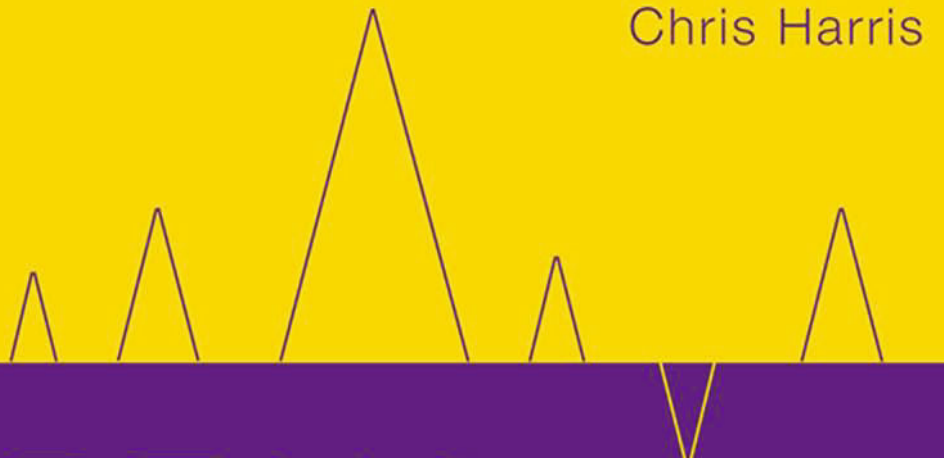


Chris Harris



PEAK
LOAD AND
CAPACITY
PRICING

Theory and Practice in Electricity



PEAK LOAD AND CAPACITY PRICING

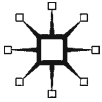
This page intentionally left blank

PEAK LOAD AND CAPACITY
PRICING

Theory and Practice in Electricity

Chris Harris

palgrave
macmillan



PEAK LOAD AND CAPACITY PRICING

Copyright © Chris Harris, 2015.

Softcover reprint of the hardcover 1st edition 2015 978-1-137-38481-2

All rights reserved.

First published in 2015 by

PALGRAVE MACMILLAN®

in the United States—a division of St. Martin's Press LLC,
175 Fifth Avenue, New York, NY 10010.

Where this book is distributed in the UK, Europe and the rest of the world,
this is by Palgrave Macmillan, a division of Macmillan Publishers Limited,
registered in England, company number 785998, of Houndmills,
Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan is the global academic imprint of the above companies
and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States,
the United Kingdom, Europe and other countries.

ISBN 978-1-349-48108-8 ISBN 978-1-137-37092-1 (eBook)

DOI 10.1057/9781137370921

Library of Congress Cataloging-in-Publication Data is available from the
Library of Congress.

A catalogue record of the book is available from the British Library.

Design by Newgen Knowledge Works (P) Ltd., Chennai, India.

First edition: April 2015

10 9 8 7 6 5 4 3 2 1

To Mum—with all my love

This page intentionally left blank

CONTENTS

<i>List of Figures</i>	ix
<i>Foreword</i>	xv
<i>Acknowledgments</i>	xvii
1 Introduction	1
2 The Modeling Framework	5
2.1 Nomenclature	5
2.2 Basic Modeling of Consumption	5
2.3 Basic Model of Physical/Production	12
2.4 Other Aspects of Modeling	14
3 The Framework and Development of Peak Load Pricing	23
3.1 Basic Peak Load Pricing Theory	23
3.2 Early Days of Tariff Evolution	28
3.3 Postwar Theory and Application of Capacity Charging—The Drèze Approach	33
3.4 Elastic Demand for Capacity—The Steiner Framework	47
3.5 Efficiency of Rationing of Demand	53
3.6 Pricing under Stochastic Demand—The Brown and Johnson Framework	59
3.7 Modeling under Variable Rationing Efficiency— The Visscher Framework	66
3.8 Demand for Capacity with Stochastic Elastic Demand—the Carlton Framework	81
3.9 Optimal Pricing, Capacity and the Technology Frontier—Crew and Kleindorfer	90
3.10 Further Development of the Price Vector— from Dansby	99
3.11 Stochastic Variation in both Production and Demand—the Chao Framework	105
3.12 Discussion of the Pricing Analyses	120

4	Relaxing the Hard Capacity Constraint	123
4.1	Introduction	123
4.2	The Hirshleifer Framework	123
4.3	Optimizing with Decreasing Returns to Scale— The Panzar Framework	126
4.4	Concluding Discussion of Soft Constraints	140
5	Modeling Capacity Using Derivatives	141
5.1	Power Markets	142
5.2	Single Period—Modeling Using Options of “European” Type	150
5.3	Modeling System Operator Option Procurement in the One-Period Setting	165
5.4	Modeling More Complex Options in the One-Period Setting	167
5.5	Many Periods—Modeling Using Options of “American” Type	169
5.6	Modeling System Operator Option Procurement in the Unrestricted Many-Period Setting	171
5.7	Modeling Many Periods under Constraint Using Options of “Swing” Type	172
5.8	Modeling System Operator Option Procurement in the Restricted Many-Period Setting	173
5.9	Modeling Real Assets	174
5.10	Modeling Prices and Price Dynamics	177
6	Capacity Mechanisms	183
6.1	Types of Capacity Mechanism	183
6.2	Development of the Key Variables in the ICAP Model	184
6.3	Development of ICAP toward the Use of Strike Prices	190
6.4	Division of the Market to Regimes	198
6.5	General Development of Capacity Mechanisms toward Energy-Only Markets	198
7	The Power Complex	201
7.1	The Demand Side	201
7.2	Transmission and Interconnection	205
7.3	Peak Load Pricing in Distribution Charging	235
7.4	Commercial Arrangements on Lost Load	235
8	Final Comments	239
	<i>Notes</i>	241
	<i>References</i>	247
	<i>Index</i>	251

FIGURES

2.1	Four useful demand functions	6
2.2	Some utility functions	7
2.3	Consumer's surplus	8
2.4	The deadweight loss of inefficient volume delivered	9
2.5	Demand shocks	10
2.6	The load duration function	11
2.7	Construction of the equilibrium price duration function	11
2.8	The standard cost model for power stations	12
2.9	(a) The available unit stack on the system; (b) The stack modeled on the technology frontier	13
2.10	Soft constraints pictured as asymptotic to hard constraints	14
2.11	Standard representation of the cost of risk	16
2.12	Cost frontier duality	21
3.1	The simplest load variation for analysing peak load pricing	24
3.2	Efficient evolution of the installed technology stack cost frontier	26
3.3	Fixed and variable tariff structures	30
3.4	Methods of measurement of peak for fixed charge	30
3.5	Tariff structures	33
3.6	Build volume in the Drèze framework	35
3.7	Probability of demand exceeding capacity	45
3.8	Approximate linearity of probability of lost load in relation to standard deviation	46
3.9	Nonlinearity of probability of lost load in relation to standard deviation	46
3.10	Periodic demand function as described by Steiner	49
3.11	Demonstration of "shifting peaks"	49
3.12	Steiner optimal pricing in the "shifting peaks" case	50
3.13	Exact recovery of total costs	50

3.14	Optimal pricing for a three-period example with shifting peaks	51
3.15	Optimal pricing in the Williamson framework with different durations of peak and off-peak	52
3.16	Pricing and capacity for two different levels of fixed cost	52
3.17	Comparison of random rationing to efficient rationing	56
3.18	Most inefficient rationing	56
3.19	Summary of rationing efficiencies	57
3.20	Purely public good with capacity less than demand	58
3.21	Depicting of rationing efficiency between consumers and by consumers	58
3.22	The inverse demand function and consumers' surplus W	60
3.23	Gross consumer surplus when capacity is sufficient, showing construction of the integral	62
3.24	Loss of consumer surplus after a shocked demand exceeding capacity, relative to the surplus at the ideal capacity level	63
3.25	Geometric representation of the Visscher framework for a linear demand function	67
3.26	Positive demand shock in the Visscher framework.	69
3.27	Welfare equivalence of demand shock and willingness to pay shock	72
3.28	The change of rationing inefficiency with amount of lost load, for random rationing	74
3.29	The impact of rationing efficiency on optimum build	78
3.30	Event ordering in the Carlton analysis	82
3.31	The demand function for different outcomes of multiplicative stochastic variable μ	82
3.32	Rationing when demand exceeds capacity	83
3.33	Producer marginal surplus and consumer surplus after payments in the Carlton framework	84
3.34	Random rationing in the Carlton analysis	85
3.35	Producer and consumer surpluses for different stochastic outcomes	87
3.36	Upward shocks only shown for three demand functions	88
3.37	Lack of symmetry in inefficiency, leading to a dependence on optimum build on the form of the utility/demand function and standard deviation	88
3.38	Producer and consumers' surplus as in figure 3.37	89
3.39	Stochastic demand as depicted in Crew and Kleindorfer	92
3.40	Plant envelope in the Crew and Kleindorfer analysis	93

3.41	Marginal generation and demand curves in six of the n periods	95
3.42	The loading of unit l in period j	101
3.43	Demand functions at different times within the period referenced against the final load duration curve	101
3.44	Depiction of reduction in run time of unit l in period j , resulting from a price increase ΔP	103
3.45	The slope ϕ of the load duration curve at the l th unit.	103
3.46	Representation of the Chao framework	105
3.47	Stochastic cycle availability vector for unit i	107
3.48	Unit operation for (a) Plant capacity sufficient; (b) Plant capacity insufficient	108
3.49	Effect of variation in the availability of a unit	108
3.50	Dependence on the sensitivity of energy delivered up to unit i to the availability of unit j (a) $j < i$, $j > i$	110
3.51	Loss of surplus in relation to loss of load.	115
3.52	Correlation between marginal and total demand depends on the demand function shape and stochastic nature	117
3.53	Effective of shock on the correlation of total and marginal demand	118
4.1	Vertical addition of demand curves for the constrained flat marginal cost and the Ricardian function	124
4.2	Hirshleifer analysis recast with a feasible production set	126
4.3	A physical representation consistent with the Panzar framework, showing unit i in period t	127
4.4	Panzar framework with and without link between units, capacities, and load	129
4.5	Unit cost envelopes	136
4.6	Optimum aggregate welfare for different percentage mixes and total size of two units	137
4.7	Allocation of unit output to deliver $Q_1 + Q_2 + Q_3$ MW in period t	137
5.1	Effective of probability distribution of price from a price cap	148
5.2	Influencers of the choice of strike offered in balancing	153
5.3	Simple variations to the option smile	154
5.4	Expectation bias of forward prices from cost of risk	156
5.5	Situation faced by producer according to different strike prices and market price outturn	157
5.6	Demonstrating an effective forward price above the cap	159
5.7	Deficiency charge in relation to its economic value	164

5.8	Probability distribution of the effective price	164
5.9	Ascending and descending clock auction to select options	165
5.10	Non-homothetic changes to the option premium-strike vector	166
5.11	(a) Continuous fair value curve for option premiums and strikes (b) Mapping of transformed premiums to strike prices, showing actual offers relative to the transformed fair value line	167
5.12	Convergence of units onto the single strike-premium vector	167
5.13	Trading book structure for options struck at the fuel price	169
5.14	Visualization of unit valuation. Area is proportional to value of service dimension	175
5.15	Visualization of the value mix of two different unit types	176
5.16	Money and power flows on plant failure	176
5.17	Hedge and delta relationships between horizon and strike	179
5.18	Simplification of supplier hedge position	179
6.1	Development of the demand for capacity function	188
6.2	The regulatory option tender by the single buyer	193
6.3	Single buyer of options showing construction of the peak price given a unit selection	193
6.4	Development of option capacity from fully administered to fully market	195
6.5	Pool price transformation in scarcity	196
6.6	Retail supplier pool cost under normal and scarcity conditions	197
6.7	Depiction of peak energy rental and scarcity rental in an options format	198
6.8	Development of the ICAP model toward an energy only model	199
7.1	Principal component shocks to the load duration function	201
7.2	Change from old world to new world production-demand paradigm	202
7.3	Composite frontier from production and demand-side management	204
7.4	Depiction of the system described	206
7.5	Cumulative and density probability functions of demand D	207
7.6	The stages of the game	207

7.7	(a) Regulator sets capacity requirement for 0 percent probability of lost load (b) Regulator sets capacity requirement for finite probability of lost load	208
7.8	Contractual energy flows for capacity sale $\Theta < K - \beta$ for the different probability domains of demand	209
7.9	Contractual energy flows for capacity sale $K - \beta < \Theta < K$	209
7.10	Contractual energy flows for capacity sale $K < \Theta < K + \beta$	210
7.11	The critical cap price for the producer to offer capacity	213
7.12	Visualization of the gas network as equivalent to a power network	222
7.13	Summary of the system as defined by Cremer, Gasmi, and Laffont cast as electricity rather than gas	223
7.14	Cost and revenue functions	223
7.15	Network cost structure envisaged by Cremer, Gasmi, and Laffont	224
7.16	Simplest formulation for a three-node constrained electrical system.	226
7.17	Cost structure for gradually increasing demand	227
7.18	(a) Generator cost and line rents as network load increases (b) Price at the demand node	229
7.19	Application of the principle of superposition to a simple network $A + B = C$	232

Note: All figures have been created by the author unless specified otherwise.

This page intentionally left blank

FOREWORD

Liberalization and restructuring of electricity industry has brought peak-load and capacity pricing to the forefront for analysis of regulatory policy and market design. In this regard, this volume brings a risk management perspective to the discussion of capacity mechanisms in electricity markets.

The perspective of risk management is needed for restructuring of the electricity industry, and liberalization of wholesale and retail markets for power. Industry restructuring introduced a new market structure in which power generators and utilities and other retailers in large regional markets managed by independent system operators (ISOs) and regional transmission operators (RTOs). It also brought a new allocation of risk bearing, in which generators initially bear investment risks, and utilities and their customers bear price risks—but then long-term contracts and financial hedges are supposed to mitigate these risks. This new scheme works well for some large industrial and commercial customers and the independent power producers with whom they contract, but beyond this, contracting has been an imperfect solution.

Resource adequacy requirements are justified by the fundamental inability of competitive markets to provide incentives for provision of sufficient capacity to ensure security of supply in liberalized wholesale markets. Many different capacity mechanisms have been implemented in various systems to fulfill the requirements. They differ mainly in whether they impose capacity obligations or subsidize investments. In both cases, the factor most critical for efficiency is a feedback mechanism that enables adjustment to changing circumstances, thus preventing under- and over-capacity. There is now some evidence of the need to impose requirements for investments in transmission and generation resources in order to provide adequate reserves, especially when bond ratings of many generators and some utilities under financial distress have deteriorated and their cost of capital has risen.

Previous regulatory compacts implemented an allocation of risk-bearing under which customers eventually bore all risks, but only

gradually, as retail rates, were adjusted slowly to recover the amortized cost of service. While restructuring has had obvious successes, including the development of regional markets and signs of improved operating efficiency, deficiencies are also evident in the form of the high costs of capital, the prevalence of financial distress among utilities, boom-and-bust cycles of investment, insufficient capacity to ensure adequate reserve margins, underdeveloped retail markets, and inadequate service differentiation. The unfilled promise of restructuring and liberalization reflects inadequate attention to the physical and financial aspects of risk management and to the consequences of restructuring for risk management by generators, utilities, and core customers. A risk management perspective that combines engineering and economic considerations is essential to resolve the issues of market design. Capacity mechanisms are needed to address the integrated resource planning mandates by vertically integrated utilities. In addition, an economic approach that draws on the vast literature of financial risk management is needed for efficient allocation of risks among generators, utilities and other retailers, and customers to lower the costs of capital, sustain investments to meet continued growth in demand, and encourage efficient demand-side usage.

It is a pleasure to commend Dr. Harris for producing a volume that provides a very valuable contribution to the theory of peak-load and capacity pricing with a risk management perspective for capacity mechanisms. This book should be read by anyone who cares about the future development of the electricity market structure.

DR HUNG-PO CHAO
Director, Market Strategy and Analysis at
Independent System Operator, New England.

ACKNOWLEDGMENTS

I would like to thank Dr. Adrian Winnett for his support for my work.

INTRODUCTION

In this book we consider the development of the theory and practice for the pricing of goods delivered by assets that do not run continuously, whether for demand reasons (periodic and/or stochastic) or production reasons (planned periodic cycling and/or technical availability). The focus is on electricity, and much of the analysis can be applied to other goods. The pricing of electricity at the peak is closely bound with the pricing of capacity.

Our practical purpose is to inform

1. efficient production and consumption choices for both private and state actors
2. efficient construction of markets, market arrangements, and capacity obligations, using market disciplines, especially those of traded derivatives
3. efficient construction of policy that explicitly recognizes the requirement to recover fixed costs and the moral hazards on the part of market actors, regulators, and governments

At some risk of oversummary, peak load pricing emphasizes long run equilibrium through the recovery of fixed costs through prices, and marginal cost pricing emphasizes short run efficiency and minimization of deadweight losses in welfare and in doing so ignores some fixed costs in price formulation.

Variable cost pricing remains the majority view in regulation, the media, politics, and commentators, and is also prevalent in the academic literature. We will show how the two methods can be reconciled in equilibrium conditions.

Consideration of peak load pricing has a long history in the theoretical literature. The recent development of thought could be viewed as having four key phases:

1. Pre-1950s: A long history of the structure of costs through moral philosophy and political economics, with key moments such as

Adam Smith's *Wealth of Nations* in 1776, the physiocrats, Dupuit and the French "econo-engineers"¹ of the mid-nineteenth century, and the growth of marginal microeconomics in the 1890s. The peak load versus variable cost debate still turns on the relative importance of sustainable equilibrium in classical economics and the marginal (and market) price in neoclassical economics.

2. 1950s–early 1970s: The theoretical foundations were developed for predominantly state-run electricity systems, commonly with tax subsidies in the fuel or other parts of the value chain. The broad academic consensus, resting on the heritage of marginal/neoclassical economics, was that most prices should be based on marginal variable costs rather than peak load pricing in which the fixed costs are loaded onto prices. It was during this period that the economic science² of peak load pricing began as an optimization problem.
3. Late 1970s–2000s: The development of liberalization, with increasing private ownership and operation and the advent of markets and market disciplines. Discrete pricing of capacity was introduced and developed, and the economics of peak load pricing matured at the beginning of this period, and essentially completed, within the standard paradigm of separable fixed and variable costs. At the end of the period, the science of peak load pricing began to borrow from the science of market derivatives.
4. 2000s and forward: Following the California electricity crisis in 2000/2001, collapse of Enron in 2001, and the ensuing demise of similar firms, and then the banking crisis of 2008, a return to state intervention and planning, this time without state ownership. The tension between consumer protection and competition in the areas of essential goods grew, with the result of price caps and protection of the socialization of capacity and public goods. In this period, the approach to capacity came under the influence of the evolving role of geopolitics, the growth of behavioural economics, experience of application of capacity mechanisms, and the integration of power systems and markets on a continental scale. In this period, there has been little debate on what was called the marginal cost controversy between the peak load and variable cost approaches.

The literature of the 2000s and beyond rests largely on the consensus for short run marginal cost pricing, which since we split into fixed and variable cost, we call variable cost pricing. Here we study the original texts, mainly in the middle period of the late 1970s to the 2000s, in order to examine the sensitivity of the conclusions to changes in explicit and implicit assumptions. We find that although there is an

apparent tension between the peak load and variable cost approaches, that under equilibrium conditions, a careful working through of the original papers shows that they can be reconciled.

Our approach here is essentially welfarist, resting on the theories of welfare economics, and in particular the second theory, that maximum welfare can be delivered by redistribution of wealth outside the microeconomy, and free market forces inside the microeconomy. In this instance, the microeconomy is the electricity sector and the redistribution of wealth is addressed in the macroeconomy of general taxation.

We work through the key elements of peak load pricing. As well as being the natural method for “energy only” markets (i.e., a normal commodity market without an imposed capacity mechanism), it also corresponds to the method for the natural development of the installed capacity obligation regulatory model—reliability options. These modelling features should therefore be attended to be the administrator/planner of the system.

THE MODELING FRAMEWORK

2.1 NOMENCLATURE

For ease of comparison with the reference works, we have generally used the same nomenclature.

Amount delivered	$S = \min(D, Z)$
Capacity	Q, q, Z
Call option premium	C
Cost	$f()$
Cumulative probability	$F()$
Demand	D
Event probability	λ
Fixed costs	B , occasionally β
Marginal probability	$f(), P(S)$
Price	P, S for forward price distribution, F for current forward price
Quantity, volume	Q, q
Rationing cost	R
Shock	u
Strike price	K
Subperiod length	w, H
Surplus	S
Variable Costs	b , occasionally γ
Welfare	W
Willingness to pay	$X^{-1}(D)$, WTP, inverse demand

“Suppliers” means retailers, rather than load serving entities

2.2 BASIC MODELING OF CONSUMPTION

Here we describe the modeling of consumption that is required to model peak load and capacity pricing.

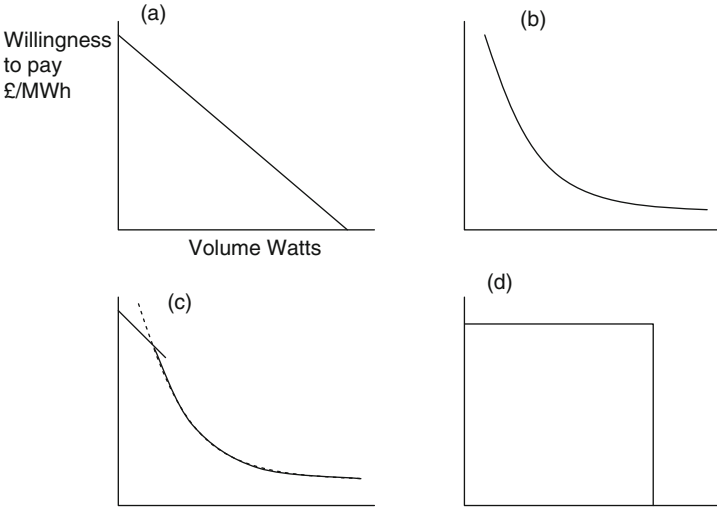


Figure 2.1 Four useful demand functions. Function C can be splined to make the slope continuous.

2.2.1 Willingness to Pay and the Demand Function

Willingness to pay is the price at which the consumer is indifferent to consuming and nonconsuming.

There are four demand functions of interest, namely: i) linear (quadratic utility in relation to volume, figure 2.1[a]), ii) linear log log (figure 2.1[b]), iii) two part (figure 2.1[c]), and iv) constant to a limit, called right angled (figure 2.1[d]). A particular challenge for electricity is that we need a demand function that can encompass a price range over at least six orders of magnitude while at the same time having a finite limit.

2.2.2 Utility

Utility is the worth of an endowment of a good to an individual, in the money metric. The willingness to pay is equal to the slope of the utility function. The main functions in use are shown in figure 2.2.

For the ex ante utility of a risky endowment, we apply the basics of the approach of Von Neumann and Morgenstern (1944) (VNM). So the ex ante utility of a total wealth that is stochastic is equal to the probability weighted average of the ex ante utilities of each wealth state. This makes a number of assumptions, the most important of which for present purposes are:

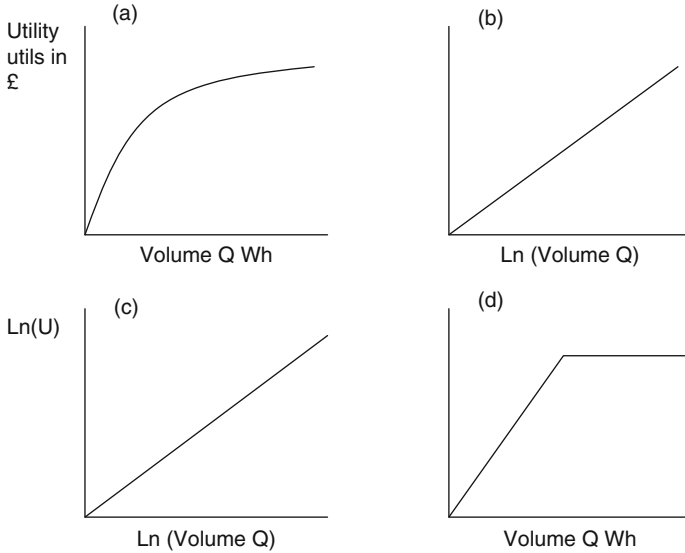


Figure 2.2 Some utility functions (a) Quadratic (b) Log (c) Log log (Cobb Douglas) (d) Linear to a limit.

1. probability distributions are stationary (constant distributional form and coefficients) for the past and future
2. probability distributions are determinable from nonparametric actuarial analysis and ideally reconcile to parametric forms constructed from the economic and engineering fundamentals and modeled using a Bayesian approach
3. probability distributions are intuitively understood even for extremely unlikely events and risk aversion increases monotonically with risk amount
4. utility is not path dependent (i.e., the level of wealth uniquely determines utility)

Each of these assumptions is fragile and important and should be controlled for where relevant.

2.2.3 Surplus

For an individual, the net surplus is equal to the utility minus the cost. It is commonly expressed as the area under the inverse demand function as shown in figure 2.3.

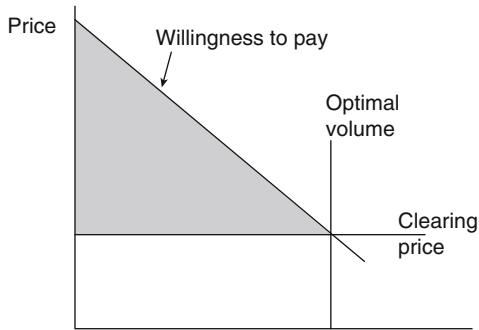


Figure 2.3 Consumer's surplus.

It is also common in economics to blur the distinction between consumer's (i.e., the individual's) and consumers' (i.e., society's) surplus. Figure 2.3 can apply to either, but we cannot regard the consumers' surplus as the function to be maximized, unless to recognize policy requirements, we apply some constraints that can be described with welfare functions.

2.2.4 Welfare

Welfare is the utility of society as a whole, taking inequality into account, and the entity that we wish to optimize through policy. In constructing a quantity for welfare, we clearly require the ability to rank any combination of endowments to different individuals in terms of societal welfare. In practice, this requires utility to be interpersonally quantitatively comparable. This point is highly contentious and here we regard it as an axiom.

What welfare can do for us is to impose restrictions on the consumers' surplus to recognize features that are additional to the optimum arrived through *tâtonnement* (a continuous auction) in the market economy. In particular, these relate to inequality and fairness.

The two extremes of welfare functions are:

1. Benthamite—societal welfare is the sum of individual utilities
2. Rawlsian—societal welfare is the lowest of individual utilities, so the objective is the maximin (maximum minimum utility)

There are various intermediate functions that for our purposes divide into two:

1. general inequality—intermediate between Benthamite and Rawlsian, with members of society differentiated by a single factor – endowment of wealth
2. lexicographical—with further characteristics applied to members of society and forming the objective function. Generally ranking by endowment amount but other weightings (e.g., by age) are possible.

Commonly we express the aggregate surplus as the “first best” entity to be maximized, but with a constraint, which makes it “second best.” So a second best optimization might be to maximize the consumers’ surplus subject to no consumer having less than a fixed amount, or half the maximum amount, or some other restriction. We also need to consider “third best” (two constraints to observe) and even “fourth best.”

2.2.5 Rationing and Deadweight Loss in the Hotelling Framework

We can see in figure 2.4 that the loss of welfare from building too little or too much volume is for small volume differences and efficient rationing proportional to the square of the volume difference. In the Hotelling framework,¹ fixed costs are regarded as sunk, and optimization is at the margin considering only short run costs.

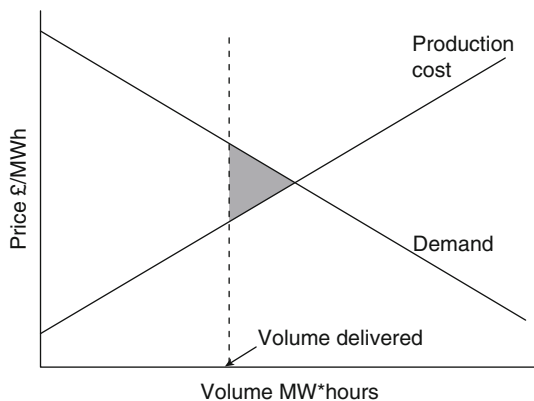


Figure 2.4 The deadweight loss of inefficient volume delivered. Shown in gray.

2.2.6 Shocks

Shocks are changes to demand or available capacity. This can be from exogenous forces, common to many or all actors, or endogenous forces, specific to the individual or production unit.

The consumer shocks are generally expressed as shocks to the demand function, which may be vertical, horizontal, or homothetic (horizontal and vertical with a movement of the demand function away from the origin).

It is important to understand the cause of the shock. It may be:

1. a change in preference (such as conversion of heating from gas to electricity or vice versa)
2. a change in need caused by a shock to endowment (such as a changed need for heat as a result of a cold weatherfront)
3. a change in aggregate wealth of the consumer, or
4. a change to relevant population.

For a linear demand function (quadratic utility), linear shocks to volume demand and to willingness to pay are geometrically equivalent, as we see in figure 2.5. Proportional shocks can also be modeled. It is generally important to reconcile the nature of the shock to a physical explanation. The proportional shock to volume can be understood in terms of population change and aggregate volume, and the proportional shock to price can be viewed as a shock to money or other good endowment.

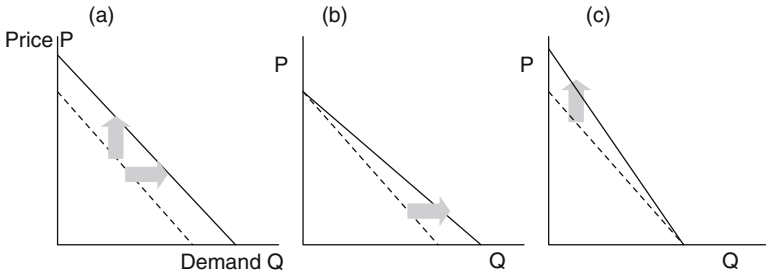


Figure 2.5 Demand shocks (a) Homothetic linear (b) Proportional shock to volume, for example, endowment (c) Proportional shock to willingness to pay, for example, the value of money

2.2.7 The Load Duration Function

The load duration function takes the system load in all subperiods in a year and rearranges them from chronological to volume order. This is shown in figure 2.6.

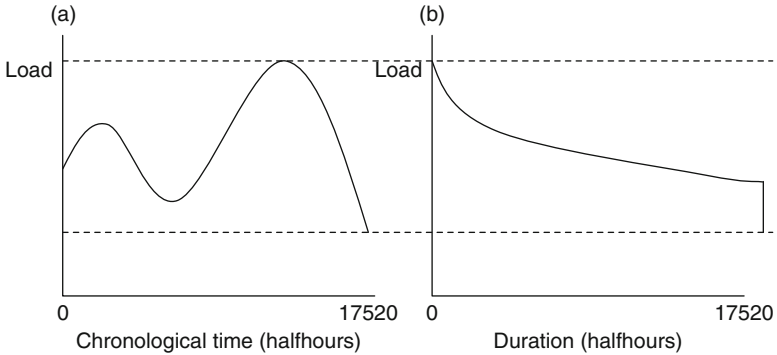


Figure 2.6 The load duration function (a) Past chronologically correct order (b) Reordered.

Source: Harris (2014)

In deterministic conditions, the load duration function and price duration function unite through the use of the production stack and a pricing algorithm, such as peak load pricing or variable cost pricing applied to inelastic demand. This is shown in figure 2.7.

Both load and price duration functions can be calibrated against ex post outcomes. Due to the uncertainty of the timing of the peaks,

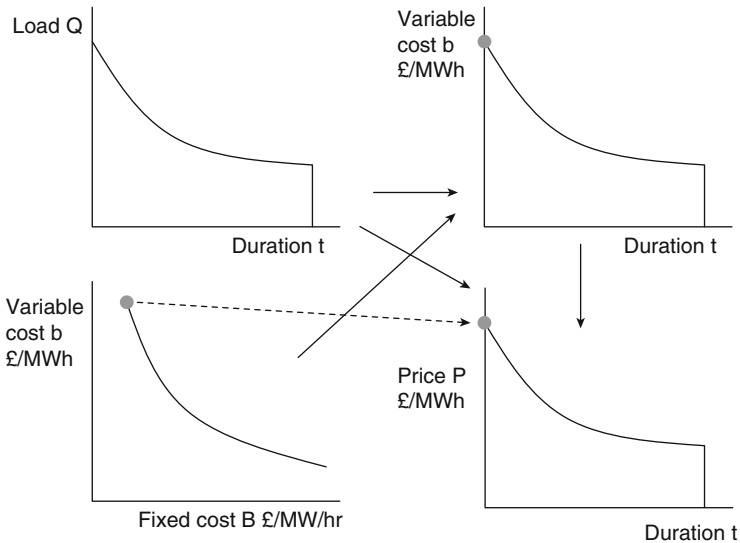


Figure 2.7 Construction of the equilibrium price duration function. The dotted line shows that the fixed costs of the peak unit play a role in forming the peak price.

the price duration function cannot be fully derived from the market forward price vectors.

Modeling shocks to the expected load duration function is essential for the consideration of capacity mechanisms and peak load pricing, and we will attend to this in section 5.7 and 7.1.1.

In common with all stack modeling that uses a load duration curve and/or which is not stochastic, the absence of state change cost modeling is a significant shortcoming of this method. While there are workarounds,² such as assuming that the timing of the system peak is fairly narrowly distributed around an expectation, resilience failures outside expected peak times are poorly catered to.

2.3 BASIC MODEL OF PHYSICAL/PRODUCTION

2.3.1 Basic Costs with Hard Constraints

The basic model for power plant is to have a fixed cost in £/MW/hr and a variable cost in £/MWh that is constant (i.e., constant returns to scale) up to the capacity limit, at which it becomes infinite.

In general, constant returns to scale in capacity are also assumed as is seen on the right in figure 2.8.

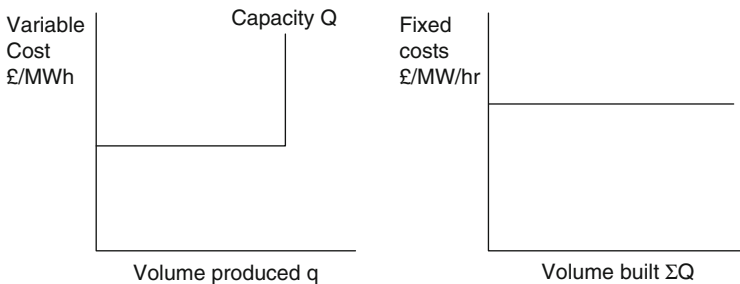


Figure 2.8 The standard cost model for power stations.

2.3.2 Costs at System Level—the Merit Order and Stack

The stack is the arrangement of all available units in “merit order” of ascending variable cost. For each variable cost, we can impute a fixed cost from the technology frontier. It is often convenient to depict the stack on the frontier as we see in figure 2.9(b). The length of the solid line denotes the installed or available volume.

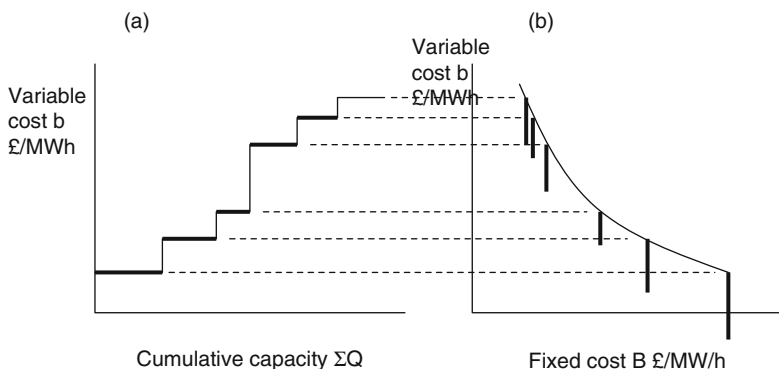


Figure 2.9 (a) The available unit stack on the system; (b) The stack modeled on the technology frontier. The length of the vertical lines represent installed or available volume and equal the length of the horizontal lines on the stack

2.3.3 Breakdown of Cost Elements

The four key costs of power stations are:

1. fuel and consumables
2. engineering, and the total cost of plant failure
3. environmental allowances and shadow costs of constraints
4. risk and finance.

Each of these has a fixed and variable element. For the majority of theoretical modeling of peak load pricing, it is sufficient to lump all fixed costs together and all variable costs together. So to fuel and consumable variable costs is added the variable engineering costs (plant life utilization constructed from the number of hours run plus the number of starts) and tradable emission allowances. The risk and finance costs are commonly treated as fixed, but as we see in section 2.4.6 and 2.4.7, fixed costs do not fall evenly over time and, in addition, can change.

2.3.4 Costs with Soft Constraints

The situation of hard constraint is the asymptotic extreme of the general case of soft constraints as shown in figure 2.10.

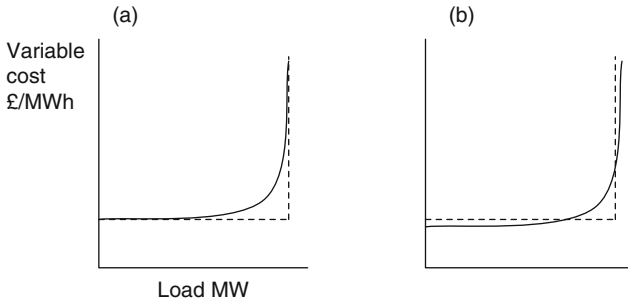


Figure 2.10 Soft constraints pictured as asymptotic to hard constraints (a) With hard constraint having cost dominance (b) No cost dominance.

2.3.5 The Energy Supply Chain and Market Arrangements

The key actors in the electricity supply chain are

1. fuel production
2. power stations
3. transmission at high voltage
4. distribution at low voltage to homes and businesses
5. retail suppliers
6. metering at the points of supply
7. wholesale markets in fuel, power, environmental allowances, and other

In the nationalized era, these were planned and managed as a complex. In the liberalization era, the sectors were “unbundled” (separated) to varying degrees, from operating level (managerial unbundling) to full ownership unbundling.

In Great Britain, the model is the Supplier Hub, in which the retail supplier pays the transmission and distribution companies and contracts with consumers. For capacity modeling, it is often useful to first model the physical supply chain using the point-to-point model (see figure 7.21 in section 7.4.2) and thence remap the commercial relationships using the Supplier Hub.

2.4 OTHER ASPECTS OF MODELING

2.4.1 Load Factor Duality—Time and Probability

In modeling terms, a single subperiod stochastic setting with n discrete probability states is equivalent to a deterministic setting with n

subperiods. Similarly, a stochastic setting with m discrete probability states and n subperiods is equivalent in modeling terms to a deterministic setting with $n * m$ subperiods. This duality is particularly useful in modeling peak load pricing.³

2.4.2 Public and Private Goods

Distinct from private goods, public goods cannot have access selectively restricted, for example, to those who can pay.

There are several key drivers to the public goods nature of electricity:

1. The nature of electrical flow, that physical demand is instantaneous and draws power from the grid regardless of contract, and in addition electricity follows the path determined by physics rather than the contract path.
2. The inadequate nature of “nonsmart” consumer metering so that the measured amount of electricity consumed commonly has a granularity of months rather than minutes, and hence it is not possible in the short term to make the good private by self-rationing, combined with ex post charging for electricity used in each short timeframe.
3. The status of electricity as an essential good with universal service and hence virtual banning of access refusal (disconnection) on grounds of nonpayment; in addition, regulatory pressure to socialize prices rather than charge on a dynamic (“time of use”) and cost reflective basis.
4. Perfectly reliable and stable transmitted electricity is a totem of a modern economy and this status is itself a public good.⁴ So, power interruption even if agreed between all actors would be a significant political issue.
5. Electricity and the economy are regarded as complementary “goods,” in which case both are public goods.

For current purposes, the two key public goods features affecting peak load pricing are rationing efficiency and regulatory suppression of economically efficient price signals.

We should note that it is technically possible to make electricity a fully private good if the amount consumed can be measured at sufficient time resolution and payment enforced and actually collected ex post. For example, if all consumers are billed on the basis of halfhourly price and consumption, then their good is to all intents and purposes purely private, as there would never be load loss from capacity

adequacy shortfall since there would always be demand-side management, commonly automated to a price trigger. Consumers who continue to consume during extreme shortage would pay extreme prices that would in practice be limited only by the willingness to accept curtailment by other consumers.

2.4.3 Risk and Cost of Risk

The standard representation of the cost of risk in terms of utility is shown in figure 2.11. This follows the approach of Pratt (1964). So if risk is applied to the quantity of endowment, we can see that both higher and lower amounts have quantities lower than the tangent to the current curve on which the expectation of utility change in relation to risk being zero. Cost of risk is just a manifestation of the utility function, and in this book we require considerably more sophistication than the standard “linear aversion with respect to variance” approach, and hence the term “cost of risk” should really be regarded as shorthand for the application of the utility function to an environment with stochastic shocks to quantity.

In this book we begin with the standard formalisms for risk and cost of risk. They are:

1. linear aversion to variance, and thence a quadratic utility function
2. no uncertainty, that is, the risks faced are well characterized
3. stationary—the risk coefficients or distributional form do not change over time
4. previsibility—the utility on arrival at a wealth state is equal to the expected utility at that state. In general, here we are faithful to the VNM theory of risk.

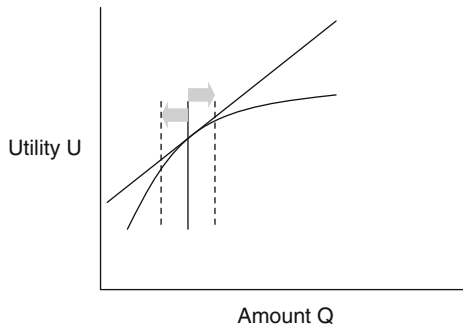


Figure 2.11 Standard representation of the cost of risk.

5. no asset portfolio—the asset owner has no other assets or indeed other risks
6. forward market drift—linear in proportion to standard deviation
7. no skew—the distribution of prices or price returns are normally distributed
8. constant cost of risk—not path dependent, and so on
9. correlations—well characterized, stationary and standard linear.

Each of these is of high practical importance in the consideration of peak load and capacity markets and we will discuss each. The largest problem in the list above is the incompatibility of items 4 and 5. To be consistent with the standard framework of derivatives, we must assume 5, and to be consistent with utility and asset theory we must assume 4. The problem is that our physical asset experiences a value drift in proportion to stochastic variance and our financial asset experiences a drift in proportion to the standard deviation (the square root of variance). While we can reconcile the two worlds by making a portfolio assumption, for example, consider that the asset owner has a portfolio of assets and thence through the capital asset pricing model experiences a cost of risk in proportion to the standard deviation not variance of the individual asset. However this framework is unwieldy and our attention here is to anchor analysis in standard theory. Therefore we have to accept a degree of self-inconsistency within our modeling. There is in fact so much uncertainty about cost of risk that this turns out not to be a significant practical modeling problem.

2.4.4 Policy Issues in Relation to Market Structure

This book explains capacity obligations and shows how capacity obligations are developing in the direction of energy-only markets, albeit that there is no certainty that this is a target or reachable destination. The development toward a pure market approach is in fact driven not at all by an ideological favor of markets but a continuous drive to efficiency based on the empirical observations of the market as it is. The policy opposition to the market approach is prevalent and there are three particular objections to the energy-only paradigm with no central intervention in relation to capacity.

These are

1. Market power—The regulator fears that a generator who at the margin can keep the lights on will charge an extortionate rent if the moment arrives.

2. Moral hazard—The generator fears that a legitimate rent will be expropriated by government when the moment arrives. With expropriation risk, the generator must plan a higher price and with the higher price the more the expropriation risk increases. The situation is commonly terminal in that the generation does not get built.
3. Regulation—There is both moral hazard, in which there is a systematic risk of expropriation through regulatory change (e.g., addition of change of price cap, ex post taxation) or simply “time inconsistency” in which adverse or beneficial effects on generators is a by-product of regulatory change

Each of these can be resolved within the market paradigm. We simply regard political and other actions as forces with a mix of exogenous and endogenous features, as well as correlations and causal links.

2.4.5 Games

In this book, we regard gaming as the behavior of rational actors, and in no pejorative or judgmental sense. There are indeed circumstances in electricity in which gaming can be, and perhaps has, accrued excess returns to market actors who have abused market power.

The key games are:

1. Stackelberg, in which a market follower takes the volume of the market leader as a given, and addresses the residual market
2. Bertrand/Edgeworth, in which the market followers assume that they have no role whatsoever in price formation and always offer at the shortest run marginal costs
3. Cournot,⁵ in which market actors effectively assume that all actors behave in the same way and know each others’ cost functions as well as the market demand function

We generally restrict ourselves to Nash equilibria, in which market actors make choices based on estimates of the behavior of other actors, and do not regret those choices after the decision uncertainties have been resolved.

There is a particular game that is of interest in the present context, which is the tension between the value of preemption and the value of optionality.

In examining real options, as we do in section 5.9, we broadly assume that our actions have limited effect on prices or behaviors, that is, we are price takers not price makers.

We also need to consider the value of preemption, in which we make a public statement about a decision made that is irrevocable. In doing so, we affect the market. The best example is the commitment to build a power station and offer at variable costs that is likely to run it at baseload. This creates a Stackelberg game, in which the residual volume addressable by the rest of the market is reduced. Note here that having an option has negative value, as, for example, the market might become Cournot rather than Stackelberg and reduce the price and/or the volume of the first actor. There is then a tension between the positive value of an option and the negative value of the market knowing that the actor has an option.

This tension plays out in the modeling of supply function equilibria, in which the elasticity of the forward market is affected by the advertising to the market immediately after a forward contract has been agreed.

The direct relevance here is in the risk management of power plant. If we are a pure price taker we should always completely ignore our forward and option contracts, and operate “live” in response to the prevailing market. However, if the installed stack is anything other than the perfect one and there are no price caps, this situation is unstable as fixed costs are not recovered.

2.4.6 Fixed Cost Allocation Over Time

The standard technique for professional traders is for the net present value of the portfolio, called a “book” to be marked to market, and for the trader to have a virtual loan at this value so that the total portfolio has a value of zero. This ensures that interest on the book value is taken into account on calculating profit and retained earnings/losses are constantly repatriated to the parent company.

The same applies to physical assets. For optimization purposes, the fixed costs of capital are set by the current value of the asset and not the purchase cost. Over a time interval then, the fixed costs are equal to the finance cost at the asset book value plus the decline in asset value.

More generally, we can deduce from this that if the only fixed costs are capital costs, then we regard these costs as incurred in accordance to marginal revenue. This is easy to understand when we have a peak period in which we make a marginal profit followed by an off-peak period in which the revenue equals the variable cost. If we reverse the chronological order of the peak and off-peak we still incur the fixed cost in the peak. Finally we can apply load factor duality as described

in section 2.4.1 to incur the cost and peak load revenue in a stochastic peak in a single period setting.

This cost allocation has a direct bearing on peak load pricing, as we will see in section 3.1.

2.4.7 Dependence of Fixed Capital Cost on Plant Value

It is apparent from section 2.4.6 that if the asset value rises, then there is an immediate profit that is repatriated, and an increase in fixed costs on the now higher value. In equity markets this is essentially the same effect as the well-known “Tobin Q.”⁶

As a further consequence, it is apparent that in the deterministic world, the profile of fixed costs in terms of capital relate directly to the asset value. It is thence obvious that for optimization purposes the fixed costs should not be allocated evenly over time but according to the profile of the asset life over time. It follows further that the fixed costs relate directly to the ongoing margin between prices and variable costs. A good example of this effect is to be found in the Steiner analysis in section 3.4.

2.4.8 Value of Lost Load—VoLL

In theory, the value of lost load (VoLL or VLL) is the average willingness to pay to avoid loss of power or, somewhat equivalently, the compensation willingness to accept in lieu of power loss. Clearly the use of an average rather than marginal rate is something of an issue in terms of efficiency.

VoLL is a huge subject in its own right. For the purpose of this book, there are certain key features that take part in the modeling:

1. the marginal VoLL implied from the demand function
2. the inefficiency of rationing, not rationing in order of willingness to pay
3. the establishment of VoLL by empirical observation by actual willingness to pay and the estimation of VoLL in the absence of observation
4. viewing voluntary acceptance of possible lost load in terms of an unconditional premium plus a conditional payment at a strike price on load loss
5. social welfare considerations, for example, not preserving power for the rich on grounds of willingness (ability) to pay more
6. systemic issues.

We see in section 3.5 that the cost of rationing in relation to total amount of rationing can be convex or concave. When we consider systemic lost load, we have to take into account broader substitution and the complementarity of electricity and the operation of society. Consider if my home loses power, I can go to my neighbor's home. I am therefore highly affected by his situation. In this sense, supply to our respective homes are complementary. As power loss widens, fundamental aspects of society start to fail, such as street and traffic lighting, pumping clean water, refrigeration of groceries for sale, and public transport. For this reason, we should regard systemic loss of supply as having a convex form in relation to amount (TWh) of lost load.

2.4.9 Cost Frontier Duality

To make modeling manageable, we need to limit the degrees of freedom for modeling power plant. In this regard, there is a duality that we find useful, relating to unit size and unit technology. Figure 2.12 shows two cost frontiers for units with convex costs (soft constraints as described in Chapter 4 and therefore no clear unit size). The first is the well-known technology frontier, which here we assume applies to different units of the same capacity. The second is the equivalent figure, for a single technology, in relation to unit size. The vertical axis is a proxy for variable costs. We can see that a high technology plant has high fixed costs and low variable costs. A small plant has low fixed costs by virtue of being small and high variable cost at a given load due to cost convexity.

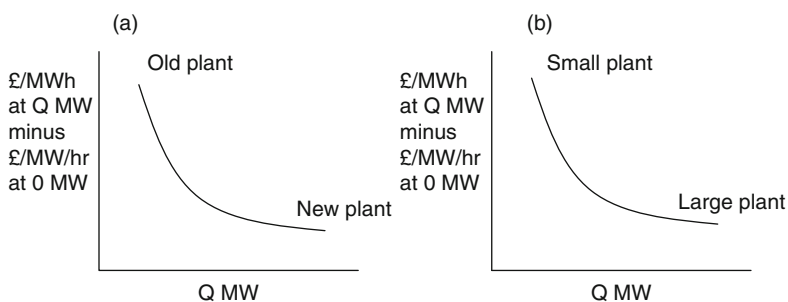


Figure 2.12 Cost frontier duality (a) Different technologies and one size (b) Different sizes and one technology.

Source: Harris (2014)

THE FRAMEWORK AND DEVELOPMENT OF PEAK LOAD PRICING

3.1 BASIC PEAK LOAD PRICING THEORY

3.1.1 Introduction

In this chapter, we describe the simplest exposition of peak load pricing.

3.1.2 Framework

1. Single epoch with no plant entry, exit, or aging.
2. The setting has two subperiods that can be of uneven length.
3. Demand is deterministic and inelastic.
4. There are two available technologies.
5. Units can be sized just before the period starts, at constant returns to scale in capacity (i.e., a single build cost per unit of capacity).
6. Returns to scale in operation are constant, that is, a single variable cost for each technology.

This is shown in figure 3.1.

3.1.3 Analysis

3.1.3.1 *The Turvey Algorithm for Efficient Build and Run*

If we have a baseload unit (baseload meaning running all the time) with (B_2, b_2) , we will only build and run the peak unit with (B_1, b_1) in the peak period of length λ_1 if

$$b_1 + \frac{B_1}{\lambda_1} < b_2 + \frac{B_2}{\lambda_1} \quad (3.1)$$

This is the basis of the Turvey algorithm.¹ Note that Turvey takes a least cost approach and does not require financial equilibrium.

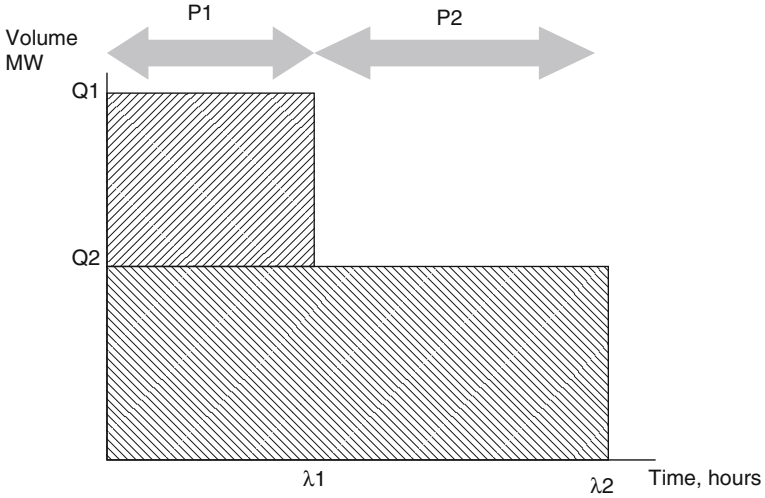


Figure 3.1 The simplest load variation for analysing peak load pricing.

If the load duration function is continuous, then the amount of time for which we run the peak unit is determined by rearranging equation (3.1). So we have:

$$\lambda_{\text{peak}} = \frac{(B_2 - B_1)}{(b_1 - b_2)} \quad (3.2)$$

The optimal build volumes can be inferred directly from this. The baseload unit build is found from the intersection of the load duration function and time λ_1 , and the peak load unit build is equal to the peak load minus the baseload capacity.

3.1.3.2 Convergence of the Turvey Inequality

Consider again discrete time intervals with a peak period length.

Turvey considers replacing a high merit unit 2 with a low merit unit 1 in the peak.

We can look at the reverse situation, where we consider replacing the high merit unit 2 by the low merit unit 1 in the off-peak. To run the off-peak unit we have;

$$b_1 + \frac{B_1}{\lambda_2} > b_2 + \frac{B_2}{\lambda_2} \quad (3.3)$$

Here, λ_2 is the time that high merit unit 2 was planning to run, in this case, unit time.

Now suppose that we have n units, and n evenly spaced subperiods of time, we have

$$\begin{aligned} (B_{i+1} - B_i) + \lambda_i(b_i - b_{i+1}) &> 0 \text{ and} \\ (B_{i+1} - B_i) + \lambda_{i+1}(b_i - b_{i+1}) &< 0 \\ (B_i - B_{i+1})\lambda_{i+1} &< (b_{i+1} - b_i)\lambda_i\lambda_{i+1} < (B_i - B_{i+1})\lambda_i \end{aligned}$$

For subperiods of even length we have

$$\lambda_{i+1} - \lambda_i = \frac{1}{n} \text{ and } \lambda_i = \frac{i}{n}$$

So,

$$\begin{aligned} (B_i - B_{i+1}) * n * \lambda_{i+1} &< (b_{i+1} - b_i)\lambda_{i+1} - (B_i - B_{i+1}) < 0 \\ (b_{i+1} - b_i)\lambda_{i+1} - (B_i - B_{i+1}) &\approx 0 \text{ as } n \rightarrow \infty \text{ and } \lambda_{i+1} \rightarrow 1 \\ \frac{db_i}{dB_i} = \frac{1}{\lambda_i} &|_{n \rightarrow \infty}. \end{aligned} \quad (3.4)$$

So the higher the divisibility of time and unit size and the higher the merit plant we are looking at, the closer the Turvey inequality is to an equality.

This is an important equation and the evolution of the technology stack works to make it true in practice.

While the installed stack does not have units of infinitely (or even very) small size, what counts here is the extent to which a lower bound on unit size or economy of scale are practically important. In practice, we find that the low slope of the load duration function for low loads means that unit size is not important, but for high loads and rare events, unit size is indeed important, and the lower bounds of practical unit size has a real effect.

Regarding the technology frontier, which we will discuss further in section 3.9, it is practically continuous even in the presence of discrete families of technologies, because of “stack evolution” in which units choose the most cost-efficient aging evolution path of fixed and variable costs.² Additionally, we can see from figure 3.2 that while technology evolution orthogonal to the frontier is expensive, evolution in the shaded regions is less so.

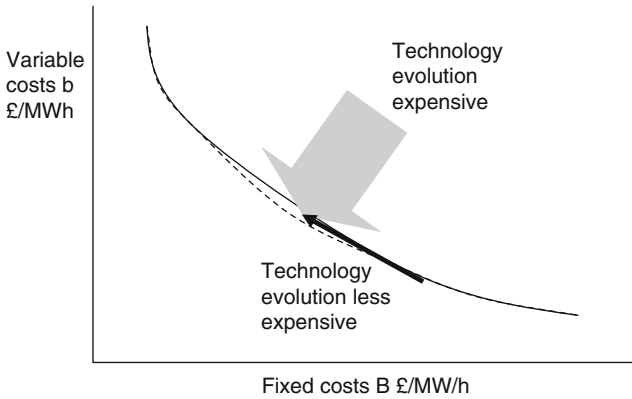


Figure 3.2 Efficient evolution of the installed technology stack cost frontier.

This effect strengthens the validity of equation (3.4) in the high- and mid-merit regions without the requirement for an infinite number of units.

A further effect strengthening equation (3.4) is the planned evolution of the unit in terms of the cost of finance, noted in section 2.4.7. In practice, the fixed cost evolution vector adjusts the fixed costs in the late stages (low merit) of plant life, making the technology frontier more continuous even for low merit units.

3.1.3.3 Cost Allocation for One Unit

Suppose the peak/off-peak load profile is delivered by one unit. Assume that the only fixed cost is cost of capital and that we price the peak at $b + B/\lambda$ and the off-peak at b . If we allocate fixed costs as described in section 2.4.6, then it is obvious that the full allocation is made to the peak period. We therefore do not regard the peak period as collecting back all the fixed costs incurred over the whole cycle, but in fact the fixed costs are incurred only in the peak and recovered in the peak.

3.1.3.4 Cost Equilibrium for Two Units

The lowest merit unit uplifts its offer above variable cost until the fixed cost is exactly covered. The second lowest merit unit then does the same, and so on.

The first test is to see if any unit makes excess profit.

For cost equilibrium, the price for the peak period is then simply;

$$P_1 = b_1 + B_1 / \lambda_1 \quad (3.5)$$

This is the essential formula for peak load pricing. For the peak unit there is an uplift above the variable cost pricing level b_1 in order to cover fixed costs.

We can now calculate the cost equilibrium price for the subperiods for which the second lowest merit unit is price setting;

$$P_2 * (\lambda_2 - \lambda_1) + P_1 * \lambda_1 = B_2 + b_2 * \lambda_2$$

Rearranging we have

$$P_2 = b_2 + \frac{B_2 - \lambda_1(P_1 - b_2)}{(\lambda_2 - \lambda_1)} \quad (3.6)$$

Now substitute equation (3.4) into equation (3.6):

$$P_2 = b_2 + \frac{(B_2 - B_1) - \lambda_1(b_1 - b_2)}{(\lambda_2 - \lambda_1)} \quad (3.7)$$

Noting equation (3.3) we substitute equation (3.4), as an equality, into equation (3.3) and arrive at

$$P_2 = b_2.$$

3.1.3.5 Consideration of a Third and More Units

Let us now consider a third unit. This is now the baseload unit. So

$$P_3 * (\lambda_3 - \lambda_2) + P_2 * (\lambda_2 - \lambda_1) + P_1 * \lambda_1 = B_3 + b_3 * \lambda_3.$$

We can then substitute in for P_1 , P_2 , and from equation (3.4) and the associated arguments, we can apply the Turvey inequality as an equality, and substitute for b_1 , B_1 , b_2 , B_2 , and B_3 . We then simplify to arrive at

$$P_3 = b_3.$$

The same approach can be taken for four or more units to give us the general

$$p_i = b_i \text{ for } i < 1, \quad (3.8)$$

where $i = 1$ represents the peak period.

This is the point of convergence of peak load pricing and marginal variable cost pricing. Note the condition for this convergence is financial equilibrium.

We can also see that even if the peak period is very short, it plays an important role in revenue generation for all units, as the shorter it is, the higher the price. Noting the duality of time and probability in the load and price duration functions, we can see similarly that price revenue during rare events is also important for all units.

We can see that if the price is capped, then we will have “missing money.”

A key area of focus of this book from much of the swathe of literature is the recognition of the fixed cost at the peak period, whether it be from i) the fixed costs incurred by readiness for demand-side management (DSM), ii) the optional fee required by any offer of load, for example, by a foreign market, or iii) a fixed fee due to consumers in return for ex ante acceptance of finite probability of lost load. These are commonly set to zero in the literature.

3.2 EARLY DAYS OF TARIFF EVOLUTION

3.2.1 Introduction

The electricity supply industry began with inventors and entrepreneurs, and we should not be surprised that tariffs at the time were as innovative as physical invention. Then, as now, tariffs were driven by social, political, and commercial realities, and were not always set rationally.³ Then, as now, tariffs were also limited by the technological capabilities of the meter and metering system.

Pricing in the early days set the precedent. Academic debate has followed development in the science of economics. The relation between practitioner debate and academic discipline has been the need for actors to “use bounded rationality to promote their own agendas.”⁴ Pricing now, as then is much influenced by relative power, ownership, control, and information, particularly in the metering sector.

3.2.2 The Early Development of Tariffs

The very earliest days were characterized by specific commercial drivers such as the need to generate demand for the product. In the relative absence of meters, the charge was essentially per outlet (down to the level of sockets when electricity arrived to individual consumers rather than municipal lighting) with broad estimates of load factor.

Of enduring fame is the Hopkinson tariff. Hopkinson⁵ (1892) proposed that consumers should pay a fixed and variable cost to ensure correct⁶ compensation for the provision of capacity. At the time, this

did not require more than basic metering technology with just the cumulative kWh consumption, because the main load was lighting, and consumers paid a capacity charge per light. Therefore the installed equipment at the consumer site determined the maximum possible (and likely⁷) consumption. If the tariff had a fixed charge per customer, in addition to the capacity and energy charges, the tariff was termed the Doherty Tariff.⁸

Arthur Wright had invented a meter that measured⁹ maximum (i.e., Watts) as well as total demand (i.e., kilowatt hours). The story goes that he met Samuel Insull during Christmas 1894 to explain the economics of fixed and marginal costs, which led to the tariff and therefore demand for the meter (for which Insull later became a shareholder). There was then a fixed and variable cost, but this time the cost could be charged according to actual (ex post) maximum consumption rather than theoretical maximum (ex ante) consumption. This was important at the time, because it was a lesser deterrent to the installation of equipment such as light bulbs than the Hopkinson tariff. Wright recognized¹⁰ that producer capacity cost depended on aggregate demand at system peak times rather than aggregate of theoretical individual maxima. The lack of available technology to measure the timing of the peak load led Wright to take a load factor approach. The implementations differed slightly. So, for example, Eisenmenger (1921) cites a charge for the first (ranked by kW rather than chronologically) few kilowatts, followed by a lower charge for the next, and so on. Nowadays a Wright tariff is generally taken to mean a tariff of this type, not recognizing the timing of the peak.

