it is sufficient to apply the standard multiplier techniques of the scalar wave equation in the string with vertex at $\mathbf{v_1} = 0$. With a little extra work the right hand side term of this inequality can be slightly weakened to yield

$$\int_0^T \left(\frac{\partial \epsilon^1}{\partial x}(0, t)\right)^2 dt \le C_T \left(\int_0^T \left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 dt + \left(y_0^1(0)\right)^2\right). \tag{27}$$

Combining (23) and (27) and the fact that

$$\int_0^T \left(\frac{\partial \phi^1}{\partial x}(0, t)\right)^2 dt \le 2\int_0^T \left(\frac{\partial y^1}{\partial x}(0, t)\right)^2 dt + 2\int_0^T \left(\frac{\partial \epsilon^1}{\partial x}(0, t)\right)^2 dt$$

and that, due to the choice of the dissipative boundary condition,

$$\int_0^T \left(\frac{\partial y^1}{\partial x}(0, t)\right)^2 dt = \int_0^T \left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 dt$$

we have

$$\mathbf{E}_{\bar{y}}^{-}(0) \le C\left[\int_0^T \left(\frac{\partial y^1}{\partial t}(0, t)\right)^2 dt + \left|y_0^1(0)\right|^2\right]. \tag{28}$$

In fact, we can remove the last term in the right hand side of (28). To do this, it is sufficient to show that

$$\left|y_0^1(0)\right|^2 \leq C_T \int_0^T \left(\frac{\partial y^1}{\partial t}(0,\, t)\right)^2 dt$$

for a positive constant C_T depending on T.

This can be done by a classical compactness-uniqueness argument using the fact that the perturbation is of rank one (and therefore compact with respect to any norm) and the fact that whenever $\partial y_1(0,\, t)/\partial t$ and $\partial y_1(0,\, t)/\partial x$ vanish for $t \in (0, T)$ during a sufficiently long time interval ($T > 2L$), then, necessarily, $y_0(0) = 0$.

In this way, we conclude that the wanted inequality (20) is true.

4.3 The Interpolation Inequality

In this subsection we recall the main ingredients of the proof of the interpolation inequality (21). Its proof uses a discrete interpolation inequality, similar to that in [5], introduced in [41] and a description of the various energies and norms entering in the estimates we have obtained so far in terms of Fourier series.

The discrete interpolation inequality reads as follows:

Let $m \in [0,\, 1)$, $0 < s < 1/2$ and assume that

$$\omega : (m,\, \infty) \to (0,\, \omega(m)) \text{ is convex and decreasing with } \omega(\infty) = 0, \tag{29}$$

$$\Phi_s : (0,\, \omega(m)) \to (0,\, \infty) \text{ is concave and increasing with } \Phi_s(0) = 0, \tag{30}$$

$$\forall t \in [1,\, \infty),\ 1 \leq \Phi_s(\omega(t))t^{2s}, \tag{31}$$

$$\text{The function } t \mapsto \frac{1}{t}\Phi_s^{-1}(t) \text{ is nondecreasing on } (0,\, 1). \tag{32}$$

Under the conditions (30)–(31), we have the following result which is a generalized Hölder's inequality, a variant of Theorem 2.1 given in [5]:

Lemma 4.2. *Let* $(\omega,\, \Phi_s)$ *be as above satisfying (29)–(31). Then for any* $f = (f_n)_{n\in\mathbb{N}^*} \in l^1(\mathbb{N}^*)$, $f \neq 0$, *we have*

$$1 \leq \Phi_s\left(\frac{\sum\limits_{n\geq 1} |f_n|\, \omega(n)}{\sum\limits_{n\geq 1} |f_n|}\right) \frac{\sum\limits_{n\geq 1} |f_n|\, n^{2s}}{\sum\limits_{n\geq 1} |f_n|}, \tag{33}$$

as soon as $(f_n\omega(n))_n \in l^1(\mathbb{N}^*)$ *and* $(f_n n^{2s})_n \in l^1(\mathbb{N}^*)$.

We now give some examples of pairs (ω, Φ_s) satisfying (29)–(32):

1. If
$$\omega(t) = \frac{c}{t^p},$$
for some $p \geq 1$, we can take Φ_s of the form
$$\Phi_s(t) = \left(\frac{t}{c}\right)^{\frac{2s}{p}}.$$
We can easily prove that (ω, Φ_s) satisfy (29)–(30) with $m = 0$ and (31)–(32).
2. If
$$\omega(t) = Ce^{-At}$$
where $A > 2(2s+1)$ and $C > 0$, we can take Φ_s of the form
$$\Phi_s(t) = \left(\frac{A}{\ln\left(\frac{C}{t}\right)}\right)^{2s}.$$
We can easily prove that $t \mapsto \frac{1}{t}\Phi_s^{-1}(t)$ is nondecreasing on $(0, 1)$ and that the pair (ω, Φ_s) satisfies (31) on $[1, \infty)$. Thus (ω, Φ_s) satisfy (29)–(32) with $m = 1/2$.

