561 LECTURE NOTES IN ECONOMICS AND MATHEMATICAL SYSTEMS

Hagen Bobzin

Principles of Network Economics

Springer

Lecture Notes in Economics and Mathematical Systems 561

Founding Editors:

M. Beckmann
H. P. Künzi

Managing Editors:

Prof. Dr. G. Fandel
Fachbereich Wirtschaftswissenschaften
Fernuniversität Hagen
Feithstr. 140/AVZ II, 58084 Hagen, Germany

Prof. Dr. W. Trockel
Institut für Mathematische Wirtschaftsforschung (IMW)
Universität Bielefeld
Universitätsstr. 25, 33615 Bielefeld, Germany

Editorial Board:

A. Basile, A. Drexl, H. Dawid, K. Inderfurth, W. Kürsten, U. Schittko

Hagen Bobzin

Principles of Network Economics

Springer

Author

Dr. Hagen Bobzin
Private Docent
University of Siegen
School of Economic Disciplines
Department of Economics
57068 Siegen
Germany
hug.bobzin@t-online.de

ISSN 0075-8442
ISBN-10 3-540-27693-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27693-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

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

Typesetting: Camera ready by author
Cover design: *Erich Kirchner*, Heidelberg

Printed on acid-free paper 42/3130Jö 5 4 3 2 1 0

To my wife

Preface

Almost all economic activities in modern societies are scattered through space and time. Transport processes, as a consequence, pervade everyday life and they have deep impact on economic and social prosperity. Today's standard of living would just be unthinkable in the absence of water or power supply systems; road, railway, and air traffic systems are virtually used by everyone; and many people even cannot imagine to live in a world without telecommunications networks — including television, telephony, and the Internet. All examples have some kind of a transport process in common, by which people, commodities, or just data are moved along the interconnections of a network.

The purpose of this text is to provide an *economic view* on basic principles of transportation related network activities. In doing so the analysis is not restricted to certain types of transport networks at the outset and this requires a relatively simple *production technology*. In order to describe the behavior of groups of actors involved in the transport process, *microeconomic theory* suggests to distinguish between the provision of networks and of network services. Consequently, the analysis refers to at least one network carrier who offers a system of network components. On the basis of this network a second group of actors produces services and supplies them to the third group – the consumers – on the respective market. Having consumer sovereignty in mind, the principal question arises as to how to adjust the production structure of networks such that they fit best to the needs of the society.

Apart from technical problems which are caused by the complexity of network problems and by technological peculiarities, various difficulties of network analysis result from *imperfections of markets* for network services. Network carriers, for example, frequently possess remarkable market power and there are numerous externalities not only external to the network but also internal to the network. It is quite obvious that simultaneous network processes share certain network resources and overstraining these capacities induces congestion at certain places in distinct periods of time. Moreover, durable and/or indivisible investment goods cause tremendous fixed and overhead costs that must be borne not only by network participants but also by persons outside of the network.

Following the concept of the *market mechanism*, where market prices coordinate economic activities in order to utilize scarce resources efficiently, the analysis starts with the core problem of traffic assignment. Pricing strategies are discussed which ensure the efficient use of given network resources in different settings. In the next step network design abandons the assumption of a fixed network. Investment problems are examined in order to adjust the network to consumer needs; they are expressed, e.g., by travel time, accessibility, and reliability. Synchronization of network processes is an additional attempt to overcome problems associated with stochastic traffic flows. Such problems are mainly reflected by congestion and queuing which considerably reduce the perceived quality of service.

The author's deep gratitude goes to Professor Dr. Walter Buhr, University of Siegen (Germany), for his advising and supporting guidance. Acknowledgment is also to be paid to Professor Dr. Karl-Josef Koch, University of Siegen (Germany), for reviewing this book. Further thanks are due to PD Dr. Thomas Christiaans, whose comments led to many improvements of the work. Most of the burden, however, was born by my wife, Dr. Gudrun Bobzin. She offered constant intellectual inspiration and valuable criticism regarding my research on transportation economics.

Hennef (Sieg), Germany

Hagen Bobzin
April, 2005

Contents

Preface VII

List of Figures XIII

List of Tables XV

List of Symbols XVII

1 An Outline of Network Economics 1

2 Fundamentals of Networks 7
2.1 The Importance of Networks to the Economy 7
2.1.1 Markets for Network Services 7
2.1.2 Network Technologies 15
2.1.3 Construction of Networks 28
2.1.4 Pricing of Network Services 31
2.1.5 Solution Methods 36
2.2 Linear and Non-linear Programming 37
2.2.1 Duality in Linear Programming 37
2.2.2 Convex Analysis 42
2.2.3 Convex Analysis Applied to Linear Programming 51
2.2.3.1 Perturbations of the Right Hand Side 51
2.2.3.2 Feasible Perturbations of the Right Hand Side 53
2.2.4 Basic Differentiation Rules 54
2.3 Transport Networks 55
2.3.1 Elementary Features of Transport Networks 55
2.3.2 Flow Conservation (Kirchhoff's Law) 63
2.3.3 Undirected Capacitated Networks 65
2.3.4 Basic Problems for Transport Networks 67
2.3.5 Decomposition of Transportation Network Planning 70
2.3.5.1 Sequential Procedure of Travel Forecasting 70

2.3.5.2 Sequential Decision Process of Travel Demand ... 74
2.4 Summary ... 76

3 Utilization of Transport Networks: Basics ... 77
3.1 Preliminary Examples ... 77
3.1.1 Max Flow–Min Cut Theorem ... 77
3.1.2 The Minimum Cost Problem for a Single Good ... 84
3.1.2.1 One Source and One Sink ... 84
3.1.2.2 Many Sources and Many Sinks ... 89
3.1.3 Multicommodity Flows ... 93
3.1.3.1 Cost Minimization ... 93
3.1.3.2 Production Technology ... 95
3.2 Cost Minimal Trip Assignment ... 98
3.2.1 Production Technology ... 98
3.2.1.1 Feasible Network Activities ... 98
3.2.1.2 Special Features of Data Networks ... 102
3.2.1.3 Aspects of Trip Assignment ... 102
3.2.2 Fixed Link Costs per Unit ... 104
3.2.2.1 System Optimality ... 104
3.2.2.2 User Optimality ... 107
3.2.3 Capacity Constraints ... 110
3.2.3.1 Fixed Capacities ... 110
3.2.3.2 Other Resource Constraints ... 120
3.2.3.3 Existence of Feasible Solutions ... 123
3.2.3.4 Revenue Maximization ... 123
3.2.4 Alternative Mathematical Formulations ... 126
3.2.4.1 Ordinary Convex Programs ... 126
3.2.4.2 Variational Inequalities ... 127
3.2.4.3 Algorithmic Remark ... 132
3.2.4.4 Nonlinear Complementary Problems ... 135
3.2.4.5 Fixed Point Problems ... 135
3.2.5 Congested Networks ... 138
3.2.5.1 System Optimality ... 138
3.2.5.2 User Optimality ... 146
3.2.5.3 Further Properties of User Optima ... 159
3.2.5.4 Routing in Data Networks ... 163
3.3 Summary ... 166

4 Utilization of Transport Networks: Extensions ... 169
4.1 Stochastic Elements of Traffic Assignment ... 169
4.1.1 Stochastic Route Choice ... 169
4.1.2 Estimates from Trip Tables ... 173
4.1.3 Estimates from Link Flows ... 177
4.1.4 Doubly Constrained Models ... 179
4.2 Traffic Assignment with Price Sensitive Demand ... 185

4.2.1 Price Dependent Demand for Traffic 185
4.2.2 Traffic Equilibria 189
4.2.3 User Optimality in Terms of a Variational Inequality Problem 191
4.2.4 Marginal Cost Pricing 200
4.2.5 Non-invertible Demand Functions 201
4.3 Special Issues 215
4.3.1 Preliminary Remarks 215
4.3.2 Bilevel Programming 216
4.3.3 Multiclass-User Traffic Assignment 218
4.3.4 Dynamic Traffic Assignment 219
4.3.5 Spatial Price Equilibrium 222
4.4 Summary 225

5 Design of Transport Networks 227
5.1 Introductory Aspects of Network Optimization 227
5.1.1 Construction of Networks 227
5.1.2 Network Improvements 232
5.1.3 The Maximum Flow Problem Revisited 235
5.1.4 Minimum Investment Cost 236
5.2 Multicommodity Flow Problem 239
5.2.1 Feasibility Conditions 239
5.2.2 Minimum Investment Cost 242
5.2.2.1 A Reference Case 242
5.2.2.2 An Application to Telecommunications Networks . 244
5.2.2.3 Traffic Diversion 246
5.2.2.4 Budget Design Problem 250
5.2.2.5 Price Sensitive Demand for Traffic 253
5.2.3 Special Cases 269
5.2.3.1 Hub-and-Spoke Networks 269
5.2.3.2 Hierarchic Networks 274
5.2.3.3 Data Networks with Queuing 275
5.2.3.4 Telephone Networks without Queuing 286
5.3 Survivability Models 295
5.3.1 Normal Operating State 295
5.3.2 Failure States of the Network 298
5.3.3 Minimum Requirements on Failure States 299
5.3.3.1 Diversification 299
5.3.3.2 Reservation 300
5.3.3.3 Path Restoration 302
5.4 Summary 303

6 Synchronization of Network Processes 305
6.1 Scheduling Discrete Event Dynamic Systems 305
6.1.1 Resource Utilization of Simultaneous Processes 305
6.1.2 Real World Applications 307
6.1.3 Auxiliary Transportation Problems 310
6.2 Max-Plus Algebra 313
6.2.1 Basic Operations 313
6.2.2 Discrete Event Dynamic Systems 316
6.2.2.1 Timetables without Delays 316
6.2.2.2 Average Interdeparture Time 318
6.2.2.3 Critical Cycles 320
6.2.2.4 Periodically Reiterated Timetables 324
6.2.2.5 Synchronization 326
6.2.2.6 Stability and Sensitivity 329
6.3 Solving the Railway Example by *MATHEMATICA* 331
6.4 Economic Theory of Synchronization 334
6.4.1 Production Technology 334
6.4.1.1 Output Measurement 334
6.4.1.2 Time Limits 337
6.4.2 Assessment of Synchronized Timetables 340
6.4.2.1 Passenger Waiting Times 340
6.4.2.2 Instruments of the Carrier 341
6.4.3 Cost Measurement 345
6.4.3.1 A Retrospect of Network Cost Evaluation 345
6.4.3.2 Prospective Approaches of Measuring Transport Costs 349
6.4.3.3 The Railway Example Revisited 354
6.5 Summary 358

7 Critical Review 361
7.1 Synopsis of Network Economics 361
7.2 Closing Comments 366

References 369

Index 386

List of Figures

2.1 An introductory example 38
2.2 Illustration of Farkas' Theorem 41
2.3 The subdifferential of f at $\bar{\mathbf{x}}$ 45
2.4 The gradient y of f at $\bar{x}$ 46
2.5 Illustration of KKT conditions 51
2.6 Basic elements of networks 55
2.7 Graphic representation of inflows and outflows 56
2.8 Representation of a node capacity 57
2.9 Example of a network 57
2.10 Cuts of a network 62
2.11 An easy case of the transportation problem 64
2.12 Sequential procedure of travel forecasting 70
2.13 Activity based traffic demand 71
2.14 Modal split 72
2.15 Sequential decision process 74

3.1 Maximum flow problem 78
3.2 Minimum cut problem 83
3.3 Relationship between link flows 84
3.4 Minimum cost problem 88
3.5 Spatial price equilibrium 90
3.6 Cost minimization 94
3.7 Wardrop's first principle 109
3.8 Traffic flow relationships 111
3.9 Variational inequality problem 128
3.10 The projection operator P_Ω 136
3.11 Overview of introduced problems 138
3.12 Total cost of transportation 148
3.13 Traffic assignments 151
3.14 Traffic assignments with an external capacity restriction 153

3.15 Illustration of the fixed point problem 160
3.16 Braess' Paradox 163

4.1 Wardrop equilibrium 192
4.2 Consumer's surplus 196
4.3 Measuring net benefit 199
4.4 Conjugate convex f^* of f 203
4.5 Equilibrium flows on route ρ_1 210
4.6 Conjugate convex g_e^* of g_e 211
4.7 Duality 215
4.8 Introducing queuing 220
4.9 Dynamic traffic flows with queuing 220

5.1 Special trees 229
5.2 Complementary investments 253
5.3 Profit maximization (1) 257
5.4 Profit maximization (2) 261
5.5 Bilevel programming 264
5.6 Maximizing the net benefit 267
5.7 Hub-and-spoke networks 269
5.8 Cost sharing 270
5.9 Hub location 272
5.10 Hierarchic network 274
5.11 Delays in a waiting system 277
5.12 Birth-death model ($M/M/1$ queue) 280
5.13 Erlang's example 286
5.14 Network survivability 299

6.1 Suburban streetcar system of the city of Stuttgart, Germany 308
6.2 Weighted digraph with loops 310
6.3 Precedence graph of $\mathbf{T}$ 311
6.4 Normalized precedence graph 322
6.5 Critical graph of $\mathbf{T}$ 324
6.6 DEDS with inputs and outputs 327
6.7 Timetable of the precedence graph 328
6.8 Petri net 335
6.9 Waiting times 340
6.10 Schedule delay costs 343
6.11 Alternative flows 351
6.12 Independent flows 352
6.13 Interdependent flows 353
6.14 Flows on a railway network 355
6.15 Alternative timetables 357

List of Tables

1.1 Introductory examples for different groups of network participants .. 3

2.1 Rules for forming the dual 39
2.2 The node/arc incidence matrix of the network in Figure 2.9. 59
2.3 The arc/path incidence matrix of the network in Figure 2.9. 60

3.1 The node/arc incidence matrix of the network in Figure 3.5. 91

6.1 Max-plus algebra 314
6.2 Max-plus matrix algebra 316
6.3 Service sojourn times 355

List of Symbols

Bold small letters ($\mathbf{a}, \mathbf{b}, \ldots, \boldsymbol{\alpha}, \boldsymbol{\beta}, \ldots$) denote column vectors.
Bold capital letters ($\mathbf{A}, \mathbf{B}, \ldots, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \ldots$) denote matrices.
Calligraphic capital letters ($\mathcal{B}, \mathcal{D}, \mathcal{E}, \ldots$) denote sets.

$<$	$\mathbf{x} < \mathbf{y} :\Longleftrightarrow$ $x_j < y_j \ (j = 1, \ldots, n)$
$\leqq$	$\mathbf{x} \leqq \mathbf{y} :\Longleftrightarrow$ $x_j \leqq y_j \ (j = 1, \ldots, n)$
$\leq$	$\mathbf{x} \leq \mathbf{y} :\Longleftrightarrow$ $[\mathbf{x} \leqq \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}]$
$\hat{}$	denotes optimal solutions
$\tilde{}$	refers to routes
$\oplus, \otimes, \oslash, \otimes$	operators in max-plus algebra; see p. 314
$*$	denotes conjugate convex functions, see p. 203
$\star$	denotes conjugate concave functions, see p. 203
$\emptyset$	empty set
$:\Longleftrightarrow$	equivalent by definition
cone	normal cone
Dom	effective domain of a proper function
n-Dom	effective domain of an n-proper function
E	expected value or mean
int	interior of a set
$\mathbb{N}$	natural numbers
$\mathbb{N}_0$	natural numbers and zero
Prob	probability
$\mathbb{R}$	real numbers
$\mathbb{R}_\oplus$	real numbers augmented by $-\infty$
rint	relative interior of a set
$\mathbf{1}$	vector with all components equal to one
∂f	subdifferential of f
∇f	gradient of f
$\nabla \mathbf{f}$	transposed Jacobian matrix $\mathbf{J_f}$
$\|\mathbf{x}\|$	Euclidean norm
$\alpha(e), \ \alpha(\boldsymbol{\rho})$	start node of arc e or path $\boldsymbol{\rho}$
$\beta_0, \ \beta_1$	parameter
$\gamma(e)$	arc weight per unit of flow $\varphi(e)$
$\gamma_e, \ \boldsymbol{\gamma}$	link prices
$\tilde{\gamma}_\rho, \ \tilde{\boldsymbol{\gamma}}$	route prices
$\delta(\cdot, \mathcal{C})$	indicator function of set $\mathcal{C}$
$\delta_e, \ \boldsymbol{\delta}$	capacity enlargements
$\boldsymbol{\Delta}$	arc/path incidence matrix defined on p. 60
Δf	superdifferential of f
ε	zero element in max-plus algebra

ζ, $\boldsymbol{\zeta}$ circuit, cycle
θ dispersion parameter
θ_e, $\boldsymbol{\theta}$ fractions of flows using link e
$\tilde{\theta}_\rho$ fractions of flows using route ρ
κ_e, κ_{ij}, $\boldsymbol{\kappa}$ capacity of arc $e = (v_i, v_j)$
κ_v, $\kappa(\mathcal{S}, \mathcal{T})$ capacity of node v or cut $(\mathcal{S}, \mathcal{T})$
$\boldsymbol{\Lambda}$ origin destination/path incidence matrix defined on p. 60
$\boldsymbol{\lambda}$ Lagrange multipliers
λ (1) arrival rate (2) eigenvalue in max-plus algebra
$\boldsymbol{\mu}$ Lagrange multipliers
μ service rate
$\mu(\boldsymbol{\rho})$ mean weight of path $\boldsymbol{\rho}$
$\mu_{\max}(\mathbf{A})$ maximum cycle mean
$\xi_{(s,t)}$, $\boldsymbol{\xi}$ demand for trips
$\pi_{(s,t)}$, $\boldsymbol{\pi}$ shadow prices (assigned to OD-pairs)
π_k probability $P\{K = k\}$
π profit
ρ, $\boldsymbol{\rho}$ chain, path, route
$\rho(\mathbf{A})$ spectral radius of $\mathbf{A}$
ϱ occupation rate
$|\boldsymbol{\rho}|_\ell$ length of path $\boldsymbol{\rho}$, see p. 312
$|\boldsymbol{\rho}|_w$ weight of path $\boldsymbol{\rho}$, see p. 312
σ_{ij}, Σ delay at node v_j that does not reach v_i
τ_e, τ_{ij}, $\boldsymbol{\tau}$, $\mathbf{T}$ travel time or distance on link e measured in seconds
$\tilde{\tau}_\rho$, $\tilde{\boldsymbol{\tau}}$ travel time or distance on route ρ measured in seconds
φ_e, φ_{ij}, $\varphi_{ij}^{(t)}$, $\boldsymbol{\varphi}$, $\boldsymbol{\Phi}$ flow on arc $e = (v_i, v_j)$ (with destination v_t)
$\varphi^{\#}$ total number of flow units
Φ set of feasible link flows
Φ^D defined on p. 61
Φ^S defined on p. 61
Φ_X^S defined on p. 191
Φ' defined on p. 179
φ_v^-, $\boldsymbol{\varphi}^-$ inflow of node v
φ_v^+, $\boldsymbol{\varphi}^+$ outflow of node v
χ entropy (χ like chaos)
$\boldsymbol{\psi}$ 0-1 variables
$\omega(e)$, $\omega(\boldsymbol{\rho})$ end node of arc e or path $\boldsymbol{\rho}$
Ω feasible sets
$\mathbf{A}$, a_{ve} adjacency matrix or arc/node incidence matrix
$\mathbf{B}$ arc/node incidence matrix
$\mathcal{B}$ set of OD-pairs
$\mathfrak{b}$ dual variable (assigned to the budget $\bar{I}$)
C cost function, context specific
c_e, $\mathbf{c}$ cost per unit on link e
c_e^f, $\mathbf{c}^f$ fixed cost per unit on link e
$\tilde{c}_\rho$, $\tilde{\mathbf{c}}$ cost per unit on route ρ
$\mathfrak{c}_e$, $\mathfrak{c}$ marginal cost per unit on link e
$\tilde{\mathfrak{c}}_\rho$, $\tilde{\mathfrak{c}}$ marginal cost per unit on route ρ
$\mathcal{D}$ set of destinations
$\mathbf{D}$ destination/path incidence matrix
$\mathbf{d}$ timetable
$\mathbf{d}_j$, $\mathbf{d}$ deadline at node v_j
e expenditure function
e, $\mathcal{E}$ edge, arc or link, set of edges ($e = 1, \ldots, m$); $e = (v_i, v_j)$
$\mathbf{e}_j$ unit vector having all components equal to zero with the exception that the j^{th} component is one; the $n \times n$ matrix $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ is called the identity matrix; $\mathbf{e}_j^{\mathsf{T}}\mathbf{x} = x_j$

f, $\tilde{f}$ total travel cost
$\mathcal{G}$, $\mathbf{G}$ graph or directed graph (= digraph)
$\mathcal{G}(\mathbf{A})$ precedence graph of $\mathbf{A}$, see p. 311
$\mathcal{G}^c(\mathbf{A})$ critical graph of $\mathbf{A}$
g, $\tilde{g} \equiv g\mathbf{\Delta}$ a line integral representing a cost term
g^*, $(g\mathbf{\Delta})^*$ conjugate convex of g and $g\mathbf{\Delta}$
h_ρ, $\mathbf{h}$ path flow on route ρ
H set of feasible route flows
H^D defined on p. 61
H^S defined on p. 61
H_X^S defined on p. 191
H' defined on p. 183
$\mathbf{H}_f$ Hessean matrix of f
I, $\bar{I}$ investment cost, budget
$\mathbf{I}$ unit matrix or identity matrix, see unit vector $\mathbf{e}_j$ redefined in max-plus algebra
$\mathcal{I}$ set of intermediate nodes
$\mathbf{J_f}$ Jacobian matrix of $\mathbf{f}$
k, $k\mathbf{\Lambda}$ a line integral representing the willingness to pay
$k^\star$, $(k\mathbf{\Lambda})^\star$ conjugate concave of k and $k\mathbf{\Lambda}$
$\mathtt{k}_e$, $\mathbf{k}$ excess capacities of arcs
$\mathtt{k}_j(t)$, $\mathbf{k}(t)$ number of events at node v_j up to time t
$\mathcal{L}$ Lagrangean functions
$\mathcal{N}$ flow-network
$\mathcal{O}$ set of origins
$\mathbf{O}$ origin/path incidence matrix
$p_{(s,t)}$, $\mathbf{p}$ minimum cost for travel from v_s to v_t
p_v^D, $\mathbf{p}^D$ consumer prices at node v
p_v^S, $\mathbf{p}^S$ producer prices at node v
$\mathtt{p}_{(s,t)}$, $\mathbf{p}$ OD-specific shadow prices
P production possibility set
P_Ω projection operator defined on p. 135
q_e, $\mathbf{q}$ price vector of capacity units
$\mathtt{q}_e$, $\mathbf{q}$ shadow prices of link capacities
$\tilde{\mathtt{q}}_\rho$, $\tilde{\mathbf{q}}$ route specific shadow prices
$\mathcal{R}(s,t)$ set of elementary routes from v_s to v_t
$\mathcal{R}$ set of all elementary routes in the network
$\mathcal{S}$ subset of nodes including the source
s_e, s_{ij}, $\mathbf{S}$ distance measured in meters
$\mathcal{T}$ subset of nodes including the sink
t_U distance or transformation function
t_e free flow travel time on link e
$\mathtt{t}_j(k)$, $\mathbf{t}(k)$ date of the k^{th} event at node v_j
$\mathtt{t}_\rho$, $\mathbf{t}$ link specific tolls
$\tilde{\mathtt{t}}_\rho$, $\tilde{\mathbf{t}}$ route specific tolls
$\mathtt{u}_e$, $\mathbf{u}$ dual variables (assigned to links)
U utility function
v, $\mathcal{V}$ node, set of nodes $(v = 1, \ldots, n)$
$\mathtt{v}_v$, $\mathbf{v}$ dual variables (assigned to nodes)
V indirect utility function
w_e, $\mathbf{w}$ link costs
$\tilde{w}_e$, $\tilde{\mathbf{w}}$ route costs
w_e, w_{ij}, $\mathbf{W}$ weight of arc $e = (v_i, v_j)$
$\mathbf{W}^\diamond$, $\mathbf{W}^+$, $\underline{\mathbf{W}}$ auxiliary matrizes derived from $\mathbf{W}$
$\mathtt{w}_v$, $\mathbf{w}$ dual variables (assigned to nodes)
$\mathbf{x}$ trip matrix in vector form

$x^D_{(s,t)}$, $\mathbf{x}^D$	demand for trips from v_s to v_t, trip table in vector form
$\mathbf{x}^S$	generated traffic in vector form
y^S_v, $\mathbf{y}^D$	nodal demand
y^S_v, $\mathbf{y}^D$	nodal supply
$\mathbf{z}$	eigenvector in max-plus algebra

1

An Outline of Network Economics

Networks are the veins between all communicating activities scattered through space and time. The wide range of network types includes power supply systems, telecommunications networks and road systems to give only a few instances. All examples have some kind of a transport process in common. Due to the specific importance of each type of network, it takes no wonder that there is a very extensive literature on the characterization, representation, and optimization of different networks. Most texts, however, draw the attention to a single type of network and then solve particular problems. For instance, the pioneering study of Beckmann et al. (1956) concentrates on highway traffic and railroad transportation. In comparison Bertsekas, Gallager (1992) discuss data network related problems. Other texts such as Bertsekas (1998) follow a more general approach with respect to network types, but consider solely network optimization problems such as shortest path, max-flow, assignment, vehicle routing, multicommodity flow, etc. These approaches are supplemented by works such as Jungnickel (1994), who studies the complexity of network problems and gives some algorithms for their solution. Last but not least even particular technical problems need deep insight because of their far reaching effects. One example is Kleinrock's (1975) work on queuing theory.

In consideration of the recent literature on *network economics* it seems to be important to delimit the above given analysis of *transport systems* in the broadest sense from another concept which is nowadays referred to as *economics of networks*; see Erber, Hagemann (2002). This concept addresses organizational phenomena including the design of economic institutions, where the nodes of a network are substituted by interacting economic agents. In doing so economic networks distinguish between the social neighborhoods of actors and the global community, and the corresponding literature analyzes the de- and re-linking of these social structures in the context of an emerging global society; cf. Erber, Hagemann (2002, pp. 236–237). At least in principle, these network types can be subsumed under the present analysis of transport networks by considering the transportation of interpersonal messages. Indeed, the importance of neighborhoods has been analyzed before in the context of so-called gravity models where, for instance, the attractiveness of retail facilities to their customers is described by some distance decay relationship. A variant of such

real economic networks is defined by Economides[1] as *virtual* network which denotes a collection of compatible goods such as VHS video players or computer software that share a common technical platform. In this sense economic networks address issues of compatibility, coordination through technical standards, interconnection, e.g., the connection of formerly separated networks, and interoperability[2] of agents' decisions as well as their effects on pricing and quality of services; see Economides (1996). All these aspects pertain also to transport networks so that we call into question that new terms such as "economic" or "virtual networks" are too useful. Moreover, with regard to large-scale physical networks all of the aforementioned issues target essentially positive network externalities on the demand side, that is network users profit from increasing networks due to a rising number of users. This observation is also valid for transport networks, however, the effect is to be contrasted with scale effects on the supply side, which will be in the center of interest in our analysis. Clearly, many transport systems exhibit economies of scale, but more important is the determination (and removal) of bottlenecks that cause remarkable congestion costs as can be experienced in the daily rush hours. The economic problem is then to design reliable transport networks which are able to cope with a fluctuating and growing demand for traffic.

The purpose of this book is to provide an economic view on transportation related network activities. As the analysis is not restricted to certain types of transport networks at the outset, the first task is to work out similarities and differences between several network types. Afterwards networks will be interpreted as firms producing multiple services which are supplied on the respective market. In doing so the network problems such as flow maximization, traffic assignment or network design are restated in terms of microeconomic theory. For instance, a network provider may pursue some optimization task while he is restricted to the production technology of his network. Similarly, the network users or, more precisely, the consumers of network services will be assumed to behave in a utility maximizing way. The difficulties of network analysis result from various additional aspects. Simultaneous processes represent composite goods that have to share certain network resources. Overstraining these capacities induces congestion at certain places in distinct periods of time. Moreover, durable and/or indivisible investment goods cause tremendous fixed and overhead costs that must be borne not only by network participants but also by persons outside of the network. Although frequently denied, such economic observations are also valid for the Internet.

The structure of this book results from the following reasoning. Starting with different network types, we can distinguish at least three groups of actors involved. The first group consists of at least one network carrier who provides a system of network components, say nodes and links. On the basis of these networks the second

[1] Economides maintains an Internet site including a dictionary of terms in network economics at `http://www.stern.nyu.edu/networks/dictionary.html`.

[2] Technical standards, interconnection and interoperability are also major problems of transport systems as can be seen from container movements that use more than one transport mode.

group produces services and offers them to the third group, namely, the consumers. Unfortunately, even simple examples show that these three groups of actors cannot always be separated accurately and that further partitions are possible. The following table refers to different transport modes, that is road, railway, and aviation, where the cited groups overlap more or less. Car drivers, for example, produce services by making use of privately owned cars and consume these services on their own. In contrast railway travelers consume services offered by some railway company which is usually the owner of the railway network. Regarding aviation it is relatively simple to tell apart the three groups of actors, but aviation networks usually dissect into airports (referring to nodes) and air traffic control (referring to links). Similar distinctions apply to other networks (gas, water, electricity, telecommunications, etc.) where additional features must be taken into consideration as will be explained in Chapter 2.

Table 1.1 Introductory examples for different groups of network participants

network type	network carrier(s) (nodes, links)	service provider(s) (additional devices)	consumers
road	private or public supply	travelers with private car ownership	
		taxi drivers/public transport	passengers
		hauliers	shippers
railway	private or public railway companies		passengers
	railway network (tracks, stations, power)	railway services (trains)	passengers
aviation	airports/air traffic control	airlines (airplanes)	passengers

Having introduced the descriptive and technical principles of network analysis – i.e., the production technology – in Chapter 2, Table 1.1 suggests a procedure which stresses at first sight the supply side of the provision of network services. Road networks, however, indicate that consumer choice cannot be neglected in the analysis of network activity patterns, particularly, when travelers decide on route choice. Section 2.3.5 considers further decisions of network users such as mode or trip choice. Moreover, it suggests a procedure that decomposes the consumer behavior into a sequential decision process. This decomposition ends up with the core problem of traffic assignment, which is to be generalized up to locational choice.

According to microeconomic theory of the firm, Chapter 3 starts with the economic analysis of networks in terms of production theory. By assumption we begin in Section 3.1 with the simplest case where the network provides one homogeneous service. The first task is to find feasible *network activities* which are technically efficient in the sense that they indicate the maximum output on a fixed network with given capacities (maximum flow problem). The next step aims at the determination of a cost minimal network activity that supports a given output (minimum cost problem). Here only transport costs are taken into consideration. Given optimal network activities with respect to one homogeneous output, we then extend the

analysis with respect to *multicommodity flows* where our attention will be drawn to a cost minimal *traffic assignment* as core problem; see Section 3.2. For the sake of simplicity the analysis starts with constant transport costs per flow unit and the demand for traffic is supposed to be fixed. However, taking *congestion* into account, average travel costs increase with network utilization. In this case users' route choice tends to be suboptimal because travelers ignore their disturbing effects on other network users. These external costs can be corrected in principle by price incentives, say Pigou taxes, such that the new traffic assignment results in an overall system optimum.

The traffic assignment problem will be generalized in Chapter 4 by relaxing several assumptions. In the first instance Section 4.1 abandons the deterministic traffic assignment by introducing *stochastic elements of route choice*. In this setting we are interested in the most likely traffic assignment, which can be found by the concept of entropy maximization. By analogy to microeconomic theory of the household, the next step is to give up a fixed demand for traffic; see Section 4.2. A *price sensitive demand* for trips implies that all effects of *traffic diversion* are now superimposed by effects of *traffic creation* or *traffic diminution*. In other words, any measure that improves traffic conditions causes two modifications of the prevailing traffic pattern: it redirects some parts of the existing flows and it induces additional flows. Other generalizations such as a dynamic traffic assignment or a multiclass-user assignment are indicated in Section 4.3.