The implementation of tariffs was widely varied¹¹ and lagged theoretical and technological development. As late as 1923,¹² consumers were still paying the capacity part of their tariffs based on the number of openings to the distribution system, and paid different rates for different kinds of rooms according to the likely load factor.

None of these two methods considered¹³ the time of day, or the consumption in relation to system peak. A meter innovation by Kapp enabled measurement of the peak demand and the time at which it occurred. Barstow of Brooklyn Edison promoted this. The Barstow tariff was used but has never been widespread¹⁴ for domestic¹⁵ use.

Figure 3.3 depicts three of these tariffs.

Figure 3.4 shows different charging methods for peak consumption.

Meters with clocks that enabled a (fixed time) peak and off-peak tariffs started to be used in the early twentieth century, but rollout was limited due to the high total costs.¹⁶

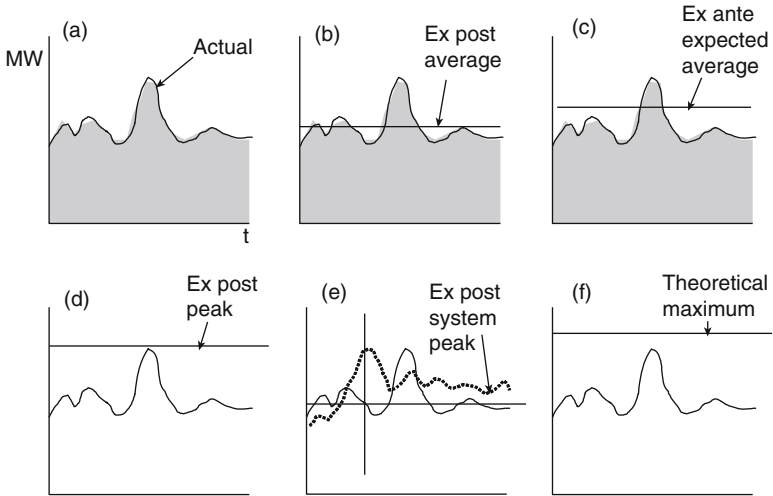


Figure 3.3 Fixed and variable tariff structures (a) No fixed charge (b) Fixed charge tied back to actual average consumption (c) Fixed charge tied to predicted average (d) Fixed charge tied to actual peak (Wright tariff) (e) Fixed charge tied to actual consumption at time of actual system peak (Barstow-Kapp tariff)¹⁷ (f) Fixed charge tied to theoretical maximum consumption (Hopkinson tariff).

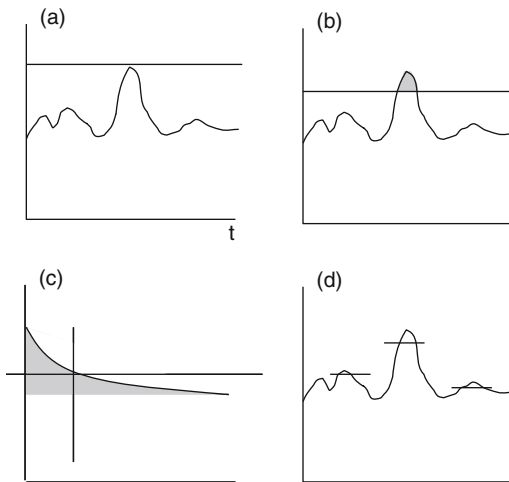


Figure 3.4 Methods of measurement of peak for fixed charge (a) Instantaneous (b) Maximum consumption in single peak in excess of designated duration (c) Load duration curve (d) Average maximum consumption in peaks separated by at least a minimum interval (the "triad").

3.2.3 Pricing in the Nationalized Era

In the early days, the growth of electricity was similar to the growth of railways, with private firms, and duplicate and incompatible infrastructure. While government and municipal involvement in infrastructure (build, planning, commitment to use and pay, etc.) had increased over the years, it was after 1945 that state control took hold by forced nationalization (1947 in Great Britain).

Over the years, there have been numerous tariff experiments. In Great Britain the seasonal (but not diurnal) “Clow tariff” was tried in 1948 and then abandoned. After that the Hopkinson two-part tariff was applied at wholesale level by the British Electricity Authority to the Area Boards. The capacity charge was based on the average of the halfhourly maximum¹⁸ demand of the last two years for that Area Board. The tariff was changed in 1957 to contain a fuel cost pass through. In 1962, the capacity charge was based on Area Board demand at the time of national peak demand and the energy charge was different in the day and the night. The Authority resisted proposals to add further diurnal resolution to tariffs.

Adjusted to money of 2014, the bulk supply tariff was £100/kW/year plus £0.1/kWh.¹⁹ For a 1kW load at 50 percent load factor, the cost was then £100 for capacity and £385 for energy. The kW level for the Area Supply Board was the average of the maximum halfhour in each year for each grid supply point in any half hour between 07:00 and 19:00 hours. Largely in response to South Eastern supply board with a peak load on Sundays from cooking, this was revised in 1950/51 to include only 07:00 to 19:00 hours on weekdays and 07:00 to 12:00 hours on Saturdays. In 1955/1956, there was an adaptation that reduced the tariff if (predominantly for weather reasons) the average countryside consumption was high. This moved some risk from suppliers to the network owner and would have had the effect of increasing transmission tariffs. In 1962/1963, a significant change was made to charge according to the average area board demand in the halfhour of system peak. This is the origin of the Triad system that prevails today. The 1960/1961 development also introduced reduced tariffs for the provision by the area boards of load interruptability.

The Netherlands also experimented with time-of-day tariffs for domestic consumers. It was in France that consumer tariff pioneering continued with the Tempo tariff in the 2000s. In the 365-days year, there was a preset number of red, white, and blue days with three respective prices, and the “color” of the day was announced in advance. Clearly this required metering and billing of daily resolution.

3.2.4 Development of Competition in the Wholesale Sector in Great Britain

The two major power station build types were coal and hydro, with nuclear picking up in the 1950s. With coal, the cost structures were broadly similar, and hence costing was similar, with consumer tariffs aiming to recover costs overall. Hydro was rather different, with high capital costs and low variable costs.

3.2.4.1 *The Pool in England and Wales*

Partly to facilitate private entry into the market, the internal “merit order” (ranking of variable costs) was formalized in England and Wales as the pool in 1990. The generator order and loading (GOAL) scheduling model took no account of fixed costs, and when the Central Electricity Generating Board was privatized, effectively used offers as a proxy for variable costs. An interesting feature of the pool was the capacity mechanism that we explain in section 5.1.2.

Pool type arrangements remain common, and in fact the trend toward real-time pricing of transmission constraints has encouraged their growth.

3.2.4.2 *Post-Pool Bilateral Markets*

The New Electricity Trading Arrangements (NETA) in 2001 (with Scotland in 2005 to make BETTA—British Electricity Trading and Transmission Arrangements). This was an energy-only market, that is, a normal market with no capacity mechanism (although a capacity mechanism was added in 2014 as part of the Electricity Market Reform).

Prices at any time are simply a function of bids and offers meeting in the market. Fixed cost recovery is driven by market forces rather than central management.

A particularly interesting feature about NETA/BETTA market is the symmetry between production and consumption. Instead of a producer market facing a nominally inelastic stochastic demand, retail suppliers procure contracts or pay a buyout “imbalance” price for the energy drawn without contract. To stimulate supplier development, the cashout prices were set initially at punitive levels. The balancing mechanism can be viewed as a capacity mechanism of sorts and is discussed in section 5.1.3.2.

3.2.5 Liberalization of the Retail Sector

The competitive retail supply market in England and Wales opened to business consumers in the 1990s, with full liberalization of the

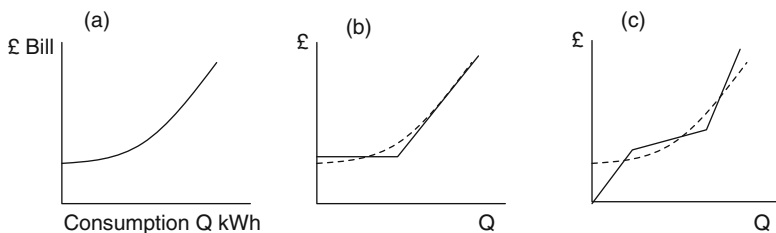


Figure 3.5 Tariff structures (a) Cost reflective (b) Approximation of cost by standing charge and single unit rate (c) Three-rate tariff.

residential sector in 1998. Under the Supplier Hub model, the transmission and distribution revenues regulated under price controls, and the actual prices subject to an element of control of charging methodologies. The supplier paid the distribution and transmission companies.

Suppliers were free to innovate in tariffs, and tended to charge a standing charge and a unit rate. With standing charges being unpopular, suppliers developed two rate tariffs with a high price primary block up to a kilowatt hour per month limit and a cheaper secondary block thereafter. The overall structure formed a proxy for a standing charge and unit rate.

Note that the most cost-reflective tariff has a standing charge and thence a convex cost with unit rate increasing with consumption. This can be approximated by a two-rate tariff. We can see this in figure 3.5.

Three-rate tariffs are favoured by some consumer advocates as they avoid the standing charge for very low consumers. Note that the tertiary rate can be regarded as a form of peak load pricing. In practice, this tariff regime is too complex to implement, particularly since the threshold amounts would ideally have a daily resolution.

3.3 POSTWAR THEORY AND APPLICATION OF CAPACITY CHARGING—THE DRÈZE APPROACH

3.3.1 Introduction

Drèze provides an excellent introduction to the practical application of theory in the early modern age of electricity, which was developed and applied mainly in France. A particularly interesting element is the movement away from the equilibrium arguments of the French econo-engineering school and thence Walrasian school, recognizing

the importance of equilibrium as well as marginal economics, to the primacy of efficiency at the margin. It provides a good example of the continued division between the approach of industry on one side, with a focus on equilibrium (covering fixed costs), and the prevailing marginalist economic theory that is easier to reconcile to the operation of short-term markets. The tension is also evident in policy formation, so here the pricing practice of the state-owned Electricité de France (EDF), recognized the need for equilibrium. The works of EDF employees Boiteux²⁰ and Massé remain seminal.

3.3.2 Framework

Our framework is assumed as follows:

3.3.2.1 Generator Cost and Pricing

1. The generator has constant variable cost returns to scale γ up to the capacity limit and constant fixed cost returns to scale β in capacity.
2. The generator is risk neutral.
3. We must make the capacity and pricing decisions prior to the resolution of demand uncertainty.
4. Prices are fixed before the resolution of uncertainty.

3.3.2.2 Consumers and Consumption

1. Consumers are homogeneous and have stochastic demand with a mix of endogenous and exogenous shocks.
2. The stochastic consumer demand functions are stationary (i.e., statistically stable) and actuarially well characterized.
3. The number of consumers is constant.

3.3.2.3 Build Volume

The stochastic consumer demand functions are of the form shown in figure 3.6. We can see in this figure that if we increase the standard deviation of demand, the inefficiency increases. If demand falls, then we have wasted capacity at a cost of $(1 - \lambda) * B$, and if demand rises (with probability 50 percent), we have wasted consumer surplus $(WTP - P) * \lambda$, where WTP is willingness to pay and P the price. In selecting our initial build volume for nonstochastic conditions, we have assumed a clearing price at the long-term equilibrium of producer costs $B + \gamma$.

If the producer opportunity cost of loss $P - (\gamma + \beta)$ exceeds the stranding cost of wasted capacity β then we build more if the standard

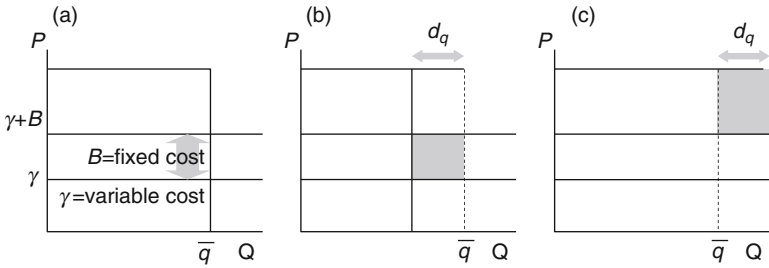


Figure 3.6 Build volume in the Drèze framework (a) Build to match deterministic demand (b) Wasted capacity cost for demand fall (c) Unsatisfied consumer surplus for demand rise dq .

deviation σ is higher. For this approach to be welfare optimal, we make the implicit assumption that the cost of lost load exceeds the cost of wasted capacity and so it is better to have too much than too little capacity.

In this simple case, we can see that provided that willingness to pay WTP exceeds variable costs plus twice fixed costs, then the optimal extra build under stochastic conditions is directly proportional to the absolute size of the demand shock dq . We can also see that for the producer to cover costs under stochastic conditions, we will need to raise the price if standard deviation increases.

We take the analysis forward by assuming that it is optimal to build an amount that exceeds the demand expectation by a factor k and the size of a statistical function ψ of demand. The combination of these two allows us to be a little more generic than having a constant willingness to pay. So, for example k can relate to the value of lost load (VOLL) and ψ can relate to the loss of load probability (LOLP). We may express $\psi = f(\sigma)$ or $\sigma = g(\psi)$.

For ease of analysis, Drèze assumes a normal distribution of aggregate demand²¹ and hence our most natural factor to use to determine build volume is the standard deviation of aggregate demand σ or the variance σ^2 . Thence for analytic simplicity, we build capacity q_c , in excess over demand expectation on a linear relationship between σ or σ^2 . So $q_c = \bar{q} + k_1\sigma$ or $q_c = \bar{q} + k_2\sigma^2$. Ideally k_1 and k_2 would be dimensionless but the exposition is easier if we make σ a dimensionless percentage and k have units of volume. Drèze expresses the probability of lost load λ . We drop λ in the equations but explore λ below.

\bar{q} is the expectation of aggregate demand. $\bar{q} = \sum_{i=1}^n \bar{q}_i$ for the n consumers.

In making a simplification to a complex demand function, we do in fact approximate to the way the industry generally manages to proxy for the VOLL and LOLP, as we describe in section 3.11 in our analysis of the Chao framework.

In shortage conditions Drèze assumes that all consumers get the same pro rata allocation q_c/q . Given that there is a single willingness to pay, there is no difference between pro rata allocation for all and complete loss for some. In this case we can allocate ex ante capacity for the purposes of charging $\bar{q} = \sum_{i=n} \bar{q}_i$. Being homogenous, all consumers have the same ex ante expectation and distribution of shock, and the same ex ante endogenous and exogenous mix of shock.

Also implicit is that the electricity is a public good and so the commitment by one consumer to pay a higher price does not assure provision of electricity. The implied arrangement is that the producer only provides the extra capacity if all consumers commit to a higher payment.

Drèze is not specific about the producer/consumer contractual relationship as he deals at the level of aggregate surplus. Nevertheless, it is consistent with his analysis to envisage a relationship. The consumers submit to the producer their unbiased ex ante probability distributions of demand. We assume no moral hazard. They then pay a capacity fee equal to the shadow cost of a statistical parameter that may, for example, be variance, which we shall examine. The producer then by agreement builds an amount that has a linear relationship to this parameter. Note that the payment of a capacity fee does engender an increase in capacity, but it does not guarantee delivery. In this respect it has some commonality with capacity obligations.

3.3.3 The Analysis

3.3.3.1 Core Equations

For N customers with normally distributed²² demand and standard linear correlation between them, we have

$$\sigma = \left(\sum_{i,j=1}^N \sigma_i \sigma_j \rho_{ij} \right)^{\frac{1}{2}}, \quad (3.9)$$

where σ_i , σ_j are the individual standard deviations and ρ_{ij} are the correlations.

$$\frac{\partial \sigma}{\partial \sigma_i} = -\frac{1}{2} \left(\sum_{i,j=1}^N \sigma_i \sigma_j \rho_{ij} \right)^{-\frac{1}{2}} \left(2\sigma_i + 2 \sum_{j=1, j \neq i}^N \rho_{ij} \sigma_j \right). \quad (3.10)$$

If $\rho_{ij} = 0$ for $i \neq j$ we have

$$\frac{\partial \sigma}{\partial \sigma_i} = -\frac{1}{2}(\sigma^2)^{-\frac{1}{2}}(2\sigma_i) = \frac{\sigma_i}{\sigma}. \quad (3.11)$$

For $\rho_{ij} = 1$ we have

$$\frac{\partial \sigma}{\partial \sigma_i} = -\frac{1}{2} \left(\sum_{i,j=1}^N \sigma_i \sigma_j \right)^{-\frac{1}{2}} \left(2 \sum_{j=1}^N \sigma_j \right)$$

If consumers are homogenous, we have $\sigma_i = \sigma_j$ for all i, j and hence

$$\frac{\partial \sigma}{\partial \sigma_i} = 1 \quad (3.12)$$

$$\frac{\partial \sigma}{\partial \sigma_i^2} = \frac{\partial}{\partial \sigma_i^2} \left(\sum_{i,j=1}^N \sigma_i \sigma_j \rho_{ij} \right)^{\frac{1}{2}}.$$

For $\rho_{ij} = 1$ we have

$$\frac{\partial \sigma}{\partial \sigma_i^2} = \frac{\partial}{\partial \sigma_i^2} (\sigma_i^2)^{\frac{1}{2}} = \frac{1}{2} (\sigma_i^2)^{-\frac{1}{2}} = \frac{1}{2} \frac{1}{\sigma_i}. \quad (3.13)$$

For $\rho_{ij} = 0$ we have

$$\sigma = \left(\sum_{i=1}^N \sigma_i^2 \right)^{\frac{1}{2}}$$

$$\frac{\partial \sigma}{\partial \sigma_i^2} = \frac{1}{2} \left(\sum_{i=1}^N \sigma_i^2 \right)^{-\frac{1}{2}} = \frac{\partial}{\partial \sigma_i^2} \sigma^2 = \frac{1}{2} \frac{1}{\sigma}. \quad (3.14)$$

3.3.4 Build and Pricing Criteria

3.3.4.1 Build in Proportion to Standard Deviation

3.3.4.1.i Cost Equations

$$f[q_c(\sigma, \bar{q}), q^*] = \beta[\bar{q} + k_1 \sigma] + \gamma q^* \quad (3.15)$$

$q_c = \bar{q} + k_1 \sigma$ is the installed capacity

$q^*(\bar{q}, \sigma, q_c, \varepsilon)$ is the quantity actually delivered and received by the consumers

ε the random term $\varepsilon = q^* - \bar{q}$. $q^* = \sum_{i=N} q_i^*$.

In this instance, we can construct the shadow costs to provide to individual consumers are as follows:

$\frac{\partial f}{\partial q_i^*} = \gamma$ is the (marginal) variable cost of energy

$$\frac{\partial f}{\partial q_i} = \beta \quad (3.16)$$

$$\frac{\partial f}{\partial \sigma_i} = \beta k_1 \frac{\partial \sigma}{\partial \sigma_i}.$$

If $\rho_{ij} = 0$ for $i \neq j$, then from equation (3.11) we have

$$\frac{\partial f}{\partial \sigma_i} = \beta k_1 \frac{\sigma_i}{\sigma}. \quad (3.17)$$

Substituting from equation (3.14), we have

$$\frac{\partial f}{\partial \sigma_i^2} = \beta k_1 \frac{\partial \sigma}{\partial \sigma_i^2} = \beta k_1 \frac{1}{2\sigma}. \quad (3.18)$$

3.3.4.1.ii Build in Proportion to Standard Deviation.

Price in Proportion to Standard Deviation

Pricing (as distinct to building) according to an ex ante probability distribution (as distinct to actual outturn) is an interesting proposition in its own right, particularly when we get to the level of individual consumer, as the only way to get an unbiased answer is to use the actuarial ex post history to date.

Now let us construct a trial marginal price function for the individual consumer from the sum of the shadow costs, assuming that these are all uncorrelated to each other. The three elements are: expected variable cost, expected capacity cost, and a charge for variability. If we use standard deviation rather than variance as a shadow cost, then, if we can ignore second-order terms and consider only small changes to consumer factors, each consumer i should be charged an (ex ante identical) expenditure E of

$$E_i(q_i^*, \bar{q}_i, \sigma_i) = q_i^* \frac{\partial f}{\partial q_i^*} + \bar{q}_i \frac{\partial f}{\partial \bar{q}_i} + \sigma_i \frac{\partial f}{\partial \sigma_i} = \gamma q_i^* + \beta \bar{q}_i + \beta k_1 \frac{\sigma_i^2}{\sigma}$$

where we used equation (3.11) for $\frac{\partial f}{\partial \sigma_i}$.

$\sigma_i \frac{\partial f}{\partial \sigma_i} = \beta k_1 \frac{\sigma_i^2}{\sigma}$ is dependent on the assumption that $\rho_{ij} = 0$ for $i \neq j$.

We will return to this below.

So consumers pay an up front capacity cost that depends both on expectation and variation.

From this we can compute the expectation of revenue for the utility from all customers. For zero correlation $\rho_{ij} = 0$ for $i \neq j$, so $\sum_{i=N}^1 \sigma_i^2 = \sigma^2$.

Our total expenditure by consumers is

$$\sum_{i=1}^N E_i(q_i^*, \bar{q}_i, \sigma_i) = \gamma \sum_{i=1}^N q_i^* + \beta \sum_{i=1}^N \bar{q}_i + \beta \frac{k_1}{\sigma} \sum_{i=1}^N \sigma_i^2 = \gamma^* q + \beta [\bar{q} + k_1 \sigma].$$

This is of course the costs as we saw in equation (3.15). It is no surprise that the utility breaks even on average. Drèze points out that this is the formula advocated by Boiteux (1951), Boiteux and Stasi (1952), Bessière (1961), and applied at EDF as shown in Boiteux (1957).

However, let us now consider for a given producer cost/consumer expenditure, the marginal price of substitution between capacity and standard deviation or variance, in comparison to the marginal cost of transformation between capacity and standard deviation or variance, according to our capacity build algorithm.

For transformation, using equations (3.16) and (3.17), we have

$$\frac{\partial f / \partial \sigma_i}{\partial f / \partial \bar{q}_i} = - \frac{d\bar{q}_i}{d\sigma_i} \Big|_{f=\text{const}} = \frac{\beta k_1 \sigma_i}{\sigma} / \beta = \frac{k_1 \sigma_i}{\sigma}$$

For substitution

$$\begin{aligned} & \frac{\partial E_i(q_i^*, \bar{q}_i, \sigma_i) / \partial \sigma_i}{\partial E_i(q_i^*, \bar{q}_i, \sigma_i) / \partial \bar{q}_i} \Big|_{E_i(q_i^*, \bar{q}_i, \sigma_i)=\text{const}} = - \frac{d\bar{q}_i}{d\sigma_i} \\ & = \beta k_1 (\lambda) \frac{\partial}{\partial \sigma_i} \left(\frac{\sigma_i^2}{\sigma} \right) / \beta \\ & = k_1 \left[\frac{2\sigma_i}{\sigma} + \sigma_i^2 \frac{\partial}{\partial \sigma_i} \frac{1}{\sigma} \right] \\ & = k_i \left[\frac{2\sigma_i}{\sigma} + \sigma_i^2 \left(-\frac{1}{2} \left(\sum_{i,j=1}^N \sigma_i \sigma_j \rho_{ij} \right)^{-\frac{3}{2}} \left(2\sigma_i + 2 \sum_{j=1, i \neq j}^N \sigma_j \rho_{ij} \right) \right) \right]. \end{aligned}$$

For $\rho_{ij} = 0$ for $i \neq j$

$$= k_1 \left[\frac{2\sigma_i}{\sigma} + \sigma_i^2 \left(-\frac{1}{2} \frac{1}{\sigma^3} 2\sigma_i \right) \right] = k_1 \left[\frac{2\sigma_i}{\sigma} + \frac{\sigma_i^3}{\sigma^3} \right].$$

Since $\sigma \gg \sigma_i$ and $\rho_{ij} = 0$ or equivalently $\frac{\partial \sigma}{\partial \sigma_i} \rightarrow 0$ we can ignore the second term, so

$$-\left. \frac{d\bar{q}_i}{d\sigma_i} \right|_{f=\text{const}} = \frac{2k_1\sigma_i}{\sigma}.$$

So, for the case $\rho_{ij} = 0$ for $i \neq j$ if we build and price according to the standard deviation regime, the marginal rate of transformation $\frac{k_1\sigma_i}{\sigma}$ of the producer is not equal to the marginal rate of substitution $\frac{2k_1\sigma_i}{\sigma}$ for the consumer, and hence the pricing regime is not efficient.

3.3.4.1.iii Build in Proportion to Standard Deviation.

Price in Proportion to Variance

The consumer expenditure is

$$E_i(q_i^*, \bar{q}_i, \sigma_i^2) = q_i^* \frac{\partial f}{\partial q_i^*} + \bar{q}_i \frac{\partial f}{\partial \bar{q}_i} + \sigma_i^2 \frac{\partial f}{\partial \sigma_i^2} = \gamma \bar{q}_i + \beta \bar{q}_i + \frac{\beta}{2} k_1 \frac{\sigma_i^2}{\sigma}.$$

So our utility revenue falls by $\frac{1}{2}\beta k_1 \sigma$ to

$$\begin{aligned} \sum_{i=1}^N E_i(q_i^*, \bar{q}_i, \sigma_i^2) &= \gamma \sum_{i=1}^N q_i^* + \beta \sum_{i=1}^N \bar{q}_i + \frac{\beta k(\lambda)}{2} \frac{1}{\sigma} \sum_{i=1}^N \sigma_i^2 \\ &= \gamma q + \beta \left[\bar{q} + \frac{k_1 \sigma}{2} \right]. \end{aligned}$$

We shortfall against fixed costs.

Let us examine the marginal rates of substitution

$$\frac{\partial E_i(q_i^*, \bar{q}_i, \sigma_i^2) / \partial \sigma_i^2}{\partial E_i(q_i^*, \bar{q}_i, \sigma_i^2) / \partial \bar{q}_i} = - \left. \frac{d\bar{q}_i}{d\sigma_i^2} \right|_{E_i(\bar{q}_i, q_i, \sigma_i) = \text{const}} = \frac{\partial}{\partial \sigma_i^2} \left(\frac{\beta}{2} k_1 \frac{\sigma_i^2}{\sigma} \right) \Big/ \beta$$

Making the same assumption on the incremental effect on aggregate standard deviation of an individual consumer, we have

$$-\frac{d\bar{q}_i}{d\sigma_i^2} \Big|_{E_i(\bar{q}_i^*, \sigma_i, \sigma_i) = \text{const}} = k_1 \left(\frac{1}{\sigma} + \frac{\partial}{\partial \sigma_i^2} \left(\frac{1}{\sigma} \right) \right) = k_1 \left(\frac{1}{\sigma} + \frac{\partial}{\partial \sigma_i^2} \left(\frac{1}{\sigma} \right) \right) \approx \frac{k_1}{2\sigma}.$$

Using equations (3.16) and (3.18) we have

$$\frac{\partial f / \partial \sigma_i^2}{\partial f / \partial \bar{q}_i} = -\frac{d\bar{q}_i}{d\sigma_i^2} \Big|_{f=\text{const}} = \frac{k_1}{2\sigma} \text{ for zero correlations.} \quad (3.19)$$

So the marginal rates of transformation and substitution are the same at $\frac{k_1}{2\sigma}$.

Therefore the variance pricing regime appears to be short-term efficient with respect to the standard deviation build regime.

So it appears that we must make a choice between equilibrium and efficiency at the margin. Drèze appears to favour efficiency at the margin.

Let us examine further to see if there is a build and pricing regime that is efficient at the margin and can sustain equilibrium.

3.3.4.2 Build in Proportion to Variance

$$f[\bar{q}_i(\sigma), q] = \beta[\bar{q} + k_2\sigma^2] + \gamma q^* \quad (3.20)$$

$$\begin{aligned} \frac{\partial f}{\partial \sigma_i^2} &= \beta k_2 \left(\frac{\partial}{\partial \sigma_i^2} \left(\sigma_i^2 + \sum_{j=N, j \neq i} \sigma_j^2 + \rho_{ij} \sigma_i \sigma_j \right) \right) \\ &= \beta k_2 \left(1 + \frac{\partial}{\partial \sigma_i^2} \left(\sum_{j=N, j \neq i} \rho_{ij} \sigma_i \sigma_j \right) \right). \end{aligned}$$

If $\rho_{ij} = 0$ for $i \neq j$ then

$$\frac{\partial f}{\partial \sigma_i^2} = \beta k_2 \quad (3.21)$$

$$\begin{aligned} \frac{\partial f}{\partial \sigma_i} &= \beta k_2 \frac{\partial}{\partial \sigma_i} \left\{ \sigma_i^2 + \sum_{j=1, j \neq i}^N \sigma_j^2 + \sigma_i \sum_{j=1, j \neq i}^N \sigma_j \rho_{ij} \right\} \\ &= \beta k_2 \left\{ 2\sigma_i + \sum_{j=1, j \neq i}^N \sigma_j \rho_{ij} \right\} \end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial \sigma_i} &= 2\sigma_i \beta k_2 \text{ if } \rho_{ij} = 0 \text{ for } i \neq j \\ \frac{\partial f}{\partial q} &= \beta.\end{aligned}\tag{3.22}$$

Now we again construct the trial marginal price from the sum of shadow costs. First, we consider pricing according to standard deviation.

3.3.4.2.i Build in Proportion to Variance.

Price in Proportion to Standard Deviation

We use the same method and find that we over-recover our costs by $\beta k \sigma^2$ and the marginal rates of substitution $4k_2$ and transformation $2k_2$ σ_i are different.

3.3.4.2.ii Build in Proportion to Variance.

Price According to Variance

Now let us price according to variance.

$$E_i(q_i^*, \bar{q}_i, \sigma_i^2) = q_i^* \frac{\partial f}{\partial q_i^*} + \bar{q}_i \frac{\partial f}{\partial \bar{q}_i} + \sigma_i^2 \frac{\partial f}{\partial \sigma_i^2} = \gamma q_i^* + \beta \bar{q}_i + \beta k_2 \sigma_i^2.$$

The utility revenue, for zero correlation, is

$$\gamma \sum_{i=1}^N q_i^* + \beta \sum_{i=1}^N \bar{q}_i + \beta k_2 \sum_{i=1}^N \sigma_i^2 = \gamma^* q + \beta [\bar{q} + k_2 \sigma^2].$$

We exactly recover our costs.

Let us now compare the marginal rates of transformation and substitution. For transformation, from equations (3.21) and (3.22),

$$\frac{\partial f / \partial \sigma_i^2}{\partial f / \partial \bar{q}_i} = - \left. \frac{d\bar{q}_i}{d\sigma_i^2} \right|_{f=\text{const}} = \frac{\beta k_2}{\beta} = k_2.$$

For substitution

$$\frac{\partial E_i(q_i^*, \bar{q}_i, \sigma_i^2) / \partial \sigma_i^2}{\partial E_i(q_i^*, \bar{q}_i, \sigma_i^2) / \partial \bar{q}_i} = - \left. \frac{d\bar{q}_i}{d\sigma_i^2} \right|_{E_i(q_i^*, \bar{q}_i, \sigma_i^2)=\text{const}} = \frac{\beta k_2}{\beta} = k_2.$$

So for variance build and pricing, we have cost equilibrium and equality of rates of substitution and transformation k_2 for a homogeneous set of consumers with no correlation.

Drèze did not examine this solution.

3.3.4.3 Further Consideration of Correlation

3.3.4.3.i Perfect Correlation—Build and Price According to Standard Deviation

A homogenous set of consumers with zero demand correlation can only be rationalized by envisaging a set of purely endogenous private events such as birthdays, as distinct to exogenous events, such as rainy days. It is as easy to imagine perfect correlation of demand.

Repeating the analysis above for this case for $\psi = \sigma$ (i.e., build according to standard deviation) and price according to standard deviation, we again have cost equilibrium and marginal rates of transformation and substitution equal at k_1 .

On reflection, the results should not be a surprise, as for homogenous consumers with zero correlation we have $\sigma^2 = N\sigma_i^2$ and for perfect correlation we have $\sigma = N\sigma_i$. Both cases boil down in effect to the treatment of the consumer community as a single consumer.

For identical consumers with normally distributed and imperfectly and linearly correlated demands, the bounds of aggregate variance are $\sqrt{N}\sigma_i^2 < \sigma^2 < N\sigma_i^2$. Hence for correlations less than perfect, for a particular consumer price function and a particular ration rate, the producer experiences what Drèze calls increased economic returns to scale, as the percentage of capacity held above expectation falls with increasing numbers of consumers. With our assumptions about the identical nature of consumers, we have been able to adjust our pricing, without worrying about the discriminatory impact on different customers. With heterogeneity of customers, this becomes more of a problem.

3.3.4.3.ii Correlation at Higher Moments

When considering security of supply of a highly secure system (as distinct to a less secure system), it is in fact not the “normal” variation of demand that is most important, it is the low-frequency high-impact events. The frequency and magnitude of these can be very sensitive both to the structure of individual demand distributions and correlations between individuals. There are several effects here

1. For low correlations, the central limit theory has the effect of attenuating the tail of the aggregate distribution, driving it toward the normal.
2. Standard normal distributions are not necessarily correlated in a linear manner.
3. Endogenous shocks are by definition uncorrelated and the correlation between exogenous shocks tends to increase with shock size, thereby “fattening²³ the tail²⁴ of the exogenous distributions.²⁵

4. Regardless of correlation, the extreme value theory and associated distribution tends to model well the tail of the aggregate distribution.

Broadly speaking, the net of all this is that the correlation is not standard linear and increases with the moment of the distribution and the size of the event. If rationing is inefficient, the demand function downward sloping, and convex with a high willingness to pay at the ordinate, then we must treat all consumers as if they have perfect correlation of demand. The build is then $\bar{q} + \sum_{k=1}^K k_k \sigma^k$, where K is some number of around five. Low correlation can be used for the low moments and high correlations for the high ones. We explore this further in the analysis of Chao.

3.3.4.3.iii Disaggregation of the Consumer Base

Drèze goes on to examine the potential to divide the consumer base. We can do these by taking our demand variation equation and separating into correlated and uncorrelated elements.

We discuss in the Chao analysis in section 3.11 how we can usefully regard our consumer population in terms of the weighted average of correlated and uncorrelated elements. So the forces on consumers are a result of purely endogenous and purely exogenous forces. The correlated component can be simply expressed as

$$\sigma^2 = \sum_{i=n} \sigma_i^2 + \sum_{i,j=n,i \neq j} \sigma_i \sigma_j \rho_{ij}.$$

Similarly, in the spirit of Modern Portfolio Theory, we can consider the correlation between the individual and the whole portfolio of individuals. Indeed we can now divide our consumer demand into 100 percent correlated and 0 percent correlated.

$\sigma^2 = \sigma_i^2 + \sigma_p^2 + 2\sigma_i \sigma_p \rho_{ip} = \sigma_i^2 + \sigma_p^2 + 2\sigma_i \beta_{ip}$ where the β in σ_p denotes portfolio, where we have used the standard nomenclature for portfolio beta.

So the impact on portfolio variance is the sum of the correlated and uncorrelated impacts.

We can regard the portfolio as a mixture of perfectly correlated and uncorrelated impacts. In simple terms then, the aggregate cost of risk will be driven by $\psi = a^* \sigma + b^* \sigma^2$, where a and b are constants.

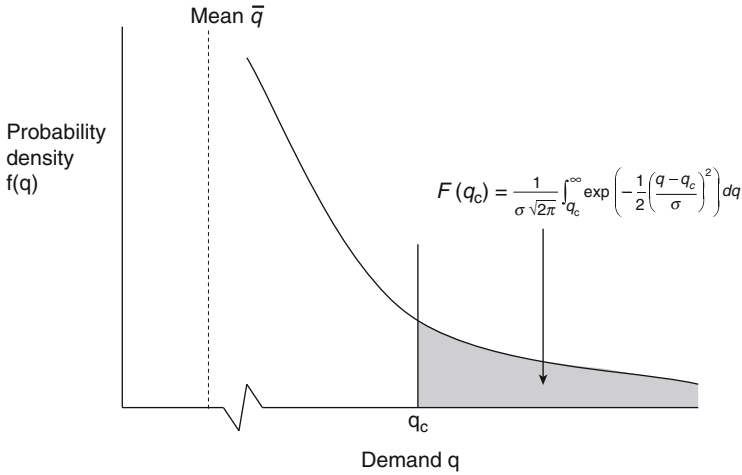


Figure 3.7 Probability of demand exceeding capacity. Normal distribution assumption.

We will see later, that since we are most interested in very low probabilities of insufficiency (i.e., high capacity margins), we can use the extreme value theory to show that for probability extremes, the normal distribution assumption is not necessary (although of course we need to characterize our distribution so that we can quantify it). The attention to the extremes is shown in figure 3.7.

3.3.4.4 Cost of Risk

To simplify the situation and address the key point, Drèze uses a highly simplified demand function and shock to it. In particular, he implicitly assumes an amount of lost load that is directly proportional to the coefficient of the statistical function ψ . If our distribution of aggregate demand is normal, then for high probabilities of lost load, with low capacity margin, the linearity is reasonable as seen in figure 3.8.

The fit is however not good for probabilities of lost load in the range that we are most interested in, as we see in figure 3.9.

This is immediately obvious to us if we consider the $\exp(-(x-\mu)^2/\sigma^2)$ form of the normal distribution,²⁶ and then consider the expansion of the exponential function $e^x = \sum_{n=1}^{\infty} \frac{x^n}{n!}$, which for lower and lower probabilities of lost load, higher and higher moments of distribution become more important.

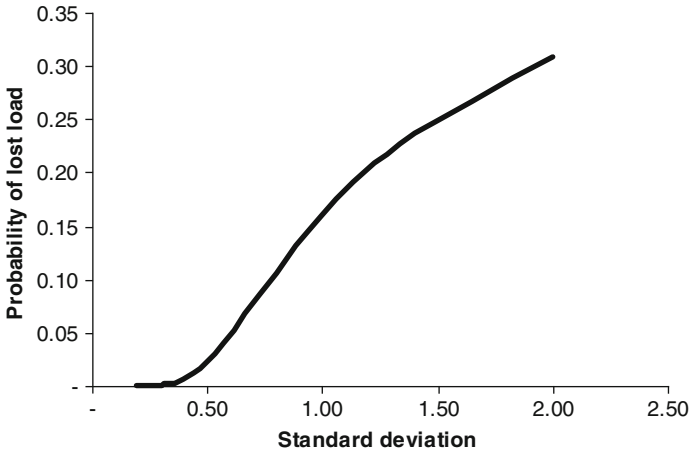


Figure 3.8 Approximate linearity of probability of lost load in relation to standard deviation—for high probabilities of lost load. Standard normal distribution. Capacity = 1. Ex ante expectation of demand = 0.25.

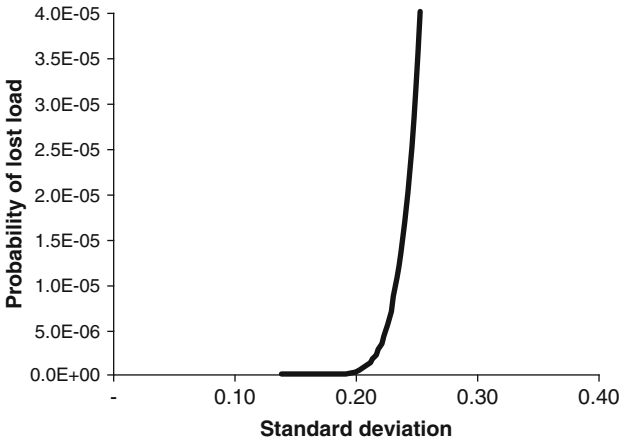


Figure 3.9 Nonlinearity of probability of lost load in relation to standard deviation—for low probabilities of lost load.

3.3.5 Demand Function and Rationing in the Drèze Framework

The Drèze analysis uses a demand function, that although stylized and still in common practical use, and sufficient to make his key points, is not realistic. To examine capacity in the context of a downward

sloping demand function, which is not only much more realistic, but very important when we consider DSM, we turn to Steiner.

3.3.6 Conclusions from the Drèze Analysis

Drèze sets efficiency at the margin as the primary objective by requiring the equation of transformation and substitution. He is prepared to sacrifice cost equilibrium to do so, allowing the generator to lose money on a continuous basis. Since we are able to show that within his analytic framework, we can satisfy equilibrium and efficiency at the margin, the real key point about his analysis is not the calculations but the preparedness to sacrifice equilibrium. It is particularly interesting to view this in the Walrasian framework (Drèze being a key author in the school called neo- or non-Walrasian) as Walras attended both the equilibrium and efficiency at the margin, placing more emphasis on equilibrium than the marginalists.

The analytics of standard deviation and variance are not really critical to the analysis, as in his solutions the simplifying assumptions on correlation render much of the workings redundant in making his key point. However these workings are very useful for various other purposes, such as the consideration of endogenous and exogenous shocks, the relationship between individual and aggregate demand, the application of the central limit theory, and the relationship between a statistical function such as the standard deviation of a normal distribution and the probability and extent of lost load.

3.4 ELASTIC DEMAND FOR CAPACITY— THE STEINER FRAMEWORK

3.4.1 Introduction

The demand function that is right angled (constant willingness to pay up to a level q and thence 0), with a lateral stochastic shock (same WTP but changing q) was adequate in the Drèze analysis to explain the reconciliation between equilibrium and the margin. It is however significantly inadequate planning for electricity. The first change that we must make is the addition of demand elasticity. This is important in particular for

1. allocation of costs to subperiods
2. consideration of rationing when capacity is less than demand at a given willingness to pay
3. the nature of stochastic shock.

Here we attend to cost allocation to subperiods.

Steiner²⁷ attended to the problem that charging the highest price in the peaks to consumers with elastic demand might actually change the timing of the peak.

Here we also introduce the concept of demand for capacity as distinct to demand for energy, which is so important in the design of capacity obligations.

3.4.2 Framework

Steiner considered the following model:

1. There is a single technology.
2. Short-term producer returns to scale are constant.
3. Short-term producer costs are nominally regarded as fuel²⁸ with cost b/MWh .
4. Long-term producer returns to scale are constant (i.e., both energy and capacity).
5. Long-term capacity cost is nominally regarded as equipment/capital with amortized cost $\frac{1}{2} \beta/\text{MW/h}$, where hours are measured in elapsed time.²⁹
6. The demand function is elastic, deterministic, and periodic (peak and off-peak).
7. The demand functions are known to the regulator.
8. The peak and off-peak durations are equal.
9. Demand functions in each period are independent (cross-elasticity is zero).
10. Demand is assured in the long term and nontrending.
11. The system is an autarky.

3.4.3 Analysis

Although the periodicity is regular and continuous as shown in figure 3.10, we can model this by a single period with a peak subperiod and an off-peak subperiod.

If we charge all the capacity in the peak period, then this may reduce demand below that in the off-peak, and thence the peak period shifts as we see in figure 3.11.

It is clear then that off-peak period 1 must make some contribution to capacity. Steiner solves this by adding together vertically³⁰ the two demand functions above the intersection with the horizontal line of

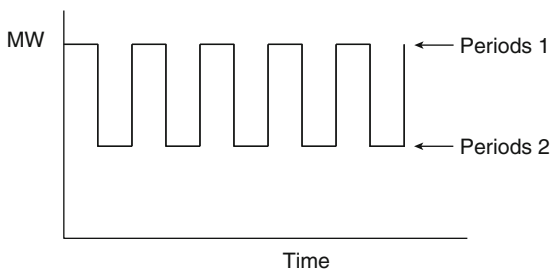


Figure 3.10 Periodic demand function as described by Steiner.

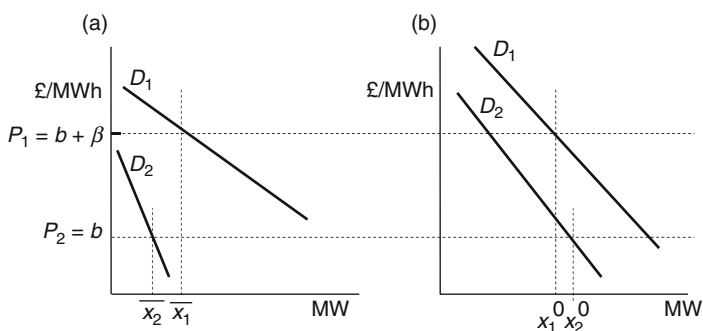


Figure 3.11 Demonstration of “shifting peaks” (a) All capacity is paid in the peak period, and peak demand remains above off-peak demand (b) Case where loading all capacity cost on the peak causes peak demand to fall below off-peak demand, and thereby shift the peak.

height b , as we can see in the figure below. The method is shown in figure 3.12 and is as follows:

1. Find the intersection of the aggregate demand for capacity, and the total cost $b + \beta$. This is the demand for both peak and off-peak. Here β is expressed in £/MWh paid over a fraction of a cycle. So if we must cover £1 for 1MW over a cycle of two periods, then $\beta = \text{£}2/\text{MWh}$. If β' is the cost over the whole cycle of unit length and w the period length, then $w * \beta = \beta'$.
2. Find the intersections between this demand and the demand curves. These are the prices.

We can see in figure 3.13 that total costs are exactly recovered, since the capacity cost saving in the peaks is exactly compensated by the capacity cost increase in the off-peaks. Indeed a geometric approach

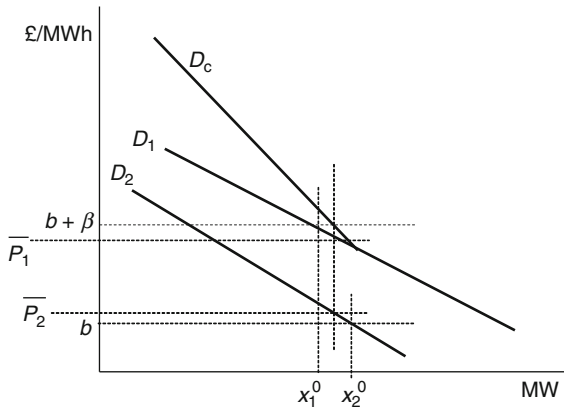


Figure 3.12 Steiner optimal pricing in the “shifting peaks” case. See text.

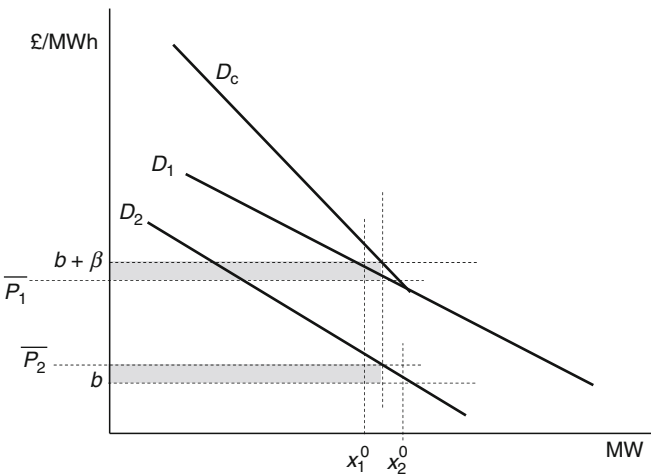


Figure 3.13 Exact recovery of total costs. The two gray areas are equal.

is very useful for consideration of much of the literature on capacity pricing.

We can extend the same argument to more periods. The three-period example for shifting peaks is shown in figure 3.14. We add the functions from their intersection with b , in ascending order.

Williamson (1966a, 1966b) extends the analysis to different lengths of peak and off-peak, and again considers the whole cycle as the appropriate period over which to consider marginal costs. While not changing the cost lines, he changes the demand for capacity by applying a

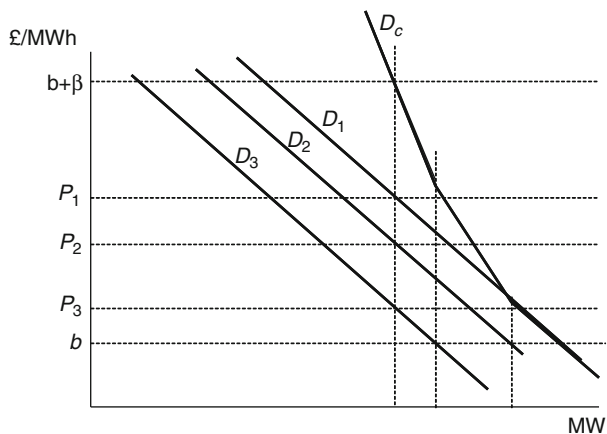


Figure 3.14 Optimal pricing for a three-period example with shifting peaks.

weighting to the demand for capacity, with the weight equal to the length of the respective demand period.

Let the demand in any period i be D_i and the period fraction of the cycle be w_i . Taking each period in isolation, and assuming that there is no demand in any other period, the price in that period that will give fixed-cost recovery over the whole period is $P_i = b + \beta/w_i$. If Q_i is the capacity demanded in any period, then the whole-cycle revenue recovered in each period is $P_i Q_i w_i$, which is equal to the costs $Q_i (b w_i + \beta)$. We now need to construct a demand for capacity curve such that at provided capacity Q_i^* , it has a price $P_i = b + \beta/w_i$. The aggregate demand for capacity curve is now formed from the time-weighted individual demand curves.

Let us suppose that we have an off-peak demand function D_1 for 8 hours and a peak demand function D_2 for 16 hours. The Williamson framework is represented in figures 3.15 and 3.16.

In the case for which both periods consume the same MW, the equation of costs and revenues give us

$$P_1 Q_1^* w_1 + P_2 Q_1^* w_2 = b Q_1^* + \beta Q_1^*$$

where

Q_1^* is the provided capacity

P_1 and P_2 are the prices charged in period 1 and period 2

w_1 and w_2 are the cycle fractions of periods 1 and 2

β is the fixed per cycle cost

b is the marginal cost.

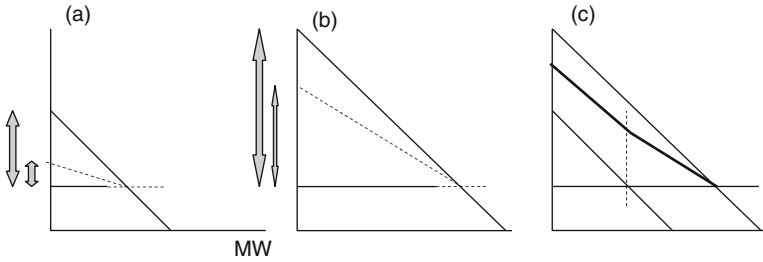


Figure 3.15 Optimal pricing in the Williamson framework with different durations of peak and off-peak (a) Off- peak. The demand curve, and the differential between the demand curve and marginal cost, multiplied by period length (b) Peak (c) Vertical addition of the demands for capacity (the dotted lines above the level b) in previous two figures.

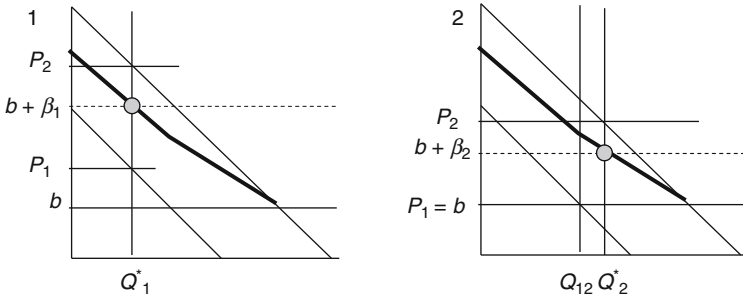


Figure 3.16 Pricing and capacity for two different levels of fixed cost (1) Example 1, consumption Q_1^* in both periods (2) Example 2, different consumption in each period.

We can rearrange as

$$P_1 = b + \frac{\beta}{w_1} - \frac{w_2}{w_1}(P_2 - b).$$

In the case where consumption is different in each period, the revenue is

$$Q_2^* P_2 w_2 + Q_1 P_1 w_1 = Q_2^* (\beta + w_2 b) + Q_1 b w_1.$$

Again, costs equal revenues.

3.4.4 Discussion of the Steiner Shifting Peaks Analysis

Steiner's starting point is a peak load rather than variable cost approach as is evident from the axiomatic recovery of fixed costs and

the reduction in consumption from the elevation of the price in the off-peak period above short run variable costs b does reduce off-peak consumption by $(P_2 - b)/\theta_2$, where θ is the slope of the inverse demand function. The variable cost approach would avoid the deadweight loss of $(P_2 - b)^2/2\theta_2$.

The deduction of the variable cost from the inverse demand function is a “demand for capacity” approach is explained by Hirshleifer in section 4.2.

The argument can be framed as a public goods argument in which the total capacity is available across all periods, and all periods may have to pay a contribution to capacity cost. We see in section 4.3 a similar approach by Panzar.

Steiner made some emphasis that his analysis introduced the importance of discriminatory pricing to the subject of peak pricing. However, it was not agreed³¹ that this pricing is actually discriminatory. The debate hinges on the definition of discriminatory for electricity, and in particular whether peak and off-peak electricity are different commodities, and whether therefore their costs are linked. We noted in section 2.4.6 that fixed cost allocation is influenced by revenue allocation.

Note that while we have not necessarily required the assessment of a utility function, we nevertheless required a fairly complete knowledge of the demand function. In theory, a demand function can be discovered by offering at different prices and finding how much is demanded, but in practice this knowledge is restricted to demand volumes close to the equilibrium volumes and we do not test willingness to pay in relation to actual loss of load.

3.5 EFFICIENCY OF RATIONING OF DEMAND

If the amount of goods produced is less than the amount demanded, then there must be some rationing. Since electricity has some public goods characteristics and because the key focus for capacity mechanisms is loss of supply, we need to examine rationing in detail.

Consider the initial deterministic situation in which each consumer consumes amount determined by their demand function and the market price P . Clearly this is not a public good situation.

Suppose now that demand suddenly rises. If we may elevate the price and in addition have sufficient time resolution on the meters to measure consumption in the period of demand elevation, again we

have a private good situation and we can achieve balance by elevating the price to the level at which aggregate demand matches supply.

Suppose, however, for practical reasons (metering, billing, collection) or social/political reasons, we may not elevate the price. Our demand now exceeds the capacity and something must happen.

One thing that can happen is to terminate power flow to whole regions in turn for a four-hour period. This “rota disconnection” is what is generally done in sustained power shortages. While in theory the concept of public goods is violated by there being a restriction, this is not a real concern in theory. We call this random rationing.

A common precursor to rota disconnection is the reduction of system voltage. In this case all consumers receive a pro rata reduction to their normal use.

With smart meters there are many more possibilities.

For example, the price can be elevated for halfhourly periods. This is efficient pricing to the extent that the consumers can respond to the price signal, for example, by automation and smart devices.

It is possible to preserve power to some individuals and disconnect everyone else at individual meter level. Even if the meter can do no more than measure, it is theoretically possible to force self-disconnection for the shortage period by sending an exceptionally high price to the meter for that period. Supposing that for social reasons power was preserved only for the poorest, we can see that we can have perfectly inefficient rationing in terms of first best welfare.

There are various other ways of rationing in a smart system. For example, smart devices can ensure that at consumer level the high-value uses (e.g., telecommunication) are preserved and the low-value uses (e.g., heating) are constrained.

Finally we need to consider when the economic efficiency in ration conditions is different to that in normal conditions. The main mechanism for this is the inability to effect a dynamic price.

An interesting situation is when we consider the same shift in aggregate demand function, but in one case arising from increased willingness to pay and in the other case an increased volume of demand. In the former case, if rationing is inefficient, increased willingness to pay will result in reduced consumption by those already consuming due to the sharing with those who are newly prepared to consume at this price. In the latter case, we would more naturally expect a widespread sharing across consumers. We examine this situation in detail in the Visscher analysis in section 3.7.

We now consider the effect of different rationing methods on efficiency.

The key methods both between consumers and by individual consumers are:

1. perfectly efficient
2. perfectly inefficient
3. random
4. pro rata
5. other.

In the most efficient rationing of demand, all those whose willingness to pay exceeds the clearing price receive the goods, and all those whose WTP does not, do not.

For pro rata rationing, in which all consumers lose the same proportion of demand, we must obviously start with how much they have before rationing. Pro rata rationing is in fact the standard method for the early stages of controlled lost load. The distribution companies reduce power draw by reducing the voltage.

To model rationing efficiency, we need to look at the level of the population (rationing some more than others) and the level of the consumer (rationing some uses more than others). In practice the literature simplifies this and we do the same.

Our characterization of consumers is:

1. homogenous with only exogenous shocks to demand
2. homogenous with only endogenous shocks to demand
3. heterogeneous
4. other.

The distribution of consumers is

1. normal distribution of the key variable (e.g., WTP at given volume q , demand at a given price, population)
2. lognormal distribution of the key variable
3. uniform distribution of the key variable
4. pyramidal (few rich, many poor)
5. other.

The main demand functions are as set out in section 2.2.1.

For the sake of easier modeling, we assume initially that each consumer has a constant willingness to pay (linear utility) up to a volume limit at which willingness to pay at the margin drops to zero. For heterogeneous consumers we then assume that the WTP for different individual consumers has a uniform distribution.

When considering the rationing of demand, it is common to model in terms of the aggregate demand function.

Figure 3.17 shows the efficiency of random rationing in relation to the efficient rationing.

Figure 3.18 shows the comparison between the most and least efficient rationing. Note that the welfare loss for efficient rationing

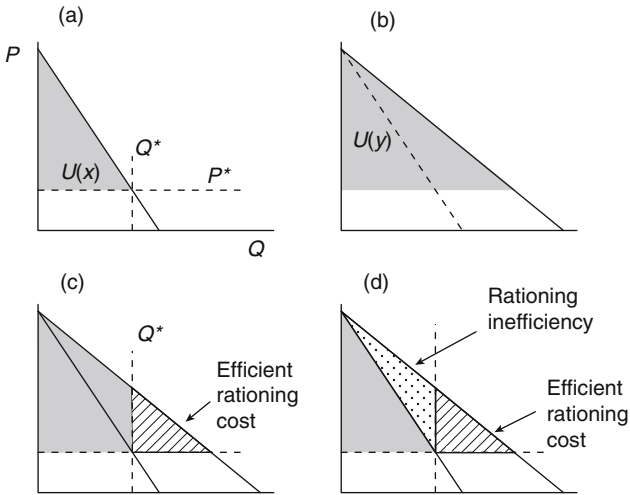


Figure 3.17 Comparison of random rationing to efficient rationing (a) Initial situation, with surplus shaded (b) Change to demand with demand satisfied (c) Change to demand with efficient rationing of unsatisfied demand (d) The same change to demand but showing the inefficiency of pro rata demand.

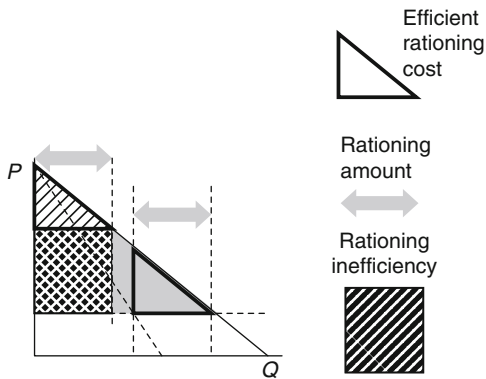


Figure 3.18 Most inefficient rationing. Serving the shaded area first. The inefficiency cost is the chequered area.

is proportional to the square of the volume lost, while the initial cost of rationing of the least efficient rationing is proportional to the volume loss.

As is pointed out by Visscher and others, the least efficient (as defined by Marshallian surplus) is quite possible. For example, suppose that the “cash rich” are also the “time poor” and energy use can be configured by spending time on it. Another reason is the social imposition, for example, all citizens receiving a certain amount of electricity, and all extra being rationed, or vulnerable consumers being prioritized for supply.

Different authors make different assumptions, which we summarize in figure 3.19. Not shown is the Drèze assumption, where since there is a single willingness to pay, the loss of welfare is directly proportional to loss of load.

For the most public good we must consider rationing even in deterministic conditions when aggregate capacity is equal to the aggregate demand at price P where P is the fully loaded (fixed plus variable) cost of generation. At aggregate level it is easiest to assume random rationing (i.e., some consumers lose everything). We see this in figure 3.20. Note that we do require demand to be finite even at zero cost.

It is in fact common to assume that the starting position is efficient and then a stochastic upward shock to demand may cause some inefficiency. This is examined in the Visscher framework in section 3.7.

Public goods are defined differently by different authors.³² Under the Samuelson³³ definition, the rationing of public goods is pro rata.

When we consider rationing in practice, it is important to recognize all of price elasticity of demand, the heterogeneity of consumers, the heterogeneity of the type of demand (heating, lighting, etc.) and

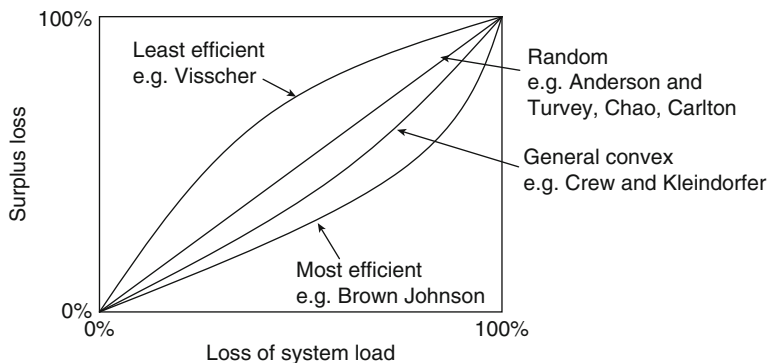


Figure 3.19 Summary of rationing efficiencies.

the difference between exogenous demand shocks that are systemic and endogenous demand shocks that are individual.

It is the heterogeneity of the use of electricity by individuals that makes all or nothing random rationing highly inefficient, and for other types of rationing, it becomes highly dependent on the consumers' ability to ration effectively, for example, maintaining computing while reducing heating. This itself is highly dependent on local storage of energy (mainly power, heat, and cold), automation, and social and technical ability to use price signals as a rationing tool. We see this in figure 3.21.