When applying this argument, $\sum_{n\geq 1} |f_n|\, \omega(n)$ will play the role of the weak energy $\mathbf{E}^-_{\bar{y}}$, $\sum_{n\geq 1} |f_n|$ the role of the standard energy $\mathbf{E}_{\bar{y}}$ and $\sum_{n\geq 1} |f_n|\, n^{2s}$ that of the norm in X_s. But for this to be done, these energies and norms have to written in a suitable discrete manner.

We explain how this can be done distinguishing each of the terms:

- **The X_s-norm.** At this level, the fact that $0 < s < 1/2$ plays a key role. The following Lemma was proved in [41]:

  Lemma 4.3. **([41])** *Assume that $(\bar{y}_0, \bar{y}_1)$ belongs to X_s, where $0 < s < 1/2$, and $(\bar{\phi}_0, \bar{\phi}_1) = (\bar{y}_0 - y_0^1(0)\bar{\gamma}, \bar{y}_1)$, where $\bar{\gamma}$ is a given smooth function such that $\gamma^1(0) = 1$ and vanishing on all other external nodes. Then there exists a positive constant* $\mathbf{C}$ *such that*
$$\left\|\left(\bar{\phi}_0, \bar{\phi}_1\right)\right\|^2_{D((-\Delta_G)^s)} + \left|y_0^1(0)\right|^2 \leq C \left\|(\bar{y}_0, \bar{y}_1)\right\|^2_{X_s},$$
  where $D((-\Delta_G)^s)$ is the domain of the operator $(-\Delta_G)^s$, which is the s-th power of the Laplacian on the graph, $-\Delta_G$, with Dirichlet boundary conditions at all exterior nodes.

  This means that it is sufficient to prove the interpolation inequality (21) with the norm in X_s replaced by $\left[\left\|\left(\bar{\phi}_0, \bar{\phi}_1\right)\right\|^2_{D((-\Delta_G)^s)} + \left|y_0^1(0)\right|^2\right]^{1/2}$.

On the other hand, the norm $\left\|\left(\bar{\phi}_0, \bar{\phi}_1\right)\right\|^2_{D((-\Delta_G)^s)}$ can be written easily in terms of the Fourier coefficients of $\left(\bar{\phi}_0, \bar{\phi}_1\right)$ in the basis of eigenfunctions of $-\Delta_G$:

$$\left\|\left(\bar{\phi}_0, \bar{\phi}_1\right)\right\|^2_{D((-\Delta_G)^s)} = \sum_{n\geq 1}\left[\mu_n^{1+s}|\phi_{0,n}|^2 + \mu_n^s|\phi_{1,n}|^2\right],$$

where $(\phi_{0,n}, \phi_{1,n})$ are the Fourier coefficients of the data $\left(\bar{\phi}_0, \bar{\phi}_1\right)$.

- **The weak energy $\mathbf{E}^-_{\bar{y}}$.** According to the results of the previous sections and, in particular, (15), the observed weak energy can be rewritten as

$$\mathbf{E}^-_{\bar{y}}(0) = \sum_{n\geq 1} c_n^2(\mu_n\phi_{0,n}^2 + \phi_{1,n}^2) + \left|y_0^1(0)\right|^2. \tag{34}$$

- **The energy $\mathbf{E}_{\bar{y}}$.** Similarly, the energy $\mathbf{E}_{\bar{y}}$ is equivalent to the discrete norm:

$$\mathbf{E}_{\bar{y}} \sim \left\|\left(\bar{\phi}_0, \bar{\phi}_1\right)\right\|^2_{\mathscr{W}} + \left|y_0^1(0)\right|^2 = \sum_{n\geq 1}\left[\mu_n|\phi_{0,n}|^2 + |\phi_{1,n}|^2\right] + \left|y_0^1(0)\right|^2.$$

Once this is done, the interpolation inequality (21) is a consequence of the abstract discrete interpolation result (33).

In the next subsection we state the main stabilization result that this analysis yields.

4.4 The Main Result

Before moving further we observe that, as proved in [41], for $0 < s < 1/2$,

$$X_s = \left(V_{\mathscr{D}} \cap \prod_j H^{1+s}(0, l_j)\right) \times \prod_j H^s(0, l_j).$$

Thus, the space X_s of smooth initial data can be identified in classical Sobolev terms.

We assume that the network is such that the weighted observability inequality (15) holds. In the previous sections we have given sufficient conditions on the network for that to hold with positive weights $c_n > 0$.