Chapter 5 continues with the cost minimal traffic assignment of Chapter 3 by giving up the assumption of a fixed network structure with given capacities. With regard to microeconomic theory this is the case where a firm is allowed to adjust all of its factors of production in the long run rather than having fixed amounts of inputs in the short run. In this sense *network design* seeks optimal *investment programs* that improve traffic conditions. This analysis is done for both a fixed and a price sensitive demand for traffic. Similar as before we start with one homogeneous flow and examine the efficient use of investment resources. This includes the maximum attainable flow given the investment costs as well as the minimum investment costs to obtain a prespecified flow. Afterwards the network design problem is discussed for multicommodity flows; see Section 5.2. In what follows selected examples point out that special network features result in specific network structures. One example results from link specific economies of scale so that network operators can take advantage of bundling flows. Moreover, a stochastic demand with queuing requires modified investment strategies which result in surplus capacities in order to avoid frictions of a standstill. Similar results are discussed with respect to susceptible network components. Excess capacities are then used to ensure a minimum operability of a network even when the failure of a network component suspends certain flows.

Besides queuing network operators may also use *schedules* to coordinate interdependent network activities on the basis of precedence constraints. In this sense Chapter 6 goes back to the assumptions of a fixed network with a given demand for traffic. The analysis aims at the determination of *feasible timetables* with a tradeoff between two objectives. On the one hand, timetables should be able to cope with

a certain amount of delays by virtue of buffer times and, on the other hand, they should minimize costs of waiting which are borne by the passengers. This last chapter is accompanied by the example of a railway system, but synchronization of network processes by timetables has a wide field of further applications; one particular example is the timing of road traffic lights.

Summarizing the work at hand is focusing on an economic view of transport networks. It delivers a consistent economic analysis of the decisions of network participants – users as well as providers – taking into consideration the costs of network design, service provision, and network usage.

2

Fundamentals of Networks

2.1 The Importance of Networks to the Economy

2.1.1 Markets for Network Services

(a) Actors Involved in the Network

The classification of transport networks starts with the market on which the respective services are traded. As will be discussed in Section 2.1.2, all these networks show their own peculiarities with regard to the used technology. Moreover, there are far reaching consequences that result from a proper network design, see Section 2.1.3. In the end economist want to determine an adequate pricing of network services. A couple of difficulties will be explained in Section 2.1.4.

Regarding markets of network services, Table 1.1 suggests to distinguish at least three groups of network participants that act on the demand and/or supply side of the respective market. It is useful to begin with a characterization of the two market sides, the market structure, and the notion of market equilibria. Afterwards, paragraph b explains what group of actors makes what network related decisions. Depending on the point of view the resulting activity pattern may be represented on different aggregation levels as can be found in paragraph c.

Demand. Transportation is a derived demand resulting from the desire to realize spatially separated activities. The most familiar case is to consume some commodity or service at a location that differs from where the good or service is produced. Shipping a freight from one place to another is the basic service produced on a transport network. In this sense (road) infrastructure may be seen as an intermediate input for other sectors (shippers). Depending on the type of freight, network outputs differ extensively. They range from physical commodities over electric currents to telecasts. Aside from freight movements, some economic activities take place at certain locations so that passengers have to make trips. Some standard examples are journeys to work, to shopping centers or for recreational purposes. In other cases networks such as telecommunications (or data) networks just connect points

of mutual interaction. A telephone call may then be interpreted as a substitute for more expensive passenger or freight (letter) movements.

In the wide field of networks most systems produce various composite services which cannot be substituted, i.e., a connection between two points is different from connecting another pair of locations. The demand for distinct services such as telephone calls is reflected by so-called multicommodity networks. Furthermore, in many real world examples there are separated networks which are interconnected so that parts of a transportation service are performed by distinct suppliers. As an example one can think of journeys which use a combination of private and public transport modes. Another case is that of telecommunications networks where roaming denotes the instance of connecting two parties which belong to different networks. In many cases, such as power supply systems, it is even possible that network services are supplied by making use of networks which belong to immediate competitors.

Supply. One important distinction on the supply side of networks refers to *network provision* and *service provision*[1]. In the ideal case the two sides can be separated so that the network is operated by the first party, say carrier, while "hauliers" supply services on that network. For many examples such as railroad systems or power supply systems it seems to be easy to "unbundle" the production of services from operating a network or parts of it. Indeed, this division is reality for air traffic where airports, airlines and air traffic control operate in vertically separated companies. The telecommunications sector, however, is characterized by several companies offering essentially the same services. In doing so they make use of their own network as well as other networks for instance by leasing lines. Operating a network *and* supplying services is referred to as *bundling* that is the tying of one service or product to the supply of others. Leasing lines or parts of a network indicates the other case: service providers who do not have their own network are sometimes referred to as switchless resellers. They just provide services on the platforms of other network providers.

Another extreme case is the Internet, a network of (data) networks. Here many subnetworks, each of which belonging to another provider, are interconnected to a common network on which services can be supplied all over the world. The Internet has a hierarchical structure similar to road systems consisting of streets, (country) roads, and highways. Aside from combining networks in the sense of complements, networks can also be interpreted as substitutes. While, in general, flights require a journey to the airport, most train journeys can also be made by private cars.

Apart from all of the aforementioned details any supplier has to make up his mind what and how much to produce. While every service provider has to decide on the quality of service delivered to his customers, the network provider must

[1] Service providers supply network services to third parties whether on their own network or otherwise. Internet service providers (ISPs) in particular offer access to the Internet either to end-users or on a wholesale basis to other service providers. A virtual network provider possesses no network but uses the platforms (hard- and software) of other network providers to offer services.

dimension and design his network. Besides the technical aspect of running the network efficiently two other important problems are to be solved. The network carrier wants to forecast the future demand for services and the corresponding capacities needed and he is interested in estimates on the capacities (or potential supply) of his competitors. The answers are particularly decisive for potential entrants to the market.

Market Structure. Following Sharkey (1987), economists of the 19th century and early 20th century spoke of natural monopoly conditions arising both from superior efficiency of single-firm production and the undesirable consequences of excessive or "destructive" competition. By the middle of the 20th century it was recognized that many network industries such as railroads or telecommunications possess to some degree the characteristics of a *natural monopoly*. It is then also "natural" to assert that these monopolies (or oligopolies) have to some degree market power, that is they are able to raise prices above the competitive level in that market for a non-transitory period. In doing so private decision-making takes inadequate account of the "public interest" which is the justification for governments all over the world to establish regulatory agencies. In many cases the governments even believe to be the only authority that can provide networks as well as network services efficiently. Some familiar examples of publicly owned networks are road systems[2], airlines, telecommunications and postal[3] networks. Since 1980 many of such networks have been privatized where the most frequently cited objectives of privatization are (1) the improvement of the economic performance and (2) the generation of public budget revenues through sale receipts; see Hanke (1987) for more details. Again regulatory authorities[4] were needed to prevent the incumbent monopolist from abusing its market power. The most important objective is now to assure that incumbents cannot illegally prevent potential rivals from market entrance so that a situation arises as if competition prevails.

Nevertheless, in many network industries overhead costs are a significant fraction of total costs and competition in such an industry is still far from being perfect. Barriers to entry indicate an additional cost which must be borne by entrants but not by firms already in the industry; or other factors, which enable an incumbent to maintain prices above the competitive level without inducing entry. Some typical strategies of the incumbents to deter rivals from market entrance include the following list.

- Prices which do deter entry by rivals with the same technology are known as *sustainable prices*. A market in which there are no barriers to entry is known as a *contestable market*.[5] If the market is contestable, the threat of potential entry

[2] On the pros and cons of a commercial provision of roads in a competitive framework see Roth (1996).

[3] Even the constitution of the United States of America includes in the first article that the Congress shall have power to establish post offices and post roads.

[4] For further details on regulating specific industries such as electricity supply, telecommunications or railway transport in Germany see Eisenkopf (2003).

[5] On the concept of contestable markets see Willig (1987).

will force the unregulated monopolist to choose from the set of sustainable prices provided that the monopolist behaves in the sense of sustainability.

- In the beginning entering competitors operate relatively small networks so that they are frequently forced to complete services by making use of other networks which belong to their rivals. The access to remote networks – say by interconnection[6] – and customers can easily be hindered by *excessive transit* or *termination fees*. The German cartel agency (Bundeskartellamt), for example, taunts the electricity supplier RWE Net AG in a warning letter of 13. August 2002 to burden other suppliers with excessive metering and billing prices, so that new suppliers in particular will be significantly hindered.
  The same reasoning holds true for a single transit which is defined as an interconnection service that involves the use of one switch but no third party. As a special case, local loop[7] unbundling was mandated by the EU in December 2000. It requires those operators designated as having significant market power to make their local networks (i.e., the telephone lines that run from a customer's premises to the local telephone exchange) available to other telecommunications companies.
- *Excess capacities* may be interpreted as a thread to potential entrants. They indicate that an increasing demand can be served by the incumbent particularly when a decline of prices occurs after the market entry of a rival. Perhaps the tremendous excess capacities in the telecommunications backbone network can be interpreted in this way. The huge capacities, however, became feasible by the use of relatively inexpensive fiber cables and are possibly just the result of miscalculating future demand.
- Regulated monopolies increasingly find themselves operating not only in regulated markets, but in competitive markets as well where they are often accused of illegally using resources from the regulated market to stifle competition. Such *cross subsidies* are often used by multiservice firms to finance costs in one market from profits made in another. This behavior includes predatory intent and attempts at transferring market power from a regulated monopoly to an unregulated market. Proving this abuse of market power is difficult because multiservice firms argue that their advantages result from economies of scope. Nevertheless, in March, 2001 the European Commission concluded its antitrust investigation into Deutsche Post AG (DPAG) with a decision, finding that the German postal operator, a beneficiary of letter monopoly, had abused its dominant position by granting fidelity rebates and engaging in predatory pricing in the market for business parcel services.

[6] Interconnection designates the physical and logical linking of telecommunications networks used by the same or a different organization in order to allow the users of one organization to communicate with users of the same or another organization, or to access services provided by another organization. Services may be provided by the parties involved or other parties who have access to the network.

[7] Local loop refers to the access network connection between a customer's premises and the local exchange. This usually takes the form of a pair of copper wires.

In the USA regulated railroads faced increasing competition from regulated and unregulated trucking, and regulated telephone companies faced increasing competition from private networks. In effect regulators were asked in what way a regulated firm should be allowed to compete with an intermodal rival or entrant. If an *intermodal competitor* enters only the most lucrative markets, he was attacked as "cream skimming" by the regulated firms and portrayed as innovative competition by the entrant.

Another barrier to entry results from barriers to exit. For instance, sunk costs are not recoverable if the activity for which they were incurred ceases.

Given inferior market results one has to think of proper forms of regulation. Two categories are to be distinguished. (1) *Ex ante regulation* includes the separation of networks and services (power supply systems, railway systems) or the unbundling of other parts of vertically or horizontally integrated firms (unbundling local loop in telecommunications systems). Other measures seek to guarantee the access of new competitors to the networks of incumbents at "fair" prices (formerly state-owned monopolies). (2) *Ex post regulation* focuses on the correction of market results. For example, some measures try to enforce marginal cost pricing, others impose price caps, and again others regulate the rate of return. In face of a long list of dissatisfactory and failing regulatory programs it is frequently postulated, not only by lobbies, to deregulate industries. By most economic measures, the industry after *deregulation* has been operating more efficiently and the average firm has increased profitability, see Breyer, MacAvoy (1987). Network users in particular profit from deregulation. For example, deregulating the German telecommunications sector has shown a drastic decline in prices and, correspondingly, an increasing demand; cf. Götz (2001).

Market Equilibrium. Having described aspects of both the demand and the supply side as well as the market structure, in the next step we are interested in the characterization of market equilibria. Note that the demand for network services refers to composite goods such as a trip from A to B involving the utilization of a sequence of roads. Each network resource in turn, that is for example one road, represents the supply side of the network, but it provides merely a fractional part of a complete network service. In accordance with Beckmann et al. (1956), the connecting link is found in the distribution of services over the network.

Given a meaningful definition of a network equilibrium, economists are interested in several of its properties. First, the *existence* of a technically feasible network activity is needed such that the output matches the demand under the prevailing circumstances. This situation is called a (static) equilibrium if none of the participants in the network has an incentive to alter its behavior. That is, all actors behave in accordance with correct expectations on the essential properties of the prevailing equilibrium. Second, in some special cases one can show that only one network equilibrium exists. This result on *uniqueness*, however, must be handled with care as there may be alternative representations of the equilibrium which are not unique. Third, the equilibrium is said to be *stable* if any but not too hard perturbation leads back to the equilibrium state. Provided the adjustment process can be modeled

correctly, stable equilibria are particularly useful for computational purposes. Fourth, *efficiency* draws the attention to the effectiveness of resource utilization at the equilibrium. Many networks are characterized by the fact that the individual behavior incurs external effects. Congestion is a typical example for inefficiency. Fifth, when inefficiency grows to an unacceptable extent, one is interested in possibilities such as road pricing to alter the results. Changing the basic conditions of a network induces a series of reactions of the participants. For example, the supply side must take into account regulatory measures as well as the reactions of the network users.

Apart from short term equilibria, long term effects of any intervention are one of the major problems in network design. Network improvements affect not only the demand patterns for network services but can change the whole market situation. For example, an additional road, bridge or airport can shift location decisions of households and firms due to new accessibility conditions. The demand thus can be expected to increase as a whole but it grows asymmetrically. These effects are less emphasized in communication and distribution systems.

(b) Decisions

In accordance with the aforementioned behavior of network participants, in more detail, network users as well as network providers have to decide on a large variety of further aspects concerning network services. As it is essential for the market result which actor can select what action, let us distinguish again between the demand side and the supply side of the market.

Demand. In accordance with standard markets, many networks require a binary choice on participation which implies to buy some further devices such as telephones, cars, radios and so on. Afterwards the client decides on the amount of service he wants to consume at the prevailing market conditions. This amount can be divisible, lumpy, discrete or even binary. Just observe electric currents, durations of telephone calls, number of trips or radio access. Furthermore, depending on the market structure the client has to choose a service provider. In power supply systems the customers have to preselect one supplier. In telecommunications the selection of a service provider can be done call by call. Of course, for monopolies such as the German letter monopoly there is no choice left to the customers.

All of the above aspects can be detected in many markets. Some specific decisions regarding transport networks include (1) *trip choice*, (2) *mode choice*, and (3) *route choice*. The first case includes the binary choice whether to make a trip or not, as well as selecting the frequency of trips. Moreover, the traveler has to decide on the destination and the departure or arrival time of each trip. The second case refers to alternative modes of transportation such as using a private car, a public bus or the subway. Routing is a special question of the type of network at hand. In some cases the traffic pattern can be regulated by some central authority. For a broad class of transportation networks, however, the travel patterns are set up by individual users. Each traveler chooses the cheapest way to arrive at his destination irrespective of the implications on some aggregate system optimum. The wide range of solving the routing problem can be illustrated by the following list:

- Routing in railway systems with relatively low connectivity is essentially a question of the network topology. The railroad carrier then coordinates the demand for trips by suggesting synchronized timetables.
- With regard to road traffic car drivers make their own route choices. This process is more flexible particularly because there is less need to synchronize individual decisions than for railways. This flexibility is contrasted with the disadvantage that drivers tend to ignore the external effects they impose on the rest of the network. Hence, the resulting user optimal traffic assignment is suboptimal compared to an overall system optimum. This gives rise for thinking about a central authority which corrects the individual behavior by some traffic guidance system.
- Airline passenger carriers and parcel delivery networks take care of routing by developing so-called hub-and-spoke networks. At first, all flows are collected via spokes at specialized nodes, say hubs. In doing so the bundling of flows on the interhub links makes more efficient transportation technologies available to the carrier.
- For telecommunications systems, routing is a question to be answered by the network provider. Possible solutions depend on the way of transmitting data. The regular voice telephone network uses circuit-switched traffic, that is all (or fixed parts of the) resources on the communication circuit are unavailable for other parties. By contrast, most data networks are based on packet-switched traffic where small units of data are routed through a network based on the destination address contained within each packet. When packets arrive at a switching node they are stored and then forwarded at the full transmission rate as soon as the communication link is free. This method provides a more efficient utilization of network resources but requires more capabilities of the hardware.
- Due to physical conditions, power supply systems operate without routes. This is an important technical problem because the voltage over the entire network including all power stations and all power points must be constant.

Supply. Besides routing network providers have to decide on the *design* of their networks. Planning the network structure must take into account a list of further aspects. (1) The formation of nodes and their connections or in short the *network topology* determines the available variety of services. In terms of geographic coverage, for instance, a network extension may have the task to offer network access to additional customers. Similarly, it determines the potential interconnection with other (competitive) networks. (2) The capacity of a network indicates the maximum feasible throughput. This measure is hard to define for multicommodity networks. On the one hand it depends on the capacities of the individual network elements as well as their combination. On the other hand, any bundle of services has its own structure and, therefore, leads to different bottlenecks indicating different capacity measurements. (3a) The design features of a network determine also the quality of service (QoS), a collective measure of the level of service delivered to the customer. Although this term is closely related to the Internet, it has corresponding significance to any other network. QoS can be characterized by several basic performance

criteria, including availability (low downtime), error performance, response time and throughput, lost calls or transmissions due to network congestion, connection set-up time, and speed of fault detection and correction. Service providers may guarantee a particular level of QoS to their subscribers. (3b) Apart from the scale of a network and QoS, accessibility[8] is another attribute determining the attractiveness of a network. Redundant connections, for instance, do not only improve the reliability but also can make the network easier to use by increasing the number of alternative routes and dispersing the traffic flows.

Interaction. The behavior of both market sides results in a list of short-run and long-run effects. Altering the features of a network changes the demand for services. This effect of traffic generation or diminution is overlapped by problems of traffic diversion. In accordance with the new network conditions, the route choice (reassignment) and/or the choice of destination (redistribution) will be modified. Hence, it is important at least for the network designer to take the response of the network users into account. The estimation of such short-run reactions is already a difficult task but it is even more complicated to forecast long-term effects. Almost all location or relocation decisions (housing, productions, warehouses, hubs, etc.) are based on accessibility criteria which are determined by the prevailing network structure. Changes in these conditions can lead to new agglomeration areas with a partially excessive utilization of network resources. Furthermore, more or less accidental events such as the politically motivated location of an airport can make network planning an absurd waste of time.

(c) Traffic Representation

Describing the process of movements on a transportation network follows the purpose to be modeled. Independent of the items carried by the transportation system two major approaches in representing traffic can be distinguished.

The *microscopic view* lays emphasis on single elements, users, and processes. Each element or part of the network has its own functionality which is in most cases restricted to the technical properties of adjacent elements. For example, capacities limit the maximum flow on links and nodes. Every network participant pursues individual tasks. Drivers, for instance, are assumed to follow utility maximizing (or expenditure minimizing) rules resulting in an individually preferred free flow speed presumably disagreed by other trip-makers. Most processes consist of a conditional sequence of steps where the failure at any stage makes the whole process unusable. When such a process arrives at a device occupied by some other process it has either to wait in a queue or it dies.

By contrast, the *macroscopic view* models traffic as a continuum akin to fluid. In road traffic, for example, one can think of two strategies to increase the throughput of a road. Either all drivers increase their speed so that followers can enter the road earlier. Or the drivers move up closer according to a compressible fluid so

[8] Accessibility typically refers to the ease with which desired destinations may be reached and is frequently measured as a function of the available opportunities moderated by some measure of impedance.

that more trip-makers can use the road at the same time. The problem is that both approaches increase traffic nuisance. This effect is referred to as congestion and requires all drivers either to increase safety distance or to slow down speed. Another phenomenon is described by shock waves. They represent the boundary between traffic of two different densities and thus of different behavior.
In between the two views lies the broad area of aggregating individual processes to operational variables such as total demand for services during the busy hour. The combination of overlapping processes, however, requires detailed work on synchronization. Besides an appropriate scheduling, some of the synchronization in waiting systems can be done by queuing. At least for loss systems queuing is ruled out.

2.1.2 Network Technologies

(a) Production

The basic ingredients of a network are nodes and links each of which having their own functionality. The combination of these network elements must take into account technical restrictions and other properties such as sharing of network resources or queuing. In the latter case the start of an activity depends on the termination of several other activities though not all activities are able to wait in a queue. Hence, the way of composing network elements determines the performance of the network as a whole.

Combinations of Network Resources. At first, observe that networks and their constitutive elements can be complements as well as substitutes.

Complements The combination of consecutive links to one path emphasizes the complementary character of arcs (and nodes). When one element fails all routes including the particular element cannot be used until it has been repaired. Similar observations are true for combinations of networks. The interconnection of telecommunications networks has already been mentioned. Joining road systems, railway systems and airlines provides another example where traffic is originated and terminated in different networks possibly using another transit network.

Substitutes Parallel links or alternative routes show that network elements or combinations of them can serve as substitutes. This property is especially useful when some parts of a network are out of service. Maintenance work, for instance, needs redundant network elements to keep the network operating. Similarly, the selection of some service provider operating on his own network shows that networks as a whole may be seen as substitutes. Another example is that of train journeys which can usually be substituted by auto trips.

As it is common to all production systems, networks of nodes and links are used to transform inputs (time, energy, etc.) into network services (flows) by some production technology the details of which are treated as a black box. Nevertheless, constructing and studying simple models requires to enumerate and define the most

important variables and to specify the relevant relationships between them. One of the most essential parts of the transportation systems studied refers to the production of services. Hence, it is useful to go through a list of aspects which will be covered by the model in question or left out.

Factors of Production. Transportation systems are typically characterized by large indispensable amounts of *fixed (lumpy or indivisible) inputs* which usually cannot be resold. Just think of railway tracks, roads, fiber cables, telephone switches and so on. That may be different for other fixed inputs such as airplanes, locomotives, cars or even copper cables. Parking (and maintaining) airplanes in a desert is a remarkable case that they are expected to be reused in the long-run. Most of the above examples require maintenance work[9] as a fixed input per period. Maintenance itself may be seen as a problem of low analytical difficulty, but its importance rises from the fact that maintaining some device usually means to take this element out of service for some time. This problem could be ignored with respect to vehicles which can be substituted easily in most cases (rent a car). Maintaining switches, however, can imply that complete subnetworks are disconnected from the rest of the network. All activities that depend on the device in question must then be stopped.
Variable inputs depend on the amount of outputs produced. The most frequently used examples are time and "work" needed to move one item of given size one unit of distance at a given level of service, say speed for instance. As the physical work, that is power times distance, performed by the system has many determinants, it is much easier to measure the consumption of gasoline which serves as a store of available energy. Notice if the level of service is measured by speed, then increasing the level of service leads to a proportional decline of time need, but one can expect that the fuel use and the risk of accidents increase progressively. Last but not least, human work (taxi drivers, air traffic controllers, technicians, etc.) is usually measured by the time, say man hours, needed to perform services.

Outputs of Networks. A first classification results from the fact that networks provide either a single good (power or water supply systems) or many commodities. In the latter case each point-to-point connection (road networks or telecommunications systems) is interpreted as a separate service that shares network resources. The complementarity of network elements is reflected by output measures such as vehicle kilometers traveled per hour or year, call minutes per day, or kilowatt hours. This way of measuring network outputs is of minor importance for the consumers. They want to consume distinct network services in the sense of bulky composite goods, all other services are irrelevant. In other words, the composition of homogeneous service units may lead to a large variety of differentiated goods.

In any case product differentiation is a common strategy of network and service providers to establish some degree of market power. Network providers can set technical standards to make their network incompatible with other networks, cf. different gauges of tracks. Service providers use similar strategies of horizontal

[9] Costs to maintain existing infrastructure involves (1) ordinary maintenance, e.g., cleaning and winter maintenance, which is independent of road use and (2) maintenance, which depends on the volume of vehicles, e.g., surface dressing.

product differentiation to tie customers to themselves, cf. the supply of natural gas with different calorific values. Other examples refer to travel time, risk, comfort, reliability, etc. Furthermore, vertical product differentiation is common to almost all networks. Passenger carriers as well as packet delivery companies offer different levels of service to particular groups of customers and grant rebates to major clients.

Apart from product differentiation network services include two major classes:[10]

- communication between individuals or groups where the service "merely" consists of establishing end-to-end connections (telephone services, broadcasting programs);
- movements of passengers or cargo through an area (journeys or trips by different transport modes, parcel delivery, postal services).

Most services differ with respect to time and space. The *service time* has two essential aspects; it is related to the duration of service (cf. telephone calls, travel times) and to the point of time when the service is made available (departure times). Busy hours and particularly peak loads put high pressure on the given capacities. Contrasted with time we have to analyse *space*. Many economic activities are tied to certain locations requiring trips of different length. Distance is also a problem of covering space. Most networks serve in the beginning only agglomeration areas and leave out the periphery so that rural areas have no or only limited access. In contrast, many regulated network companies are obliged by regulators to provide universal services all over the space, cf. Deutsche Post AG. As a service in turn, these companies are protected against rivals which are not willing or able to provide universal services.

In the above sense *accessibility* can be added to the aforementioned classes of network outputs. Although there are various definitions of accessibility two major categories of indicators can be distinguished, cf. van Wee et al. (2001). (1) Infrastructure related approaches focus on the characteristics of infrastructure and on the flexibility of its use. For example, the density of networks is measured by total length of a network in relation to space, and coverage indicates the fraction of space or the nuumber of potential customers having access. Flexibility refers, for instance, to choice of departure time, speed on motorways, the opportunity to avoid congestion and so on. (2) Activity related indicators draw attention to needs such as living, working, recreation or shopping reachable within a certain time limit or distance.[11] One measure may be the population density, i.e., the number of inhabitants per square kilometer. The worth of accessibility becomes clear if we think of some network provider who rents out parts of his network with accessibility as one performance criterion.

[10] Nagurney, Dong (2002c) add economic and financial networks as a third class besides transportation networks and communication networks. They discuss two examples of supply chain networks, namely, the online grocer Tesco and the book retailer Amazon.com.

[11] Gravity models are used to include time and distance properties; they combine the attractiveness of some point or areal zone with a distance decay functions in order to explain spatial interaction patterns for example between central business districts and their periphery.

Some further aspects on the description of network outputs result from the fact that network services often represent *composite goods*. In most cases network services are the result of many sequential or concurrent steps. For example, a journey may consist of several subsequent trips possibly by different transport modes. It is then plausible when analyzing parts of a network to speak of intermediate goods. A railway station, for instance, can be used to initiate or to terminate a trip. But it also serves as a point of interconnecting trains in which case arriving and departing trains are interpreted as inputs and outputs, respectively. The arriving train provides a transport service needed to complete a set of different journeys. This observation leads to another problem when calculating the contribution of one network element to the provision of one unit of a final network service.

2.1 Remark (Quality of Service, QoS) Competition in the provision of communication services has forced the providers to measure their quality of service in an operationally meaningful manner and to guarantee predefined levels of service to their customers. The following list of QoS criteria applies to (tele-)communications systems (see Hardy (2001)), but it has corresponding meanings for other networks such as road systems, where trip times, risk of accidents, and reliability are of major importance.

Latency is the delay in a transmission path or in a device within a transmission path. In a router, latency is the amount of time between when a data packet is received and when it is retransmitted.

Jitter refers to the distortion of a signal as it is propagated through the network, where the signal varies from its original reference timing and packets do not arrive at their destination in consecutive order or on a timely basis, i.e., they vary in latency. In packet-switched networks, jitter is a distortion particularly damaging to multimedia traffic.

Bandwidth indicates the theoretical maximum transmission capacity of a computer channel, communications line, or bus. As the theoretical bandwidth is approached, negative factors such as transmission delay or signal-to-noise ratio can cause deterioration in quality, which corresponds to congestion in road traffic.

Reliability addresses the need of error-free communication, where each signal produces noise in its surroundings and deranges other signals. These errors can be corrected up to a certain amount. Beyond this point messages cannot be guaranteed to be delivered without errors.

QoS differs from grade of service (GoS) which addresses accessibility and availability of network services. Accessibility refers to the possibility of conditions that make it impossible to set up end-to-end connections normally supported through a telecommunications service. While accessibility is a measure of perceived QoS, availability measures intrinsic QoS which is defined from the viewpoint of the service provider rather than the service user. Availability is a ratio of uptime to total time. Network performance signifies not only QoS such as overall system delays but also throughput, both of which depend among others on the speed on the communication lines, number of hops, and occupation rates. □

Production Technologies. Networks exhibit positive and negative consumption and production *externalities*. A positive consumption or network externality signifies the fact that the value of a unit of the network commodity increases with the (expected) number of units sold; see Katz, Shapiro (1985) or Economides (1996). The benefits of a telecommunications network rise with the number of customers having access. This positive externality of an extensive network utilization is contrasted with congestion effects when the network is used more intensively. Almost all flow units disturb the traffic in their immediate surrounding. As traffic becomes heavier increasing flows use the same resources at the same time so that the mutual disturbances cannot be evaded. When congestion predominates, the corresponding flows are blocked and the movements are stalled. Congestion is usually a temporary problem due to substantial variability in demand. During most of the operational time of the system, that is outside the relatively short periods with peak loads, one has considerable excess capacities.

Both cases, congestion and excess capacities, indicate an inefficient resource utilization although sharing of a network resource is in principle possible. However, even uniform flows can cause severe problems when concurrent processes do not match. A typical strategy to *coordinate* road traffic flows is to control junctions by traffic lights and to limit the speed on the connecting roads. In an ideal (deterministic) world this procedure avoids congestion effects and on average all drivers can realize a certain speed. Another strategy is *synchronization*. In railway systems schedules represent the conditioning of events. A departing train must wait for certain incoming trains so as to allow passengers to change. By contrast, limiting capacities bind the number of trains waiting at a station. The preceding approaches show that stopping processes and storing them in waiting rooms until they are served cannot be avoided in any case. However, we have to distinguish

loss systems where telephone calls get lost if they are blocked by occupied facilities, and

waiting systems where incoming jobs are able to wait in queues. Customers may be patient and willing to wait but others may be impatient and leave after a while. For example, in call centers, customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while.

When *queuing* as a symptom of congestion occurs, the operator has to think about strategies how to minimize the waste of time of his clients.

- Queuing systems are understood as (1) a link connecting a service station with a subsequent (2) waiting room, where the job has to wait until it is handled by the following (3) server. Accordingly the time need consists of a propagation delay, a queuing delay, and a service time. As the propagation delay usually cannot be influenced, the waiting time in the queue plus the service time are referred to as sojourn time or lifetime of the job.
- When jobs (passengers, data packets, etc.) arrive they may have to wait behind other jobs. The arrival process introduces random delay. The length of the queue grows if the number of incoming jobs exceeds the capacity of the server. As the devices have finite buffering, arriving jobs are lost when the queue buffer is full.

- Queue disciplines determine the order in which waiting jobs enter service. Two examples are first in first out (FIFO) and service in random order (SIRO). Furthermore, priorization is often used to improve efficiency. Rights of way rule road traffic. In data networks different jobs can be privileged to use certain parts of a capacity. Jobs can even be rejected to reserve capacities for other jobs of higher priority.
- The relevant *performance* (or delay) *measures* in the analysis of queuing models are the distribution of the waiting time and the sojourn time of a customer.

Regarding data networks several *congestion management tools* may be implemented to handle an overflow of arriving traffic. They sort the traffic and determine some method of prioritizing it onto an output link. Analogous concepts are used in other types of transport networks.

First In First Out (FIFO) Queuing in its simplest form involves storing packets when the network is congested and forwarding them in order of arrival when the respective network element becomes free. FIFO is the default method of queuing in many instances, but it has several shortcomings. Most importantly, FIFO queuing makes no decision about packet priority; the order of arrival determines promptness and resource allocation. Moreover, it does not provide protection against ill-behaved jobs. Bursty sources can cause high delays in delivering time-sensitive application traffic, and potentially to network control and signaling messages. FIFO queuing was a necessary first step in controlling network traffic, but today's intelligent networks need more sophisticated approaches.

Priority Queuing (PQ) ensures that important traffic gets the fastest handling at each point where it is used. It was designed to give strict priority to important traffic. PQ can flexibly prioritize according to traffic mode, incoming interface, packet size, source/destination address, and so on. In PQ each packet is placed in a separate queue based on an assigned priority. During transmission, the algorithm gives higher-priority queues absolute preferential treatment over low-priority queues. This is a simple and intuitive approach, where higher-priority traffic is usually rate limited to protect lower-priority traffic against distortion or disruption.

Custom Queuing (CQ) guarantees bandwidth. CQ was designed to allow various applications or organizations to share the network among applications with specific minimum bandwidth or latency requirements. In these environments, bandwidth must be shared proportionally between applications and users.