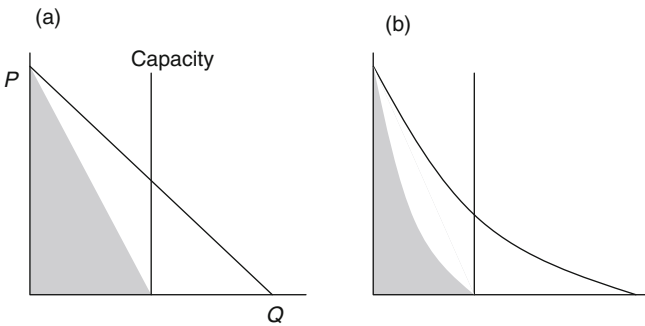


Figure 3.20 Purely public good with capacity less than demand. Shaded area shows consumers' surplus (a) Linear demand function (b) Convex demand function.

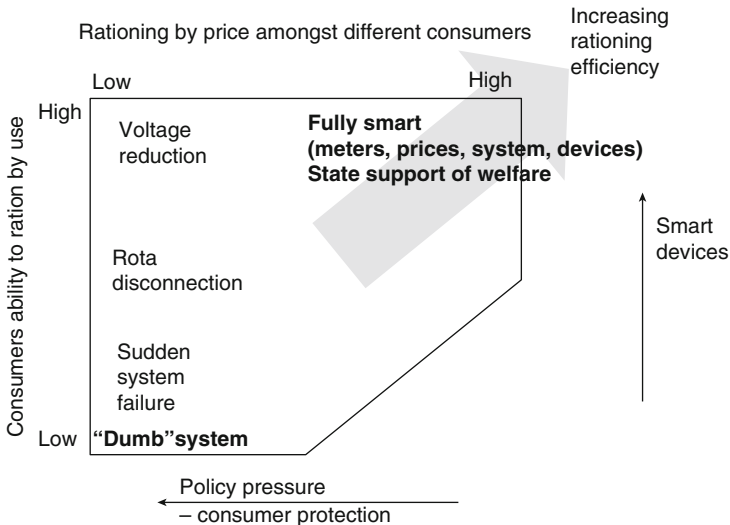


Figure 3.21 Depicting of rationing efficiency between consumers and by consumers.

3.6 PRICING UNDER STOCHASTIC DEMAND— THE BROWN AND JOHNSON FRAMEWORK

3.6.1 Introduction

As we have noted, the axis of the debate between peak load and marginal cost pricing revolves around the uplift of fixed costs in prices.

The Brown and Johnson (BJ, 1969) framework and analysis provides key insights into the debate as it stood.

BJ arrive at the conclusion that for deterministic demand, prices are fully loaded ($P = b + \beta$) and fixed costs are recovered, but for stochastic demand, prices are set at short run variable costs ($P = b$), and fixed costs (β) are not recovered. This is an important conclusion that is frequently quoted in the literature.³⁴

We examine their analysis here, particularly so that we can consider how the specific decision framework can be generalized and to find the sensitivity of the conclusions to the assumptions.

3.6.2 Framework

1. There is one technology.
2. There is a single peak and an off-peak subperiod that may be of different lengths.
3. Returns to scale in capacity β and operation b are constant.
4. There is a hard capacity constraint (variable costs infinite above it).
5. The inverse demand function is downward sloping (elastic).
6. Demand is stochastic with a linear symmetrical shock.
7. Not noted by BJ, but the implication of the shock is that the demand function is linear.
8. Rationing is efficient. We can do this either by changing the price to allow consumers to self-ration, or through knowledge of the individual demand functions (Kaldor-Hicks efficiency).

3.6.3 The Analysis

3.6.3.1.i Optimum Size in the One Subperiod Deterministic Setting
The welfare equation used to optimize capacity Q is

$$W = \int_0^{Q'} X^{-1}(Q) dQ - (b + \beta)Q,$$

where $X^{-1}(Q)$ is the inverse demand function.

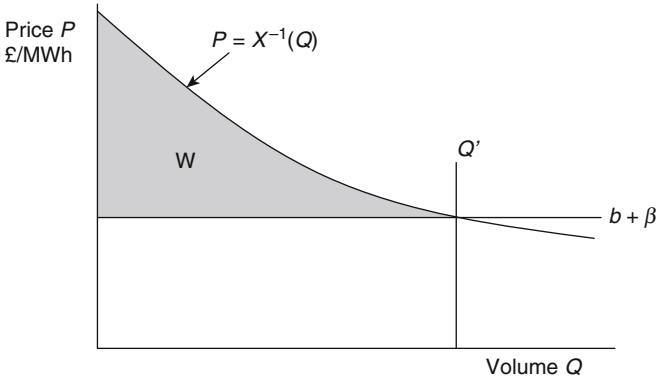


Figure 3.22 The inverse demand function and consumers' surplus W .

We can regard this either as the total surplus, being gross consumers surplus minus producer costs, or net consumers' surplus after paying the set price $b + \beta$. Figure 3.22 shows the net surplus.

3.6.3.1.ii The Two Subperiod Deterministic Setting

BJ then extends the analysis to the Williamson framework (see section 3.4) with a peak and off-peak periods. So we have

$$W = w_1 \int_0^{Q_1'} X^{-1}(Q_1) dQ_1 + w_2 \int_0^{Q_2'} X^{-1}(Q_2) dQ_2 - w_1 b Q_1' - w_1 b Q_2' - \beta Q_2',$$

where period 2 is the peak period³⁵ of length w_2 . Period 1 is the off-peak, of length w_1 . $w_1 + w_2 = 1$.

In the case where demand elasticity is such that demand in peak and off-peak period is the same (the shifting peaks that we see in section 3.8), that is, $Q = Q_1' = Q_2'$, then

$$\frac{\partial W}{\partial Q} = P_1^* w_1 + P_2^* w_2 - (b + \beta) = 0$$

We know from the Williamson framework that in this case $P_1^* = b + x$, so $P_1^* = P_2^* = b + \beta - x$, where x is as shown in the Steiner/Williamson framework in section 3.4, and hence fixed costs are exactly recovered.

Note that in the one-period case, grouping together the cost terms b and β , we assume that the incurrence of short-term variable costs

necessarily incurs the long-term capacity costs. In other words, the owner makes the capacity and operating decision simultaneously at build time. We optimize aggregate welfare at build time and we do not reoptimize at run time. In the Hotelling framework, we would indeed reoptimize, and then set the clearing price at $P = b$. This is outlined in the discussion on Steiner, in section 3.4. The subtlety here is that BJ not only assume a deterministic demand function, but also deterministic demand satisfaction.

BJ follow the same line for the case of different demand in both periods (one demand function intersects the production cost function on its vertical section, the other on the horizontal section), so for off-peak period 1 we have $P_1 = b$ and for peak period 2 we have $P_2 = b + \frac{\beta}{w_2}$. Again, fixed costs are exactly recovered.

3.6.3.2 Stochastic Demand

BJ then consider a single period with stochastic demand.

First we consider a well-behaved symmetrical function that is additive to demand. $D = X(P) + u$, where D is the quantity demanded at price P , and u a stochastic variable. In the Von Neumann Morgenstern framework, this is a simple shock to commodity endowment.

The firm must make the price \bar{P} and capacity Z decision before the resolution of the uncertainty u . We assume that we can ration by willingness to pay and hence rationing is efficient.

3.6.3.2.i Welfare

BJ then construct total welfare as for the deterministic case, but this time use (unconditional) expectations. So,

$$\text{Total welfare } W = E\{\text{GCS}\} - E\{\text{PC}\}, \quad (3.23)$$

where GCS is the gross consumer surplus and PC is total producer cost. If now we have to ration, then BJ express the welfare as:

Expectation of net consumers' surplus if there were no rationing
 Plus expectation of revenue paid by consumer to producer if there were no rationing
 Minus expectation of the loss of net consumers' surplus from rationing if there is rationing
 Minus expectation of the loss of producer revenue from rationing
 Minus expectation of variable costs

Minus fixed capacity costs.

So, to optimize, we will set $\frac{\partial W}{\partial P} = 0 \quad \frac{\partial W}{\partial Z} = 0$.

Let us consider the terms one by one:

$$E\{GCS_{nr}\} = \int_{-\infty}^{\infty} f(u) \left[\int_{\bar{P}}^{X^{-1}(-u)} [X(P) + u] dP + \bar{P}(X(\bar{P}) + u) \right] du,$$

where the suffix *nr* denotes no rationing

If *u* is symmetrically distributed, then $\int_{-\infty}^{\infty} \bar{P}(X(\bar{P}) + u) du = \bar{P}X(\bar{P})$.

So,

$$E\{GCS_{nr}\} = \int_{-\infty}^{\infty} f(u) \int_{\bar{P}}^{X^{-1}(-u)} [X(P) + u] dP du + \bar{P}X(\bar{P}) \quad (3.24)$$

The build up of the integral is shown in figure 3.23.

The loss of consumer surplus after a positive demand shock such that volume requested at price *P* exceeds capacity *Z*, relative to what that surplus would have been if the capacity had increased such that demand is exactly satisfied, is shown in figure 3.24

Adding together we have,

$$\begin{aligned} E\{GCS\} &= \int_{-\infty}^{\infty} f(u) \int_{\bar{P}}^{X^{-1}(-u)} [X(p) + u] dP du + \bar{P}X(\bar{P}) \\ &\quad - \int_{Z-X(\bar{P})}^{\infty} f(u) P [X(P) + u - Z] dP du \quad . \\ &\quad - \int_{Z-X(\bar{P})}^{\infty} f(u) P [X(P) + u - Z] du \end{aligned}$$

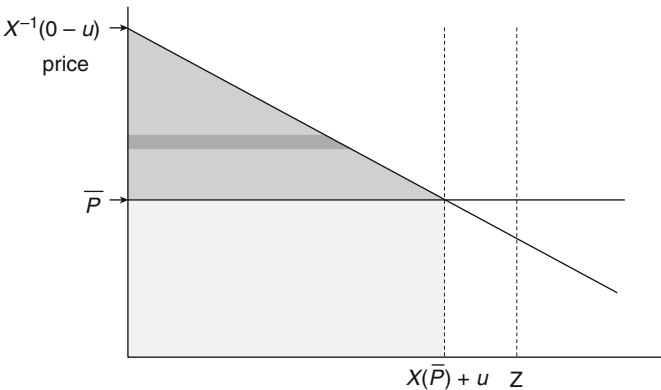


Figure 3.23 Gross consumer surplus when capacity is sufficient, showing construction of the integral.

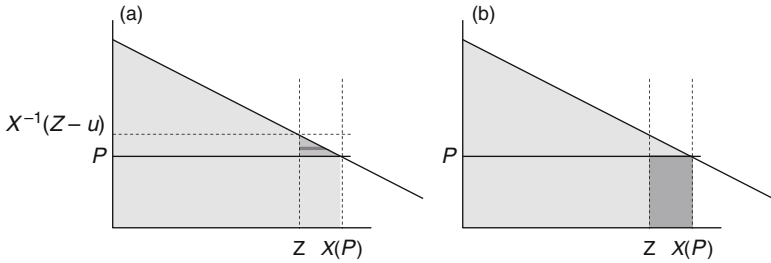


Figure 3.24 Loss of consumer surplus after a shocked demand exceeding capacity, relative to the surplus at the ideal capacity level (a) Net consumer surplus (b) Consumer cost. The addition of the two areas is the gross consumer surplus differential from this shock.

The expectation of sales volume is simply the expectation without constraint, minus the expectation of constraint. So,

$$E\{\text{Sales}\} = \int_{u=-\infty}^{\infty} f(u)[X(\bar{P}) + u] du - \int_{Z-X(\bar{P})}^{\infty} f(u)[u - (Z - X(P))] du. \tag{3.25}$$

This gives us the expected variable costs $b * E\{\text{Sales}\}$ plus $\beta * Z$ and the expected producer revenue $b * P$. So the expected producer costs are”

$$E\{\text{PC}\} = b \left\{ \begin{aligned} & X(P) - \int_{Z-X(P)}^{\infty} uf(u) du \\ & + [Z - X(P)] \int_{Z-X(P)}^{\infty} f(u) du \end{aligned} \right\} - \beta Z. \tag{3.26}$$

The total welfare expectation, which we will need to differentiate to arrive at the optimum price and capacity, is the net of the expectations of costs and gross consumer surplus.

$$\begin{aligned} W = & \int_{-\infty}^{+\infty} f(u) \left[\int_{\bar{P}}^{X^{-1}(-u)} [X(P) + u] dP \right] du + \bar{P}X(\bar{P}) \\ & - \int_{Z-X(\bar{P})}^{\infty} f(u) \left\{ \int_P^{X^{-1}(Z-u)} [X(P) + u - Z] dP + P[X(P) + u - Z] \right\} du. \\ & - b \left\{ X(P) - \int_{Z-X(P)}^{\infty} uf(u) du + [Z - X(P)] \int_{Z-X(P)}^{\infty} f(u) du \right\} - \beta Z \end{aligned}$$

3.6.3.2.ii Optimizing Price

$$\begin{aligned} \frac{dW}{d\bar{P}} &= \bar{P}X'(\bar{P})F[Z - X(\bar{P})] \\ &\quad - bX'(\bar{P})F[Z - X(\bar{P})] - \beta \frac{\partial Z}{\partial \bar{P}} = 0. \end{aligned} \quad (3.27)$$

BJ omit the $\partial Z/\partial P$ term, effectively assuming that the capacity decision is taken before the pricing decision. In practice we would not necessarily expect this to be the case.

We can see that if $\frac{\partial Z}{\partial \bar{P}} = 0$, then $\bar{P} = b$.

However if $\frac{\partial Z}{\partial \bar{P}} = X'(\bar{P})F[Z - X(\bar{P})]$ then $\bar{P} = b + \beta$. (3.28)

$F[Z - X(\bar{P})]$ is the probability that the capacity exceeds the volume demanded at price \bar{P} (i.e., the probability of satisfaction of demand with no rationing) and $X'(\bar{P})$ is the slope of the demand function at price \bar{P} .

As before in the deterministic case, this depends on the capacity decision. If we optimize capacity and price at the same time, then the optimum price is the fully loaded cost. If we optimize aggregate welfare after the capacity decision and the resolution of uncertainty, then the optimum price is the variable cost. This is precisely the peak load versus marginal cost dilemma. With the former, we make our investment plans with sight of the risks. With the latter we first plan ignoring the risks and then replan after the realization of the risks.

Since we conclude, unlike BJ, that $\partial Z/\partial P \neq 0$, then we should look at how optimum price and installed capacity are affected by the distribution of u .

3.6.3.2.iii Expectation of Loss

For a negative outcome of u , the cost of wasted capacity relative to the perfect hindsight vision case is βu , and for a positive outcome of u , the welfare cost is $\frac{1}{2}u^2\theta$, where θ is the slope of the demand function. If, as we do in the Chao analysis in section 3.11, we simplify the distribution to a simple binomial one, then let us assume that we will see a positive shock of \bar{u} , with a probability of 50 percent, and a negative shock of $-\bar{u}$, with a probability of 50 percent. So our expectation

of welfare loss is $\frac{1}{2}(\beta\bar{u} + \frac{1}{2}\bar{u}^2\theta)$. If we allow a distribution to the shock u , then the expectation of loss is

$$E\{\text{Loss}\} = \beta E\{u|u < 0\} + \frac{1}{2}\theta E\{u^2|u > 0\}.$$

We can use the symmetry of the distribution to further simplify to

$$E\{\text{Loss}\} = \beta^* \text{USE} * \text{LOLP} + \frac{1}{2}\theta\sigma^2,$$

where USE is the conditional expectation of lost energy, LOLP is the probability of losing energy, and σ^2 is the variance of u .

Now let us consider adding a volume q to the riskless optimum Z in order to maximize welfare. So our expectation of loss is now

$$E\{\text{Loss}_q\} = \frac{1}{2}\left[\beta(u+q) + \frac{1}{2}(u-q)^2\theta\right]$$

$$E\{\text{Loss}_q\} - E\{\text{Loss}_{q=0}\} = \frac{1}{2}(q - \theta(2qu + q^2)).$$

It is clear from this that depending on the distribution, it may be better to increase or decrease capacity. The more convex the demand function, the smaller the variance at which it is best to increase capacity.

For large variance, the optimal installed capacity increases monotonically with variance. Note the similarity to the Drèze conclusion.

The argument for price is similar. For large variance, we install a large amount of capacity. Let us for a moment forget the problem of negative demand for a symmetric distribution, and consider again a simple binomial distribution with two outcomes. Suppose that the shock is equal to the current deterministic equilibrium demand. Let us also suppose for simplicity that the optimal installed capacity in this stochastic situation is exactly equal to that for deterministic demand. The aggregate welfare is then

$$W = -Z_d\beta - \frac{1}{2}bZ_d + \frac{1}{2}\frac{1}{2}Z_d^2\theta.$$

Z_d is the deterministic demand.

There is no direct dependence on P , but there is a dependence on Z , which depends on P .

There is another feature of the BJ framework that merits consideration. BJ advocate a price of b . Let us consider two cases for build volume. First, a build volume that is commensurate with satiating of demand at price b . There is no rationing, although the long-term economic inefficiency is $\frac{1}{2}\beta^2/\theta$. Suppose instead that the build volume is

commensurate with a price of $b + \beta$. We must then have rationing if we set the price at b . BJ assume that this rationing is efficient. This is consistent with the BJ framework, but when we include rationing inefficiency, it means that if we build a volume consistent with a price of $b + \beta$, then we must consider rationing inefficiency if we have an offer price below this. We will see this in the Visscher framework below.

3.6.4 Discussion of the Brown and Johnson Framework

We can regard the BJ framework as a stochastic application of the Hotelling framework in which a volume decision is made first and independently of the pricing decision. Once built, the fixed costs play no further part in the analysis and thence the price gravitates to variable costs.

We can see from equation (3.26) that when we recognize pricing at the time of build, we return to equilibrium pricing in which fixed costs are recovered.

3.7 MODELING UNDER VARIABLE RATIONING EFFICIENCY—THE VISSCHER FRAMEWORK

The BJ framework in section 3.6 above is highly dependent on efficient rationing. Since in practice rationing is inefficient, we need to examine the sensitivity to rationing efficiency. The Visscher framework allows us to do this. We are then able to show that the optimal volume may increase or decrease with standard deviation of demand shock, according to the nature of the demand function and shock.

While being faithful to the analysis, we here take a geometric approach to the calculus and in doing so can address some other effects, and in particular consideration of a shock to the demand function either in terms of increased volume of demand or of increased willingness to pay.

3.7.1 Visscher's Framework and the Implied Utility Function

1. There is one technology.
2. There is a single subperiod.
3. Returns to scale in capacity β and operation b are constant.
4. There is a hard capacity constraint (variable costs infinite above it).
5. The inverse demand function is downward sloping (elastic).
6. Demand is stochastic with a linear symmetrical shock to the inverse demand function that can be viewed as upward (willingness to pay) or rightward (volume).
7. Rationing can be efficient or inefficient.

Visscher considers welfare in terms of the probability weighted expectation of welfare delivered in all outcomes. In doing so, the welfare of all consumers (born and unborn, domestic and foreign, planning and unplanned) are considered equally in our objective function.

3.7.2 The Analysis

The gross rationing cost to the consumer is the sum of the efficient rationing costs L_1 plus the maximum inefficiency L_2 as shown in figure 3.25. Visscher's formula is

$$E\{L_1 + L_2\} = \int_{Z-X(P)}^{\infty} f(u) \left\{ \begin{array}{l} \int_{X^{-1}(-u)-X^{-1}(Z-u)+P}^{X^{-1}(-u)} \\ [X(P) + u] dP \\ + [X^{-1}(-u) - X^{-1}(Z-u) + \bar{P}] \\ [X(\bar{P}) + u - Z] \end{array} \right\} du.$$

\bar{P} is the preset price.

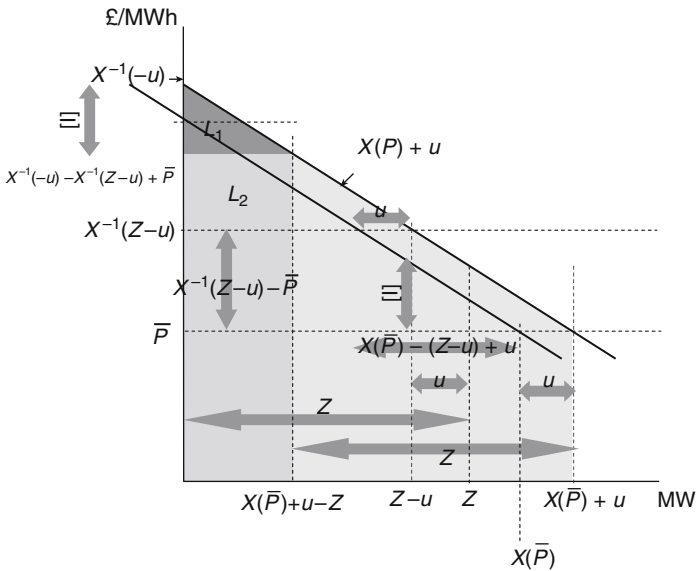


Figure 3.25 Geometric representation of the Visscher framework for a linear demand function.

The welfare formula is then found by finding the total consumer surplus assuming no rationing, subtracting the gross rationing cost, and subtracting the variable cost saving from rationing. To demonstrate our workings, we use slightly different nomenclature to Visscher.

$$\begin{aligned}
 W &= \int_{-\infty}^{\infty} f(u) \int_{\bar{P}}^{X^{-1}(-u)} [X(P) + u] dP du + \bar{P} X(\bar{P}) \\
 &- \int_{z-X(\bar{P})}^{\infty} f(u) \int_{X^{-1}(-u)-X^{-1}(Z-u)+\bar{P}}^{X^{-1}(-u)} [X(P) + u] dP du \\
 &- \int_{z-X(\bar{P})}^{\infty} f(u) [X^{-1}(-u) - X^{-1}(Z-u) + \bar{P}] [X(\bar{P}) + u - Z] du \\
 &- b \left\{ X(\bar{P}) - \int_{z-X(P)}^{\infty} u f(u) du + [Z - X(\bar{P})] \int_{z-X(P)}^{\infty} f(u) du \right\} - \beta Z.
 \end{aligned} \tag{3.29}$$

The first term is the gross consumer surplus without rationing, the second and third terms³⁶ on the second line of the equation are the gross consumer rationing costs, and the fourth and fifth terms on the third line of the equation are the variable and fixed costs respectively. The fourth term is comprised of the variable cost with no rationing minus the variable cost saving from rationing.

While Visscher remains general with respect to demand function, the nature of the shock does lend itself naturally to an assumption of linearity.

3.7.2.1*i* Varying Rationing Efficiency

Visscher allows for the complete spectrum of rationing efficiency from perfect (as in BJ) to perfectly inefficient. We see in figure 3.26 that for perfect efficiency the welfare loss from rationing is proportional to the square of volume difference from the ideal and the additional loss from perfect inefficiency is proportional to the volume difference.

If we denote the welfare loss from rationing inefficiency by W_e , then, for quadratic utility, and beginning with a build amount z that is optimized for the deterministic case,

$W_e = (X^{-1}(0) - P) \int_{u=0}^{u=\infty} g(u) du$ where $g(\cdot)$ is the probability density, and here we have looked only at positive shocks to demand (negative shocks to endowment).

Expectation of the rationing cost is the sum of efficient rationing plus the inefficiency:

$$\begin{aligned}
 W_r &= \int_{u=0}^{u=\infty} g(u) (X^{-1}(u) - X^{-1}(0)) du \\
 &+ \alpha (X^{-1}(0) - P) \int_{u=0}^{u=\infty} u^* g(u) du
 \end{aligned}$$

$$= \frac{1}{2} \frac{\bar{u}^2}{\theta} + \alpha (X^{-1}(0) - P) * \bar{u} \text{ Where } \bar{u} = E\{u \mid u > 0\},$$

where α is the rationing inefficiency. Note that we have the BJ framework as the special case for $\alpha = 0$, and the Visscher case for $\alpha = 1$. We can in this equation clearly see the quadratic term for perfect efficiency and the linear term for inefficiency.

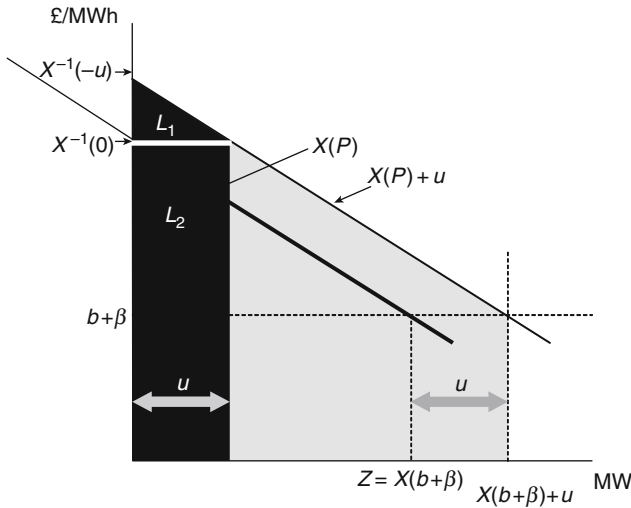


Figure 3.26 Positive demand shock in the Visscher framework. Efficient rationing (L1) plus rationing inefficiency (L2) for a linear demand function and an additive shock to demand.

3.7.2.1.ii Comparison of Horizontal and Vertical Shifts to the Demand Function

The same shock to the demand function can be regarded as a pure shift to volume (e.g., the number of consumers) or a pure shift in willingness to pay (e.g., due to shock to wealth or endowment of substitute goods), or a combination. We will show how these are related by the inefficiency.

Let us then start with the trial solution, price equals fully loaded cost $b + B$, estimate the expectation loss of surplus from stochastic shock, and then see how this changes with a small increment and decrement of capacity. Finally we could, with no loss of generality for this one period one technology setting, set variable cost $b = 0$, and then add b to the optimum price at the end. We do not do this here, in order to make the diagrams similar to those of Visscher.

The central case is shown first, for a positive shock u , which has a 50 percent probability. This is shown in figure 3.26.

The gross consumer surplus for $u = 0$ is

$$\begin{aligned} & [(b + \beta) + \frac{1}{2}(X^{-1}(0) - (b + \beta))][(X(b + \beta))] \\ & = \frac{1}{2}[X^{-1}(0) + (b + \beta)][(X(b + \beta))]. \end{aligned}$$

The total cost for $u = 0$ is

$$[(b + \beta)][(X(b + \beta))].$$

The extra gross consumer surplus for an increase in demand of magnitude u , with no rationing is

$$X^{-1}(-u) * u - \frac{1}{2}u * [X^{-1}(-u) - X^{-1}(0)] = \frac{1}{2}u * [X^{-1}(-u) + X^{-1}(0)].$$

The reduction in gross consumer surplus for a decrease in demand of magnitude u , with no allocative inefficiency is

$$X^{-1}(0) * u - \frac{1}{2}u * [X^{-1}(0) - X^{-1}(u)] = \frac{1}{2}u * [X^{-1}(0) + X^{-1}(u)].$$

which for a linear demand function is

$$= \frac{1}{2}u * [X^{-1}(-u) + X^{-1}(0)].$$

The extra cost, were it possible to deliver this load with no change in capacity is

$$u * b.$$

If we had known in advance about the (certain) shock u , then the spend on extra capacity u would be

$$u * \beta.$$

The gross consumer cost of rationing is

$$u * [X^{-1}(0) + \frac{1}{2}(X^{-1}(-u) - X^{-1}(0))] = \frac{1}{2}u * [X^{-1}(-u) + X^{-1}(0)].$$

Using the assumed linearity of the demand function, we can express

$$X^{-1}(-u) - X^{-1}(0) = \theta * u,$$

where θ is the negative of the slope of the demand function

$$X^{-1}(-u) + X^{-1}(0) = 2X^{-1}(0) + \theta^* u.$$

So the total welfare for positive shock is equal to the welfare with no shock, plus the increase in gross consumer surplus with no rationing, minus the total consumer welfare loss from rationing minus the increase in cost.

The increase in gross consumer surplus with no rationing is equal to

$$X^{-1}(0)^* u + \frac{1}{2} u^* [X^{-1}(-u) + X^{-1}(0)].$$

The gross cost of rationing is equal to the cost of efficient rationing plus the inefficiency

$$\frac{1}{2} u^* [X^{-1}(-u) + X^{-1}(0)] + \alpha^* u^* [X^{-1}(0) - (b + \beta)].$$

There is no change in cost.

So,

$$W_{+u, Z=X(b+\beta)} = W_{0, Z=X(b+\beta)} + X^{-1}(0)^* u - \alpha^* u^* [X^{-1}(0) - (b + \beta)].$$

For the efficient case $\alpha = 0$, we have the welfare after positive shock u

$$W_{+u, Z=X(b+\beta)} = W_{0, Z=X(b+\beta)} + X^{-1}(0)^* u.$$

For the inefficient case $\alpha = 1$ we have

$$W_{+u, Z=X(b+\beta)} = W_{0, Z=X(b+\beta)} + u^* (b + \beta).$$

The cost of perfect inefficiency is then $u^* (X^{-1}(0) - (b + \beta))$.

We will now examine why this expression is also equal to the welfare gain from a vertical shock (i.e., a shock to willingness to pay).

We can see from the figure 3.27 that the maximum inefficiency following a horizontal demand shock is equal to the surplus gain from a willingness to pay (vertical) shock giving rise to the same shock to the demand function.

We can see that

$$\theta = -\frac{X^{-1}(0) - (b + B)}{X(b + B)},$$

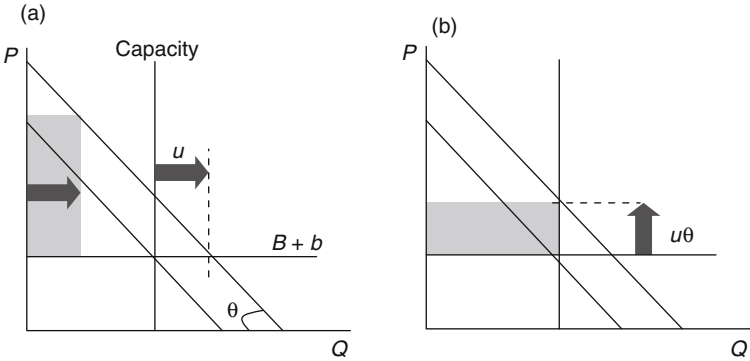


Figure 3.27 Welfare equivalence of demand shock and willingness to pay shock (a) Demand shock with perfect inefficiency (b) WTP shock with constrained capacity.

and that the horizontal demand function shock u is equivalent to a vertical shock $u\theta$.

For the former, the inefficiency is

$$[X^{-1}(0) - (b + B)] * u.$$

For the latter, the surplus gain is

$$X(b + B) * u * \theta.$$

Substituting for θ , we can see in figure 3.27 that the areas are the same.

3.7.2.2 Interpretation of the Nature of the Demand Function Shock

It is worth considering how we might interpret an increase in welfare without an increase in load delivered, due to the constraint in capacity.

Let us first suppose that it is a result of increased willingness to pay. For a linear demand function, a horizontal movement of u has the same effect as a vertical one of $u * \theta$.

We can see by inspection that an increase in willingness to pay with no change in load for anyone results in an increase in welfare of

$$\begin{aligned} Z(X^{-1}(-u) - X^{-1}(0)) &= Z\theta u = Z \frac{X^{-1}(0) - (b + \beta)}{Z} u \\ &= u [X^{-1}(0) - (b + \beta)]. \end{aligned}$$

A vertical movement in the demand function would require no reallocation to consumers, and hence the rationing is efficient. The welfare gain is precisely equal to the inefficiency of the change in demand curve. For a linear demand function, we cannot tell from the movement of the function whether the shock was vertical or horizontal. However, while a vertical movement would not cause reallocation, a horizontal change would.

The inefficiency of investment (i.e., the welfare relative to the optimal investment case with perfect foresight for a one-off demand shock) is $\frac{1}{2}u^2\theta$.

If instead we suppose that consumers are heterogeneous, with inelastic demand to a willingness to pay threshold and add u consumers, with a high willingness to pay between $X^{-1}(u)$ and $X^{-1}(0)$, none of whom get served, then it is easy to see that our allocation is perfectly inefficient, and our welfare is unchanged. If the allocation is perfectly efficient, then we redistribute allocation from the consumer with WTP in the range $b + \beta$ to those in the range $b + \beta + (X^{-1}(u) - X^{-1}(0))$. We can see by inspection that this redistribution delivers a utility

$$u[X^{-1}(0) - (b + \beta)].$$

We can therefore see why the welfare gain from vertical demand shock with perfect efficiency is equal to the efficiency cost of perfect inefficiency.

3.7.2.3 Estimation of the Inefficiency α for Random Rationing

Changes in the number of consumers in society gives us problems, for example, regarding the tax base and the welfare of visitors. Let us then suppose that the horizontal shock to the demand function is a result of shock to endowment. Let us suppose first that all consumers are homogenous and hence the downward slope of the aggregate demand function is entirely due to the individual slopes and not at all due to the heterogeneity of society. Our inefficiency now is not a result of withdrawing power from those with highest willingness to pay (as all consumers have the same WTP), but the welfare inefficiency of withdrawing a large amount from a small number of people (as happens in practice) than a small amount of load from a large number of people.

For a system-wide loss λ percent of satiated demand, or $\lambda Z(b + \beta)$ MW, the most efficient way is to spread evenly across all consumers if they are homogenous, that is, all consumers lose λ percent rather than λ percent losing all consumption. For maximum efficiency, we assume

that the consumers can ration use according to value. For n homogeneous consumers, with pro rata loss that is efficient at the level of the individual, the net consumer welfare loss is

$$\frac{1}{2} * \lambda^2 * n * (Z(b + \beta))^2 * \theta,$$

where θ is the negative of the slope of the demand function.

If instead some consumers lose all their power, then the net consumer welfare loss is

$$\lambda * n * \frac{1}{2} * [X^{-1}(0) - (b + \beta)][Z(b + \beta)].$$

If it were the case that either i) all consumers lost some power, and that the mechanics of utilization were that this were the most valuable power, or ii) some consumers lose all power and these have the highest WTP, then the net consumer welfare loss is

$$\lambda * n * [X^{-1}(0) - (b + \beta)][Z(b + \beta)]. \text{ This is the } \alpha = 1 \text{ case.}$$

So,

$$\alpha_{rr} = \frac{\lambda * n * \frac{1}{2} * [X^{-1}(0) - (b + \beta)][Z(b + \beta)] - \frac{1}{2} * \lambda^2 * n * (Z(b + \beta))^2 * \theta}{\lambda * n * [X^{-1}(0) - (b + \beta)][Z(b + \beta)] - \frac{1}{2} * \lambda^2 * n * (Z(b + \beta))^2 * \theta},$$

where α_{rr} denotes random rationing.

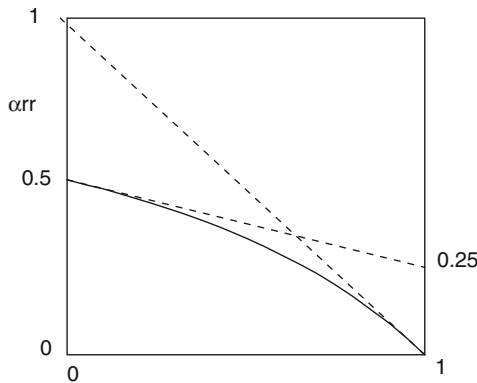


Figure 3.28 The change of rationing inefficiency with amount of lost load, for random rationing.

Noting that $\theta = \frac{X^{-1}(0) - (b + B)}{Z^*(b + B)}$ and then simplifying, we have

$$\alpha_{rr} = \frac{1 - \lambda}{2 - \lambda}.$$

We can see that for small losses $\lambda \rightarrow 0$, then $\alpha_{rr} \approx \frac{1}{2}$, and for large losses $\lambda \rightarrow 1$ then $\alpha_{rr} \rightarrow 0$.

The form is shown in figure 3.28. Being of concave form, the inefficiency pervades to fairly high amounts of lost load.

Let us turn to efficient build and pricing under stochastic conditions of an additive shock.

3.7.2.4 *The Dependence of the Efficient Build Volume on Rationing Efficiency*

We assume initially that for the deterministic case $u = 0$, there is no allocative inefficiency for consumers. Allocative efficiency for this case is something we examine later.

For a positive shock, the welfare is the gross consumer surplus with no rationing, minus the efficient rationing cost minus the cost minus the net inefficiency

$$\begin{aligned} W_{+u}(Z = X(b + \beta)) &= \left[\frac{1}{2}(X^{-1}(0) + (b + \beta)) \right] [(X(b + \beta)) \\ &\quad + \frac{1}{2}u^* [X^{-1}(-u) + X^{-1}(0)] \\ &\quad - \frac{1}{2}u^* [X^{-1}(-u) + X^{-1}(0)] \\ &\quad - [(b + \beta)][(X(b + \beta))] \\ &\quad - u^* \alpha^* [X^{-1}(0) - (b + \beta)]. \end{aligned}$$

If the shock is negative, and there is no productive allocative inefficiency/rationing cost, there is a cost saving of $u^* b$.

$$\begin{aligned} W_{-u}(Z = X(b + \beta)) &= \left[\frac{1}{2}(X^{-1}(0) + (b + \beta)) \right] [(X(b + \beta)) \\ &\quad - \frac{1}{2}u^* [X^{-1}(0) - X^{-1}(u)] \\ &\quad - [(b + \beta)][(X(b + \beta)) + u^* b]. \end{aligned}$$

Here we use the sign convention that u is always positive, and is multiplied by +1 for a positive shock and -1 for a negative shock.

The expectation is

$$E\{W_{\bar{u}}\} = \frac{1}{2}W_{+u} + \frac{1}{2}W_{-u}$$

$$\begin{aligned}
2E\{W_{\tilde{u}}\} &= 2 * E\{W_{\tilde{u}}(Z(b + \beta))\} \\
&= 2\left[\frac{1}{2}(X^{-1}(0) + (b + \beta))\right][(X(b + \beta)) - 0 \\
&\quad - \frac{1}{2}u * [X^{-1}(-u) - X^{-1}(0)] \\
&\quad - 2[(b + \beta)][(X(b + \beta)) + u * b - u * \alpha * [X^{-1}(0) - (b + \beta)]].
\end{aligned}$$

Here we have assumed that since we have a linear demand function and an additive shock,

$$X^{-1}(-u) - X^{-1}(0) = X^{-1}(0) - X^{-1}(u).$$

The deterministic case was

$$E\{W_{\tilde{u}=0}\} = \left[\frac{1}{2}(X^{-1}(0) - (b + \beta))\right][(X(b + \beta))].$$

So the expectation of welfare loss from the addition of stochasticity is

$$\begin{aligned}
E\{W_{\tilde{u}=0}\} - E\{W_{\tilde{u}}\} &= -\frac{1}{2}u * [X^{-1}(-u) - X^{-1}(0)] \\
&\quad - \frac{1}{2}u * b + \frac{1}{2}u * \alpha * [X^{-1}(0) - (b + \beta)].
\end{aligned}$$

Let

$$[X^{-1}(-u) - X^{-1}(0)] = \theta * u.$$

So,

$$X^{-1}(0) = \theta * X(b + \beta) + (b + \beta).$$

So the expectation of loss from the addition of the stochastic shock \tilde{u} is

$$E\{W_{\tilde{u}=0, Z}\} - E\{W_{\tilde{u}, Z}\} = \frac{1}{2}u * [-\theta * u - b + \alpha * X(b + \beta)].$$

Now let us add a small amount of capacity z , where $z < u$.

For zero shock, the addition of capacity does not result in an increase of delivered load, so the welfare change from the increase in demand is simply $-\beta * z$.

For a positive shock to demand, let us visualize the aggregate demand curve as being represented as a set of heterogeneous consumers, each with a single willingness to pay, and each of whom experience the same shock.

Relative to the $Z = X(b + \beta)$ case, we look at the changes to i) gross surplus with no rationing, ii) efficient rationing cost, iii) cost of rationing inefficiency, and iv) delivery cost.

The no-rationing welfare is unchanged relative to the $z = 0$ case.

The efficient rationing cost component of welfare changes from

$$-\frac{1}{2}u^* [X^{-1}(-u) + X^{-1}(0)] \text{ to } -\frac{1}{2}u^* [X^{-1}(-u + z) + X^{-1}(0)].$$

The delivery cost component of welfare changes from

$$-[(b + \beta)][X(b + \beta)] \text{ to } -[(b + \beta)][X(b + \beta) + z].$$

The inefficiency component of welfare changes from

$$-u^* \alpha^* [X^{-1}(0) - (b + \beta)] \text{ to } -(u - z)^* \alpha^* [X^{-1}(0) - (b + \beta)].$$

So, for positive shock u , the total difference in welfare from the increase in capacity is

$$\begin{aligned} & -\frac{1}{2}u^* [X^{-1}(-u + z) - X^{-1}(-u)] \\ & - (b + \beta)^* z + z^* \alpha^* [X^{-1}(0) - (b + \beta)] \\ & = -\frac{1}{2}u^* z\theta - (b + \beta)^* z + z^* \alpha^* Z\theta \\ & = z(\frac{1}{2}u^* \theta - (b + \beta) + \alpha^* Z\theta). \end{aligned} \quad (3.30)$$

For a negative demand shock, then relative to the $Z = X(b + \beta)$ case, there is still no rationing, and the cost is increased by $\beta^* z$.

$$\begin{aligned} E\{W_{\tilde{u}, Z+z}\} - E\{W_{\tilde{u}, Z}\} &= \frac{1}{2}\{z(\frac{1}{2}u^* \theta - (b + \beta) + \alpha^* Z\theta) - \beta^* z\} \\ \frac{\partial [E\{W_{\tilde{u}, Z+z}\} - E\{W_{\tilde{u}, Z}\}]}{\partial z} &= \frac{1}{2}\{(\frac{1}{2}u^* \theta - (b + 2\beta) + \alpha^* Z\theta)\}. \end{aligned}$$

For $u \rightarrow 0$, $\alpha = 0$, and $z > 0$,³⁷ then this is just the expected increase in costs $\frac{1}{2}b + \beta$. Remembering that we limited $z < u$, this is the initial slope of the quadratic form of the deadweight loss.

For $\alpha > 0$ and $u \rightarrow 0$, we add a benefit of $\frac{1}{2}(\alpha^* Z\theta)$.

The size of u at which it becomes beneficial to add capacity is

$$u = \frac{2(b + 2\beta)}{\theta} - 2^* \alpha^* Z. \text{ This is depicted in figure 3.29.}$$

The introduction of rationing inefficiency therefore causes us to reduce the size of u for which it is beneficial to add capacity. The

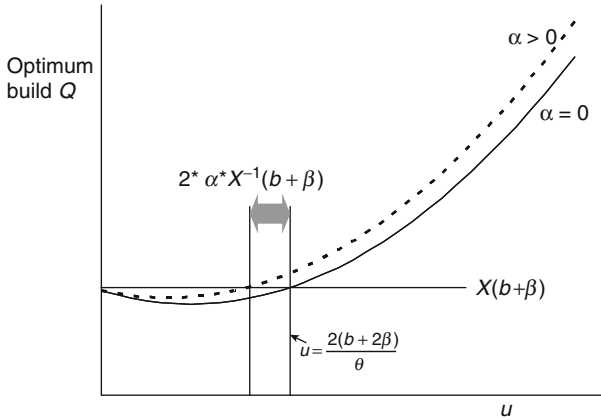


Figure 3.29 The impact of rationing efficiency on optimum build. $\alpha > 0$ represents inefficiency.

optimum capacity for the inefficient rationing case is at least equal to that for the efficient rationing case.

So for linear demand functions and constant returns to scale, inefficiency in rationing never decreases the optimal capacity, and always decreases the stochastic shock size at which capacity addition is welfare optimal.

For small u , we have seen that the optimum capacity for the stochastic case can be less than for the deterministic case.

3.7.2.5 Price and the Financial Position of the Generating Unit

In this book we have made much of the Pareto optimality requirement of the generator not to make losses on an ex ante expectation benefit, and of the competitive market force to drive positive generator profits to zero.

The main argument about the value of $\partial Z/\partial P$ in welfare equation (3.29) is the same as BJ and not repeated here. There is more flexibility in the Visscher framework to consider the additional effect of inefficient rationing.

We can see that if we set price at the full cost $b + \beta$ and then introduce stochasticity of demand, that if the generator builds an amount $X^{-1}(b + \beta)$, then it has an expectation loss of $\frac{1}{2} * u * \beta$, since there is no revenue change for positive shock, but there is a net loss of revenue of $u * \beta$ for negative shock. The same is true for efficient or inefficient rationing, given that $P = b + \beta$, although we should remember that willingness to pay does depend on rationing efficiency.

To restore economic equilibrium and ensure the build of plant, the unit must make the same profit under positive shock as loss under negative shock.

Let us assume initially that u is large enough for the demand with positive shock u and price $P + \Delta P$ exceeds the demand with no shock and price P . So $u > \frac{\Delta P}{\theta}$ where θ is the negative of the slope of the demand function $\theta > 0$

The demand is not met and build volume $X(b + \beta)$ is delivered. So with offer price $b + \beta + \Delta P$, we have an expectation of profit:

$$E\{\pi_{\Delta P}\} = \frac{1}{2} * [X(b + \beta)] * (b + \beta + \Delta P - b) \\ + \frac{1}{2} [(X(b + \beta + \Delta P) - u)] * (b + \beta + \Delta P - b) - \beta * X(b + \beta).$$

We know that

$$X(b + \beta + \Delta P) = X(b + \beta) - \frac{\Delta P}{\theta}.$$

So,

$$E\{\pi_{\Delta P}\} \Big| u > \frac{\Delta P}{\theta} = \left[X(b + \beta) - \frac{1}{2} \frac{\Delta P}{\theta} - \frac{1}{2} u \right] * (\beta + \Delta P) - \beta * X(b + \beta) \\ E\{\pi_{\Delta P}\} \Big| u > \frac{\Delta P}{\theta} = \Delta P * X(b + \beta) - \frac{1}{2} \left[\frac{\Delta P}{\theta} + u \right] * (\beta + \Delta P). \quad (3.31)$$

The term on the left is the profit increase from the price increase if there is no change in demand.

We can break down the term on the right as follows:

$$\frac{1}{2} u * \beta + \frac{1}{2} \Delta P * \left(\frac{\beta + \Delta P}{\theta} + u \right)$$

If $u < \frac{\Delta P}{\theta}$, then we always have excess capacity and so,

$$E\{\pi_{\Delta P}\} = \frac{1}{2} * [(X(b + \beta + \Delta P) + u)] * (b + \beta + \Delta P - b) \\ + \frac{1}{2} [(X(b + \beta + \Delta P) - u)] * (b + \beta + \Delta P - b) - \beta * X(b + \beta) \\ = [(X(b + \beta + \Delta P))] * (\beta + \Delta P) - \beta * X(b + \beta) \\ = [(X(b + \beta + \Delta P))] * (\Delta P) - \beta * \frac{\Delta P}{\theta}.$$

Within the range, there is no dependence on u , since there is a symmetrical gain if demand increases (since the producer has spare capacity and can deliver it) and if demand decreases.

Clearly, profit is 0 when $\Delta P = 0$. We can also see that if the producer can price above fully loaded costs (implying market power), then he can make more, with the optimum price determined by the level at

$$\text{which } \frac{\partial [E\{\pi_{\Delta P}\}]}{\partial [\Delta P]} = 0.$$

Now in a competitive market for the deterministic case, a positive ΔP is unsustainable, as it attracts new production. In the stochastic case, the equilibrium ΔP is set by the level at which ex ante profit expectation is zero (ignoring cost of risk for the moment).

$$E\{\pi_{\Delta P}\} = 0 = \left[-\frac{1}{2} \frac{\Delta P}{\theta} - \frac{1}{2} u \right] * (\beta + \Delta P) + \Delta P * X^{-1}(b + \beta).$$

This is a rather awkward quadratic equation, but it does have a solution.

$$\text{The net consumer welfare change from the price rise is } \frac{-\frac{1}{2}(\Delta P)^2}{\theta}.$$

If we build a lesser amount $X(b + \beta) - \Delta Q$, and correspondingly price at $X^{-1}(X(b + \beta) - \Delta Q) = b + \beta + \theta^* \Delta Q$, then our expectation profit is

$$\begin{aligned} E\{\pi\} &= \frac{1}{2}(b + \beta + \theta^* \Delta Q - b) * [(X(b + \beta) - \Delta Q) - u] \\ &\quad + \frac{1}{2}(b + \beta + \theta^* \Delta Q - b) * [(X(b + \beta) - \Delta Q) + 0] \\ &\quad - \beta^* [(X(b + \beta) - \Delta Q)] \\ E\{\pi\} &= \frac{1}{2}(\beta + \theta^* \Delta Q) * [(X(b + \beta) - \Delta Q) - u] \\ &\quad + \frac{1}{2}(\beta + \theta^* \Delta Q) * [(X(b + \beta) - \Delta Q)] \\ &\quad - \beta^* [(X(b + \beta) - \Delta Q)] \\ &= [(X(b + \beta) - \Delta Q)][(\beta + \theta^* \Delta Q - \beta)] - \frac{1}{2}(\beta + \theta^* \Delta Q) * u \\ &= [(X(b + \beta) - \Delta Q)][(\theta^* \Delta Q)] - \frac{1}{2}(\beta + \theta^* \Delta Q) * u \end{aligned}$$

This is maximized by

$$0 = \frac{\partial E\{\pi\}}{\partial \Delta Q} = -[(\theta^* \Delta Q)] + [(X(b + \beta) - \Delta Q)][(\theta)] - \frac{1}{2}(\theta) * u,$$

$$\Delta Q = \frac{X(b + \beta)}{2} - \frac{1}{4}u.$$

If we have constrained the capacity to generate a higher producer profit for the deterministic case (with producer entry barriers), then the addition of stochasticity increases the optimum build, since an increment of build generates a marginal profit of $\beta + \theta \Delta Q$ for positive shock and only a loss of β for negative shock.

In the absence of a price increase, the addition of the stochastic term decreases the optimal installed volume from the producer perspective. This is due to the wasted capacity and lost revenue on a negative demand shock, with no opportunity for gain on a positive demand shock.

In the absence of entry barriers or market power, then the solution drives to $E\{\pi\} = 0$.

3.7.3 Discussion of the Visscher Analysis

We can regard the Visscher framework as a generalization of the BJ framework for a range of rationing inefficiency.

The sensitivity of price to fixed costs is apparent from differentiating equation (3.29) with respect to price and the arguments for including or omitting the fixed cost term are the same as for BJ.

Interestingly, as is common with many authors, Visscher does not regard the covering of fixed costs as paramount, and indeed he views the uplift of price from variable to full costs as a function of rationing inefficiency.

The framework is very useful for examining more general results in relation to rationing inefficiency, such as the effect on optimal capacity, the loss of welfare. The use of a linear demand is particularly useful, as horizontal and vertical shocks look the same but actually can be quite different in terms of rationing. We see in the analysis that it can be sensible to assume efficient rationing in static conditions and either efficient or inefficient rationing in dynamic conditions, and this depends on the driver to the shock to the demand function.

3.8 DEMAND FOR CAPACITY WITH STOCHASTIC ELASTIC DEMAND—THE CARLTON FRAMEWORK

So far we have introduced the demand for capacity, stochastic demand functions, rationing, efficiency at the margin, and equilibrium.



Figure 3.30 Event ordering in the Carlton analysis.

The work of Carlton (1977) allows us to bring some of these together in a more general manner, with a less restricted demand function and shock to it, albeit still with a stochastic demand function with a maximum willingness to pay in all circumstances. The event order is shown in figure 3.30.

3.8.1 Framework

1. a one period setting
2. a single unit with fixed costs β , variable costs b , and constant returns to scale in capacity and operation
3. a hard constraint at installed capacity z
4. a convex demand function $D(p) = x(p)$, finite at price $p = 0$
5. demand shock is a dimensionless stochastic variable u , so $D(p) = x(p)u$,
6. u may represent number of consumers.
7. Correspondingly $pu = x^{-1}(D)$.

For maximum consistency we imagine the following:

1. We have a finite population of n people (in fact the most consistent form of shock here is shock to population).
2. They each have an identical utility function.
3. They each have a stochastic endowment of a substitute good.
4. The endowment has digital stochasticity (zero or nominally infinite).

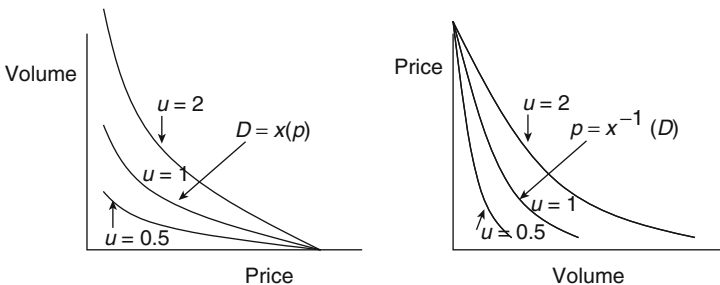


Figure 3.31 The demand function for different outcomes of multiplicative stochastic variable u .

5. Each has the same ex ante probability of endowment.
6. The probability distribution of the number of those with a single demand function as depicted below is therefore Poisson, which we may variously simplify to binomial or normal for convenience.
7. There is a maximum willingness to pay, which we see in figure 3.31.

3.8.2 Analysis

We choose to invest a total of z MW of capacity. If the realization of demand $u * x(p)$ exceeds z then we have to ration. Carlton assumes random rationing (for an n percent rationing overall, this means that n percent of consumers receive no product).

The rationing is shown in figure 3.32. We demand $x(P)$ at price P , receive z , and so the rationing ratio defined as $s = \frac{z}{x(p)}$. The assumptions on rationing are described further down.

The expectation of (Marshallian) surplus to society is equal to the unconditional expectation of surplus for no rationing ($u < 1$), plus the (unconditional) expectation of surplus for rationing ($u > 1$), minus the expectation of the cost of rationing. Note that Carlton makes a key assumption here that society is a consistent entity. For example, if we really do regard society as having a varying number of members that we maximize the welfare of this varying number, rather than, for example, the current number, which may correspond to our tax base.

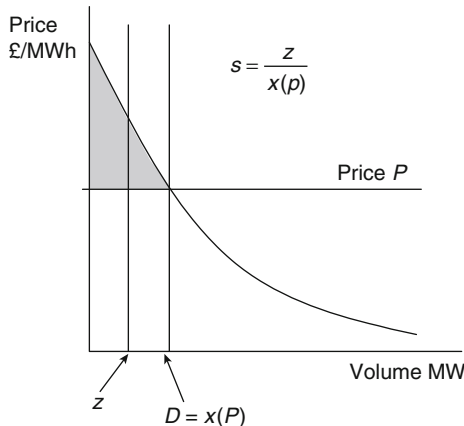


Figure 3.32 Rationing when demand exceeds capacity. The shaded area is consumers' surplus before rationing.

We should note that when the shock factor u is multiplicative $u \cdot x(p)$ rather than additive $u + x(p)$, an increase in u will increase the mean, and we must adjust for this. If u is normally distributed then this is straightforward and the adjustment is $-\frac{1}{2}u^2$

3.8.2.1.i *The Stochastic Term*

For $\tilde{u} = 1$, the surplus, if unconstrained, and with no rationing efficiency costs is

$$S = \left[\int_0^{x(p)} x^{-1}(q) dq - bx(p) \right] - \beta z,$$

where b is the variable cost and β is the fixed cost. We can see the net surpluses in figure 3.33.

For $\tilde{u} \neq 1$ then, for the unconstrained case, given the nature of the stochastic variation, we simply multiply the marginal surplus by u .

If unconstrained, then the surplus is

$$S = u \left[\int_0^{x(p)} x^{-1}(q) dq - bx(p) \right] - \beta z.$$

If constrained, then the surplus is

$$S = s \left[\int_0^{x(p)} x^{-1}(q) dq - bx(p) \right] - \beta z.$$

Putting these together in the stochastic world, the expectation of surplus S is

$$S = \int_0^{u=s} u \left[\int_0^{x(p)} x^{-1}(q) dq - bx(p) \right] dF(u) + \int_{u=s}^{u=\infty} s \left[\int_0^{x(p)} x^{-1}(q) dq - bx(p) \right] dF(u) - \beta z, \tag{3.32}$$

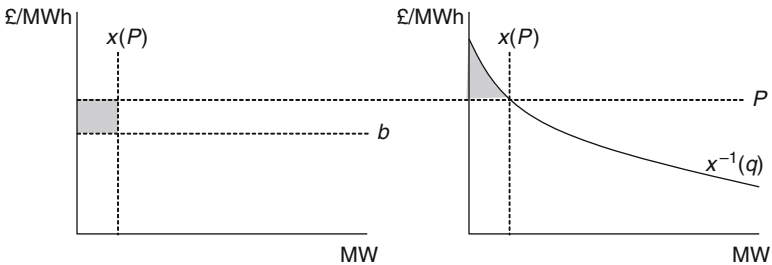


Figure 3.33 Producer marginal surplus and consumer surplus after payments in the Carlton framework.

where

p is the price

F is the cumulative probability density of \tilde{u} .

On the right hand side, the surplus before rationing is the surplus for $u = 1$ times u , and the amount of rationing is $\frac{z}{x(p)}$. The effect of rationing is shown in figure 3.34.

For a given z , to find the price for ex ante maximum welfare, we set the first derivative of welfare/surplus with respect to price to zero.

The chain rule is

$$\frac{d}{dx} \int_0^x g(y)dy = g(x) \text{ and } \frac{d}{dx} \int_0^{f(x)} g(y)dy = g(f(x))f'(x),$$

So, $\frac{\partial}{\partial p} \int_0^{x(p)} x^{-1}(q)dq = x'(p) * p$ where $x'(p) = \frac{\partial x(p)}{\partial p}$ and $x^{-1}(x(p)) = p$,

$$\begin{aligned} \frac{dS}{dp} &= \int_0^{\tilde{u}=s} \tilde{u} [x'(p)(p - b)] dF(u) + \int_{\tilde{u}=s}^{\infty} s [x'(p)(p - b)] dF(u) \\ &- \beta s x'(p) = 0. \end{aligned} \tag{3.33}$$

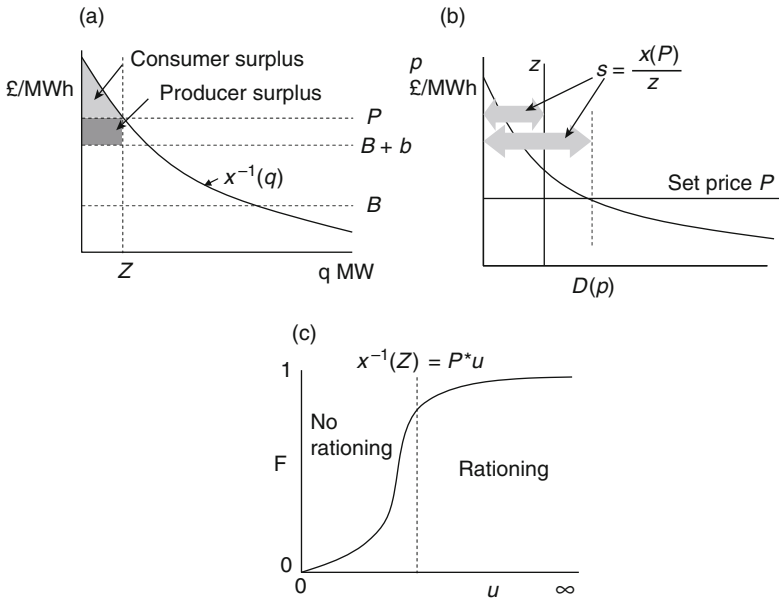


Figure 3.34 Random rationing in the Carlton analysis a) Net surpluses for $u = 1$ b) Application of ratio for surplus under rationing for $u > 1$ c) Probability domains.

Dividing through by $x'(p)$, which is nonzero, we have

$$(p - b) \left[\int_0^{\tilde{u}=s} \tilde{u} dF(\tilde{u}) + \int_{\tilde{u}=s}^{\infty} s dF(\tilde{u}) \right] - \beta s = 0 \quad (3.34)$$

If we set the capacity z just at the point at which we have to ration and assume a very tiny stochastic variation³⁸ in u from $u = 1$, then this equation resolves to

$$(p - b) \left[\int_0^{\tilde{u}=s-\delta} \tilde{u} dF(\tilde{u}) + \int_{\tilde{u}=s-\delta}^{\tilde{u}=s} \tilde{u} dF(\tilde{u}) \right. \\ \left. + \int_{\tilde{u}=s+\delta}^{\infty} s dF(\tilde{u}) + \int_{\tilde{u}=s+\delta}^{\infty} s dF(\tilde{u}) \right] - \beta^* \mathbf{1} = 0, \\ (p - b) \left[0 + \frac{1}{2} * \mathbf{1} + \frac{1}{2} * \mathbf{1} + 0 \right] - \beta = 0.$$

$P = b + \beta$. So for the deterministic case, costs are exactly recovered.

We can rearrange equation (3.34) to

$$(p - b) \left[\int_0^s \tilde{u} dF(\tilde{u}) + s - \int_0^s s dF(\tilde{u}) \right] - \beta s = 0.$$

If $\int_0^s (s - u) dF(u) > 0$ and $p > b$ then $p > b + \beta$. This is true for both deterministic and stochastic conditions.

3.8.2.2 Discussion of the Optimum Price

As with the other frameworks, if our capacity commitment is made first and then we commit to price, then $\frac{\partial z}{\partial p} = 0$ and thence $p = b$.

The elevation of p above $b + \beta$ does not give us any particular indication for optimal build. The elevation caters for the fact that under stochastic conditions, units will be idle, but costing money for part of the time.

We are then expecting consumers to pay extra to producers for having the ability to have varying demand satisfied by available capacity.

As for the deterministic case, we have co-optimized capacity and price, and we have also assumed that our rationing regime is such that it is more expensive in welfare terms to have too little capacity than too much. Since we can underuse capacity but not overuse it, we would expect stochasticity to cause an increase in price charged by the producer. Where u is multiplicative, it is not easy to work out whether producers break even on average, because an increase in σ increases the

average level of demand. We will examine in a few paragraphs the case where u is additive (i.e., quadratic utility with endowment shock).