The main stabilization result is as follows:

Theorem 4.4. *Assume that the weighted observability inequality (15) holds for every solution of (2) with* $\liminf_{n\to\infty} c_n = 0$ *and* $c_n \neq 0$ *for all* $n \in \mathbb{N}^*$. *Let* ω *be defined by a lower envelope of the sequence of weights* (c_n^2) *satisfying (29). Assume that the initial data* $(\bar{y}_0, \bar{y}_1)$ *belong to* X_s *where* $0 < s < 1/2$. *Let* Φ_s *be a function* *such that the pair (ω, Φ_s) satisfies (29)–(32). Then there exists a constant $C > 0$ such that the corresponding solution $\bar{y}$ of (17) verifies*

$$\forall t \geq 0,\ \mathbf{E}_{\bar{y}}(t) \leq C\Phi_s\left(\frac{1}{t+1}\right) \|(\bar{y}_0, \bar{y}_1)\|^2_{X_s}. \tag{35}$$

We see that the decay rate of the energy directly depends on the behavior of the interpolation function Φ_s near 0 and thus of ω and of the weights c_n^2 as $n \to \infty$.

Using (35) and making a particular and explicit choice of the concave function Φ_s, we obtain a more explicit decay rate. To be more precise, we set

$$\forall t > 0,\ \varphi(t) = \frac{\omega(t)}{t^2}.$$

Then there exists a constant $C > 0$ such that for any initial data $(\bar{y}_0, \bar{y}_1) \in X_s$ $(0 < s < 1/2)$, the corresponding solution $\bar{y}$ of (17) verifies

$$\forall t \geq 0,\ \mathbf{E}_{\bar{y}}(t) \leq \frac{C}{\left(\varphi^{-1}\left(\frac{1}{t+1}\right)\right)^{2s}} \|(\bar{y}_0, \bar{y}_1)\|^2_{X_s}. \tag{36}$$

We refer to [7, 41] for explicit examples of networks in which explicit estimates on the rate of vanishing on the weights (c_n^2) and, accordingly, of the decay rate of the energy for the dissipative system are given.

5 Further Comments and Open Problems

As we have mentioned throughout the article, there are many interesting questions (most of them are difficult) to be investigated in connection with the topics we have addressed here and some other closely related ones, in connection with PDE in networks. We mention here some of them. Of course the list is non exhaustive. We refer to [22], for instance, for a recent survey on this area.

- **Lower bounds on the weights.** As we have seen, the weights entering the observability inequalities, and, more precisely, their decay at high frequencies, play a key role when identifying the control/observation spaces and also the decay rates on the dissipative framework. It would be very interesting to analyze how the degeneracy of these weights at high frequencies depends on the properties of the network under consideration.
- **Wave equations with potentials.** We have considered here the pure wave model. What happens when the equations are perturbed by lower order terms? In the context of the equation in domains of the Euclidean space, it is well known that these lower order perturbations do not matter in the sense that they add compact

perturbations that can be get rid-of by a compactness-uniqueness argument. But the situation is different in networks because the best expected observability results are weak and require the loss of at least one derivative. This derivative is precisely the one that the zero order potentials allow gaining, but it is not enough to ensure the compactness of the perturbations. Note moreover that this happens in very special situations where the diophantine theory can be applied. But, in general, the loss of derivatives can be arbitrary. Thus, the problem of whether the observability/controllability/stabilizability properties we have proved here are preserved when one adds arbitrary bounded potentials on the various strings of the network is open.

- **Wave equations with variable coefficients.** The same problem above can be formulated for wave equations on 1-d networks with variable and sufficiently smooth coefficients (say BV-ones). Note that, even for the 1-d wave equation on an interval the BV-regularity is the minimal one required for the observability property to holds [6].
- **Semilinear wave equations.** Similar issues arise for semilinear wave equations. In the case of domains of the Euclidean space, sharp estimates on the cost of controlling wave equations with potentials, together with fixed point techniques, allow proving the controllability of semilinear wave equations, under suitable growth conditions on the nonlinearity at infinity. This is an open issue in the context of networks, the first difficulty being, as mentioned above, that of dealing with wave equations with potentials.
- **Transmutation.** As we have mentioned above, most of the analysis of control problems on networks has been developed for the wave equation. Then, the obtained result, using the method of transmutation based on Kannai's transform, leads to null control results for the heat equation. This can also be done establishing a continuity result on the property of null controllability between the wave and the heat equation through the damped wave equation (see [28]).

  But, in the case of the heat equation in bounded domains of the Euclidean space, the corresponding observability inequalities are often obtained applying Carleman inequalities directly to the heat model. This is still to be done in the context of the heat equation on graphs. Note however that the evidence that the expected results need to depend on the topological and number theoretical properties of the network makes this method very hard to be applied in this context. In any case the issue of applying Carleman inequalities to obtain directly observability inequalities for PDE in networks is widely open.
- **Multipliers.** The results in this paper were obtained using a fine analysis of the propagation properties of waves along the network. However, in the context of the wave equation in the Euclidean space, relevant results can be obtained much more easily by using the method of multipliers (see [31]). It would be interesting to explore if the observability results for waves on networks (other than the one guaranteeing the observability of the energy of a tree-like network when measurements are done on all but one external node) can be obtained by the method of multipliers.