Weighted Fair Queuing (WFQ) is a queuing method for situations in which it is desirable to provide consistent response time to heavy and light network users without adding excessive bandwidth. It is a flow-based queuing algorithm that does two things simultaneously: it schedules interactive traffic to the front of the queue to reduce response time, and it fairly shares the remaining bandwidth between high bandwidth flows. WFQ ensures that queues do not starve for bandwidth, and that traffic gets predictable service. Low-volume traffic streams – which comprise the majority of traffic – receive preferential service, transmitting

their entire offered loads in a timely fashion. High-volume traffic streams share the remaining capacity proportionally between them.

Each queuing algorithm was designed to solve a specific network traffic problem and has a particular effect on network performance.

Examples. The extremely different inputs, outputs, and production technologies can be explained best by giving some examples.

Water resource planning or *hydraulic systems* represent a relatively easy case in the sense that they distribute[12] one homogeneous and perfectly divisible good (water, gas, oil). Hence, routing has no importance; merely sources and sinks are of interest. Notice, however, that natural gas with different calorific values cannot be mixed. The nodes of such networks correspond to production platforms of limited number, pumping stations, reservoirs or lakes, which are connected by pipelines. Typically these networks have low connectivity implying that failures of single elements may require to shut down large parts of the network for a relatively long time; that may be weeks for gas networks.

At first glance *power supply systems* seem to be very similar to hydraulic systems. The decisive difference, however, is that electricity cannot be stored without very inefficient conversion. Furthermore, there are various technologies using different energy sources such as gas, fossil fuels, nuclear power, or hydroelectric power. Gas turbines in particular have low capital but high generation costs, while nuclear power stations have the converse properties. Because nuclear power systems are inflexible in the adaption of a varying demand, gas turbines are mainly used for peak loads. Note also that the use of hydroelectric systems – with low generation and capital costs – depends strongly on the multi-purpose of water and on water inflows. Power supply systems usually consist of many power stations and the network is highly meshed so that failures have on average less drastic consequences.

Television or *radio* companies broadcast a variety of programs over their networks. The transmission requires either a network of cables or follows wireless via air or satellite. In principle many customers receive the same service from one sender. Of course, the quality can vary depending on the hardware used. Radio networks are typical loss systems in the sense of queuing theory, although complete transmissions can be stored by using recorders.

Telecommunications systems are multicommodity networks as they connect a large variety of origins and destinations. In a communication network, nodes represent origin and destination stations for messages (also computers, satellites, switches, etc.), and arcs represent transmission lines (copper cables, fiber optic links, microwave links, etc.). Flows include voice messages, data, video transmissions, etc. As has been mentioned above, the regular voice telephone network uses circuit-switched traffic, but modern solutions use more and more often store-and-forward switching which admits a more efficient utilization of network resources. In any case routing and coordination of traffic are tasks of the system operator. Customers merely await that telephone calls will be established and held over a certain period of time, where the quality of transmission must be acceptable. In data networks, transmission

[12] Water supply systems are to be supplemented by sewage disposal systems.

has different further requirements on quality: completeness, transmission speed, minimum bandwidth, maximum delay, reliability, security, etc.

Railway systems are multicommodity networks in the above sense of telecommunications systems. Moreover, product differentiation leads to many levels of service: first and second class, slow and fast connections, etc. The nodes represent stations, switches or rail-yards and the arcs are tracks. (The transport network of *inland navigation* mainly consists of ports and shipping canals.) The flows correspond to passengers or cargo loaded on trains with maximum capacity. As railway systems have low connectivity there is only a limited number of meaningful routes for each origin destination pair. At the same time the concurrent and sequential processes require high efforts of synchronization. Some flexibility remains with regard to waiting systems. Even impatient passengers are able and willing to wait as long as there are no other more attractive alternatives.

The *airline industry* is very similar to a railway system. What makes them different in reality is the market structure. Most airports are under the control of special companies selling – among others services – time slots in which this network node can be used by the respective carrier or airline. Furthermore, the air traffic control coordinates the aircrafts using the same airspace at the same time to ensure safety.

Road networks are much more flexible. At least partially they can be used simultaneously by many different transport modes (pedestrians, bicycles, cars, buses, trucks, etc.) with extremely different characteristics. The nodes are given by junctions, parking places, service stations and so on. Trip-makers individually decide on modes, routes, departure times and speed including take overs possibly at the risk of oncoming traffic. Disturbances are omnipresent and a particular problem at junctions with self-coordination or traffic lights. A long list of rules exists to ensure safety (e.g., drive on the "right" side) besides operational motives.

Theoretical Implications. All networks have in common that they consist of nodes which are connected by links. These links may be undirected edges or oriented arcs. Both, nodes and edges, are usually characterized by capacities imposing upper bounds on the respective flow per period. Furthermore, weights such as travel time or length are assigned to the branches. Given a predefined level of service, the weights together with the utilization of capacities will determine the link cost per flow unit.

The main problems result from the fact that many processes are performed by the network at the same time. Some of these processes pass successively, while others occur simultaneously. (1) Splitting a process into a sequence of consecutive jobs requires that certain conditioning activities are finished; initiating a job also requires free and operating facilities; and eventually the output of a process needs either free storage capacities or access to the subsequent device. (2) Many processes such as trips overlap so that concurrent jobs may use the same resource at the same time (passengers in a train). When the device in question can handle only one job at a time, jobs have to wait in queues until they can be served one after the other (landing aircrafts). Such processes need typically some sort of cooperation (safety distance) or coordination (drivers at junctions). When the restrictions due to overlapping become stronger more effort is needed for the synchronization of processes (time tables for trains).

Modeling the production structure is the first major problem in network analysis. In view of microeconomic theory the task is to develop a construct similar to the "production function" of a firm producing a single homogeneous good. Such a production function describes technically efficient activities where the maximum output is achieved at given inputs and/or the given output is produced by minimum inputs. Although it can even be difficult to detect efficient activities for firms producing one homogeneous commodity, there are a lot of essential obstacles regarding network technologies which require a broader approach. Most networks produce a large variety of outputs by a large number of concurrent processes. A substantial part of the factors of production is indivisible or at least lumpy; take, for example, a bridge of fixed length where the capacity can be varied within limits by the number of lanes. Durable inputs require far reaching investment decisions resulting among others in huge fixed capacities which are idle except for temporary peak loads. This is true for almost all parts of networks, cf. cars, fiber cables or airports. Besides investments and a substantial variability in demand further aspects of time grow in importance. When many interacting processes make use of the same network they need synchronization in order to avoid a waste of resources. Furthermore, technical progress is hard to realize in the sense that networks with an old technology may be incompatible with recent developments. It is not unusual to observe a mixture of old and new techniques (telecommunications systems) or incompatible systems side by side (railway systems). In view of all these network attributes it takes no wonder that many efficiency problems cannot be solved particularly for large scale networks.

Throughout the book multicommodity flow problems will be in the center of interest even though networks producing a single output are discussed for explanatory purposes. Multicommodity networks generate a wide range of services where each commodity will be interpreted as connecting pairs of not necessarily adjacent nodes. For example, in road traffic a trip from one location to another leading through different roads and junctions represents one commodity and the collection of all commodities is referred to as trip table. In telecommunications systems each commodity corresponds to a so-called end-to-end connection.

As technical progress goes beyond the scope of this book, the analysis focuses on the factors of production needed to generate a bundle of services at a given technology. Three major classes of inputs will be distinguished:

fixed inputs with regard to the topology of the network, e.g., roads, bridges, airports, switches, routers, etc.;

fixed inputs referring to the provision of services, e.g., cars, trains, airplanes, etc.;

variable inputs depending on the amount of service, e.g., fuel, electric currents, time, etc.

The first two classes indicate the *investment problems* of the network and the service providers. The main problem, however, is to determine the contribution of the above inputs to the services produced. This may be relatively easy in the case of variable inputs but at least for network components it is hard to assign them to specific services. As most network components accommodate many processes and every process uses a sequence of components it is useful to think of intermediate goods

as a fourth class of inputs. Then, each commodity can be interpreted as a composite good consisting of a series of services produced by the individual components of the network in question.

Decomposing networks this way allows to analyze the performance of each network element, cf. the capacity of a single road. The problem is now to go one step back and to analyze the cooperation of all elements constituting the network possibly with respect to some kind of "optimal" network utilization. One example is to detect the maximum throughput of a network when all link capacities are known. By intuition, the maximum flow is determined by the some "bottleneck" usually consisting of more than one network element. Just think of several parallel congested bridges connecting two parts of a town limiting the flow from one district to the other.

Flows on a network are analyzed similar to networks and its components. Each flow unit passes through sequences of consecutive links and nodes where many routes may be available. In doing so it disturbs other flow units in its immediate surrounding. Although services supplied by road systems are sometimes referred to as public goods[13] having the property of non-rivalry, networks show a limited capability of serving flow units which simultaneously consume the same network service. Furthermore, flow units produce "noise" in the sense of external effects. When traffic grows heavier the disturbing effects induce congestion such as heat increases the resistance of a conductor. Depending on the type of network, processes in a network can be synchronized in different ways. (1) The individual flow units coordinate the behavior on their own, cf. road traffic with individual choice of routes, safety distance, speed and so on. (2) A carrier organizes flows in order to improve the utilization of given capacities, cf. railway systems with appropriate time tables or telecommunications systems using routers for an appropriate traffic assignment.

Finally, the extent of a network can change its functionality. For example, expanding a network may lead to a more uniform utilization of indivisible resources so that excess capacities can be reduced. *Economies of scale* are also realized by hub-and-spoke networks or other hierarchical networks. The idea is to bundle flows on the interhub links so that carriers can use for instance larger aircrafts. *Other* so-called *network effects* go beyond pure technical aspects. Network utilization frequently makes only sense if there is a minimum number of processes. Two examples illustrate this network characteristic. (1) The benefits of using a network increase with the number of customers or with the extent of the network. This is the case of an increasing number of clients in a telephone network. *Metcalfe's Law*[14] suggests that the usefulness, or utility, of a network equals the square of the number of users. According to Varian (1999), consumers would generally like to be connected to as large a network as possible. This implies that if there are several different providers of networks, then it is very advantageous to consumers if they interconnect. Of course, there is a countereffect when too many network users have to share limited resources

[13] Diewert (1986) defines infrastructure services to be public goods when all users have access to them at no charge.

[14] Robert Metcalfe founded 3Com Corporation and designed the Ethernet protocol for computer networks.

so that congestion takes place. (2) Many networks such as railway systems require an extremely large capital stock so that excess capacities decline to an acceptable degree only at a very intensive utilization.

(b) Valuation of Inputs and Outputs

Having discussed the production of transport services we need prices for inputs and outputs to calculate the corresponding costs and revenues. The classification of network providers and service providers suggests to decompose network costs as follows. Due to huge capital stocks a substantial part of cost of a network provider is fixed. We shall come back to this point when discussing network design and pricing network services. Instead we focus here on the variable and fixed costs of a service provider.

Cost. This distinction is immediate for road traffic, where the cost of making trips by car includes two components. The variable cost of driving some distance results from inputs such as fuel and travel time multiplied by their respective prices. By contrast, the decision to participate in the network causes fixed cost in the form of purchasing a driver's license and buying a car. Furthermore, insurance is usually compulsory for every driver. Similar observations are valid for railway systems. Time, however, involves here the working time of conductors as well as the travel time of customers including waiting times and delays. Both groups of people are probably related to different prices of time. This problem becomes even more severe in air traffic where airlines have to buy time slots to use the resources of an airport.

While many inputs can be evaluated (e.g., human work, fuel, electricity) at market prices, it is harder to determine the cost per unit of output with respect to durable inputs. With regard to one period of time, vehicles may have different capacities to transport passengers or other freight at different levels of service. So what is the adequate price of one capacity unit for peak and off-peak periods? Assume we can find appropriate valuations of these restrictions, then further problems must be taken into account. Many networks suffer from congestion due to negative external effects. For a constant level of service this fact does not simply mean that the capacity of a network element is used up, but that flow units tend to disturb other units in their surroundings. In doing so the level of service goes down as the traffic grows heavier. Economists are then interested in an appropriate system of prices which reflect the external effects ignored by the individual network users. These external effects include also the risk of accidents[15] that drivers seem to underestimate systematically. For road networks in particular one further cost category results from road damage that is strongly related to axle weight. Thus user classes should not only pay for the distance traveled but also for the class specific damages. In other words, the costs of road maintenance increase with traffic volumes and especially with the number of heavy weight trucks.

[15] As pointed out by Vickrey (1968), the National Safety Council estimated the cost of accidents in the United States to be roughly the same magnitude as the entire annual cost of providing the roadways on which these accidents occur.

Apart from external effects within the network (congestion) there are other external effects emanating from the network on the rest of the embedding economy (pollution). Running a network makes use of environmental resources without explicit property rights, that is at zero cost. Aside from prohibitions by the government there are different instruments to internalize these costs to the network in question. For example, protective measures against noise pollution can be imposed on the emitter by the government, cf. aircrafts, cars and so on. Of course, one can think also of solutions by private negotiations when allocating the respective entitlements by different legal rules. According to the polluter-pays-principle, taxation can also be an appropriate measure if environmental pollution depends on the amount of network services consumed. For example, a tax increasing the fuel price directly relates the used environmental resources to fuel consumption. Observe, however, that many social costs are highly localized and are amenable to local pricing rather than to a network wide cost-matching scheme. Another measure to allocate property rights by the state are licenses. In telecommunications systems, for instance, there is only a limited number of frequency ranges each of which can be used by one supplier. The state solves this problem by generating property rights through licenses which may be sold by auction or awarded in some other way; cf. licenses for Universal Mobile Telecommunications Systems (UMTS) on the market for high-quality wireless multimedia communications.

Revenue. In accordance with the wide range of network types and network services there are many ways of billing customers. Ignoring periods of peak demand, easy cases are that of water or power supply systems with a standard quality of the homogeneous commodity in question. All we need to know are the volumes delivered to the customers and the price per unit of the good.

In contrast, all multicommodity networks have the problem of specifying a particular service. Besides volumes (number of persons, freight, data), distance and duration grow more important; just compare the price of a railway ticket with the price of a telephone[16] call. Furthermore, the point of space and time and the level of service determine what we call a unit of output. Many service providers demand higher prices during periods of peak load (time slots at airports) to account for the fact that peak flows impose the most pressure on local capacities. In other cases they reduce prices for periods with low demand and corresponding excess capacities (night fares). Supplying different levels of service are common practice by almost all passenger carriers (tickets for the first or second class) and package delivery networks (express or ordinary delivery).

Broadcasting network and road systems have a further problem in common: at least to some degree it is difficult to exclude customers from consuming their services. There are different solutions for television or radio companies to earn revenues. Apart from selling time slots for advertising purposes, they can be allowed by the state to charge duties as soon as a household acquires a receiving facility or they rent out special decoding hardware to assure the excludability of their services.

[16] Telecommunication companies establishing end-to-end connections also have to decide whether the calling or the called party pays.

The same problem applies to road traffic. Excludability may be guaranteed for taxis, buses or subways with corresponding fares, but the general utilization of most public roads is usually free. Of course, one can restrict the usage of turnpikes by (expensive) tollgates, nevertheless it is common practice to finance publicly provided roads by different tax instruments. Besides the general tax income, there are special motor vehicle taxes as well as taxes on gasoline. The first act like a permission to use the road system with the vehicle at hand, while the latter are oriented to the amount of consumed services.

Notice that taxes on transport services pursue also other tasks besides realizing tax income. In the presence of external effects Pigou taxes are used to implement incentives which make network users behave in some optimal way. We shall come back to this point below.

(c) Objectives

The aspects so far observed are now summarized by different aims pursued by the network and service providers. All of these tasks have in common that they require the efficient usage of scarce resources. That is, we are only interested in economic activities which can unambiguously be regarded as non-wasteful. Some exemplary objectives refer to routing or traffic guidance

- minimum distances or times between distinct sinks and sources (shortest paths), minimum number of nodes (min-hop routing), or minimum overall delay,
- maximum throughput for distinct origin-destination-pairs or for the network as a whole.

Having selected technically efficient "production plans" out of the class of feasible activities, the next step is to choose one activity which is optimal in an economic sense.

In privately owned networks it is usually assumed that providers pursue the task of profit maximization. This necessarily requires an efficient production plan realizing, for instance,

- minimum total transport cost for given traffic volumes or
- maximum revenue given the inputs.

Bear in mind that the network services must match the needs of the customers. In particular competition forces carriers to provide reasonable levels of service. As an example, telephone systems seek to minimize the blocking of calls. Data networks are particularly interested in a high tolerance on transmission errors. Package distribution companies compete by "transmission times", cf., e.g., guaranteed on-schedule delivery. Passenger carriers try to minimize the overall system delay, where the major components of delay to aircrafts or rail cars in passing through airports or yards are waiting for classification and connection to other means of transportation. One major task is to determine time tables that are optimal with respect to minimizing passenger waiting times.

Publicly owned network monopolies work in principle in the same way. They seek for some efficient activity, however, with the objective to maximize welfare.

Introducing a social planner endowed with coherent objectives in the form of a social welfare function has at least two aspects which should not be mixed up. *First*, there are real world examples (road systems, postal services, power supply systems) where the state operates networks because he believes that he can provide solutions being superior to any private outcome. One argument is that the supply of material infrastructure as a precondition for other economic activities is "too low" due to several external effects. Another argument refers to the huge capital stocks that cannot be installed privately but by the state. Furthermore, many networks have the characteristics of a natural monopoly and the state claims to be the only who that protects the customers from abusing market power. An assessment whether these arguments hold true or not goes beyond the scope of this book. Instead we shall focus on the *second* aspect of welfare maximization. Constructing ideal models of networks allows to derive reference cases which can be compared to alternative network structures. If the individual behavior prevents the realization of a system optimal solution, the reference case can be used to analyze reasons for the differences and to find ways and means to remedy possible faults.

To sum up the aforementioned arguments of assessing network processes we refer to Friedman, Boorstin (1996, p. 228): *At first glance, it seems hardly possible that the apparently trivial problem of how to charge people for the highway services they use is a key to the whole problem of how to plan and pay for better highways; yet it is just that. This fact cannot be too strongly emphasized. It is a key not only for a system that would involve operation of roads by private enterprise but equally for the present system of public operation. Should a particular road be built? How should it be built? How should it be financed? Should an existing road be maintained, improved, or allowed to deteriorate? If we could charge directly for the service of the road, we could answer those questions whether under private or public ownership.*

2.1.3 Construction of Networks

The explanations above have drawn the attention to the provider of network services and neglected the problems of "operating a network". This approach has the advantage that the consumed resources are relatively closely related to the produced services so that the variable cost of producing one unit of output can be calculated. Although service providers do have fixed costs they are of minor importance in relation to the fixed cost of network carriers which are mainly characterized by large capital stocks. This observation can be explained best when interpreting networks as a collection of nodes and links connecting them. The corresponding capital stocks then consist of nodes such as railway stations, airports, ports, switches, routers as well as connections such as roads, bridges, subway and railway track, shipping canals, cables. The main problem is now to determine the contribution of a network as a whole to one unit of output produced on the network in question. We shall come back to this point in the next section with regard to an appropriate pricing of network services.

Network design is a question of what services, respectively their volumes and qualities, the network should be able to produce. Multicommodity networks, for in-

stance, provide a wide range of connections between nodes of origin and destination. The flow on each connection is usually restricted by the capacities of the network elements involved. Moreover, many flows have to share the same network resources. Notice that quality such as speed or reliability of transmission is determined by features of the network as well as the flow volumes accommodated by the network at the same time. When congestion takes place, most networks give modified answers to the question of shortest, fastest, cheapest or most reliable routes.

One of the most important problems when adjusting networks to continuously varying flows is that almost all network elements, namely, nodes and links, are indivisible. In words of Baumol (1987), *to carry any freight from New York to Chicago a railroad must lay at least two rails, each about 1000 miles long.* Nevertheless, the capacity of such a connection can vary within limits by designing it for express or ordinary traffic. These observations suggest to distinguish at least two classes of network design problems, see Bell, Iida (1997, Chap. 9).

1. Discrete network design lays emphasis on indivisible projects. It is concerned with the topology of a network, that is the number of nodes and how to connect them (minimum spanning trees, Steiner trees, fully meshed networks, etc.).

- In a greenfield or, more drastically, scorched earth approach, network design starts at least theoretically from scratch. Some tasks are to find the optimal number of nodes, their optimal locations as well as their optimal connections. As an example one may think about the installation of new communication systems such as UMTS.[17] Of course, scorched earth models can also be used to revise existing networks.
- Within an existing network incremental nodes or links may lead to the provision of new transport services. This is obvious with regard to additional airports, roads, rail tracks, tunnels or bypasses. A less apparent fact is that existing flows will be redirected because more efficient routes become available (traffic diversion) and that additional traffic is loaded to the old network (traffic creation).

Bear in mind that the way of connecting nodes – chains, stars, rings, traverses, fine or big mesh – determines much of the functionality of a network. Among other aspects, discrete network design must be aware of failing network elements. Hence, a redundant bypass may serve not only as a short cut but also as a safeguard to guarantee the survivability of a network. One result of discrete network design is an hierarchical structure of almost all network classes, where parts of the network suit to special functional requirements. That may be the provision of different levels of service (high-speed vs. low-speed connections), bundling of flows (hub-and-spoke networks), failsafe performance (special security measures for backbone networks), or areal particularities (local area networks (LANs) vs. wide area networks (WANs)).

2. Continuous network design takes the network topology as given and is concerned with the parameterization of the network. Parameters include link or node

[17] With the development of wireless telecommunication networks further problems arise due to the permanently changing topology. Think, for example, of festivals where lots of customers meet for a relatively short period with a peak demand at that location.

capacities as well as user charges. One task is to maximize the network throughput at a given level of service by removing bottlenecks or synchronization of concurrent and sequential processes. Specifying the number of lanes or signal timing are two examples regarding road traffic. Sensitivity analysis helps to determine effects on traffic equilibria due to small changes on the supply side of the network.

Both categories, discrete and continuous network design, include quality improving and cost reducing investments. The projected Gotthard railway tunnel, for example, has the tasks (1) to increase the capacity of the existing connection, (2) to reduce the travel time between Zurich and Mailand from four to three hours, (3) to reduce travel cost by avoiding the current difference in altitude, and last but not least (4) to increase quality of life in the narrow valleys of Switzerland by redirecting freight traffic from road to rail.

A cost minimal network design must not deceive the designer that a trustworthy network operability requires usually measures against unpredictable failures of parts of the network. An appropriate network design (including redundant infrastructure) may help to increase network reliability by reducing the consequences of random failures of network components. One typical cause of failures is long-term wear and tear not only for road systems (implying road works) but also for telecommunications and electronic networks. *Network reliability* aims at the modeling of failure mechanisms in order to come up with failure rates that can be expected to increase in time and traffic intensity, see also Ball et al. (1995a). Vulnerability of a network refers also to the design of *survivable networks* that are still functional after the failure of certain network components so that the surviving traffic can be rerouted, see Grötschel et al. (1995). Network failures do not necessarily mean that equipment gets broken. An ordinary example for network components being out of service is periodical maintenance work. Observe, however, that disasters such as floods or earth quakes can destroy large parts of a network, where further strategies are needed to regain a minimum network performance. A less drastic example is snowfall, where road carriers are interested in optimum routes for clearing vehicles to obtain a minimum functionality within a given time limit.

To sum up, transport networks accommodate the demand for transportation derived from economic activities that are scattered mainly through space but also through time. In the short run it is reasonable to assume that the spatial activity pattern as the origin of traffic demand and the network topology, which determines the supply side, remain constant. Accordingly, short-term network design problems essentially deal with the question of how to adjust capacities such that the drop in performance due to congestion during peak-periods can be reduced to an acceptable limit. This is in fact the quest for an efficient production activity. Note that particularly indivisible network components can imply substantial excess capacities for off-peak periods. On the opposite side, we have to account for long-term problems, where the structure of activity patterns can change. For example, one might think of an expansion of suburban living due to increased mobility that follows on the one hand from network improvements and requires on the other hand further investments in road infrastructure. Consequently, network designers want to forecast land use patterns including location decisions of households and firms in order to predict

future traffic demand. This is necessary in order to determine appropriate network adjustments that suit best to the requirements of demand for transportation. In privately owned networks, all aspects of network design are in the responsibility of the network carrier. For publicly owned networks, however, official decision-makers have to determine the economically justified needs for road construction or other investment.

The problem what the highway network ought to be in order that it is of maximal benefit to the public, is a long-run economic problem with a high order of complexity. The main difficulty is that the demand responds to the properties of the network including the prevailing traffic conditions. Thus almost all network modifications induce reactions of customers such as traffic diversion (choice of alternative routes or transport modes) and traffic generation (introduction of a new fast transit line). This problem of transportation planning is by far too complex in order to solve it in one step. Thus, network designers are interested in decomposing that problem into smaller but interdependent problems. One approach which is referred to as bilevel programming will be introduced in Section 4.3.2. The idea is to determine – in a first step – optimal network improvements given the demand for transportation. The second step computes the responses of the demand side. With that we go back to the first step in order to revise the investment program. The aim is to find a computational procedure that converges to a unique statement on how to expand the network.

A final remark underscores the problem of financing network infrastructure. Even the development of existing network structures can easily exceed the financial capabilities of a single network provider due to enormous indivisible investments. This is particularly the case where today's cost face uncertain future revenues. Till this day the supposed weak capabilities of private firms serves as an argument for governments to be the only instance that can provide a sufficient amount of infrastructure. Nevertheless, the governments' notorious lack of resources led to several forms of outsourcing, commercialization, and privatization; see Gibson, Pavlak (2002). To give an example, Build-Operate-Transfer (BOT) is a special form of private public partnership where a private enterprise builds the infrastructure, operates it for a specified period, and then hand the ownership of the facility over to the public sector. Such investments are usually paid for by direct tolls on users. In an alternative setting, however, one might also think of some governmental agency that pays so-called "shadow" tolls to the private road-builder by means of motor fuel tax based on achieved traffic counts. Contrasted with BOT, the IRT and BMT subway lines in New York City were paid for by the city authorities and then turned over to private companies.

2.1.4 Pricing of Network Services

(a) Competitive Pricing

The determination of prices for network services is basically the same as in any other market, where the pricing mechanism coordinates demand and supply of a good in order to encompass market clearing. In the well behaved case of an ordinary

market for network services, competition makes sure (1) that services are produced without waste of resources, (2) that network performance suits to the needs of its customers (consumer sovereignty) and (3) that customers pay the full cost of consumed services. A (partial) market equilibrium may then be identified as an outcome of a market economy in which each agent is doing as well as he can given the actions of all other network participants including network carriers, service providers and consumers. One prominent counterexample for cooperation of network and service providers is interconnection, where parts of one service are produced by competitors. Furthermore, backlashes to other markets may be severe, especially when considering network services as intermediate inputs for other activities, but they are ignored for the moment.

Competition may be interpreted as the most important concept to discover methods of how to make efficient use of scarce resources. Regarding network markets, competition takes place between various groups of agents and institutions. There are network carriers, service providers, and consumers all of which strive for something that all of them cannot obtain. As an example, one may think of telecommunications companies that compete, on the one hand, for a limited number of frequency ranges and, on the other hand, for customers, where technological progress[18] is one instrument to improve the level of service delivered to the customers. For urban transportation networks, competition is reflected by several choices left to the travelers. Having chosen their locations[19] on the network at hand, residents decide whether or not to make trips at what frequency. Although most trips are induced by journeys to work at a given location, many trips require a choice between competing destinations, cf., e.g., shopping centers. For many journeys, different transport modes – such as bicycle, auto, bus, train, aircraft – are available; they may be related to each other as substitutes or complements. At least trips by bicycle or car require a decision on route choice, where especially motorists compete for local network resources they want to use at the same time. For further details see Section 2.3.5.

(b) Network Related Difficulties

Real world examples, however, indicate that many network markets suffer from missing basic conditions that are necessary for the market's ability to operate. It is then common practice to accuse the "market" of failing not to find an acceptable solution. A brief list of known problems is given as follows.

- Publicly owned road networks have the deficiency that it is hardly possible to *exclude free riding persons* from utilization at reasonable cost. Today, however, the technical difficulty of charging for the use of roads may be seen as dissolved in the light of modern electronic systems of toll collection. Nevertheless, the

[18] In principle there can be many networks with different technical standards. Such technical features may be used by incumbent firms to protect market segments against potential rivals. In many cases, however, networks using different techniques can be interconnected, see for example fixed-line and cellular telecommunications networks.

[19] Competition among districts or regions for inhabitants or firms is not in the center of interest for this book.

political resistance against such systems seems to be unbroken even though several attempts of road pricing exist. Observe that in networks such as railway systems or telecommunications networks where the exclusion of free riders can be managed, the problem of pricing network services has been solved long ago.

- Although it is well known that road damage depends heavily on axle weights, the political process is not willing or able to introduce a distance-weight metering for road networks usually with the spurious argument not to compromise the competitiveness of hauliers.
- Transport networks show at least two classes of (negative) *external effects* the costs of which are borne by third parties.

  User-upon-user externality is reflected by traffic congestion where each driver imposes – time and space dependent – reductions in the level of service that can be consumed by other network users. Bear in mind that pure rivalry in consumption (i.e., consumption of network resources) is common to almost all markets, but limiting the performance of the network as a whole indicates an externality internal to the network. Such externalities are not road specific, for almost all networks have to reduce their level of service as traffic grows heavier. Observe for road traffic that not only delays but also accidents are more likely to occur when traffic intensifies.

  User-upon-non-user externalities refer to environmental damages such as noise, vibration or air pollution imposed by network participants on non-traveling residents of an area. This is essentially a question of adequately defined property rights. Without enforceable property rights, Pigouvian taxes are one way to oblige consumers to internalize their externalities. As has already been mentioned, one disadvantage with these taxes is that many social costs are highly localized; such neighborhood effects refer to drivers vs. residents or airports vs. their environments. Consequently, we would need a local rather than a network wide pricing scheme.

  Of course, any method of optimizing user-upon-user externalities may also have the desirable side effect of reducing air pollution and noise.
- As has been mentioned above, many network industries possess to some degree the characteristics of a natural monopoly. A single-firm production is advantageous for consumers as long as they can profit from the firm's superior efficiency. Nevertheless, it is likely that the monopolist uses its market power to exploit its consumers by supplying a lower level of service at increased prices compared to the competitive market outcome. Excessive transmission fees can also be used to discriminate (potential) rivals, where two examples among others refer to the market for telecommunications services and to electricity markets.
- There is a variety of public instruments (and mixtures of them) to impinge on the pricing of network services. In the easiest case such as an urban subway, passengers pay *fares* to travel on a public (or private) conveyance in order to finance the supply of that service. For road systems, where excludability of road usage is hard to enforce, *tolls* must be paid for the privilege to use particular roads or bridges. The rest of the network is financed by *taxes* such as motor fuel taxes, vehicle taxes and other taxes accounting for the government's general tax

income. Although network users should be charged according to the services they consume, merely motor fuel taxes are related to the distance traveled. Even this instrument ignores weight dependent road damage, it does not distinguish between distinct network resources, and the travel time has no importance. When public networks are financed by general tax income, even those residents who do not consume network services have to pay for them. Nevertheless, it is likely that especially these persons suffer from environmental damages.

An approach that solves most of the above problems would require some sort of *element based charging* where consumers pay for services according to the full cost of consuming a sequence of network resources. With special regard to external effects of transportation, one such distance and time based road pricing trial in Dublin is reported in O'Mahony et al. (2000), where the potential user response to road pricing is evaluated for a sample of individuals drawn randomly from a group of volunteers. The road use pricing charge levels were selected to match the marginal external costs of car transport, i.e., those costs not currently paid by the car user. Such costs include marginal external costs of congestion, air pollution and noise.

Apart from private trip costs (mainly caused by the consumption of fuel and time) and external costs, network users must also pay for the provision of networks. This statement refers not only to construction costs but also to network operating costs including costs of maintenance, cf. road damage, sweeping, snow removal, etc. One problem of network design is then to assign the tremendous costs of durable and indivisible network components with excess capacities during most of the time to distinct network services. This is no network specific difficulty, but its importance results from the fact that users do not perceive those costs they incur. The European Commission (1995, 1998) concluded indeed that *one of the important reasons why imbalances and inefficiencies have arisen is because transport users have not been adequately confronted with the full costs of their activities ... As prices do not reflect the full social cost of transport, demand has been artificially high. If appropriate pricing and infrastructure policies were to be pursued, these inefficiencies would largely disappear over time.*

As individual trip related decisions depend on perceived travel cost we shall distinguish two approaches, namely, the superior situation of *system optimality* and a restricted solution called *user optimality*.