Let us consider the intuition on this, simplifying with a single willingness to pay. First, suppose we invest in the amount of capacity that is optimal for deterministic conditions. The surpluses in relation to the realization of u is shown in figure 3.35. The ex post investment inefficiency, that is, the surplus we would have got if we had predicted the outcome of u , minus the actual surplus, is shown in figure 3.35(c).

So, for a simple single willingness to pay, and with constant returns to scale in capacity and operations, we have symmetry (i.e., independence of optimal installed volume with respect to standard deviation), if $\beta = WTP - b - \beta$, that is, $WTP = b + 2\beta$. The problem is that we should expect that if the willingness to pay is calculated at the same horizon time, then, at least for the non-stochastic case, we would expect $WTP = b + \beta$, as otherwise more and more production would arrive until the capacity cost has risen.

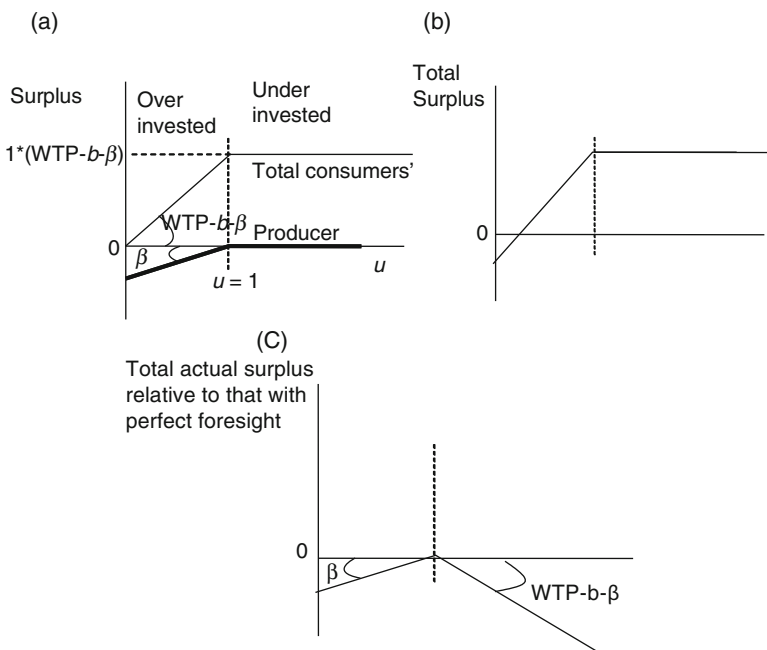


Figure 3.35 Producer and consumer surpluses for different stochastic outcomes. (a, b) following capacity investment at the volume that is optimal for the deterministic case. The situation with constant willingness to pay. (c) aggregate surplus relative to that with perfect foresight in demand outcome. Consumers's surplus shown linear for simplicity.

3.8.2.2.i Consideration of Different Forms of Demand Function

With a multiplicative shock, and hence a linear shock to the logarithm, the natural demand function candidate is similar to the logarithmic, although, recalling that the willingness to pay is equal to the slope of the utility function, we must take care to have finite willingness to pay and welfare for vanishingly small consumption.

Three interesting functions shown in figure 3.36 are i) constant willingness to pay, ii) quadratic utility (linear demand function), and iii) exponential.

The welfare outturn compared to that for perfect demand foresight is shown in figure 3.37.

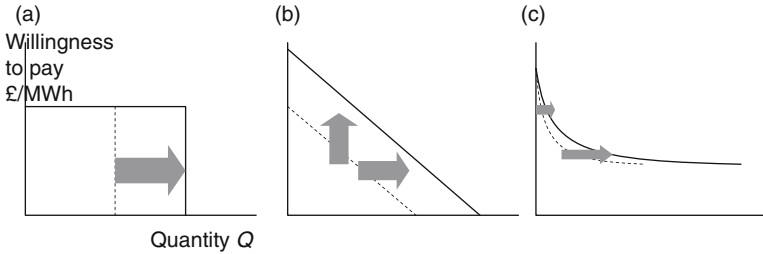


Figure 3.36 Upward shocks only shown for three demand functions (a) Constant willingness to pay (b) Linear (c) Exponential.

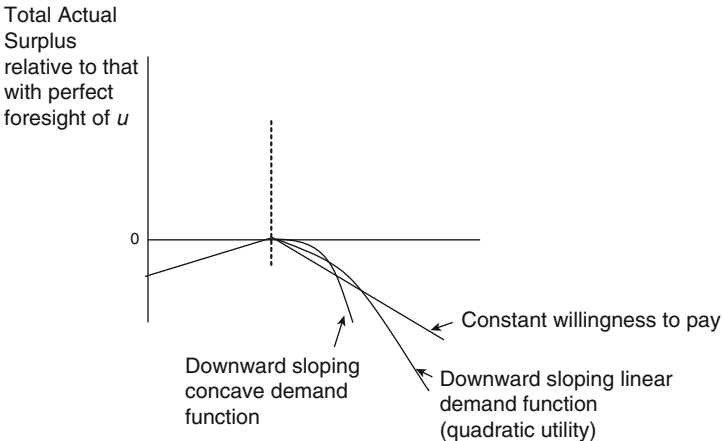


Figure 3.37 Lack of symmetry in inefficiency, leading to a dependence on optimum build on the form of the utility/demand function and standard deviation.

We can see by inspection, that our optimal capacity, relative to the $u = 0$ case, is decreased for small u and increased for large u . We noted a similar effect in our analysis of the Visscher framework, and also showed with the Visscher framework how the combination of inefficiency rationing and quadratic utility can combine to give an approximately linear aggregate utility function (i.e., constant willingness to pay).

The loss of surplus on the right-hand side of the figure in outcome where demand exceeds capacity is proportional to the square of the shortfall. The slope of the expectation-of-loss function on the right-hand side is dependent on the probability function. For example, for a very simple uniform distribution, the slope is proportional to the capacity shortfall and hence is zero at the equilibrium point.

We notice here the similarity of surplus profile of the formula $\int_0^s (s - \tilde{u}) dF(\tilde{u}) > 0$ to that of a put option with strike s and underlying price \tilde{u} . The expectation loss of surplus in relation to the variance of \tilde{u} can then be found with standard option analytics.

The producer has zero probability of making a profit and a finite probability of making a loss and hence must raise his price above $b + \beta$ to maintain ex ante cost equilibrium.

The aggregate surplus argument for raising the price above $b + \beta$ is in the effect on demand. By raising the price above $b + \beta$ the demand expectation drops. Since we did not change capacity z , the expected amount of rationing also drops. So in effect we are rationing by price, since consumers will expect to consume less and will choose to forego the least valuable demand. We can regard this, instead of maintaining z and raising price, as providing spare capacity for a self-rationed demand.

The effect of the benefit of foresight of demand is shown in figure 3.38.

We therefore conclude that whether the addition of stochasticity to the demand function should increase or reduce capacity, is

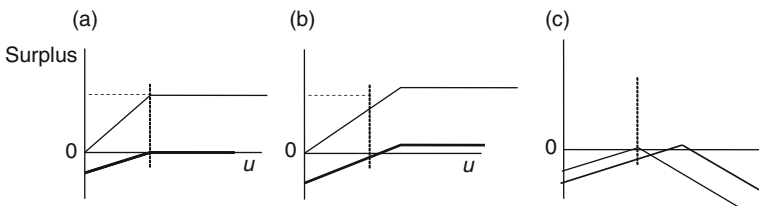


Figure 3.38 (a) Producer and consumers' surplus as in figure 3.37 (b). As (a) but with increased capacity build (c) Surplus relative to "perfect foresight" surplus for optimal build and extra build.

indeterminate in the general case. The raising of price above $b + \beta$ is simply a result of raising price to encourage consumers with downward sloping demand functions, to self-ration.

Another way of looking at this is the loading of capacity cost into price—a demand for capacity.

Note that asymmetry of the inefficiency function causes us to have a capacity bias relative to the deterministic case that increases with distributional variance. The multiplicative nature of the stochastic demand suggested by Carlton would normally suggest an asymmetric distribution, such as lognormal, although as we have seen, a symmetric distribution is technically possible. The reason that we consider it unlikely is that to rationalize the framework, we applied a nominally infinite endowment of a substitute good. We would not expect this to have 50 percent probability as we require for a symmetric distribution, but would instead have a very low probability and hence a highly skewed distribution.

We conclude then that addition of stochastic shock to demand could either decrease or increase optimal capacity, according to the relative slopes of demand and cost functions.

3.8.3 Discussion of the Carlton Analysis

The Carlton analysis confirms the BJ and Visscher result for optimal price of marginal cost b given a capacity decision but is more emphatic on the optimal price of the fully loaded cost $b + B$ when price and volume are co-optimized. The result is not dependent on an excessively simple assumption of demand function and shock to it, although the assumption on rationing efficiency is less flexible than Visscher's.

Having now modeled single assets, we now move closer to modeling a whole installed stack of power station units.

3.9 OPTIMAL PRICING, CAPACITY AND THE TECHNOLOGY FRONTIER—CREW AND KLEINDORFER

So far we have considered the one-period stochastic setting. We now consider the situation where we have many subperiods and many power station units. The demand in each period is stochastic and the function being continuous, we cannot have easy recourse to the duality between stochastic and deterministic load factor that we described in section 2.4.1. If there were m stochastic states per period, then we could model as deterministic with $m * n$ subperiods.

Crew and Kleindorfer (CK) pursue an interesting discussion on generation diversity modeled as a spectrum of fixed/variable costs. We characterize the problem as follows, which is consistent with the CK analysis.

3.9.1 Framework

1. Over the cycle considered, there are n periods of equal length, with a total length of 1.
2. We have an initial candidate installed generation stack, and the number of technologies is limited to those in the candidate stack.
3. The units that have arbitrarily small capacity (i.e., they are infinitely divisible from a practical perspective), and they are available in infinite volume at investment time.
4. Generating units never fail, have zero start costs, and have clearly delineated fixed and variable costs. There is no step change in costs as load rises from zero.
5. Generating units can operate at all levels from zero to full load with constant returns to scale in the short and long run (i.e., energy and capacity).
6. All generating units are owned and operated by a benign monopolist.
7. Prices, which are periodic, are set before the cycle begins, may not be changed subsequently and must be the same to all consumers.
8. Demand is periodic and elastic with zero cross-elasticity between periods. The demand functions are continuously differentiable.
9. The stochastic disturbance to demand is a linear addition or reduction in volume rather than price.
10. When demand exceeds capacity, load is rationed. The welfare loss is consistent with random rationing.
11. Demand rationing may be associated with costs to generator efficiency—for example, from the appearance of transmission constraints under increased flow on peak days.

We make some relatively minor changes to the nomenclature of CK, to add clarity for the purposes of our arguments.

3.9.1.1.i Consumer Characteristics

The demand function and inverse demand functions associated with this is shown in figure 3.39. The demand function is depicted as convex to maintain generality, though we note that a function that is

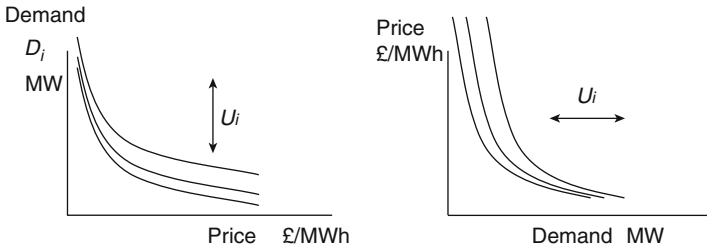


Figure 3.39 Stochastic demand as depicted in Crew and Kleindorfer.

highly convex at low loads, with an additive quantity shock do not go well together for electricity. For a relatively modest change in heat load, there is a huge change to the VOLL (the intersect with the ordinate)

3.9.2 The Analysis

Let $x = (x_1, x_2, \dots, x_n)$ denote before any shock the vector of quantities of consumption of periods $i = 1, \dots, n$ and let $p = (p_1, p_2, \dots, p_n)$ be the corresponding vector of prices.

$D_i(p_i, \tilde{u}_i) = X_i(p_i) + \tilde{u}_i$. Here the stochastic demand experiences an additive shock

Here $X_i(p_i)$ represents the mean demand in period i for price p_i and the disturbance term \tilde{u}_i has an expected value $E[\tilde{u}_i] = 0$.

We consider power plant as follows:

- b_l is the variable operating cost of the l th plant, in £/MWh
- β_l is the capacity cost of the l th plant in £/MW/cycle
- q_l , also noted \underline{q}_l is the capacity of the l th plant in MW
- q_{li} is the actual output of the l th plant in the i th period, in MWh.

We index the plants such that $0 < b_1 < b_2, \dots, b_m$ for all periods, and since returns to scale are constant and we would not wish to run inferior,³⁹ plant we can state that $\beta_1 > \beta_2 > \dots > \beta_m > 0$.

The peak period is period m and it is a reasonable assumption that the first unit runs baseload.

3.9.2.1.i The Installed Stack and Available Technology Frontier

Figure 3.40 shows the construction of the plant cost envelope. In this case, the continuous envelope is constructed from the four actual

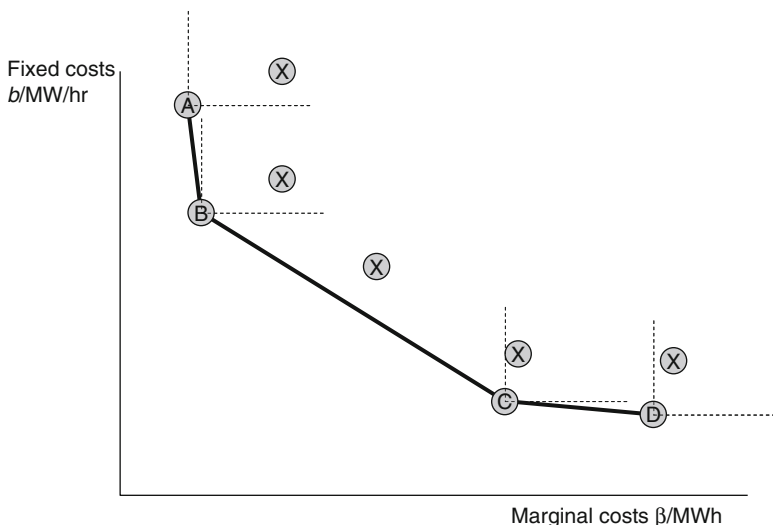


Figure 3.40 Plant envelope in the Crew and Kleindorfer analysis.

plants ABCD. Under deterministic conditions, and with divisibility, the performance of intermediate plant on the connecting lines can be achieved by weighted mixtures of these plants. The plants annotated “X” are subordinate plants that are not invested in. It is obvious from figure 3.40 that the plant technology envelope under the assumptions given must be downward sloping and non-concave at any point.

Plant is fully characterized by three parameters, capacity q , fixed cost β /kW/period, variable cost b /MWh. Since we have assumed full divisibility and constant returns to scale in energy (marginal costs) and capacity (fixed costs), q is not an important parameter for individual plant, though there may be some limit on total MW of plant available of any type.

We need to consider the CK assumption that the optimum plant envelope can be constructed by linear interpolation of lower bound points. This makes interlinked assumptions of i) divisibility and ii) constant returns to scale for both short run and long run marginal costs. Assume that we have two plants a 500 MW plant of £40/kw/year and £30/MWh and a 500 MW plant of £44/kw/year and £28/MWh. CK assume that we can deliver 500 MW at £42/kw/year and £29/MWh by delivering 250 MW at £40/kw/year and £30/MWh and 250 MW at £44/kw/year and £28/MWh.

3.9.2.1.ii Unit Operation

The stack that runs is then all “in-merit” (in variable cost order) plant running at full load, and the marginal plant running at part load.

We can represent the energy produced in period i by plant l formally by

$$q_{li}(x_i, \bar{q}) = \min \left[\left(x_i - \sum_{k=1}^{l-1} q_{ki}(x_i, \bar{q}) \right), \bar{q}_l \right],$$

where $\bar{q} = (\bar{q}_1, \dots, \bar{q}_m)$ represents the vector of installed capacities of plant types 1 to m . This is shown in the figure 3.41 below.

Note that this equation caters for the possibility of part load. In reality, the construction of the framework does not require part loading. This is due to the discretization of the stack and the time periods, and the constant returns to scale.

3.9.2.1.iii Stochastic Demand

$$D_i(p_i, \bar{u}_i) = X_i(p_i) + \bar{u}_i.$$

3.9.2.1.iv Build Up of the Welfare Equation

The output of the l th unit in period i is $q_{li}(D_i(p_i, \bar{u}_i))$.

Here \bar{u} is an outcome of the stochastic variable \bar{u} , which has actuarial parameters that are known and constant. u is the more generic term for the variation.

We assume efficient allocation for $u = 0$. There is no running in excess of capacity, so for peak period $i = n$ $\sum_{l=1}^{m-1} \bar{q}_l \geq D_i(p_n, u_n)$ as we see in figure 3.41.

Demand is periodic, and since we are not concerned with start costs or capacity changes during the total period, and assume zero cross-elasticity between periods and independent stochastic shock to periods, we can rank in a load duration curve.

We denote the total installed capacity by z , where $z = \sum_{l=m}^1 \bar{q}_l$.

Let us specify the shortfall of generation capacity in any period by $D_i(p_i, \bar{u}_i) - z = \bar{q}_{si}$.

We denote the total output in period i by S_i . This is the minimum of demand and capacity

$$S_i(p_i, u_i, z) = \min [D_i(p_i, u_i), z].$$

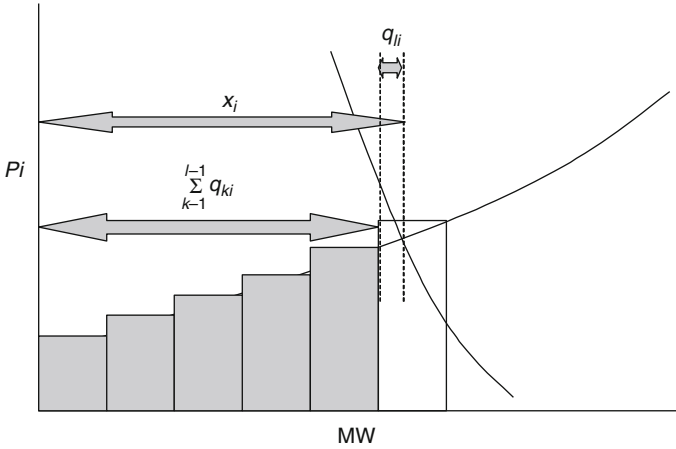


Figure 3.41 Marginal generation and demand curves in six of the n periods.

The terms added to arrive at the total welfare $W(u, p, \bar{q})$ are i) gross consumers' surplus, ii) variable cost, iii) fixed cost, and iv) rationing cost (efficient cost plus inefficiency).

The consumer surplus, before payment, and ignoring rationing, over all periods is

$$\sum_{i=1}^n \int_0^{S_i(u_i, u_i, z)} P_i(x_i - u_i) dx_i \tag{3.35}$$

The producer cost is

$$-\sum_{i=1}^n \sum_{l=1}^m b_l q_{li} (D_i(p_i, u_i), \bar{q}) - \sum_{l=1}^m \beta_l \bar{q}_l,$$

where \bar{q} is the installed capacity vector.

CK assume that the surplus loss under rationing is a generalized upward sloping linear function:

$$\bar{R} = \sum_{i=1}^n r_i(\bar{q}_{si}) = \sum_{r=1}^n r_i(D_i(P_i, \bar{u}_i) - z).$$

Collecting all of these terms, we have the welfare equation

$$W(u, p, \bar{q}) = \sum_{i=1}^n \int_0^{S_i(u_i, u_i, z)} P_i(x_i - u_i) dx_i \\ - \sum_{i=1}^n \sum_{l=1}^m b_l q_{li} (D_i(p_i, u_i), \bar{q}) - \sum_{l=1}^m \beta_l \bar{q}_l - \sum_{i=1}^n r_i (D_i(p_i, u_i) - z)$$

To arrive at the first term on the right-hand side, we do the following:
Using the chain rule

$$\frac{\partial W}{\partial P} = \frac{\partial W}{\partial S} \frac{\partial S}{\partial P}.$$

For a regular integral, we may interchange the order of differentiation and expectation.

$$\frac{\partial S}{\partial P} = \frac{\partial}{\partial p_i} \left(\sum_{i=1}^n \int_0^{S_i(p_i, u_i, z)} P_i(x_i - u_i) dx_i \right) \\ = P_i(S_i(p_i, u_i, z) - u_i) \frac{\partial}{\partial p_i} S_i(p_i, u_i, z) \\ = p_i X'_i(p) \text{ if } u_i < z - X_i(p_i) \\ = 0 \text{ if } u_i > z - X_i(p_i)$$

3.9.2.1.v Derivatives of the Welfare Formula

$$\bar{W}(p, \bar{q}) = E_{\bar{u}} [W(\bar{u}, p, \bar{q})].$$

This is what we wish to maximize with respect to q_k and p_i . So we solve the installed volume from $\frac{\partial W}{\partial q_k} = 0$, and we solve the optimal price vector from $\frac{\partial \bar{W}}{\partial p_i} = 0$.

3.9.2.2 Capacity Optimization

Here we have used the above definition of S_i and $P_i = X_i^{-1}$.

The results are, for the capacity of the k th unit under given prices:

$$\frac{\partial W}{\partial q_k} = \sum_{i=1}^n \left\{ \int_{z - X_i(p_i)}^{\infty} P_i(z - u_i) dF_i(u_i) - b_k [1 - F_i(\bar{Q}_k - X_i(p_i))] \right\} \\ + \sum_{l=1}^m b_l [F_i(\bar{Q}_l - X_i(p_i)) - F_i(\bar{Q}_{l-1} - X_i(p_i))] \\ - \beta_k + E_{u_i} \{ r'_i(X_i(p_i) + \bar{u}_i - z) \}. \quad (3.37)$$

So here we optimize the installed capacity and running load, a unit at a time, in relation to the demand vector and the available frontier of m units. This is a variant of the Turvey approach. We can simplify the equation above by equating the number of units and the number of subperiods and ensuring that we have enough capacity for no rationing.

$$\begin{aligned} \frac{\partial W}{\partial q_k} &= \sum_{i=1}^n \left\{ \int_{z-X_i(p_i)}^{\infty} P_i(z-u_i) dF_i(u_i) - b_k [1 - F_i(\ddot{Q}_k - X_i(p_i))] \right\} \\ &+ \sum_{i=1}^n b_l [F_i(\ddot{Q}_l - X_i(p_i)) - F_i(\ddot{Q}_{l-1} - X_i(p_i))] - \beta_k. \end{aligned}$$

3.9.2.3 Price Optimization

For the price in the i th period, under given capacities

$$\begin{aligned} \frac{\partial \bar{W}}{\partial p_i} &= p_i X'_i(p_i) F_i(z - X_i(p_i)) - \sum_{l=1}^m b_l X'_i(p_i) \left[\begin{array}{l} F_i(\ddot{Q}_l - X_i(p_i)) \\ - F_i(\ddot{Q}_{l-1} - X_i(p_i)) \end{array} \right] \\ &- X'_i(p_i) E_{\bar{u}_i} [r'_i(X_i(p_i) + \bar{u}_i - z)]. \end{aligned} \quad (3.38)$$

Here

$$X'_i(p_i) = \frac{\partial X_i(p_i)}{\partial p_i}$$

and \ddot{Q}_l is the capacity up to plant l $\ddot{Q}_l = \sum_{k=1}^l \bar{q}_k$, $l = 1, \dots, m$. So $z = \ddot{Q}_m$.

For simplification of notation, CK define

$F_i^l = F_i(\ddot{Q}_l - X_i(p_i)) = \Pr\{D_i(p_i, \bar{u}_i)\} \leq \ddot{Q}_l$ for all i, l where F as usual denotes the cumulative probability distribution function.

They also define $F_i^m = F_i(z - X_i(p_i))$.

We also note that $F_i(\ddot{Q}_l - X_i(p_i))$ is the probability that the capacity \ddot{Q}_l of the first l units exceeds the demand $X_i(p_i)$.

Setting $\partial W_i / \partial p_i = 0$, we have

$$\begin{aligned} p_i X'_i(p_i) \Pr\{D_i(p_i, \bar{u}_i) < z\} &= \sum_{l=1}^m b_l X'_i(p_i) \Pr\{\bar{Q}_{l-1} \leq D_i(p_i, \bar{u}_i) \leq \bar{Q}_l\} \\ &+ X'_i(p_i) E_{\bar{u}_i} \{r'_i(D_i(p_i, \bar{u}_i) - z)\} \end{aligned} \quad (3.39)$$

The left-hand side is the expected marginal benefits in the form of revenue and consumer surplus at price p_i and the right-hand side is the sum of expected marginal operating costs and rationing costs.

There is some resonance here with the way Chao breaks down his similar equation. The apparent absence of fixed costs in the equation above is described partly by our analysis of Chao in section 3.11 and partly by our analysis of Dansby in section 3.10. In brief, the fixed and variable costs are so bound together that we could replace the term b_i by $b_i(\beta_i)$

So in equation (3.39), we divide all terms by $X_i'(p_i)$ and by F_i^m to get

$$p_i = \sum_{l=1}^m \gamma_{li} b_i + E \left\{ \gamma_i'(X_i(p_i) + \tilde{u} - z_i) \right\} / F_i^m \text{ for } i = 1, \dots, m, \quad (3.40)$$

where $0 < \gamma_{li} < 1$ is defined by

$$\gamma_{li} = (F_i^l - F_i^{l-1}) / F_i^m = \Pr \left\{ \ddot{Q}_{t-1} \leq D_i(p_i, \tilde{u}_i) \leq \ddot{Q}_t \mid D_i(p_i, u_i) \leq z \right\}$$

We can therefore see that the price in each period is the conditional expectation of variable operating costs of the marginal unit plus rationing costs. The condition is that capacity exceeds demand.

3.9.3 Discussion of the Crew and Kleindorfer Analysis

CK set the optimum price at the variable cost of the marginal unit plus an allowance for rationing that effectively treats rationing as an extra cost that can be regarded as a power station unit of infinite size and no fixed costs or demand response.

In nonequilibrium conditions, this may give a small excess profit to the nonmarginal unit but this disappears as the divisibility of unit size and pricing period increases.

CK are explicit in not requiring units to recover their costs. Though we showed in equation (3.8) in section 3.1.3.5 that this pricing regime can ensure the recovery of all costs, provided the peak unit covers cost or DSM can be used as virtual production. In fact, the recovery of fixed cost is implicit in the CK analysis as a version of the Turvey algorithm appears in equation (3.37). It is nevertheless of interest that CK do not place high importance on financial equilibrium.

Divisibility of time period and unit size is clearly important, especially in the peak. We showed in section 3.1.3 that the peak price plays an important role in fixed-cost recovery for all units. This is not covered in the CK analysis, but is in the Dansby analysis.

3.10 FURTHER DEVELOPMENT OF THE PRICE VECTOR—FROM DANSBY

The absence of fixed-cost terms in optimum price formulae of CK, Chao, and others might lead a reader to infer that these analyses correctly indicated marginal cost pricing. This is not the case. The analysis of Dansby (1978) in a quasi-deterministic framework provides a useful setting to discuss this.

3.10.1 Framework

The Dansby framework is:

1. constant returns to scale in fixed and variable costs
2. a continuous load duration curve (i.e., infinite number of subperiods for demand)
3. a fixed number n of subperiods j for the purposes of pricing, although not necessarily the same length
4. deterministic elastic demand
5. a discrete number m of technologies l on the technology frontier.

3.10.2 The Analysis

3.10.2.1 Welfare

Dansby then takes total welfare as the net consumer surplus and total costs, and applies a Lagrangean function to maximize welfare, subject to price in each pricing period, with the constraint that load is fully served in each period. That is, if demand at price P is $Q(P)$, then there is sufficient capacity to deliver $Q(P)$.

$$W = \sum_{j=1}^n \int_{\tau_{j-1}}^{\tau_j} B_j(P_j, t) dt - \sum_{l=m}^1 \left[\beta_l K_l + \sum_{j=n}^1 b_l \int_{\bar{K}_{l-1}}^{\bar{K}_l} H_j(P_j, x) dx \right] - R$$

W is the total welfare over the period

$B_j(P_j, t)$ is the gross consumer welfare rate (i.e., £/hr) at time t , if the price is P_j , and assuming that the load is delivered

\bar{K}_l is the capacity of the l th unit

\bar{K}_l is the total capacity of all units up to and including the l th unit

b_l, β_l are the variable and fixed costs of the l th unit. $b_{l+1} > b_l, \beta_{l+1} > \beta_l$

P_j is the single price across the whole j th subperiod

$H_j(P_j, x)$ is the amount of time in period j for which the demand at price P_j exceeds a level x

R is the rationing cost, which Dansby sets to zero by ensuring sufficient capacity.

3.10.2.2 *Optimal Pricing with Full Divisibility*

The Lagrangean application to the welfare equation, with the constraint that capacity is not less than demand, is:

$$L = \sum_{j=n}^1 \int_{\tau_j}^{\tau_{j+1}} B_j(P_j, t) dt - \sum_{l=m}^1 \left[\beta_l K_l + \sum_{j=n}^1 b_l \int_{\check{K}_{l-1}}^{\check{K}_l} H_j(P_j, x) dx \right] - \mu_1 (\check{K}_m - Q_j(P_j)).$$

We have assumed that capacity is a public good across all periods, and therefore the constraint only bites in the peak subperiod (subperiod 1).

Dansby then optimizes using the Kuhn Tucker conditions applied in effect as a generalization of the BJ method. Of key importance for us is that in Dansby's analysis the term $\frac{\partial K_l}{\partial P_j} = 0$, for all l and j , where

K_l is the capacity of unit l and P_j is the price of period j . So the price is optimized on the basis of a given installed stack, and the capacity is optimized on the basis of least cost optimization to a given load profile. For off-peak periods we have

$$W_j = \int_{\tau_j}^{\tau_{j+1}} B_j(P_j, t) dt - b_l \int_{\check{K}_{l-1}}^{\check{K}_l} H_j(P_j, x) dx - \sum_{l=m}^1 [\beta_l K_l].$$

Let us simplify by assuming that the demand function has the same slope θ in all periods, that is, quadratic utility. Then using the standard Hotelling analysis for deadweight loss, we have

$$\frac{\partial W}{\partial P_j} \Big|_{l=1} \sum H_l < t < \sum_l H_l = \frac{\int_{\frac{1}{2} \sum_l H_l}^{\frac{1}{2} \sum_l H_l} \Delta P_t dt}{\sum_l H_l - \sum_{l=1}^1 H_l} * \theta * H_l,$$

where ΔP_t is simply the price above or below the optimal price, which is the variable cost of the marginal unit at time t , and the first expression on the right-hand side of the equation is the average ΔP_t over the time for which unit l would set the price if the subperiod lengths were infinitely small.

For infinitely small periods, then the price in each subperiod is exactly equal to the variable cost of the marginal unit, and $\frac{\partial W_t}{\partial P_t} = 0$ at all times, that is, pricing is optimal.

The averaging of prices for finite ΔP_t can be seen to be a result of the inefficiency of the timing of the subperiods. We also note that if the technology frontier is continuous, then it is inefficient for the number of units not to equal the number of subperiods, that is, for optimality $n = m$.

3.10.2.3 *Optimal Pricing without Full Divisibility*

Dansby maintains a high degree of generalization throughout, and we briefly summarize his analysis below. Figure 3.42 shows the continuous inelastic load duration as it intersects the installed power generation stack

Figure 3.43 shows the adaptation of figure 3.42 to accommodate price elasticity. To simplify we have depicted a single slope for all demand functions.

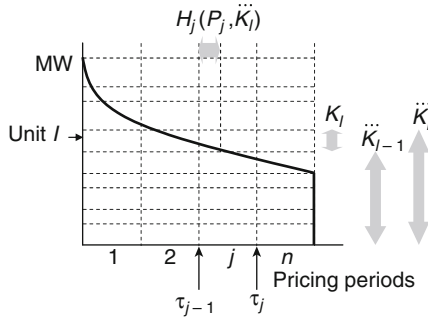


Figure 3.42 The loading of unit l in period j .

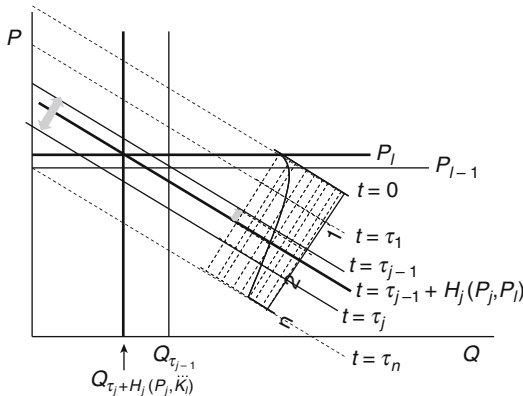


Figure 3.43 Demand functions at different times within the period referenced against the final load duration curve.

Given the construction of the Kuhn Tucker conditions, the constraint is slack in all but the peak period, which gives the optimal price equation for all periods but the peak period. Setting $\frac{\partial L}{\partial P_j} = 0$, we have

$$P_j^* = \sum_l b_l \alpha_{jl}$$

where P_j^* is the welfare optimal price in subperiod j

l is the index number of the lowest merit unit that runs in period j

$$\alpha_{jl} = \frac{\int_{x=\check{K}_{l-1}}^{x=\check{K}_l} \frac{\partial H_j}{\partial P_j} dx}{\int_0^{x=\check{K}_m} \frac{\partial H_j}{\partial P_j} dx}$$

As we have expressed in this formula,⁴⁰ the numerator and denominator have the units of time, and the denominator is equal to the subperiod length.

For all periods in which the unit is not loaded for part of the period (i.e., for those in which it is fully loaded or not loaded), then $\frac{\partial H_j}{\partial P_j} = 0$ and for the part loaded periods, then $\frac{\partial H_j}{\partial P_j}$ is the slope of the

load duration curve and $\int_{x=\check{K}_{l-1}}^{x=\check{K}_l} \frac{\partial H_j}{\partial P_j} dx$ is the time spent loaded. The denominator is the period length.

Expressed crudely, we have, as depicted in figure 3.44,

$$\frac{\Delta H_{lj}}{\Delta P_j} = \frac{\Delta H_{lj}}{\Delta Q_j} * \frac{\Delta Q_j}{\Delta P_j},$$

$\frac{\Delta Q_j}{\Delta P_j}$ is the negative of the reciprocal of the slope of the demand function, that is, $-1/\theta$

$\frac{\Delta H_{lj}}{\Delta Q_j}$ is the negative of the reciprocal of the slope of the load duration curve at the last moment of operation of unit l , which is at the margin in period l . Let us call the negative of the slope φ and φ_l at the last moment of the loading of unit l . This is shown in figure 3.45.

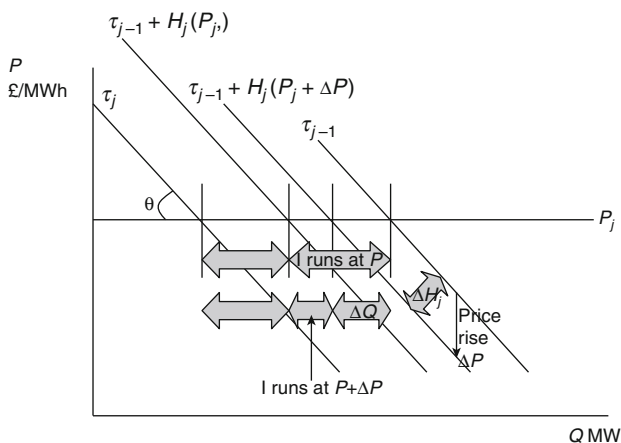


Figure 3.44 Depiction of reduction in run time of unit l in period j , resulting from a price increase ΔP .

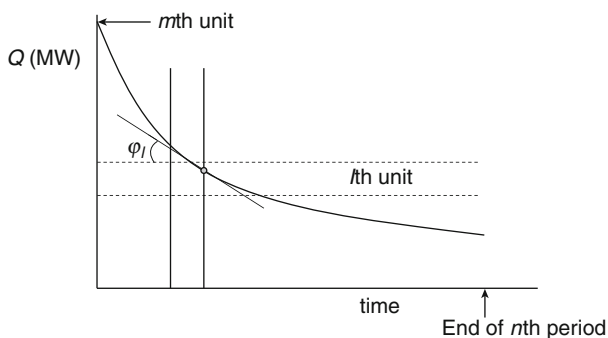


Figure 3.45 The slope ϕ of the load duration curve at the l th unit. See text.

So,
$$\frac{\Delta H_{lj}}{\Delta P_j} = \frac{1}{\theta \phi_l} .$$

Slightly more formally we have

$$\frac{\partial H_{lj}}{\partial P_j} = \frac{\partial H_{lj}}{\partial Q_j} * \frac{dQ_j}{dP_j} \Big|_{t=\tau_{j-1}+H_j} .$$

We have expressed the demand functions in terms of a family in which since the implied utility function is quadratic, we can regard the intrinsic utility function as identical in all periods, and the inter-period differences are explained by endowment differences. Using the

familiar form of the demand function $P_j = a_j - \theta Q_j$, if we designate q_j as the intercept of the demand function with the abscissa in period j , then we can see that $a_j = q_j \cdot \theta$, so q_j gives us the relative endowment in the period. Then φ is a proxy for the slope of the q_j duration curve.

So the price in all off-peak periods (all periods being off-peak except for the single peak period) is a convex combination of the variable costs of the different units. If the number of units and periods is the same and nominally infinite, then this resolves to variable cost pricing in all but the peak period, which is infinitesimally short.

Let us now turn to the peak period. The averaging process across the subperiod is similar to that for off-peak prices. The welfare optimal price of the peak period, given the installed stack, is

$$P_1^* = \sum_{l=m}^1 b_l \alpha_{il} + \beta_m \left[\frac{\frac{\partial Q_1}{\partial P_1}}{\int_0^{\tau_1} \frac{\partial Q_1}{\partial P_1} dt} \right] \tag{3.41}$$

This is the key equation. On the left-hand side we see that all units that are marginal in the price period have a role in price setting, and the indivisibility causes a cost shortfall for the lowest merit units, as the price falls below their variable costs.

The right-hand side shows the fixed cost contribution. It is highly dependent on unit divisibility. If the inverse demand function is flat in the peak period, and only the peak unit runs in the peak period, then the peak price is $P_l = b_l + \beta_l$.

In optimizing, we find a version of the Turvey equation:

$$\frac{\beta_l - \beta_{l+1}}{b_{l+1} - b_l} = \sum_j H_j(P_j, \bar{K}_l).$$

3.10.3 Discussion of the Dansby Analysis

Dansby considers a discrete stack facing a continuous load duration curve. In requiring the optimum price to be a combination of the variable costs of the units that set price at any time in subperiod, he therefore prices below variable costs for the marginal unit. As divisibility of the stack increases, and the pricing subperiods shorten to match, the theoretical problem of not covering fixed costs disappears.

The Dansby approach is quite different to that of Crew and Kleindorfer and allows us to model divisibility discretely. However,

as with CK, we find in the analysis a version of the Turvey algorithm, and in doing so implicitly recognize the consideration of fixed costs at the time of optimization.

3.11 STOCHASTIC VARIATION IN BOTH PRODUCTION AND DEMAND—THE CHAO FRAMEWORK

To a great extent, we can regard this framework as the conclusion of the peak load pricing debate, before moving to reliability options, the liberalization agenda, and thence to capacity and security of supply as a public good. The main paper draws together the threads of the literature to date, adds the final ingredient of plant reliability, and demonstrates variable cost pricing in a manner that is consistent with peak load pricing. There are additional results relating to lost load that we will examine.

In modeling the probability and impact of production shortfall, we commonly treat only production or only demand as stochastic, modeling the total effect by assuming similar stochasticity and loading all variance into either production or demand. So, production failure is treated as demand increase. For some applications we need to look in more detail into the specifics of one or other (as we did in rationalizing demand stochasticity in the Carlton framework), perhaps because our regulatory measures require separate treatment. Also, if we need to take a more sophisticated approach to consumer utility and rationing, we need to model stochastic demand and production functions separately.

Chao (1983) considers a single period in order to establish optimum pricing and capacity in the face of stochastic demand as well as power station failure, in circumstances where we cannot change prices following the resolution of uncertainty, and hence need to ration. In doing so, he allows us to investigate and test the sensitivity of welfare optimization to the specifics of the consumer demand function, and, by implication, the stochastic utility function. The sequence of decisions is shown in figure 3.46.

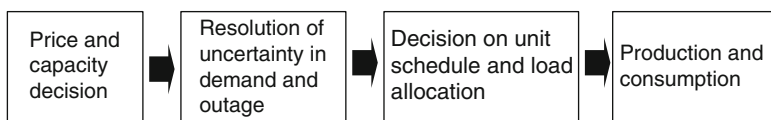


Figure 3.46 Representation of the Chao framework.

3.11.1 Framework

The framework is:

1. single period
2. multiple plant installation choice, on the technology frontier, with a finite number of units
3. constant returns to scale for capacity (and perfect divisibility)
4. constant returns to scale for energy
5. downward sloping convex stochastic demand function
6. stochastic availability of units⁴¹
7. specified rationing efficiencies.

3.11.2 Analysis

We write the demand

$$\tilde{D}_\epsilon(P) = \theta E \{ \tilde{D}(P) | \xi \},$$

where $\tilde{D}(P)$ is a random variable and the suffix ϵ the exogenous variable.

θ is the period length

ξ is a random event $\xi \in \Omega$, with Ω being the sample set

E is the expectation operator.

Chao assumes that the plant offer stack is the same as the variable cost stack, and there is no uplift added for fixed cost recovery. In doing so, he does not require all units to cover costs.

We define the total benefit (i.e., area under the marginal utility curve for a given ξ) of consumption (before payment) by $U(q, \xi)$ or simply $\tilde{U}(q)$. At this point there is no restriction to the form of this stochastic function other than the assumption that is always concave for all ξ and q , and later for the multisubperiod version, that the chronological order of the load duration curve does not change.

For a consumption volume $\tilde{D}_\epsilon(P)$, the marginal willingness to pay is the derivative of the benefit $P = \frac{\partial \tilde{U}(\tilde{D}_\epsilon(P))}{\partial P}$.

On the production side, we have γ_i MW installed for each unit with fixed costs k_i and variable costs C_i , so our total capacity cost over a unit time interval is $\sum_{i=1}^n k_i \gamma_i$. We can install as much or as little of any plant on the fixed marginal cost frontier with constant returns to scale.

Due to the impact of forced outages, our available capacity for each unit is a random variable \tilde{Y}_i , as we see in figure 3.47.

We assume that at the beginning of the period θ , but before the units are stacked for the schedule decision, that the available unit capacity is resolved. There is no restriction on the sophistication of this function.

The stochastic available capacity for unit i over the cycle is then found by integrating the marginal probability function over the capacity, so

$$\tilde{Y}_i = \tilde{Y}_i(\mathcal{Y}_i) = \int_0^{\mathcal{Y}_i} \tilde{y}_i(z) dz_i,$$

$\tilde{y}_i(z)$ is a stochastic factor with a uniform distribution between 0 and 1, and mean a_i . In the Chao analysis, it has the dimension of MW. For ease of exposition in this long analysis, we assume that all units are the same size and interpret $\tilde{y}_i(z)$ as dimensionless. This does not affect the result.

Our available capacity of our plant stack up to unit i is then

$$\tilde{Z}_i = \sum_{j=1}^i \tilde{Y}_j$$

So our power actually supplied up to any technology i , given the capacity constraint is

$\tilde{Q}(P, Z_i) = \min\{\tilde{D}(P), \tilde{Z}_i\}$ for all $i = 1, \dots, n$ (where the n th plant is the lowest merit unit required⁴²).

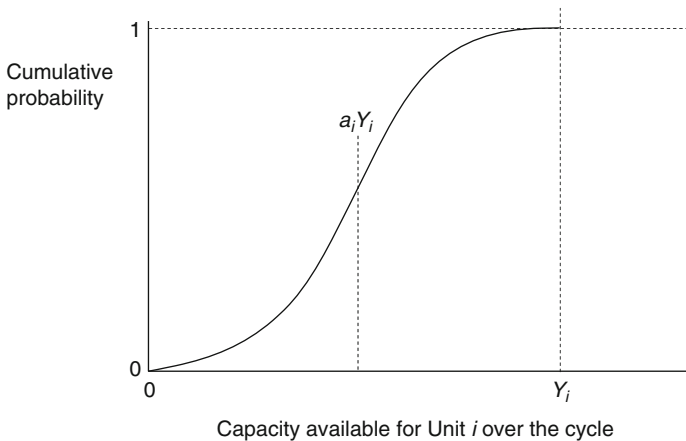


Figure 3.47 Stochastic cycle availability vector for unit i .

So the expected amount supplied by each technology is

$$\theta E \{ \tilde{Q}(P, Z_i) - \tilde{Q}(P, Z_{i-1}) \}$$

In section 3.9, we showed a similar representation of unit operation under the Crew and Kleindorfer analysis. Figure 3.48 shows the downrating total capacity from generator unavailability.

Figure 3.49 shows the effect of the position of the unit in the merit order on the effect of its failure on the relevant section of the cost function.

Our expected total variable running cost is

$$\sum_{i=1}^n \theta C_i E \{ \tilde{Q}(P, Z_i) - \tilde{Q}(P, Z_{i-1}) \}.$$

We have seen that there are numerous possible rationing schemes. In general, the expected loss of welfare in a period resulting from

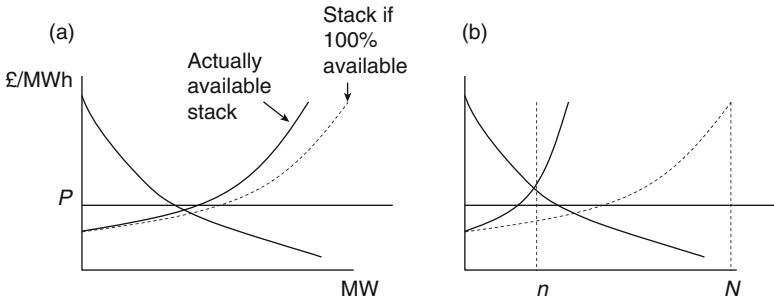


Figure 3.48 Unit operation for (a) Plant capacity sufficient; (b) Plant capacity insufficient.

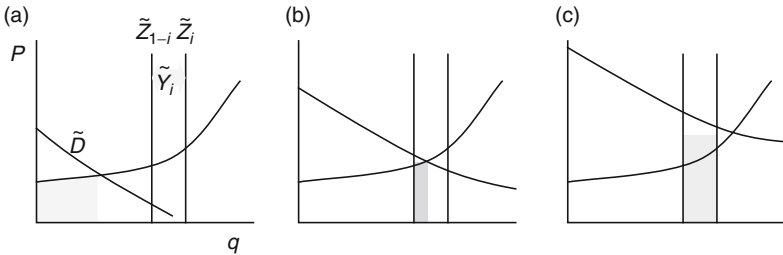


Figure 3.49 Effect of variation in the availability of a unit (a) Out of merit unit; (b) Marginal unit; (c) In merit unit.

outages is a function of the energy supplied and demanded. So we represent this stochastic cost \tilde{S} by

$$\tilde{S}(\tilde{Q}_e(P, Z_n), \tilde{D}_e(P)) \text{ where } \tilde{Q}_e(P, Z_n) \triangleq \theta E\{\tilde{Q}(P, Z_n) | \mathcal{E}_n^e\}. \quad (3.42)$$

We wish to maximize the expectation of social welfare, which is gross consumer benefit, assuming no rationing, minus capacity cost minus operating cost minus outage cost, by optimizing the price and the capacity levels of each technology.

We must first codify the shortage cost. Chao initially assumes a linear cost in relation to volume loss.

$\theta b E\{\tilde{D}(P) - \tilde{Q}(P, Z_n)\}$, where $\tilde{D}(P)$ is the quantity demanded, $\tilde{Q}(P, Z_n)$ and b is the multiplier arising from the demand function and the rationing regime. In effect b is the VOLL, and moreover we shall see that the problem takes the form of a system with no loss of load, but containing a unit of infinite size, zero fixed costs, and variable costs of b .

Using this, Chao builds the total expected social welfare from gross consumer benefit, minus the sum of capacity cost, operating cost, and outage cost. It is a minor point here but important in considering the value of “available capacity” (ACAP) obligations, that Chao does not scale down our fixed costs according to unit availability. So $\partial k_i / \partial a_i = 0$ and $\partial \sigma(k_i) / \partial a_i = 0$, where $\sigma(k_i)$ is the standard deviation of the stochastic availability. We will see in equation (3.43) the stochastic version of the Turvey equation that the fixed cost does actually get normalized by the availability.

$$W = E\{\tilde{U}(\tilde{D}_e(P))\} - \sum_{i=1}^n k_i \gamma_i - \sum_{i=1}^n \theta C_i E\{\tilde{Q}(P, Z_i) - \tilde{Q}(P, Z_{i-1})\} \\ - \theta b E\{\tilde{D}(P) - \tilde{Q}(P, Z_n)\} \quad (3.43)$$

Equation 3.43 Four terms of aggregate welfare. Gross consumer surplus, fixed costs, variable costs, rationing costs.

We can see from this equation that the consumer who loses load is effectively treated as a power generator with marginal cost of b , and infinite capacity.

To optimize the installed capacity, we need to set $\frac{\partial W}{\partial \gamma_i} = 0$ for all $i \in 1, \dots, n$, where n is the number of units installed.

3.11.2.1 Optimization of Capacity

If unit i is in merit (the offer price is less than the clearing price, with offers being made in variable cost order), the differential of energy delivered up to unit i , with respect to the availability of unit j is

$$\frac{\partial \tilde{Q}(P, Z_i)}{\partial \gamma_j} = \tilde{y}_j \text{ if } \tilde{D}(P) > \tilde{Z}_i \text{ and } j \leq i$$

$$\frac{\partial \tilde{Q}(P, Z_i)}{\partial \gamma_j} = 0 \text{ if } \tilde{D}(P) < \tilde{Z}_i \text{ or } j > i$$

As we can see in figure 3.50, the availability of the unit is only important if it is merit.

Since we hold the price constant, the volume delivered is then the lower of the stochastic demand at price P and the available capacity with variable cost less than P . We assume independence of stochastic forces, and so for an individual unit, the expectation of delivery is then the expectation of availability multiplied by the probability of demand exceeding the plant position in the merit order.

$$E\{\tilde{Q}(P, Z_i)\} = \min\{\tilde{D}(P), \tilde{Z}_i\}$$

So,

$$\frac{\partial E\{Q(P, Z_i)\}}{\partial \gamma_j} = E\{\tilde{y}_j \mid \tilde{D}(P) > \tilde{Z}_i\} \Pr\{\tilde{D}(P) > \tilde{Z}_i\} \text{ if } j \leq i \quad (3.44)$$

$$\frac{\partial E\{Q(P, Z_i)\}}{\partial \gamma_j} = 0 \text{ if } j > i \quad (3.45)$$

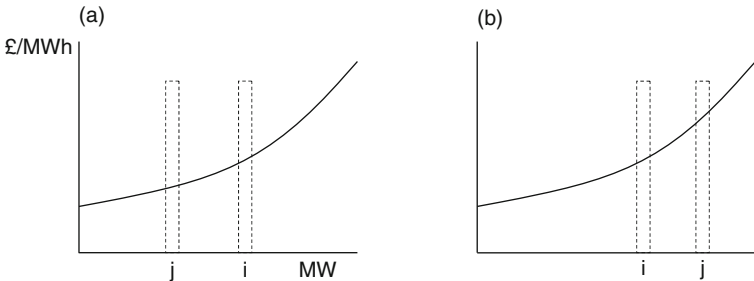


Figure 3.50 Dependence on the sensitivity of energy delivered up to unit i to the availability of unit j (a) $j < i$, $j > i$. See text.

The random variable \tilde{y}_j is assumed independent of $\tilde{D}(P)$ and \tilde{Z}_i , so the conditional expectation can be unconditional.

So equations (3.44) and (3.45) become

$$\frac{\partial E\{\tilde{Q}(P, Z_i)\}}{\partial \Upsilon_j} = a_j \Pr\{\tilde{D}(P) > \tilde{Z}_i\} \text{ if } j \leq i \quad (3.46)$$

$$\frac{\partial E\{\tilde{Q}(P, Z_i)\}}{\partial \Upsilon_j} = 0 \text{ if } j > i \quad (3.47)$$

where a_j denotes the availability of unit j , that is, the ex ante probability of being able to run when called.⁴³

So the derivative of the first part of the third term of equation (3.43) is

$$\frac{\partial}{\partial \Upsilon_j} \sum_{i=1}^n \theta C_i E\{\tilde{Q}(P, Z_i)\} = \sum_{i=j}^n \theta C_i [a_j \Pr\{\tilde{D}(P) > \tilde{Z}_i\}].$$

Now, turning to the second part of the third term in equation (3.43),

$$\frac{\partial E\{\tilde{Q}(P, Z_{i-1})\}}{\partial \Upsilon_j} = a_j \Pr\{\tilde{D}(P) > \tilde{Z}_{i-1}\} \text{ if } j \leq i-1$$

$$\frac{\partial E\{\tilde{Q}(P, Z_{i-1})\}}{\partial \Upsilon_j} = 0 \text{ if } j > i-1$$

Hence,

$$\begin{aligned} \frac{\partial}{\partial \Upsilon_j} \sum_{i=1}^n \theta C_i E\{\tilde{Q}(P, Z_{i-1})\} &= \sum_{i=j+1}^n \theta C_i [a_j \Pr\{\tilde{D}(P) > \tilde{Z}_{i-1}\}] \\ &= \sum_{i=j}^n \theta C_{i+1} [a_j \Pr\{\tilde{D}(P) > \tilde{Z}_i\}] \\ &= \sum_{i=j}^{n-1} \theta C_{i+1} [a_j \Pr\{\tilde{D}(P) > \tilde{Z}_i\}] + a_n \theta C_n \Pr\{\tilde{D}(P) > Z_n\}. \end{aligned}$$

So the whole of the derivative of the third term is,

$$\begin{aligned} &\frac{\partial}{\partial \Upsilon_j} \sum_{i=1}^n \theta C_i E\{\tilde{Q}(P, Z_i) - \tilde{Q}(P, Z_{i-1})\} \\ &= \theta a_j \sum_{i=j}^n (C_{i+1} - C_i) [\Pr\{\tilde{D}(P) > \tilde{Z}_i\}] + a_n \theta C_n \Pr\{\tilde{D}(P) > Z_n\}. \end{aligned}$$

Now, turning to the fourth term of equation (3.43), we have,

$$\frac{\partial(\theta b E\{\tilde{D}(P) - \tilde{Q}(P, Z_n)\})}{\partial \Upsilon_j} = 0 - a_i b \theta \Pr\{\tilde{D}(P) > \tilde{Z}_n\} \text{ for } j \leq i.$$

So finally we have, for all four terms in equation (3.43)

$$\begin{aligned} \frac{\partial W}{\partial \Upsilon_i} = 0 &= -k_i + a_i \sum_{j=i}^{n-1} \theta(C_{j+1} - C_j) \Pr\{\tilde{D}(P) > \tilde{Z}_j\} \\ &+ a_i \theta (b - C_n) \Pr\{\tilde{D}(P) > \tilde{Z}_n\}. \end{aligned} \quad (3.48)$$

We can regard the right-most term of this equation as denoting the aforementioned unit of infinite capacity and variable cost b .

In this equation, we are looking at the total welfare relative to the availability of the i th unit on the stack. The second term refers to the j th unit running under a particular outcome of $\tilde{D}(P)$.

For $i = n$,

$$\frac{\partial W}{\partial \Upsilon_n} = 0 = -k_n + 0 + a_n \theta (b - C_n) \Pr\{\tilde{D}(P) > \tilde{Z}_n\}. \quad (3.49)$$

Rearranging, we have,

$$\Pr\{\tilde{D}(P) > \tilde{Z}_n\} = \frac{k_n / a_n}{\theta(b - C_n)}. \quad (3.50)$$

For $i = n - 1$, we have,

$$\begin{aligned} 0 &= \frac{\partial W}{\partial \Upsilon_{n-1}} = -k_{n-1} + a_{n-1} \theta (C_{j+1} - C_j) \Pr\{\tilde{D}(P) > \tilde{Z}_j\} \\ &+ a_{n-1} \theta (b - C_n) \frac{k_n / a_n}{\theta(b - C_n)} \\ 0 &= -\frac{k_{n-1}}{a_{n-1}} + \theta (C_{j+1} - C_j) \Pr\{\tilde{D}(P) > \tilde{Z}_j\} + \frac{k_n}{a_n} \end{aligned}$$

Substituting the above equation into the one above, we have:

$$\Pr\{\tilde{D}(P) > \tilde{Z}_j\} = \frac{\frac{k_n}{a_n} - \frac{k_{n-1}}{a_{n-1}}}{\theta(C_{j+1} - C_j)} \text{ for } i = 1, \dots, n-1 \quad (3.51)$$

Note that the fixed cost k_n is normalized by the availability a_n .

Chao's interpretation of this is a stochastic version of the Turvey equation (3.2) described in section 3.31.

If demand and supply are certain, then the optimal running time for technology i is, from the Turvey formula,

$$\frac{(k_i - k_{i+1})}{\theta(C_{i+1} - C_i)}. \quad (3.52)$$

Now suppose that a fully operational ($a = 1$) unit of technology $i + 1$ is substituted by one of technology i . Multiply both sides of equation (3.51) by $\theta(C_{i+1} - C_i)$ giving us

$$\Pr\{\tilde{D}(P) > \tilde{Z}_i\} \theta(C_{i+1} - C_i) = (k_i / a_i) - (k_{i+1} / a_{i+1}). \quad (3.53)$$

We now interpret the left-hand side as the expected variable savings from the substitution, and the right-hand side as the capacity cost increase. Here we have weighted the capacity cost by availability. This is consistent with a forward-looking view of capacity in a competitive market, on the assumption that failure is a completely random event and that the welfare cost of rationing is linear.

The optimal running time is of some interest as it informs us of load factor, but since we always run plant in variable cost merit order, there is no real decision to make once we have installed the technology. The equation does, however, guide us to the optimal installation.

Having concluded a long work through of this framework, now let us turn to the key point of interest for us, which is the sensitivity to the form of the utility function.

3.11.2.2 LOLP, Optimum Capacity, and the Utility Function

Noting that, at least in the one period setting, $\Pr\{\tilde{D}(P) > \tilde{Z}_n\} = \text{LOLP}$, where LOLP is, as usual, the loss of load probability, then our formula for optimum LOLP_o is

$$\text{LOLP}_o = \frac{k_n / a_n}{\theta(b - C_n)}. \quad (3.54)$$

Now, we can express USE,⁴⁴ the expectation of loss, divided by the expectation of demand

$$\text{USE} = E\{\tilde{D}(P) - \tilde{Q}(P, Z_n)\} / E\{\tilde{D}(P)\}.$$

So the loss of load expectation, conditional on there being a loss of load is:

$$\text{LOLE} = \frac{\text{USE}}{\text{LOLP}} = E \left\{ \tilde{D}(P) - \tilde{Q}(P, Z_n) \mid \tilde{D}(P) > \tilde{Z}_n \right\} / E \left\{ \tilde{D}(P) \right\}$$

Chao is interested to show that LOLP is an oversimplistic objective measure when rationing costs are convex and we now develop this theme. We do this in stages.

1. simplification of LOLP and VOLL
2. application of simplified LOLP and VOLL with an extra term in the rationing polynomial.

3.11.2.2.i Simplification of LOLP and VOLL.

Chao initially simplifies with a single failure event possibility. So we make a simplifying equation of a single event losing L/M with probability $\text{LOLP} = \lambda M$. Formally, the conditional variance of LOLE is zero, and M represents a change of measure.⁴⁵

In essence, we have assumed a very simple distribution, that demand is fully satisfied with a probability $1 - \lambda$, or that there is, with a probability λ , some loss of load, and that the conditional expectation of loss of load is L , and the conditional variance is zero. This is similar to the assumptions used by Drèze and Brennan. If the distribution changed so that with a probability λM , we had a conditional expectation of losing L/M , then for a given USE the economics would be sensitive to the size of M .

In equation (3.43), the linearity of the rationing cost is implicit. We have noted that we can regard rationing as having DSM units with a variable cost b . So, when we need to calculate the conditional expectation of loss, we are not presented with complications. We consider that we lose LOLE with a probability of LOLP.

The motivation for using LOLP as an objective function is that it is practically feasible to estimate it. LOLE and USE are much harder. In this simple case, we can use the Black (1976) option formula to relate standard deviation, LOLP, LOLE, and USE together in a straightforward manner.

For policy, simplification, or other reasons, we can adjust M . Indeed, this is exactly what regulators do. So, for example, in England and Wales, LOLP was set artificially high, and VOLL artificially low. Two reasons for this were i) the political externality of the impact of lost load and ii) the use of VOLL as a regulatory price cap. In the

single Irish electricity market, there is an explicit “power factor”⁴⁶ that reduces LOLP in the peak period.

3.11.2.2.ii Addition of the Second Term to the Rationing Polynomial

Particularly since we recognize the application of $M \neq 1$, then it is apparent that in adjusting VOLL (and therefore LOLP), we are making some adjustment for the convexity of rationing costs. We can make a better adjustment by explicitly recognizing this convexity.

With convex rationing costs, which may arise, for example, from quadratic utility, the distribution is more important. For example, we would be sensitive to the size of M in the previous paragraph.

This next simplest polynomial generalization of our initial formula, which allows for demand convexity and/or rationing efficiency is

$$\tilde{S}(x, y) = b_1(y - x) + b_2 \frac{(y - x)^2}{y}.$$

Here we have dropped Chao’s $\frac{1}{2}$ multiplier on the right-hand side, as we can absorb it into b_2 . The range is depicted in figure 3.51.