- **Thermoelasticity.** In the context of PDE in domains of the Euclidean space one can combine the theory of the wave and heat equations to obtain results on the controllability of several relevant systems, including the system of thermoelasticity (see [26]).
- **Hyperbolic–Parabolic systems.** Recently, motivated by problems of fluid-structure interaction, there has been work done for models coupling a wave and a heat equation along an interface. The coupling turns out to be quite weak so that the corresponding system does not even decay uniformly exponentially [43]. Similar issues could be considered on networks where, in principle, one could choose arbitrarily the location of the heat and wave equations. In the context of the control of those systems in 1-d (a wave equation and a heat equation coupled through a point-wise interface) it is well known that the controllability properties depend on the location of the controller. In particular, the system is much more easily controllable when the control is on the external boundary of the wave domain than in the one of the heat domain (see [44]). Using the methods in [44], which combine sidewise energy estimates with known controllability results on the heat equation, and the results on the heat equation that one can derive from the results on the wave equation we have presented here by transmutation, one could prove controllability results on general networks provided: (a) All the wave components are located on external segments so that the system under consideration is the heat equation on a graph surrounded by external vibrating controlled strings; (b) the resulting heat-like configuration is controllable. But all the other situations are still to be investigated.
- **Other joint conditions.** All the results presented here refer to the Laplacian on networks defined through the so-called Kirchhoff conditions. But the systems under consideration have a physical meaning and are well-posed for other joint conditions. In particular we could assume that the external and/or internal nodes contain point masses. Very likely similar results will hold in that case but, even in the case of two strings connected by a point-mass it is well-known that the control theoretical properties change dramatically because of the presence of the mass. In particular it is well known that, in those cases, the observability/controllability spaces are asymmetric to both sides of the point-mass (see [17]). Similar asymmetry properties may be expected in the case of networks with point masses on the joints.
- **Switching control.** Recently, a theory of switching controls has been developed for PDE with various actuators or controllers. This is particularly suited for networks endowed with different controllers, located in various nodes (internal or external ones). It would be interesting to analyze systematically the possibilities of controlling networks (in particular for the heat equation in which the time-analiticity of solutions can be guaranteed) by means of switching controls (see [46]). The same can be said in the context of stabilization, in which the various feedback controllers are requested to be activated in a switching manner. At this respect, the work [11] is worth mentioning. There the authors consider a star-like network composed by M strings endowed with M feedback controllers on the exterior nodes, each of which can be deactivated by a time-dependent

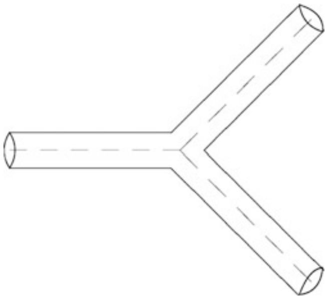


Fig. 3 A graph-like thin manifold or a $3 - d$ branching-domain

switching law. They provide conditions on the switching laws guaranteeing that the network can be stabilized asymptotically to rest.

- **Infinite networks of finite length.** It would be interesting to investigate the possible extension of the results of this paper to networks involving an infinite number of strings, but of finite total length.
- **Optimal placement of controllers.** We have discussed here the problem of observation, control or stabilization from a given external vertex. But it would be of interest to discuss the problem of the choice of the optimal placement of the controller. This is a widely open subject. We refer to [20] for some of the few existing results in context of the string equation on a segment.
- **Graph-like thin manifolds.** In these notes we have considered the control and stabilization of the wave equation on 1-d networks. We have also discussed similar issues for other models as the heat or Schrödinger equations. It is very natural to analyze the same issues in thin $2 - d$ domains obtained by simply adding a thickness of size ε to the network on the perpendicular direction to each string (Fig. 3).

  The control of PDE's in thin cylinders is reasonably understood. In the case of the wave equation, due to the existence of trapped rays in the perpendicular directions, the wave and plate process can not be controlled from the lateral boundary and the filtering of the high frequency trapped rays is needed to get uniform controllability results [14]. To the contrary, in the case of the heat equation, the intrinsic strongly dissipative effect damps out the high frequency components that the added dimension generates, and the limit of null controls in thin domains is a null control in the limit cross section (see [8, 45]).

  It would be natural and interesting to analyze similar questions in the context of "thick networks" when the thickness tends to zero. The subject will however be more complex than in domains of the Euclidean space since, as the results concerning 1-d networks show (see [7]), the results one has to expect when passing to the limit, necessarily, will depend on the number theoretical properties of the lengths of the edges of the network.

  We refer to [9] for recent results on the behavior of the spectrum of the Laplacian under this singular perturbation and to [12] for a recent survey on the subject.