System Optimality A system optimal traffic pattern represents the global viewpoint of a network provider who is interested in the efficient usage of network resources. In doing so a private or public network carrier would arrange goods movements – possibly by virtue of some sort of peak load pricing – so as to internalize external costs of congestion. Nevertheless, negative effects external to the network are not necessarily taken into account. Thus, a first-best traffic pattern requires further instruments to internalize the latter external effects.

User Optimality A user optimal traffic pattern refers to the individual point of view of each motorist who tends to be unaware of, or at least to be unresponsive to, certain costs he imposes on others. Hence, the driver will rationally not take into account these costs in his travel related decisions. Drivers are only concerned

with the private cost they must bear themselves such as petrol costs, the time costs of making the trip and maintaining their vehicle. Consequently, motorists tend to underestimate the social cost of trips that should include all impacts of their activities on others.

(c) Road Pricing

One of the best-studied applications of Pigouvian taxes is road pricing. The economic fundamentals were laid out by Pigou (1920), Knight (1924), Walters (1961), and Vickrey (1969), among others. Pigou (1920) introduced the idea of a congestion toll by presenting the famous two-road example. He postulates that one road is wide enough to serve the whole traffic without congestion, but is also rough and slow whereas the other is narrow and likely to be congested, but smooth and fast. In accordance with what we know as Wardrop's first principle, Pigou argues that traffic will distribute itself among the two alternatives routes until the travel time is the same for each user. Imposing a toll on the traffic using the congested road would allow to divert excessive traffic from that road onto the wide but slow road. Total travel time would be lowered and social welfare enhanced by the imposition of such a Pigouvian tax. Moreover, this solution is Pareto-efficient since the individual travel cost is neither increased nor reduced by the toll. All drivers switching to the uncongested slow road can do that at constant cost. Remaining on the disburdened fast road imposes also constant cost on the drivers due to the corrective Pigouvian tax. Hence, the toll revenues correspond to an increase in the overall surplus.

Pigou implicitly assumed public roads without private property rights so that overloads arise on comfortable roads. The need for government intervention as reflected by imposing Pigouvian taxes can be avoided, however, by allocating private property rights. Following Knight (1924), if competitive pressure is present then self-interested road owners (or network providers) would charge trip-makers the same differential toll. In this competitive setting, the price which maximizes the revenue of the owner is equal to the difference of the marginal and average costs of a trip, where average costs are immediately borne by the motorist. Similar observations would hold true for privately owned networks where service providers without market power set up some sort of marginal cost pricing. The difference is "merely" to find an appropriate system of internal prices that reflect the concept of element based charging. Moreover, this view opens up a new perspective since the network provider accounts also for the costs involved in network construction including expansions, improvements, and maintenance. This indicates a close relation between congestion pricing and optimal network investment.

Vickrey (1969) supplements the congestion problem by concomitant phenomena of queuing. In doing so, he assumes that traffic congestion takes the form of cars queuing behind a bottleneck of fixed flow capacity. Arriving drivers have to wait in queues of demand dependent length until their accommodation. This effect is captured by waiting costs. Due to a substantial variability in demand particular rush hours imply delays so that individuals have to adjust their departure times if they want to arrive at their destination in time. The problem is that not all commuters

can arrive at work at the same time because of the finite capacity of the bottleneck. Hence, some commuters must arrive early and others late, which is reflected by so-called "schedule delay costs". An equilibrium requires that no commuter can reduce private costs from unilaterally changing his departure time.

Vickrey proposed a variable toll system which is in many aspects very similar to the Pigouvian solution. Each toll rises from zero during off-peak hours to reach a maximum in the middle of the rush hour. In contrast with a Pigouvian tax, these tolls comprise two effects. On the one hand, they reflect the relative scarcity of fixed capacities in the presence of a variable demand. On the other hand, they internalize the individually ignored cost of delays that are imposed by any commuter on the drivers enqueued behind him. With this pattern of tolls, the commuters are just well off paying the variable toll and having no (or a shorter) queue as they were before with no toll but with an equivalent queue. There is no change in the arrival pattern at the city center. Summing up, Vickrey (1969, p. 255) emphasizes: *the revenue derived from the charges thus represents a clear gain. We thus have an example of tax revenue that not only has no excess burden, it has no burden at all!* Furthermore, Vickrey proves that a transport investment may be worthless without an appropriate congestion pricing. The expansion of a bottleneck route then has the only effect to attract traffic from slow but un-congested alternative routes to the expanded route until the total travel costs per vehicle are the same as before.

In a first-best world, many models, both static and dynamic, end up on a Pareto-efficient outcome: a tax revenue that has no burden. Nevertheless, the real world differs from Pareto's idealized world, and introducing more reality in the congestion models appears to be a necessity.

2.1.5 Solution Methods

The analytical treatment of network problems requires a sound mathematical framework. Many problems can be tackled on the basis of *linear* and *non-linear programming*, the fundamentals of which can be found in Section 2.2. It gives a brief overview on duality theory and introduces basic results of sensitivity analysis. Although one might expect to start with graph theory, which is closely related to combinatory optimization, we shall see that graph theoretic problems with relevance for our purpose have equivalent convex programming formulations; cf., for example, the max-flow/min-cut theorem. The advantage of convex programming is its analytical strength, e.g., to prove the existence of optimal solutions and to characterize their properties. One substantial result with regard to economic analysis is the relationship between optimally chosen primal quantity variables and the corresponding system of shadow prices indicating the worth of restrictions.

Nevertheless, we strongly recommend not to lose sight of *graph theory*, which emphasizes the algorithmic point of view. It is by far not enough to detect the existence of an optimal solution but to determine at least approximately optimal solutions in an efficient manner. Not the pure extent of large scale network problems but their complexity prevents even high-performance computers from solving such problems. In cases where even brute force is not able to determine optimal solutions

(or to verify the optimality), we need approximation algorithms, the quality of which is measured by the goodness of computed solutions in relation to the effort needed to obtain these solutions.

Section 3.2 underscores that convex programming heavily relies on the existence of some convex objective function with a symmetric Jacobian matrix. Regarding aggregate problems such as welfare maximization such an assumption turns out to be more of a hindrance than a help. We introduce therefore in Section 3.2.4.2 the more general concept of *variational inequality problems*, which can handle asymmetric problems that have no aggregate objective function due to missing integrability conditions. Duality theory of conjugate functions though returns to the symmetric case in order to derive alternative classes of more regular problems that have the same task of efficient resource utilization. The corresponding results are given in Section 4.2.5.

Chapter 5 starts with the aforementioned analytical framework and explaines then how to split certain complex problems into two smaller but interdependent problems. Such *bilevel programming* problems have appealing interpretations in conjunction with game theory. Afterwards the analysis is supplemented with basic approaches taken from *queuing theory*, see Section 5.2.3.3. This is necessary whenever (stochastic) arrivals of flows overwhelm the capacities of the network so that queuing delays reduce the quality of service delivered to the customers. Queuing is only one instrument to coordinate network processes. An alternative approach will be discussed in Chapter 6, where interacting processes are recorded by a *discrete event dynamic system* (DEDS). The analysis aims at the synchonization of flows by optimized timetables that reveal at the time the bottlenecks for a faster schedule. This framework is similar to Petri networks and makes use of the so-called max-plus algebra. Several modules are provided that can be used to solve DEDSs with the aid of *MATHEMATICA*®. Although simulation is neglected throughout the book, this may be one starting point for future research.

2.2 Linear and Non-linear Programming

2.2.1 Duality in Linear Programming

Before going into the details of network analysis we need some fundamentals of convex and particularly linear programming. The definitions and results collected in this section brief comprehensive works such as Dantzig (1963), Rockafellar (1972), or Bazaraa et al. (1993) and serve as reference points for the following analysis. Readers who are familiar with duality theory are suggested to jump over this section and to come back when needed. Nevertheless bear in mind the following notation.

Bold letters ($\mathbf{a}, \mathbf{b}, \ldots$) designate always column vectors, e.g., $\mathbf{x} = (x_1, \ldots, x_n)^\top$, and bold capital letters ($\mathbf{A}, \mathbf{B}, \ldots$) indicate matrices (all of appropriate size). We make use of the following notational convenience.

$$\begin{aligned}
\mathbf{x} > \mathbf{y} &:\Longleftrightarrow x_j > y_j \qquad j = 1,\ldots,n;\\
\mathbf{x} \geqq \mathbf{y} &:\Longleftrightarrow x_j \geqq y_j \qquad j = 1,\ldots,n;\\
\mathbf{x} \geq \mathbf{y} &:\Longleftrightarrow [\mathbf{x} \geqq \mathbf{y} \quad \text{and} \quad \mathbf{x} \neq \mathbf{y}].
\end{aligned}$$

It is assumed that terms such as point (or direction), line, plane, hyperplane, half space, convex set, cone, etc. are known to the reader. To give a comprehensive example look at the following graphic representation. It includes a feasible set given by $\mathbf{Ax} \leqq \mathbf{b}$, a tangent (hyper-)plane determined by $\mathbf{p}^\top\mathbf{x} = \alpha$, the corresponding closed half space satisfying $\mathbf{p}^\top\mathbf{x} \leqq \alpha$, a normal vector $\mathbf{p}$ (originated at the vertex $\hat{\mathbf{x}}$) to that halfspace and the normal cone $C(\hat{\mathbf{x}})$ to the feasible set at $\hat{\mathbf{x}}$.

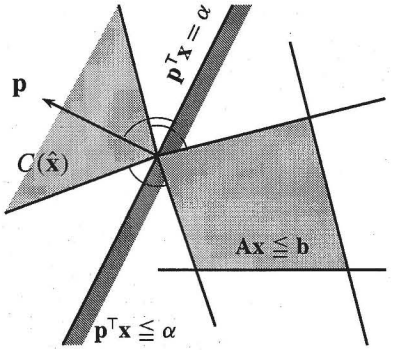


Fig. 2.1
An introductory example
The *point* $\hat{\mathbf{x}}$ is not explicitely indicated. It can be found where the *direction* $\mathbf{p}$ starts. Notice that the feasible set, the cone, the tangent hyperplane and the halfspace are all convex sets.

A linear program has the form as given in (2.1.P) or (2.1.D). We shall refer to (2.1.P) as the *primal program* with the primal variables x_1, x_2, and x_3. Each linear program is assigned to a *dual program* of the form (2.1.D) with the dual variables u_1, u_2, and u_3.

(2.1)

$$\begin{array}{llll}
\text{(P)} & \min & p_1x_1 + p_2x_2 + p_3x_3 & \\
& \text{s.t.} & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \geqq b_1 & \\
& & a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 & \\
& & a_{31}x_2 + a_{32}x_2 + a_{33}x_3 \leqq b_3 & \\
& & x_1 \qquad\qquad\quad \geqq 0 & \\
& & \qquad x_2 \qquad\quad \text{unrestr.} & \\
& & \qquad\qquad x_3 \leqq 0 &
\end{array}$$

$$\begin{array}{llll}
\text{(D)} & \max & b_1u_1 + b_2u_2 + b_3u_3 & \\
& \text{s.t.} & a_{11}u_1 + a_{21}u_2 + a_{31}u_3 \leqq p_1 & \\
& & a_{12}u_1 + a_{22}u_2 + a_{32}u_3 = p_2 & \\
& & a_{13}u_1 + a_{23}u_2 + a_{33}u_3 \geqq p_3 & \\
& & u_1 \qquad\qquad\quad \geqq 0 & \\
& & \qquad u_2 \qquad\quad \text{unrestr.} & \\
& & \qquad\qquad u_3 \leqq 0 &
\end{array}$$

The principles of constructing the dual program are summarized in the following Table 2.1. It is to be read line by line. In contrast to the above programs the table assumes n primal variables x_j, $j = 1,\ldots,n$, and m dual variables u_i, $i = 1,\ldots,m$, where in general $m \neq n$. The number of primal variables m is equal to the number of dual restrictions, and vice versa. Hence, the number of coefficients on the right hand side (RHS) of the primal restrictions matches the number of coefficients in the dual objective function, and vice versa. Notice that there is a one-to-one correspondence between the two programs, that is the dual of the dual program yields the primal

program. A linear program in its most general form is given in (2.2.P), where the

Table 2.1 Rules for forming the dual

Primal Problem	Dual Problem
minimize primal objective	maximize dual objective
objective coefficients	RHS of dual restrictions
RHS of primal restrictions	objective coefficients
coefficient matrix	transposed coefficient matrix
primal relation	dual variable ($i = 1,\ldots, m$)
(ith) inequality $\geqq$	$u_i \geqq 0$
(ith) inequality $\leqq$	$u_i \leqq 0$
(ith) equation $=$	u_i unrestricted in sign
primal variable ($j = 1,\ldots, n$)	dual relation
$x_j \geqq 0$	(jth) inequality $\leqq$
$x_j \leqq 0$	(jth) inequality $\geqq$
x_j unrestricted in sign	(jth) equation $=$

primal variables are collected by three vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. The reader may verify that the dual program (2.2.D) with the dual variables $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ is derived by the rules of Table 2.1.

(2.2)

$$
\begin{array}{llll}
\text{(P)} & \min & \mathbf{p}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} + \mathbf{r}^T\mathbf{z} & \\
& \text{s.t.} & \mathbf{A}_1\mathbf{x} + \mathbf{A}_2\mathbf{y} + \mathbf{A}_3\mathbf{z} \geqq \mathbf{a} & \\
& & \mathbf{B}_1\mathbf{x} + \mathbf{B}_2\mathbf{y} + \mathbf{B}_3\mathbf{z} = \mathbf{b} & \\
& & \mathbf{C}_1\mathbf{x} + \mathbf{C}_2\mathbf{y} + \mathbf{C}_3\mathbf{z} \leqq \mathbf{c} & \\
& & \mathbf{x} \qquad\qquad \geqq \mathbf{0} & \\
& & \mathbf{y} \qquad \text{unrestr.} & \\
& & \mathbf{z} \leqq \mathbf{0} &
\end{array}
\qquad
\begin{array}{llll}
\text{(D)} & \max & \mathbf{a}^T\mathbf{u} + \mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{w} & \\
& \text{s.t.} & \mathbf{A}_1^T\mathbf{u} + \mathbf{B}_1^T\mathbf{v} + \mathbf{C}_1^T\mathbf{w} \leqq \mathbf{p} & \\
& & \mathbf{A}_2^T\mathbf{u} + \mathbf{B}_2^T\mathbf{v} + \mathbf{C}_2^T\mathbf{w} = \mathbf{q} & \\
& & \mathbf{A}_3^T\mathbf{u} + \mathbf{B}_3^T\mathbf{v} + \mathbf{C}_3^T\mathbf{w} \geqq \mathbf{r} & \\
& & \mathbf{u} \qquad\qquad \geqq \mathbf{0} & \\
& & \mathbf{v} \qquad \text{unrestr.} & \\
& & \mathbf{w} \leqq \mathbf{0} &
\end{array}
$$

The task to find an optimal primal solution is closely related to the existence of dually feasible solutions. In what follows we state conditions on feasible solutions which permit us to assess whether or not we have found an optimal solution.

2.2 Theorem (Weak Duality) *If* $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ *is a feasible vector of the primal problem* (2.2.P) *and* $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ *a feasible vector of the dual problem* (2.2.D), *then*

$$\mathbf{p}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} + \mathbf{r}^T\mathbf{z} \geqq \mathbf{a}^T\mathbf{u} + \mathbf{b}^T\mathbf{v} + \mathbf{c}^T\mathbf{w}.$$

Proof: The rows of the following table correspond to the restrictions of the primal problem, while the columns refer to the constraints of the dual problem.

$$\begin{array}{lllll}
\mathbf{u}^T\mathbf{A}_1\mathbf{x} + & \mathbf{u}^T\mathbf{A}_2\mathbf{y} + & \mathbf{u}^T\mathbf{A}_3\mathbf{z} & \geqq \mathbf{u}^T\mathbf{a} & (\mathbf{u} \geqq \mathbf{0}) \\
+ & + & + & & \\
\mathbf{v}^T\mathbf{B}_1\mathbf{x} + & \mathbf{v}^T\mathbf{B}_2\mathbf{y} + & \mathbf{v}^T\mathbf{B}_3\mathbf{z} & = \mathbf{v}^T\mathbf{b} & \\
+ & + & + & & \\
\mathbf{w}^T\mathbf{C}_1\mathbf{x} + & \mathbf{w}^T\mathbf{C}_2\mathbf{y} + & \mathbf{w}^T\mathbf{C}_3\mathbf{z} & \geqq \mathbf{w}^T\mathbf{c} & (\mathbf{w} \leqq \mathbf{0}) \\
\leqq \mathbf{p}^T\mathbf{x} & = \mathbf{q}^T\mathbf{y} & \leqq \mathbf{r}^T\mathbf{z} & & \\
(\mathbf{x} \geqq \mathbf{0}) & & (\mathbf{z} \leqq \mathbf{0}) & &
\end{array}$$

Summing up the fourth row and the fourth column yields the statement of the theorem. ■

An immediate consequence of Theorem 2.2 is

2.3 Theorem (Unbounded-Infeasible Relationship) *If* (2.2.P) *is unbounded, then* (2.2.D) *is infeasible, and vice versa.*

Theorem 2.2 can be transformed into a stronger statement when feasible solutions do exist. For a proof see, e.g., Collatz, Wetterling (1971, p. 56 f.).

2.4 Theorem (Strong Duality) *A feasible vector* $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ *of the primal problem* (2.2.P) *is a minimum solution to* (2.2.P) *if and only if a feasible vector* $(\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ *of* (2.2.D) *exists with*

$$\mathbf{p}^T\hat{\mathbf{x}} + \mathbf{q}^T\hat{\mathbf{y}} + \mathbf{r}^T\hat{\mathbf{z}} = \mathbf{a}^T\hat{\mathbf{u}} + \mathbf{b}^T\hat{\mathbf{v}} + \mathbf{c}^T\hat{\mathbf{w}}\,.$$

In this case $(\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ *is a maximum solution of the dual problem* (2.2.D). – *A similar statement results if we start with* (2.2.D) *instead of* (2.2.P).

Sometimes we know or assume properties of primal or dual optimal solutions. In order to derive the corresponding properties of the complementary solution, it is useful to define primal slack variables $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ and dual slack variables $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$.

$$\begin{aligned}
\boldsymbol{\sigma}_1 &:= \mathbf{A}_1\mathbf{x} + \mathbf{A}_2\mathbf{y} + \mathbf{A}_3\mathbf{z} - \mathbf{a} \geqq \mathbf{0} \\
\boldsymbol{\sigma}_2 &:= \mathbf{C}_1\mathbf{x} + \mathbf{C}_2\mathbf{y} + \mathbf{C}_3\mathbf{z} - \mathbf{c} \leqq \mathbf{0} \\
\boldsymbol{\tau}_1 &:= \mathbf{A}_1^T\mathbf{u} + \mathbf{B}_1^T\mathbf{v} + \mathbf{C}_1^T\mathbf{w} - \mathbf{p} \geqq \mathbf{0} \\
\boldsymbol{\tau}_2 &:= \mathbf{A}_3^T\mathbf{u} + \mathbf{B}_3^T\mathbf{v} + \mathbf{C}_3^T\mathbf{w} - \mathbf{r} \leqq \mathbf{0}
\end{aligned}$$

2.5 Theorem (Complementary Slackness) *A feasible vector* $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ *of the primal problem* (2.2.P) *is a minimum solution to* (2.2.P) *if and only if a feasible vector* $(\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}})$ *of* (2.2.D) *exists with*

$$\begin{array}{lll}
\hat{\mathbf{u}}^T\hat{\boldsymbol{\sigma}}_1 = 0, & \mathbf{B}_1\hat{\mathbf{x}} + \mathbf{B}_2\hat{\mathbf{y}} + \mathbf{B}_3\hat{\mathbf{z}} - \mathbf{b} = \mathbf{0}, & \hat{\mathbf{w}}^T\hat{\boldsymbol{\sigma}}_2 = 0, \\
\hat{\mathbf{x}}^T\hat{\boldsymbol{\tau}}_1 = 0, & \mathbf{A}_2^T\hat{\mathbf{u}} + \mathbf{B}_2^T\hat{\mathbf{v}} + \mathbf{C}_2^T\hat{\mathbf{w}} - \mathbf{q} = \mathbf{0}, & \hat{\mathbf{z}}^T\hat{\boldsymbol{\tau}}_2 = 0.
\end{array}$$

Regarding the signs of the variables, the complementary slackness theorem requires particularly

$$\begin{aligned}
\hat{\mathbf{u}}^T\hat{\boldsymbol{\sigma}}_1 = 0 \quad &\Longleftrightarrow \quad \hat{u}_i\,\hat{\sigma}_{1i} = 0 \;\; \forall\, i \\
\hat{\mathbf{w}}^T\hat{\boldsymbol{\sigma}}_2 = 0 \quad &\Longleftrightarrow \quad \hat{w}_j\,\hat{\sigma}_{2j} = 0 \;\; \forall\, j
\end{aligned}$$

$$\hat{\mathbf{x}}^T\hat{\boldsymbol{\tau}}_1 = 0 \iff \hat{x}_k\,\hat{\tau}_{1k} = 0 \;\forall k$$
$$\hat{\mathbf{z}}^T\hat{\boldsymbol{\tau}}_2 = 0 \iff \hat{z}_l\,\hat{\tau}_{2l} = 0 \;\forall l$$

Farkas' theorem below is concerned with the feasibility of the following primal program. Using $\mathbf{p} = \mathbf{0}$ means that any feasible primal solution is worth the same as any other solution.

$$\begin{array}{lll} \text{(P)} & \min\ \mathbf{p}^T\mathbf{x} & \text{(D)} \quad \max\ \mathbf{b}^T\mathbf{u} \\ & \text{s.t.}\ \ \mathbf{A}\mathbf{x} = \mathbf{b} & \text{s.t.}\ \ \mathbf{A}^T\mathbf{u} \leqq \mathbf{p} \\ & \quad\ \mathbf{x} \geqq \mathbf{0} & \quad\ \mathbf{u}\ \text{unrestr.} \end{array}$$

2.6 Theorem (Farkas)[20] *One and only one of the two systems* $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geqq \mathbf{0}$ *and* $\mathbf{A}^T\mathbf{u} \leqq \mathbf{0}$, $\mathbf{b}^T\mathbf{u} > 0$ *has a solution.*

Proof: Let $\hat{\mathbf{x}}$ be a solution to the first system and $\hat{\mathbf{u}}$ such that $\mathbf{A}^T\hat{\mathbf{u}} \leqq \mathbf{0}$. Then $\mathbf{b}^T\hat{\mathbf{u}} = \hat{\mathbf{x}}^T\mathbf{A}\hat{\mathbf{u}} \leqq 0$ and, therefore, the second system has no solution. Now, let $\hat{\mathbf{u}}$ be a solution to the second system. Then an α exists such that $\mathbf{b}^T\hat{\mathbf{u}} > \alpha > 0$. Every solution $\hat{\mathbf{x}}$ to the first system requires, on the one hand, $\mathbf{b}^T\hat{\mathbf{u}} = \hat{\mathbf{x}}^T\mathbf{A}^T\hat{\mathbf{u}} > \alpha > 0$. On the other hand, we have $\mathbf{x}^T\mathbf{A}^T\hat{\mathbf{u}} \leqq 0$ for every $\mathbf{x} \geqq \mathbf{0}$. Both conditions contradict each other so that the first system has no solution. ■

One can show that Gordan's Theorem (one and only one of the two systems $\mathbf{A}\mathbf{x} = \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{A}^T\mathbf{u} < \mathbf{0}$ has a solution) is an immediate consequence of Farkas' Theorem. The figure below illustrates both cases of Farkas' Theorem. Recall that $\mathbf{A}\mathbf{x}$ represents a linear combination of the columns $\mathbf{a}_j$ of $\mathbf{A}$ with non-negative weights collected by $\mathbf{x}$. With regard to Figure 2.2, any non-negative solution of $\mathbf{a}_1x_1 + \mathbf{a}_2x_2 + \mathbf{a}_3x_3 = \mathbf{b}$ thus can be represented by the vector $\mathbf{b}$ lying in the convex cone spanned by $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$.

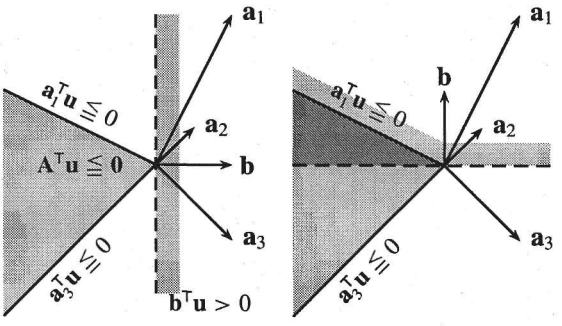


Fig. 2.2
Illustration of Farkas' Theorem
In the left part the first system of Farkas' Theorem has a solution but the second system has none. The opposite is true for the other part.

Suppose now that the optimal primal solution is known. As suggested by Figures 2.1 and 2.2 one can show that the optimal value is always attained at some vertex of the feasible set. The next step is to ask for the effects when we are allowed – probably

[20] Cf. Bazaraa et al. (1993, p. 46). A similar version of the theorem together with the corresponding proof can be found in Vanderbei (1998, p. 159).

at some cost – to change the parameters of the linear program in question, that is for example

$$\begin{array}{rcl} \min\ \mathbf{p}^\top\mathbf{x} & \rightarrow & \min\ (\mathbf{p}+\mathrm{d}\mathbf{p})^\top\mathbf{x} \\ \text{s.t.}\ \mathbf{A}\mathbf{x} \leqq \mathbf{b} & \rightarrow & \text{s.t.}\ (\mathbf{A}+\mathrm{d}\mathbf{A})\,\mathbf{x} \leqq (\mathbf{b}+\mathrm{d}\mathbf{b}) \\ \mathbf{x} \geqq \mathbf{0} & \rightarrow & \mathbf{x} \geqq \mathbf{0} \end{array} \tag{2.3}$$

As can be found in Neumann, Morlock (1993, p. 118–135), two approaches are to be distinguished.

1. *Sensitivity analysis.* Modify the parameters of the problem such that the optimal solution does not change qualitatively (same vertex of the feasible region).
2. *Parametric optimization.* Modify the parameters of the problem, where we allow for switches to new vertices of the feasible region as optimal solution.

Although the dual program of (2.3) shows non-positive dual variables $\mathbf{u} \leqq \mathbf{0}$, we shall discuss some answers on the basis of the Lagrangean function $\mathcal{L}$ using non-negative Lagrange multipliers $\boldsymbol{\lambda} \geqq \mathbf{0}$ associated with the primal restrictions.

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{p}^\top\mathbf{x} + \boldsymbol{\lambda}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$$

The problem is now to find a saddle point $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ of $\mathcal{L}$, that is

$$\mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}) \geqq \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) \geqq \mathcal{L}(\hat{\mathbf{x}}, \boldsymbol{\lambda}) \quad \forall\boldsymbol{\lambda},\ \forall\mathbf{x},$$

which can be described by so-called Karush-Kuhn-Tucker (KKT) conditions

$$\begin{array}{lll} \mathbf{p} + (\hat{\boldsymbol{\lambda}}^\top\mathbf{A})^\top \geqq \mathbf{0}, & \hat{\mathbf{x}} \geqq \mathbf{0}, & \mathbf{p}^\top\hat{\mathbf{x}} + \hat{\boldsymbol{\lambda}}^\top\mathbf{A}\hat{\mathbf{x}} = 0, \\ \mathbf{A}\hat{\mathbf{x}} - \mathbf{b} \leqq \mathbf{0}, & \hat{\boldsymbol{\lambda}} \geqq \mathbf{0}, & \hat{\boldsymbol{\lambda}}^\top\mathbf{A}\hat{\mathbf{x}} - \hat{\boldsymbol{\lambda}}^\top\mathbf{b} = 0. \end{array}$$

If we identify $\boldsymbol{\lambda} = -\mathbf{u}$, the results so far obtained fit in perfectly with the KKT conditions. The left conditions correspond to the dual and primal restrictions. The right conditions reflect the complementary slackness of Theorem 2.5. Furthermore, when the KKT conditions are satisfied, then $\mathbf{p}^\top\hat{\mathbf{x}} = -\mathbf{b}^\top\hat{\boldsymbol{\lambda}} = \mathbf{b}^\top\hat{\mathbf{u}}$, as has been stated by the strong duality theorem.

Before going into aspects of sensitivity analysis and parametric optimization, let us review some basic results of the more general convex analysis.

2.2.2 Convex Analysis

In this section we are concerned with *ordinary convex programs* of the form

$$\inf_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leqq \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0} \right\}, \tag{2.4}$$

where f and g_i, $i = 1,\ldots, m$, are convex functions on $\mathcal{X}$, and h_k, $k = 1,\ldots, p$, are affine functions on $\mathcal{X}$, and $\mathcal{X}$ is a non-empty convex set in $\mathbb{R}^n$; cf., e.g., the non-negative orthant $\mathcal{X} = \mathbb{R}^n_+$. The convexity of the restrictions implies the *feasible set*

$$\Omega := \{\mathbf{x} \in \mathcal{X} \mid \mathbf{g}(\mathbf{x}) \leqq \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$$

to be convex. With that $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \Omega\}$ is frequently used as a synonym for (2.4). A vector $\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2)$ is called a *Karush-Kuhn-Tucker (KKT) vector* for (2.4) if $\hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}$ and the infimum of the Lagrangean function for (2.4)

$$\mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}) := f(\mathbf{x}) + \hat{\boldsymbol{\lambda}}_1{}^{\mathsf{T}}\mathbf{g}(\mathbf{x}) + \hat{\boldsymbol{\lambda}}_2{}^{\mathsf{T}}\mathbf{h}(\mathbf{x})$$

is finite[21] and equal to the optimal value in (2.4).

The associated *perturbation function* (or *envelope function*) results from[22]

$$\hat{F}(\boldsymbol{\beta}) := \inf_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leqq \boldsymbol{\beta}_1,\ \mathbf{h}(\mathbf{x}) = \boldsymbol{\beta}_2\}.$$

This is the optimal value function of (2.4) perturbed by the amount $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. We now direct our attention to the properties of $\hat{F}$ around $\boldsymbol{\beta} = \mathbf{0}$. Notice that $\hat{F}(\mathbf{0})$ is the optimal value in the unperturbed problem (2.4). Of course one can think of more general perturbations of all parameters involved, see for example Tobin (1986). A standard reference on that topic is Fiacco, McCormick (1968).

Assume that we are allowed to change the right hand side of the restriction $\mathbf{0}$ to any $\boldsymbol{\beta}$ that we please, except that we must pay for the change, the price being q_i per unit of perturbation variable β_i. Then, for any perturbation $\boldsymbol{\beta}$, the minimum cost we can achieve in the perturbed problem plus the cost of $\boldsymbol{\beta}$ will be

$$\hat{F}(\boldsymbol{\beta}) + \mathbf{q}^{\mathsf{T}}\boldsymbol{\beta}.$$

A perturbation will be "worth buying" if and only if this quantity is less than $\hat{F}(\mathbf{0})$, the optimal value of the unperturbed problem. Alternatively, suppose $\hat{F}(\mathbf{0})$ is finite. Then no perturbation whatsoever is worth buying at non-negative prices $\mathbf{q}$ if

$$\hat{F}(\mathbf{0}) \leqq \hat{F}(\boldsymbol{\beta}) + \mathbf{q}^{\mathsf{T}}\boldsymbol{\beta} \qquad \forall\, \boldsymbol{\beta}. \tag{2.5}$$

Now we compare $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2)$ to a KKT vector $\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2)$. When $\hat{F}(\mathbf{0})$ is finite, a KKT vector $\hat{\boldsymbol{\lambda}}$ requires

$$\hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0} \quad \text{and} \quad \hat{F}(\mathbf{0}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}). \tag{2.6}$$

In order to compare (2.5) and (2.6) we rewrite (2.5) as follows

---

[21] Declare $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = +\infty$ whenever $\mathbf{x} \notin \Omega$.