Note that the second term on the right hand side contains $\frac{(y - x)^2}{y}$ rather than $(y - x)^2$ to preserve the correct dimensions of the equation, as the units of b are £/MWh and y and x are MW * hr.

$$\text{Our loss in } \pounds \text{ is } b_1 * \frac{L}{M} * \lambda * M + b_2 * \left(\frac{L}{M}\right)^2 * \lambda * M \equiv b * \frac{L}{M} * \lambda * M.$$

$$\text{So } b \equiv b_1 + b_2 \frac{\text{USE}}{\text{LOLP}}.$$

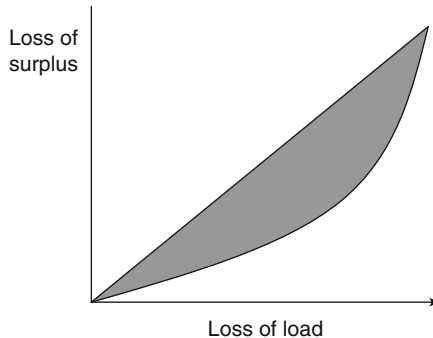


Figure 3.51 Loss of surplus in relation to loss of load. Feasible region between linear and quadratic shaded.

We substitute for this b in equation (3.50) and get,

$$\text{LOLP}_o = \frac{k_n / a_n}{\theta(b_1 + b_2 \frac{\text{USE}_o}{\text{LOLP}_o} - C_n)}$$

which simplifies to

$$(b_1 - C_n)\text{LOLP}_o + b_2\text{USE}_o = k_n / a_n \theta$$

So the reliability index that we need to optimize is a weighted average of LOLP and USE. Diversity improves both but has a much greater effect on LOLP. If we use lognormal distribution of demand increase plus production shortfall, then we can use the European option formula in section 5.2 to establish some simple relationships.

Let us first give some consideration to the VOLL. The Chao framework makes it quite compatible for us to regard lost load in terms of alternative production rather than actual loss. Indeed, for electricity, a finite value of VOLL is somewhat predicated on this. The alternative technology must have a fixed cost associated with it. So, for example, if it is a diesel generator, fully dedicated to electricity standby, then we must, if we do not include the unit in the production stack, load its fixed costs into its short-term price offer. If it has shared use (candles, flashlight, etc.), then we apply hedonic pricing to isolate that part of fixed costs that should be applied to electricity substitution. Finally, regarding load that was lost with no replacement, the consumer will do something else with the money saved from spend on electricity consumption spend. The activity would have entailed a fixed cost somewhere along the way.

3.11.2.3 Optimum Price

$$\frac{\partial \tilde{Q}(P, Z_i)}{\partial P} = \tilde{D}'(P) \text{ when } \tilde{D}(P) < \tilde{Z}_i \text{ for } i = 1, \dots, n, \text{ and } 0 \text{ for } \tilde{D}(P) > \tilde{Z}_i.$$

So,

$$\frac{\partial E\{\tilde{Q}(P, Z_i)\}}{\partial P} = E\{\tilde{D}'(P) | \tilde{D}(P) < \tilde{Z}_i\} \Pr\{\tilde{D}(P) < \tilde{Z}_i\}.$$

So, to optimize, we differentiate

$$\frac{\partial W}{\partial P} = 0 = \theta P D'(P) - \sum_{i=1}^n \theta C_i E \left\{ \tilde{D}'(P) \mid Z_i > \tilde{D}(P) \geq \tilde{Z}_{i-1} \right\}$$

$$\Pr \left\{ \tilde{Z}_i > \tilde{D}(P) \geq \tilde{Z}_{i-1} \right\} - \theta E \left\{ b \tilde{D}'(P) \mid \tilde{D}(P) > \tilde{Z}_n \right\} \Pr \left\{ \tilde{D}(P) > \tilde{Z}_n \right\}.$$

We can see in the second and third terms on the right-hand side that this requires a knowledge of the joint probability distribution of the marginal demand $\frac{\partial \tilde{D}(P)}{\partial P}$ and the total demand $\tilde{D}(P)$.

Accordingly, Chao takes two alternative demand specifications. First, as Chao expresses it, marginal demand is uncorrelated to total demand, and second, it is perfectly correlated. We can regard this as a development of the Drèze analysis, with more sophisticated production and demand functions, and distributional forms.

By representing two stochastic outcomes of the demand function, as shown in figure 3.52, we can see that the correlation between marginal demand and total demand is a function of the stochastic form of the demand curve. So if the shock u is additive to demand, then there is no correlation between marginal and total demand, but if the shock u is multiplicative to demand, then there is 100 percent correlation between them.

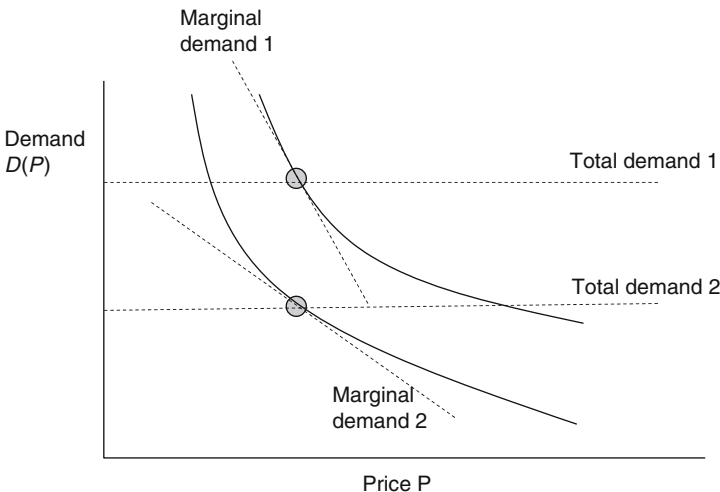


Figure 3.52 Correlation between marginal and total demand depends on the demand function shape and stochastic nature.

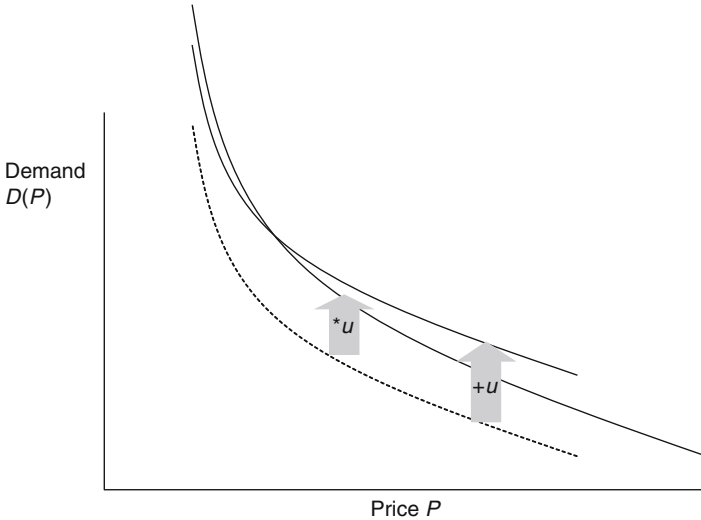


Figure 3.53 Effective of shock on the correlation of total and marginal demand. Additive (=u) with zero correlation and multiplicative (*u) with 100 percent correlation.

Figure 3.53 shows that the correlation between marginal and total demand is dependent on the form of the shock.

For zero correlation of total and marginal demand, the formula, which is a generalization of the CK formula is,

$$P_l = C_l + \sum_{i=1}^{n-1} (C_{i+1} - C_i) \Pr \{ \tilde{D}(P) > \tilde{Z}_i \} + (b - C_n) \Pr \{ \tilde{D}(P) > \tilde{Z}_n \}$$

Under this condition, the long run price is apparently lower than the total cost, and Chao expressly states that expected profits are negative (and hence fixed costs are not covered). Our rationalization of this follows the same logic as for the CK result and the Dansby result. In brief, under long-term equilibrium, the evolution of the stack is such that profits become zero.

Next we assume perfect correlation of marginal and total demand. Uncertainty now appears in multiplicative form, and in this case Chao states that the long run profit could be positive or negative according to rationing. The profits are zero with pro rata rationing, negative for perfect rationing, and positive if those who have lowest willingness to pay get served first.

More generally, we can express the optimum price as a weighted average of the correlated and uncorrelated components.

3.11.2.4 *Extension to Multiple Subperiods*

Finally, Chao extends the analysis to multiple periods, reasoning that the solution reduces to that of the single period. This effectively makes us of load factor duality described in section 2.4.1.

3.11.3 **Concluding Remarks on the Chao Analysis**

In summary, Chao then addressed a number of key features of the problem in hand, and his framework is sound, robust, and flexible. The elegance of the formal formulation does make the analysis rather intractable to the layman, and the extent of ground covered in a single paper precludes Chao from developing the theory to an examination to the sensitivity to probability distributions and utility functions. Nevertheless, he conclusively proves that while we can simplify a demand distribution with a single VOLL, that LOLP is even then an inappropriate objective measure where the circumstances of demand loss are such that utility concavity and/or rationing convexity/efficiency are a practical issue. The development of his theory, using USE as an objective function, has the same conclusion for the use of USE as a sole objective measure. Given that VOLL and USE are the dominant objective measures used in practice, these conclusions are important.

In reviewing the Chao framework, we reveal a key asymmetry in the analytics of production and consumption in the provision of energy in rare events. In assigning a de facto economic VOLL, which we interpret as voluntary DSM, we have created a virtual producer. This producer has no fixed costs, which, though a standard framework, is problematic. For the simplest analysis with a constant willingness to pay b , we can simply model the variable cost b as $b = b + B/\lambda$ where λ is the probability of lost load (with zero conditional variance) and B is a proxy for the fixed costs of maintaining alternatives to networked electrical consumption. A simple example would be having a portable generator.

Using the Chao framework, a more sophisticated analysis for a downward sloping demand function is possible, thereby creating a range of virtual producers, which can all be regarded as real producers.

This however does not optimize the “DSM stack.” This is because we have not at this point assigned a fixed cost to having the capability to lose networked power without harm. By simple assignment of fixed cost to DSM, we can use the Chao generalization of the Turvey

algorithm to determine the optimum amount of build and run of the lowest merit technology.

3.12 DISCUSSION OF THE PRICING ANALYSES

We find by working through the details of all papers in the canon that none contradict the theory of peak load pricing, in which we recover the fixed costs in the peak period/s. While many and perhaps most authors regard the recover of fixed costs as unnecessary or low priority, we can in all cases find either an implicit recognition of fixed cost in optimizing the power generation stack through a version of the Turvey algorithm or an explicit decision to make price optimization decisions regarding fixed costs as sunk and irrelevant.

The array of model features is rich, and each feature is important for practical modeling. These are:

1. periodicity (one or many periods)
2. discretization of the periods and a discontinuous technology frontier
3. returns to scale in capacity (discontinuity at zero load, slope, convexity)
4. returns to scale in operation (discontinuity at zero load, slope, convexity)
5. price cap in the highest peak period
6. dimensions in the technology stack (variable costs, fixed costs, divisibility, flexibility start costs)
7. plant aging characteristics (aging profiles, stack stationarity, temporal indivisibility)
8. the application of fixed costs to variable plant value (so-called Tobin Q)
9. consumer utility function (Marshallian or various concave forms)
10. producer utility function (weighting relative to consumers', cost of risk)
11. drivers of stochastic demand (number of consumers, demand of individual consumers)
12. homogeneity/heterogeneity of individual consumers
13. boundaries of the economy (visitors, immigrants, unborn, nonconsumers)
14. producer response timeframe (stack evolution, investment, price offer strategy, price offer)
15. actuarial consistency of stochastic variations (stationary or not, risky or uncertain)

16. chronological order of the load duration curve and ‘principal component’ shock to this
17. the complexity of the demand shock vector across different periods
18. the individual or joint optimization of price and of capacity build
19. consumer response timeframes
20. the degree to which electricity can be a private good
21. the rationing method
22. gaming (Cournot, market power, stationarity of unit offer structures).

We established that if all units regard themselves as price takers rather than price makers, and hence we can ignore gaming, evolution of the installed generation stack in terms of fixed and variable cost and volume installed, tends to self-optimize toward equilibrium, with both an engineering and finance (fixed cost) response.

The treatment of lost load is somewhat problematic, as the ideal is to view lost load as a series of DSM contracts that are like virtual power stations, with both fixed and variable costs. In fact there is minimal empirical observation.

We have largely ignored gaming considerations. In general, the analysis is robust with respect to gaming. For example, with the optimal price in a subperiod in equilibrium being the variable cost of the marginal unit, the Stackelberg game is in operation with no distortion. By far the most important game is the feared or actual expropriation of the rent of peaking units by the state and the reciprocal fear by the state that peaking units will lever market power.

RELAXING THE HARD CAPACITY CONSTRAINT

4.1 INTRODUCTION

Most authors in the canon have simplified plant costs by assuming a single variable cost up to a hard capacity limit at which variable costs become infinite. The cost of capacity then tends to have constant returns to scale. If we have decreasing returns to scale in operation, then the concept of capacity becomes harder to define. Since decreasing returns to scale in operation is a reality, we must attend to this.

4.2 THE HIRSHLEIFER FRAMEWORK

Hirshleifer (1958) offers a tantalizing view of this in a paper that, in the author's view, had the potential to change the development of the canon on capacity costing, pricing, optimisation, and regulation, but did not do so. We now use his framework to develop the argument.

Hirshleifer first begins with the Steiner framework that we define as

1. a setting with two equal length period
2. constant returns to scale in capacity, starting at the origin and with no upper limit. There is no "fixed operational cost," that causes a step change as operation increases from zero.
3. constant returns to scale in operation, starting at the origin,¹ up to the capacity limit
4. full divisibility (this is a direct consequence of the above two statements)
5. downward sloping linear demand curves that are deterministic in the setting considered, but not necessarily known before the capacity build decision.

Because the Steiner diagrammatic representation of demand for capacity cannot be used for relaxation of the hardness of the capacity constraint, Hirshleifer first presents the Steiner framework in a different diagrammatic representation. In particular, Hirshleifer shows the actual demand function rather than the demand function net of variable costs (the so-called demand for capacity).

This is shown in figure 4.1. For ease of reference, we have used the same nomenclature for fixed and marginal costs as he did. b per period are the fixed costs and β the variable costs. In the first instance, Hirshleifer supposes that the capacity decision has been made prior to the resolution of the demand curve, and/or that the installed capacity Q_A is less than would optimally be installed with knowledge of the demand function. In this case, Hirshleifer states that the prices in the peak and off-peak are represented by P and S . We discuss the validity of this statement below. Hirshleifer then supposes that the demand function is known prior to capacity commitment. The optimum build is then Q_B , the optimum pricing is β in the off-peak, and $\beta + 2b$ in the peak. We ignore peak shifting here.

Let us consider the optimum pricing and build under hard constraint. These prices are T and V . Since the two-period pricing with full capacity cost loaded into the peak is somewhat trivial, we wish to confirm that the Hirshleifer framework is equivalent to the Steiner framework in the shifting peak case, in which the off-peak period should pay above variable costs.

Hirshleifer follows convention in pricing at the intersection between the demand function and the vertical constraint, but this pricing in practice makes an assumption on build optimization, and therefore should not be considered as optimal short run pricing. In particular,

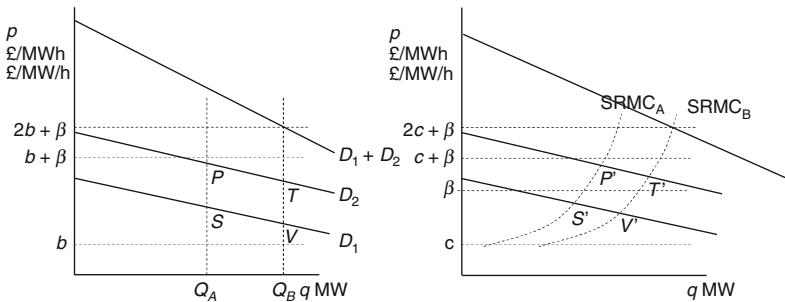


Figure 4.1 Vertical addition of demand curves for the constrained flat marginal cost and the Ricardian function. Adapted from Hirshleifer (1958). b, c are marginal costs, β is marginal costs, Q_B is the installed sufficient capacity, and Q_A a smaller capacity.

we should not regard the vertical section of the cost function as in any sense representing a part of the short run cost function whose intersection with the demand curve can be used for optimal pricing, but rather the intersection with the vertical line represents the price at which we most efficiently ration the demand. The price, as long as it exceeds variable costs, has no effect on aggregate production.

Hirshleifer now states that the Steiner framework with a hard constraint is simply a special case of a more general framework, which we here call the Hirshleifer framework, where the constraint is softer. Hirshleifer posits a curve that (probably asymptotically) approaches the right-angled function at low loads and at nominal capacity. The Hirshleifer framework is then presented, in diagrammatical form as shown in figure 4.1, as a generalization of the Steiner framework.

First, under hard constraint, let us again assume that we have made the capacity decision prior to the resolution of demand uncertainty, and that this capacity turns out lower than the optimum that would have been built if we had known the demand function. Hirshleifer states that the peak and off-peak prices should be P and S as shown in figure 4.1. He then allows capacity build after the resolution of demand uncertainty and arrives at the prices T and V .

Now, under soft constraint, suppose again that we have somehow ended up with less capacity than we would have built with knowledge of the demand function. Hirshleifer proposes prices of P' and S' in the peak and off-peak periods respectively. While being presented as essentially the same as the Steiner framework, we note i) that different prices now engender different produced volumes and ii) that we have explicitly ignored fixed costs. This is definitively short run pricing.

Now suppose that we know the demand function in advance and accordingly build optimally. Hirshleifer again presents a curve to use as a generalization of the right-angled function. Note however, that we now make explicit the cost dominance of the larger production volume. This was the case with the right-angled function, but less obviously so. It is quite clear that if the larger production capacity has a cost dominance in variable costs, it must have higher fixed costs, or it would not be on the technology frontier that we choose from. It is now obviously incorrect to have a single fixed cost c in the low-build and high-build cases. Further, the returns to scale in capacity cannot be linear from the origin, and must either start above the origin, or be convex, or both.

So, while not following Hirshleifer's conclusion that short run pricing is best after all, we do use his work to demonstrate the key relationship between short and long run costs.

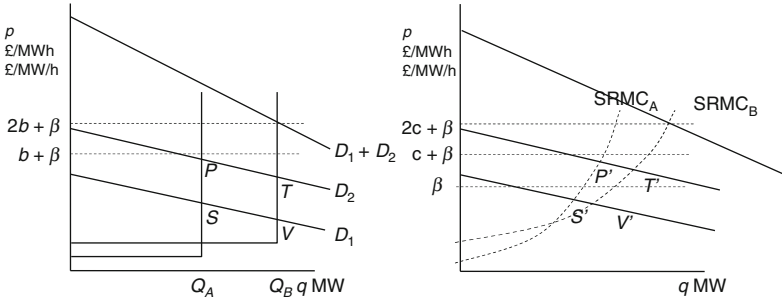


Figure 4.2 Hirshleifer analysis recast with a feasible production set.

The Hirshleifer example, recast using a Crew and Kleindorfer feasible plant frontier, is shown in figure 4.2.

In modeling, decreasing returns to scale with production, and thereby reducing the hard capacity constraint, we must ensure a technology frontier by allowing the fixed/variable cost curves of units to cross. In doing so, we demonstrate that peak pricing and not variable cost pricing is correct. This conclusion reinterprets the work of Hirshleifer.

We now turn to a more formal treatment of the soft capacity constraint by using decreasing returns to scale.

4.3 OPTIMIZING WITH DECREASING RETURNS TO SCALE—THE PANZAR FRAMEWORK

The Hirshleifer framework attempts to draw together the short- and long-term cost structures of a plant into a single convex upward sloping curve. Panzar does something rather similar and uses his framework to make some conclusions about the pricing of periodic demand and the cost recovery by generators. He posits a convex short run cost, and a constraint that is limited by the spend on capacity.

Steiner showed that if elasticity is such that demand management in the peak can bring the peak price below the off-peak price, then both periods should pay some contribution to capacity. Panzar (1976) goes further and suggests that all periods should pay some capacity cost. We examine this here.

Panzar used “neoclassical” (upward sloping quasi-convex) marginal cost curves. He noted that the conclusions of much of the work to date rested on the assumption of constant returns to scale (“fixed

proportions” of fixed and marginal costs). He concluded that under an assumption of $e_f < 1$, where e_f is the elasticity of scale of short run costs, that i) even under deterministic conditions, all periods should pay a contribution to capacity costs, not just the peak period and ii) it is efficient under the $e_f < 1$ assumption not to use full capacity at any time.

Panzar used a purely mathematical approach with little reference to physical terminology, and, as ever, the conclusion is dependent on the assumptions. Let us first rationalize the description in terms of power stations.

4.3.1 Framework

n equal length subperiods from $t = 1, \dots, T$

m units

Decreasing returns to scale in operation (convex variable cost)

Decreasing returns to scale in capacity/size (convex fixed cost). Here “size” and “capacity” could be regarded differently. Capacity refers to an absolute constraint and size conforms to the running load at maximum average efficiency (i.e., power output divided by fuel input).

Perfect reliability

The analysis is not readily interpretable in a physical sense but figure 4.3 shows one possibility.

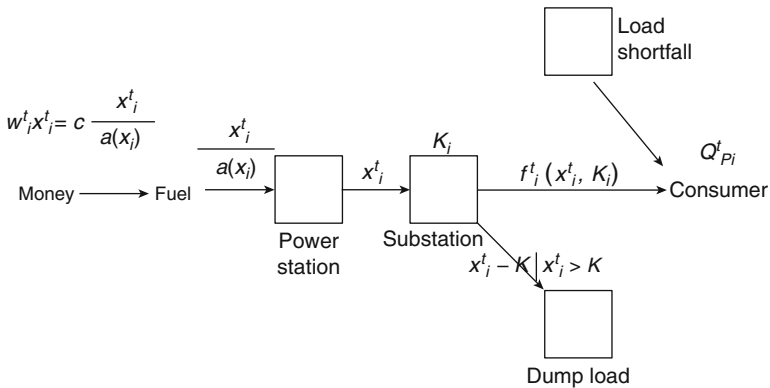


Figure 4.3 A physical representation consistent with the Panzar framework, showing unit i in period t .

K is regarded as a capacity limit that we will describe

$Q_p^t = f(x^t, K^t)$ is the MWh electrical output produced from the power station in period t , which is determined only by physical production and configuration, and is entirely unaffected by whether it is requested by consumer. We imagine that there is a physical vehicle, such as pumped storage to which we can dump unwanted power. The suffix “p,” which Panzar does not use, denotes production.

x^t is the factor input in period t , which we shall interpret as running level in MW that is sent (from the station transformer²) to the substation to the grid.

The marginal efficiency of the j th unit at load Q_p in period t is $a(x_j) = \frac{\partial Q_p^t}{\partial x_j^t}$ if the j th unit substation is unconstrained by its capacity, that is, $x_j^t \leq K_j$ and x_j^t is the output of the j th unit in time interval t .

We treat constraint of the j th as the point at which $\frac{\partial Q_p^t}{\partial x_j^t} = 0$ if the j th unit is constrained and thence the effect of the constraint limit is $\frac{\partial Q_p^t}{\partial K_j} = 1$ if the j th substation is constrained and $\frac{\partial Q_p^t}{\partial K_j} = 0$ if it is not.

K^t is the vector of capacities in period t . We regard this as substation capacity.

We suppose that the first element of running corresponds to the first element of capacity. Physically we imagine that K corresponds to the thermal capacity of the transmission line sending electricity out of the power station, and that there is some entity on the substation that dumps excess power to avoid overloading the lines. We could, if we chose, regard the forced dumping of power as a zero marginal efficiency of the station, once the substation capacity is reached. We suppose no electrical losses from the station transformer to the substation busbars (the exit point to the grid) $K^t < K$. $K = (K_{n+1}, \dots, K_{n+m})$ where K is a scalar quantity equal to the sum of unit substation capacities. The vector is described in the stylized manner beginning at $n + 1$, where we interpret n as the number of units, because Panzar strictly decouples the variable costs from the capacity vector, which is regarded as a public good. It is a public good in the sense that the capacity mobilization in any period is limited to the sum of capacities and that the cost of capacity mobilization in one period is entirely unaffected by the mobilization in any other period. In our stylization, $m = n$,

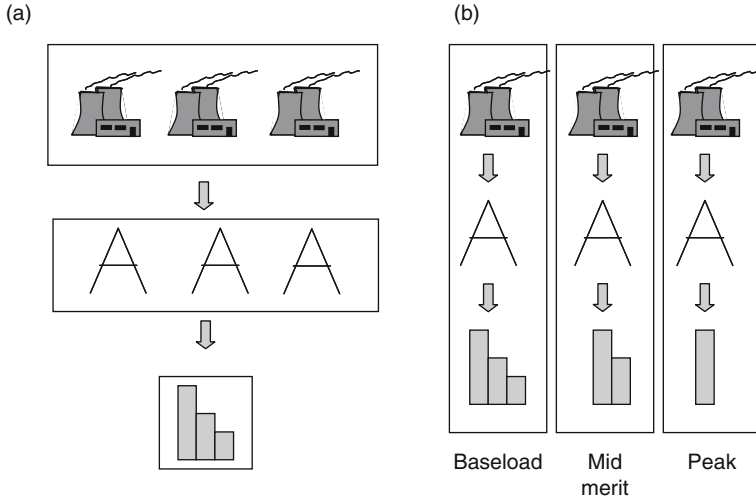


Figure 4.4 Panzar framework with and without link between units, capacities, and load (a) without (b) with (as used in this analysis).

although Panzar does not use a physical representation and does not note this, and in fact regards capacity as being a public good across space (i.e., all units) as well as time (i.e., all periods). Panzar regards the mobilization of both the private good (fuel into units) and the public good (capacity) as necessary for production, but does not bring these together, for example, by defining a variable cost function for a unit that becomes infinite at its capacity, either by asymptote or discontinuity. The two possible physical interpretations of the Panzar framework are shown in figure 4.4, with (a) being the most faithful to the equations, while (b) being more faithful to the standard paradigm and normal physical reality.

a is the constant ratio of inputs to outputs, which we can regard as efficiency. It is consistent with Panzar's analysis, to regard efficiency as a function of load $a(x^t)$

4.3.2 The Analysis

We now assume that because of the form of vectors x and K , that $f(x, K)$ is a continuously differentiable quasi-concave function; that is, we have continuously decreasing short run returns to scale. Panzar explicitly states that long run returns to scale are constrained to be decreasing, and we can rationalize this by assuming that the

energy and capacity vectors can expand at constant or declining cost if they are given time, but at an increasing cost if they are not given enough time.

Short-term decreasing returns to scale are represented by short run

$$\text{elasticity of scale being less than unity } e_s = \frac{\sum_{j=1}^n x_j f_j(x, K)}{f(x, K)} < 1.$$

We then add some boundary conditions to which we can apply physical descriptions. First, standard fuels work in standard power stations, so $\frac{\partial f}{\partial K_k} > 0$ for all $k = 1, \dots, n + m$, subject to $x > 0$. Second, either no fuel or no capacity results in no output, but small finite capacity and fuel result in small but finite output, so $f(0, K) = f(x, 0) = 0$. Hence the cost curves pass through the origin.

The key analysis in this framework is expounded in a multiperiod setting with deterministic periodic inelastic demand. Under a stylized neoclassical cost function, the key Panzar conclusions that are relevant to this book are i) even in the highest demand period, it is not economic for all units to run at full capacity and ii) full cost equilibrium is only optimal in the special case where cost elasticity = 1.

We will see below that the fixed costs of the fleet are related to the capacity, and hence K refers to unit capacity, not unit running level.

Before moving on, we must characterize our cost frontier. In particular, we need to decide whether to have units of variable size, variable technology, or both. There are too many degrees of freedom to solve the Lagrangean to have both, and hence we must have one or the other. The duality of the cost frontier is described in section 2.4.9.

In constructing the Lagrangean to minimize the cost of delivering the inelastic load in all periods, Panzar ignores the welfare cost of rationing in the event of insufficient capacity, since his Lagrangean condition ensures sufficient capacity. Under the neoclassical cost model with a soft capacity constraint then, capacity is infinite.

$$\text{The full cycle average}^3 \text{ cost is } \sum_{t=1}^T \sum_{j=1}^n w_j x_j^t + \sum_{k=1}^n \beta_k K_k,$$

where $w_j = w_j(x_j)$ is the average variable cost, which we interpret as fuel, and $\beta_k = \beta_k(K_k)$ is the average fixed cost, which we interpret as substation capacity. Note that Panzar makes no link between the fixed and variable costs and hence has no technology frontier. Our interpretation of total average cost is the cost per MW per unit period to satisfy a δ MW increment of load in each and every period, that is, a δ increment of baseload.

Note that we have not used the summation terminology of Panzar of capacity mobilization $n + 1, \dots, n + m$, but we use the index k instead of j in order to maintain the decoupling of running and capacity that Panzar envisages. It is not efficient to have more technologies than time periods (since this would cause running of units that are subordinate in cost terms) or less technologies than time periods (since we would prefer more discrete technologies on the technology frontier). Hence we can state that the number of technologies should equal the number of time subperiods.

The capacity constraint is $f^t - Q_d^t \geq 0$ where $f^t \equiv f(x^t, K)$ and Q_d^t is the inelastic and exogenously determined demand in period t . The suffix "d," which Panzar does not use, denotes demand.

So our Lagrangean is

$$L_t = -\sum_{t=1}^T \sum_{j=1}^n w_j x_j^t - \sum_{k=1}^n \beta_k K_k + \sum_{t=1}^T \mu^t (f^t - Q_d^t) \quad (4.1)$$

$P^t = \mu^t$ price = shadow cost

$w_j^t = \int_0^{Q_j^t} x_j^t dq_j^t$ for the j 'th unit in the t 'th period

$\beta_j = \int_0^{Q_j} f_k dq_k$ for the j 'th unit.

To find the optimum deployment of short- and long-term resource, we solve for the Karush Kuhn Tucker (KKT) conditions. The KKT conditions for multiple constraints (capacity sufficient in each and every period), tell us that either the Lagrangean multiplier is zero or the associated constraint is binding.

The first condition is

$$\mu_t \geq 0; f^t - Q_d^t > 0; (f^t - Q_d^t) \mu_t > 0$$

Thence,

$$\frac{\partial L}{\partial x_j^t} = 0 \text{ For all } t = 1, \dots, T, \text{ and } j = 1, \dots, n \text{ and}$$

$$\frac{\partial L}{\partial K_k} = 0 \text{ For all } k = 1, \dots, n.$$

Marginal variable cost is $\bar{w}_j = \frac{\partial w_j(x_j) x_j}{\partial x_j}$. Similarly for $\bar{\beta}_k$ for marginal

fixed costs.

For $j, t, k=1$ we have

$$0 = \frac{\partial L}{\partial x_1^1} = -\bar{w}_1 - \frac{\partial K_1(x_1^1)}{\partial x_1^1} + \mu^1 \frac{\partial f_1^1}{\partial x_1^1} = -\bar{w}_1 + \mu^1 f_{j=1}^1$$

$$\text{where } f_j^t \equiv \frac{\partial f(x^t, K)}{\partial x_j}$$

$$0 = \frac{\partial L}{\partial K_1} = -\bar{\beta}_1 - \frac{\partial x_1^1(K)}{\partial K} + \mu^1 \frac{\partial f_1^1}{\partial K_1} = -\bar{\beta}_1 + \mu^1 f_{k=1}^1$$

$$\text{where } f_k^t \equiv \frac{\partial f(x^t, K)}{\partial K_k}$$

Here we have added the term $\frac{\partial x_1^1(K)}{\partial K}$ to the Panzar term and then set it to zero to be consistent with Panzar.

The addition of the two terms stated, in the partial differential equations, is a matter of judgment, and depends on whether the relationship, for example, between x and K is first order (e.g., as in the Crew and Kleindorfer technology frontier) or second order. We can interpret this in two ways as follows.

First, we can, as Panzar states, regard K_k as a public good, which is a pure capacity limit, more or less unrelated to any unit. This corresponds to the Ramsey analysis in this most general form, but to make it more meaningful in the present context, we require the total cost recovery to relate to the total capacity. In this case there can be no relationship between the cost structure of any individual unit and therefore $\frac{\partial x_1^1(K)}{\partial K} = 0$.

Second, we can regard K_k as broadly relating to the “size” of the individual unit to which the capacity cost is directly associated, but with no hard constraint. Spend on capacity then reduces the marginal and average variable cost at high loads. In this case, at the optimum for the unit, for getting an increment of load δ MW out of the unit, we are indifferent between running it beyond its normal limit and spending more on its size to increase its nominal limit.

Clearly, if $\frac{\partial x_1^1(K)}{\partial K} \neq 0$, then $\frac{\partial K}{\partial x_1^1(K)} \neq 0$. Our interpretation of

$\frac{\partial K}{\partial x_1^1(K)} \neq 0$ is that we make the capacity decision and the running

decision at the same time. This point was discussed in section 3.6 on the Brown and Johnson framework and in 3.8 in the Carlton framework.

The subtlety of the neoclassical cost function is that it is curved, so while we would expect for a linear cost function for only one cost to be tangent to the constraint and thence bind, a curved function could more easily be tangent to many constraints.

Panzar invokes KKT further:

$$-\bar{w}_j + \mu^t f_j^t \leq 0, \quad x_j^t \geq 0, \quad x_j^t (-w_j + \mu^t f_j^t) = 0 \text{ for all } j = 1, \dots, n$$

and $t = 1, \dots, T$

Under KKT, either the delivered volume of unit j must exceed zero, or its cost structure must affect the optimum welfare, or both.

$$\frac{w_1}{f_{j=1}^1} = \mu^1 = \frac{\beta_k}{f_{k=1}^1}$$

$$\frac{f_{k=1}^1}{f_{j=1}^1} = \frac{\bar{\beta}_k}{\bar{w}_1}$$

We can interpret this variously. If we tie the investment to the unit, such that the increase in size decreases its variable costs, then this represents the conditions at optimum size and load.

Panzar again invokes KKT to arrive at

$$-\beta_k + \sum_{t=1}^T \mu^t f_k^t \leq 0; \quad K_k \geq 0; \quad K_k \left[-\beta_k + \sum_{t=1}^T \mu^t f_k^t \leq 0 \right] = 0 \text{ for all } k$$

Panzar then states that all Lagrangean multipliers μ are positive, and the constraints must be nonbinding, that is, there must be excess capacity in all periods.

Our understanding of the construction of KKT is that either the capacity of unit K is used, or its cost structure must affect the optimum welfare, or both.

Clearly this conclusion is dependent on the statement that the constraints bind in all periods.

We noted earlier that Panzar regards the production system as being comprised of n private goods (power generating units) and m public goods (which we picture as substations). He derives the KKT conditions by differentiating the Lagrangean with respect to both unit

cost and capacity. However, as we showed above, if we are to invoke technological efficiency in the form of the technology frontier, then we must regard each public good as being directly associated with its corresponding private good. So in equating the number of public and private units to the number of time periods, we have initially twice as many constraint terms as time periods in the KKT, but then the terms involving the private good become redundant as the constraints do not bite. In doing so, we lose Panzar's requirement for all unit constraints to be finite and thence lose the condition that all capacity constraints are slack.

Panzar's conclusion that all units must run in all periods is correct only in a particular theoretical construction of the physical configuration of the industry, and becomes invalid when that configuration is provisionally optimized by removal of entities not on the technology frontier.

4.3.3 Physical Interpretation

Let us consider this in physical terms, beginning with the peak period. Note that in our physical representation of the Panzar framework, we not only require the n substations to match the n units, but also require n periods with a different unit being the highest merit in each. In Panzar terminology, $n = m$.

If we build our plant stack according to the technology/size frontier, then we will choose and build N units and not build the $N + 1$ th. By definition we run the N th unit in the peak period and by definition we do not build the $N + 1$ th.

If it is not optimal for any unit to run harder in the off-peak than in the peak, then there should be no spare capacity in the peak. Let us suppose this is the case.

We will have spare substation capacity in all periods but the peak period.

Now let us consider the units. While the technology frontier is constructed with a single degree of freedom in plant characteristic, that of "size," there is no limit to the running of any unit. Provided that there is sufficient substation capacity for each unit, then depending on the level and convexity of the variable cost curves, all variations are possible, from only one unit being optimal in any timeframe, to all units running even in the lowest demand period.

It is not however correct, as Panzar does, to couple the unit and substation together and state that the capacity limit of the joint entity is at the same level as an asymptote of the variable cost curve. This is

incorrect in two ways, one of which is important and one less so. First, the less important aspect is that there is a clear coupling between the substation and the unit, and hence capacity is not public in the way expressed by Panzar. This aspect has been incorporated in our analysis by setting $m = n$. The second is more important. The substation capacity is a hard and not soft limit to unit output. By dumping load, we have represented it as a sharp reduction of marginal efficiency to zero at the capacity constraint. There is no connection to the unit itself and hence no reason to assume an asymptotic approach of marginal unit efficiency to zero (or, equivalently asymptotic approach of marginal variable costs to infinity as the capacity limit). Indeed there is no connection, other than a practical optimization of asset sizes, to connect the substation limit to the unit size. As with the Boiteux analysis, the unit has no size in the sense of a capacity limit. Size is simply a parameter denoting increasing efficiency with increasing spend.

Let us consider the practicalities of the Panzar framework. If returns to scale are negative (i.e., cost elasticity < 1) in both energy and capacity, then it is clear that we would ideally build lots of very small units. This restores the situation of constant returns to scale, and therefore to restore the Panzar framework, we must place a limitation on this. One way is to add indivisibility. We cannot do this because the calculations require the cost functions to be twice differentiable. The alternative is to provide a plant choice frontier that makes small units prohibitively expensive.

Let us now successively add shape to the load duration curve. Initially we start with baseload at a particular MW. We choose the plant with the lowest cost to serve the baseload. This is the plant with the lowest $b + \beta$, where b are variable costs and β are fixed costs. Note that both are load dependent but we drop the suffixes for convenience.

Now let us add more load in the peak. To simplify, we assume that the peak period is half the total period, and the load added is equal to the previous baseload level. Let us now add a unit with the minimum $\frac{1}{2}b + \beta$. This is represented in figure 4.5 in two ways. One, which is most consistent with practical application, has units with fixed and variable operating costs,⁴ and the other, which is most consistent with the Panzar framework (and similar to the Williamson framework), for which short-term costs are zero at, and also slightly above, zero load.

We can add more and more periods to the load duration curve using the same method. For example, if we add a super peak to this example, of duration $\frac{1}{4}$ of the whole period, then we use the unit with the lowest $\frac{1}{4}b + \beta$.

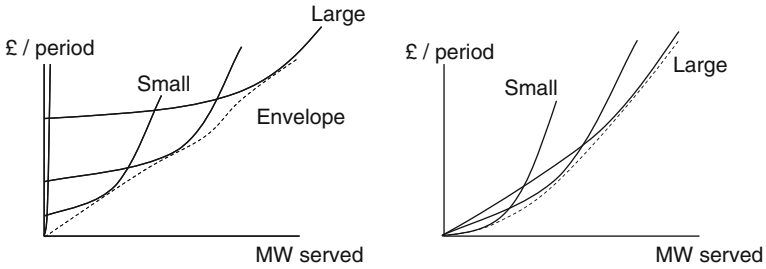


Figure 4.5 Unit cost envelopes.

Our definition of costs is as in Panzar. That is, large units have higher per MW capacity costs than small units, and unit capacity limits unit output.

We can check that our solution is optimal by

1. rebalancing the relative running volumes of the two units
2. replacing the high merit unit with a different unit
3. replacing the low merit unit with a different unit.

The first of these represents the application of the Turvey algorithm, described in section 3.1.3.1, and the latter two are both unit replacements, visualized as beginning the solution with a set volume of one unit and zero of a different unit, and seeing if a nonzero volume of the second unit is optimal.

The question on hand is whether in any of these possibilities, we have a nonbinding capacity constraint.

We can represent the optimum as shown in figure 4.6. For different installed capacity ratios, scheduling optimally for each, we can observe the aggregate welfare in relation to total installed capacity.

At the optimum, in each period, the marginal variable cost must be the same for all units running.

In addition, to meet the equilibrium load, the capacity cost saving from disinvesting an infinitesimal amount of capacity for any unit must equal the variable cost increase for that unit to meet the same load profile.

Brief consideration shows that even if marginal costs become asymptotically infinite, we do not necessarily run all available units. In figure 4.7 we can see that for a delivered demand in period t of $Q_1 + Q_2 + Q_3$ with the first three highest merit units, we would not run the fourth even if it were installed. The criterion for running the fourth

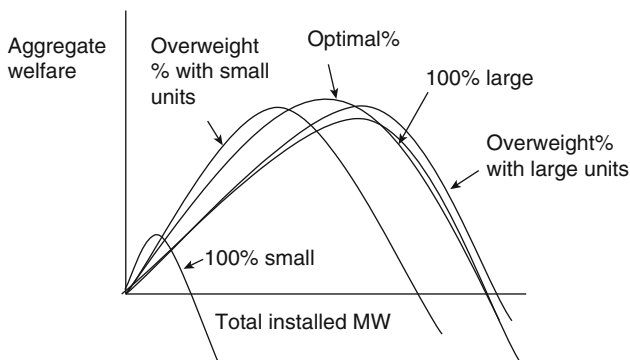


Figure 4.6 Optimum aggregate welfare for different percentage mixes and total size of two units.

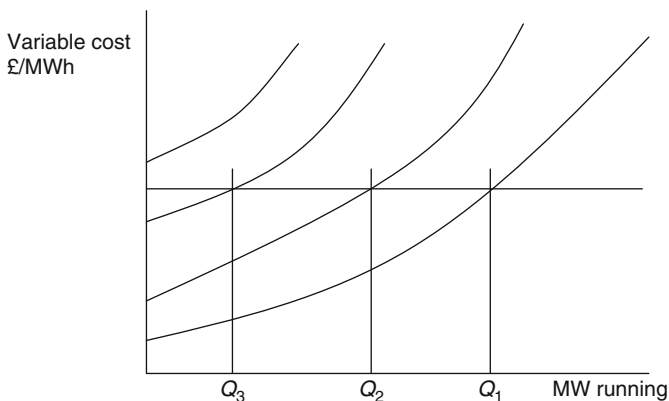


Figure 4.7 Allocation of unit output to deliver $Q_1 + Q_2 + Q_3$ MW in period t .

installed unit is represented by the point at which the horizontal line intercepts the fourth unit variable cost function at the ordinate. Note that since the market is orderly (centrally managed), once total delivered volume in the period is affected by unit fixed costs, the allocation of this volume is determined only by variable costs.

In deciding whether to build unit 4, we look at the peak period and the loads delivered by the other three units in total equilibrium, and then test to see whether the total cost of unit 4 in the period (i.e., variable cost, plus fixed costs load into that period), is less than the equilibrium clearing price in the period. Supposing that figure 4.7 represents the peak, then the clearing price is found by adding the

fixed costs of the first three units, loading into the periods in which they run.

For deterministic demand, fixed costs should only be loaded into the period for which units run. Once built, units do not necessarily run in all periods. These conclusions differ to Panzar's.

4.3.4 The Relevance of Returns to Scale

Panzar also arrives at the very interesting conclusion that there is cost equilibrium only in the special case when cost elasticity is equal to 1 (i.e., constant returns to scale). Let us consider if this can be applied in our physical situation.

The key equation here is

$$TR - TC = \sum_{t=1}^T \mu^t f^t (1 - e_l^t)$$

where TC is total costs, TR is total revenue, μ is the Lagrangean multiplier, and e_l^t is the variable cost elasticity of unit l in period t where the cost elasticity in period t is defined as

$$e_l^t = \frac{\sum_{j=1}^n x_j^t f_j^t + \sum_{k=1}^m K_k f_k^t}{f_t}. \text{ Note that we are assigning a capacity cost}$$

to a period and therefore must have a cost allocation rule

$$TC = \sum_{t=1}^T \sum_{j=1}^n w_j x_j^t + \sum_{k=1}^m \beta_k K_k, \text{ where } w_j \text{ are average variable costs (4.2)}$$

From the KKT conditions we have

$$x_j^t w_j = x_j^t \mu^t f_j^t \text{ for all } j, t, \text{ and}$$

$$K_k \beta_k = K_k \mu^t f_k^t$$

So, substituting into equation (4.2), we have

$$TC = \sum_{t=1}^T \mu_t \left\{ \sum_{j=1}^n x_j^t f_j^t + \sum_{k=1}^m K_k f_k^t \right\}$$

So,

$$TC = \sum_{t=1}^T \mu_t f_t e_l^t$$

Our Lagrangean to maximize total surplus subject to the constraint of satisfying demand is

$$L = \sum_{t=1}^T \int_0^{Q^t} P^t(Q) dQ - \sum_{t=1}^T \sum_{j=1}^n w_j x_j^t - \sum_{k=1}^n \beta_k K_k - \sum_{t=1}^T \mu^t (f^t - Q^t)$$

For this to represent aggregate surplus, $P^t(Q)$ must be the willingness to pay and hence the first term is the gross Marshallian consumer surplus. The last term (the Lagrangean constraint) is the condition of no rationing once prices are set.

Panzar then sets

$$TR = \sum_{t=1}^T P_t Q_t = \sum_{t=1}^T \mu_t f^t$$

Here P_t must be the price received by the producer.

So,

$$TR - TC = \sum_1^T \mu_t f^t (1 - e_t^f)$$

Hence for $e_t^f = 1$, $TR = TC$

What is more interesting is the cases where $e_t^f \neq 1$. While this is in accordance with conventional analysis for short run equilibrium, ignoring fixed costs, Panzar presents this in the context where fixed costs are included. Our argument has been that $TR = TC$ for all elasticities.

From a variable cost basis, it is easy to see how units with an upward sloping (neoclassical/Ricardian) variable cost function make money in the short term, even if the price equals the marginal variable cost. This is because the price is above the average variable cost. The total revenue will only equal the total cost and the surplus in each period is exactly matched by the fixed cost. It is quite possible for net of revenue and cost, that is, the profit to be negative, zero, or positive. For a given combination of technology frontier, load duration curve, rationing characteristics, and stochastic features, if we constrain no unit to have negative profits, then one will have zero profits and the rest will have positive profits. Evolution of the technology frontier and the so-called Tobin Q effect described in section 2.4.7 drives all unit profits to zero.

4.4 CONCLUDING DISCUSSION OF SOFT CONSTRAINTS

In practice, the relaxation of hard constraint to more accurately model capacity and peak load pricing is problematic. To make the analysis tractable we either end up with an infeasible set of costs, as with Hirshleifer in which we have dominant or subordinate units, or we create false constraints, as with Panzar, in which the connection between cost and capacity is lost.

For this reason the analytics here and in the literature use hard constraints as an essential simplification.

There are nevertheless some important results that come from soft constraint modeling, for example, the theory that most or even all units can set price simultaneously.

MODELING CAPACITY USING DERIVATIVES

The development of peak load pricing theory, with something of a conclusion in the Chao framework of the early 1980s, had for its context a world of central planning for power generation and scheduling.

The early 1980s ushered in the world of privatization and private access to wholesale markets. Development in the finance markets was broadly similar to the explosion of the growth of derivative trading. At this time, management science involving the value of choice and planning for different outcomes in a probabilistic manner, and the development of mathematics and formalism of derivatives converged in the form of “real options,” which, in their “no arbitrage” form, synthesized the holding of a physical asset through the holding of a derivative contract.

While capacity mechanisms of sorts did exist earlier in privatized models (e.g., the England and Wales pool) and capacity pricing did exist (e.g., in France as noted in the Drèze framework), it was in the 2000s that the new breed of capacity mechanisms grew. The form that they have developed to, called reliability options, are very similar to traded derivatives. With increasing concerns about the capacity adequacy aspect of security of supply, and power station capacity in particular, it is now essential to view electricity markets in terms of derivatives. In addition to this, the complexities of power plant, particularly the interaction of market prices, plant engineering, and reliability costs, environmental constraints, plus the “dimensions of service” of power plant in terms of products more complicated than “off” and “full load” are highly amenable to a real option approach. Finally the literature on derivatives is vast, and provides a ready-made tested library of functions for our use.

To use this modeling suite, we have to start by making some assumptions, namely:

1. single service—plant is running at full load or off
2. zero state-change costs—in particular the engineering cost of a power station start
3. perfect reliability and zero engineering costs
4. unconstrained cost-free transmission of electricity
5. constant price for fuel and a market of infinite depth (elasticity)
6. constant price for environmental allowances per MWh of production.

All these can be unraveled, and must be in practice.

The logical flow for approaching this subject is

1. power markets
2. spot and forward contracts
3. European options
4. American and swing options
5. real options.

We will then examine the next level of complication, such as plant reliability, nonfirm contracts, price dynamics, the value and probability of lost load, and demand-side management as virtual production.

5.1 POWER MARKETS

There are some aspects of market structures that have an important effect on the design and efficacy of capacity mechanisms. The three main models are i) fully administered, ii) pool, and iii) bilateral.

The optimal market structure can evolve from any of administered, pool, and bilateral markets.

5.1.1 Administered Regimes

Administered regimes tend to have some common features, being

1. treating of demand as inflexible and with a halfhourly variable forecast
2. plant build made on a strategic basis, commonly state-owned
3. plant scheduling done using a merit order of variable cost

4. ex ante and ex post adjustments to the administrative schedule, such as modeling some plant as “must run,” and adjusting the initial schedule to take account of factors such as transmission constraint and adequate system reserve.

5.1.2 Pool Markets

Pools come in a variety of models and share certain key features that are relevant for this book.

1. Demand forecast by the system operator
2. scheduling based on the legacy models of administered regimes, treating unit offers as the proxy for variable costs
3. a set of rules for bidding (offering) plant, such as allowing or not prices that are profiling across time and load bands, withdrawal and resubmission of bids, penalty on failure to deliver
4. a resulting family of indices, for example, a day-ahead profile of 48 halfhourly prices, constructed from the intersection of the offer “stack” and the demand forecast
5. a contracting suite of ancillary services offered by the monopsony system operator
6. schedule instructions from the system operator
7. governance arrangements regarding market abuse
8. mandation of offers, such as frequency response, or offering of available capacity.
9. price limits
10. a variety of locational structures.

5.1.2.1 *Locational Structure and Transmission Charging in the Pool*

At one extreme we have postage stamp pricing in which there is a single price across the whole control area. Annual transmission charges may have both a regional structure, varying cost split between production and demand, and a temporal structure (e.g., the triad system in which the whole year’s charges are based on three halfhourly peaks).

At the other extreme, in the location marginal pricing (LMP) model, each node has a different price at any time according to the status of transmission constraint. A node is one or a family of Grid Supply Points.

There are many intermediate models. For example, in markets that are coupled and split, the price is the same across all zones for which

transmission constraint does not bite between the zones, and the zones have different prices when interzonal flow is constrained.

5.1.2.2 Value of Lost Load in the Pool

All markets apply a value of lost load (VOLL that is nominally a proxy for the price at which consumers would be prepared to accept loss of load. In practice this is a regulatory construct with only a tenuous link to the economic VOLL.

5.1.2.3 Short-Term Capacity Payments in the Pool

The system marginal price (SMP) is found by the intersection of the generation offer stack and the demand forecast. To this is added a capacity amount equal to $(VOLL - SMP) * LOLP$. All plants get this whether or not it runs. The units with accepted offer therefore get in total the pool purchase price $PPP = SMP + \text{capacity}$. The suppliers pay an additional uplift related to system costs.

In fact, LOLP was artificially elevated, and VOLL, though approximately based on empirical estimates, was artificially depressed. The significance of lowering VoLL while increasing LOLP to maintain a constant $VOLL * LOLP$ is explored in section 3.11 within the Chao framework.

5.1.3 Bilateral Markets

Bilateral markets have three key features:

1. the market in physical notifications (PNs)
2. the imbalance “cashout”
3. the balancing mechanism.

There are in all markets numerous adjustments. In BETTA in Great Britain, one of most interest is the balancing service use of system (BSUoS) charge in which the system operator charges an ex post amount to participants, which varies halfhourly, and for the moment at least has no regional variation.

5.1.3.1 Physical Notification and Forward Contracting in the Bilateral Markets

In the BETTA bilateral market in Great Britain, one party sells a PN to the buyer. The trade is notified to the system operator.

The PN is commonly described as being physical. This is a reasonable description in normal circumstances, but this representation

breaks down in conditions of tight margins, load shedding, options, and contracting for capacity. Since these conditions are of great interest to us in this book, we must consider the nature of PNs in more detail.

The PN is not in any sense a right to physical power, and the purchase of a PN has no effect on physical delivery to the consumer. In this sense it has public goods characteristics.

We can conveniently regard a PN as three things:

1. an intent to produce/consume at the date/time specified, notified to the system operator
2. a financial arrangement with the market operator, offsetting the volume that is cashed out as imbalance
3. the potential for compensation for lost load, although this is not currently a core market feature.

After “gate closure,” currently an hour ahead of real time, no more PNs can be submitted and all subsequent transactions are with the system/market operator.

5.1.3.2 *Balancing in Bilateral Markets*

Balancing contracts are options provided to the system operator by generators or suppliers who can adjust load or have it adjusted by the system operator. There is no premium. The balancing mechanism can be viewed as the monopoly/monopsony market immediately after gate closure.

Balancing mechanism prices differ, but the one of most interest here is the “pay as bid,” in which the participants pay/receive the bid amount if they are called to change volume. We will see in section 5.2.3 that balancing offers can be viewed as options with no premium.

The system operator will accept balancing bids even when the system is at national balance, for example, to resolve transmission constraints, maintain adequate reserve, or manage more complex issues such as reactive power and frequency. This does not affect our core analysis.

It is worth noting that there is a public goods elements to balance.

1. If the system is in balance, then the cost to an individual supplier of their imbalance is much reduced.
2. If a supplier is out of balance, for correlation reasons, it is likely that the system is out of balance.

3. Even if a supplier is in balance, they pay for system imbalance, the mechanism in Great Britain being the BSUoS charge, which is halfhourly ex post.
4. The bulk of demand is not metered on a halfhourly basis, the rest of demand being settled centrally using central estimates of consumption profiles. The imbalance is charged in relation not to the actual demand but the profile estimated demand.

5.1.3.3 *Imbalance in Bilateral Markets*

If the actor consumes or produces more or less than the PN, then the difference is cashed out at the imbalance price. Commonly there is a different price for extra spill (energy into the system) and draw (energy from the system), and they depend on whether the actor's individual imbalance made the system imbalance better or worse.

For analytic convenience we will generally assume that the imbalance price is equal to the volume weighted average of accepted balancing offers. We will also ignore all balancing acceptances for system reasons, such as reserve and transmission constraint, and finally assume that there is no simultaneous positive and negative imbalance of individual power station units and suppliers. While unrealistic in normal circumstances, it is less so for the situations of tight capacity that we consider. It is also relatively easy to relax the assumption. For example, we might net all imbalances that help the system (i.e., excess production or less consumption when the system is short overall) at a balancing offer price.

5.1.4 Forward Contracting

5.1.4.1 *Definitions*

Spot contracts—A spot contract is the sale by one counterparty to another of commodity in a standard form. A spot contract will typically have “weight-rate-date” specification, that is, how much, at what price and at what time. Electricity contracts have a maximum resolution that is equal to the balancing period of the market. For example, for a market with halfhourly balancing, the contract will specify a MWh total for the period. There is in general no specification for the MW profile within the halfhourly period, although in practice the producer may be limited by the grid code.

Forward contracts—Forward contracts are exactly the same as spot contracts and apply to fixed dates/times in the future. Since electricity has three periodic cycles (daily, weekly, and seasonal), these can be “sliced” in different ways.¹ Contracts involving simultaneous purchase

of one contract and sale of another (which may be a different commodity or a different delivery time) are termed spread contracts.

Over-the-counter (OTC) contracts—OTC contracts are between two counterparts and can be of any legal nature that they wish. Commonly OTC contracts become fungible on exchanges by an “exchange for physical” process; in the current case, an OTC PN must become a formal PN to have any value.

Futures contracts—Futures contracts are very similar to forward contracts and are traded on exchanges. To access the greatest liquidity, their delivery specifications can be broader than forward contracts. The differential between futures and forward prices, particularly if the “cheapest to deliver²” mechanism embeds a high degree of basis differentials³ on the forward contracts.

Forward and futures contracts can be described as weight-rate-date.

1. Weight commitment volume—A total MWh per commitment period and a MW.
2. Rate—Strike price—If there is a strike price, then the seller sells energy at this price. Otherwise the seller can sell at the market price, which may be floating or have a cap.
3. Date—This has several dimensions: the commitment period being the first and last dates/times of the whole commitment (e.g., a calendar year), the notice period before delivery of energy, a weighting of payment, or deficiency in relation to time of day or week or season.

A key definition is firmness. If the commitment is firm, then some form of preagreed penalty or liquidated damage is paid for failure to deliver. Ideally for firm markets, this is the prevailing offer price in the energy market. We will see that a PN in the bilateral market is firm to the system operator and nonfirm to the buyer.

For a nonfirm contract, there is no penalty for failure. An example is the submission of an offer to the England and Wales pool. There was no penalty for nondelivery by a scheduled unit.

In a pool market, the forward contracts for difference (CFD), for example, are OTC and nominally⁴ entirely unconnected to system scheduling. They simply result in the buyer/seller paying to the seller/buyer an amount equal to contract volume times the difference between the CFD price and the outturn of the index.

Finally, a physical activity that is not optimized in the market is called nonruthless. The term ruthless in some circumstances mean

that market optimization is out of keeping with the spirit of the contract. Burning one's house down for the insurance, if not in violation of contract or illegal, would be an example of a ruthless activity.

5.1.5 Price Limits

All or almost all markets have price limits that are either enshrined in regulation or de facto, in that offers/bids higher/lower than the maximum are not accepted by the portals, for example, due to a limit in the number of digits enterable.

In pool markets, the OTC CFD has a practical cap created by the maximum pool price index. In the bilateral market the practical cap is associated with the imbalance price cap.

The price limit of interest to us is the maximum price in the balancing mechanism (creating an arbitrage limit in the multilateral market) or pool market, and the equivalent limits in the reserve markets.

Price limits are commonly set well below the peak price needed for equilibrium and this causes the problem known as “missing money” for peaking units.

Normally the balancing price cap exceeds the imbalance price cap, which exceeds the market price cap.

The way that a price cap affects the probability distribution is not straightforward. By simple preservation of unit total probability, we arrive at figure 5.1, which decreases the price expectation. However,

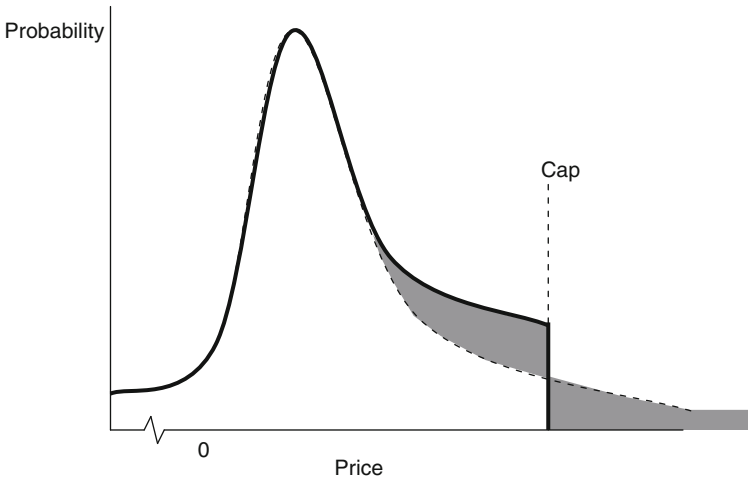


Figure 5.1 Effective of probability distribution of price from a price cap.

the price cap acts as a focal point that has the effect of flattening the distribution at high prices and restoring the expectation price to its precap level.

5.1.6 Reserve and Ancillary Markets

These are contracted on a monopsony basis with the system operator.

The key ancillary contract types are i) black start, in which a unit can start and deliver load to a grid that has stopped, ii) reactive power to stabilize voltage and phase of the alternating current, and iii) reserve. Our interest here is in reserve.

Reserve is essentially the same product as bilateral option contracting except for time, in particular, i) temporal resolution within the main pricing period and ii) shorter option notice that can be managed in the multilateral market. Reserve is ultimately a monopsony market with the system operator, although there can be a secondary market.

Commonly a reserve contract with the system operator would have the same weight-rate-date delivery components as option contracts. The weight is the MW or MWh according to contract, the rate is the price in £/MWh for volume and £/MW/hr for capacity, and the date is the delivery period.

In practice not only does the system operator have the ability to foreclose the forward market with the reserve market but often has a commercial incentive to do so. There is a balance to be struck between having the confidence that markets can self-balance power in short timeframes and recognizing that managing the very short term, resolving transmission constraints, and technical aspects such as frequency, phase, and voltage are the natural province of an expert and accountable monopoly/monopsony.

5.1.7 Arbitrage between Markets

The three markets of forwards (PNs, CFDs), balancing and imbalance are partly fungible with each other. For example, a supplier can choose to buy a PN or get cashed out in imbalance. This causes convergence of the forward price, the expectation of the average long and short imbalance price, and the average balancing acceptance price.

This is relevant from the perspective of price expectations and price caps in each market.

The arbitrage exists at the level of options as well as forward contracts, and here the substantial potential for foreclosure of the option markets through the monopsony reserve market is highly relevant.

There is also arbitrage between interconnected markets, particularly where there exist price caps and cross-border contracts, such as importable/exportable capacity.

5.1.8 Modeling Similarities between Pool and Bilateral Markets

In this book we use pool markets for some models and bilateral markets for others. In fact, the two markets are very similar in modeling terms and differ in practice mainly through different emphases. For example, in bilateral markets, the market operator exerts considerable emphasis on the construction of balancing and imbalance prices and in addition can charge or credit a halfhourly and locational amount evenly or unevenly to all actors, thereby having the same effect as transforming the pool price.

Consider some situations:

1. For a small player, selling a PN at price F has a very similar outcome to selling a CFD at F and offering at zero price in the pool.
2. For any player, buying no PN and thereby paying the imbalance price on all power drawn is similar to paying the pool selling price.
3. For any player selling a PN at F and having a strategy to buy back if the forward price falls below variable cost b is similar to selling a CFD at F and offering into the pool at b .
4. Selling no PN and having a strategy to offer in balancing at K is similar to offering into the pool at K .

5.2 SINGLE PERIOD—MODELING USING OPTIONS OF “EUROPEAN” TYPE

The holder of a call/put option has the right, but not the obligation to buy/sell from/to the option grantor at a predetermined price formula, usually a fixed strike price. The buyer pays a premium for this right.

5.2.1 Definitions for Contract Terms for Options of European Type

Call (vertical) spread—the purchase of a European call, accompanied by the sale of a European call identical in all respects except strike price, which is higher.