- **Numerics.** In recent years the problem of numerical approximation of control problems, especially for waves, has been the object of intensive research (see [49]). But very little is known in the context of networks. The abstract results in [40] can be applied in this context and we can obtain controllability results for time-discrete wave equations on networks, provided the high frequency components are appropriately filtered out. But the analysis of space-discretizations is a widely open subject.
- **Strichartz inequalities.** There are other interesting features of PDE on domains of the Euclidean space that are badly understood in the context of networks. That is for instance the case of the dispersive or Strichartz estimates for the Schrödinger equation. This issue is still to be investigated in a systematic manner in the context of networks. We refer to [19] for the first results in this direction in the case of some particular tree like infinite networks.

  Note also that these dispersive estimates play a key role when analyzing the solvability of the corresponding nonlinear problems and in their numerical approximation [19].
- **Inverse problems.** Inverse problems for waves on networks are intimately related to the control problems we have considered in this paper. The issue consists roughly on determining the topological and geometric properties of the network through measurements done on the exterior vertices. We refer to the recent paper [2] for the analysis of tree-like networks through the so-called boundary-control-approach developed in [3] and to the references in [22]. These kind of problems are widely open in the case of networks containing circuits.

Acknowledgements The author has been partially supported by Grant MTM2008-03541 of the MICINN, Spain, the ERC Advanced Grant FP7-246775 NUMERIWAVES, the ESF Research Networking Program OPTPDE and the Grant PI2010-04 of the Basque Government.

References

1. K. Ammari, M. Tucsnak, Stabilization of second order evolution equations by a class of unbounded feedbacks. ESAIM Control Optim. Calc. Var. **6**, 361–386 (2001)
2. S. Avdonin, G. Leugering, V. Mikhaylov, On an inverse poblem for tree-like networks of elastic strings. ZAMM J. Appl. Math. Mech. **90**(2), 136–150 (2010)
3. M. I. Belishev, Boundary spectral inverse problem on a class of graphs (trees) by the BC method, Inverse Probl. **20**, 647–672 (2004)
4. C. Bardos, G. Lebeau, J. Rauch, Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary. SIAM J. Control Optim. **30**, 1024–1065 (1992)
5. P. Bégout, F. Soria, An interpolation inequality and its application to the stabilization of damped equations. J. Differ. Equ. **240**(2), 324–356 (2007)
6. G. Castro, E. Zuazua, Concentration and lack of observability of waves in highly heterogeneous media. Arch. Ration. Mech. Anal. **164**(1), 39–72 (2001)
7. R. Dáger, E. Zuazua, in *Wave Propagation, Observation and Control in* 1*-d Flexible Multi-structures*. Mathématiques & Applications (Springer, Berlin, 2006)
8. L. de Teresa, E. Zuazua, Null controllability of linear and semilinear heat equations in thin domains. Asymptot. Anal. **24**, 295–317 (2000)