[22] The perturbation function can be written as $\hat{F}(\boldsymbol{\beta}) = \inf_{\mathbf{x}} F(\mathbf{x}, \boldsymbol{\beta})$, where $F(\cdot, \boldsymbol{\beta}) := f(\cdot) + \delta(\cdot \mid \Omega(\boldsymbol{\beta}))$ is a *convex bifunction* and δ denotes the indicator function of the set $\Omega(\boldsymbol{\beta}) := \{\mathbf{x} \in \mathcal{X} \mid \mathbf{g}(\mathbf{x}) \leqq \boldsymbol{\beta}_1,\ \mathbf{h}(\mathbf{x}) = \boldsymbol{\beta}_2\}$.

In the special case of $f(\mathbf{x}) = \mathbf{p}^{\mathsf{T}}\mathbf{x}$ the function $\rho_*(\mathbf{p} \mid \Omega) = \inf_{\mathbf{x}} \{\mathbf{p}^{\mathsf{T}}\mathbf{x} - \rho(\mathbf{x} \mid \Omega)\}$ is called the concave conjugate function of $\rho(\cdot \mid \Omega)$ with $\rho(\cdot \mid \Omega) = -\delta(\cdot \mid \Omega)$. This function corresponds to the (reciprocal) support function of Ω. In the case that Ω depends on $\boldsymbol{\beta}$ one may write $\rho_*(\mathbf{p} \mid \Omega(\boldsymbol{\beta})) = \hat{F}(\boldsymbol{\beta} \mid \mathbf{p})$.

$$\begin{aligned}\hat{F}(\mathbf{0}) &\leqq \inf_{\boldsymbol{\beta}} \left\{ \hat{F}(\boldsymbol{\beta}) + \mathbf{q}^{\mathsf{T}}\boldsymbol{\beta} \right\} \\ &= \inf_{\boldsymbol{\beta}} \inf_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \mathbf{q}^{\mathsf{T}}\boldsymbol{\beta} \mid \mathbf{g}(\mathbf{x}) \leqq \boldsymbol{\beta}_1,\ \mathbf{h}(\mathbf{x}) = \boldsymbol{\beta}_2 \right\} \\ &= \begin{cases} \inf_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \mathbf{q}_1{}^{\mathsf{T}}\mathbf{g}(\mathbf{x}) + \mathbf{q}_2{}^{\mathsf{T}}\mathbf{h}(\mathbf{x}) \right\} & \text{if } \mathbf{q}_1 \geqq \mathbf{0}\text{; see (2.6)} \\ -\infty & \text{otherwise.} \end{cases}\end{aligned}$$

Under the assumption of $\hat{F}(\mathbf{0})$ being finite we gain

$$\hat{F}(\mathbf{0}) = \inf_{\mathbf{x}\in\mathcal{X}} \left\{ f(\mathbf{x}) + \mathbf{q}_1{}^{\mathsf{T}}\mathbf{g}(\mathbf{x}) + \mathbf{q}_2{}^{\mathsf{T}}\mathbf{h}(\mathbf{x}) \right\} = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{q}). \tag{2.7}$$

Thus, when $\hat{F}(\mathbf{0})$ is finite and $\hat{\boldsymbol{\lambda}} = \mathbf{q}$, the statements of (2.5) and (2.6) are equivalent.

2.7 Summary The Lagrange multipliers $\hat{\boldsymbol{\lambda}}$ may be seen as equilibrium prices $\mathbf{q}$. Given these prices neither a change in $\hat{\mathbf{x}}$ nor any perturbation $\boldsymbol{\beta}$ will improve the optimal value $\hat{F}(\mathbf{0})$. □

The Lagrangean duality deals with the following problem referred to as the dual problem of (2.4).

$$\sup_{\boldsymbol{\lambda}_1 \geqq \mathbf{0}} \widehat{\mathcal{L}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \quad \text{with} \quad \widehat{\mathcal{L}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) := \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \tag{2.8}$$

This is the point to restate the results of linear programming in terms of convex analysis. The theorems are taken from Bazaraa et al. (1993, Chapter 6). We state them without proof.

2.8 Theorem (Weak Duality) *Let* $\mathbf{x}$ *be a feasible solution to* (2.4)*, that is* $\mathbf{x} \in \Omega$*. Also let* $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ *be a feasible solution to* (2.8)*, that is* $\boldsymbol{\lambda}_1 \geqq \mathbf{0}$*. Then* $f(\mathbf{x}) \geqq \widehat{\mathcal{L}}(\boldsymbol{\lambda})$.

Some implications are worth to be mentioned.

- $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \Omega\} \geqq \sup_{\boldsymbol{\lambda}_1 \geqq \mathbf{0}} \widehat{\mathcal{L}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$
- If $f(\hat{\mathbf{x}}) = \widehat{\mathcal{L}}(\hat{\boldsymbol{\lambda}})$, where $\hat{\mathbf{x}} \in \Omega$ and $\hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}$, then $\hat{\mathbf{x}}$ and $\hat{\boldsymbol{\lambda}}$ solve the primal and dual problems, respectively.
- If $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \Omega\} = -\infty$, then the dual problem has no feasible solution, that is $\widehat{\mathcal{L}}(\boldsymbol{\lambda}) = -\infty$ for every $\boldsymbol{\lambda}_1 \geqq \mathbf{0}$.
- If $\sup_{\boldsymbol{\lambda}_1 \geqq \mathbf{0}} \widehat{\mathcal{L}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = +\infty$, then the primal problem is infeasible.

The weak duality property $f(\mathbf{x}) \geqq \widehat{\mathcal{L}}(\boldsymbol{\lambda})$ permits us to state how far a feasible solution is from optimality. In the case of strict inequality the term $[f(\mathbf{x}) - \widehat{\mathcal{L}}(\boldsymbol{\lambda})]/\widehat{\mathcal{L}}(\boldsymbol{\lambda}) \leqq \alpha$ gives an upper bound on the relative deviation from optimality. This observation is particularly useful in practice to formulate stopping rules by which an algorithm stops if the degree of suboptimality falls short of some threshold value.

2.9 Theorem (Strong Duality) *Let* f *and* $\mathbf{g}$ *be convex and let* $\mathbf{h}$ *be affine. Suppose that the following constraint qualification holds true. There is an* $\bar{\mathbf{x}} \in \mathcal{X}$ *such that* $\mathbf{g}(\bar{\mathbf{x}}) < \mathbf{0}$, $\mathbf{h}(\bar{\mathbf{x}}) = \mathbf{0}$, *and* $\mathbf{0} \in \operatorname{rint}\{\mathbf{h}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$. *Then*

$$\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \Omega\} = \sup_{\boldsymbol{\lambda}_1 \geqq \mathbf{0}} \widehat{\mathcal{L}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2).$$

Furthermore, if the infimum is finite, the supremum is achieved at $\hat{\boldsymbol{\lambda}}$ *with* $\hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}$. *That is a KKT vector exists for* (2.4). *If the infimum is achieved at* $\hat{\mathbf{x}}$, *then* $\hat{\boldsymbol{\lambda}}_1{}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{x}}) = 0$.

The complementary slackness theorem has also a counterpart in convex analysis – namely, the KKT conditions of Theorem 2.19 below. Before proceeding with this result we need some further instruments characterizing the properties of the functions involved.

2.10 Definition (Subgradient) *A vector* $\mathbf{y}$ *is said to be a* subgradient *of a convex function* f *at a point* $\bar{\mathbf{x}}$ *if*[23]

$$f(\mathbf{x}) \geqq f(\bar{\mathbf{x}}) + \mathbf{y}^{\mathsf{T}}(\mathbf{x} - \bar{\mathbf{x}}) \qquad \forall \mathbf{x}. \tag{2.9}$$

The set of all subgradients $\mathbf{y}$ *is called the* subdifferential[24] *of* f *at* $\bar{\mathbf{x}}$ *and is denoted by* $\partial f(\bar{\mathbf{x}})$. *Reversing the direction of the inequality in* (2.9) *with respect to a concave function* f *yields the* superdifferential $\Delta f(\bar{\mathbf{x}})$ *of* f *at* $\bar{\mathbf{x}}$ *and* $\mathbf{y}$ *is called a* supergradient.

From Rockafellar (1972, Theorem 25.1) we know, that the *gradient* $\nabla f(\mathbf{x})$ is the unique subgradient of f at $\mathbf{x}$, provided f is differentiable at $\mathbf{x}$. In this case we have $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

2.11 Remark (Geometric Meaning) Suppose f is finite at $\bar{\mathbf{x}}$. The subdifferential $\partial f(\bar{\mathbf{x}})$ corresponds to a convex cone which is normal to the level of f having $\bar{\mathbf{x}}$ on its boundary, see Figure 2.3.

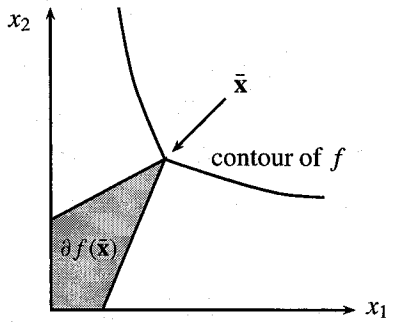


Fig. 2.3
The subdifferential of f at $\bar{\mathbf{x}}$

With respect to the subgradient inequality (2.9), the graph of $h(\mathbf{x}) = f(\bar{\mathbf{x}}) + \mathbf{y}^{\mathsf{T}}(\mathbf{x} - \bar{\mathbf{x}})$ represents a non-vertical supporting hyperplane to the convex epigraph epi$f$ at the point $\binom{\bar{\mathbf{x}}}{f(\bar{\mathbf{x}})}$. To see this, rewrite $h(\mathbf{x})$ as

$$\begin{pmatrix} \mathbf{y} \\ -1 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} \mathbf{x} \\ h(\mathbf{x}) \end{pmatrix} = \mathbf{y}^{\mathsf{T}}\bar{\mathbf{x}} - f(\bar{\mathbf{x}}) = \text{ const.}$$

[23] The subgradient inequality (2.9) must be satisfied for all $\mathbf{x} \in \mathbb{R}^n$. This is useful because $f(\mathbf{x}) = +\infty$ for every $\mathbf{x}$ not belonging to the effective domain of f, that is $\mathbf{x} \notin \text{Dom } f$. Introducing the feasible set Ω, the inequality (2.9) must either hold good for all $\mathbf{x} \in \Omega$ or we redefine f such that $f(\mathbf{x}) = \infty$ if $\mathbf{x} \notin \text{Dom } f \cap \Omega$. This can be done by $F(\cdot) = f(\cdot) + \delta(\cdot|\Omega)$.

[24] The subdifferential is a closed convex set, which may be empty or it may consist of exactly one vector. For further properties of that generalized gradient of f see Clarke (1975).

The vector $\binom{y}{-1}$ is normal to epi f at the point concerned.

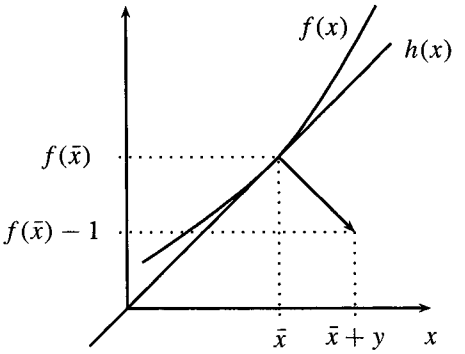


Fig. 2.4
The gradient y of f at $\bar{x}$

□

An immediate consequence of the definition of a subgradient is the following: a necessary and sufficient condition for a point $\hat{\mathbf{x}}$ to belong to the minimum set of f is $\mathbf{0} \in \partial f(\hat{\mathbf{x}})$. This relationship reduces to $\mathbf{0} = \nabla f(\hat{\mathbf{x}})$ if f is differentiable at $\hat{\mathbf{x}}$.

Recall that $\boldsymbol{\lambda}$ is a KKT vector if the optimal value $\hat{F}(\mathbf{0})$ of (2.4) is finite and

$$\hat{F}(\mathbf{0}) \leqq \hat{F}(\boldsymbol{\beta}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\beta} \qquad \forall\, \boldsymbol{\beta}.$$

This inequality (2.5) describes $-\boldsymbol{\lambda}$ as a subgradient of $\hat{F}$ at $\boldsymbol{\beta} = \mathbf{0}$, that is $-\boldsymbol{\lambda} \in \partial \hat{F}(\mathbf{0})$.

Some important results on the *relationship between convex functions and its subgradients* can be found in Bazaraa et al. (1993), Chapter 3. They are cited here for the sake of convenience.

2.12 Theorem *Let Ω be a nonempty convex set in $\mathbb{R}^n$, and let $f\colon \Omega \to \mathbb{R}$. Suppose that for each point $\bar{\mathbf{x}} \in \Omega$ there exists a subgradient* $\mathbf{y}$ *such that* (2.9) *holds good. Then f is convex on* int Ω.

2.13 Theorem *Let Ω be a nonempty open convex set in $\mathbb{R}^n$, and let $f\colon \Omega \to \mathbb{R}$ be differentiable on Ω. Then f is convex if and only if*

$$f(\mathbf{x}) \geqq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) \qquad \forall\, \mathbf{x} \in \Omega$$

for any $\bar{\mathbf{x}} \in \Omega$.

Another convexity criterion depends on the Hessean matrix of f. Note that a quadratic matrix $\mathbf{A}$ is said to be *positive semidefinite* if $\mathbf{y}^{\mathrm{T}}\mathbf{A}\,\mathbf{y} \geqq 0$ for all $\mathbf{y}$.

2.14 Theorem *Let Ω be a nonempty open convex set in $\mathbb{R}^n$, and let $f\colon \Omega \to \mathbb{R}$ be twice differentiable on Ω. Then, f is convex if and only if the* Hessean matrix $\mathbf{H}_f := \left[\partial^2 f(\mathbf{x})/\partial x_i \partial x_j\right]_{i,j=1,\ldots,n}$ *is positive semidefinite at each point* $\mathbf{x}$ *of Ω.*

The proof of this theorem refers to the mean value theorem by which there is an $\hat{\mathbf{x}} = \lambda\mathbf{x} + (1 - \lambda)\bar{\mathbf{x}}$ for some $0 < \lambda < 1$ such that

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{H}_f(\hat{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}).$$

If $\mathbf{H}_f$ is positive semidefinite then f is convex by Theorem 2.13.

The relationship between convex functions and its subgradients can now be continued with respect to ordinary convex programs and subgradients of the objective function f at an optimal solution $\hat{\mathbf{x}}$. Furthermore, in "regular" cases we can expect to find optimal solutions at some boundary point of Ω so that Figure 2.3 applies for some additional feasible set Ω touching the contour of f at $\bar{\mathbf{x}}$.

2.15 Theorem *Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex function, and Ω be a nonempty convex set in $\mathbb{R}^n$. The point $\hat{\mathbf{x}} \in \Omega$ is an optimal solution to*

$$\min\{f(\mathbf{x}) | \mathbf{x} \in \Omega\} \tag{2.10}$$

if and only if f has a subgradient $\mathbf{y} \in \partial f(\hat{\mathbf{x}})$ such that

$$\mathbf{y}^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \qquad \forall \mathbf{x} \in \Omega. \tag{2.11}$$

The convexity of f is needed to show that (2.11) implies (2.10). The next theorem does not assume f to be convex.

2.16 Theorem[25] *Let $\hat{\mathbf{x}}$ be a solution to the optimization problem* (2.10)*, where f is continuously differentiable and Ω is a nonempty closed convex set in $\mathbb{R}^n$. Then $\hat{\mathbf{x}}$ is a solution of the (variational) problem*[26]

$$\nabla f(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \qquad \forall \mathbf{x} \in \Omega. \tag{2.12}$$

2.17 Theorem[27] *If f is a convex function and $\hat{\mathbf{x}}$ is a solution to* (2.12)*, then $\hat{\mathbf{x}}$ is a solution to the optimization problem* (2.10)*.*

Proof: Use the inequality of Theorem 2.13 with $\bar{\mathbf{x}} = \hat{\mathbf{x}}$. Assuming (2.12) holds good, we have $f(\mathbf{x}) \geqq f(\hat{\mathbf{x}})$ for all $\mathbf{x} \in \Omega$. This is (2.10). ■

Gradients are not only special cases of subdifferentials but also closely related to so-called directional derivatives. They indicate the change of f when starting at a feasible $\mathbf{x}$ and going into some direction $\mathbf{d}$. As many algorithms make use of this observation, we give a brief introduction into that concept.

2.18 Remark (Directional Derivative) Let $\mathbf{x}$ be a point where the function f is finite. The *(one-sided) directional derivative* of f at $\mathbf{x}$ with respect to a vector $\mathbf{d}$ is defined to be the limit

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{\lambda \downarrow 0} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda}$$

if it exists. The directional derivative is *two-sided* if and only if $f'(\mathbf{x}; -\mathbf{d})$ exists and $-f'(\mathbf{x}; -\mathbf{d}) = f'(\mathbf{x}; \mathbf{d})$. If f is differentiable at $\mathbf{x}$, all directional derivatives are finite and two-sided, and one has $f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^{\mathsf{T}}\mathbf{d}$.

---

[25] Nagurney (1999, p. 6), Prop. 1.2.

[26] Variational inequalities will be introduced in Section 3.2.4.2.

As problem (2.12) does not say anything about convexity of f it suggests to define a new class of functions. A differentiable function f is said to be *pseudoconvex* if for each $\mathbf{x}$, $\hat{\mathbf{x}} \in \Omega$ with $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0$ we have $f(\mathbf{x}) \geqq f(\hat{\mathbf{x}})$.

[27] Nagurney (1999, p. 6), Prop. 1.3.

Let $\mathbf{x}$ be a point where the convex function f is finite. Without differentiability of f, Rockafellar (1972, Theorem 23.1) states that

$$-f'(\mathbf{x}; -\mathbf{d}) \leqq f'(\mathbf{x}; \mathbf{d}) \quad \forall \mathbf{d}.$$

Rockafellar (1972, Theorem 23.2) proves $\mathbf{y}$ to be a subgradient of f at $\mathbf{x}$ if and only if the directional derivative satisfies

$$f'(\mathbf{x}; \mathbf{d}) \geqq \mathbf{y}^{\mathsf{T}}\mathbf{d} \quad \forall \mathbf{d}.$$

The minimum set of f consists of a unique vector $\hat{\mathbf{x}}$ if and only if the conjugate[28] convex function f^* defined by $f^*(\mathbf{y}) := \sup_{\mathbf{x}} \{\mathbf{y}^{\mathsf{T}}\mathbf{x} - f(\mathbf{x})\}$ is differentiable at $\mathbf{y} = \mathbf{0}$ and $\hat{\mathbf{x}} = \nabla f^*(\mathbf{0})$; Rockafellar (1972, Theorem 27.1(e)). □

The subsequent results are taken from Rockafellar (1972, §§ 28, 29). The idea is to look at (2.6) with regard to a saddle-point condition. A pair $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ is said to be a *saddle-point* of the Lagrangean function $\mathcal{L}$ if

$$\mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}) \geqq \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) \geqq \mathcal{L}(\hat{\mathbf{x}}, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda}, \ \forall \mathbf{x}.$$

This is equivalent to having

$$\inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}}) = \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) = \sup_{\boldsymbol{\lambda}} \mathcal{L}(\hat{\mathbf{x}}, \boldsymbol{\lambda}).$$

2.19 Theorem (Saddle Point Optimality) *In order that $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{m+p}$ be a KKT vector for* (2.4) *and $\hat{\mathbf{x}} \in \mathbb{R}^n$ be an optimal solution to* (2.4)*, it is necessary and sufficient that $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ be a saddle-point of the Lagrangean function $\mathcal{L}$ of* (2.4)*. Moreover, this condition holds true if and only if $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ satisfies the following* ***Karush-Kuhn-Tucker (KKT) conditions.***[29]

$$\mathbf{0} \in \partial f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \hat{\lambda}_i \partial g_i(\hat{\mathbf{x}}) + \sum_{i=m+1}^{m+p} \hat{\lambda}_i \partial h_i(\hat{\mathbf{x}}), \tag{2.13a}$$

$$\mathbf{g}(\hat{\mathbf{x}}) \leqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}_1{}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{x}}) = 0 \tag{2.13b}$$

$$\mathbf{h}(\hat{\mathbf{x}}) = \mathbf{0} \tag{2.13c}$$

The first condition is frequently written in the form $\mathbf{0} \in \partial_{\mathbf{x}} \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$. Moreover, when all functions involved are differentiable at $\hat{\mathbf{x}}$ we gain

$$\nabla_{\mathbf{x}} \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) = \nabla f(\hat{\mathbf{x}}) + \nabla \mathbf{g}(\hat{\mathbf{x}})\hat{\boldsymbol{\lambda}}_1 + \nabla \mathbf{h}(\hat{\mathbf{x}})\hat{\boldsymbol{\lambda}}_2 = \mathbf{0}.$$

Be aware that the *Jacobian matrix* of $\mathbf{g}$ and $\mathbf{h}$ are given by $\mathbf{J}_{\mathbf{g}}(\hat{\mathbf{x}}) = (\nabla \mathbf{g}(\hat{\mathbf{x}}))^{\mathsf{T}}$ and $\mathbf{J}_{\mathbf{h}}(\hat{\mathbf{x}}) = (\nabla \mathbf{h}(\hat{\mathbf{x}}))^{\mathsf{T}}$, respectively. That is, the columns of $\nabla \mathbf{g}(\hat{\mathbf{x}})$ and $\nabla \mathbf{h}(\hat{\mathbf{x}})$ are the gradients of the respective restrictions.

---

[28] Pairs of conjugate functions and results obtained from duality theory will be in the center of interest in Section 4.2.5.

[29] To emphasize that Karush independently derived the same conditions in 1939 (that is 12 years before Kuhn and Tucker), one speaks today of Karush-Kuhn-Tucker (KKT) conditions.

In the special case of $\mathcal{X} = \mathbb{R}^n_+$ this preceding condition goes over to

$$\nabla_{\mathbf{x}}\mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) \geqq \mathbf{0}, \quad \hat{\mathbf{x}} \geqq \mathbf{0}, \quad \hat{\mathbf{x}}^{\top}\nabla_{\mathbf{x}}\mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) = 0.$$

The condition $\hat{\mathbf{x}} \in \Omega$ is referred to as *primal feasibility* condition, whereas (2.13a) together with $\hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}$ is named *dual feasibility* condition. As in linear programming, $\hat{\boldsymbol{\lambda}}_1{}^{\top}\mathbf{g}(\hat{\mathbf{x}}) = 0$ is called the *complementary slackness* condition.

Theorem 2.19 presupposes that an optimal solution $\hat{\mathbf{x}}$ together with a KKT vector $\hat{\boldsymbol{\lambda}}$ do exist. On the necessary conditions, where $\hat{\mathbf{x}}$ implies the existence of a KKT vector $\hat{\boldsymbol{\lambda}}$, see Bazaraa et al. (1993, Theorem 4.3.7). Sufficient conditions, where $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ satisfying the KKT conditions imply $\hat{\mathbf{x}}$ to be an optimal solution, are given in Bazaraa et al. (1993, Theorem 4.3.7).

2.20 Theorem (Karush-Kuhn-Tucker Theorem) *Assume that the optimal value of* (2.4) *is finite and that the constraint qualification of Theorem 2.9 holds good. In order that a given vector* $\hat{\mathbf{x}}$ *is an optimal solution to* (2.4), *it is necessary and sufficient that a vector* $\hat{\boldsymbol{\lambda}}$ *exists such that* $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ *is a saddle-point of the Lagrangean function* $\mathcal{L}$ *of* (2.4). *Equivalently,* $\hat{\mathbf{x}}$ *is an optimal solution if and only if there is a vector* $\hat{\boldsymbol{\lambda}}$ *of Lagrange multipliers which, together with* $\hat{\mathbf{x}}$, *satisfies the KKT conditions for* (2.4).

The calculation using the result in (2.7) suggests an alternative definition of the Lagrangean function,

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) &:= \inf_{\boldsymbol{\beta}} \left\{\boldsymbol{\lambda}^{\top}\boldsymbol{\beta} + F(\mathbf{x}, \boldsymbol{\beta})\right\} \\ &= \inf_{\boldsymbol{\beta}} \left\{\boldsymbol{\lambda}^{\top}\boldsymbol{\beta} + f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leqq \boldsymbol{\beta}\right\},\end{aligned}$$

where F is the bifunction with $F(\mathbf{x}, \boldsymbol{\beta}) := f(\mathbf{x}) + \delta(\mathbf{x} \mid \Omega(\boldsymbol{\beta}))$. Taking the infimum with respect to $\mathbf{x}$ yields

$$\inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{\beta}, \mathbf{x}} \left\{\boldsymbol{\lambda}^{\top}\boldsymbol{\beta} + F(\mathbf{x}, \boldsymbol{\beta})\right\} = \inf_{\boldsymbol{\beta}} \left\{\boldsymbol{\lambda}^{\top}\boldsymbol{\beta} + \hat{F}(\boldsymbol{\beta})\right\}.$$

From (2.6) we know that $\hat{\boldsymbol{\lambda}} \geqq \mathbf{0}$ is a KKT vector if $\hat{F}(\mathbf{0})$ is finite and $\hat{F}(\mathbf{0}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\boldsymbol{\lambda}})$. (Notation: $\widehat{\mathcal{L}}(\hat{\boldsymbol{\lambda}}) = \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}}) = \hat{F}(\mathbf{0})$) Thus, if the optimal value in the convex program (2.4) is finite, the KKT vectors for (2.4) are precisely the vectors $\hat{\boldsymbol{\lambda}}$ such that $-\hat{\boldsymbol{\lambda}}$ is a subgradient of $\hat{F}$ at $\boldsymbol{\beta} = \mathbf{0}$, see Rockafellar (1972, Theorem 29.1).[30]

$$-\hat{\boldsymbol{\lambda}} \in \partial\hat{F}(\mathbf{0}). \tag{2.14}$$

This is a special case of the *envelope theorem* which deals with the more general case of $f(\mathbf{x}, \boldsymbol{\beta})$ and $g(\mathbf{x}, \boldsymbol{\beta}) \leqq \mathbf{0}$.

The program (2.4) has a unique KKT vector $\hat{\boldsymbol{\lambda}}$ if and only if the function $\hat{F}$ is differentiable at $\boldsymbol{\beta} = \mathbf{0}$, in which case $\hat{\boldsymbol{\lambda}}$ is given by the formula (Rockafellar (1972, Corollary 29.1.3))

[30] To put it in other words a necessary and sufficient condition for $\boldsymbol{\beta} = \mathbf{0}$ to be in the minimum set of $\hat{\boldsymbol{\lambda}}^{\top}\boldsymbol{\beta} + \hat{F}(\boldsymbol{\beta})$ is $\mathbf{0} \in \partial(\hat{\boldsymbol{\lambda}}^{\top}\boldsymbol{\beta} + \hat{F}(\boldsymbol{\beta}))$. Relation (2.14) results from $\boldsymbol{\beta} = \mathbf{0}$.

$$-\hat{\boldsymbol{\lambda}} = \nabla \hat{F}(\mathbf{0}). \tag{2.15}$$

As a notational remark observe that (2.15) is the same as $-\hat{\lambda}_i = \partial \hat{F}(\mathbf{0})/\partial \beta_i$ for all i.

Without going into further details, the corresponding statements with respect to a marginal change in the dual variables $\boldsymbol{\lambda}$ can be found in Bazaraa et al. (1993, Theorems 6.3.3 and 6.3.4), that is

$$\mathbf{g}(\hat{\mathbf{x}}) \in \partial \widehat{\mathcal{L}}(\hat{\boldsymbol{\lambda}}) \quad \text{or} \quad \mathbf{g}(\hat{\mathbf{x}}) = \nabla \widehat{\mathcal{L}}(\hat{\boldsymbol{\lambda}}) \tag{2.16}$$

Notational remark: if $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ attains its minimum at $\hat{\mathbf{x}}$, then $\nabla \widehat{\mathcal{L}}(\hat{\boldsymbol{\lambda}}) = \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$.

Although (2.15) and (2.16) are derived independently, they are closely related. According to Theorem 2.9 we have $\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{g}(\hat{\mathbf{x}}) = 0$. This equation is also part of the KKT conditions.

2.21 Remark (Geometric meaning of the KKT conditions) The idea is to show that if f is convex, and if f attains a minimum at $\hat{\mathbf{x}}$, then the KKT conditions are satisfied.

If all functions are differentiable at $\hat{\mathbf{x}}$, the KKT conditions (2.13) for (2.4) *without equality constraints* are

$$\nabla f(\hat{\mathbf{x}}) + \nabla \mathbf{g}(\hat{\mathbf{x}}) \hat{\boldsymbol{\lambda}} = \mathbf{0},$$
$$\mathbf{g}(\hat{\mathbf{x}}) \leqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}} \geqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{g}(\hat{\mathbf{x}}) = 0.$$

Let $\mathcal{B}$ be the index set of all restrictions g_i being active at $\hat{\mathbf{x}}$ (in linear programming the set $\mathcal{B}$ denotes the basis of a solution). Then

$$C(\hat{\mathbf{x}}) = \left\{ \mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} = \sum_{i \in \mathcal{B}} \lambda_i \nabla g_i(\hat{\mathbf{x}}), \quad \lambda_i \geqq 0 \, \forall \, i \in \mathcal{B} \right\}$$

is a polehydral cone generated by the gradients of the restrictions g_i being active at $\hat{\mathbf{x}}$, cf. Figure 2.5. For convenience let $\mathbf{A}$ designate this collection of gradients such that $C = \left\{ \mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} = \mathbf{A}\boldsymbol{\lambda}_{\mathcal{B}}, \ \boldsymbol{\lambda}_{\mathcal{B}} \geqq \mathbf{0} \right\}$. Now $-\nabla f(\hat{\mathbf{x}}) \in C$ if and only if the ***first system*** $-\nabla f(\hat{\mathbf{x}}) = \mathbf{A}\boldsymbol{\lambda}_{\mathcal{B}}, \boldsymbol{\lambda}_{\mathcal{B}} \geqq \mathbf{0}$ has a solution, say $\hat{\boldsymbol{\lambda}}_{\mathcal{B}}$. This is true by Farkas' Theorem 2.6 if and only if the ***second system*** $\mathbf{A}^{\mathrm{T}}\mathbf{z} \leqq \mathbf{0}$, $\nabla f(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{z} < 0$ has no solution. Assume f attains a minimum on Ω at $\hat{\mathbf{x}}$. Then Theorem 2.16 states that

$$\begin{aligned} & \nabla f(\hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{z} - \hat{\mathbf{x}}) \geqq 0 \qquad && \forall \, \mathbf{z} \in \Omega \\ \Longleftrightarrow \quad & \nabla f(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{z} \geqq 0 && \forall \, \mathbf{z} \in \Omega - \hat{\mathbf{x}}. \end{aligned}$$

With that any solution to the second system must be outside $\Omega - \hat{\mathbf{x}}$ and at the same time within the cone[31] $K(\hat{\mathbf{x}}) := \left\{ \mathbf{z} \mid \mathbf{A}(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{z} \leqq \mathbf{0} \right\}$. In the next step one can show that $\Omega - \hat{\mathbf{x}} \subseteq K(\hat{\mathbf{x}})$ or, equivalently, $\Omega \subseteq K(\hat{\mathbf{x}}) + \hat{\mathbf{x}}$. Thus the second system has no solution and, therefore, the first system has a solution $\hat{\boldsymbol{\lambda}}_{\mathcal{B}}$. Setting $\hat{\lambda}_i = 0$ for all $i \notin \mathcal{B}$ implies that $(\hat{\mathbf{x}}, \hat{\boldsymbol{\lambda}})$ is a solution to the KKT conditions.[32]

[31] The cone $K(\hat{\mathbf{x}}) \setminus \{\mathbf{0}\}$ correspond to the set of feasible directions in Bazaraa et al. (1993, p. 140 ff.).

[32] Some authors prefer the polar cone C° of C which coincides with $-K(\hat{\mathbf{x}})$; cf., e.g., Elster et al. (1977, p. 151).