Calendar (horizontal) call spread—the purchase of a European call, accompanied by the sale of a European call identical in all respects except expiry date, which is nearer.

Caplets—are European call options—struck on short periods (e.g., halfhours).

Caps—these are a continuous series of caplets, struck at the same strike, independently exercisable.

Declaration—the decision declared by the option owner on the option grantor. If not declared the option is abandoned on the expiry date.

Delivery date—the delivery date of the forward contract that is called.

Exercise for financial delivery—when the forward contract declared is immediately sold at the prevailing price, resulting in a cash settlement in favor of the option buyer.

Exercise for physical delivery—when declaration of the option causes the buyer to have a forward contract.

Expiry date—the date and time on which the option expires.

Ladder—a series of call options struck at different levels.

Strike price K —the preagreed fixed or variable price at which the commodity is bought if the option is exercised.

Swaption—A family of caplets that may only be exercised together.

5.2.2 Modeling Definitions

Average volatility—the average volatility over a period, being either the average historic, or the average implied volatility, assuming constant volatility over the averaging period.

Cost of risk—the amount that someone will pay to avoid the risk for the period in question.

Current/instantaneous volatility—the current volatility of a forward contract, whether historic or implied.

Delta— $\Delta = \partial C / \partial P$. The increase in option value for a unit increase in underlying forward price, thereby representing the option amount

of forward contract to sell to minimize risk to an option holder with no other positions.

Drift—the change in value of a market forward price over time is the superposition of random change and deterministic drift. The drift is a direct function of cost of risk.

Extrinsic value—time value

Gamma— $\Gamma = \partial\Delta/\partial P = \partial^2 C/\partial P^2$ the increase in option delta for a unit increase in underlying forward price, thereby causing option holders to sell into a rising market and buy into a falling one.

Historic volatility—the volatility of a forward contract calculated actuarially from the price history.

Implied volatility—the forward contract volatility implied from the actual current traded price of options.

Intrinsic value—the value of the option with volatility is set to zero.

In the money—a call option with strike price below the forward price, which is therefore likely to be exercised.

Kappa—the exposure of the value of the option to changes in volatility. $\kappa = \partial C/\partial\sigma$. Also called vega.

Out of the money—a call option with strike price above the forward price, which is therefore unlikely to be exercised.

Price returns— dP/P . The change in price dP over time interval dt , divided by price P .

Risk—a parameter denoting ex ante variation of a quantity over time. Usually standard deviation or variance.

Stochastic—varying over time, usually according to defined coefficients of a defined probability distributional form.

Tenor—the time from now to the period in question. Also called horizon.

Term structure of volatility—at an instant in time, the structure of the (current or average) volatilities of forward prices of different tenors.

Time value—the value of an option that relates to volatility. Intrinsic value plus time value = total value.

Volatility— σ , the annualized deviation of the price returns or equivalently of the logarithm of price. For low volatility, this can be viewed as a percentage of price per year.

5.2.3 Regarding Balancing Mechanisms as No Premium Options

The rationale for making a commitment with no immediate recompense is that balancing is a market foreclosed by the monopoly

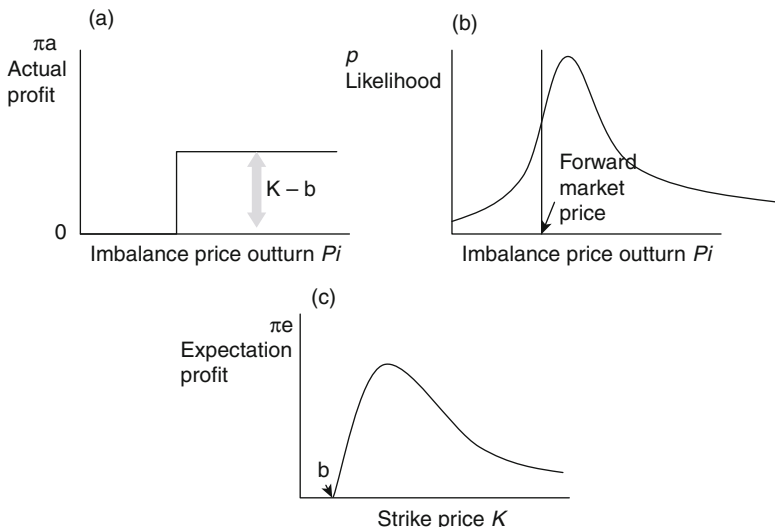


Figure 5.2 Influencers of the choice of strike offered in balancing (a) Payoff following offer in balancing at K (b) Probability distribution of imbalance price (c) Optimization of balancing price offer K .

operator. In the absence of any other route to market, a unit that offers the option with a strike K above the variable cost b , such that $K > b$, has a finite probability of making a profit. Since no offer will with certainty deliver zero profit, offering balancing is a dominant strategy relative to not offering, even in the absence of a premium.

The value of selling the balancing option at $K < b$ has only negative value and hence is a subordinate strategy. There is then the question of which strike to sell. The profit is zero if $K = b$, but the exercise probability tends to zero as $K \gg b$. There is then an optimum. While it is quite possible to do regression analysis on forward price at gate closure, imbalance price and balancing acceptance probability according to submitted price, in practice there are so many system idiosyncrasies that this is a matter of judgment more than statistics. The optimization is represented in figure 5.2.

5.2.4 Modeling Probability Extremes Using Options

First we do this assuming zero cost of risk, constant volatility, and zero interest rate. The value of the option must be $C = \int_K^\infty P(S)(S - K)dS$.

A rearrangement gives the market price conditional of exercise as $E\{S|S > K\} = K + \frac{C}{P(S > K)}$.

We can then “bootstrap” a probability distribution, using the log-normal as a basis, starting the calibration with high strike price options and moving down.

In practice the off-the-shelf techniques available from the derivatives world are much more amenable to working upward from the “at the money” option with $S = K$.

The simplest method is to evaluate the premium for any strike by using the Black⁵ formula for option pricing $C = S * N(d_1) - K * N(d_2)$. We can see by inspection that $N(d_2)$ is the probability of exercise.

$$d_1 = \frac{1}{\sigma\sqrt{t}} \ln \left[\left(\frac{S}{K} \right) + \frac{1}{2} \sigma^2 t \right]$$

where t is the time to exercise and deliver and σ is the lognormal standard deviation, called volatility, and for low volatilities can be expressed as a percentage of price.

$$d_2 = d_1 - \sigma\sqrt{t}$$

The standard method of expressing the vector of call option premiums in relation to strike price is through the implied volatility, taking the probability distribution to be normal. The shape of this vector

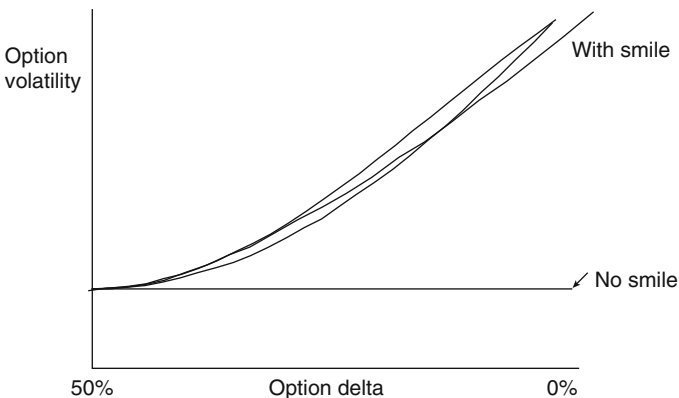


Figure 5.3 Simple variations to the option smile.

is commonly called “smile” and can be modeled in different ways. About the simplest way to model smile with one degree of freedom while observing boundary conditions such as monotonic decrease in value with increasing strike is by a sine function as depicted in figure 5.3.

More generally we can perfectly model any strike/volatility vector with a polynomial, although this is in practice highly unstable. We could instead model the actual distribution with a polynomial. This is mathematically more pure, although in practice it is even more unstable than modeling the smile with a polynomial.

With the strike/premium vector, we can build a probability distribution.

When modeling cost of risk, it can be useful to have the conditional variance.

The formula is⁶

$$V = F^2 \left[\exp(\sigma^2 t) N(d_3) - (N(d_1))^2 \right] - N(-d_2) K [2FN(d_1) - KN(d_2)]$$

$$\text{or } V = F \left[\exp(\sigma^2 t) N(d_3) - N(d_1) \right] - C(K + C),$$

$$\text{where } d_3 = (\ln(F / K) + \frac{3}{2} \sigma^2 t) / \sigma \sqrt{t}$$

5.2.5 Cost of Risk Bias

Cost of risk in power is complex. In the long term, there are competing forces. On one hand, the stock market is generally averse to positions that are negatively correlated to the oil price. The power price is broadly correlated with oil in the long term, mainly via the gas price. On the other hand, generators are long of power and seek to reduce this. Overall the net effect is a positive cost of risk, that is, the market is long and the forward price is lower than the expectation of the spot price, that is, it has a downward bias. In the short term, both generators and suppliers avoid being caught short and hence the forward price is an upward biased estimator of spot price. The bias is reduced by price caps.

Figure 5.4 shows this term structure. It also shows the term structure of the price conditional on option exercise. The residual risks of high and low merit units are different. High merit plants have more forward hedges in place, especially if failure risk is low. Low merit plants have little or no forward hedges in place.

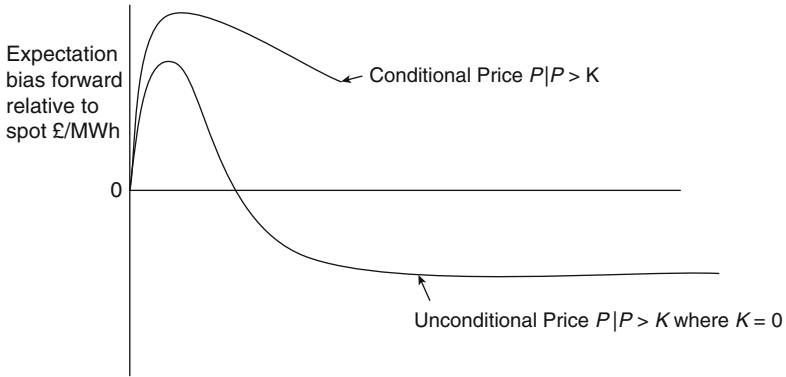


Figure 5.4 Expectation bias of forward prices from cost of risk.

5.2.6 Selling Options at Variable Cost

We will examine real options in section 5.9. For the moment we assume that our perfect asset has a fixed cost B £/MW/hr and a variable cost b £/MWh.

If the asset sells a call option at strike K for premium C , where $K = b$ and $C = B$, then with certainty the unit earns a revenue of $C - B = 0$ and with likelihood λ , the unit earns an additional revenue of $K - b = 0$, where λ is the likelihood of exercise.

The situation in which cash flows are certain is called “no arbitrage” and competitive conditions drive the net present value to zero.

The situation where $C > B$ for $K = b$ or $C = B$ with $K > b$ generates an excess return equivalent to the alpha⁷ α of capital market theory.

For simplicity we focus on the situation with zero α .

The cost of risk can influence what strike to sell at and whether or not to sell an option. Generally speaking, a reliable high merit generator will be fully hedged. A reliable low merit generator can capture the value of cost of risk premium to price by selling something (generally an option) forward.

5.2.7 Selling Options above or below Variable Cost

Suppose that a generator has a variable cost b and sells at strike K . What issues does it cause if $K \neq b$?

If $K < b$, then the generator will lose whenever called, but will, if the options are fairly priced, recover this ex ante in an option premium

that is higher for $K < b$ than for $b = K$. There is a maximum loss per period of $((K - b) - C) * Q$, where Q is the capacity and C is the total premium for the period. A simple utility analysis of the cost of risk shows us that the generator should charge a fairly high-risk premium for this profile, with a maximum at $K = 0$.

If $K > b$, then the generator will gain whenever called, but will have reduced premium. This is a step toward having no option at all, since this can be represented by $K = \infty$. In practice $K > b$ allows for a degree of uncertainty of variable costs (the uncertainty arising both from general uncertainty and also from the division of fixed and variable costs, which is in turn partly theoretical. In terms of risk profile, the worst case is no exercise and hence the minimum net revenue is C . This is a benign risk profile and hence will not be accompanied by a high-risk premium charged by the generator.

Figure 5.5 shows these different situations.

If we set aside issues such as gaming, general uncertainties, cost of risk bias of prices, and plant failure, the optimum strategy for a risk-averse producer is to sell an option struck at variable costs. We will examine this in the no arbitrage approach.

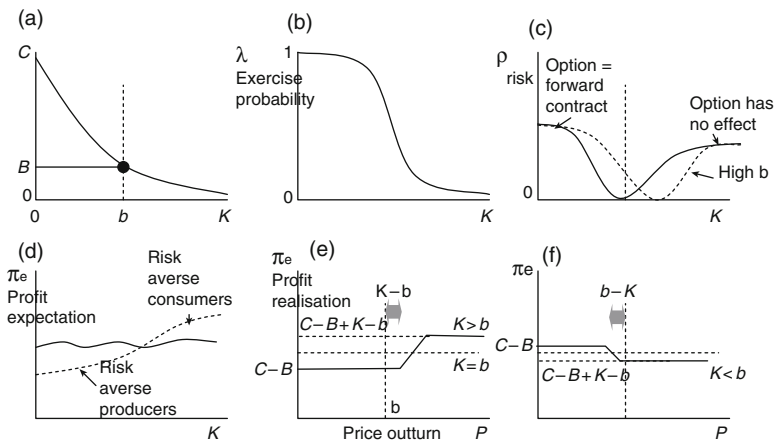


Figure 5.5 Situation faced by producer according to different strike prices and market price outcome (a) Call option versus premium, showing the position of one generator (b) Call exercise probability in relation to strike (c) Risk in relation to strike (d) Profit expectation in relation to strike (e) and (f) Profit realization versus outcome price for strikes above and below variable costs.

5.2.8 Options Struck at the Market Price Cap $K = P_{\text{mcap}}$

It is not straightforward to apply a price cap to OTC trades not on a formal market, but in most markets there is a portal for trades, on which price caps can be applied.

5.2.8.1.i $K = P_{\text{mcap}}$: $P_{\text{balcap}} = P_{\text{imbalcap}} = P_{\text{mcap}}$

Let us firstly consider the situation in which the PN market, balancing and imbalance caps are equal.

An option purchase struck at the cap has no guarantee of physical delivery to the buyer and no prospect of profit. Therefore the value of the option to the buyer is zero.

An option sale struck at the cap results in a PN, if exercised. Failure to deliver against the PN is punished only at the imbalance price. Therefore the maximum cost to the seller is zero.

In the absence of any other consideration, there is then no incentive to pay more than zero and there is no cost of sale. The contract is essentially irrelevant.

5.2.8.1.ii $K = P_{\text{mcap}}$: $P_{\text{balcap}} = P_{\text{imbalcap}} > P_{\text{mcap}}$

Now the balancing and imbalance cap exceed the market cap. This now has potential value to the buyer, since if they arrive at gate closure with a short position and no PN offers in the market, they face the imbalance cost. Knowing that they can achieve the balancing price, the sellers may withhold volume from the PN market and instead offer into balancing to capture the missing money lost by the market price cap. The likelihood of this strategy is enhanced by the fact that the balancing cap acts as a focal point.

Note from figure 5.4 in section 5.2.5 that the forward actual or implied price exceeds the spot/imbalance price expectation and hence the unit will prefer to sell an option than wait.

The option price will move up and down as the expectation of balancing and imbalance prices change.

Now, if we consider how the option value rises and falls with the prevailing forward price, we can use basic option theory to construct an implied forward price. The logic is as follows:

1. We can model the probability distribution of the forward price for any future observation date, such as contract maturity.
2. We can model the relationship between today's forward price and today's vector of option prices with different strike prices.

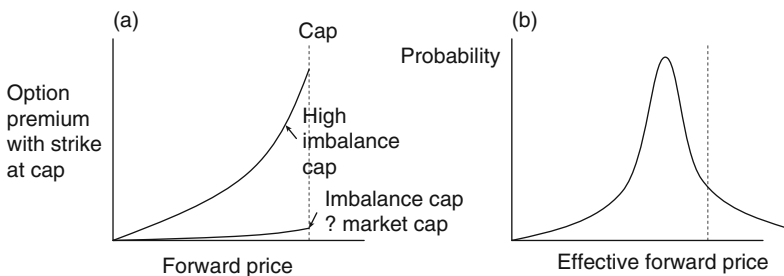


Figure 5.6 Demonstrating an effective forward price above the cap.

3. For any option, we can, using the strike and the premium, impute an effective forward price at option exercise, conditional on option exercise
4. For the option vectors at the current and different today’s forward prices, we can model the probability distribution of the effective forward price.

We now have an effective forward price distribution as shown in figure 5.6. Note in particular that options have circumvented the price cap. In essence, the option market has restored the economic signals denied by the cap.

5.2.8.1.iii $K = P_{\text{mcap}}; P_{\text{balcap}} > P_{\text{imbalcap}} > P_{\text{mcap}}$

If, as we expect, the balancing cap exceeds the imbalance cap, then the maximum revenue is attained by offering in balancing. Broadly speaking, if the market arrives overall short at gate closure, then few players will be long.

In the absence of the option, the supplier has a worst-case cost of P_{imbalcap} . Through the purchase of the option, this is reduced to $PPN_{\text{cap}} + C$, with C being the call premium.

The buyer (generally retail supplier) has several choices:

1. intentionally overcontract to be expected to be long at gate closure—this is consistent with risk averse behavior
2. knowingly undercontract—this may attract censure by the market operator or regulator but can happen if the participant baulks at paying a rising price and “chases the market” eventually ending up short
3. undercontract due to lack of liquidity in the PN market as the market demand forecast rises—this would be quite normal

4. buy an option well in advance—this is the optimum behaviour,⁸ although there is no empirical evidence since no market is this developed
5. buy an option if and when PN liquidity dries up at the cap price—this will depend on the relativities of the imbalance and balancing price caps and the concomitant probability distributions of the two prices.

The issue now is that the maximum revenue for the generator is found by withholding capacity from the option market and offering in the balancing mechanism instead. The maximum revenue for the generator also exceeds the maximum cost for the supplier. The actual strategy will depend on the value and probabilities.

Broadly speaking, we would expect generators to prefer to sell the option than wait for balancing for two reasons: i) lower cost of risk and ii) lower likelihood of censure, expropriation of revenues, or other regulatory intervention that would compromise the benefit of being called at the balancing cap. Clearly the relative expectations of the two strategies will play a role. In addition to this, the price bias effect of cost of risk depicted in figure 5.4 means that a reliable generator selling forward or options can capture an expectation of profit from the bias.

$$5.2.8.1.iv \quad K = P_{\text{mcap}}: P_{\text{imbalcap}} > P_{\text{balcap}} > P_{\text{mcap}}$$

Now the imbalancing cap exceeds the balancing cap, which exceeds the market cap, but the strike is only at the market cap. This is not a normal situation, as the monies received from imbalance would exceed the monies paid for balancing, and these would need recycling to participants.

However, it is an interesting situation when we consider capacity mechanisms, and worth pursuing here.

Now the maximum imbalance cost of the supplier (or generator) exceeds the maximum balancing revenue of the generator (or supplier).

With respect to worst/best case, the impetus of the supplier to buy an option struck at the market price cap is greater than the impetus of the generator to sell. This then drives up the premium. We would expect a market to develop.

5.2.9 Options Struck at the Imbalance Price Cap $K = P_{\text{imbalcap}}$

First, we must consider how this can come about since options deliver to PNs and the PNs have a price cap. One possibility is that the PN is tagged in some way so that it may exceed the market price cap.

We can see that there is never either physical surety or financial value to the supplier.

If the balancing cap exceeds the imbalance cap, the generator may prefer to withhold volume

We would therefore expect no trades.

5.2.10 Options Struck at the Balancing Price Cap $K = P_{\text{balcap}}$

$$5.2.10.1.i \quad K = P_{\text{balcap}}: P_{\text{balcap}} = P_{\text{imbalcap}} > P_{\text{mcap}}$$

There is no physical surety or financial value to the supplier.

There is no cost to the generator of failing to deliver or benefit in excess of offering in balancing.

The contract is therefore irrelevant.

$$5.2.10.1.ii \quad K = P_{\text{balcap}}: P_{\text{balcap}} > P_{\text{imbalcap}} > P_{\text{mcap}}$$

There is no value to the supplier.

The maximum profit in selling the option is less than the maximum profit in balancing.

We therefore expect no trades.

$$5.2.10.1.iii \quad K = P_{\text{balcap}}: P_{\text{imbalcap}} > P_{\text{balcap}} > P_{\text{mcap}}$$

With this ranking of caps, the situation is the same as for the strike equal to the market cap, that is, a market should develop.

5.2.11 Generator Reliability Considerations

With the imbalance cap set very high, the greatest fear is imbalance. The generator will fear short-term failure and the supplier will fear shortfall in demand forecast. As a result of this, both may be conservative. The supplier may enter gate closure with a long position, the generator may withhold capacity when the risk is highest, that is, when the system is tightest. In tight-system high-price conditions, the generator will then spill power and the retail supplier will draw less than the PN. Both get the imbalance cashout for long players. Seeing the high implied forecast from the PN, the low generation from the PN, the system operator may call on reserve. This displaces other power and hence the spill price received by the supplier and generator are low.

Given the inefficiency of using costly reserve power that displaces market PNs, it may be efficient for all actors to limit their imbalance cost exposure by collectively commissioning reserve capacity.

A very important feature in this scenario is the probability profile of failure. Plant status (in a fail or not failed state) has relatively high

“persistence,” that is, if it is not in a failed state now, it will probably not be in a failed state tomorrow. This means that while a unit may not wish to sell high strike options a long term ahead, for fear of imbalance cost on failure, she may be happy to do so a few hours or even days ahead as the plant status is well characterized.

Let us then consider the near horizon when the plant is in an unfailed state. There is a high driver for the supplier to buy as this limits the imbalance cost. There is a high driver for the generator to sell as the premium is captured, and in addition the implied price expectation captured conditional on exercise may exceed the expectation without the option. Note that the more units get called on through option declaration the less the imbalance and thence the less prospect of high returns in balancing.

5.2.12 Raising the Imbalance Cap in Times of System Tightness

We can envisage three situations:

1. normal
2. scarcity
3. actual loss of load.

Scarcity is actually a commonly used term. For now, we will consider only loss of load.

Let us cap the imbalance charge at P_{imbalcap} in normal times and set a charge of $P_{\text{imbalvoll}}$ for times of actual loss of load on the system.

We can immediately see from the analysis above that options at any strike below $P_{\text{imbalvoll}}$ has a finite value.

This could have adjustments, for example, the imbalance rate on loss of load could be different for suppliers and generators, or for generators who failed unexpectedly, and so on.

5.2.13 Ex Post Settlement of the Imbalance Charge on Lost Load

A simple mechanism is to charge for each halfhourly period, a VOLL (which could be both cyclic and stochastic) times the lost load in GWh and charge ex post pro rata to all market participants according to their imbalance.

There are some particular problems with this approach.

1. If load is actually interrupted involuntarily then we have no way to know *ex ante* or *ex post* what the real VOLL is.
2. Particularly, given point (i) above, the actors risking this charge have no real way to estimate VOLL *ex ante*.
3. The charge can be very large, and being charged *ex post* is accompanied by substantial risk of nonpayment.
4. The actors have no real view of the loss of load probability (LOLP).
5. Knowing neither the volume nor the price of imbalance, each individual supplier might act conservatively to such a degree that the aggregate contracted capacity is excessively conservative as diversification of risk is not taken into account.
6. Credit risk, that a supplier will exit under extreme prices, and thence default on their obligations.

Nevertheless we can see that the mechanism can work crudely. In practice, it is the credit issue that is the largest.

We can see how this mechanism can stimulate demand-side management. If regulatory VOLL is set very high, then individual consumers may offer demand-side management with a strike lower than VOLL. While they get less than VOLL in the event of system failure, there are times when they get K when the system is whole. In fact, the more consumers recognize this, the more they compete for load loss. The initial offer would be at just below $P_{\text{imbalvoll}}$ or otherwise at the highest price allowed.

5.2.14 Ex Ante Settlement of Imbalance

Suppose that the issues for *ex post* imbalance settlement on lost load are considered too great. We can reduce the issues, while creating others, by an *ex ante* method.

We could do this by requiring each supplier to have a demand forecast and then purchase in advance some form of capacity (forward, options with potential value, options with no potential value) to an amount equal to the demand forecast plus a capacity margin Q percent.

There is then an element of public good since the purchase of capacity by any supplier reduces the LOLP for the system and thence all other suppliers. Suppliers may then choose to make a rule that everyone, including them, must buy capacity.

5.2.15 Deficiency

To have ex ante settlement of failure against any obligation we need a censure mechanism. One mechanism is an infinite fine, but this raises uncertainty costs. It may then make sense to have a regulatory backstop, commonly called a “buyout,” in which the obligation can be discharged by a penalty payment, in this case a deficiency charge. This would be at some level above the lower of the cost to mobilize

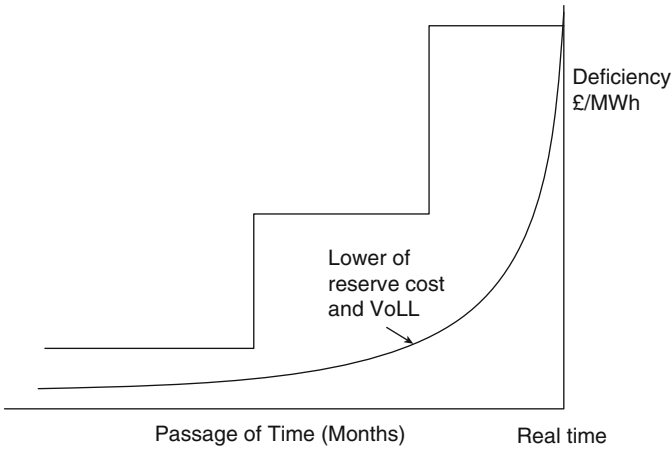


Figure 5.7 Deficiency charge in relation to its economic value.

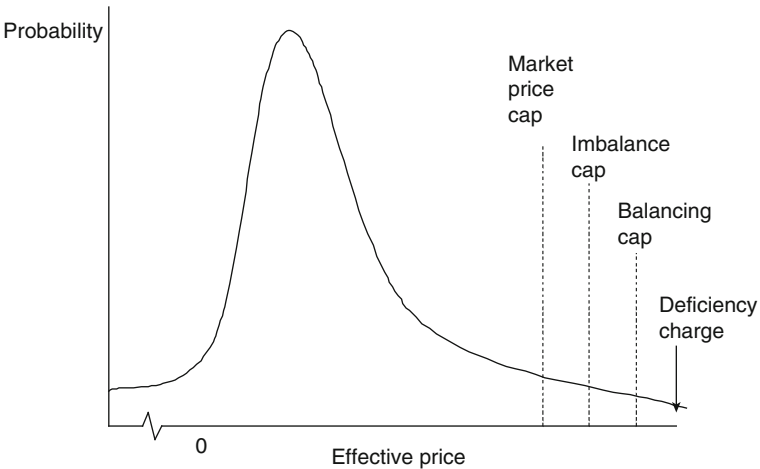


Figure 5.8 Probability distribution of the effective price.

reserve or the VOLL, otherwise it has no purpose and does not act to resolve the issues with ex post imbalance charges in times of scarcity. An example of such a function is shown in figure 5.7. The stepwise nature of the function acts to ease the administration and concentrate the liquidity in the secondary market.

Putting all the price caps and deficiency charge together, we can see that the market price cap is largely circumvented and we have a reasonably continuous distribution of the effective price, as we see in figure 5.8

5.3 MODELING SYSTEM OPERATOR OPTION PROCUREMENT IN THE ONE-PERIOD SETTING

Here we consider a one-period setting and suppose that the system administrator/operator views the system entirely as a collection of firm European options. Demand is assumed stochastic and inelastic and that there is a single system price. We assume either demand-side management at VOLL or an infinite generation capacity at VOLL, so we can regard demand as always satisfied.

The initial distribution of price can be found in a number of ways. For example, we assume that all options are backed with units that have well-defined views of their fixed and variable costs, and then offer calls with strike K at variable cost b and premium C and fixed costs B .

We can then apply some premium-strike vector that moves to give the system operator least cost. So all units below the line get paid premium. Lost load is treated as a cost at VOLL. Figure 5.9 shows that the descending clock method can start with excess plant and then

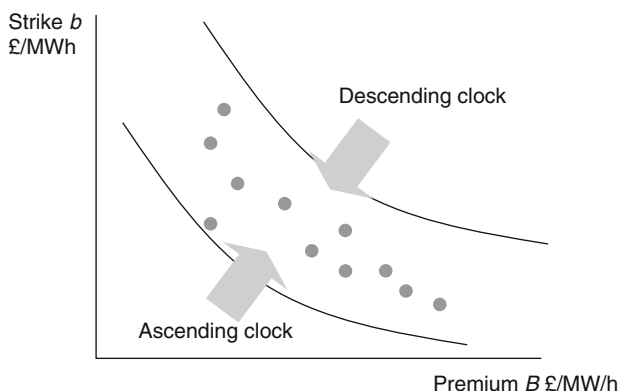


Figure 5.9 Ascending and descending clock auction to select options.

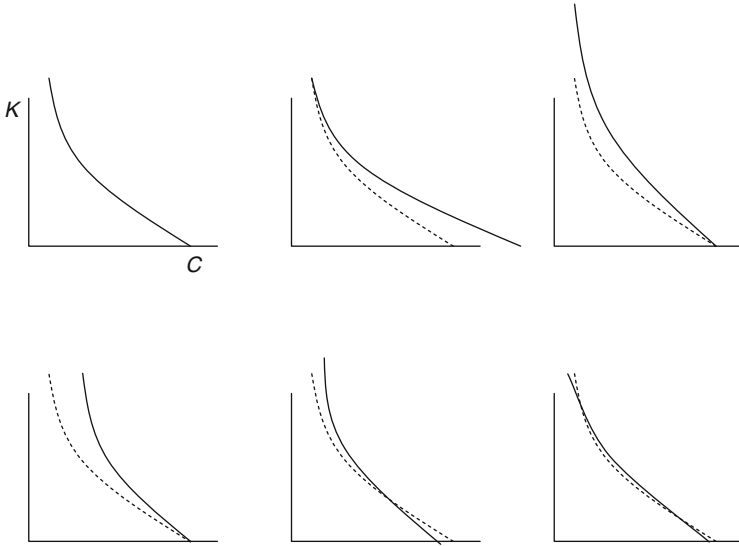


Figure 5.10 Non-homothetic changes to the option premium-strike vector.

reduce the vector until the requirement is just satisfied. The ascending clock works in reverse.

Figure 5.10 shows some non-homothetic shifts to the strike-premium vector used for the descending clock auction.

We could instead start with best-fit vector through the points, and by assuming that all of the strikes are set at variable costs move the vector rightward until least cost is achieved.

Another approach is to construct the stochastic equivalent of the load duration function and apply the Turvey algorithm starting with the highest merit unit.

Finally, there can be a tâtonnement in which the strike-premium vector moves continuously and units can see if they qualify or not, before the process closes. There would then be a moving strike-premium vector and units can see if they are above or below the line as shown in figure 5.11.

All methods essentially end up in the same place, which is the single optimum selection of units.

There is now a significant complication, as summarized in section 2.4.7, that fixed costs are dependent on plant value. A plant not selected loses value and therefore fixed costs. Similarly a plant recognizing that it could offer higher and still be selected can do so, increase its value, and therefore its fixed costs. There is then a convergence of

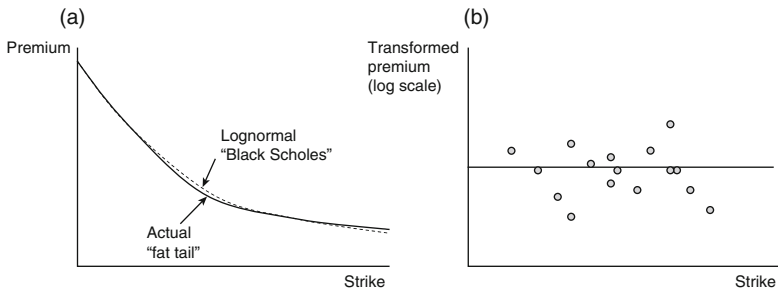


Figure 5.11 (a) Continuous fair value curve for option premiums and strikes (b) Mapping of transformed premiums to strike prices, showing actual offers relative to the transformed fair value line.

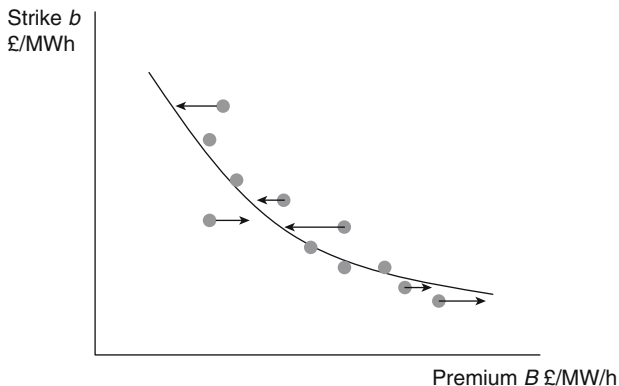


Figure 5.12 Convergence of units onto the single strike-premium vector.

units toward a single vector, which we may regard as an expression of the law of one price. This is further enhanced by the temporal allocation of fixed costs as described in section 2.4.6, and physical choices available for different stack evolutions. The convergence is shown in figure 5.12.

5.4 MODELING MORE COMPLEX OPTIONS IN THE ONE-PERIOD SETTING

For more complex modeling we require commodity spread options, contingent claim options, average rate options, and various derivatives that can be constructed from components of the options mentioned.

Fortunately not only is there a vast literature on traded derivatives, but there is a variety of off-the-shelf solutions that can be applied directly.

Whilst fuel price variations are in principle a significant analytic hurdle to surmount, in practice when we consider capacity obligations, we can ignore all variables except plant availability (which is technology specification) and variation in demand.

5.4.1 Option with Strike Indexed to Fuel Cost

We include this section for completeness as the analysis of peak load and capacity obligations has to be done first assuming static fuel and environmental costs, and then the practical reality of volatility in these factors must be addressed.

The standard approach to complex derivatives, which we can regard power stations as, has several steps:

1. Split the contract out into discrete components. The orthogonal “dimensions of service” approach described in section 5.9.2 does this, while recognizing the conditionality of exercise of some options on exercise or nonexercise of others
2. Make simplifying approximations or boundary conditions where possible. An example is the simplification of swing options as Bermudan flexicaps
3. Hedge out the main risks into separate trading “books” as shown in figure 5.13. The three key risks are volatility (κ and/or γ), forward hedges (δ), and correlation. In practice, correlation instruments are very limited and any correlation hedges are highly approximate.

Correlation is in fact very important when considering capacity mechanisms and high strike options, noting in particular that correlations in extreme events are very different indeed to those in normal events. So some correlations go from ~ 0 to ~ 100 percent and some from $c100$ to $c0$ percent.

In very general terms, these rules of thumb work

1. Electricity does not matter to oil. So oil prices drive electricity prices and not vice versa.
2. “Clean spark spreads” (power minus gas minus CO_2) and “clean dark spreads” (power minus coal minus CO_2), at the relevant standard power station efficiencies are primary state variables

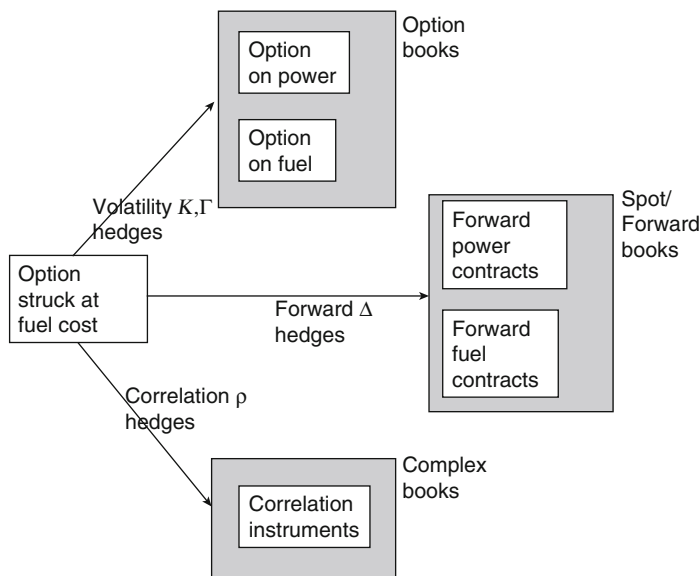


Figure 5.13 Trading book structure for options struck at the fuel price.

commonly with lower volatilities than the underlying commodities, with “power” here being either baseload, peak, or off-peak according to what plant is at the margin.

3. Gas is highly correlated to oil. Coal price is driven by global demand and has a flat term structure of volatility. CO_2 prices must be treated in swing terms and are highly dependent on how much time remains in the compliance period.
4. The volatility of the transaction currency should be taken into account, although this can generally be ignored when considering capacity and high strike price options.

5.5 MANY PERIODS—MODELING USING OPTIONS OF “AMERICAN” TYPE

Peak load and capacity pricing relies heavily on modeling shocks to the price and load duration functions. In doing so, vital information on the timing of events is discarded.

Electricity prices are exposed to seasonal factors, and these are subject to shocks of phase.

The result of these two facts means that we have to model the derivatives in their “American” form. This greatly complicates the analysis but it is necessary.

We should note the modeling commonality between European and American options for electricity, through the consideration of time-probability load factor duality described in section 2.4.1. On one hand, we can model a single period and consider the ex ante (and ex post verified) probability distribution of load or prices. On the other hand we can consider first, a many-period deterministic load duration function and second, a many-period stochastic load duration function, as described in section 2.4.1. The single-period stochastic version is pure European in modeling. The many-period deterministic version is pure American although not stochastic. What we are interested in is the ex ante expectation of the load (or price) duration curve and the stochastic shocks to it. For this we need to do swing modeling as we see in section 5.7.

5.5.1 Definitions of American Options

American options—The contractual definitions of American options are very similar to European options. The critical difference is that while the forward contract to which the European option refers to is constant, the forward contract delivery date to which the American option refers to is tied to the declaration date. So, for example, a December delivery European option contract expiring in August will always deliver in December if it is exercised. If the corresponding American option is exercised in March, then it will deliver in March, and if exercised in April will deliver in April, and so on. For deliverables with zero or near-zero cost of physical storage, such as gold or money, this difference is not critical. For electricity, the nonstorability of the commodity makes the American option a very complex one.

5.5.2 Simplification of American Options as “Bermudan”

Bermudan⁹ options—these are American options for which the opportunity to exercise is at discrete occasions, rather than at any time/date prior to expiry. The difference in value is generally relatively small, and modeling American options as Bermudan even if this does not match the contract, makes modeling much simpler.

5.5.3 Modeling Bermudan Options as European

Consider an option of Bermudan type, in which there are two delivery date possibilities A and B, where B is after A. We can treat this as the sum of

1. a European option with delivery date A, expiring on date A
2. a European option on the price differential between contracts A and B, with expiry date A
3. a European option on contract B, expiring on date B, with a strike price of the value of B, conditional on no exercise of the first two options.

This method is called semianalytic. In practice this technique is sufficiently reliable because the more complex models that model the single instrument are limited in the degree to which they can model the principal components of the forward price vector and any incremental gain from modeling perfection is lost in increased opacity and problems in calibrating the coefficients.

5.6 MODELING SYSTEM OPERATOR OPTION PROCUREMENT IN THE UNRESTRICTED MANY-PERIOD SETTING

The system operator can in theory work out the required option portfolio at system level, for the stochastic multiperiod setting. We can view this as the option version of the Crew and Kleindorfer model.

Initially assuming deterministic inelastic demand, the market operator experiences a cost of

$$\text{Cost} = \sum_{n,m} Q_i C_{ij} + \sum_n Q_i \rho_{ij} K_{ij},$$

where C_{ij} is the premium that the i th of the n units accepted offers in the j th of the m subperiods, K_{ij} and ρ_{ij} are the strike price and load factor of the i th unit in the j th period. Q_i is the capacity of the i th unit. For convenience, we have assumed that there is no part loading.

In practice we have two problems with modeling as a series of caps:

1. cross-elasticity of demand between subperiods is complex
2. power stations commonly have some restrictions of the swing type.

We therefore need to use swing option models.

5.7 MODELING MANY PERIODS UNDER CONSTRAINT USING OPTIONS OF “SWING” TYPE

American options refer to options with a single exercise.

A swing option allows the buyer to buy any amount of commodity on any date/time, subject to a maximum volume overall and a maximum “take” on any one date/time. There is one strike price.

Modeling of swing options is essential in both gas and power markets and is of particular relevance to the value of capacity and capacity obligations.

Unfortunately there are considerable modeling difficulties.¹⁰ Even the flexicap, which is the well characterized analogue in the money markets is highly complicated when the market moves in a more complex way than the simple one factor.

Swing options are also essential to model in upstream gas contracts, and industrial and commercial gas and power supply contracts. Much of the analytics in the literature refer to the former and the analysis for capacity obligations and the analysis for upstream gas can be regarded as two branches of the same tree. Here we focus on the version most appropriate to capacity obligations.

5.7.1 Definitions

Swing—a general form of contract at a fixed delivery price in which there are minimum and maximum volumes over the contract period.

Flexicap—a cap in which the total number m of caplets exercisable is less than the total number n of caplets in the cap. Can be viewed as a swing contract in Bermudan form and with only a maximum take.

Take or pay—Below a minimum volume “take” the commodity is paid for whether or not delivered. There is a limit to the maximum take at the take or pay price. Commonly there would be daily and monthly minimum and maximum takes.

5.7.2 Modeling of Swing Options

The complexity of swing options is such that we make model simplifications to suit specific applications. For the consideration of capacity, the approach that we take is stochastic shock to the load duration curve.

There are different ways to do this:

1. spot price trajectories
2. spot price probability trees
3. forward price vector analysis
4. load/price duration shocks.

In spot price trajectory modeling, jump diffusion models are used to simulate spot price trajectories. These models commonly accommodate “regime switching” (e.g., normal and scarcity conditions). The backward induction method is readily applicable for American options but not so for swing. Spot price models are generally less amenable for calibration to market instruments than other methods.

Spot price probability trees (“forests” for swing options) are the standard method for swing modeling. However the restrictions on both the volatility structure in time (term structure) and price (modeling skew to the lognormal distribution) are too excessive for capacity modeling.

Forward price vector shocks are the standard methods from advanced derivative markets. However they are commonly much more flexible in relation to volatility term structure than volatility strike structure (smile, etc.) and in practice not useful for capacity modeling.

Here we use load factor shocks as described in section 7.1.1.

We can now see the expectation load factor, and all other statistical coefficients of any unit in the stack based on the assumption that it will run in merit. We can incorporate demand-side management and estimate the probability of lost load.

We can now work out the equilibrium price duration curve in the following way:

1. For the lost load period, set the price to equal VoLL.
2. For the peak unit we know the probability profile of loading. Noting that the peak price is VoLL we raise the second peak price P_1 to the level at which unit is at ex ante financial equilibrium, taking cost of risk into account if required.
3. We now go the next unit and keep going until all units cover their costs.
4. We now have a stochastic price duration function. In the same way that we did with European options in a single period setting in figure 5.12, in section 5.3 we can now adjust the fixed costs of the units and iteratively optimize.

5.8 MODELING SYSTEM OPERATOR OPTION PROCUREMENT IN THE RESTRICTED MANY-PERIOD SETTING

We will see in the section on real options (section 5.9.1) that while treating units as caps is a good starting point, it is better to model

them as flexicaps with a limit on the number of exercises over the year. If the variation in the percentage of caplet exercise (equivalent to the load factor) is low, then the restriction in exercises makes little difference to value.

5.9 MODELING REAL ASSETS

We have noted that there are three key risk factors for plant, them being fuel, environmental, technical (mainly failure), and the main nonfuel costs are capital and engineering. These can all be modeled effectively and in a single framework.

5.9.1 Real Options

Over the last 30 years, the science of optimizing both the provision and the exercise of choice in business decisions has advanced considerably. The factors driving our choices are extremely varied, such as consumer demand/trends, politics, natural/weather events, and market changes. If the choices can be synthesized by tradable instruments then the physical/decision options are regarded as the real options of their synthetic counterparts. Insurance and reinsurance cover a wider variety of risks, which can be synthesized with varying difficulty.

Our interest here is in a family of real options called no arbitrage. For example, we saw in section 5.2.6 in the simplified example that a plant that sells a call option struck at variable cost is financially indifferent to the outturn of the market price. If the total portfolio of asset, market instrument, decision framework are assembled such that the financial outcome is independent of market outcome, then we have a no arbitrage situation.

5.9.2 Complex Real Options and Dimensions of Service

For the majority of theoretical modeling of peak load pricing, it is sufficient to consider two plant states “full load” and “off.” For more general practical and theoretical modeling of power plant it is important to recognize further states such as minimum stable generation, the ability to switch fuels, the ability to load cycle on a planned and unplanned basis, and the provision of short-term reserve (more or less load) either by storing energy in the plant or by incurring more engineering damage. Additionally, there are further dimensions of service such as “black start” (starting without grid power), reactive power (a form of power essential for grid stability), maximum generation

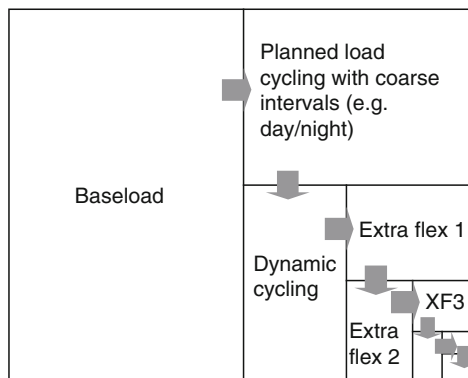


Figure 5.14 Visualization of unit valuation. Area is proportional to value of service dimension.

Source: Harris (2014).

(“maxgen”) above normal capacity, and movement of planned outage dates (essential for capacity planning).

Figure 5.14 shows the hierarchy of services that do not conflict (in derivative terms, they are “orthogonal”) if offered in the correct order and the values are additive. Not shown here is the fuel switching family (e.g., coal to gas, coal to biofuel co-fire or dedicated biofuel, gas to distillate). Ignoring environmental restrictions for convenience, the sequence is:

1. sell baseload power, and buy the fuel
2. buy back power in the off peaks, and sell back the fuel
3. offer to increase load in the off peaks and decrease in the peaks, at the behest of the option buyer
4. offer the remaining flexibility options “live” in the market rather than prearranged by option sales. For example, if the plant is partially loaded, it can offer upward reserve. There is a series of further options according to plant status, for example, very short-term reserve in the form of frequency response.

We note that conceptually this is a form of the principal component method and that each subarea is orthogonal to all others. Figure 5.15 shows a schematic view of the values of the different dimensions for two plant types.

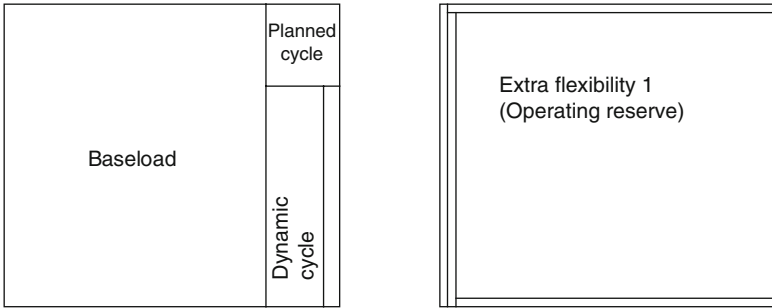


Figure 5.15 Visualization of the value mix of two different unit types (a) New build CCGT (b) OCGT.

Source: Harris (2014).

5.9.3 Plant Reliability and Nonfirm Contracts

Our core modeling assumes perfect plant reliability. However since plant availability is such an integral plant of security of supply modeling and the construction of capacity obligations, we cannot ignore it.

Here we briefly summarize an approach.

Where a plant is selling forward contracts, as distinct to option contracts, the internal market model can operate. This is shown in figure 5.16. “RCo” acts as an internal insurer and buys power according to fail likelihood. The unit sells firm power, and in the event of failure RCo sells power to the unit at its marginal cost, so that it can honor its contracts with no financial loss.

The flows of fuel and environmental allowances are not shown, nor are the engineering costs. We can see that this arrangement has

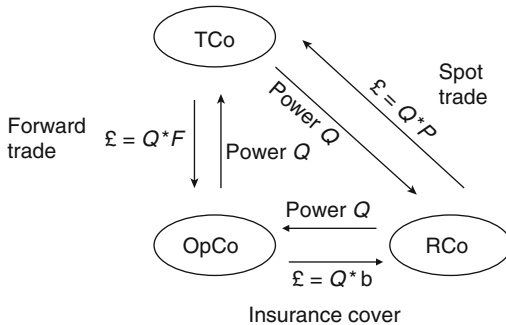


Figure 5.16 Money and power flows on plant failure.

the net effect that the contracts from OpCo from TCo are firm even when the plant fails, and on failure the insurer RCo loses the difference between market price P and variable cost b on the volume. The plant is financially unaffected by the failure.

The model can work in principle for option contracts. The core mechanism is essentially the same, in that RCo acts as an insurer. There are significant complications.

1. Instead of the prevailing forward price, RCo must model the forward price conditional on exercise.
2. Due to the heterogeneity of units called on, the market price is affected significantly by the failure of the unit, and this must be factored in.
3. The correlation between failure events is complex as they can be dependent on logistics, technologies, natural events, and human events.
4. In failure events, not only are the units called on heterogeneous, but they are few, and technology frontier characterization is not straightforward.

These complications are broadly manageable in modeling terms, although for actual detailed planning of system security, the idiosyncrasies must be taken into account.

If, as we expect there would be, there is a correlation between plant failure and other plant failure on the system, this introduces convexity to the loss on exercise of insurance cover.

RCo can in principle provide insurance for options. Here the concavity of the risk is severe because not only is the market price conditional on exercise high, but correlation between failures of different units increases in scarce conditions. Though the RCo method can be used to evaluate the commercial aspects of failure (whether to contract, how much to spend on reliability, etc.), it is not a useful product for internal market transactions. In practice, power stations rely on system reserve to cater for short notice failure.

5.10 MODELING PRICES AND PRICE DYNAMICS

5.10.1 Modeling the Stack as European Call Options—Caps

Since power stations can be expressed as options, the power system can be modeled “virtually.” What is particularly useful here is that we

can ignore failure, and real plant failure can be readily handles with the RCo modeling above.

In fact the situation in which we have a series of options with strike prices and premiums is practically identical to that of the modeling environment of Crew and Kleindorfer.

We now apply the no arbitrage¹¹ condition and have a series of options, each with premium equal to the fixed cost and a strike price equal to the variable cost. Each option premium must be at its fair value,¹² equal to the probability weighted conditional payoff.

We now ignore the physical characteristics of the plant and consider the development of price.

With stochastic inelastic demand, or demand with an inelastic and elastic component, price is now determined only by option holders. There are many possibilities.

One possibility is that each option holder does nothing until dispatch day and offers power at variable cost b . The market will clear at the intersection of the demand and the option offer stack, and provided that the lowest merit unit offers at $b + B/\lambda$, that is, with a fixed cost uplift, then all units will run in merit, and option holders will have an expectation of profit equal to the option premia.

5.10.2 Hedging

The finite cost of risk of option holders causes them to hedge their positions in the forward market. Let us consider the main effects to see if they make a difference to outturn prices.

First, the option delta ideally requires a forward hedge that matches the tenor of the forward contract that underlie the option.

Second, cost of risk considerations gives more impetus to forward “delta” hedge a low strike “in the money” option than a high strike “out of the money” option. This is shown in figure 5.17(a).

Third, the further out in time horizon the option declaration is, the closer the delta is to 50 percent.¹³ This is shown in figure 5.17(b).

The overall result is this:

1. Hedge selling generally increases continually with time, as seen in figure 5.17(a).
2. If the forward price is low then the passage of time may cause hedge selling, as is evident from figure 5.17(b).
3. As prices rise, not only do all traders hedge sell (the “gamma”), but the increased impetus shown in figure 5.17(a) increases this effect as forward price P rises relative to strike price K .

The only people to sell to are the suppliers, so we must consider their position. We can generally simplify the supplier delta with two parameters:

1. the latency—broadly corresponding to the time taken for consumers to switch supplier once they have decided to do so
2. the persistence—broadly corresponding to the intersupplier switching rate for consumers and the number of suppliers in the market. Low switching and few suppliers give rise to higher persistence. Market volatility reduces persistence.

These are shown in figure 5.18.

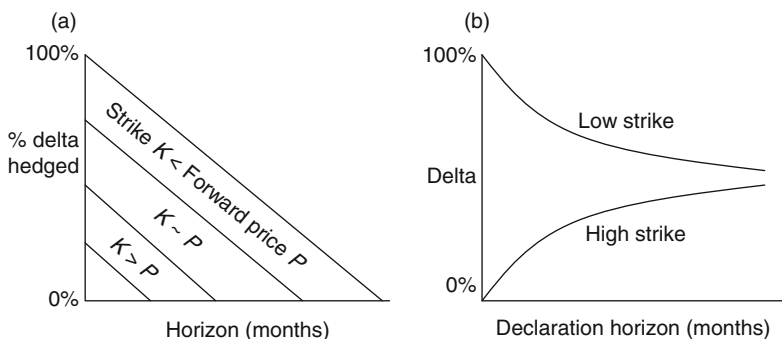


Figure 5.17 Hedge and delta relationships between horizon and strike.

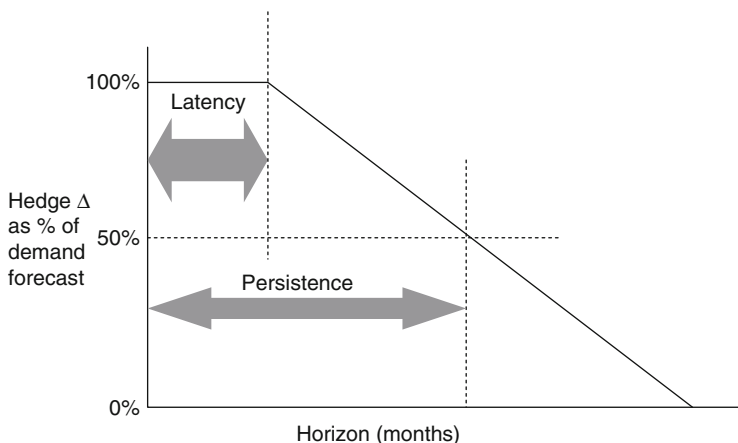


Figure 5.18 Simplification of supplier hedge position.

Suppliers then buy as time progresses, but the core delta does not change with commodity price.

A key issue now is that the supplier hedge tenor is much shorter than the producer (and option holder) hedge tenor. This is the case in all commodity markets.

One producer tactic is to stack hedge, in which a shorter-term liquid contract is hedged as a proxy for a longer-term contract. In practice this is not prevalent, both because a producer is risk averse to the “naked” short exposure in the stack hedge, and second, because stack hedging by producers would involve selling forward more volume than the system demand forecast, and hence liquidity is very limited.

While this is an important effect, we will set it aside for the moment.

Another important effect that we will set aside is that retail suppliers do in fact have concave risk positions that can be expressed, and indeed hedged, by options. The generation stack cost function is convex, so prices rise in a convex manner in relation to demand. Small increases in demand relative to expectation increase supplier profits, but the slope decreases until the point is reached where the wholesale price has risen such that variable costs exceed consumer tariff unit rates. Further demand increases then causes a stronger and stronger effect with the convex stack, and the profits fall sharply from there. In the absence of option trading between producers and suppliers, there is no particular economic reason for the supplier profit concavity in relation to prices to correspond to the producer convexity, and indeed the situation in gas is much complicated by gas-fired power stations that consume gas. Broadly speaking however, the match is a reasonable one, although in practice highly occluded by a plethora of other factors.

5.10.2.1 *Price Dynamics with Call Option Holdings*

In the risk neutral world, the expectation payoff of the option holder is zero. In the absence of hedging, they either just lose the premium or gain a payoff minus the premium. Since the Black option formula is $C = S * N(d_1) - K * N(d_2)$ then we can see by inspection that the probability of exercise $\lambda = N(d_2)$. Non-lognormality affects this but not enough to affect the logic.

By hedging, the actual payoff becomes much closer to zero. In fact the genius of the Black Scholes formulation is that by hedging continuously, under specified conditions such as constant volatility, the payoff is actually riskless.

Let us consider what happens after the financial player has purchased the option expiring a matter of hours before prompt. If the option seller is called, then she calls on the power station unit.

Consider the purchase of a high strike option (i.e., a mid-merit or peak unit, most relevant to the consideration of system capacity). The call option is out of the money. The buyer hedge sells the delta. Now, if the forward price falls, the holder buys back part of the hedge, and if the price rises, she sells. Initially then, she drives the forward price away from the strike price. Only when the market actually crosses the strike does it act as an attractor, with the option holder buying below and selling above. We can see first that dispatch is efficient, as the unit will run if the market price exceeds variable cost (the strike price), but that the forward or spot prices are not driven to the variable cost (which would cause variable cost pricing).

The result of this is that not only do all units therefore run in merit, as the option holders will only declare the option if $P > (K = b)$, but the hedging behaviour, acting as an attractor, will drive the outturn price to the variable cost of the marginal unit.

5.10.2.2 *Market Completeness*

Markets obey the law of one price, meaning that there can only be one price in the market. There can also only be one probability distribution implied from all market instruments, including options at all strikes. Provided that there is a liquid forward market, and that volatility is constant, the riskiness of the arbitrage in buying one apparently overpriced option and buying an underpriced one is much reduced by the ability to delta hedge. So we can consider that there is a single price distribution implied from all options.¹⁴

CAPACITY MECHANISMS

6.1 TYPES OF CAPACITY MECHANISM

There are numerous ways to secure capacity. These are summarized below. The countries listed should be viewed as case examples to review rather than a strict categorization.

1. central planning (e.g., Japan)
2. long term power purchase (e.g., Finland)
3. mandatory vertical integration (e.g., Greece)
4. strategic reserve (e.g., Australia, Ireland, New Zealand, France, Sweden, Norway, Finland)
5. strategic capacity payments (e.g., China)
6. generation contracting by the system operator (e.g., Great Britain, Norway, Sweden, Germany, Netherlands)
7. day-ahead capacity payments (e.g., England and Wales pre-2,000, Chile, all Ireland, Argentina, South Korea, Spain, Peru, Colombia, Bolivia)
8. mandatory forward contracting (e.g., Brazil, Chile, Guatemala, Nicaragua, Panama, Honduras, Costa Rica)
9. mandatory option contracting (e.g., Brazil)
10. installed capacity (ICAP) type obligation (e.g., control areas in the United States, being ISO-NE, CAISO, NY-ISO, PJM)
11. direct demand side contracts (e.g., ERCOT in the United States)
12. virtual power plant (e.g., France, Spain)
13. pure market, called “energy only” (nowhere).

These can be grouped into the following core methods:

1. strategic capacity paid by the taxpayer or consumers, and administered by government or the system operator
2. leave it to the market to decide

3. require retail suppliers to procure energy or options in advance
4. capacity incentive embedded in the short-term price
5. development of capacity obligations of the ICAP type.

The key axis is between strategic capacity decided between government and the system operator, and pure market. The actual mechanisms tend at all times toward one or the other, but this direction changes according to events and the prevailing politics.

We are particularly interested in the ICAP model. In this model, suppliers are required to buy capacity certificates with the general aim that the aggregate purchased volume exceeds the system peak by a reserve margin. Capacity certificates confer no benefit to the supplier other than avoiding penalties.

We will examine the features of this model, and in doing so show that in practice and in theory, and especially in the absence of price caps and the presence of the physical and political ability to make electricity a private good, the natural evolution is toward an energy-only market model.

6.2 DEVELOPMENT OF THE KEY VARIABLES IN THE ICAP MODEL

6.2.1 Ex Post or Ex Ante Methods

In section 5.2.13 we noted that having an imbalance price that differed according to whether the prevailing regime was normal or scarce (possibly with load loss) creates two specific difficulties, namely credit risk and excessive uncertainty to the parties. We showed that it may be in the interests of all parties to set up a rule that they will all be required to purchase capacity ex ante. In this way the aggregate knowledge may be used more effectively to make the system more efficient overall.

The ex post approach is undoubtedly closer to the pure market approach and more efficient, as private knowledge is used more effectively, if the problems of credit and fragmentation of knowledge can be solved.

Attending first to the issue of credit, electricity markets commonly have a credit cover arrangement in which all actors are required to post credit in advance, whether by cash or by letter of credit or by other guarantee. It is therefore quite possible for the market operator to track prospective cash flow from imbalance in relation to credit cover posted, and to call for credit on a daily or intraday process as required. In practice, market operators and regulators have been

reluctant to legislate, enshrine and enact regimes, whereby a failure to post sufficient credit results in timely and firm action in the way that happens on traded exchanges. While exchanges have the facility to forcibly close out trades when the initial and variation margins are insufficient, the action in retail supply has to be much further as there is no asset (in the case of the exchanges, and “in the money” contract) to call against. The supply license has to be withdrawn and an immediate takeover of the company, or the consumers (by a “supplier of last resort” process that incumbent suppliers are required to participate in) switched to a new supplier who takes on the settlement liabilities.

To date then, it appears that the *ex post* method, while the most efficient, is institutionally unworkable at this point. ICAP does in fact generally work on an *ex ante* basis.

The *ex ante* system does have a complete spectrum. So, for example, an *ex ante* capacity requirement that is set on a day-ahead basis is very close to an *ex post* requirement. An *ex ante* requirement set years ahead leaves the obvious question about what to do about changing size of the supplier in terms of customer numbers.

Since electricity has seasonal demand, the optimum horizon for *ex ante* capacity mechanism is sufficient time ahead of the next seasonal peak.

6.2.2 Firmness

In the most basic model—the UCAP unforced capacity model—retail suppliers buy capacity, generators sell capacity, and there is no penalty for actual generator failure, although there may be some *ex ante* generator testing.

The simplest way to allow for generator failure is to apply a percentage uplift to total capacity requirement. One development can be to leave the capacity requirement as it is but downlift generator certificates by an amount that could be related to historic failrates.