9. P. Exner, O. Post, Convergence of graph-like thin manifolds. J. Geom. Phys. **54**(1), 77–115 (2005)
10. J. von Below, F. Ali Mehmeti, S. Nicaise (eds.), in *Partial Differential Equations on Multistructures*. Lecture Notes in Pure and Applied Mathematics, vol. 219 (Marcel Dekker, New York, 2001)
11. M. Gugat, M. Sigalotti, Star-shaped string networks: Switching boundary feedback stabilization. Networks Heterogeneous Media, **5**(2), 299–314 (2010)
12. D. Grieser, Spectra of graph neighborhoods and scattering. Proc. London Math. Soc. **97**(3), 718–752 (2008)
13. C. Fabre, J.P. Puel, Pointwise controllability as limit of internal controllability for the wave equation in one space dimension. Portugal. Math. **51**, 335–350 (1994)
14. I. Figueiredo, E. Zuazua, Exact controllability and asymptotic limit of thin plates. Asymptot. Anal. **12**, 213–252 (1996)
15. J-M. Coron, G. Bastin, B. D'Andréa-Novel, Lyapunov stability analysis of networks of scalar conservation laws. Netw. Heterogeneous Media **2**(4), 749–757 (2007)
16. J-M. Coron, G. Bastin, B. D'Andréa-Novel, On lyapunov stability of linearised saint-venant equations for a sloping channel. Netw. Heterogeneous Media **4**(2), 177–187 (2009)
17. S. Hansen, E. Zuazua, Exact controllability and stabilization of a vibrating string with an interior point mass. SIAM J. Control Optim. **33**(5), 1357–1391 (1995)
18. A. Haraux, S. Jaffard, Pointwise and spectral control of plate vibrations. Rev. Matemática Iberoamericana **7:1**, 1–24 (1991)
19. L. Ignat, *Strichartz estimates for the schrödinger equation on a tree and applications*, SIAM J. Math. Anal. Appl. **42**(5), 2041–2057 (2010)
20. A. Henrot, K. Ammari, M. Tucsnak, Asymptotic behaviour of the solutions and optimal location of the actuator for the pointwise stabilization of a string. Asymptot. Anal. **28**(3–4), 215–240 (2001)
21. H. Koch, E. Zuazua, in *A Hybrid System of PDE's Arising in Multi-structure Interaction: Coupling of Wave Equations in n and $n - 1$ Space Dimensions*. Recent Trends in Partial Differential Equations. Contemporary Mathematics, vol. 409 (American Mathematical Society, Providence, 2006), pp. 55–77
22. P. Kuchment, in *Quantum Graphs: An Introduction and a Brief Survey*. Analysis on Graphs and Its Applications. Proceedings of the Symposium on Pure Mathematics, vol. 77 (American Mathematical Society, Providence, 2008), pp. 291–312
23. J.E. Lagnese, in *Recent Progress and Open Problems in Control of Multi-link Elastic Structures*. Contemporary Mathematics, vol. 209 (American Mathematical Society, Providence, 1997), pp. 161–175
24. J.E. Lagnese, G. Leugering, *Domain Decomposition Methods in Optimal Control of Partial Differential Equations*, vol. 148 (Birkhäuser, Basel, 2004)
25. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, in *Modelling, Analysis and Control of Multi-link Flexible Structures*. Systems and Control: Foundations and Applications (Birkhäuser, Basel, 1994)
26. G. Lebeau, E. Zuazua, Decay rates for the linear system of three-dimensional system of thermoelasticity. Arch. Ration. Mech. Anal. **148**, 179–231 (1999)
27. V. Lescarret, E. Zuazua, *Numerical schemes for waves in multi-dimensional media: convergence in asymmetric spaces*, preprint (2009)
28. A. Lopez, X. Zhang E. Zuazua, Null controllability of the heat equation as a singular limit of the exact controllability of dissipative wave equations. J. Math. Pures et Appl., **79**, 741–809 (2000)
29. J.-L. Lions, Contrôlabilité exacte de systèmes distribués. C. R. Acad. Sci. Paris Sér I **302**, 471–475 (1986)
30. J.-L. Lions, *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, vol. I (Masson, Paris, 1988)
31. J.L. Lions, Exact controllability, stabilizability and perturbations for distributed systems. SIAM Rev. **30**, 1–68 (1988)

32. G. Lumer, in *Connecting of Local Operators and Evolution Equations on Networks*. Lecture Notes in Mathematics, vol. 787 (Springer, Berlin, 1980), pp. 219–234
33. F. Ali Mehmeti, A characterization of a generalized C^{∞}-notion on nets. Integral Equ. Operator Theory **9**(6), 753–766 (1986)
34. S. Micu, E. Zuazua, An introduction to the controllability of linear PDE. in *Quelques questions de théorie du contrôle*, ed. by T. Sari. Collection Travaux en Cours (Hermann, Paris, 2005), pp. 69–157
35. L. Miller, Geometric bounds on the growth rate of null-controllability cost for the heat equation in small time. J. Differ. Equ. **204**, 202–226 (2004)
36. S. Nicaise, Spectre des réseaux topologiques finis. Bull. Sci. Math. **111**, 401–413 (1987)
37. S. Nicaise, in *Polygonal Interface Problems*. Methoden und Verfahren Math. Physiks, Frankfurt/M., Berlin, Bern, New York, Paris, Wien, vol. 39 (Peter Lang, 1993)
38. Yu.V. Pokornyi, O.M. Penkin, E.N. Provotorova, in *On a Vectorial Boundary Value Problem*. Boundary Value Problems. Interuniv. Collect. Sci. Works, Perm' (1983), pp. 64–70
39. D.L. Russell, A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. Stud. Appl. Math. **52**, 189–221 (1973)
40. Ch. Zheng, S. Ervedoza, E. Zuazua, On the observability of time-discrete conservative linear systems. J. Funct. Anal. **254**(12), 3037–3078 (2008)
41. J. Valein, E. Zuazua, Stabilization of the wave equation on 1-d networks. SIAM J. Control Optim. **48**(4), 2771–2797 (2009)
42. J. von Below, A characteristic equation associated to an eigenvalue problem on c^2-networks. Linear Algebra Appl. **71**, 309–325 (1985)
43. X. Zhang, E. Zuazua, Polynomial decay and control for a 1-d model of fluid-structure interaction. C. R. Acad. Sci. Paris, Sér. I **336**, 745–750 (2003)
44. E. Zuazua, Null control of a 1-d model of mixed hyperbolic-parabolic type. in *Optimal Control and Partial Differential Equations*, ed. by J.L. Menaldi et al. (IOS Press, Amsterdam), pp. 198–210
45. E. Zuazua, *Null controllability of the heat equation in thin domains*, Equations aux dérivées partielles et applications. Articles dédiés à Jacques-Louis Lions, Proc. Sympos. Pure Math. vol. 77, Gauthier-Villars, pp. 787–801 (1998)
46. E. Zuazua, Switching control. J. Eur. Math. Soc., 2011 **13**, 85–117. doi: 10.4171/JEMS/245
47. E. Zuazua, in *Some Problems and Results on the Controllability of Partial Differential Equations*. Progress in Mathematics, vol. 169 (Birkhaüser, Basel, 1998), pp. 276–311
48. E. Zuazua, Controllability of partial differential equations and its semi-discrete approximations. Discrete Contin. Dyn. Syst. **8**, 469–513 (2002)
49. E. Zuazua, Propagation, observation, and control of waves approximated by finite difference methods. SIAM Rev. **47**(2), 197–243 (2005)
50. E. Zuazua, *Controllability and observability of partial differential equations: Some results and open problems*. in *Handbook of Differential Equations: Evolutionary Equations*, vol. 3, ed. by C.M. Dafermos, E. Feireisl (Elsevier Science, 2006), pp. 527–621