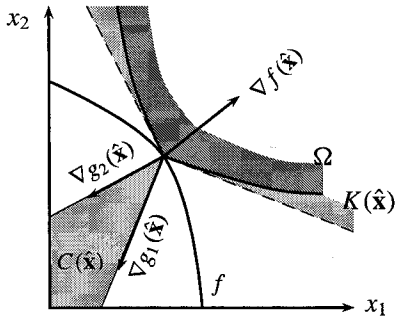


Fig. 2.5
Illustration of KKT conditions

□

Although an optimal solution to the ordinary convex program satisfies $f(\hat{\mathbf{x}}) = \hat{F}(\mathbf{0})$, the two statements $\nabla f(\hat{\mathbf{x}}) = -\nabla \mathbf{g}(\hat{\mathbf{x}})\hat{\boldsymbol{\lambda}}$ and $\nabla \hat{F}(\mathbf{0}) = -\hat{\boldsymbol{\lambda}}$ have a completely different meaning. Nevertheless, if $\nabla \mathbf{g}(\hat{\mathbf{x}})$ is an invertible $n \times n$-matrix, then

$$\nabla \hat{F}(\mathbf{0}) = -\hat{\boldsymbol{\lambda}} = (\nabla \mathbf{g}(\hat{\mathbf{x}}))^{-1} \nabla f(\hat{\mathbf{x}}). \tag{2.17}$$

2.2.3 Convex Analysis Applied to Linear Programming

2.2.3.1 Perturbations of the Right Hand Side

Going back to the criterion (2.5), we are now ready to discuss the linear program (2.3) with respect to changes of $\mathbf{b}$. These parameters are chosen optimally if the subsequent condition (2.19) holds good. Regarding sensitivity analysis, the perturbation $d\mathbf{b} = \boldsymbol{\beta}$ is restricted such that the basis of the unperturbed problem remains feasible and thus optimal. Perturbations which are feasible in this sense are discussed in the next section. The ordinary convex program given in (2.4) now takes the form of a linear program

$$\min_{\mathbf{x} \in \mathbb{R}^n_+} \left\{ \mathbf{p}^\top \mathbf{x} \mid \mathbf{A}^\top \mathbf{x} \leqq \mathbf{b} \right\}$$

so that a perturbation $\boldsymbol{\beta}$ yields

$$\begin{aligned} &(\mathrm{P}) \quad \min \ \mathbf{p}^\top \mathbf{x} && (\mathrm{D}) \quad \max \ (\mathbf{b} + \boldsymbol{\beta})^\top \mathbf{u} \\ &\quad \text{s.t.} \ \mathbf{A}\mathbf{x} \leqq \mathbf{b} + \boldsymbol{\beta} && \quad \text{s.t.} \quad \mathbf{A}^\top \mathbf{u} \leqq \mathbf{p} \\ &\qquad \mathbf{x} \geqq \mathbf{0} && \qquad \mathbf{u} \leqq \mathbf{0}. \end{aligned} \tag{2.18}$$

If the primal variables $\mathbf{x}$ are unrestricted in sign, the dual restrictions become $\mathbf{A}^\top \mathbf{u} = \mathbf{p}$ and the above KKT conditions (2.13) apply. The difference between the preceding Lagrange multipliers $\boldsymbol{\lambda}$ and the dual variables $\mathbf{u}$ is just $\boldsymbol{\lambda} \geqq \mathbf{0}$ and $\mathbf{u} \leqq \mathbf{0}$. The gradient of the primal objective function $f(\mathbf{x}) = \mathbf{p}^\top \mathbf{x}$ is $\nabla f(\mathbf{x}) = \mathbf{p}$. The gradients of the restrictions $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ follow from $\nabla \mathbf{g}(\mathbf{x}) = \mathbf{A}^\top$.

No perturbation $\boldsymbol{\beta}$ is worth buying at market prices $\mathbf{q}$ if

$$\begin{aligned} &\mathbf{p}^\top \hat{\mathbf{x}}(\mathbf{0}) \leqq \mathbf{p}^\top \hat{\mathbf{x}}(\boldsymbol{\beta}) + \mathbf{q}^\top \boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \\ :\Longleftrightarrow \quad &-\mathbf{q} \in \partial \hat{F}(\mathbf{0}), \end{aligned} \tag{2.19}$$

where $\hat{\mathbf{x}}(\boldsymbol{\beta})$ is an optimal solution to the perturbed problem and $\hat{\mathbf{x}}(\mathbf{0})$ denotes a corresponding solution to the unperturbed problem. (Under appropriate, but restrictive assumptions one would obtain $-\mathbf{q} = \nabla\hat{F}(\mathbf{0}) = -\hat{\boldsymbol{\lambda}} = \hat{\mathbf{u}} = (\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{p}$ as has been mentioned in (2.17).) If there is a solution $\hat{\mathbf{x}}(\mathbf{0})$ such that $\hat{F}(\mathbf{0}) = \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\mathbf{0})$ is finite, then, by the strong duality theorem 2.4,

$$\mathbf{b}^{\mathsf{T}}\hat{\mathbf{u}}(\mathbf{0}) \geqq (\mathbf{b}+\boldsymbol{\beta})^{\mathsf{T}}\hat{\mathbf{u}}(\boldsymbol{\beta}) + \mathbf{q}^{\mathsf{T}}\boldsymbol{\beta} \qquad \forall\,\boldsymbol{\beta}.$$

Notice that the right hand side is either finite with $\hat{F}(\boldsymbol{\beta}) = (\mathbf{b}+\boldsymbol{\beta})^{\mathsf{T}}\hat{\mathbf{u}}(\boldsymbol{\beta})$ or infinite in the case in which the perturbation prevents the perturbed problem from having a feasible solution. Rearranging the inequality yields

$$\mathbf{b}^{\mathsf{T}}\big(\hat{\mathbf{u}}(\mathbf{0}) - \hat{\mathbf{u}}(\boldsymbol{\beta})\big) \geqq \boldsymbol{\beta}^{\mathsf{T}}\big(\hat{\mathbf{u}}(\boldsymbol{\beta}) + \mathbf{q}\big) \qquad \forall\,\boldsymbol{\beta}.$$

Assuming that the basis of the primal solution does not change, we have $\hat{\mathbf{u}}(\mathbf{0}) = \hat{\mathbf{u}}(\boldsymbol{\beta})$. Thus,

$$\hat{\mathbf{u}}(\mathbf{0}) = \hat{\mathbf{u}}(\boldsymbol{\beta}) = -\mathbf{q}$$

if all components of $\boldsymbol{\beta}$ are allowed to be positive or negative. Regarding the objective function the perturbation $\boldsymbol{\beta}$ may have a positive effect, i.e., $\mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\boldsymbol{\beta}) < \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\mathbf{0})$, but the cost of perturbation $\mathbf{q}^{\mathsf{T}}\boldsymbol{\beta}$ always overrides this improvement.

Choosing $\boldsymbol{\beta} = \mathbf{e}_i\beta_i$ that is $\Delta b_i = \beta_i \neq 0$ such that the basis of the primal solution does not change, we obtain, on the one hand, $\hat{\mathbf{u}}(\mathbf{0}) = \hat{\mathbf{u}}(\boldsymbol{\beta})$ and, on the other hand,

$$\Delta\hat{F}(\mathbf{0}) = \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\boldsymbol{\beta}) - \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\mathbf{0}) = \beta_i\hat{u}_i(\boldsymbol{\beta}) \geqq -\beta_i q_i.$$

Thus,

$$\frac{\Delta\hat{F}(\mathbf{0})}{\Delta b_i} = \frac{\mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\boldsymbol{\beta}) - \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\mathbf{0})}{\beta_i} \geqq -q_i \quad \text{if } \beta_i > 0$$

$$\frac{\Delta\hat{F}(\mathbf{0})}{\Delta b_i} = \frac{\mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\boldsymbol{\beta}) - \mathbf{p}^{\mathsf{T}}\hat{\mathbf{x}}(\mathbf{0})}{\beta_i} \leqq -q_i \quad \text{if } \beta_i < 0$$

As has been noted by (2.15), these inequalities describe the effect of a marginal change in b_i, i.e., $\Delta b_i \to 0$. The corresponding uniqueness of the KKT vector may be stated as follows. *Suppose the optimal value $\hat{F}(\mathbf{0})$ of the program* $\inf\{\mathbf{p}^{\mathsf{T}}\mathbf{x} \mid \mathbf{A}\mathbf{x} \geqq \mathbf{b},\ \mathbf{x} \geqq \mathbf{0}\}$ *is finite. Then this program has a unique KKT vector $\hat{\boldsymbol{\lambda}} = -\hat{\mathbf{u}}$ if and only if the function $\hat{F}$ is differentiable at $\boldsymbol{\beta} = \mathbf{0}$, in which case $\hat{\mathbf{u}}$ is given by the formula*

$$\frac{\partial\hat{F}(\mathbf{0})}{\partial b_i} = \hat{u}_i = -q_i \qquad i = 1,\ldots,m. \tag{2.20}$$

For the sake of notational convenience it is useful to redefine $\hat{F}$ being a function of $\boldsymbol{\beta}$. Let

$$Z(\tilde{\mathbf{b}}) := \min\{\mathbf{p}^{\mathsf{T}}\mathbf{x} \mid \mathbf{A}\mathbf{x} \geqq \tilde{\mathbf{b}},\ \mathbf{x} \geqq \mathbf{0}\},$$

where $\tilde{\mathbf{b}} := \mathbf{b} + \boldsymbol{\beta}$. Then $\hat{F}(\boldsymbol{\beta}) = Z(\tilde{\mathbf{b}})$ and (2.20) becomes

$$\frac{\partial Z(\mathbf{b})}{\partial \tilde{b}_i} = \hat{u}_i = -q_i, \qquad i = 1,\ldots,m. \tag{2.21}$$

2.2.3.2 Feasible Perturbations of the Right Hand Side

In what follows we are interested in perturbations $\mathbf{db} = \boldsymbol{\beta}$ which leave the characteristics of the former solution untouched. By intuition, we require that the new solution is attained at the same vertex of the feasible set as before. In doing so a characterization of a vertex by basis and non-basis variables is needed. The standard form of (2.18) after introducing slack variables $\mathbf{w}$ follows from renaming the variables and parameters.

$$\begin{array}{lllll}
\min\ \mathbf{p}^{\mathsf{T}}\mathbf{x}+\mathbf{0}^{\mathsf{T}}\mathbf{w} & \rightarrow & \text{(P)}\ \min\ \tilde{\mathbf{p}}^{\mathsf{T}}\mathbf{y} & \text{(D)}\ \max\ \mathbf{b}^{\mathsf{T}}\mathbf{u} \\
\text{s.t.}\ \mathbf{A}\mathbf{x}+\ \mathbf{I}\mathbf{w}=\mathbf{b} & \rightarrow & \text{s.t.}\ \mathbf{C}\mathbf{y}=\mathbf{b} & \text{s.t.}\ \mathbf{C}^{\mathsf{T}}\mathbf{u} \leqq \tilde{\mathbf{p}} \\
\mathbf{x} \geqq \mathbf{0} & \rightarrow & \mathbf{y} \geqq \mathbf{0} & \mathbf{u}\ \text{unrestr.} \\
\mathbf{w} \geqq \mathbf{0} & & &
\end{array}$$

Let $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_{\mathcal{B}}, \hat{\mathbf{y}}_{\mathcal{N}})$ be a primal solution with the basis variables $\hat{\mathbf{y}}_{\mathcal{B}} \geqq \mathbf{0}$ and the non-basis variables $\hat{\mathbf{y}}_{\mathcal{N}} = \mathbf{0}$. The primal restrictions then take the form $\mathbf{C}_{\mathcal{B}}\hat{\mathbf{y}}_{\mathcal{B}} + \mathbf{C}_{\mathcal{N}}\hat{\mathbf{y}}_{\mathcal{N}} = \mathbf{C}_{\mathcal{B}}\hat{\mathbf{y}}_{\mathcal{B}} = \mathbf{b}$, where the basis matrix $\mathbf{C}_{\mathcal{B}}$ is invertible, $\hat{\mathbf{y}}_{\mathcal{B}} = \mathbf{C}_{\mathcal{B}}^{-1}\mathbf{b}$. The complementary slackness requires for the corresponding dual restrictions $\mathbf{C}_{\mathcal{B}}^{\mathsf{T}}\hat{\mathbf{u}} = \tilde{\mathbf{p}}_{\mathcal{B}}$ and $\mathbf{C}_{\mathcal{N}}^{\mathsf{T}}\hat{\mathbf{u}} \leqq \tilde{\mathbf{p}}_{\mathcal{N}}$. *As long as a perturbation of* $\mathbf{b}$ *does not change the basis* $\mathcal{B}$, *the former solution remains optimal* and the dual solution is determined by $\hat{\mathbf{u}}(\mathbf{0}) = \hat{\mathbf{u}}(\boldsymbol{\beta}) = (\mathbf{C}_{\mathcal{B}}^{-1})^{\mathsf{T}}\tilde{\mathbf{p}}_{\mathcal{B}}$. The optimal values of the programs, however, differ

$$\mathbf{b}^{\mathsf{T}}(\mathbf{C}_{\mathcal{B}}^{-1})^{\mathsf{T}}\tilde{\mathbf{p}}_{\mathcal{B}} = \hat{\mathbf{y}}_{\mathcal{B}}(\mathbf{0})^{\mathsf{T}}\tilde{\mathbf{p}}_{\mathcal{B}} \quad \text{and} \quad (\mathbf{b}+\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{C}_{\mathcal{B}}^{-1})^{\mathsf{T}}\tilde{\mathbf{p}}_{\mathcal{B}} = \hat{\mathbf{y}}_{\mathcal{B}}(\boldsymbol{\beta})^{\mathsf{T}}\tilde{\mathbf{p}}_{\mathcal{B}}.$$

The primal solution requires $\hat{\mathbf{y}}_{\mathcal{B}}(\boldsymbol{\beta}) = \mathbf{C}_{\mathcal{B}}^{-1}(\mathbf{b}+\boldsymbol{\beta})$ and $\hat{\mathbf{y}}_{\mathcal{B}}(\boldsymbol{\beta}) \geqq \mathbf{0}$. Since the solution to the unperturbed problem is $\hat{\mathbf{y}}_{\mathcal{B}}(\mathbf{0}) = \mathbf{C}_{\mathcal{B}}^{-1}\mathbf{b}$, we gain $\hat{\mathbf{y}}_{\mathcal{B}}(\boldsymbol{\beta}) = \hat{\mathbf{y}}_{\mathcal{B}}(\mathbf{0}) + \mathbf{C}_{\mathcal{B}}^{-1}\boldsymbol{\beta} \geqq \mathbf{0}$ as a condition for the optimal basis to be unchanged. Perturbing one particular b_i the feasible range of β_i is

$$\min_{\gamma_{ji}<0} \frac{-(\hat{y}_{\mathcal{B}}(\mathbf{0}))_j}{\gamma_{ji}} \geqq \beta_i \geqq \max_{\gamma_{ji}>0} \frac{-(\hat{y}_{\mathcal{B}}(\mathbf{0}))_j}{\gamma_{ji}} \qquad (j \in \mathcal{B}),$$

where $\gamma_{ji} = (\mathbf{C}_{\mathcal{B}}^{-1})_{ji}$; cf. Dantzig, Thapa (1997, p. 191). Similarly one can think of a given vector $\boldsymbol{\beta}$ and a parameter ζ so that $\zeta\boldsymbol{\beta}$ is the perturbation at hand. The range of ζ is then

$$\min_{\gamma_j<0} \frac{-(\hat{y}_{\mathcal{B}}(\mathbf{0}))_j}{\gamma_j} \geqq \zeta \geqq \max_{\gamma_j>0} \frac{-(\hat{y}_{\mathcal{B}}(\mathbf{0}))_j}{\gamma_j} \qquad (j \in \mathcal{B}),$$

where $\gamma_j = (\mathbf{C}_{\mathcal{B}}^{-1}\boldsymbol{\beta})_j$.
As long as $\boldsymbol{\beta}$ satisfies these restrictions the optimal basis does not change and the envelope function takes the value

$$\hat{F}(\boldsymbol{\beta}) = \tilde{\mathbf{p}}^{\mathsf{T}}\hat{\mathbf{y}}(\boldsymbol{\beta}) = \tilde{\mathbf{p}}_{\mathcal{B}}^{\mathsf{T}}\mathbf{C}_{\mathcal{B}}^{-1}(\mathbf{b}+\boldsymbol{\beta}).$$

We have in particular $\hat{F}(\mathbf{0}) = \tilde{\mathbf{p}}^{\mathsf{T}}\hat{\mathbf{y}}(\mathbf{0}) = \tilde{\mathbf{p}}_{\mathcal{B}}^{\mathsf{T}}\hat{\mathbf{y}}_{\mathcal{B}}(\mathbf{0}) = \tilde{\mathbf{p}}_{\mathcal{B}}^{\mathsf{T}}\mathbf{C}_{\mathcal{B}}^{-1}\mathbf{b}$ and $\hat{\mathbf{u}} = (\tilde{\mathbf{p}}_{\mathcal{B}}^{\mathsf{T}}\mathbf{C}_{\mathcal{B}}^{-1})^{\mathsf{T}} \in \partial F(\mathbf{0})$.

2.2.4 Basic Differentiation Rules

Matrix operations and differentiation rules can be found in a number of mathematical handbooks. Nevertheless, it turns out to be useful having a brief list of basic differentiation rules with respect to single-valued or vector-valued functions of several variables. The gradients of a (differentiable) single-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$ with $\mathbf{x} \mapsto y = f(\mathbf{x})$ and of a vector-valued function $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ with $\mathbf{x} \mapsto \mathbf{y} = \mathbf{f}(\mathbf{x})$ are denoted by

$$\nabla f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} \\ \vdots \\ \frac{\partial f_1}{\partial x_n} \end{pmatrix} \quad \text{and} \quad \nabla \mathbf{f} = (\nabla f_1, \ldots, \nabla f_n) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

This convention implies $\mathbf{J_f} = (\nabla \mathbf{f})^\mathsf{T}$ for the *Jacobian matrix* $\mathbf{J_f}$ of $\mathbf{f}$.

With that the reader may verify the following differentiation rules for single-valued functions with $\mathbf{a}$ and $\mathbf{A}$ being a column vector and a matrix of constants, respectively.

$$\begin{aligned}
f(\mathbf{x}) &= \mathbf{a}^\mathsf{T}\mathbf{x} && \nabla f(\mathbf{x}) = \mathbf{a} \\
f(\mathbf{x}) &= \mathbf{a}^\mathsf{T}\mathbf{A}\mathbf{x} && \nabla f(\mathbf{x}) = \mathbf{A}^\mathsf{T}\mathbf{a} \\
f(\mathbf{x}) &= \mathbf{a}^\mathsf{T}\mathbf{g}(\mathbf{x}) && \nabla f(\mathbf{x}) = \nabla \mathbf{g}(\mathbf{x})\,\mathbf{a} && = \mathbf{J_g}(\mathbf{x})^\mathsf{T}\mathbf{a} \\
f(\mathbf{x}) &= \mathbf{x}^\mathsf{T}\mathbf{g}(\mathbf{x}) && \nabla f(\mathbf{x}) = \nabla \mathbf{g}(\mathbf{x})\,\mathbf{x} + \mathbf{g}(\mathbf{x}) && = \mathbf{J_g}(\mathbf{x})^\mathsf{T}\mathbf{x} + \mathbf{g}(\mathbf{x}) \\
f(\mathbf{x}) &= g(\mathbf{A}\mathbf{x}) && \nabla f(\mathbf{x}) = \mathbf{A}^\mathsf{T}\nabla g(\mathbf{A}\mathbf{x}) \\
f(\mathbf{x}) &= g(\mathbf{h}(\mathbf{x})) && \nabla f(\mathbf{x}) = \nabla \mathbf{h}(\mathbf{x})\,\nabla g(\mathbf{h}(\mathbf{x})) && = \mathbf{J_h}(\mathbf{x})^\mathsf{T}\nabla g(\mathbf{h}(\mathbf{x}))
\end{aligned}$$

Regarding vector-valued functions we shall need particularly the following four rules.

$$\begin{aligned}
\mathbf{f}(\mathbf{x}) &= \mathbf{A}\mathbf{x} && \nabla \mathbf{f}(\mathbf{x}) = \mathbf{A}^\mathsf{T} \\
\mathbf{f}(\mathbf{x}) &= \mathbf{A}\,\mathbf{g}(\mathbf{x}) && \nabla \mathbf{f}(\mathbf{x}) = \nabla \mathbf{g}(\mathbf{x})\,\mathbf{A}^\mathsf{T} \\
\mathbf{f}(\mathbf{x}) &= \mathbf{g}(\mathbf{h}(\mathbf{x})) && \nabla \mathbf{f}(\mathbf{x}) = \nabla \mathbf{h}(\mathbf{x}) \nabla \mathbf{g}(\mathbf{h}(\mathbf{x})) \\
\mathbf{f}(\mathbf{x}) &= \mathbf{A}\,\mathbf{g}(\mathbf{h}(\mathbf{x})) && \nabla \mathbf{f}(\mathbf{x}) = \nabla \mathbf{h}(\mathbf{x}) \nabla \mathbf{g}(\mathbf{h}(\mathbf{x}))\,\mathbf{A}^\mathsf{T} \iff \\
& && \mathbf{J_f}(\mathbf{x}) = \mathbf{A}\,\mathbf{J_g}(\mathbf{h}(\mathbf{x}))\,\mathbf{J_h}(\mathbf{x})
\end{aligned}$$

We shall use the above rules without further reference.

2.3 Transport Networks

2.3.1 Elementary Features of Transport Networks

Analyzing networks needs some basic notation in terms taken from graph theory. The fundamental ingredients of a network are *nodes* or *vertices* v_j, $j = 1,\ldots,n$, and *edges*, *branches*, or *links* e_i, $i = 1,\ldots,m$, connecting pairs of nodes. Regarding transportation networks it is frequently assumed that every edge can only be used in one direction. Such directed edges are called *arcs*. Thus arcs are ordered pairs of nodes, for example $e_1 = (v_1, v_2)$. Loops of the form (v_1, v_1) are almost always ruled out. The first node determines the origin (or tail) of the link at hand, $\alpha(e_1) = v_1$. The second node designates the destination (or head) of the arc, $\omega(e_1) = v_2$. The two adjacent endpoints of a link are called *neighbors* and the neighborhood of a vertex is the collection of all the nodes adjacent to it. The arcs (v_1, v_2) and (v_2, v_1) are said to be antiparallel. Depending on the type of network there may be parallel connections like lanes, which are combined in one arc. Multiple parallel edges constitute a *multigraph*. In contrast a *simple graph* has no loops or parallel edges.

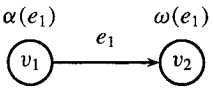


Fig. 2.6
Basic elements of networks
The two adjacent nodes v_1 and v_2 are connected by the arc e_1 which is said to be incident with the nodes.

The set of nodes[33] and the set of arcs are denoted $\mathcal{V} = \{v_1,\ldots, v_n\}$ and $\mathcal{E} = \{e_1,\ldots, e_m\}$, respectively. Similarly, the *order* $\#\mathcal{V}$ and the *size* $\#\mathcal{E}$ of a graph are the number of vertices and edges it has. If each arc is an ordered pair of $\mathcal{V} \times \mathcal{V}$, then $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ is called a *directed graph* or *digraph* in order to distinguish it from undirected graphs. In accordance with Neumann, Morlock (1993), a *network* is a digraph $\mathbf{G}$ without isolated nodes, that is nodes without any connections. The assumption of directed links is improper when edges can be used in both directions. The corresponding modifications will be discussed in Section 2.3.3.

For the sake of convenience it is useful to abuse the above graph theoretical notation somewhat. Let $v = 1,\ldots, n$ and $e = 1,\ldots, m$ be two indices representing nodes and arcs, respectively. The advantage is that the following vector notation can be handled much easier.

The next step is to describe movements of traffic on a given digraph. Let φ_e, $e \in \mathcal{E}$, be the (directed) non-negative flow on link e. It describes the number of flow units traveling from the start node $\alpha(e)$ to the end node $\omega(e)$ possibly with respect to a given period of time. Throughout the book it is assumed that a flow entering an arc equals the flow leaving the arc. Indeed, this assumption on conservation of

[33] Subsets of nodes are called *centroids* in transport models.

flows rules out many examples where links lose some of their flow, for example, power transmission through electric lines, pipelines with a seepage, or transportation of perishable commodities. All link flows are collected in the link flow vector $\boldsymbol{\varphi}$. In general each link $e \in \mathcal{E}$ is able to manage a maximum flow κ_e which is assumed to be fixed, possibly with respect to a predefined level of service delivered to the flow units. The link flow vector in such a *capacitated network* $(\mathbf{G}, \boldsymbol{\kappa})$ is then restricted by the corresponding vector of link capacities $\boldsymbol{\kappa}$, that is $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$. If a link e consists of parallel arcs each having the same capacity, then κ_e represents a multiple of this capacity. Notice that $\boldsymbol{\varphi}$ and $\boldsymbol{\kappa}$ have the same dimension.

$$[\varphi_e] = [\kappa_e] = \text{flow units on edge } e \text{ (per period)}$$

Apart from capacity constraints one of the most important restrictions to flows on a network refers to node flows instead of link flows. The idea is that no flow unit appears or vanishes accidentally. Every flow unit is initiated at one vertex, runs through various other nodes, and finishes its travel at a node. The corresponding trip may be described by a sequence of leaving and entering certain nodes. Hence, each node $v \in \mathcal{V}$ is characterized by its *inflow* φ_v^- and its *outflow* φ_v^+, where the inflow is the sum of all link flows ending at v and the outflow is the sum of all link flows starting at v.

$$\varphi_v^+ = \sum_{e:\, \alpha(e)=v} \varphi_e \quad \text{and} \quad \varphi_v^- = \sum_{e:\, \omega(e)=v} \varphi_e \tag{2.22}$$

Corresponding to Figure 2.6, the properties of nodes are described by the following figure.

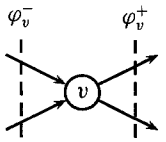


Fig. 2.7
Graphic representation of inflows and outflows

With Figure 2.7 we distinguish three types of nodes

- source or supply nodes when $\varphi_v^- < \varphi_v^+$,
- sink or demand nodes when $\varphi_v^- > \varphi_v^+$, and
- transshipment (transit or relay) nodes with $\varphi_v^- = \varphi_v^+$.

If the network consists only of transit nodes, it satisfies the condition of *flow conservation*:

$$\varphi_v^+ = \varphi_v^- \qquad \forall\, v \in \mathcal{V}. \tag{2.23}$$

Similar to links the flow through a node may be restricted to a certain maximum. To model this feature the node at hand can be split into two nodes. The inflows enter the first node and the outflows leave the second node. The capacity of the initial node is now represented by the capacity of the arc connecting the two auxiliary nodes.

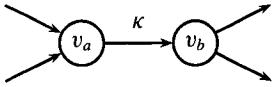


Fig. 2.8
Representation of a node capacity

Depending on the type of network different nodal capacity restrictions can be found. For example, Beckmann (1999, p. 157) suggests an upper bound on the node flow of the form $\varphi_v^- + \varphi_v^+ \leqq \kappa$ which is realistic for an airport. In general, we shall take link capacities as constants. Node capacities, however, determine an upper bound on the sum of incoming flows. Thus, if flow units cannot exit a link due to a congested node, the flow on that link cannot be increased although it has an excess capacity. Road junctions may constitute another important example for varying link capacities. Due to signalization conditions link capacities can change in time.

As a first example for a network consider Figure 2.9 including three types of nodes: a unique source node v_1 which has no inflow, i.e., $\varphi_1^- = 0$, a unique sink node v_4 which has no outflow, i.e., $\varphi_4^+ = 0$, and two transit nodes v_2 and v_3. The arcs are represented by the solid vectors. The dotted link is an artificial feedback edge which is ignored for the moment.

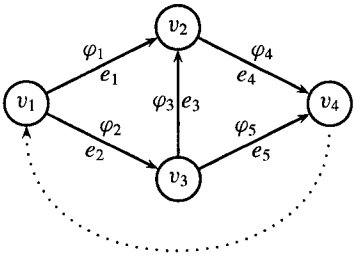


Fig. 2.9
Example of a network
The source is v_1 and the sink is v_4. The dotted vector indicates an artificial link.

As can be seen in Figure 2.9 each flow unit traveling from the source v_1 to the sink v_4 may choose one of three distinct *routes*, namely, $\rho_1 = (e_1, e_4)$, $\rho_2 = (e_2, e_5)$, and $\rho_3 = (e_2, e_3, e_4)$ but not (e_1, e_3, e_5). As the edges have a definite direction, each arc must end at a node where the succeeding arc starts, i.e., for instance $\omega(e_1) = v_2 = \alpha(e_4)$ with respect to the first route. Regarding digraphs a *path*, which is a synonym for route, denotes a forward directed sequence of consecutive arcs. In order to exclude implausible[34] paths in more complex networks we restrict our attention to *elementary routes*[35] which traverse no node and, therefore, no edge twice, except

[34] The enumeration of admissible routes that lead from a certain source to a certain sink may be a very complex task. For example, think of intersections in an urban transport network at which all left turns are prohibited. One technique to overcome this problem is based upon a *dual representation of transportation networks*, see Añez et al. (1996).

[35] Contrasted with elementary paths are *simple paths* which have no repeated arcs. A *closed path* or *cycle* terminates at its initial nodes. Furthermore, acyclic routes include no cycles so that any acyclic route cannot end where it starts.

that the start node and the end node could be the same. The set of all elementary routes from v_i to v_j is denoted $\mathcal{R}(i, j)$. For two routes ρ_1 and ρ_2 joining the same pair of nodes (v_i, v_j) we write $\rho_1 \sim \rho_2$. Notice that each route ρ has a unique origin $\alpha(\rho) = v_i$ and a unique destination $\omega(\rho) = v_j$, where (v_i, v_j) is called an *origin destination (OD) pair*. In comparison to link flows $\boldsymbol{\varphi}$, the flow on route ρ is denoted h_ρ or in vector notation $\mathbf{h}$. Route flows may also have an upper bound, think, for example, of a train of given size commuting along a given route. Notice also that different routes may use the same link. Thus link flows are related to route flows by summing up all route flows which use the particular arc e.

$$\varphi_e = \sum_{\rho:\, e \in \rho} h_\rho \qquad \forall e \in \mathcal{E} \tag{2.24}$$

With regard to the example of Figure 2.9 we have for instance $\varphi_4 = h_1 + h_3$.

According to Figure 2.9 it is sometimes useful to introduce an artificial arc e_ℓ, $\ell \equiv m + 1$, leading back from the sink v_4 to the source v_1 as it is indicated by the dotted vector. Now we can think of all units which have been moved from v_1 to v_4 as if they virtually go back to v_1 via the artificial link. This concept is important with respect to flow conservation, which will be discussed in Section 2.3.2 below. Moreover, the above network together with the artificial edge constitutes a *strongly connected network*[36], where each ordered pair of nodes is joined by a (directed) route.

Multicommodity flows: The number of equal sized "flow units" with the desire to make a trip from v_1 to v_4 is called the demand $x^D_{(1,4)}$. Having more than one OD-pair, where $\mathcal{B}$ is the set of all OD-pairs, this demand for travel is described by a (*multicommodity*) vector $\mathbf{x}^D$. Each component (or commodity) $x^D_{(i,j)}$ denotes[37] the demand for transportation from node v_i to node v_j. In general $\mathbf{x}^D$ is called a *spatial interaction pattern*. Regarding traffic networks, the vector $\mathbf{x}^D$ corresponds to the *trip table*. For telecommunications networks $\mathbf{x}^D$ is called the end-to-end matrix or just demand. In the example of Figure 2.9 the demand $x^D_{(1,4)}$ is divided into three route flows, $x^D_{(1,4)} = h_1 + h_2 + h_3$. For multicommodity networks we have to sum up all route flows connecting the same OD-pair.

$$\sum_{\rho \in \mathcal{R}(i,j)} h_\rho = x^D_{(i,j)} \qquad \forall (v_i, v_j) \in \mathcal{V} \times \mathcal{V}. \tag{2.25}$$

If each OD-pair is connected by a unique route then $\mathbf{h} = \mathbf{x}^D$. In fact (2.25) represents the *(path) flow conservation* in multicommodity networks. Each flow unit generated at the source v_s ends up in the destination node v_t and no unit is lost in between.

Going back to equation (2.22), the relation between node flows and the demand vector $\mathbf{x}^D$ is given by

[36] Undirected graphs as well as digraphs are said to be connected when every pair of nodes is connected by a *chain*, i.e., an undirected sequence of edges or arcs.