Since i) unavailability rates of power stations are very high compared to other assets, ii) security of supply events are commonly associated with generator failure, iii) there can be systemic generator failures, and iv) there is insufficient incentive to maximize reliability in critical times, the UCAP method is considerably exposed to systemic events.

6.2.3 Deficiency

Given that i) the key advantage of the *ex ante* method is the additional certainty to market participants, ii) the *ex ante* method requires a

penalty mechanism, and iii) a solution for generator failure is required, we need a deficiency mechanism.

In the simplest form, an actor who has failed against an obligation, whether this be *ex ante* (e.g., not buying enough certificates) or *ex post* (e.g., a generator failing), there is a buyout charge, which, if not constant, is reasonably estimable in advance.

It is the most pure in market terms for deficiency to be “ruthless,” that is, the payment of a deficiency charge not be regarded as a regulatory censure.

The deficiency charge acts i) as an incentive not to be deficient, ii) to provide funds, for example, for lost load compensation or the purchase of reserve.

6.2.4 Secondary Markets

We saw in section 5.3 that we either need well-crafted models for the initial strike-premium vector and its movement or an effective market *tâtonnement* for the optimum strike-premium vector to be achieved. For the *tâtonnement* to be effective we need a liquid market for options and this may not be realistic.

Accordingly one method of approach is to have an initial auction with the monopsony system/market operator followed by trading in the secondary market. This concentrates liquidity and can allow a first-pass schedule. Then using the price (premium) vector, the market has something to start with.

A good example of a market that can arrive at an efficient run by *tâtonnement* is the bilateral market, and one that relies on a central schedule is the pool.

6.2.5 Requirement Setting for Retail Suppliers

The aggregate requirement ideally relates to the optimal security requirement, equating benefit at the margin to cost at the margin. In practice there are several issues with security requirements:

1. They are commonly unclear in terms of quantity, for example, a capacity margin may include an assumed requirement for reserve that may be substantial (approximately 10 percent).
2. The amount of reserve needed depends on many factors such as transmission constraint, reactive power, and power station inertia.
3. They are commonly unclear in terms of formal function, for example, LOLP must be over a time interval of designated length and apply clearly to either system or grid supply point.

4. They commonly do not recognize amount of load lost, that is, recognizing only LOLP and not either expectation or conditional expectation of GW or TWh.
5. They are commonly biased, the England and Wales pool being an example.

In practice the only way to apply the security requirement is by setting an aggregate capacity requirement, and this is set somewhat arbitrarily using a view of an acceptable LOLP, unconditional loss of load expectation (LOLE), conditional LOLE, and estimated reserve requirement.

Accordingly then, the retail supplier requirement normally refers to their market share of the system peak and not their peak or average demand. The normal way to do this is to look back at the prior year/s and use market share at the peak/s. Clearly there are a number of issues to consider here, such as whether last year's peak was at an unusual time, whether the period of most risk is actually at the system peak, the supplier customer base or market share may be changing, and so on.

6.2.6 Demand-Side Provision of Capacity

The key for the demand side is the reference point. So if a consumer consumed Q MW at last year's peak and has a contract that limits her consumption to Q MW for the same period this year, then the provision of D MW demand response is meaningful.

If however, as is usual, the consumer has a full requirement contract that allows any amount of consumption then there is no reference point from which to offer demand response.

The supplier can however benefit from demand-side management (DSM) that is not submitted for capacity. If this year's requirement is based on last year's demand at system peak, then any DSM has a direct saving for the next year's obligation.

6.2.7 Price Formation for ICAP

We noted in section 5.3 that the system operator can run an auction with a vector that is defined. It was very clear that the auction has to be iterated as offers refine.

A standard method for capacity obligations is to set an official demand for capacity function and then run a descending clock auction, so starting with a high price, the price is gradually lowered until only the required amount of capacity is tendered. The demand

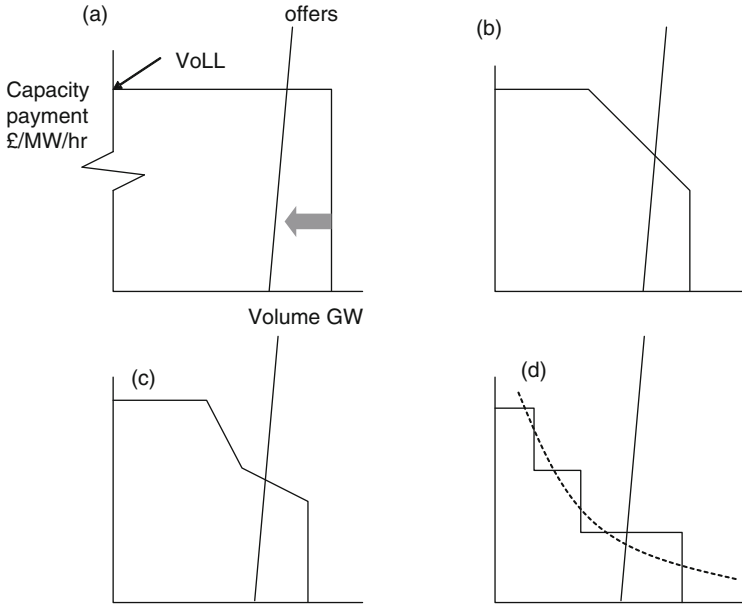


Figure 6.1 Development of the demand for capacity function.

for capacity function is refined over the years. This is shown in figure 6.1. Figure 6.1(d) is conceptual, as in practice, figure 6.1(c) is used.

Figure 6.1(c) has four coordinates. The key one is the peak demand for capacity. This is set by the production side rather than the consumption side, with the logic being essentially as laid out in section 5.2.15, that given sufficient time, the demand for capacity should be set by the lower of production cost and willingness to pay. A common level to use is twice the fixed cost of new entrants (CONE) for peaking plant.

A specific problem here is that in the situation where the production and demand functions are both near vertical and maximum demand is uncertain, that price is very unstable.¹ This requires the demand for capacity to have an artificially low price and flat slope and hence the mechanism caters poorly for demand-side response.

6.2.8 System/Market Operator Setting of Price Caps

We have noted that the setting of price caps is driven largely by the minimization of moral hazard. We have seen also that caps can be

circumvented, and indeed the California crisis² of 2000/2001 provides a good example of such arbitrage in action.

We can apply some broad rules of thumb to the levels of price caps:

1. They should not be too low or too high. A rough range would be within £1,000 and £10,000/MWh. Commonly they are set low.
2. The ideal ordering by size is market, balancing, imbalance, scarcity, VOLL, and in practice this order of balancing and imbalance in particular is difficult.
3. The price caps of each should not be so different as to make arbitrage an attractive proposition to enact frequently.

6.2.9 Interconnection

At high level, there are two distinct approaches:

1. with a focus on resilience, decentralization of electricity production, and DSM, a drive for energy balance at the peak at a local level
2. with a focus on adequacy and long-distance energy flows from primary sources to demand, a drive for interconnectivity and two-way adequacy arrangements between control areas or countries.

Broadly speaking, it is the latter method that prevails and which we will discuss. It should be noted that in practice, network resilience and the reactive power issues that arise from long-distance flows, need to be addressed. In addition, the increased decentralization of production and increased participation in small-scale DSM are likely to change the dynamics of security of supply.

The main focus on interconnection from an ICAP perspective is the recognition of export from the control area as a risk factor and import to the control area as a risk mitigator.

For current purposes there are two considerations.

1. The sale of a capacity certificate requires the ability to deliver electricity in the control area, whether it be direct generation, direct demand response, or import.
2. To guarantee import, any electricity purchased out of state must be firm (e.g., not “recallable”) and in addition must be importable (in particular the interconnector transmission must be secured).

6.2.10 Locational Issues within a Control Area

Where there are local constraints, or wider constraints that due to loop flow have the effect of local constraints, some control areas use the local installed capacity obligations (LICAP) mechanism.

6.3 DEVELOPMENT OF ICAP TOWARD THE USE OF STRIKE PRICES

There are numerous problems with having a capacity obligation with no strike price, particularly,

1. they are public goods that poorly represent the actual lost load demand function
2. their presence forecloses the option market
3. being administered, they cannot respond efficiently to market signals
4. they encourage the lowering of price caps, thereby further denying demand-side participation.

Many of these problems are avoided by having strike prices. This brings many benefits and the principal change is that an option with a strike price has a potential value as a private good.

In the closest model to the energy-only model, the regulator requires all retail suppliers to have not a no-strike capacity certificate (ICAP) but an option, with the rules on volume requirement and ex ante deficiency management the same as for ICAP.

However, given that the presence of ICAP suppresses the option market, it is a large step to go straight from no-strike to any-strike obligation. The market must therefore evolve.

We first consider some actual proposals and mechanisms with strikes.

6.3.1 Transformation of Prices by the Single Buyer and Market Operator

This ability by the market operator in the pool model facilitates the use of dividing the market in regimes of “normal” and “scarcity;” in addition, it allows different prices for generators and retail suppliers at the same time.

The same effect can be achieved in the bilateral market by ex post levies and rebates. While some levies are indeed charged ex post on a halfhourly basis, such as the Balancing Services Use of System charge

in Great Britain, they are less in keeping with the ethos of bilateral markets, which is closer to “energy only” than the pool.

In the pool, the single buyer pays for all power generation as a monopsony and sells to the retail suppliers as a monopoly. Generators and suppliers can trade contracts for differences (CFDs) with each other.

The single buyer has discretion in the construction and transformation of the prices for generators and suppliers. For example, in the England and Wales pool, the Pool Purchase Price (PPP) paid to generators was a construction from the System Marginal Price formed from the generator offers, and the Pool Selling Price for suppliers had an uplift from PPP that could have been adjusted.

6.3.2 Effect of Pool Price Transformation on Hedges

Since CFDs are settled against pool prices, it is important to recognize the effect of pool price transformation on these. For example, if a retail supplier buys an option from the generator that is financially settled against the ex post pool index, and the power is drawn as expected, and the market operator transforms the prices in such a way that the “basis” difference between the pool price index and the price paid by the retail supplier changes, then basis risk is introduced. Suppose the PPP used for the option is not transformed, the price received by the generators is transformed relative to the PPP, and the pool selling price paid by suppliers follows the transformed PPP, then risk is introduced to the retail supplier.

6.3.3 Development of ICAP by Lowering the Option Strike below the Cap

Oren (2005) proposed a weaning off from ICAP capacity obligations by the inclusion of a strike price below the market cap in the obligation.

He also suggests a ban on short selling by generators, so if they sell ICAP certificates, they must offer physical plant to support the sale, rather than simply plan to buy back financial power to support the contract. The same logic would preclude short selling in the secondary market

We can approach the lowering of strike this in three ways:

1. ICAP becomes a private good to the retail supplier.
2. ICAP remains a public good forcibly paid for by the supplier; the option is owned by the system operator.

3. The intermediate case in which the option is a private good with part of its value surrendered to the market operator.

In the former case we are effectively allowing the presentation of an option in lieu of a capacity certificate. We can regard the purchase of a capacity certificate as the purchase of an option struck at $P_{\text{imbalvollcap}}$, that is, the highest possible price, as described in section 5.2.9. The valuation of lower strike options is then as described in section 5.2. As described in sections 5.2.6–5.2.10, the development of the option market depends largely on the setting of price caps. Additionally, the deficiency mechanisms can make a substantial difference.

With regard to the second case, where the option is a purely public good, there is no incentive for the retail supplier to buy an option of lower strike than is mandated. The market development will be minimal.

In the third case, we can regard this as the purchase of an option from the generator plus the surrender of a call spread to the market operator. The value of this spread is closely related to the peak energy rent operating in some markets.

6.3.4 Strike Prices and the Regulatory Option Tender

Vázquez, Rivier and Pérez-Arriaga (VRP, 2002) proposed what they call a “market approach,” which we will call, for taxonomical convenience, a “regulatory option tender.” The essence of the model is that the regulator should require the system operator to purchase options from generators, possibly via a market.

VRP recognize the difficulty of having an auction with a continuum of strike prices, and propose that the regulator specify the volume of each tranche of options. We saw from section 5.2.7 that it is not hugely problematic for generators to offer options at a designated strike (here being that of the tranche) rather than exactly at variable costs.

Bidwell (2005) proposes a similar solution. The generators sell options with strike prices, and these options are called by the system/market operator. The strike prices can be above the market cap.

Figure 6.2 shows a generalization of the model in which there is a regulatory option tender in which the single buyer buys all options in the market.

The system operator ends up with a stack that can be represented as in figure 6.3. The vertical thickness of the area in gray in figure 6.3(a)



Figure 6.2 The regulatory option tender by the single buyer.

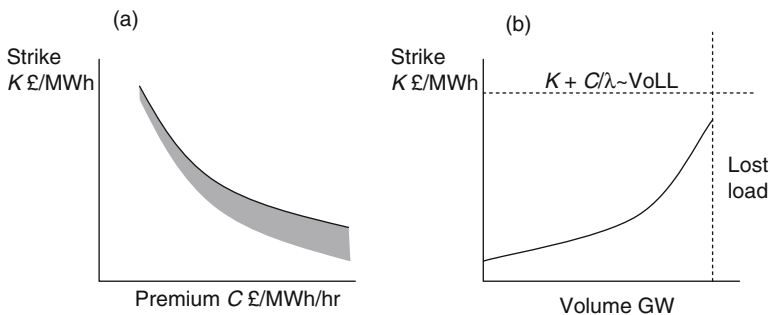


Figure 6.3 Single buyer of options showing construction of the peak price given a unit selection.

represents the installed volume at any particular strike/premium pair. This translates to figure 6.3(b) in which we can see that the probability of lost load is finite. We can see the convergence of VOLL and the peak load price for the peak unit.

If the single buyer buys all the options, then to cover her costs she must charge the retail suppliers according to peak load pricing.

The suppliers are then exposed to price risk. They cannot hedge with the generators, physically or financially, because the generators are already hedged with the single buyer. The single buyer however has a risk that broadly matches the suppliers and hence forms a natural counterpart for forwards and options. In the extreme, all of the options are sold and then the single buyer has no risk and the retail suppliers have their risks greatly mitigated by their options.

The market in options is completely foreclosed, but in theory at least, for inelastic stochastic demand, the aggregation of knowledge with the system operator means that the aggregate option purchase is optimal.

No formal capacity mechanism is required, but in its absence there does exist the credit risk that an unhedged retail supplier defaults, when, to satisfy their demand, they must buy power at peak load pricing in the peaks but has to sell at retail tariff.

Figure 6.4 shows how the market can develop in 6 stages:

1. The system operators buys all options and sells power in real time according to peak load pricing as above. The system operator funds the extrinsic value via a capacity obligation levy—essentially an enforced deficiency payment.
2. The system operator buys the options and offers forward hedges to retail suppliers, continuously or in auction rounds.
3. The system operator offers options to retail suppliers, and option purchases relieve the suppliers of part of their capacity obligation.
4. The generators offer options to both the system operator and suppliers, and demonstration of option holdings relieves the suppliers of capacity obligations.
5. The system operator participates only as a backstop, taking deficiency payments for ex ante shortfalls in option holdings and using these to fund option purchases, effectively bringing on “difficult” capacity (reducing unreliability, running units beyond maximum generation, delaying close, relaxing environmental limits, accelerating new peaker build, securing transmission and generation import, etc.).
6. The market operator treats deficiency purely as credit risk. Retail suppliers are required to post collateral (cash, letter of credit, parent guarantee, etc.) in relation to deficiency. Deficiency can be defined variously such as having insufficient options and similar contracts.

6.3.5 Discussion of the Introduction of a Strike Price

The best energy-only solution is for retail suppliers to have an ex ante requirement to purchase options in volume relating to a volume corresponding to the demand forecast expectation at the expected system peak, with a series of deficiency auctions. An additional comfort factor is commonly added, and 10 percent would be normal. The system operator then ensures that all deficiency payments (plus/minus a residual payment that flows back to suppliers over a period of time) fund the purchase of adequate reserve.

What starts as the other extreme is when the system operator is a single buyer monopsony and then disburses options, and/or capacity certificates, and levies deficiency charges.

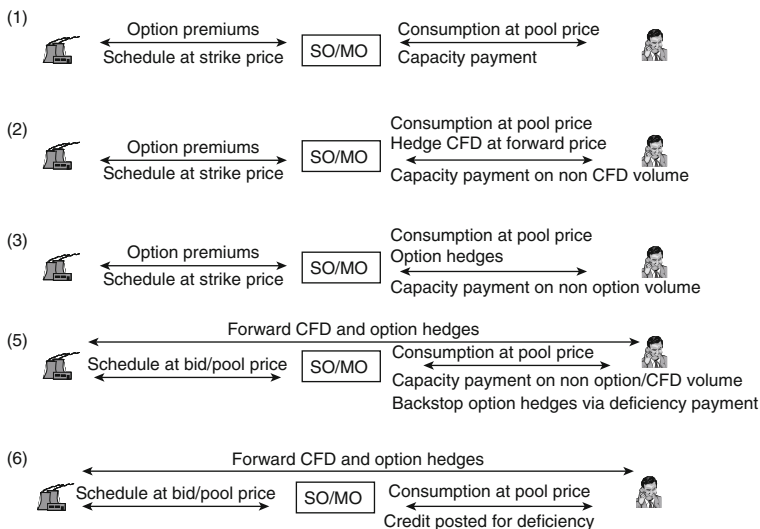


Figure 6.4 Development of option capacity from fully administered to fully market.

The challenge of the energy-only method is getting it going with the market prices capped too low and options foreclosed by the existing capacity mechanism. By gradually raising the cap and allowing options in lieu of capacity certificates, the market can develop.

The challenge of the single buyer method is in gradually allowing direct purchase by suppliers. This can be achieved by the system operator gradually increasing the lowest strike that it purchases and opening the low strike sector to the market. Option purchase would need to be notified to the system operator so that she can tally the total capacity.

6.3.6 Forms of Peak Energy Rent Remission

We showed in section 5.1.2.3 how in a pool market the market operator has the capability to transform prices. We see an example in the Forward Capacity model.

Cramton and Stoft (2006) (CS) note that while many origins and designs of the capacity models start in different places, their adaptations are making them converge. They note the problem of fixed cost recovery—the “missing money,” particularly in the presence of price caps, and propose a solution.

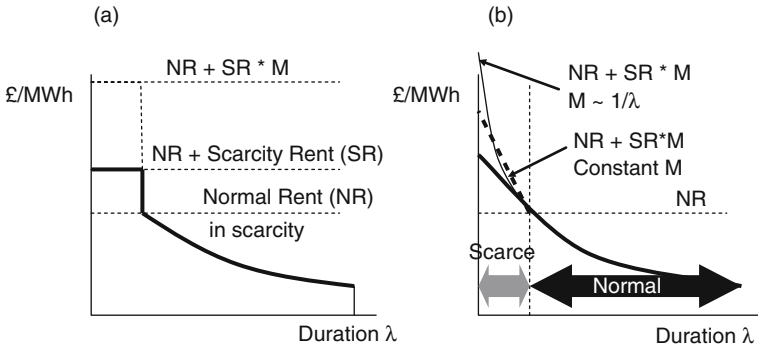


Figure 6.5 Pool price transformation in scarcity.

Here we follow a generalization of the CS approach.

Figure 6.5 shows a simplified variable cost generation stack, approximately conforming to the American model in which peaking plant is built for peaking and is of uniform technology. The scheduling of peakers effectively defines scarcity conditions. The pool price is uplifted as shown in figure 6.5. The key here is

1. while generators do earn scarcity rent, the moral hazard on the generator side is reduced by the administratively determined value of the rent
2. retail suppliers pay the Normal Rent (NR) on the day, but fund ex ante the generator Scarcity Rent (SR) by an ex ante charge.

We can see here that there is discretion on the multiplier M of the scarcity rent. It can be constant ex ante, or variable according to circumstance.

Consider now the position of the retail supplier. As we can see in figure 6.6, the risk profile is that of a call option, with the premium being the capacity payment.

From the generator perspective, if all units receive $NR + SR$, then if the scarcity rent is equal to the variable cost of the peakers plus the fixed costs divided by the ex ante peaker load factor, that is, , we can see that we have peak load pricing.

An alternative model is in which the retail supplier buys options from the generator but is required to forego part of the payoff. We can regard this as the purchase of a call option and the surrender of a call spread.

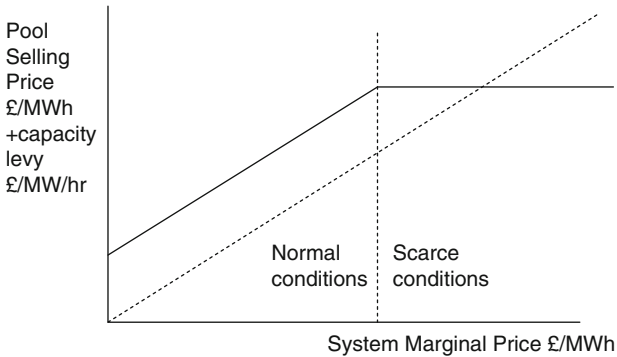


Figure 6.6 Retail supplier pool cost under normal and scarcity conditions.

It is of course important that the retail supplier receives an appropriate amount of hedge benefit from the option purchase. This however happens to some degree automatically. For example, if the option has a strike b then if the volume purchased covers physical demand that is notified, then the option potentially saves cashout at the imbalance price. If however the supplier does not need to volume, then they sell the PN into the market and gain a payoff of $(S - b)$ where S is the prevailing market price that may be at the cap. The additional payoff of up to $P_{\text{imbalcap}} - P_{\text{cap}}$ is automatically lost.

If the capacity certificate is a public good to the retail supplier but an option with a strike for the system operator, then the peak energy rent goes to the system operator.

Consider, for example, the purchase of an option at strike, b being the variable cost of the generating unit. Figure 6.7 shows the various possibilities of surrender of call spread parts of an option struck at the generator variable cost b .

Obvious shortcomings of the hybrid models are i) they are highly contrived, ii) the monopoly/monopsony situation creates conflict of interest for a system operator, iii) the price lacunae are complex and opaque, iv) there is a quasi-market above the cap that is opaque and controlled, and v) the determinants such as whether the system operator decides that the market is tight are opaque and subject to conflict of interest.

At best these models can act as temporary patches to ICAP models in which excess or insufficient rent for generators or cost to suppliers has become manifest.

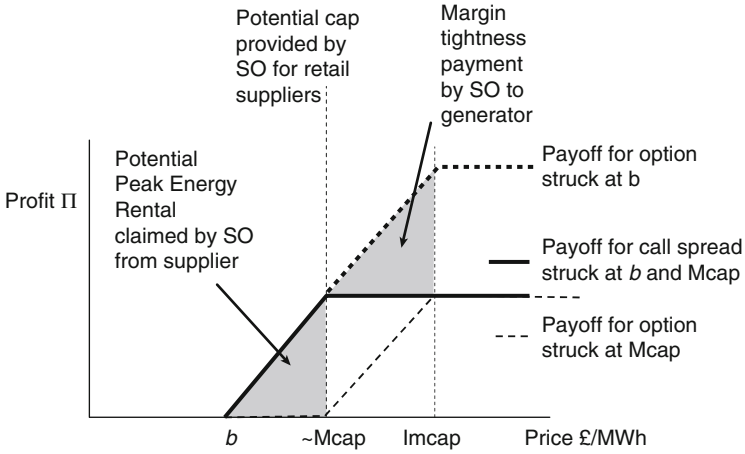


Figure 6.7 Depiction of peak energy rental and scarcity rental in an options format.

6.4 DIVISION OF THE MARKET TO REGIMES

We have seen that the system operator can divide the market into regimes. Examples are:

1. transformation of the pool price under scarcity conditions
2. elevation of the imbalance price under scarcity conditions.

It is almost impossible at one time to obey the “market-first” principle and to participate at points of market failure. The standard method is to limit system operator intervention only to peak production, generally by controlling a unit but only running it when there are no other units available.

Almost by definition, the system operator will either pollute the market with bids subsidized by socialized costs, or pay more for peak capacity than consumers are willing to pay.

Where we can apply a specific regime is when there is actual lost load on the transmission system.

6.5 GENERAL DEVELOPMENT OF CAPACITY MECHANISMS TOWARD ENERGY-ONLY MARKETS

We can characterize the general development of capacity mechanisms toward energy-only markets with a single theme, which is the gradual movement from ex ante administered estimates to live values as expressed by the market, and settled ex post.

The three key developments are:

1. strike prices—conversion of options/capacity to private goods
2. generator deficiency—ex post deficiency charging for actual failure at prevailing prices
3. supplier deficiency—conversion of ex ante requirement to the posting of credit.

This presents an interesting dilemma to a country with a developed market in energy and a nascent market in options. Should they i) create a capacity mechanism and develop that over time so that in the end it resembles an energy only mechanism? or ii) do they incentivize development of the option market, for example, by accepting option notifications in lieu of capacity certificates? This is shown in figure 6.8.

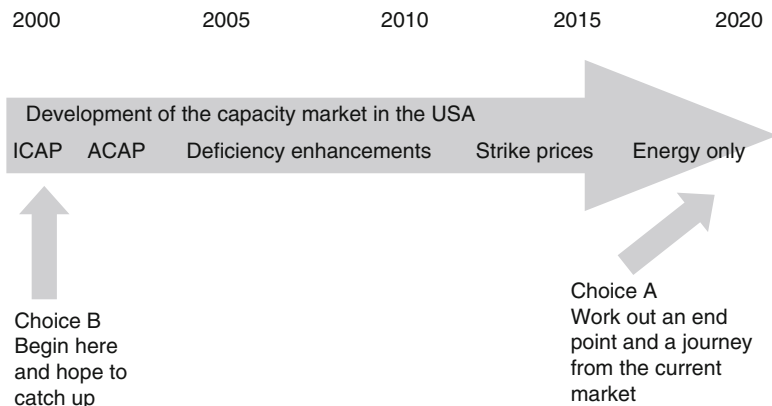


Figure 6.8 Development of the ICAP model toward an energy only model.

THE POWER COMPLEX

The modeling to this point has been based on the old-world paradigm of treating consumption as stochastic and inelastic and having an administered view of the value of lost load (VOLL).

We have also paid scant attention to the importance of transmission constraints, neighbor markets and the physical interconnection to them, and consideration of the complexities of loop flow in networks.

We now address these.

7.1 THE DEMAND SIDE

7.1.1 Modeling Shocks to the Load Duration Curve

We have shown that largely due to the stochasticity of the phase of demand and also the cross-elasticity of demand across time periods, load factor modeling is essential for peak load and capacity modeling.

One reasonably straightforward way to do this is using principal components (PCs). All PC shocks are orthogonal, in that they do not affect one another. This is represented in figure 7.1.

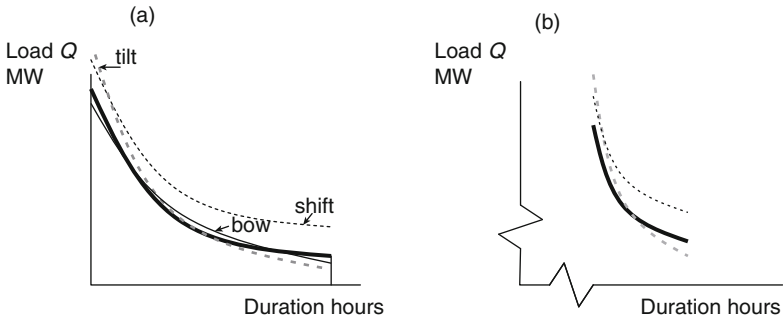


Figure 7.1 Principal component shocks to the load duration function (a) Whole function (b) Peak area of the function.

We are particularly interested in the peak, hence for consideration of peaks only we just shock this area.

By applying one, two, and three standard deviation shocks to the first four PCs we now have for any duration the probability profile of load with 25 points. Similarly for any load we have the duration profile.

7.1.2 The Changing Paradigm of Decentralized Energy

The England and Wales pool typifies the old-world paradigm of treating generation as predictable (reliability being well-characterized in short horizons) and flexible and consumption being stochastic and inelastic.

As we see in figure 7.2, the new-world paradigm is the reverse, mainly due to

1. general movement from fossil fuel power stations with discretionary load to low-carbon generation, which is either “must run” with energy flow determined by nature (wind, sun, etc.) or designed largely for baseload operation (nuclear)
2. modern power stations being designed in recognition of the need for flexibility but in reality so finely tuned that there is less inherent flexibility
3. generation units decentralized and therefore much smaller and with less sophisticated control systems
4. potentially large increase in power consumption from the electrification of heat and power in which the time of user demand may be inflexible

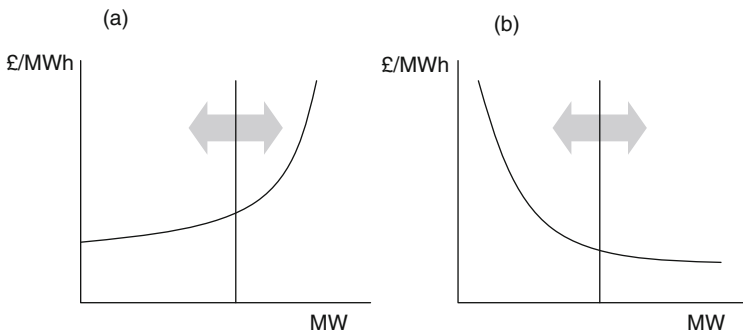


Figure 7.2 Change from old world to new world production-demand paradigm (a) Old world with flexible generation and inflexible stochastic demand (b) New world with flexible demand and inflexible stochastic generation.

5. the necessity of demand-side management of the load of items such as heat pumps and electric vehicles, in order to minimize transmission and distribution network and generation capacity build
6. the general change in demand patterns and the response of consumers to price, meaning that the system operator's demand forecast will become increasingly inaccurate.

7.1.3 "Smart" in Metering and Other Areas

The market arrangements for consumption have been largely driven by a single factor, namely the extraordinarily poor temporal resolution of the measurement of energy flow compared to that needed by consumers and the system.

Fortunately this is changing in this decade, and we can expect the meters to have halfhourly resolution of energy, polled on a regular basis. This commercially enables a vast and still-to-be discovered infrastructure of smart, being grids (distribution systems), devices, applications ("apps"), algorithms, consumption, communication, heuristics, tariff structures, and so on.

What counts for present purposes is that with a universal smart system, electricity, including security of supply, becomes a private good. Smart meters can reduce/prevent the flow of electricity, and even if there is not the regulatory functionality to do so, it can be effected by the consumer setting the algorithm to stop consuming on receipt of a price signal and thence the sending of the signal.

7.1.4 Fixed and Variable Costs of the Demand Side

We have seen that it is very common to treat demand side management as a single VOLL with no fixed costs. We have also seen that it is quite possible for entities that have fixed costs to load the fixed costs onto the price received if there is no premium in advance.

This approach is inadequate for integrating the demand side into peak load and capacity modeling for the following reasons:

1. Generation and demand can be harmonized if both use the mechanism of a premium (for fixed costs) and a strike price (for variable costs).
2. Demand-side contingency for lost load has in practice a fixed cost, for example, a portable generator.
3. In the absence of a premium, the uplift applied to VOLL for low probability events is too extreme for the valuation to be practical.

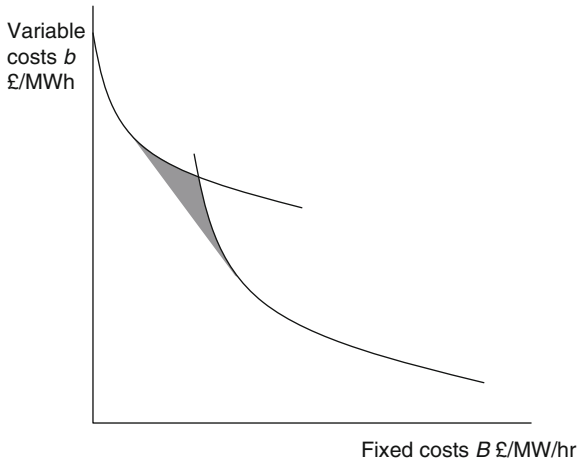


Figure 7.3 Composite frontier from production and demand-side management.

We must therefore treat the demand side as having fixed costs, variable costs, and capacity. We can see this in figure 7.3. The gray area shows where the frontier ends in theory if the volume of certain areas of the production and consumption “stacks” can be expanded.

7.1.5 Consumer Protection

We have at the same time two paramount requirements:

1. The ability of consumers to participate in the market, both out of democratic fairness in terms of access to market and policy reasons for managing the system in the new-world paradigm.
2. The need to protect from high prices those consumers who cannot access the market for reasons of vulnerability.

These can be achieved together and indeed it is the ability of the active consumers to respond that provides both the security at a reasonable price that vulnerable consumers require and the cost efficiency that allows the general lowering of bills.

However it is essential to ensure proactive consumer protection at the level of the individual.

The welfarist approach to this is to find (each and every) individuals who are disadvantaged by inability to respond effectively to price signals without hardship and provide energy efficiency and demand-side

management free or heavily subsidized, or apply financial compensation directly to the bill.

7.2 TRANSMISSION AND INTERCONNECTION

7.2.1 Policy Modeling of Capacity Obligations in a Networked Island Economy

We have examined in some detail the nature of consumers and consumption, and the significance of the ordering and timing of stages, such as capacity build, pricing, and uncertainty resolution. We can now look at capacity from a whole system perspective, adding in considerations of a connected foreign market, and domestic and international transmission constraints.

We can take a gaming analysis to the efficacy of capacity obligations under a set of simplifying assumptions. The analysis here closely follows the calculus of Creti and Fabra (2004) (CF), which was inspired by the Pennsylvania–New Jersey–Maryland (PJM) market, although there are some material differences in description in order to apply to an unbundled market.

This framework is particularly useful because it has the key ingredients needed to model the efficacy of a capacity obligation and the flexibility to relax the key assumptions.

7.2.1.1 *Characterization*

1. Single demand period.
2. Demand—right-angled demand function with willingness to pay v . Stochastic with zero probability above a maximum. Stochastic form is twice differentiable in all domains with probability density $g(D)$ and cumulative $G(D)$, as shown in figure 7.5.
3. Domestic transmission—perfectly reliable. Fixed and variable costs nominally zero up to capacity limit K at which variable costs become infinite.
4. Interconnector to foreign market—perfectly reliable. Fixed and variable costs nominally zero up to capacity limit β in both directions, at which variable costs become infinite.
5. Production—fixed and variable costs nominally set to zero. Perfectly reliable. Capacity $K > \beta$ and $K + \beta \geq D_{\max}$ that is, demand can always be fulfilled. Infinitely flexible with zero state change cost. Infinite life.
6. Retail supply market— n actors of equal size. Regulated.

7. Market structure—fully competitive and monopoly both examined. No trading of energy or capacity between suppliers or between generators.
8. Domestic regulator—benign optimizer of welfare. Assigns an attenuation factor of α to generator welfare. The regulator sets a price cap $P < f$, a capacity certificate price cap C , and mandatory capacity purchase Θ across the retail suppliers.
9. Foreign market—always at price $f < v$ and infinitely elastic.

The limit of demand to $K + \beta$ is somewhat contrived. We must assume that generation and transmission capacity evolved to meet this limit. In doing so however, we must have made assumptions about the probability distribution of demand, since the probability of demand at the near the $K + \beta$ limit must be sufficient to merit the build.

The sale of a capacity certificate requires the generator to provide power on demand, whether by generating without selling abroad, or by importing. The purchase of a capacity certificate gives the retail supplier the license to operate. We can treat this as having zero commercial value but an infinite fine for not having sufficient capacity to redeem against the regulatory obligation.

The overall system is shown in figure 7.4.

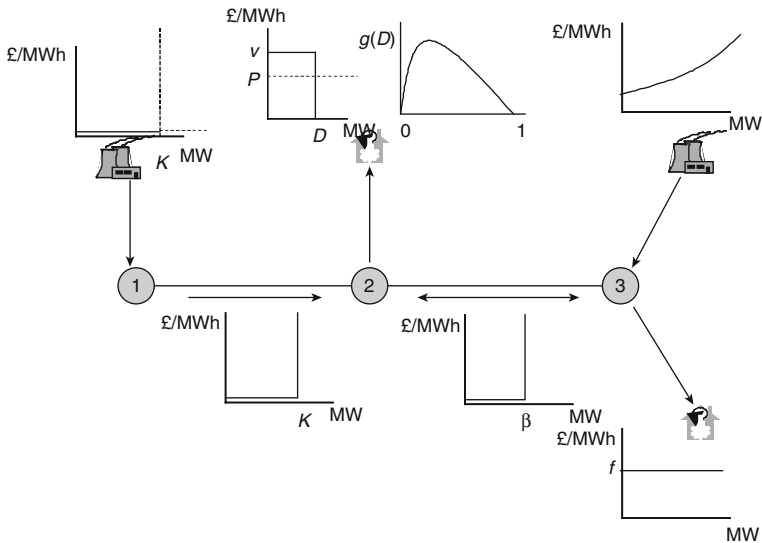


Figure 7.4 Depiction of the system described. Nodes 1 and 2 are domestic and node 3 is foreign.

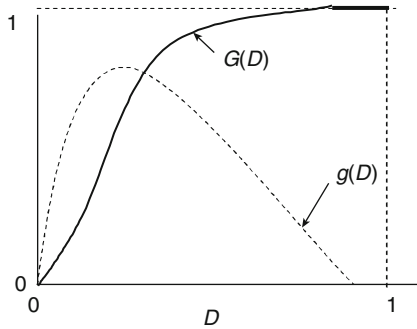


Figure 7.5 Cumulative and density probability functions of demand D .

7.2.1.2 The Game

We consider a two-period setting with one consumption period. In the first period, suppliers have unbiased estimates of expected consumption and buy capacity according to regulations. The regulator also has perfect and unbiased estimates of ex ante demand distributions in aggregate and by supplier. In the second period, the consumption uncertainty is resolved and the energy flows.

We assume that build can be executed after the capacity order and price cap setting by the regulator and before any reforecast of demand becomes material.

We now solve by backward induction with the sequence shown in figure 7.6. We begin with the energy market competition game, and then move to the capacity market. Finally, we analyze the regulator’s problem, who has to set the capacity obligation Θ , the cap C of the capacity price, and cap P of the market price.

We will assume that the regulator can guarantee 0 percent probability of lost load as we see in figure 7.7(a).

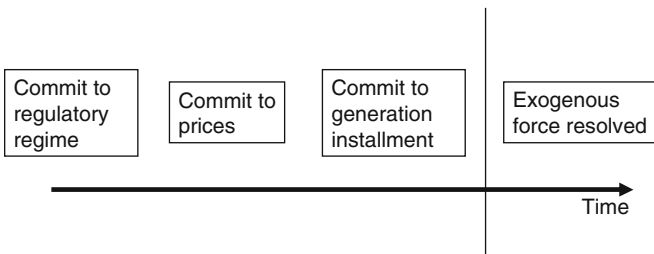


Figure 7.6 The stages of the game.

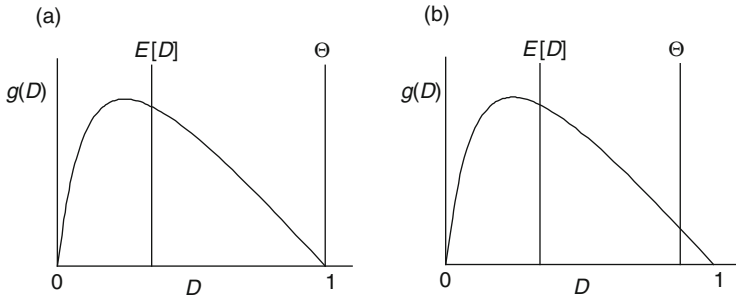


Figure 7.7 (a) Regulator sets capacity requirement for 0 percent probability of lost load (b) Regulator sets capacity requirement for finite probability of lost load.

7.2.1.2.i Example 1—The Monopoly Generator with Regulated Prices
 Since demand for energy is inelastic, the monopoly generator optimizes her revenue in the energy market by offering at the cap P . In addition, if the regulatory environment is such that demand for capacity is inelastic, then she also optimizes her revenue by offering capacity at the cap C .

Her decision is how much capacity Θ to offer. Since $f > P$, she exports all uncommitted energy to the full capacity of the link. This leaves her with nonexportable capacity $K - \beta$.

If she has sold capacity, the cost of default is high (it should be equal to v). In this case, since we assumed that the probability of demand exceeding domestic capacity plus transmission constraint is zero, she can go to the foreign market and buy energy at f and sell it in the domestic market at P , and hence we simply set the default rate to be greater than f .

Let us first consider the case for the sale of capacity at below the nonexportable limit $\Theta < K - \beta$.

We can see from figure 7.8 that the optimum energy sale level by the generator is unrelated to the capacity sale Θ , for all $\Theta < K - \beta$, and therefore the generator will sell at least $K - \beta$ to gain the capacity income $C * (K - \beta)$. The profit expectation is

$$\begin{aligned} \pi_m(\theta) | \theta < K - \beta &= C\theta + \int_{D=0}^{D=K-\beta} [PD + f\beta] dG(D) \\ &\quad + \int_{D=K-\beta}^{D=D_{\max}} [P(K - \beta) + f\beta] dG(D) \end{aligned}$$

Figures 7.8–7.10 show the actions for the probability domains. The generator exports, even when the supplier imports, netting off the flows.

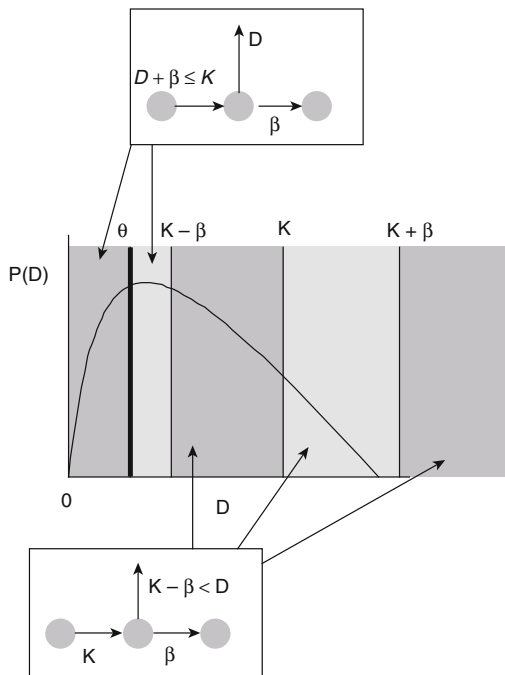


Figure 7.8 Contractual energy flows for capacity sale $\Theta < K - \beta$ for the different probability domains of demand.

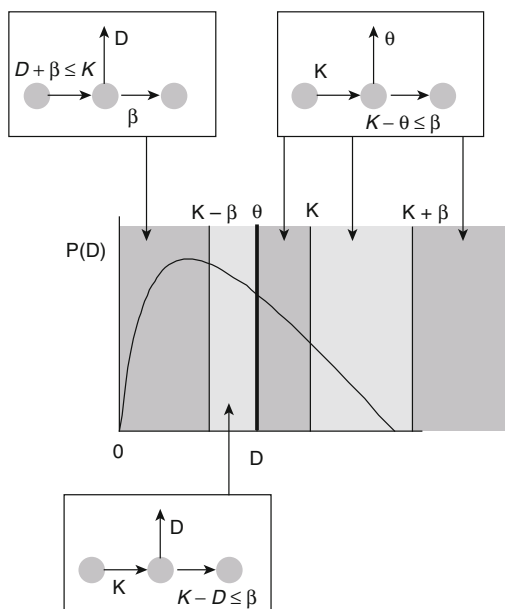


Figure 7.9 Contractual energy flows for capacity sale $K - \beta < \Theta < K$.

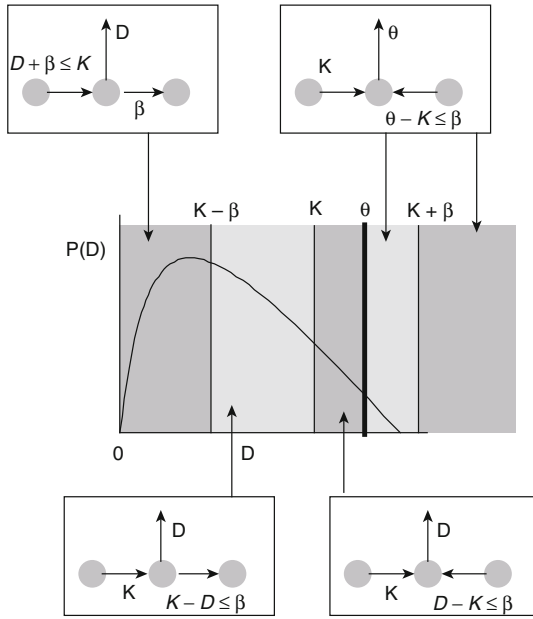


Figure 7.10 Contractual energy flows for capacity sale $K < \Theta < K + \beta$.

We generalize slightly from the CF framework, allowing a finite probability λ of a demand outturn $D_{\max} > K + \beta$

So the profit, including the capacity income is

$$\begin{aligned} \pi_m(\theta) | K - \beta < \theta < K &= C\theta + \int_{D=0}^{D=K-\beta} [PD + f\beta] dG(D) \\ &+ \int_{D=K-\beta}^{D=\theta} [PD + f(K - D)] dG(D) \\ &+ [P\theta + f[K - \theta]] \int_{D=\theta}^{D=K+\beta} dG(D) \\ &+ \lambda [P\theta - f(K - \theta)]. \end{aligned}$$

To find the optimum Θ , conditional on it being within a specified range, we differentiate the conditional profit with respect to Θ .

If the capacity price is above a critical level C_m , we will commit some level Θ rather than $K - \beta$.

$$C_m \Theta + \pi_m(\Theta) \geq C_m * (K - \beta) + \pi_m(K - \beta)$$

$$C_m = \frac{(f - P) \int_{K-\beta}^{K+\beta} D dG + (f - P)(\beta - K) \int_{K-\beta}^{K+\beta} dG}{2(\Theta - (K - \beta))},$$

where C_m denotes monopoly.

Note the resonance between the numerator and the Chao analysis in section 3.1.1. We have a term on the right-hand side that is only related to the probability of demand exceeding the nonexportable amount $K - \beta$, and a term on the left-hand side, which is related to the probability distribution of demand. The key term for the matter in hand is the term on the left-hand side. Depending on the shape of the distribution, Θ could be anywhere from $K - \beta$ to $K + \beta$.

To understand this equation, let us rearrange it. To simplify, we assume that $\Theta = K + \beta$.

$$C_m = (f - P) \frac{1}{4\beta} \int_{K-\beta}^{K+\beta} [D - (K - \beta)] dG.$$

At the extremes, $D = K + \beta$ and $D = K - \beta$, and correspondingly

$$C_m = \frac{(f - P)}{2} \text{ and } C_m = 0 \text{ respectively.}$$

This has the form of the call option equation, only this time we have a fixed payoff with variable volume, rather than vice versa. For lognormally distributed demand, we have off-the-shelf solutions for this, and the noninfinite upper limit of the integral is easily handled by regarding the option as the difference between two options with an infinite limit (i.e., a call option spread).

We will in a moment consider the position of the regulator, but first let us examine the position of the generator facing perfect competition.

7.2.1.2.ii "Perfect Competition" in Generation

The model that CF are looking for is the sale of domestic energy at a price that does not exceed variable costs. We can arrive at this construction in a number of ways. One way is an infinite number of competitors who, while managing to come to an arrangement in sharing the profitable export market, have not done the same in the domestic market, and instead offer at variable costs. Another way is for the regulator to proscribe internal sales above variable costs, and allow a sharing of the export market.

Similarly, the capacity market is perfectly competitive and hence clears at the opportunity cost of the marginal seller of capacity.

We must now take a moment to consider what the build algorithm is. If build has already been done, and capacity is in excess, then given that energy sales cannot be above variable cost, the policy of offering capacity must dominate the policy of not offering capacity. However, in the event of allowed collusion or Cournot competition in capacity, it might be optimal for producers to offer less capacity than would be offered from a competitive market and an upward sloping capacity cost curve.

Our profit equation is similar to before, only now the (ex ante) marginal profits for energy sold in the domestic market are zero.

We commit Θ rather than $K - \beta$ if

$$c * \Theta + \pi_c(\Theta) \geq c * (K - \beta) + \pi_c(K - \beta),$$

where now c is the clearing price in the capacity market.

Rearranging and substituting as before we have

$$c \geq C_c = \frac{\pi_c(K - \beta) - \pi_c(\Theta)}{\Theta - (K - \beta)}$$

where C_c denotes competition

After some more calculus we have

$$C_c = \frac{f \int_{K-\beta}^{K+\beta} D dG + f(\beta - K) \int_{K-\beta}^{K+\beta} dG}{2(\Theta - (K - \beta))}$$

We can compare the critical capacity price for perfect competition C_c to that for monopoly C_m .

$$C_c = \frac{f}{f - P} C_m.$$

So the critical capacity price in competition exceeds that in monopoly.

We now have the capacity price bounds between the extremes of monopoly and perfect competition, and can then, in theory at least, use this to estimate the use of market power. This may be exercised in part to ensure fixed cost recovery.

The relationship between capacity price cap and the capacity commitment is shown in figure 7.11.

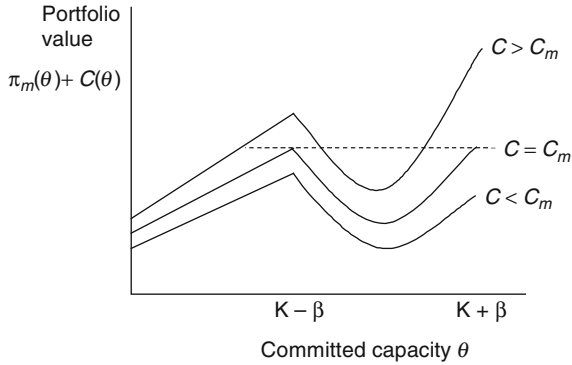


Figure 7.11 The critical cap price for the producer to offer capacity.

7.2.1.2.iii The Regulator

Finally we consider the position of the regulator. His decisions are

1. whether or not to introduce a capacity obligation and a capacity market
2. at what level to set P , C , and Θ .

Let us think initially why a regulator would want to intervene between a willing buyer and willing seller. It can only be because some form of coordination between producers results in a volume of capacity offering that is less than the amount that they would offer without coordination, or because either a buyers consortium is impractical or illegal or with conditions that cannot be enforced.

We begin with the monopoly case.

We, (i.e., Creti and Fabra) will show that

1. There is some level of welfare level $\underline{v}_m(\alpha, f, P, K - \beta)$ above (and only above) which it is optimal for the regulator to impose a capacity obligation.
2. For $v > \underline{v}_m$ and where $\theta_m < K + \beta$, the optimal choice of capacity obligation is maximum possible demand $\Theta_m^* = 1$.
3. For $v > \underline{v}_m$ the optimal choice of capacity price cap equals the minimum value at which the market clears $C_m^* = C_m(1, K - \beta, f - P)$.

Again we work by backward induction. First, we examine the optimal conditions if there is a capacity obligation, and then we compare

this to the welfare with no capacity obligation to see if it worth having an obligation.

Creti and Fabra ignore the case of $\Theta < K - \beta$, arguing that regardless of a capacity market, the generator will install this amount of capacity without being compensated for it in advance, as he can export at a price higher than his costs.

We can also ignore $C < C_m$ since below this price, as we saw above, the generator will only offer nonexportable capacity.

So, for $\Theta > K - \beta$ and $C > C_m$ we consider the total welfare under monopoly and in the presence of a capacity market, W_m^c , for a given Θ , C by differentiating welfare with respect to capacity obligation.

CF give us

$$\begin{aligned} \frac{\partial W_m^c(\Theta, C_m)}{\partial \Theta} &= [v - f][1 - G(\Theta)] \\ &+ [1 - \alpha][f - P] \frac{[K - \beta]}{\Theta - [K - \beta]} \int_{K - \beta}^{\Theta} \left[\frac{D - (K - \beta)}{\Theta - (K - \beta)} \right] dG(D). \end{aligned}$$

Both terms exceed zero and hence within the specified range $\Theta < 1$ it is optimal to increase Θ . Hence the optimum mandatory regulatory capacity $\Theta_m^* = 1$.

7.2.1.2.iv Benefit of Capacity Obligations

So, under monopoly conditions, the benefit of having a capacity obligation is;

$$\begin{aligned} W_m^c(1, C_m) - W_m^{NC} &= \left[[v - P] - [f - P] \frac{1 - \alpha(K - \beta)}{1 - (K - \beta)} \right] \\ &\int_{K - \beta}^1 (D - (K - \beta)) dG(D). \end{aligned}$$

So, there are conditions under which it is welfare optimal for the regulator to apply a capacity obligation.

We can see that the right hand side of the above equation is positive if and only if

$$v_m = v > \underline{v}_c(\alpha, f, K - \beta) = f \frac{1 - \alpha(K - \beta)}{1 - (K - \beta)}$$

For $\alpha = 1$, this simplifies to $\underline{v}_c = f$

So if VoLL is less than the foreign price there should be no capacity obligation

For the case of $\alpha = 1$, let us simplify equation (8) above.

$$W_m^C(1, C_m) - W_m^{NC} = (v - f) \int_{K-\beta}^1 (D - (K - \beta)) dG(D)$$

We note the similarity of this equation to the critical capacity price equation and the similarity of form to the option equation. We can see by inspection, that the expectation of saving from having a capacity obligation is simply equal to the payoff to the consumer of being able to import volumes above the nonexportable capacity, rather than lose load.

So, there is indeed a welfare benefit from having a capacity obligation, but this is in the context of a specific institutional constraint—capping the price P at below the foreign price f . In effect we are facilitating the sale of energy at price f by the producer, but bypassing the rules by loading the cost of this purchase into the capacity payment rather than the energy payment. Looking at it this way, we can see why the equation has option form—the producer simply weights the capacity payment (the premium) against the expectation payoff. It is particularly useful then to consider the producer welfare weighting α , in circumstances where the capacity has been built and the producer is exposed to moral hazard on the part of the regulator.

7.2.1.3 *Conclusions and Dependence on Assumptions*

Creti and Fabra make key conclusions that are consistent with the analysis presented above. The lemma's are, in summary:

1. With a retail supplier monopoly, if there is a capacity obligation with capped price, then it is optimal for the generator to offer sufficient capacity, above the nonexportable capacity, only if the capacity price cap exceeds the free market clearing price for capacity.
2. The same conclusion as above for a retail supplier market that is competitive in terms of the description.
3. Under monopoly, it is optimal for the regulator to introduce capacity obligations if and only if the VOLL exceeds a particular level, which is a function of α , f , P and $K - \beta$. With the demand limit, it is optimal for the generator to set the capacity at this limit, and the price at the minimum price at which the market would clear this volume.
4. The same conclusion for a retail supplier market that is competitive in terms of the description.
5. In the monopoly case, the price cap should remain lower than the foreign price.

So, in summary, for a high VOLL, it is optimal in this model for the regulator to introduce a capacity obligation.

This begs the question of why an intervention should be beneficial—that is, why not let producers and consumers make their own arrangements for capacity. Indeed if electricity were a private good under all circumstances, then there would be no benefit to force consumers to do what they would anyway. The benefit of the obligation is a direct result of the public goods nature of electricity in conditions of shortage.

The model made a number of assumptions that may be important. We examine this now to see whether their relaxation alters the conclusions.

They are

1. price cap
2. producer welfare weighting
3. interconnector constraint
4. domestic transmission constraint
5. transmission costs
6. power station failure
7. fixed costs
8. limit on demand
9. demand management
10. rationing
11. infinite capacity at fixed price in the foreign market
12. VOLL
13. retail competition
14. many periods
15. constant willingness to pay.

7.2.1.3.i Price Cap

We have shown that the reason that it is indeed optimal for a regulator to intervene in the partially regulated bilateral arrangements between a willing buyer and willing seller is that the capacity obligation facilitates a partial undoing of a prior restriction by the regulator on bilateral engagements—namely the imposition of a price cap.

The price cap limits the generator revenue. This limit is circumvented by providing a capacity payment to the generator.

In almost all markets with price caps, there are various mechanisms that circumvent the cap, generally in the reserve markets.

7.2.1.3.ii Producer Welfare Weighting

The use of α to weigh welfare is a useful construct that is important in practice. Since any deviation from $\alpha = 1$ essentially violates the market paradigm (as all producers are consumers and vice versa), any modeling of $\alpha < 1$ should be examined carefully, for example, rational behavior by generators.

We showed that for $\alpha = 1$, if VOLL is less than the foreign price, there should be no capacity obligation. While we would indeed expect VOLL to be far above any normal market price, the treatment of the foreign market here is essentially that of reserve power. Indeed we also saw that for $\alpha = 1$ the expectation of saving from having a capacity obligation is simply equal to the payoff to the consumer of being able to import volumes above the nonexportable capacity, rather than lose load.

7.2.1.3.iii Interconnector Constraint

We can view the interconnector in two ways. First, we can assume that any capacity build applies equally to flow in either direction. At the other extreme, we have to build one for import and one for export. While import and export constraints tend to be similar, we do need to consider who pays for the interconnector. Since the interconnector is heavily used for export to the foreign market, the cost should be borne by the domestic generator and the foreign consumer. We can view the Panzar or Steiner analysis to take a view of whether the domestic consumer should pay a contribution to interconnector capacity cost or whether it is a free public good for him.

In this model, since the foreign price is high, we would in fact expect both generator and interconnector build to increase until the point that the generator only provided power to the domestic consumer when the domestic price reaches the foreign price.

If the interconnector is constrained to zero, we have an autarky.

Partial constraint is covered by CF. We are interesting in the sensitivity to the constraint.

The greater K the greater the domestic generator will export. The sensitivity of the benefit from the obligation, under monopoly conditions, is

$$\frac{\partial}{\partial K} \left\{ \left[[v - P] - [f - P] \frac{1 - \alpha(K - \beta)}{1 - (K - \beta)} \right] \int_{K - \beta}^1 (D - (K - \beta)) dG(D) \right\}.$$

For $\alpha = 1$ we have

$$[v - f] \frac{\partial}{\partial K} \int_{K-\beta}^1 (D - (K - \beta)) dG(D).$$

We can see that this is a version of our option formula.

7.2.1.3.iv Domestic Transmission Constraint

If the domestic transmission is completely constrained, then all power must be imported, and hence the domestic price is always f .

The depression of the domestic price below the foreign price relies on the interconnector constraint and indeed we would expect interconnector capacity to expand until foreign and domestic prices converge.

The domestic generator, domestic consumer, and foreign importer all rely on the domestic transmission.

We would expect domestic transmission to get built up to the capacity of the interconnector.

7.2.1.3.v Transmission Costs

Broadly speaking we can treat transmission as having zero variable cost. We can take a peak load approach to transmission costs. This can be done with the standard Lagrangean method. The way to calculate the allocation of costs is first to discretize the probability domains as seen in figures 8.5–8.7 and then to use load duration duality as described in section 2.4.1. We can then apply peak load pricing.

7.2.1.3.vi Power Station Failure

This can be taken into account by assuming that each power station has an individual probability of complete failure of λ_i . In the simplest case we assume no correlation between the failures of any power stations or between failure and demand. For an infinite number of very small power stations, the aggregate failure distribution is normal. To model the pricing we can use load duration duality and model two subperiods, one with the power station failed and one unfailed.

7.2.1.3.vii Fixed Costs

CF set fixed costs of the generator to zero, but at the same time create a proxy for fixed costs by considering the export earnings potential of the generator. If we take as a base case the situation in which the generator is built and gains an export revenue exactly equal to fixed plus

variable costs, then $f = B + b$. If the generator stops exporting then it loses $B = f - b$. We can with no loss of generality set variable costs $b = 0$ and in this instance $B = f$.

Suppose that we built the power station before the interconnector, and with the domestic market in mind. Then $B = b + B$. If we have a technology frontier with B and b as the axes, then we will choose an optimal technology mix. If we assign a cost of risk to the producer, this will drive the built mix toward the higher fixed cost end of the frontier.

7.2.1.3.viii Limit on Demand

We have seen that the calculus of the CF model can relatively easily accommodate an increase in maximum demand and for this to tend to infinity. It is the physical model and cost assumptions that start to break down.

Given that there is a single willingness to pay, we do need to worry about degrees of loss of load. We can simply assume that there is a conditional amount of lost load, with a probability λ . This was explored in the Chao analysis in section 3.11.

The way to analyze this is to assume that demand is distributed as we have described, but instead of having a cumulative probability of 1 for some level of demand $K + \beta$, the probability of demand being below $K + \beta$ is $(1 - \lambda)$. The demand, conditional on it exceeding $K + \beta$, is some higher level $D\gamma$.

The combination of the limit on demand and the constant willingness to pay means that we omit the effect of steeply downward sloping demand curves. With the limit on demand there is no rationing since we build $K - \beta$ or Θ and nothing in between, so if we have some capacity build there is no rationing and hence rationing method is irrelevant. If we have no limit on demand, or a high limit, then it may be that consumers are prepared to pay for the common pool good of capacity, as was described in the simple framework in section 5.2

7.2.1.3.ix Infinite Capacity at Fixed Price in the Foreign Market

The assumption of a foreign market with infinite elasticity at price f is critical. We assumed that $v > f > P$, so the domestic price effectively flips from P to f during times of shortage, and the interconnector energy flow reverses. The foreign market is effectively providing a call option, struck at price f , with infinite volume, for zero premium. A national operator in the foreign market would instead offer to the domestic market at v rather than f , even if he truly has the infinite

volume option struck at f , and hence we can approximately value the option at $(v - f)$ times the expectation of reverse flow. To a degree, we can imagine the existence of an option struck at f , if foreign operators have more access to these options than exist in the domestic market, and there are many competing neighbor countries, or full competition in one country with no national regulation. Finally, we have assumed in the reversal of flow in conditions of tightness that the foreign system does not experience the same tightness from the same causes. Either way, our model is so heavily dependent on the foreign system that we should not regard it as self-contained, and we should recognize significant physical and political heterogeneity in the b_i or multinational complex. To avoid this heterogeneity requires us to set $f = P$, in which the foreign system becomes part of the domestic system. Alternatively we create an autarky by setting $\beta = 0$.