List of Participants

1. Atalla Shadi
   Politecnico Torino, Italy
2. Barbagallo Annamaria
   University of Catania, Italy
3. Bayen Alexandre
   UC Berkeley, USA
4. Bigolin Francesco
   University of Trento, Italy
5. Blandin Sebastien
   UC Berkeley, USA
6. Bonaccorsi Stefano
   University of Trento, Italy
7. Bretti Gabriella
   University of La Sapienza, Roma, Italy
8. Caravenna Laura
   SISSA, Italy
9. Cavalletti Fabio
   SISSA, Italy
10. Cazacu Mihai
    BCAM, Spain
11. Coclite Giuseppe
    University of Bari, Italy
12. Colombo Rinaldo M.
    University of Brescia, Italy
13. Cristiani Emiliano
    University of Salerno, Italy
14. Dinler Ali
    Istanbul Technical University, Turkey
15. Donadello Carlotta
    Northwestern University, USA

L. Ambrosio et al., *Modelling and Optimisation of Flows on Networks*, Lecture Notes
in Mathematics 2062, DOI 10.1007/978-3-642-32160-3,

16. Facchi Giancarlo
    Penn State University, USA
17. Fang Guoliang
    Penn State University, USA
18. Festa Adriano
    University of La Sapienza, Roma, Italy
19. Frasca Paolo,
    IAC-CNR, Italy
20. Garavello Mauro
    University of Piemonte Orientale, Italy
21. Goatin Paola
    University of Sud Toulon Var, France
22. Goettlich Simone
    TU Kaiserslautern, Germany
23. Guerra Graziano
    University of Milano-Bicocca, Italy
24. Khabbaz Muhannad
    University of Salzburg, Austria
25. Khe Alexander
    Lavrentyev Institute of Hydrodynamics, Russia
26. Ligabò Marilena
    University of Bari, Italy
27. Losada Chaya
    Kent University, UK
28. Maldarella Dario
    University of Ferrara, Italy
29. Marcellini Francesca
    University of Milano Bicocca, Italy
30. Marigonda Antonio
    University of Verona, Italy
31. Marino Francesco
    University of Roma 2, Italy
32. Martin Stephan
    TU Kaiserslautern, Germany
33. Maurizi Amelio
    University of Aquila, Italy
34. Mercier Magali
    University of Lyon1, France
35. Mishkovski Igor
    Politecnico di Torino, Italy
36. Monti Francesca
    University of Milano-Bicocca, Italy
37. Najjar Atieh
    Salzburg University, Austria

38. Popova Evdokia
    Saint-Petersburg State University, Russia
39. Priuli Fabio
    SISSA, Italy
40. Roberto Mecca
    University of Roma-Sapienza, Italy
41. Rosini Massimiliano
    Institute of Mathematics Polish Academy, Poland
42. Sepe Alice
    University of Bari, Italy
43. Spinolo Laura
    Centro De Giorgi, Pisa, Italy
44. Spirito Stefano
    University of Aquila, Italy
45. Thakur Mohit
    Munich University of Technology, Germany
46. Tien Khai Nguyen
    University of Padova, Italy
47. Torres Ramiro
    Escuela Politecnica Nacional, Spain
48. Tosin Andrea
    Politecnico di Torino, Italy
49. Work Daniel
    UC Berkeley, USA
50. Zhang Dongmei
    Penn State University, USA
51. Zhang Tianyou
    Penn State University, USA
52. Ziegler Ute
    TU Kaiserslautern, Germany

LECTURE NOTES IN MATHEMATICS Springer

Editorial Policy (for Multi-Author Publications: Summer Schools / Intensive Courses)

1. Lecture Notes aim to report new developments in all areas of mathematics and their applications - quickly, informally and at a high level. Mathematical texts analysing new developments in modelling and numerical simulation are welcome. Manuscripts should be reasonably selfcontained and rounded off. Thus they may, and often will, present not only results of the author but also related work by other people. They should provide sufficient motivation, examples and applications. There should also be an introduction making the text comprehensible to a wider audience. This clearly distinguishes Lecture Notes from journal articles or technical reports which normally are very concise. Articles intended for a journal but too long to be accepted by most journals, usually do not have this "lecture notes" character.