[37] If needed, we may introduce an index for the OD-pairs. Given n numbered nodes, there are $n(n-1)$ ordered pairs of the form (v_i, v_j) with the index $k = ni + j$, where $i, j = 1, \ldots, n$ and $i \neq j$. Then, $x^D_{(i,j)} \to x^D_k$.

$$\varphi_i^+ = \sum_j x_{(i,j)}^D \quad \forall i \in \mathcal{V} \qquad \text{(net) outflow or production of trips} \tag{2.26a}$$

$$\varphi_j^- = \sum_i x_{(i,j)}^D \quad \forall j \in \mathcal{V} \qquad \text{(net) inflow or attraction of trips} \tag{2.26b}$$

Observe that supply nodes (or sources) generate flow units while demand nodes (or sinks) attract flow units. The total number of trips, which is referred to as *network throughput*, results from

$$\varphi^{\#} := \sum_{i \in \mathcal{V}} \varphi_i^+ = \sum_{j \in \mathcal{V}} \varphi_j^- = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} x_{(i,j)}^D = \sum_\rho h_\rho .$$

In what follows we introduce some matrices in order to describe the topological properties of a network. With respect to linear (and convex) programming these matrices will simplify the notation greatly.

- the node/arc incidence matrix $\mathbf{A}$ ($\mathbf{a}_e = e^{\text{th}}$ column of $\mathbf{A}$)
- the arc/path (or link/route) incidence matrix $\mathbf{\Delta}$
- the OD/path incidence matrix $\mathbf{\Lambda}$ ($\mathbf{\Lambda}_\rho = \rho^{\text{th}}$ column of $\mathbf{\Lambda}$)
- the origin/path incidence matrix $\mathbf{O}$
- the destination/path incidence matrix $\mathbf{D}$

The node/arc incidence matrix $\mathbf{A}$ describes at which nodes the arcs $e \in \mathcal{E}$ of the network start and at which node they end.

$$\mathbf{A} = (a_{ve}) \quad \text{with} \quad a_{ve} = \begin{cases} +1 & \text{if } v = \alpha(e) \\ -1 & \text{if } v = \omega(e) \\ 0 & \text{otherwise} \end{cases}$$

Given the above example we denote $e_1 = (v_1, v_2)$, $e_2 = (v_1, v_3)$, ...so that the matrix $\mathbf{A}$ has the form of Table 2.2 with $\ell = 6$ indicating the artificial link. (The variables u_e and w_v are needed for later purposes.) Each column belongs to an arc and each row refers to a node such that $a_{16} = -1$ and $a_{46} = 1$ represents the artificial link e_ℓ leading back from v_4 to v_1.

Table 2.2 The node/arc incidence matrix of the network in Figure 2.9.

A		arcs						
	a_{ve}	e_1	e_2	e_3	e_4	e_5	$e_\ell = e_6$	
nodes	v_1	1	1	0	0	0	−1	w_1
	v_2	−1	0	−1	1	0	0	w_2
	v_3	0	−1	1	0	1	0	w_3
	v_4	0	0	0	−1	−1	1	w_4
		u_1	u_2	u_3	u_4	u_5	u_6	

Substituting (2.22) into (2.23), flow conservation becomes in vector notation

$$\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}. \tag{2.27}$$

As pointed out before this equation describes the condition that the total inflow of a specific node equals the outflow.

In the next step we consider that different routes may use the same link. According to (2.24) arc flows $\boldsymbol{\varphi}$ are related to path flows $\mathbf{h}$ by

$$\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}, \tag{2.28}$$

where the arc/path incidence matrix $\boldsymbol{\Delta}$ is defined as follows:

$$\boldsymbol{\Delta} = [\Delta_{e\rho}] \quad \text{with} \quad \Delta_{e\rho} = \begin{cases} 1 & \text{if arc } e \text{ is part of route } \rho \\ 0 & \text{otherwise.} \end{cases}$$

The three paths $\rho_1 = (e_1, e_4)$, $\rho_2 = (e_2, e_5)$, and $\rho_3 = (e_2, e_3, e_4)$ as indicated by Figure 2.9 are then represented by the matrix of Table 2.3.

Table 2.3 The arc/path incidence matrix of the network in Figure 2.9.

$\boldsymbol{\Delta}$		paths		
	$\Delta_{e\rho}$	ρ_1	ρ_2	ρ_3
arcs	e_1	1	0	0
	e_2	0	1	1
	e_3	0	0	1
	e_4	1	0	1
	e_5	0	1	0

The third matrix account for multicommodity flows and is called the OD/path incidence matrix $\boldsymbol{\Lambda}$. Corresponding to (2.25) it relates the trip table $\mathbf{x}^D$ to the vector of path flows by

$$\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D, \tag{2.29}$$

where the elements of $\boldsymbol{\Lambda}$ result from[38]

$$\boldsymbol{\Lambda} = [\Lambda_{(i,j),\rho}] \quad \text{with} \quad \Lambda_{(i,j),\rho} = \begin{cases} 1 & \text{if route } \rho \text{ leads from } v_i \text{ to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Regarding traffic networks, (2.29) is frequently described as an identity. Each trip from an origin to a destination constitutes a part of the corresponding route flow. Nevertheless, (2.29) has a different interpretation when taking carriers into account. From an economic point of view $\mathbf{x}^S := \boldsymbol{\Lambda}\,\mathbf{h}$ may be seen as the (supplied) output of the network. As the network provider cannot sell more than its demand, we have $\mathbf{x}^S \leqq \mathbf{x}^D$. Supposing that the total demand $\mathbf{x}^D$ must be served implies $\mathbf{x}^S \geqq \mathbf{x}^D$. Both inequalities then yield $\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D$.

[38] With regard to Figure 2.9 the matrix $\boldsymbol{\Lambda}$ reduces to a row vector as there is only one OD-pair.

A route flow $\mathbf{h}$ is called *demand feasible*, if it is an element of the set[39]

$$H^D := \left\{ \mathbf{h} \mid \mathbf{h} \geqq \mathbf{0},\ \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D \right\}.$$

In accordance with (2.28) an arc flow $\boldsymbol{\varphi}$ is called *demand feasible*, if it is an element of the set

$$\Phi^D := \left\{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h},\ \mathbf{h} \in H^D \right\}.$$

For the sake of analytical convenience we shall ignore in almost all cases the fact that the number of trips must be integer. On the one hand, this simplification is not too hard when the number of movements is large. On the other hand, there is a great advantage assuming a strongly connected network. In this case both feasible sets are non-empty and convex. To show that Φ^D is convex we need

$$\boldsymbol{\varphi}^1, \boldsymbol{\varphi}^2 \in \Phi^D \implies \lambda\boldsymbol{\varphi}^1 + (1-\lambda)\boldsymbol{\varphi}^2 \in \Phi^D \quad \forall \lambda \in [0,1].$$

The left hand side implies that there is an $\mathbf{h}^1$ and an $\mathbf{h}^2$ with $\mathbf{\Delta}\,\mathbf{h}^1 = \boldsymbol{\varphi}^1$ and $\mathbf{\Delta}\,\mathbf{h}^2 = \boldsymbol{\varphi}^2$. Defining $\bar{\mathbf{h}} = \lambda\mathbf{h}^1 + (1-\lambda)\mathbf{h}^2$ yields $\mathbf{\Lambda}\,\bar{\mathbf{h}} = \mathbf{x}^D$. Thus, H^D is convex and

$$\bar{\boldsymbol{\varphi}} = \lambda\boldsymbol{\varphi}^1 + (1-\lambda)\boldsymbol{\varphi}^2 = \lambda\mathbf{\Delta}\,\mathbf{h}^1 + (1-\lambda)\mathbf{\Delta}\,\mathbf{h}^2 = \mathbf{\Delta}\,\bar{\mathbf{h}}$$

so that $\bar{\boldsymbol{\varphi}} \in \Phi^D$. Moreover H^D is compact because every route flow h_ρ connecting v_s and v_t is restricted to $0 \leqq h_\rho \leqq x^D_{(s,t)}$.

While demand feasibility emphasizes the minimum flow needed to serve $\mathbf{x}^D$, each link e may be limited to a maximum flow, which is called the capacity κ_e of that link. An arc flow $\boldsymbol{\varphi}$ is called *supply feasible*, if $\boldsymbol{\varphi}$ is an element of the convex and compact set

$$\Phi^S := \left\{ \boldsymbol{\varphi} \mid \mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa} \right\} \quad \text{or} \quad H^S := \left\{ \mathbf{h} \mid \mathbf{0} \leqq \mathbf{\Delta}\,\mathbf{h} \leqq \boldsymbol{\kappa} \right\}.$$

Of course an appropriate redefinition of the matrix $\mathbf{\Delta}$ can be applied to describe other input requirements by a flow h_ρ. If one trip along route ρ needs $\Delta_{e\rho}$ units of resource e, then $\mathbf{\Delta}_\rho h_\rho$ denotes the amount of inputs consumed by h_ρ. For example, airline passenger carriers or package distribution companies use this fact to calculate the consumed capacities of their vehicles.

The arc/path incidence matrix $\mathbf{\Delta}$ is also useful to compute the travel costs per route from the travel costs per arc. Let c_e be the cost to move one unit along link e with given properties such as distance, so that $\mathbf{c}$ is the vector of link costs per unit. The route cost per unit $\tilde{c}_\rho$ results from summing up all link costs comprised by the route.

$$\tilde{c}_\rho = \sum_{e \in \rho} c_e \quad \forall\, \rho \iff \tilde{\mathbf{c}} = \mathbf{\Delta}^{\mathsf{T}}\mathbf{c} \tag{2.30}$$

[39] The vertices of the set of demand feasible route flows H^D correspond to all-or-nothing assignments. That is, for one route in $\mathcal{R}(s,t)$ we have $h_\rho = x^D_{(s,t)}$ and $h_\rho = 0$ for the other routes.

In the example of Figure 2.9 the cost of traveling via the third route is $\tilde{c}_3 = c_2 + c_3 + c_4$. The same calculation applies to other *edge weights* or *values*, e.g., travel time or distance. Let $\boldsymbol{\tau}$ be the travel time per arc and let $\tilde{\boldsymbol{\tau}}$ be the travel time per route. Then $\tilde{\boldsymbol{\tau}} = \boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\tau}$. Accordingly $\tilde{c}_\rho$ or $\tilde{\tau}_\rho$ is sometimes called the value or weight of route ρ. A collection of the form $(\mathbf{G}, \mathbf{c})$ designates a *weighted digraph*.[40]

Eventually the use of the origin/path incidence matrix $\mathbf{O}$ and the destination/path incidence matrix $\mathbf{D}$ simplifies (2.26) in accordance with (2.25), such that

$$\mathbf{O}\,\mathbf{h} = \boldsymbol{\varphi}^{+} \quad \text{and} \quad \mathbf{D}\,\mathbf{h} = \boldsymbol{\varphi}^{-}\,. \tag{2.31}$$

Be aware that the two fixed vectors $\boldsymbol{\varphi}^{+}$ and $\boldsymbol{\varphi}^{-}$ reveal less information than a given trip table $\mathbf{x}^{D}$, which includes origin destination specific data. Together with the route flow $\mathbf{h}$, however, (2.31) incorporates the same information as $\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^{D}$.

Up until now the flows in a network are described by three concepts, (1) the inflows φ_v^- and outflows φ_v^+ for every node $v \in \mathcal{V}$, (2) the link flows φ_e for all arcs $e \in \mathcal{E}$, and (3) the path flows h_ρ for all routes ρ. However, in order to determine the bottlenecks of a network a further concept is needed – namely, the *cut of a network*.

A *cut* of a *flow network* $\mathcal{N} = ((\mathcal{V}, \mathcal{E}), \kappa, v_s, v_t)$ is a partition[41] $(\mathcal{S}, \mathcal{T})$ of the set of nodes $\mathcal{V}$ such that the start node and the end node satisfy $v_s \in \mathcal{S}$ and $v_t \in \mathcal{T}$; see Figure 2.10 with $v_s = v_1$ and $v_t = v_4$. The dashed curves just indicate the four possible decompositions of $\mathcal{V}$ into $(\mathcal{S}, \mathcal{T})$.

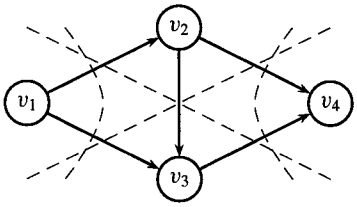


Fig. 2.10
Cuts of a network
The dashed lines indicate different cuts of the example network.

The *capacity of the cut* $(\mathcal{S}, \mathcal{T})$ is the sum of all capacities of links which start in $\mathcal{S}$ and end in $\mathcal{T}$.

$$\kappa(\mathcal{S}, \mathcal{T}) := \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \kappa_e$$

The first cut in Figure 2.10 with $\mathcal{S} = \{v_1\}$ and $\mathcal{T} = \{v_2, v_3, v_4\}$ then yields $\kappa(\mathcal{S}, \mathcal{T}) = \kappa_1 + \kappa_2$. Similarly, the *flow value (or weight) of the cut* $(\mathcal{S}, \mathcal{T})$ results from

$$\varphi(\mathcal{S}, \mathcal{T}) := \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \varphi_e.$$

A set of cuts may be used to partition a network (or area) into nonempty disjunct subsets of nodes, so-called centroids. This allows to study intra-regional and inter-regional traffic relationships, for instance, in the analysis of spatial price equilibria.

[40] Some authors require a network to be a weighted digraph.

[41] A *partition* or decomposition of a set $\mathcal{V}$ is a disjoint family of nonempty subsets of $\mathcal{V}$, the union of which is $\mathcal{V}$.

2.3.2 Flow Conservation (Kirchhoff's Law)

The flow conservation is one of the most important rules to be satisfied by a flow on a network. It requires that no flow unit entering the network or generated at a source of the network is lost within the network. It either leaves the network at some node or is attracted by some destination node. Regarding transshipment nodes, which neither attract nor generate flow units, we require that the sum of units entering that node equals the sum of units leaving it. This is equivalent to Kirchhoff's (Current) Law. Correspondingly, we require that all flow units generated by a source are eventually attracted by a sink. Now we can take advantage of the artificial link as introduced in Figure 2.9. The link flow $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_5, \varphi_\ell)^\top$ must hold the *(node) flow conservation rule*, which has been introduced in (2.23) as

$$\varphi_v^+ = \varphi_v^- \qquad \forall\, v \in \mathcal{V}.$$

Using (2.22), this expression can be rewritten in terms of a link flow vector $\boldsymbol{\varphi}$. If $\boldsymbol{\varphi}$ satisfies the above balance equations, it is called a *circulation*. The corresponding *(link) flow conservation* is given by (2.27), that is

$$\mathbf{A}\,\boldsymbol{\varphi} = \mathbf{0}.$$

The last column of the matrix $\mathbf{A}$ (see Table 2.2) refers to the artificial link e_ℓ. In accordance with the example of Figure 2.9 the first and the last condition denote the source ($v_s = v_1$) and the sink ($v_t = v_4$), respectively. The link flow restrictions and their equivalent node flow restrictions are

$$\begin{aligned} \varphi_1 + \varphi_2 - \varphi_6 = 0 &\iff \varphi_1^+ = \varphi_1^- \\ -\varphi_4 - \varphi_5 + \varphi_6 = 0 &\iff \varphi_6^- = \varphi_6^+ \end{aligned}$$

where the artificial link $\ell = 6$ requires $\varphi_6 = \varphi_1^- \equiv \varphi_4^+$. This is Kirchhoff's law. No flow unit generated at v_1 is lost in the network and is finally attracted by v_4, that is $\varphi_1^+ = \varphi_4^-$. In other words the system $\mathbf{A}\,\boldsymbol{\varphi} = \mathbf{0}$ is redundant. Summing up over the first m equations yields the flow φ_6 on the artificial edge. Ignoring the link capacities, we can fix φ_6 at an arbitrary value, say $\bar{b}$. Now take the last column of the matrix $\mathbf{A}$ and define $\mathbf{b} := \mathbf{a}_\ell \bar{b}$, where particularly $\varphi_s^+ = b_s = \bar{b} = -b_t = \varphi_t^-$ for the source v_s and the sink v_t. To be more concrete, in the example we have $\mathbf{b} = (\bar{b}, 0, 0, -\bar{b})^\top$. The system $\mathbf{A}\,\boldsymbol{\varphi} = \mathbf{0}$ can now be rewritten as

$$\mathbf{B}\,\boldsymbol{\varphi} = \mathbf{b}. \tag{2.32}$$

Contrasted with $\mathbf{A}$, the matrix $\mathbf{B}$ represents the node/arc incidence matrix without any fictitious link. In general b_v designates the net flow at node v, $b_v = \varphi_v^+ - \varphi_v^-$. The interpretation in detail is as follows:

- The supply generated at the source node $v_s = v_1$ is $b_1 > 0$.
- The supply arriving at the sink node $v_t = v_m$ is $b_m < 0$.
- All other nodes are transshipment nodes with $b_v = 0$.

The flow vector $\boldsymbol{\varphi}$ in (2.32) satisfies Kirchhoff's law if

$$\mathbf{1}^{\mathsf{T}}\mathbf{b} = \sum_{v\in\mathcal{V}} b_v = 0 \quad \text{or} \quad \sum_{v\in\mathcal{V}} \varphi_v^+ = \sum_{v\in\mathcal{V}} \varphi_v^- \quad \text{or} \quad \mathbf{1}^{\mathsf{T}}\boldsymbol{\varphi}^+ = \mathbf{1}^{\mathsf{T}}\boldsymbol{\varphi}^-.$$

The same concept applies to the case with more than one source or sink. Let $\mathcal{O}$, $\mathcal{I}$, $\mathcal{D}$ be a partition of the set of nodes $\mathcal{V}$.

$$\mathbf{B}\,\boldsymbol{\varphi} = \mathbf{b} \quad \text{with} \quad b_v \begin{cases} > 0 & v \in \mathcal{O} \text{ is a source } (\varphi_v^+ > \varphi_v^-) \\ = 0 & v \in \mathcal{I} \text{ is an intermediate node (flow conservation)} \\ < 0 & v \in \mathcal{D} \text{ is a sink } (\varphi_v^- > \varphi_v^+) \end{cases}$$

Of course, this generalization reduces to the previous case if we connect all sources to a super-origin v_s and all sinks to a super-destination v_t so that all sources and destinations become technically transshipment nodes; see Figure 2.11.[42]

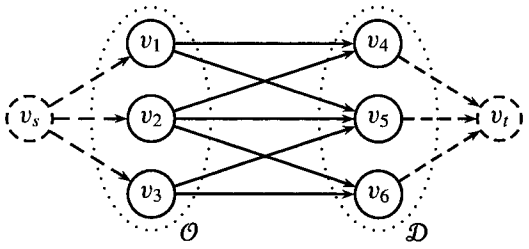


Fig. 2.11
An easy case of the transportation problem

The example shows a case with three supply sources $\mathcal{O} = \{v_1, v_2, v_3\}$, no transit node $\mathcal{I} = \emptyset$, and three consumer destinations $\mathcal{D} = \{v_4, v_5, v_6\}$.

Another version of flow conservation refers to the flow value of a cut.

2.22 Theorem (Cut Flow Conservation) *If the link flow vector $\boldsymbol{\varphi}$ is a* circulation, *that is the flow conservation* (2.23) *or* (2.27) *holds true, then*

$$\varphi(\mathcal{S}, \mathcal{T}) = \varphi(\mathcal{T}, \mathcal{S}).$$

Proof: The two relations of (2.22) give

$$\sum_{v\in\mathcal{S}} \varphi_v^+ = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \varphi_e + \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{S}}} \varphi_e,$$

$$\sum_{v\in\mathcal{S}} \varphi_v^- = \sum_{\substack{e:\ \alpha(e)\in\mathcal{T} \\ \omega(e)\in\mathcal{S}}} \varphi_e + \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{S}}} \varphi_e.$$

The sums on the left hand side are equal by the assumption of a circulation. The two middle sums yield the statement of the theorem. ■

[42] Figure 2.11 suggests a new class of graphs. If the set of nodes $\mathcal{V}$ can be partitioned into two disjoint subsets $\mathcal{V}_1$ and $\mathcal{V}_2$ such that every arc of $\mathcal{E}$ connects an element of $\mathcal{V}_1$ with one of $\mathcal{V}_2$ and the other way around, then $((\mathcal{V}_1, \mathcal{V}_2), \mathcal{E})$ is called *bipartite*.

One cut in the example of Figure 2.9 on p. 57 is $\mathcal{S} = \{v_1, v_3\}$ and $\mathcal{T} = \{v_2, v_4\}$. All flow units starting in v_1 and ending in v_4 must switch anywhere from $\mathcal{S}$ to $\mathcal{T}$ and in doing so they have to cross either e_1, e_3 or e_5. Assuming a circulation $\boldsymbol{\varphi}$, the theorem says that $\varphi_1 + \varphi_3 + \varphi_5 = \varphi_\ell$, where φ_ℓ is the return flow along the dotted arc.

In multicommodity networks each node can be a source, a sink and a transshipment node at the same time. One way to ensure the flow conservation anyhow has been introduced in (2.25), that is $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$ in comprehensive form. This *(path) flow conservation* states that each flow unit $x^D_{(s,t)}$ generated at the source v_s ends up in the destination node v_t. This multicommodity flow concept has an equivalent node flow formulation (2.34) as introduced in the section below.

2.23 Remark (Neural networks) Throughout the thesis we neglect an important class of "dynamic" networks that might be used to model the design of transport networks in consideration of its utilization. A neural network consists[43] of nodes that correspond to neurons and arcs that correspond to synaptic connections in the biological metaphor. Each node has a neural state described by a variable, say x_v. In the brain, this could be the potassium level; in computing applications, it could be anything the modeler chooses. Each arc has a weight w_e that affects the state of its neighboring nodes when firing. If the weight is positive, it is said to be excitatory; if it is negative, it is inhibitory. The neural states $\mathbf{x}$ change by some differential (or difference) equation depending on prevailing arc weights $\mathbf{w}$, say $\dot{\mathbf{x}} = F(\mathbf{x}, \mathbf{w}, t)$. Typically (but not necessarily), $-F$ is the gradient of an energy function (in keeping with the biological metaphor), say $F(\mathbf{x}, \mathbf{w}, t) = -\nabla_{\mathbf{x}} E(\mathbf{x}, \mathbf{w}, t)$, so that $\mathbf{x}(t)$ follows a path of steepest descent towards a minimum energy state. A learning mechanism L could consist of equations to change the weights: $\dot{\mathbf{w}} = L(\mathbf{x}, \mathbf{w}, t)$. Various learning mechanisms are represented this way, including a form of supervised learning that uses a training set to provide feedback on errors. Other elements can be learned besides the arc weights, including the topology of the network. □

2.3.3 Undirected Capacitated Networks

Up until now the analysis refers to digraphs $\mathbf{G}$ where all arcs are represented by ordered pairs of nodes. In telecommunications networks, however, edges such as copper cables or fiber cables are able to transmit data in both directions, that is $\{v_i, v_j\}$ and $\{v_j, v_i\}$ denote the same connection of two adjacent vertices. (Loops $\{v_i, v_i\}$ are ruled out.) Similarly a point-to-point communication follows some route (or many routes in the case of data packets) in the network with a mutual exchange of data. Nevertheless each flow unit or data bit moves into a certain direction. Thus every edge needs two variables describing the *bi-directional* flow. Regardless of their direction all flow units together determine the flow on a link which is restricted to the respective capacity. In road traffic networks one can think of two antiparallel lanes having a common capacity restriction.

[43] There is a vast literature on neural networks where the following exposition is given in Greenberg (1996–2003).

Let φ_{ij} denote the flow from node v_i to an adjacent node v_j. The non-negative total flow on the link $e = \{v_i, v_j\}$ is given by

$$\varphi_e = \varphi_{ij} + \varphi_{ji} \geqq 0. \tag{2.33}$$

Although appealing, a net-flow concept with $\varphi_{ij} - \varphi_{ji}$ is in almost all cases useless. For example, think of data traffic during a telephone call. Of course the capacity κ_e restricts the total flow on link e.

$$\varphi_e \leqq \kappa_e \qquad \forall\, e \in \mathcal{E}$$

In contrast to (2.22), the inflow φ_v^- and the outflow φ_v^+ of a node $v \in \mathcal{V}$ is given by

$$\varphi_v^- = \sum_{i \in \mathcal{V}} \varphi_{iv} \quad \text{and} \quad \varphi_v^+ = \sum_{i \in \mathcal{V}} \varphi_{vi},$$

where $\varphi_{ij} = 0$ if the two end nodes are not adjacent. Again sources and destinations satisfy $\varphi_v^- < \varphi_v^+$ and $\varphi_v^- > \varphi_v^+$, respectively. Notice that both types of nodes can also serve as transshipment nodes. The flow conservation requires $\varphi_v^- = \varphi_v^+$ for pure transshipment nodes.

Multicommodity flows: In accordance with undirected edges all routes joining an OD-pair (v_s, v_t) connect also (v_t, v_s). Hence, the flow on that route must be distinguished by the direction of flow units. In doing so we may split each path ρ into two directed routes ρ' and ρ'' such that $\rho' \in \mathcal{R}(s,t)$ and $\rho'' \in \mathcal{R}(t,s)$. These routes are again encoded by the OD/path incidence matrix $\mathbf{\Lambda}$ and we regain

$$\mathbf{x}^D = \mathbf{\Lambda}\,\mathbf{h} \quad \Longleftrightarrow \quad x^D_{(s,t)} = \sum_{\rho \in \mathcal{R}(s,t)} h_\rho \quad \forall\, (s,t) \in \mathcal{B}.$$

Note with respect to links that $e \in \rho'$ if and only if $e \in \rho''$. This property must be taken into account by the edge/path incidence matrix $\mathbf{\Delta}$.

$$\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h} \quad \Longleftrightarrow \quad \varphi_e = \sum_{\rho:\, e \in \rho} h_\rho \quad \forall\, e \in \mathcal{E}$$

According to (2.33) the sum includes both $h_{\rho'}$ and $h_{\rho''}$. Eventually the capacity restrictions become again $\mathbf{\Delta}\,\mathbf{h} \leqq \boldsymbol{\kappa}$.

Some authors prefer an immediate relationship between the trip table $\mathbf{x}^D$ and the link flows $\boldsymbol{\varphi}$. Unfortunately in almost all cases the matrix $\mathbf{\Delta}$ is not invertible so that $\mathbf{h}$ cannot be eliminated directly from the above system. In order to retain the flow conservation, assume that all flow units carry their destination address v_t. Then, $\varphi_{ij}^{(t)}$ is the flow from v_i to v_j with destination v_t and the flow conservation of node v_s becomes

$$\sum_{j \in \mathcal{V}} \varphi_{sj}^{(t)} = x^D_{(s,t)} + \sum_{j \in \mathcal{V}} \varphi_{js}^{(t)} \qquad \forall\, s \in \mathcal{V},\ s \neq t. \tag{2.34}$$

The left hand side gives the outflow which equals the inflow on the right hand side plus the generated traffic. That is, some flow units with destination v_t are originated

in v_s and other units use v_s as transshipment node. In the special case of $s = t$ one can define $-x^D_{(t,t)} = \sum_{j \in \mathcal{V}} \varphi^{(t)}_{jt} \geqq 0$ as the amount of traffic that leaves the network at sink v_t.

2.24 Example Consider the directed network in Figure 2.9 on p. 57 and let $\mathbf{B}$ be the corresponding node/arc incidence matrix in Table 2.2 *without* the artificial link ℓ. Assume now that each link is bi-directional. That means, every link has two flows in opposite directions with the destination $v_t = v_4$. For each of the $m = 5$ edges this aspect can be expressed by two flow variables $\varphi^{(t)}_e = \varphi^{(t)}_{ij}$ and $\varphi^{(t)}_{m+e} = \varphi^{(t)}_{ji}$. By analogy the expanded node/arc incidence matrix is defined as $\bar{\mathbf{B}} := (\mathbf{B}, -\mathbf{B})$. Now consider that each node can generate trips $x^D_{(i,t)} \geqq 0$ $(i \neq t)$ with destination v_t. When $x^D_{(t,t)} \leqq 0$ denotes the traffic exiting at v_t the nodal flow conservation becomes

$$\bar{\mathbf{B}}\, \boldsymbol{\varphi}^{(t)} = \mathbf{x}^D_{(t)},$$

where $\sum_{i \in \mathcal{V}} x^D_{(i,t)} = 0$. In a multicommodity network this condition must be satisfied for every destination $t \in \mathcal{V}$. The capacity restriction requires for the total flow on any link $e = \{v_i, v_j\}$

$$\varphi_e = \sum_{t \in \mathcal{V}} \varphi^{(t)}_{ij} + \varphi^{(t)}_{ji} \leqq \kappa_e.$$

These restrictions will be important for the routing problem of Section 3.2.5.4. □

The system $\boldsymbol{\Delta}\, \mathbf{h} \leqq \boldsymbol{\kappa}$, $\boldsymbol{\Lambda}\, \mathbf{h} = \mathbf{x}^D$, $\mathbf{h} \geqq \mathbf{0}$ can also be applied to data transmission in computer networks. The idea is to segment messages into packets and send them along feasible routes to their destination. One possible objective is to minimize the average delay of data packages. Furthermore, in this setting nodes are usually able to store packets in queues for each outgoing channel until they become free. See Bertsekas, Gallager (1992) for detailed information on these so-called store-and-forward networks. One way to consider node capacities has been shown in Figure 2.8.

2.25 Summary The topology network is encoded in the edge/path incidence matrix $\boldsymbol{\Delta}$ and the OD/path incidence matrix $\boldsymbol{\Lambda}$. The only difference between directed and undirected graphs is here the appropriate definition of available routes. Regarding link capacities the direction of flow units has no importance. One of the most important features of transport networks is the conservation of flows. □

2.3.4 Basic Problems for Transport Networks

Network analysis deals with a great number of questions economists want to answer. Although most of these problems cannot be solved independent of the others it is useful to introduce some classifications revealing the characteristics of transport networks. For instance, Chapter 3 examines processes on networks with a given structure, that is in particular the number of nodes and links as well as the corresponding capacities are held fixed. Aspects of "favorable" variations of network structures and their repercussion on network processes will be discussed with respect to network design in Chapter 5.

Regarding a given network, it is meaningful to distinguish between path related and flow related problems followed by numerous further reaching questions.

1. Routing problems seek for "favorable" paths or *cycles*[44] in a network.
One prevalent class, with many variations, is vehicle routing. Here the task is to find optimal delivery routes from one or more depots to a set of geographically scattered points (e.g., population centers). A simple case is detecting proper routes for snow removal, garbage collection, or street sweeping. Without complications, this is akin to the perhaps easiest routing problem: the *shortest path problem* is to determine paths of minimum length, minimum travel time, minimum travel cost, or of maximum reliability. The underlying networks may have a unique source and a unique sink or many OD-pairs.
In its most complex form, the vehicle routing problem is a generalization of the *traveling salesman problem* (TSP), as it can include additional time and capacity constraints, precedence constraints, plus more. The TSP is to find a tour of minimum total cost, where a tour is a permutation of n given points in consideration of a link cost vector $\mathbf{c}$. The nodes can be cities and the permutation corresponds to the visit of each city exactly once before returning to the first city (called home). The cost of a tour is the sum of its link costs. The *Chinese postman problem* differs from the TSP in that the postman can traverse an edge more than once.

2. Flow problems have two aspects, namely, the quantities of flows as well as their transport costs.
The *maximum flow problem* deals with the question of how much can be transported from a unique source to a unique sink on a capacitated network.[45] We shall see that this model is complementary to the shortest path problem. The *multicommodity flow problem* abstracts from maximizing flows between many OD-pairs. Instead it determines whether many simultaneous flows of given size can be realized on a network with limited feasibility. This model identifies each OD-pair with one commodity.
Having detected feasible flows, the next step is to determine "favorable" flows. The *transportation problem* is based on one set of supply nodes and one set of demand nodes (bipartite network). It then searches for a cost minimal flow that ships from warehouses to customers. Regarding the more general multicommodity flow problem all nodes can be origins and destinations of different commodity flows at the same time. The task is again to find a demand and supply feasible flow that ships from origins to destinations at minimum transportation cost.

3. Location problems take into account that many network participants need to concentrate their actions around a few if not one center. Thus, when market activities are performed at dispersed points in space, firms and households are interested in finding "favorable" places for their residence. The *facility location problem* in

[44] A *cycle* is a closed path where the initial and the terminal node coincide. Thus a loop is a cycle of length one.