Let us then consider the autarky ($\beta = 0$). First, let us take the generation capacity K as given and assume that maximum demand = K . Let us initially assume a highly simplified demand structure such that demand is zero with probability $(1 - \lambda)$ and is K with a probability λ . If the consumer is forced by capacity if offered, then the generator profit is $\pi = -\lambda bK - \beta K + \lambda PK + CK$. We can see by inspection that we will wish to build K if $\lambda(P - b) > C + \beta$.

We can now model the foreign market by considering the autarky and adding a generator with zero fixed costs and variable costs of f . Clearly, with zero fixed costs, the generator will build as much as possible. Pursuing this analysis we should then add a fixed cost to the generator, and then make the build decision as described in the previous paragraph. This tells us how much capacity payment we should really pay to the foreign market.

Alternatively, the foreign market could charge f for exports and some price less than f for its own domestic trading. In effect this simply becomes a method of circumventing the domestic price cap.

7.2.1.3.x Value of Lost Load

It is obvious that if $v = 0$ then no capacity is required and if $v \rightarrow \infty$ then the capacity requirement must be the maximum possible demand.

With the right-angled demand function, if the shock to demand is lognormal then this is very convenient analytically, since we arrive at a version of our option formula, transposing the payoff and the volume.

This makes modeling of lost load more straightforward.

A downward sloping demand function is more problematic in this framework, and we must then use some of the more sophisticated frameworks such as those of Crew and Kleindorfer, or Chao.

7.2.1.3.xi Retail Competition

If we add retail competition then we must consider consumer switching. We must then apply the capacity obligation to market share. To apply a deficiency penalty, we must either define market share according to actual flow or anticipated flow. If we define according to actual flow, then our system is not really a capacity obligation but a cashout at VOLL. If we define according to anticipated flow, then we have the practical problem of calculating this in the light of ongoing switching, of possible gaming of the system to increase flow after the capacity auctions.¹

7.2.2 Modeling Capacity Obligations in a Simple Networked System

At this point, we have considered a simple system with a single node and a uniform generation fleet. Before going into more detail with this model, we briefly examine how adding system complexity, in particular the consideration of transportation, changes our analysis. Here we consider a simple system with generators and a transmission system constraint. We use the formalism of Cremer, Gasmi, and Laffont (CGL) as a basis. As the authors point out, the approach uses a combination of location theory and equilibrium economics in the “economics of spatial equilibrium” as well as operations research and computational economics that developed the field of transportation economics. It is particularly useful for us because the work of Laffont in industrial economics is very important.

We now have six entities, one consumer, two producers, and three transporters. The transporters effectively convert production at one point to production at a different point. This is the equivalent of converting production in one period to production in another. We ignore capacity limits for the producers but consider them for the transporters. We borrow extensively from the analytics of the 2003 paper, commenting as we go on elements that are particularly important for the discussion in hand.

The paper is posed as a gas transportation problem. We pose this as an electricity capacity problem. Gas is driven by pressure differentials and direct current is driven by voltage differentials, and we can interpret Kirchhoff’s voltage law as having a single pressure at any node

and Kirchhoff's current law as having zero total net flow to any node. We can visualize the equivalence between gas and power as shown in figure 7.12. The power network has no resistance except for a resistor in the middle of each line. Power is lost through the resistors and power flow is determined by the relative resistances. The gas network is a frictionless pipe with an aperture in the middle of each pipe. Gas is lost at the apertures by using gas to drive motors to compress the gas to drive it through the pipes. Gas flow is determined by the aperture sizes.

The system set up is as follows:

1. There is one producer at each of two nodes, and one consumer at the third node.
2. The producers have zero (or ignored) fixed costs and (different) variable costs. Returns to scale on variable costs are constant.
3. Generation offers at variable costs, either because it is forced to by regulation (the Hotelling model) or because the prevailing many-player Cournot game is to offer at variable costs.
4. The network is unconstrained except between the consumer and producer 2.
5. The network has both fixed and variable costs.
6. The social planner can set all prices, dispatch volumes, and has sight of all cost functions.
7. Demand is elastic and deterministic.
8. We initially examine a single period.

Figure 7.13 shows a possible representation of the system.

Producers 1 and 2 have variable cost functions C_1 and C_2 respectively, which we state as inverse supply functions $p_1(q_1)$ and $p_2(q_2)$ respectively for outputs q_1 and q_2 .

The inverse demand function is denoted as $p_3(d_3)$ where d_3 is the demand.

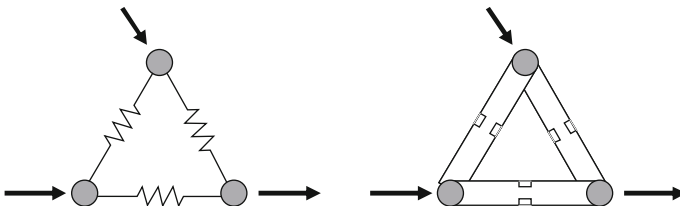


Figure 7.12 Visualization of the gas network as equivalent to a power network.

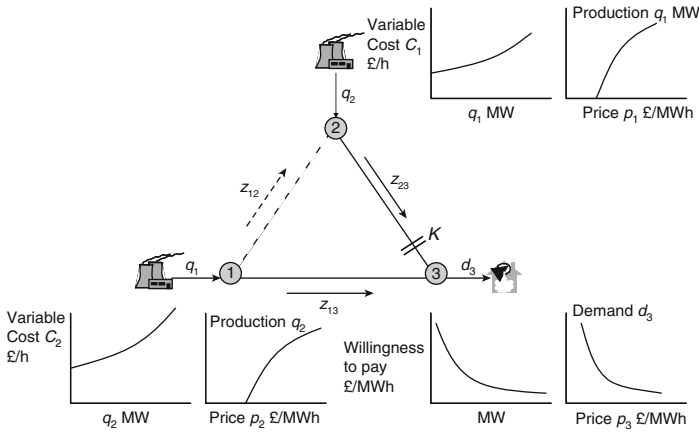


Figure 7.13 Summary of the system as defined by Cremer, Gasmı, and Laffont cast as electricity rather than gas.

The marginal cost for the network owner operator to get electricity from i to j is $C_{ij}(z_{ij}, l_{ij})$, where z_{ij} is the flow and l_{ij} is some other cost determinant that we notionally regard as network length,² so we assume that $C_{ij}(z_{ij}, l_{ij}) = c_{ij}(z_{ij})l_{ij}$.

The cost and revenue functions are shown in figure 7.14.

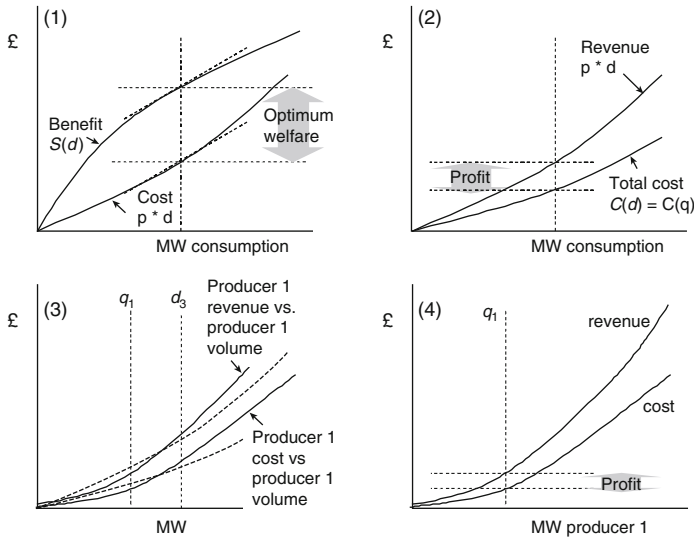


Figure 7.14 Cost and revenue functions (1) Consumer welfare (2) Cost and revenue for single producer (3) Cost and revenue for producer 1 (with lower costs than producer 2 for lower loads) (4) Profit for producer 1 if both producers available.

7.2.2.1.i The Construction of the Aggregate Welfare Function

In the first instance, let us assume that network capacity (i.e., the cost/volume relationship) is given.

Our net social welfare SW is then the sum of the surpluses for the producers, the network owner-operator, and the consumer.

$$\begin{aligned}
 SW_3 &= S(d_3) - p_3(d_3)d_3 && \text{---consumer, (7.2)} \\
 SW_1 &= p_1(q_1)q_1 - C_1(q_1) && \text{---producer 1,} \\
 SW_2 &= p_2(q_2)q_2 - C_2(q_2) && \text{---producer 2,} \\
 SW_n &= p_3(d_3)d_3 - p_1(q_1)q_1 - p_2(q_2)q_2 - c_{12}(z_{13})l_{13} \\
 &\quad - c_{21}(z_{21})l_{21} - c_{23}(z_{23})l_{23} - H
 \end{aligned}$$

The term $p_3(d_3)d_3 - p_1(q_1)q_1 - p_2(q_2)q_2$ is the network rent, which for zero losses, we can write $p_3(d_3)(q_1 + q_2) - p_1(q_1)q_1 - p_2(q_2)q_2$. We can see that the network rent is constructed from the nodal prices.

Note that the linear relationship between money paid and social welfare, implies risk neutrality for the consumer.

Here we assume that the network operator n acts economically by buying electricity at one node, selling at another, and honoring the contract by transportation³. H is the fixed inescapable cost of the network owner operator. Note that the units of this are in £/period and not £/MW/period. Note also that we are treating the generator and network capacity costs quite differently in recognizing the network fixed costs and not the generator fixed costs.

The cost structure is shown in figure 7.15.

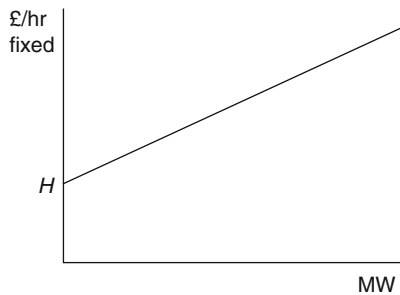


Figure 7.15 Network cost structure envisaged by Cremer, Gasmi, and Laffont.

Now let us assume that the line linking nodes 2 and 3 has a maximum capacity of K . We shall later consider how we determine the optimum capacity K .

Let us briefly look at the structure of H . Let us first generalize it slightly in a manner consistent with CGL. $H \equiv H(K) + 'H$. Note that $'H$ exists as a constant in all states of the world as defined. It is therefore a lump sum that is exogenously determined and is entirely unrelated to build or running costs. This is in accordance with the Ramsey framework, and should not be applied to consumption in an optimizing economy.

We know from the conservation of flow that $q_2 = z_{21} + z_{23}$, $z_{21} + q_2 = z_{13}$, $z_{13} + z_{23} = d_3$ and $q_1 + q_2 = d_3$.

The social planner must then solve:

$$\begin{aligned} \max(z_{13}, z_{23}, z_{21})SW = & S(z_{13} + z_{23}) - C_1(z_{13} - z_{21}) - C_2(z_{21} + z_{23}) \\ & - c_{13}(z_{13})l_{13} - c_{21}(z_{21})l_{21} - c_{23}(z_{23})l_{23} - H \end{aligned}$$

subject to $z_{23} \leq K$.

CGL use the usual Lagrangean method to solve for optimum pricing of the constrained link. Noting the equation of price to the differential of surplus and the pricing of transmission as the difference of price between two nodes, CGL show that the optimum pricing for transmission for the unconstrained link is equal to the variable costs, and for the constrained link, there is an additional cost. This is consistent with the analysis in the rest of this book, and the solution for pricing for three constraints can follow the Carlton method described in section 3.8. While this is quite consistent with Ramsey pricing, we have shown that for a (Walrasian) balancing economy, there should be no volume-independent term in H . By taking the simplest case of $H(K) \equiv H_K * K$, we can in this one-period deterministic setting load this cost into the variable cost c , and the Ramsey term disappears.

So far so good. Where the CGL analysis is useful for the matter in hand is in the implied conclusion that in the deterministic case, the decentralized market correctly optimizes and hence there is no need for regulatory intervention in prices or volumes. CGL state that "The optimal allocation can be decentralized with transportation charges equal to the (short-run) marginal cost of transportation plus a Ramsey term (if any), supplemented by an ex ante sale of capacities priced at marginal cost, followed by competitive secondary markets."

7.2.2.1.ii Modeling the Decentralized Economy with Decentralized Transmission

Let us take the simplest cost structures that we can, while being sufficiently general to attack the nub of the problem—the efficacy of decentralization to maximize aggregate welfare. Since we have a deterministic one-period setting, with inelastic demand and fragmented production, there is no separate identity of fixed and variable costs. To allow for different cost structures, we do need to consider the “fixed heat” required for the first MW of load delivered. The simplest formulation is shown in figure 7.16.

Let us suppose that we have direct rather than alternating current, that the three major lines 12, 23, 13 have equal resistance.

Let us initially assume that the line resistances represent an insignificant efficiency cost. Since there is an efficiency/capital cost tradeoff, we simply imagine all lines to be short.

In the solution above, we effectively assumed that the transmission system was made up of large number of parallel lines, so that we can control the flow using switches. Let us now consider the reality of an electrical transmission system.

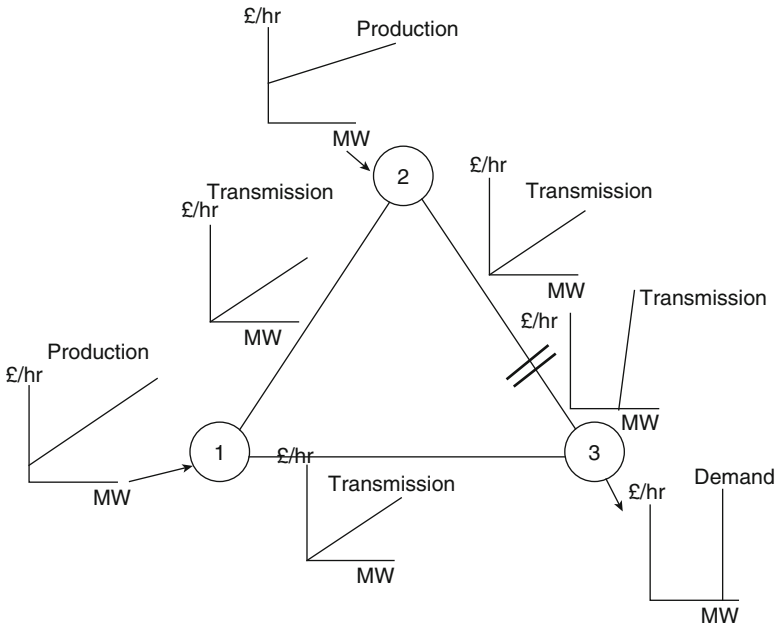


Figure 7.16 Simplest formulation for a three-node constrained electrical system. The transmission line cost slopes are near zero in the calculations.

We must of course remember Kirchhoff's laws. In this instance, once constrained, an increment of demand requires not only an increase in production from unit 1 but a decrease in production from unit 2.

The progression of optimal loading as demand at point 3 rises is shown in figure 7.17. The rationale is as follows:

First (build and) load the cheapest unit at low load. At the point of crossover of unit cost curves, we switch from unit 1 to unit 2. We increase the load on unit 2 until flow on line 23 hits the constraint. At this point, for every MW of load, we must add 2 MW from unit 1 and deload⁴ unit 2 by 1 MW. This we do until the unit 2 loading is zero.⁵ In the first instance, let us suppose that the line constraint is absolute, and cannot be resolved at any cost.

Let us change the framework slightly to see if the decentralized market will optimize the complex. We can divide the units, so that even under a Cournot game, they do not offer above costs. Conceptually, we can do the same with transmission lines, dividing them into parallel lines. Let us assume constant returns to scale in commodity and capacity, zero variable costs at zero and near-zero loads, and a downward sloping demand function.

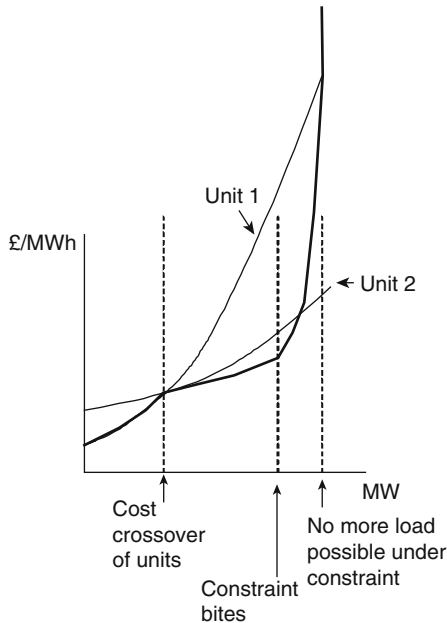


Figure 7.17 Cost structure for gradually increasing demand.

Our question then is whether Pareto optimization would arrive at the optimum solution. Let us consider how a decentralized solution would gradually change the nodal prices as demand increases. The first increments of load are straightforward. Unit 1 undercuts unit 2, and then unit 2 undercuts unit 1.

Let us turn to the decentralization of nodal pricing. This works by buying and selling nodes, and “financial transmission” occurs by buying power at one node from the system operator and selling it at an adjacent node. We now simplify power generation costs as being infinitely elastic at the nodes at variable cost b_1 and b_2 .

Given our constant returns to scale in generation, zero fixed costs, and no generation capacity constraint, for the volume range in which the line constraint bites, the prices at nodes 1, 2, and 3, must be b_1 , b_2 , and $2b_1 - b_2$ respectively.

The required reconfiguration of power station load to relieve network constraint is described in section 7.2.3.3. Here we note briefly that if line 23 exceeds the constraint level by amount δ , we must reduce the load of high merit (cheaper) unit 2 by δ and increase the load of low merit unit 1 by 2δ . The marginal cost, and thence the price at demand node 3 is then $2b_1 - b_2$.

With a constraint on line 23 of z_{23} , we have three load regimes,

$$\frac{2}{3}Q < z_{23} \quad p_3 = b_2$$

$$z_{23} < \frac{2}{3}Q < 2z_{23} < p_3 = 2b_1 - b_2$$

$$\frac{2}{3}Q > 2z_{23} \quad \text{inadmissible due to constraint violation.}$$

For load $Q = \frac{3}{2}z_{23} + \delta$ where $\delta \rightarrow 0$;

The total consumer cost is $Q(2b_1 - b_2)$.

The total producer cost is Qb_2 .

Hence the available rent to the transmission system is $2Q(b_1 - b_2)$.

If line 23 is given 1 MW more capacity, then (assuming as before equal resistance/impedance on each line), unit 2 can increase by 3 MW and unit 1 decrease by 3 MW. We use the principle of superposition to add the flows from node 1 to/from the load and the flows from node 3 to/from the load. This reduces the cost to deliver the same amount of energy to the consumer by $3(b_1 - b_2)$. The changes in line rents are $-2(b_1 - b_2)$, $-(b_1 - b_2)$, and $+(2b_1 - 2b_2)$ on 12, 13, and 23 respectively. The sum of these is an aggregate fall of $(b_1 - b_2)$ but line 23 expansion is funded.

7.2.2.1.iii Self-Dispatch into a Passive Network

Here we assume that there are no switches in the network and that load is controlled only by pricing signals at the nodes.

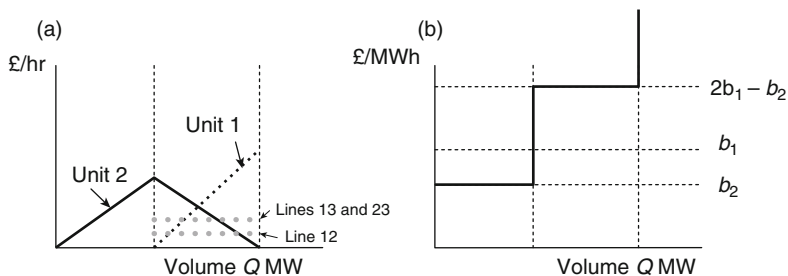


Figure 7.18 (a) Generator cost and line rents as network load increases (b) Price at the demand node.

In a decentralized market, in the region where line 23 is constrained but no constraints are violated, we would expect rents for lines, 12, 13, and 23 to be $\frac{1}{3}Q(b_1 - b_2)$, $\frac{2}{3}Q(2b_1 - b_2 - b_1) = \frac{2}{3}Q(b_1 - b_2)$, and $\frac{1}{3}Q(2b_1 - b_2 - b_2) = \frac{2}{3}Q(b_1 - b_2)$ respectively. This adds up to $\frac{5}{3}Q(b_1 - b_2)$, which exceeds the available rent.

In the absence of nodal pricing or switches, unit deliveries would increase until the constraint was exceeded and line 23 burnt out⁶; it would continue to increase until lines 12 and 13 burn out. The order depends on their respective capacities and resistances.

While generators can affect prices at loads through the effects of changing their offers at the margin, networks can affect the difference in prices of adjacent nodes. The question is whether the simple expedient of price signals can drive the most efficient generation, flow, and demand satisfaction.

The particular challenge here is that the whole network has to be modeled in one go. For example, the appearance of a constraint a long way away in the network can have a significant difference on local flows. This is called “loop flow.” The location marginal pricing model, described briefly in section 7.2.3.3 explains this. From this, it is clear that in a passive network the so-called loop flow acts as an externality to the commercial driver to get load from one node to another.

In summary—while price differences between local nodes can indeed be influenced by the transmission line between the nodes, and the transmission build between nodes can indeed be influenced by the nodal price differences, it is very hard if not impossible for nodal prices to develop purely organically, with generator and line build driven by these prices. In practice, a system operator is needed to construct the nodal pricing system across the whole network.

As a brief aside, it is worth mentioning that private lines do get built.⁷ However these are single entities that can to all intents and purposes be regarded as interconnectors. In commercial terms, loop flow along parallel routes to the interconnector are ignored and thence the net system saving from interconnection arrives at the same interconnector value as treating each zone as a node with one price.

The capacity of transmission lines, financial transmission rights, and location marginal pricing, are subjects very closely related to our subject here, but they are out of scope for further analysis in this book, and we predominantly model on the assumption of an unconstrained zero-cost network.

7.2.2.1.iv Risk Aversion and Stochastic Load

CGL introduce consumer risk aversion by the utility function $\frac{\partial^2 U(SW)}{\partial q^2} < 0$ and used a linear rationing cost. The importance of the structure of the utility function and rationing cost were considered in detail with the Chao framework analysis in section 3.11.

CGL then show that in their framework risk aversion increases optimum build. While on the basis of comparative statics, if we change nothing else but consumer risk aversion, then this is true. However, we should not forget that i) there is no reason to suppose that the normalized risk aversion of producers is less than that of consumers and ii) increased risk (as distinct to increased risk aversion) could make optimum capacity higher or lower than the deterministic case.

The assertion that risk aversion on the part of the consumer increases optimum build is correct within the CGL framework but cannot be generalized without some care. If the net (i.e., after variable costs) VOLL exceeds the cost of capacity (as is true in the CGL framework), and producer cost of risk is zero, then indeed, consumer risk aversion increases the optimum build.

7.2.2.2 Conclusions of the Network Model

For the purposes of considering capacity payments, the analysis of single interconnectors is very similar to that of generating units. The essential difference is that networks have high fixed costs and low variable costs.

However, where networks are complex enough to have loop flow, decentralized scheduling does not work, and hence capacity payments cannot be applied to individual lines.

7.2.3 Peak Load Pricing in Transmission Charging

Transmission networks, and high voltage direct current (HVDC) interconnectors in particular have very low variable costs in relation to the construction costs. In modeling terms it is reasonable to model them as having zero variable costs.

The recognition of the need to recover fixed costs therefore represents a useful precedent. As is obvious from section 7.2.3.3, the flow through transmission networks generally has to be managed centrally and only in specific circumstances can the primary influence on the nodal price differentials at either end of the line be by the line owner. An example would be a single interconnector between two isolated systems.

The triad charging system in Great Britain, depicted in figure 3.4 and noted in section 3.2.3 is a form of peak load pricing for transmission.

7.2.3.1 *Consideration of Incremental Load*

If we are to charge peak load pricing for transmission, we need to consider who to charge. There are three relevant theories to use.

The first is Kirchhoff's laws. These are uncontroversial and straightforward to understand. Kirchhoff's voltage law (KVL) says that the voltage at a node must be equal to the sum of voltages applied to it and the current law says that the current through a node must be the sum of currents flowing through it.

The second is the principle of superposition, which is a representation of Kirchhoff's laws, saying that the total voltage and current in a system can be represented as the sum of voltage and current of flows that observe Kirchhoff's laws. This is shown in figure 7.19. We see that figure 7.19(c) can be viewed as the sum of (a) and (b)

This is done with direct current load flow (DCLF) models for the "active" power that is used by consumers. Given the importance of reactive power in security of supply incidents, it is important for system operation to use alternating current load flow (ACLF) models to run the system. This is complicated for the use of pricing, not least because the ratio of capacitance to impedance of transmission lines decreases as load increases.

So far so good, but suppose that all we can see is figure 7.19(c) and we need to allocate the constraint cost on one (or both) of the two lines bearing a load of 3 GW. In the absence of electrical losses or regulatory constraints on charging a specific percentage to production and generation, allocation between the generation and demand

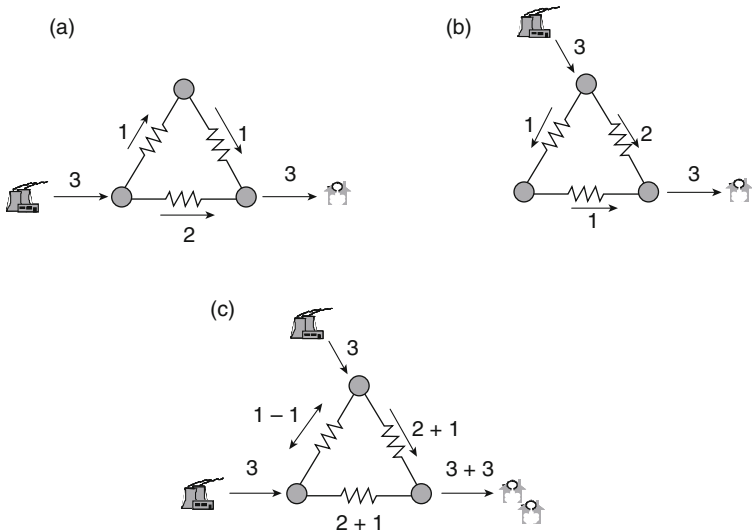


Figure 7.19 Application of the principle of superposition to a simple network $A + B = C$.

sectors is simple. Since 1 GW of demand requires 1 GW of production, we can charge production and demand equally.

The third principle is the principle of additionality.

To allocate cost within the production or demand sectors, we need to rely on the principle of additionality. This principle assumes that we can chronologically rank the individual flows. In figure 7.19 this means that if we first built and ran the power station in A, then all of the constraint costs should be charged to B.

Consider however some scenarios:

1. Power complex B was planned a long time in advance and committed to paying for the transmission. Power complex A arrived quicker and had no contracts.
2. Power complex A has a certain demand and production of 3 GW and power complex B has a production demand of 3 GW plus or minus 1 MW for half the time each.
3. Power complex B has a certain demand and production of 3 GW and power complex A has a production demand of 3 GW plus or minus 1 MW for half the time each.
4. Transmission was built under contract with complex B but without contract for complex A, with A paying live for transmission.
5. There are no contracts and complex B reduces its load relative to its original plan in order to deconstraint all transmission lines.

What becomes obvious here is that what counts in charging for transmission is actual flow and actual commitment and intent. We apply the electrical flow using the principle of superposition not according to the arrival chronology of production and demand but according to commitment first and intent second.

Suppose now that we are building a distribution system in an area of growing population. Networks cannot easily be expanded incrementally and have to be done in large chunks in intervals that are commonly decades long. We size the “wires” to accommodate the anticipated growth and have a distribution charging profile over time that recovers the cost in a manner that recognizes that some but not all of the load growth is caused by the existing population on the network. A newly built factory then argues that the spare network capacity should be allocated to it for free, and when the line constrains following population growth, it will pay some cost then.

Before moving on, it is worth noting that the cost of electrical losses adds an interesting extra dimension to this problem. Because the losses are not linearly proportional to power flow but the square of the flow, the allocation of losses is an interesting problem that can use the methods above.

7.2.3.2 *Charging according to Peak Load*

If we consider a time period of, say, a year, in which there is no effect of long-term trend of production and consumption change, then the allocation of cost has three elements to it.

First, we use load factor duality with a discretization of probability of events to m events in each of the m subperiods (commonly half-hourly), to construct a single deterministic load duration function for the year, with $m * n$ elements.

Second, we consider the events for which constraint is exceeded. For each of these we consider two things i) reschedule power stations, as we see in section 7.2.3.3 and ii) drop load voluntarily or by the distribution system operator. The constraints are now resolved.

Third, we apply both a fixed and variable cost to each deconstraint scenario.

Now we can charge system users using both extremes of method.

If all charging is fixed then we charge users ex ante fixed costs (irrespective of actual outturn) according to their contribution to the constrained scenarios. The Panzar method can be used for this.

If all charging is variable then we charge users on arrival at each constraint scenario using peak load pricing.

In practice there can be a mixture.

A good example is the triad system in Great Britain. The total cost recovered by the transmission network is set in advance. The share of this cost is allocated according to the average market share across the actual three peaks (the halfhourly peak in its peak, which is separated by a fortnight).

A good example of live pricing of constraint is the location marginal pricing described in section 7.2.3.3.

In practice the regulatory culture tends to be one of socialization of prices and avoidance of peak load pricing. Hence much of network pricing i) is charged per MWh over the year rather than actual GW peak and ii) has a total system charge over the year that is constant (“capacitized”) rather than variable (“commoditized”).

7.2.3.3 Location Marginal Pricing

Location marginal pricing is a pioneering model for pricing electricity at nodal level at high temporal resolution. It has been in operation in PJM since 1998.

In this model bids and offers are made at each node, which we can regard here as a grid supply point (GSP), and the price at the node depends on the transmission constraint. If we have a generator at cost X that is “constrained down” to resolve constraint and generator Y is dispatched instead, then the cost at the node that cause the constraint is not Y but $Y + (Y - X)$.

We can see this in figure 7.20. If the bottom line is constrained to 2 GW, then to satisfy a 1 GW increase in demand we have to load unit B and deload unit A. The cost of the 1 GW is equal to $2 * £120 + 2 * 100 - 3 * £100 = £140/\text{MWh}$.

Note that even if only one watt of load is added under conditions of constraint, the marginal price at the demand node is still $£140/\text{MWh}$.

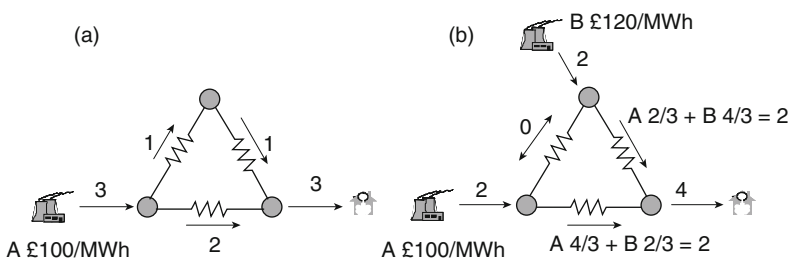


Figure 7.20 Location marginal pricing. Increasing demand under conditions of constraint can cause a deload of the cheaper unit.

We can see from the above that the total cost for the 4 MWh is £440 and the total revenue is £560/MWh.

7.3 PEAK LOAD PRICING IN DISTRIBUTION CHARGING

Increasingly the distribution networks are likely to become constrained as electrification of heat and transport take hold. A challenge in distribution is that upgrade of capacity is expensive and cannot be done in increments. The infrastructure can be in place for at least 50 years.

If the population of individuals over the full period of investment and use is unchanged and the interest costs to the people, the state and the network company are the same, then allocating the cost is straightforward. We simply form a load duration function of n periods per year times y years times m stochastic states and then apply peak load pricing.

However there are numerous practical problems:

1. emigration from the area
2. immigration to the area
3. net population change
4. moral hazard that the regulator will abrogate commit to refund investment through network charges
5. moral hazard that the network company really has lower costs of capital or operation than recognized by the regulator
6. the incentive of the network to “gold plate” the assets, for reasons of capital advantage [the Averch and Johnson (1962) effect], operational advantage or avoidance of opprobrium on lost load
7. the risk of stranded assets, that is, assets that turn out not to be required and hence no consumers to fund them.

For the reasons above we need to charge consumers in the early rather than load period of use. The location marginal pricing model is useful in this regard, not so much in relation to the solving the economics of loop flow but for the principal of live pricing of network constraint. For example, users could pay according to their incremental effect on the cost of reinforcement.

7.4 COMMERCIAL ARRANGEMENTS ON LOST LOAD

Politics aside, a mature electrical system should allow for lost load and have a nonpunitive level of compensation.

We have concentrated mainly on the relationship between producer, consumer, supplier, and system/market operator/regulator.

We can now consider the wider commercial arrangements on lost load. In particular it is important to recognize that a network failure in the peak can deny the generator essential revenue.

7.4.1 Generator Inadequacy in an Integral Network

If the supplier has a short imbalance position, then they must pay $P_{\text{imbalvollcap}}$ on the excess load provided and VOLL to consumers for whom the loss of load is out of keeping with the contract.

If the supplier is in balance and load is lost, then he must pay his consumers VOLL and this must come from the system operator, as the PN purchased from generators are not really physical, as described in section 5.1.3. The system operator must collect the money. There are here several scenarios.

1. Suppliers bought capacity certificates and the total demand exceed the total certificates.
2. Suppliers bought capacity certificates and the total demand did not exceed the total certificates but generators failed.
3. No capacity certificates, and the total demand exceeded the total physical notifications by suppliers.
4. No capacity certificates, and the total demand did not exceed the total physical notifications by suppliers but generators failed.

In the first case the system operator must pay (via the market operator) the suppliers and in practice would do this by a mix of “recovery” mechanism in which he pays now but increases ongoing supplier charges to recover the loss, and an incentive mechanism in which the system operator takes part of the loss.

In the second and fourth cases the generators pay $P_{\text{imbalvollcap}}$ to the system operator who pays the suppliers.

In the third case the suppliers must pay $P_{\text{imbalvollcap}}$ on the extra power provided and VOLL to their consumers on the balance.

7.4.2 Transmission Failure

Now the generator and consumer are both denied volume. The transmission company must pay $P_{\text{imbalvollcap}}$ to both generator and supplier. Indeed it may be that they have to pay VOLL to the supplier.

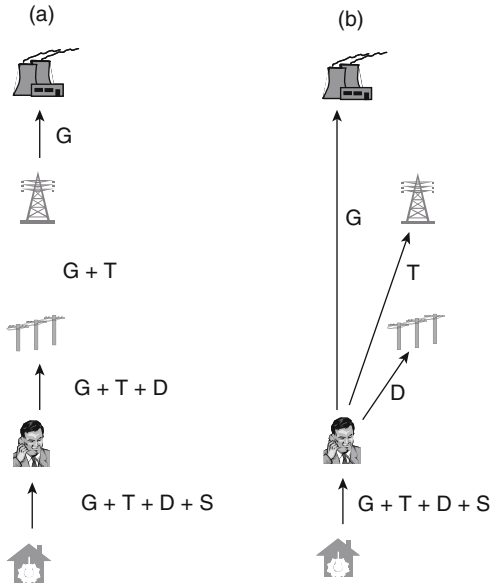


Figure 7.21 The point-to-point and supplier hub market models (a) Point-to-point showing generation, transmission, distribution, supply, and consumption (b) Supplier hub.

In practice we would expect generators to have balancing offers, and hence transmission failure would result in the denied unit being sold power at its bid in balancing and a unit that has been called on, if there is one, paid its balancing offer.

A transmission failure would in general be highly correlated to high demand.

When modeling arrangements between generation, transmission, and distribution, it is commonly helpful first to consider these using the point-to-point business model shown in figure 7.21 and then map to the supplier hub if required.

We work out the compensation economics using the point-to-point model, and then the actual payment path through the supplier hub model.

7.4.3 Distribution Failure

A distribution failure is generally regarded as a local and not a system event.

In theory the distribution system operator (DSO) should pay the generators (via the market operator) for lost opportunity, although this does not happen in practice.

The DSO should pay the suppliers the VOLL to go to consumers.

Distribution failures are in general lowly correlated to system demand, although there may in future be a stronger relationship between local loading and local failure.

FINAL COMMENTS

There remains a debate about the relative merits of peak load pricing, in which the covering of fixed costs is specifically recognized as a required uplift of price above variable costs, and variable cost pricing, which advocates market clearing at variable costs, in order to maximize economic efficiency at the margin.

While the rhetoric of variable cost pricing remains dominant, both in the literature and in the formation of policy, we found by following the development of the canon that in fact the two alternatives converge under equilibrium conditions in which there have been some discretionary choices, for example, in technology selection and the degree of investment in maintenance.

This conclusion is robust with respect to the alteration of a number of modeling variables, such as divisibility in time of the pricing period, divisibility of asset size, the form of the demand function and the shocks to it, varying plant reliability, and other features.

We find that the status of public goods is important for electricity in many dimensions, including physical capability, social requirements, and the general structure of metering, billing, and payment for energy. While electricity delivered can become a private good, and may become so in the “smart” system, there is little evidence of this at this point. As a result, it is important to be able to model the degree of rationing efficiency in conditions of scarcity, in order to optimize the system.

We then showed how power stations can be represented as financial options, first, in the most simple European call option, but also with much more complex modeling, for example, flexibilities in load level and other types, and constraints such as environmental limits. In addition, the whole system can be modeled as a family of options.

We showed that capacity obligations can be represented in terms of call options, with extra features such as nonfirmness, mandatory surrender of part of the peak energy rent value in terms of a call spread.

We also showed that while the ICAP installed capacity model begins very crudely as an option with no strike or value to the buyer and an option with no cost to the seller (even on plant failure), the continuous refinement of the mechanism introduces more and more market disciplines. The logical conclusion, albeit somewhat distant at present, is for a mature energy-only market with liquid forward and option trading to develop.

Finally we model capacity on a wider landscape by considering consumers, complex transmission networks, and distribution systems. We find that at this point at least, it seems that network management, including network capacity, is a matter of a central system operator, market operator, and capacity planner.

Much market design remains constrained by moral hazard. On the part of the generators there is the fear that excess rent will be extracted through market power. On the part of the regulator and government there is the fear that fair rent from infrequent or other events will be expropriated. There is in addition a related risk that while the most efficient capacity obligation may be simply an ex post charge for imbalance, the same system/market event can cause both the requirement to pay and the inability to pay. Credit risk is therefore a key feature that drives capacity obligation models.

Cost of risk has a significant effect on the models shown. Separately, and in addition, there is the cost of uncertainty, such as in moral hazard. This is a large subject worthy of further exploration.

NOTES

1 INTRODUCTION

1. See Ekelund and Hébert (1999).
2. For the early period, see, for example, Baumol and Bradford (1970), Buchanan (1966), Clemens (1964), and for the later period, Joskow (1976).

2 THE MODELING FRAMEWORK

1. After Hotelling (1939).
2. An example of a workaround is the operation of Lagrangean optimization by iteration. The state change costs that are loaded into standardized variable costs can be adjusted mid-routine according to the development of the load duration curve.
3. This is described in more detail in Harris (2014).
4. See, for example, Abbot (2001).
5. See Rosellón (2005) for a Cournot analysis in relation to the offering of capacity.
6. After Tobin (1969) and Brainard and Tobin (1968).

3 THE FRAMEWORK AND DEVELOPMENT OF PEAK LOAD PRICING

1. Turvey (1968a) and Turvey and Anderson (1977). See also Turvey (1968b, 1968c, 1969, 2000).
2. See Harris (2014) for a detailed description.
3. See Yakubovich, Granovetter, and McGuire for an illuminating description of price formation in the early days.
4. From the study of early days pricing by Yakubovich, Granovetter, and McGuire (2005)
5. Consulting engineer to First Edison and subsequently professor of Electrical Engineering at King's College, London.
6. The motivation was more to optimize the respective applications of electricity and gas for different consumer segments than for price regulation.

7. The working assumption was that maximum load would be in times of fogs and therefore consumer peak demands were coincident.
8. See Eisenmenger (1921).
9. The technology was such that the measurement was over about ten minutes, rather than being instantaneous.
10. Wright (1896).
11. In 1900 examples, in modern terminology are flat rate, uniform meter rate, nonlinear meter rate, peak capacity triad rate, Wright tariff, and two-rate meter. Source: Doherty (1900).
12. Lyndon (1923).
13. In his 1892 paper, Hopkinson refers to the problem and the ideal solution, but does not propose a practical application.
14. The same method is still used today. The “triad” transmission charging method in Great Britain uses this method.
15. In England and Wales, the transmission cost paid by suppliers to the transmission operator for industrial consumer load is equal to the averaged maximum demand on the three “triad” periods of maximum system demand. As of 2014, this is under sporadic review to increase the number of periods. The French “critical periods” tariff contains features of the Barstow tariff.
16. Byatt (1963) quoting H. H. Perry in 1913 at the Manchester Section of the Institution of Electrical Engineers.
17. The width of the peak can be quite important, particularly in a system with hydrogeneration (Turvey 1968).
18. Excluding nights between 19:00 pm and 7:00 hours. In 1950, Saturday afternoons and Sundays were also excluded. Source: Meek (1963).
19. There was additionally a fuel price adjustment.
20. See Boiteux (1949, 1956a, 1956b, 1957, 1960a, 1960b).
21. The central limit theory is invoked. This requires homogeneity of consumers.
22. We assume a probability of demand falling below zero or twice the mean to be sufficiently low as not to affect the analysis. This is an important assumption when using a power law demand function.
23. “Fat tail” is a common colloquial description of a distribution that is near normal for low excursions from the mean and much more probable for high excursions than is predicted by the normal distributions.
24. For empirical application in electricity, see Weron and Simonsen (2005).
25. This is due to the wide reach of high impact low probability events. Using the language of reinsurance, the facultative commonality of claims increases with the maximum limit of the retrocession band.
26. Here we resort to the extreme value theory, rather than the approximation of the binomial distribution by the normal distribution, and hence are untroubled by high moments in the form of fat tails.
27. It later transpired that Steiner’s 1957 analysis had been anticipated by Boiteux in the appendix to his 1949 paper. While observing the proper

accreditation, in this book we use only the work by Boiteux that was known in the British academic community. We do this because the British economists arrived at essentially the same models and conclusions as Boiteux, and because the evolution of the associated ideas is accessible to sole Anglophones. Similarly, Crew and Kleindorfer (1995) note that Ault and Ekelund (1987) cite Bye (1926) as an early author. Ekelund and Hebert (1999) thence provide the path back to the French engineers.

28. Steiner is not so specific, and uses the term “operating cost.”
29. Steiner is not specific on the time denominator. However, implicit in his analysis is the fact that there are only two time periods, and they are of equal length. Since he assumes that charging β/MWh for half the time covers the fixed costs, we can imply that the capacity cost is $\frac{1}{2} \beta/\text{MWh}$ when measured in elapsed time.
30. This technique was already in use at the time. See, for example, the Samuelson (1955) application from Lindahl (1919).
31. See, for example Hirshleifer (1958).
32. For example, Weisbrod (1964).
33. Samuelson (1954,1955).
34. It is commonly quoted as if it were correct. Crew and Kleindorfer point to the problem of the discontinuity of the the optimal price as demand variance rises incrementally above zero, and propose a resolution for the error. Our exposition of the problem with the BJ result differs to that of Crew and Kleindorfer.
35. This differs from most of our analysis, where the convention is that period 1 is the peak. The BJ convention is followed for easier cross reference to the literature.
36. We can see in figure 3.25 that for the division between L1 and L2 to be at, as specified by Visscher, then we need $u = 0$.
37. The slope for $z < 0$ is the opposite of this. The derivative is discontinuous at $z = 0$.
38. The probability function for \tilde{u} is a singularity, or a Dirac delta, centered at $u = 1$.
39. CK assume only two cost dimensions to plant, fixed and variable, and inherently assume deterministic and constant demand, by ignoring cost functions that cross each other twice.
40. Dansby uses a shorthand by integrating across t not x in the denominator, and does not make the assumption that the $(m + 1)$ th unit is not installed (not being used).
41. Although the framework allows for partial failure of individual units, given the perfect divisibility, we do not need to attend to the problems of part loading.
42. The loading order is prespecified from the variable cost stack. n is determined following the resolution of the uncertainties of demand and availability.

43. We assume here a zero correlation between being called to run and available to run. In practice the correlation is high.
44. Note that Chao uses the acronym LOEP.
45. This is a common device in the calculus of traded derivatives. An excellent introduction to the principles can be found in Baxter and Rennie (1996).
46. The power factor is unrelated to the power factor used for reactive power estimation.

4 RELAXING THE HARD CAPACITY CONSTRAINT

1. There is no “fixed operational cost,” that causes a step change as operation increases from zero.
2. For simplicity, we have simplified the terminology. Here the station transformer is the only connection between station and grid, and on this station there is only one unit.
3. Panzar is unclear about whether this is average or marginal. We choose it to be average, for best consistency, with the Lagrangian constraint.
4. This is the Ruggles (1949) framework, described in Harris (2014), where fixed costs are only incurred during operation.

5 MODELING CAPACITY USING DERIVATIVES

1. For example, daily peaks, weekdays, quarterly baseload.
2. Pool models generally use the cheapest to deliver method on location, that is, all locations get the same price. In the LMP model this applies for supply points in the same node
3. A “basis” risk is the difference in value between the market reference contract (here the cheapest to deliver) and the actual physical delivery (here the grid supply point).
4. In practice they are closely related due to gaming effects.
5. See Black (1976).
6. See Harris (2006).
7. Technically the alpha is the excess return over the fair value return adjusted for volatility and correlation. Here we regard α in a simple sense ignoring these factors.
8. The retail supplier risk profile is concave on a profit versus price axis. The optimum hedge is to buy convex instruments, that is, options.
9. Called Bermudan because they are between the European and American type.
10. See in particular Clewlow and Strickland (2000), Eydeland and Wolniec (2003), and Geman (2005). The practicalities of pricing swing options are described in Harris (2006), p. 349.
11. The no arbitrage method is the foundation of derivative pricing. Here we have used the concept, such that a generation unit that has sold an option experiences no future net cash flows.

12. Whether a risk weighting is applied depends on numeraire asset and the alternative strategy. In this setting we assume the presence of a forward market, which will be risk adjusted, and provided that forward prices are used for valuation, no further adjustment for market price risk (as distinct to volatility risk and risk that is nonlinear with respect to variance) need be applied.
13. This is true for low volatilities and sufficiently true generally for the purposes of this argument
14. The breakdown of the Black Scholes formalism for a non-lognormal distribution has only a minor effect on this. In practice the volatility of volatility has a sufficiently large value bias (“kappa convexity”) and cost of risk bias that it is does need to be corrected for.

6 CAPACITY MECHANISMS

1. This is explained in Weitzman (1974).
2. See Borenstein et al. (1999), Joskow (2001), Joskow and Kahn (2001, 2002a, 2002b).

7 THE POWER COMPLEX

1. This might seem unlikely, but in fact the large industrial market is fast moving, and it is quite credible for the aggregate defined supplier capacity to be less than the aggregate anticipated energy flow by holding back on booking these contracts.
2. National cost models vary. Postage stamp pricing (see Harris 2006) is used in Denmark, Finland, Sweden, and Spain. Distance-related pricing is used for national transmission in Austria, France, Germany, Netherlands, and Spain. In Great Britain, national transmission is distance-related for system entry to the national balancing point (NBP), and distance-related (with short-haul exceptions) from the NBP. In Great Britain, the distribution charge uses sum of notional distance, with small pipes carrying a higher distance weighting per physical length.
3. This is a “commodity chain” view of the world rather than an “unbundled third party access” view of the world. However for a centrally managed system or a benign monopoly, this does not make any difference.
4. This is easily calculated in this example using Kirchoff’s laws. This is in electrical textbooks and briefly summarized in Harris (2006), pp. 66, 493.
5. Note that it can be cost-effective to dump load. This is not considered further here.
6. “Burnt out” is a simplifying metaphor. Probably the line would sag and be exposed to contact with vegetation or other objects that casuse a short circuit.
7. See, for example, Littlechild and Skerk on the building of the “fourth line” in Argentina.

REFERENCES

- Abbot, M. (2001). "Is Security of Supply a Public Good." *The Electricity Journal*, 14(7): 31–33.
- Ault, R. W., and Ekelund, R. B. Jr. (1987). "The Problem of Unnecessary Originality in Economics." *Southern Economic Journal*, 53(3): 650–651.
- Averch, H., and Johnson, L. L. (1962). "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, 52: 1052–1069.
- Baumol, W., and Bradford, D. (1970). "Optimal Departures from Marginal Cost Pricing." *American Economic Review*, 60(3): 265–283.
- Bidwell, M. (2005). "Reliability Options: A Market-Oriented Approach to Long-Term Adequacy." *The Electricity Journal*, 18(5): 11–25.
- Black, F. (1976). "The Pricing of Commodity Contracts." *Journal of Financial Economics*, 3: 167–179.
- Boiteux, M. (1949). "La Tarification des Demandes en Pointe: Application de la Théorie de la Vente au Coût Marginal." *Revue Générale de l'Electricité*, 58: 321–340; trans. as "Peak-Load Pricing." *Journal of Business*, 33: 157–179 (1960). Repr. Nelson (1964).
- . (1956a). "La vente au coût marginal" in *Revue Française de l'Energie* (December 1956)—in Nelson (1964).
- . (1956b). "Sur la Gestion des Monopoles Publics Astreint à l'Equilibre Budgétaire." *Econometrica*, 24: 22–40. Trans. as "On the Management of Public Monopolies Subject to Budgetary Constraints." *Journal of Economic Theory*, 3: 219–240.
- . (1957). "The 'Tarif Vert' of Electricité de France." *Revue Française de l'Energie*. Trans. in Nelson (1964).
- . (1960a). "Peak Load Pricing." *Journal of Business*, 33: 157–159.
- . (1960b). Originally delivered at Ecoles des Mines, appeared as "L'énergie électrique: données, problèmes et perspectives," in *Annales des Mines* (October 1960) and repr. in *Revue Française de l'Energie* Nov 1960). Trans. as "Electric Energy: Facts, Problems and Prospects" in Nelson (1964).
- Boiteux, M., and Stasi, P. (1952). "Sur la Détermination des Prix de Revient de Développement dans un Système Interconnecté de Production-Distribution." Repr. in G. Morlat et F. Bessière (Eds.) *Vingt-Cinq ans d'Economie Electrique*. Paris: Dunod, pp. 361–400.

- Borenstein, S., Bushnell, J., and Wolak, F. (1999 and August 2000). *Diagnosing Market Power in California's Deregulated Wholesale Electricity Market*. Berkeley, CA: University of California Energy Institute.
- Brainard, W. C., and Tobin, J. (1968). "Pitfalls in Financial Model Building." *American Economic Review*, 58(2): 99–122.
- Brennan, T. (2004). "Market Failures in Real-Time Metering" *Journal of Regulatory Economics*, 26(2): 119–139.
- Brown, G., and Johnson, M. B. (1969). "Public Utility Pricing and Output under Risk." *American Economic Review*, 59 (March): 119–128.
- Buchanan, J. M. (1966). "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics*, 80: 463–471.
- Bye, R. T. (1929). "Composite Demand and Joint Supply in Relation to Public Utility Rates." *Quarterly Journal of Economics*, 44(November): 44–62.
- Carlton, D. W. (1977a). "Peak Load Pricing and Stochastic Demand." *American Economic Review*, 67 (December): 1006–1010.
- . (1977b). "Uncertainty, Production Lags, and Pricing." *American Economic Review*, 67(1) (February): 244–249.
- Chao, H. (1983). "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty." *The Bell Journal of Economics*, 14(1): 179–190.
- Clemens, E. W. (1964). "Marginal Cost Pricing: A Comparison of French and American Industrial Power Rates." *Land Economics*, 40: 389–404.
- Clewlow, L., and Strickland, C. (2000). *Energy Derivatives: Pricing and Risk Management*. London: Lacima Publications.
- Cramton, P., and Stoft, S. (2005). "A Capacity Market that Makes Sense." *Electricity Journal*, 18(7): 43–54.
- Creti, A., and Fabra, N. (February 2004). *Capacity Markets for Electricity*. Center for the Study of Energy Markets (CSEM), University of California Energy Institute. Authors from University of Toulous and Universidad Carlos III de Madrid.
- Crew, M. A., and Kleindorfer, P. R. (1995). "The Theory of Peak Load Pricing: A Survey." *Bell Journal of Regulatory Economics*, 8: 215–248.
- Dansby, R. (1978). "Capacity Constrained Peak Load Pricing." *Quarterly Journal of Economics*, 92: 387–398.
- Ekelund, R. B., and Hébert, R. F. (1999). *Secret Origins of Modern Microeconomics. Dupuit and the Engineers*. Chicago, IL: Chicago Press.
- Eydeland, A., and Wolniac, K. (2003). *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*. Hoboken, NJ: Wiley Finance.
- Geman, H. (2005). *Commodities and Commodity Derivatives*. Chichester, UK: Wiley Finance, John Wiley.
- Harris, C. (2006). *Electricity Markets*. Chichester, UK: Wiley.
- . (2014). *Fixed and Variable Costs—Theory and Practice in Electricity*. New York: Palgrave.
- Hirshleifer, J. (1958). "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics*, 72: 451–62.

- Hogan, W. W. (2006). "Electricity Market Restructuring: Successful Market Design: Electricity Deregulation Six Years Later-The Solution or the Problem?" NECPUC Conference, Rockport, ME, June 12, 2006.
- Hopkinson, J. (1892). "On the Cost of Electric Supply. Presidential Address to the Junior Engineering Society." In *The Development of Scientific Rates for Electricity Supply*. Detroit, MI: The Edison Illuminating Company of Detroit.
- Hotelling, H. (1939). "The Relation to Marginal Costs in an Optimum System." *Econometrica*, 7(2) (April): 151–155.
- Joskow, P. L. (1976). "Contributions to the Theory of Marginal Cost Pricing." *The Bell Journal of Economics*, 7(1): 197–206.
- . (2001). "California's Energy Crisis." *Oxford Review of Economic Policy*, 17(3): 365–388.
- Joskow, P., and Kahn, E. (2001). "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market during Summer 2000." January 15. web.mit.edu/pjoskow, updated in the *Energy Journal* (2002).
- . (2002a). "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market during Summer 2000." *The Energy Journal*, 23(4): 1–35.
- . (2002b). "A Quantitative Analysis of Pricing Behavior in California's Wholesale Market during Summer 2000." *The Energy Journal*, 23(4): 1–35.
- Lindahl, E. (1919). *Die Gerechtigkeit in der Besteuerung*. Lund: Hakan Ohlssons Buchdruckerie.
- Lyndon, L. (1923). *Rate Making for Public Utilities*. New York: Mc-Graw Hill.
- Meek, R. L. (1963). "The Bulk Supply Tariff in Electricity." *Oxford Economic Papers*, 15(2): 107–123.
- Oren, S. S. (2000) "Capacity Payments and Supply Adequacy in Competitive Electricity Markets." In *Proceedings of the VII Symposium of Specialists in Electric Operational and Expansion Planning*, Curitiba (Brasil), May 21–26, 2000.
- Panzar, J. C. (1976). "A Neoclassical Approach to Peak Load Pricing." *The Bell Journal of Economics*, 7(2): 521–530.
- Pratt, I. W. (1964). "Risk Aversion in the Small and the Large." *Econometrica*, 32(1, 2): 122–136.
- Rosellón, J. (2005). *Different Approaches to Supply Adequacy in Electricity Markets*. Mexico and Harvard: Centro de Investigación Docencia Económicas (CIDE) and Harvard University
- Ruggles, N. (1949). "Recent Developments in the Theory of Marginal Cost Pricing." *Review of Economic Studies*, 17: 107–126
- Samuelson, P. A. (1954). "A Pure Theory of Public Expenditure." *Review of Economics and Statistics*, 36: 387–389.
- . (1955). "Diagrammatic Exposition of a Theory of Public Expenditure." *Review of Economics and Statistics*, 37: 350–356.

- Tobin, J. (1969). "A General Equilibrium Approach to Monetary Theory." *Journal of Money, Credit and Banking*, 1(1): 15–29.
- Turvey, R. (1968a). *Optimal Pricing and Investment in Electricity Supply*. London: George Allen and Unwin.
- . (1968b). "Peak Load Pricing." *Journal of Political Economy*, 76 (February): 101–113.
- . (1968c). *Public Enterprise: Selected Readings*. Harmondsworth: Penguin.
- . (1969). "Marginal Cost." *The Economic Journal*, 79(314): 282–299
- . (2000). "What are Marginal Costs and How to Estimate Them." University of Bath, Centre for the Study of Regulated Industries.
- Turvey, R., and Anderson, D. (1977). *Electricity Economics*. Baltimore: Johns Hopkins University Press.
- Vázquez, C., Rivier, M., and Pérez-Arriaga, I. J. (2002). "A Market Approach to Long-Term Security of Supply." *IEEE Transactions on Power Systems*, 17(2): 349–357
- Von Neumann, J., and Morgenstern, O. (1944). *The Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.
- Weisbrod, B. A. (1964). "Collective-Consumption Services of Individual-Consumption Goods." *Quarterly Journal of Economics*, 78: 471–477.
- Weitzman, M. (1974). "Prices vs. Quantities." *Review of Economic Studies*, 41(4): 477–491.
- Weron, R., and Simonsen, I. (2005). "Blackouts, Risk and Fat-Tailed Distributions." Hugo Steinhaud Center for Stochastic Methods, Wrocław University of Technology, Poland and Department of Physics, NTNU Trondheim, Norway. arXiv.physics/0510077v1.
- Williamson, O. E. (1966a). "Peak-Load Pricing and Optimal Capacity Under Indiv.isibility Constraints." *American Economic Review*, 56: 810–827.
- . (1966b). "Peak-Load Pricing—Some Further Remarks." *Bell Journal of Economics and Management Science*, 5(1): 223–228.
- Wright, A. (1896). *Cost of Electricity Supply*. In Edison (1915), "The Development of Scientific Rates for Electricity Supply." Printed for private circulation only. The Edison Illuminating Company of Detroit. Minutes of Municipal Electrical Association.
- Yakubovich, V., Granovetter, M., and McGuire, P. (2005). "Electric Charges: The Social Construction of Rate Systems." *Theory and Society*, 34: 579–612.

INDEX

- alternating current load flow (ACLF), 231
- ancillary markets, 149
- arbitrage, 149
- available capacity (ACAP), 109

- balancing mechanism, 145
- Barstow tariff, 29, 30
- Bentham, Benthamite, 8
- Bertrand game, 18
- Bessière, 39
- Bidwell, 192
- Boiteux, 34, 39
- British Electricity Transmission and Trading Arrangements BETTA, 32, 144
- Brown and Johnson framework, 59

- California crisis, 2
- call option spreads, 151
- capacity mechanisms, 183
- caplets, 151
- caps, 151
- Carlton framework, 81
- Chao framework, 105
- Clow tariff, 31
- consumer protection, 204
- consumer's surplus, 8
- contracts for difference (CFD), 147
- cost frontier
 - duality, 21
- Cournot game, 18
- Crampton and Stoft, 195
- Cremer, Gamsi and Laffont (CGL), 221
- Creti and Fabra, 205

- Crew and Kleindorfer framework, 90, 171

- Dansby framework, 99
- deadweight loss, 9
- decentralised energy, 202
- deficiency, 164, 185
- demand side management (DSM), 28
- dimensions of service, 174
- direct current load flow (DCLF), 231
- distribution charging, 235
- Doherty tariff, 29
- Drèze framework, 33
- Dupuit, 2

- econo-engineers, 2
- Edgeworth, 18
- Eisenmenger, 29
- Electricité de France (EDF), 34
- energy only markets, 3, 198
- Enron, 2

- firm contracts, firmness, 147, 185
- fixed cost
 - allocation, 19, 26
 - fixed capital cost, 20
- Forward Capacity Model (FCM), 195
- forward contracts, 146
- futures contracts, 147

- generation order and loading (GOAL), 32

- hedging, 178
- Hirshleifer framework, 123

- Hopkinson tariff, 28, 29, 30, 31
hotelling framework, 9
- imbalance, 146
installed capacity mechanism (ICAP), 184
- Kaldor Hicks efficiency, 59
Kapp meter, 29
Karush Kuhn Tucker (KKT), 133
Kirchhoff's laws, 231
- lexical. *See* lexicographical
lexicographical ordering, 9
liberalization, of retail, 32
load duration function, 10
load factor duality, 14
location marginal pricing (LMP), 143
loss of load expectation (LOLE), 114
loss of load probability (LOLP), 65, 113–16, 144
- market power, 17
Massé, 34
merit order, 12
Modern Portfolio Theory, 44
moral hazard, 18
- Nash equilibrium, 18
New Electricity Trading Arrangements (NETA), 32
non firm contracts, 147
- option
American, 169
Bermudan, 170
declaration, delivery date, delta, exercise, expiry date, extrinsic value, intrinsic value, kappa, ladder, strike price, swaption, tenor, volatility, 151
European, 150
real, 174
swing, 172
- Oren, 191
over the counter (OTC) contracts, 147
- Panzar framework, 126
Pennsylvania, New Jersey, Maryland (PJM), 205
physical notification (PN), 145
plant life utilization, 13
point to point regulatory model, 14
pool, 141, 143
the England and Wales, 32, 141
preemption, 19
principle components (PC), 201
public goods, 15, 57
- rationing of demand, 53, 57, 68, 73–8
Rawls, Rawlsian, 8
reliability, plant, 161, 176
reserve markets, 149
risk
cost of, 16, 155
- scarcity rent, 196
secondary markets, 186
shocks, 10, 201
smart, 203
Smith, Adam, 2
spot contracts, 146
stack evolution, 25
Stackelberg game, 18
Stasi, 39
stationary probability distribution, 7
Steiner framework, 47
Supplier Hub regulatory model, 14, 33
supply chain, 14
system marginal price (SMP), 144
- Tempo tariff, 31
Tobin Q, 20
triad, 30, 31
Turvey algorithm and inequality, 23–5

- unbundling, 14
- Unserved Energy (USE), 65, 113–16
- value of lost load (VoLL, VLL), 20, 35, 144
- variable cost pricing, 1
- Vázquez, Rivier and Pérez-Arriaga (VRP), 192
- Visscher framework, 66
- Von Neumann and Morgenstern (VNM), 6, 16
- Walras, 47
- Wealth of Nations, 2
- Williamson, 50, 60
- Wright meter, 29, 30