2. In general SUMMER SCHOOLS and other similar INTENSIVE COURSES are held to present mathematical topics that are close to the frontiers of recent research to an audience at the beginning or intermediate graduate level, who may want to continue with this area of work, for a thesis or later. This makes demands on the didactic aspects of the presentation. Because the subjects of such schools are advanced, there often exists no textbook, and so ideally, the publication resulting from such a school could be a first approximation to such a textbook. Usually several authors are involved in the writing, so it is not always simple to obtain a unified approach to the presentation.
   For prospective publication in LNM, the resulting manuscript should not be just a collection of course notes, each of which has been developed by an individual author with little or no coordination with the others, and with little or no common concept. The subject matter should dictate the structure of the book, and the authorship of each part or chapter should take secondary importance. Of course the choice of authors is crucial to the quality of the material at the school and in the book, and the intention here is not to belittle their impact, but simply to say that the book should be planned to be written by these authors jointly, and not just assembled as a result of what these authors happen to submit.
   This represents considerable preparatory work (as it is imperative to ensure that the authors know these criteria before they invest work on a manuscript), and also considerable editing work afterwards, to get the book into final shape. Still it is the form that holds the most promise of a successful book that will be used by its intended audience, rather than yet another volume of proceedings for the library shelf.

3. Manuscripts should be submitted either online at www.editorialmanager.com/lnm/ to Springer's mathematics editorial, or to one of the series editors. Volume editors are expected to arrange for the refereeing, to the usual scientific standards, of the individual contributions. If the resulting reports can be forwarded to us (series editors or Springer) this is very helpful. If no reports are forwarded or if other questions remain unclear in respect of homogeneity etc, the series editors may wish to consult external referees for an overall evaluation of the volume. A final decision to publish can be made only on the basis of the complete manuscript; however a preliminary decision can be based on a pre-final or incomplete manuscript. The strict minimum amount of material that will be considered should include a detailed outline describing the planned contents of each chapter.
   Volume editors and authors should be aware that incomplete or insufficiently close to final manuscripts almost always result in longer evaluation times. They should also be aware that parallel submission of their manuscript to another publisher while under consideration for LNM will in general lead to immediate rejection.

4. Manuscripts should in general be submitted in English. Final manuscripts should contain at least 100 pages of mathematical text and should always include

   – a general table of contents;
   – an informative introduction, with adequate motivation and perhaps some historical remarks: it should be accessible to a reader not intimately familiar with the topic treated;
   – a global subject index: as a rule this is genuinely helpful for the reader.

   Lecture Notes volumes are, as a rule, printed digitally from the authors' files. We strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. To ensure best results, authors are asked to use the LaTeX2e style files available from Springer's web-server at
   ftp://ftp.springer.de/pub/tex/latex/svmonot1/ (for monographs) and
   ftp://ftp.springer.de/pub/tex/latex/svmultt1/ (for summer schools/tutorials).
   Additional technical instructions, if necessary, are available on request from:
   lnm@springer.com.

5. Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online. After acceptance of the manuscript authors will be asked to prepare the final LaTeX source files and also the corresponding dvi-, pdf- or zipped ps-file. The LaTeX source files are essential for producing the full-text online version of the book. For the existing online volumes of LNM see:
   http://www.springerlink.com/openurl.asp?genre=journal&issn=0075-8434.
   The actual production of a Lecture Notes volume takes approximately 12 weeks.

6. Volume editors receive a total of 50 free copies of their volume to be shared with the authors, but no royalties. They and the authors are entitled to a discount of 33.3 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

7. Commitment to publish is made by letter of intent rather than by signing a formal contract. Springer-Verlag secures the copyright for each volume. Authors are free to reuse material contained in their LNM volumes in later publications: a brief written (or e-mail) request for formal permission is sufficient.

Addresses:
Professor J.-M. Morel, CMLA,
École Normale Supérieure de Cachan,
61 Avenue du Président Wilson, 94235 Cachan Cedex, France
E-mail: morel@cmla.ens-cachan.fr

Professor B. Teissier, Institut Mathématique de Jussieu,
UMR 7586 du CNRS, Équipe "Géométrie et Dynamique",
175 rue du Chevaleret,
75013 Paris, France
E-mail: teissier@math.jussieu.fr

For the "Mathematical Biosciences Subseries" of LNM:

Professor P. K. Maini, Center for Mathematical Biology,
Mathematical Institute, 24-29 St Giles,
Oxford OX1 3LP, UK
E-mail : maini@maths.ox.ac.uk

Springer, Mathematics Editorial I,
Tiergartenstr. 17,
69121 Heidelberg, Germany,
Tel.: +49 (6221) 4876-8259
Fax: +49 (6221) 4876-8259
E-mail: lnm@springer.com