[45] *Circulation problems* are concerned with networks having only transshipment nodes, see Figure 2.9 with an existing arc from v_4 to v_1. One task is to detect flows that are compatible with lower and upper flow limits on every arc.

particular is to specify the location and sizing of facilities (plants, warehouses, hospitals, etc.) at some nodes of a network. One objective is to minimize the sum of fixed charges for opening various plants and the transportation costs, possibly with respect to a given location pattern of competitors; cf., e.g., Gabszewicz, Thisse (1986) for further details.

It is clear that almost all solutions to the above cited problems depend on the given structure of the underlying network. Consequently, we have to ask whether "favorable" modifications of the network exist so that the processes can be performed in a less restrictive way. Unfortunately, network adjustments must be aware of a long list of dynamic effects.

(a) Former solutions respond to new restrictions determined by the adjusted network structure. For example, some drivers may switch to other preferred routes or additional drivers may enter the improved network.
(b) A growing demand for network services requires a continuous or at least discrete upgrading of the network (e.g., growing cities).
(c) Location decisions indicate that concentrations of economic activities may shift to new places (e.g., location of an airport).

To sum up, forecasts of future travel demand depend on future economic activities scattered through space (e.g., land-use activities) and future transport supply, i.e., the future transport system. Transport planning aims at the design of a transport system to serve the future pattern of transport demand.

4. Several classes of decision-makers determine what network processes take place. Consequently, economists wants to know who bears the cost of network related decisions. At least four categories of cost must be distinguished.

1. Variable costs directly related to trips or other network services are generally borne by the trip-makers. Typical examples are time needs and fuel consumption per trip.
2. Fixed travel costs following, for example, from car ownership are born by the respective vehicle owner. That may be a private household or a firm providing network services. In the latter case these costs are directly shifted to consumers via fares.
3. Fixed costs that are related to the construction (roads, cables, airports, etc.) and maintenance of an infrastructure network generally fall to the network carrier who is paid for the network provision in different ways. The easiest case refers to privately owned networks, where customers pay for services. Regarding publicly owned networks, taxpayers usually bear at least parts of the costs even when they do not consume any services.
4. External costs which are related to the infrastructure or the use of it, must frequently be borne by third parties. As an example, we refer to traffic noise and other types of pollution that affect persons external to the network. Another case is concerned with delay costs due to queuing that varies with the capacity utilization of each arc, i.e., congestion. Delays may also pertain to waiting times of passengers at railway stations.

2.3.5 Decomposition of Transportation Network Planning

2.3.5.1 Sequential Procedure of Travel Forecasting

Following Stopher, Meyburg (1975), the conventional transportation planning process as depicted in Figure 2.12 consists of four interacting models of travel demand, which may be preceded by a further stage of land-use forecasts. Because of the importance of traffic conditions – essentially times and costs of travel – to traffic demand there should be a feedback loop joining the steps which ensures that the sequential results are internally consistent and give an equilibrium solution. The corresponding behavioral approach should then represent the decisions that network users make when confronted with alternative choices, cf. Domencich, McFadden (1975). The subsequent step of transport planning is also neglected; it aims at the design of transport systems to serve future patterns of transport demand and focuses on the instruments that network providers are able to control and that are likely to affect all of the decisions of the network participants. Bear in mind that the four-step procedure has the importance of a pedagogical device rather than a profound hierarchical background. This point is underscored in Boyce (2002) and we shall come back to it after some explanations on the submodels.

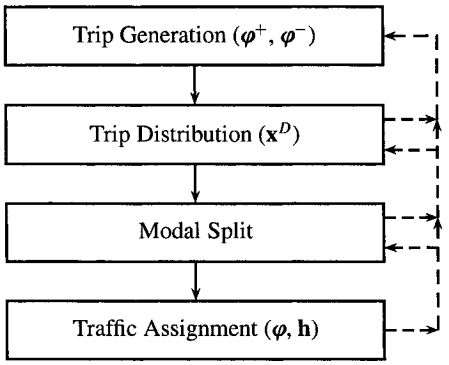


Fig. 2.12
Sequential procedure of travel forecasting
The dashed lines indicate that the sequential procedure is to be solved with feedback.

Trip generation essentially aims at the estimation of traffic volumes. Firstly, we try to obtain the number of trips originated at any node (or areal unit) $v \in \mathcal{V}$. The result is a vector, namely, the *trip production* φ^+. Secondly, we need the number of trips terminated at any node $v \in \mathcal{V}$. This vector represents the *trip attraction* φ^- and is of particular interest for competing destinations such as shopping centers. Thirdly, a check for consistency

$$\sum_{v \in \mathcal{V}} \varphi_v^+ = \sum_{v \in \mathcal{V}} \varphi_v^- = \varphi^\#$$

yields the total number of trips observed. Generated and attracted trips are usually set up as descriptive or mechanical rather than behavioral functions of socioeconomic and locational structure. Such an approach is non-behavioral as long as it does not

allow for travelers' decisions concerning travel frequency. We shall see indeed that the costs of traveling – which reflect the traffic conditions obtained from the transportation system, the traffic assignment as well as administrative measures such as congestion pricing or parking charges – have strong impact on the frequency of trips.

Trips may be classified by different categories of purpose. With regard to an activity based travel demand, Ben-Akiva, Bowman (1998) motivate the daily activity pattern by survival, social encounters and ego gratification. As almost all trips return daily to the home base for rest and personal maintenance, dwelling may be seen as the most important pole in Figure 2.13. Most trips are caused by commuters (journeys to work) who can try to combine different activities in one trip so that all poles are connected by corresponding links. As it is not difficult to encounter numerous examples of different tours, take Figure 2.13 as it is and bear in mind that travel demand is derived from activity demand, the analysis of which goes beyond the scope of this book. One further aspect, however, must be mentioned. Economic

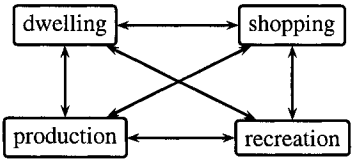


Fig. 2.13
Activity based traffic demand

activities are not only scattered through *space* but also through *time*. Temporal aspects of traveling such as the departure (or arrival) time and the duration of a trip cannot be neglected because peak periods are the major causes of congestion where simultaneous trips overwhelm the network capacities. Although congestion delays can at least partially be avoided through appropriate rescheduling of traffic, most network designers are particular interested in the traffic volumes during busy hours to plan capacities.

Trip distribution has the task to analyze the trip production φ^+ and the trip attraction φ^- with respect to the number of OD-trips $x^D_{(s,t)}$ originated at v_s and destined for v_t. The result is a *trip table* $\mathbf{x}^D$ being restricted to (2.26). According to the *gravity model*, trip distribution is often made on the basis of a trade-off between relative attractiveness of potential destination points and the costs or impedance of travel. The alternative approach of *entropy maximization* goes back to Wilson (1970); it describes in a non-causal manner the most likely trip distribution and ignores any behavioral aspects. This problem dissolves into approaches concerning *random utility maximization*, where in words of Ben-Akiva, Lerman (1985b) the utility of any alternative is, from the analysist's point of view, best viewed as random variable due to a variety of observational deficiencies. Domencich, McFadden (1975) dinguish between (binary or multinomial) linear, logit, and probit models, where the authors focus on the analytically more convenient logit approach. Corresponding probit models are discussed in Daganzo (1979).

A special problem of trip distribution is the case where different types of network users – fast passenger cars versus slow trucks for instance – share the same network

resources so that the assumption of a homogeneous flow must be given up; cf. Dafermos (1972). The decisive fact on *multiclass traffic* is that the demand for trips of different classes possess different functional relationships which are represented by class specific demand functions as well as class specific travel cost functions; cf. Nagurney, Dong (2002a). Due to the distinct transport requirement, priorization may be used to improve network performance, e.g., trucks are not allowed to use certain roads at certain times.

Modal split is directed at the distribution of OD-specific traffic demands $x^D_{(s,t)}$ among competing traffic modes, e.g., private cars or public transit. Mode choice may be seen as a function of the level of service[46] (including the accessibility of a destination) delivered by alternative modes to their customers. In this sense modes may be seen as *substitutes*, although changes in transportation policy variables may have little impact on transport modes, but substantially affect the distribution of trips. As an example Domencich, McFadden (1975, p. 8) refer to a downtown parking charge that may divert far fewer travelers from private car to public transit than from downtown to competing shopping locations. Furthermore, a trip can include different *complementary* modes. A traveler may switch at a certain station from his private car to public transportation and vice versa. For computational purposes it may be convenient to split such a combined journey into separate trips. As a consequence we have to bear in mind that some trips cause others after they are finished, see Figure 2.14. The need for synchronization of network processes will be discussed with respect to discrete event dynamic system (DEDS), see Chapter 6.

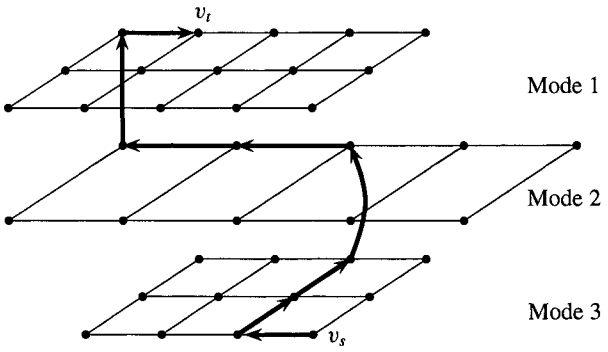


Fig. 2.14
Modal split

Modal split particularly takes into account that trip-makers have the opportunity to switch between alternative modes. Against that multiclass user models focus on the sharing of network elements with different characteristics of the classes. Observe, however, that different modes also can make use of the same network (e.g., busses and cars). Nevertheless, some nodes or links can be dedicated to special modes (e.g., bus stations or bus lanes). In other cases networks just overlap, as is the case of

[46] Wilson (1976) and Deneubourg et al. (1979) propose for a dynamic model using bifurcation analysis that describes the fluctuations between choices of tranportation modes as a function of mode attractiveness.

road and streetcar systems. Figure 2.14 represents the case where each mode has its own network (e.g., roads, railroads, airways). Certain points of interconnection allow to switch between the modes. Interconnection is also an important question with regard to other markets of network services. Telecommunications networks, for example, can belong to different competing carriers who may agree to interconnect their networks at certain nodes, say peers. Using these contacts, carriers extent the scope of services offered to their customers. A service can then start in one network, pass another one, and terminate in a third network just as indicated in Figure 2.14.

Trip assignment is a procedure to forecast the traffic loads on a network of transportation facilities given predicted travel demands comprised in the trip table $\mathbf{x}^D$. Each number of trips $x^D_{(s,t)}$ must be assigned to the alternative routes connecting the origin v_s with the destination v_t. The result is twofold. While path flows $\mathbf{h}$ are of particular interest from the travelers' point of view, the network designer is usually more interested in the corresponding link flows $\boldsymbol{\varphi}$ to detect bottlenecks. Nevertheless, bottlenecks cannot be dissolved by separating singular links or nodes from the rest of the network. In fact trip assignment distinguishes two procedures:

System optimality The OD-trips $\mathbf{x}^D$ are allocated to available routes with the objective to minimize total travel cost, where interactions between flows sharing the same links or nodes are taken into account.

User optimality Travelers choose the routes on their trying to minimize their private travel costs.[47] Contrasted with system optimality, individual network users tend to ignore their own impacts on other trip-makers.

If paths partially overlap the trip-makers have the possibility to switch to other routes at diversion nodes depending on events occurring during the trip. This aspect will almost always be neglected throughout the book; an exception is discussed with regard to survivable networks, where failures of network elements require rerouting. The term "rerouting" refers to communications networks where traffic assignment is called *routing*. It is performed by the network provider such that the distinction between system and user optimal traffic assignment becomes obsolete.

According to Hillier, Lieberman (1995, p. 212), operations research uses a slightly different denomination of problems that are related to trip distribution. One requirement of the *transportation problem* is that the way in which units are distributed from each origin to each destination is known in advance, so that the corresponding trip cost per unit can be determined. Sometimes, however, the best distribution is not clear because of the possibility of transshipment, whereby shipments can go through intermediate transfer points. The extension of the transport problem to include the routing decisions is referred to as the *transshipment problem*.

Locational decisions have been neglected so far what may be justified for the transportation analysis on a day-to-day basis. In the long run, however, decisions on

[47] The concept of *multicriteria* traffic network distinguishes usually two criteria, namely, travel time and travel cost. A subjective weighting of these criteria then gives rise to introduce a multiclass, multicriteria network equilibrium as can be found in Nagurney (2000b, 2002a).

the location or relocation of economic activities change the demand for transportation services. Households decide on residental and work location and firms choose places for their production facilities. Similarly, network carriers adjust the structure of the transport network so that future land use and the corresponding travel behavior are hard to predict. As an example, think of the introduction of a new rapid transit line modifying the supply side of the transport markets by substitution effects. Due to improved accessibility, it is then likely that different economic activities are located along this line.

2.3.5.2 Sequential Decision Process of Travel Demand

A network equilibrium on the market for trips is not easy to define because the demand refers to trips, and capacity (or network performance in condensed from) refers to link loads. Nevertheless, Beckmann et al. (1956) demonstrate how to solve this problem by means of a user optimal traffic assignment. An equilibrium in a transport system has been arrived when the prevailing demand for transportation gives rise to traffic conditions, i.e., times and costs of trips that will maintain that same demand. Taking the network performance as given, let us go back to Figure 2.12 in order to provide a model of individual travel demand. The revised behavioral decision sequence may take the form as depicted in Figure 2.15; cf. also Beckmann, Golob (1974) and Domencich, McFadden (1975). It represents the idea that the choice on each decision level is conditioned on fixed preceding decisions and optimal succeeding decisions.

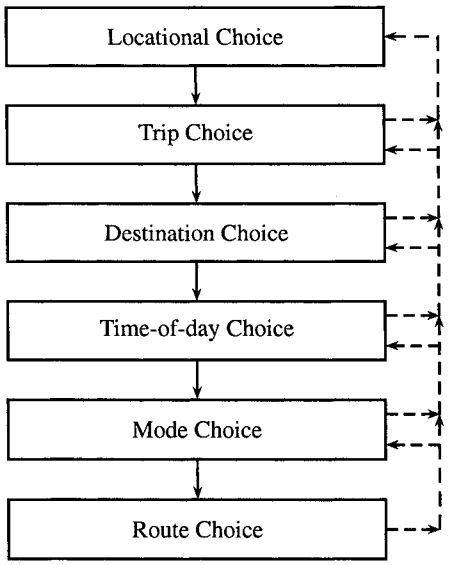


Fig. 2.15
Sequential decision process
The dashed lines indicate the flow of information that is needed to guide optimal choices at each level. Moreover, tâtonnement processes towards an equilibrium may require to revise earlier decisions.

This sequence differs from Figure 2.12 particularly by substituting trip generation and trip distribution by behavioral approaches representing the travelers' trip

and destination choice. Network users decide whether or not to make a trip at what frequency. Afterwards they decide where and when to realize the underlying activities. As an example, one can imagine households that decide to go shopping at off-peak hours, while the different shopping centers compete with each other for these customers.

Given the route choice in a network equilibrium, no traveler has an incentive to change today (not retrospective) any of his anterior decisions with respect to the fixed network properties. This travel behavior can be observed by the network designer so that he is able to compute the utilization of network resources including the dimension of congestion at bottlenecks. Provided the network carrier is able to forecast the result of the sequential decision process in Figure 2.15, he should also be able to predict the consequences of network modifications on network loading and throughput. In this ideal case an optimal network design seems to be within reach and future travel demand would face a well-suited transport network.

In the light of Figure 2.15, Boyce (2002) suggests to start with the equilibrium route choice problem as the *core problem* of the whole travel forecasting process. Traffic assignment represents the utilization of network resources, say road congestion, assuming all other determinants of a network equilibrium to be fixed. The corresponding equilibrated path flow pattern $\hat{\mathbf{h}}$ emphasizes the demand side and can be converted into a link flow pattern $\hat{\boldsymbol{\varphi}}$ stressing the supply side. Link flows, in turn, determine the travel costs and the travel times per link (as functions of link utilization) and thus per trip. Afterwards the model must be resolved when relaxing the supposition of a fixed demand, i.e., the trip table $\mathbf{x}^D$, by a price sensitive demand with regard to mode choice, time-of-day choice, and origin-destination choice. The next step would be to tackle the long-term problem of locational choice. On the basis of this predicted aggregate behavior of all network participants, investment planning takes place in order to adjust or to improve the fixed network structure. This procedure is also pursued in the present book. We start in Section 3.2 with a given network and a fixed trip table and solve a sequence of traffic assignment problems by relaxing one assumption after the other. Section 4.2 introduces a price sensitive demand and Chapter 5 abandons the assumption of a fixed network.

In doing so we present also several combined models, the origins of which can be found in the literature on transportation economics. We mention here only a few of them. One of the first combined approaches is given in Evans (1976). With reference to her Ph.D. thesis (1973), she assumes fixed trip generation (i.e., $\boldsymbol{\varphi}^+$ and $\boldsymbol{\varphi}^-$) and uses an entropy model to solve the problems of traffic assignment and trip distribution in one step; see also Florian et al. (1975). Lam, Huang (1992) present such a model with the addition of multiple user classes. Florian, Nguyen (1978) expand trip assignment by the submodels of trip distribution and modal split. Abdulaal, LeBlanc (1979b) combine route choice with mode choice. Similarly, Dafermos (1982) integrates traffic assignment, mode choice and a price sensitive travel demand which includes trip choice as well as destination choice. The corresponding sensitivity analysis with respect to small changes on the supply side of the network is proposed in Dafermos, Nagurney (1984). An "all in one" model for large-scale networks is discussed by Safwat et al. (1988). Here, the trips

generated from every origin and attracted by every destination are described as a function of socioeconomic variables and of accessibility, which is some composite measure of perceived utility when traveling to some destination. The trip distribution follows a logit model. According to a Wardrop user equilibrium model, the authors assume that each user chooses the mode and route combination that minimizes total perceived cost from node of origin to node of destination. Although more recent approaches exist, we terminate the enumeration by referring to Chen (1999) who puts the submodels of travel choice into a dynamic context.

2.26 Remark The estimation of trip matrices usually rely on link traffic counts. Sherali et al. (1994) enumerate three approaches, namely, statistical estimation methods, models based on maximum entropy theory, and network equilibrium based techniques. In particular the maximum entropy (Definition 4.3) approach attempts to determine the most likely OD-trip table which regenerates the observed link flows in a non-causal manner. In contrast, the behavioral approach based on the traffic equilibrium principles of Wardrop regenerates observed costs. Fisk (1989) proposed a combined distribution and assignment technique, which also relies on link flows; see Section 4.2.3. Furthermore, Fisk has shown that the network equilibrium approach, the maximum entropy approach, and the combined distribution assignment formulation can be expected to produce the same results under network congestion conditions with respect to an equilibrium flow pattern. □

2.4 Summary

The present chapter has rendered essential preparatory work for the subsequent network analysis. Starting with nodes and edges as integral components of all transport systems, Section 2.3 has shown how to describe the structure of networks by corresponding incidence matrices. Given a network, we have learned to distinguish at least three classes of flows, namely, link flows, path flows, and origin-destination (OD) flows. All of these concepts describe different aspects of network activities, where conservation of flows (Kirchhoff's current law) is the most important rule for consistency. The problem of scarce network resources that are shared by several flows becomes apparent by introducing capacitated networks. This restricts the set of feasible network activities and leads to the core problem of traffic assignment which will be discussed in detail in the next two chapters. In doing so the technical framework will be linear and non-linear programming as has been introduced in Section 2.2. These instruments allow to characterize optimal solutions and, furthermore, to perform sensitivity analysis for basic settings. Given a capacitated network we shall start with finding technically efficient network activity patterns. Afterwards the analysis is refined to cost minimal activities that conform to different concepts of network equilibria. The assumption of a fixed network structure with given capacities for the network components will be relaxed in Chapter 5. Here, the task is to find an appropriate network design that fits to the needs of traffic demand.

3

Utilization of Transport Networks: Basics

3.1 Preliminary Examples

3.1.1 Max Flow–Min Cut Theorem

In this chapter we analyze aspects of a technically and economically efficient usage of a given transport network and its implications for pricing transport services. Here, technical efficiency refers to output efficiency: a vector of outputs is said to be technically output efficient for a given input vector when no output can be increased without raising any input. One example for determining technically efficient activities is the following problem of finding the maximum flow on a given network.

Having detected output efficient activities the next task is to produce a given output at minimum cost. In other words, we seek for the cheapest of all input efficient activities (with respect to variable inputs) that produces at least a given level of output. That is the minimum cost problem in the section below. For further details, applications, and algorithms see Ahuja et al. (1993).

Contrasted with this short term problem, Chapter 5 discusses aspects of network design. Essentially, network optimization is the task of accommodating a given demand for transportation at minimum inputs, i.e., input efficiency. One problem will be that the networks of interest mainly consist of fixed components which cannot be adjusted arbitrarily. As explained before, network optimization is in the end a problem of determining a cost minimal investment such that a given output can be realized.

The economic analysis of networks in terms of production theory starts with a distinct class of networks. The idea is to concentrate on capacitated directed networks $\mathcal{N} = ((\mathcal{V}, \mathcal{E}), \kappa, v_s, v_t)$ producing one and only one output. Similar to Figure 2.9 the networks concerned are restricted to having a unique source node v_s with no inflow and a unique sink node v_t with no outflow. Thus, the source and the sink offend flow conservation. This is the point where the artificial arc $e_\ell = (v_t, v_s)$ comes in. Based upon the augmented set of arcs $\mathcal{E}' = \mathcal{E} \cup \{e_\ell\}$ we seek a flow vector

$\varphi = (\varphi_1, \ldots, \varphi_m, \varphi_\ell)^\top$ such that every node including the sink and the source satisfies the flow conservation.

The output of this circulation is now measured by flow units moving along the artificial arc. As the capacities κ_e, $e \in \mathcal{E}$, of the (existing) links are fixed, the flow φ_ℓ cannot be increased arbitrarily.

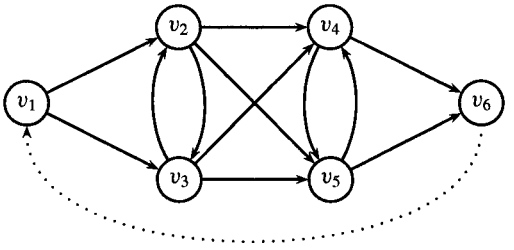


Fig. 3.1
Maximum flow problem
The capacities κ of the arcs are not indicated. The source of the flow network is v_1 and the sink is v_6.

Regarding the arc e_ℓ, we make the frequently used

3.1 Assumption *The capacity κ_ℓ of the artificial arc e_ℓ leading from the sink v_t to the source v_s is sufficiently high, so that $\varphi_\ell < \kappa_\ell$. When needed, this capacity can be increased at zero cost. Moreover, the virtual transportation along this arc can be done at zero cost.*

Let $\kappa = (\kappa_1, \ldots, \kappa_m, \kappa_\ell)^\top$. Then the *flow network* is the collection $\mathcal{N} = ((\mathcal{V}, \mathcal{E}'), \kappa, v_s, v_t)$. Starting with a flow network $\mathcal{N}$, the maximum flow problem requests the maximum number of units which can be transported from the source to the sink. With the link e_ℓ the problem can be written as follows, where the first restriction reflects the flow conservation of all nodes.

$$\begin{aligned} &\max \ \varphi_\ell \\ &\text{s.t. } \varphi_v^+ = \varphi_v^- \qquad \forall v \in \mathcal{V} \\ &\qquad 0 \leqq \varphi_e \leqq \kappa_e \quad \forall e \in \mathcal{E}' \end{aligned}$$

Switching to vector notation the corresponding linear program (3.1.P) is called the *maximum flow problem.* Theorem 3.2 will show that the dual program (3.1.D) seeks the minimum capacity over all cuts which have been illustrated in Figure 2.10 on p. 62. This is the reason to name it the *minimum cut problem.* Example 3.3 below will explain the interrelationship between the two problems on the basis of numerical values.

(3.1)
$$\begin{array}{llll} \text{(P)} & \max & \varphi_\ell & \\ & \text{s.t.} & \mathbf{A}\varphi = \mathbf{0} & \\ & & \varphi \leqq \kappa & \\ & & \varphi \geqq \mathbf{0} & \end{array} \qquad \begin{array}{llll} \text{(D)} & \min & \kappa^\top \mathbf{u} & \\ & \text{s.t.} & \mathbf{A}^\top \mathbf{w} + \mathbf{u} & \geqq \mathbf{e}_\ell \\ & & \mathbf{w} & \text{unrestr.} \\ & & \mathbf{u} & \geqq \mathbf{0} \end{array}$$

In accordance with (2.27), $\mathbf{A}\varphi = \mathbf{0}$ describes the conservation of flow. The primal objective function may be written as $\mathbf{e}_\ell^\top \varphi$, where $\mathbf{e}_\ell$ is the ℓ^{th} unit vector appearing

on the left hand side of the dual restriction. As the corresponding components of $\boldsymbol{\varphi}$ and $\boldsymbol{\kappa}$ have the same dimension, i.e., flow units on the edge concerned, the respective dual variable u_e may be interpreted as *relative* shadow price which is assigned to the capacity of edge e.

$$[u_e] = \frac{\text{flow unit on edge } \ell}{\text{flow unit on edge } e}$$

In an optimal solution, however, each flow unit on edge e is eventually led back via edge ℓ. Thus there is no need to differentiate between units on edge e and units on edge ℓ. The flow units may be interpreted as one homogeneous good. Consequently, u_e has no dimension. Indeed, the following analysis shows that the optimal value $\hat{u}_e$ turns out to be either zero or one. Hence, increasing κ_e either raises φ_e (and in the end φ_ℓ) on a one-to-one basis or it has no effect by the respective complementary slackness condition. This argument becomes void in cost minimization. In this case the value of a particular flow unit depends on the link where it is observed.

While the dual variable u_e is assigned to edge e, the dual variable w_v refers to node v. The same reasoning as before says that the dual variable $\mathbf{w}$ as well as the elements of $\mathbf{A}$ are dimensionless. Bear in mind, however, that $\mathbf{A}$ transforms link flows into node flows. As can be seen from the matrix $\mathbf{A}$ in Table 2.2 on p. 59, which includes also the variables $\mathbf{u}$ and $\mathbf{w}$, each column describes the properties of the respective link. Similarly, each row shows the properties of a node.

Let $\boldsymbol{\varphi}$ be a feasible flow and $(\mathbf{w}, \mathbf{u})$ be a feasible solution to the minimum cut problem. Then the weak duality (Theorem 2.2) states $\varphi_\ell \leqq \boldsymbol{\kappa}^\mathsf{T}\mathbf{u}$, which is especially valid for an optimal primal solution $\hat{\varphi}_\ell = \mathbf{e}_\ell^\mathsf{T}\hat{\boldsymbol{\varphi}}$ and an optimal dual solution $(\hat{\mathbf{w}}, \hat{\mathbf{u}})$.

The corresponding strong duality theorem is reflected by the Max Flow–Min Cut Theorem given below. It states that the maximum flow problem (3.1.P) has a finite optimal solution, if and only if there is an optimal solution to the minimum cut problem (3.1.D). The maximum flow equals the minimum cut.

$$\hat{\varphi}_\ell = \boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}}$$

The discussion of an optimal primal solution $\hat{\boldsymbol{\varphi}}$ and an optimal solution $(\hat{\mathbf{w}}, \hat{\mathbf{u}})$ starts with an analysis of the dual restrictions. The main properties of $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{w}}, \hat{\mathbf{u}})$ will be summarized in (3.2). Notice that the dual variables $\mathbf{w}$ are unrestricted with respect to their sign. According to Table 2.2, p. 59, the extensive form of the dual restrictions is

$$\begin{aligned} w_{\alpha(e)} - w_{\omega(e)} + u_e &\geqq 0 \qquad e \in \mathcal{E}, \\ w_n \quad - w_1 \qquad\quad &\geqq 1 \qquad e = \ell. \end{aligned}$$

Evaluating the complementary slackness conditions regarding all ordinary arcs yields

$$\begin{aligned} (\hat{\varphi}_e - \kappa_e)\,\hat{u}_e &= 0 \qquad e \in \mathcal{E} \quad \text{(primal restriction)}, \\ (\hat{w}_{\alpha(e)} - \hat{w}_{\omega(e)} + \hat{u}_e)\,\hat{\varphi}_e &= 0 \qquad e \in \mathcal{E} \quad \text{(dual restriction)}. \end{aligned}$$

By Assumption 3.1, the remaining artificial link satisfies $\kappa_\ell > \varphi_\ell$, so that

$$\begin{aligned} &\hat{\varphi}_\ell < \kappa_\ell \implies \hat{u}_\ell = 0, \\ &(\hat{w}_n - \hat{w}_1 - 1)\,\hat{\varphi}_\ell = 0. \end{aligned}$$

In the next step we have to distinguish three cases with respect to $0 \leqq \hat{\varphi}_e \leqq \kappa_e$ for all $e \in \mathcal{E}$.

$$\begin{aligned} 0 = \hat{\varphi}_e < \kappa_e &\Longrightarrow \hat{u}_e = 0,\ \hat{w}_{\alpha(e)} \geqq \hat{w}_{\omega(e)} \\ 0 < \hat{\varphi}_e < \kappa_e &\Longrightarrow \hat{u}_e = 0,\ \hat{w}_{\alpha(e)} = \hat{w}_{\omega(e)} \\ 0 < \hat{\varphi}_e = \kappa_e &\Longrightarrow \hat{u}_e \geqq 0,\ \hat{w}_{\alpha(e)} \leqq \hat{w}_{\omega(e)} \end{aligned}$$

Inverting the argument of the third case yields on the one hand

$$\left[\hat{u}_e > 0 \Longrightarrow \hat{\varphi}_e = \kappa_e > 0 \Longrightarrow \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}\right] \Longrightarrow \hat{w}_{\alpha(e)} < \hat{w}_{\omega(e)}.$$

On the other hand, we have $\boldsymbol{\kappa} > \mathbf{0}$ so that the value of the objective $\boldsymbol{\kappa}^{\mathrm{T}}\mathbf{u}$ cannot take a minimum for $\hat{u}_e > \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)} > 0$. Hence,

$$\hat{w}_{\alpha(e)} < \hat{w}_{\omega(e)} \Longrightarrow \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)} > 0.$$

The preceding two statements together give

$$\hat{w}_{\alpha(e)} < \hat{w}_{\omega(e)} \Longleftrightarrow \left[\hat{u}_e > 0 \quad \text{and} \quad \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}\right].$$

Now suppose $\hat{w}_{\alpha(e)} > \hat{w}_{\omega(e)}$. This inequality has two implications.

1. It contradicts $\hat{u}_e > 0$ so that

   $$\hat{w}_{\alpha(e)} > \hat{w}_{\omega(e)} \Longrightarrow \hat{u}_e = 0.$$

   In summary, link values $\mathbf{u}$ are connected to node values $\mathbf{w}$ by

   $$\hat{u}_e = \max\left\{\hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}, 0\right\} \qquad e = 1,\ldots,m.$$

2. The assumption $\hat{w}_{\alpha(e)} > \hat{w}_{\omega(e)}$ together with a positive flow $\hat{\varphi}_e > 0$ is contradicted by

   $$\hat{\varphi}_e > 0 \Longrightarrow \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)} < 0.$$

To sum up, one has for all links $e \in \mathcal{E}$

(3.2a) $\quad \hat{w}_{\alpha(e)} > \hat{w}_{\omega(e)} \Longrightarrow 0 = \hat{\varphi}_e < \kappa_e \Longrightarrow 0 = \hat{u}_e > \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}$

(3.2b) $\quad 0 < \hat{\varphi}_e < \kappa_e \Longrightarrow 0 = \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}$

$$\Longrightarrow \hat{w}_{\omega(e)} = \hat{w}_{\alpha(e)}$$

(3.2c) $\quad \hat{w}_{\alpha(e)} < \hat{w}_{\omega(e)} \Longrightarrow 0 < \hat{\varphi}_e = \kappa_e \Longrightarrow 0 < \hat{u}_e = \hat{w}_{\omega(e)} - \hat{w}_{\alpha(e)}.$

Note that the case

$$\hat{w}_{\alpha(e)} = \hat{w}_{\omega(e)},\ \hat{u}_e = 0,\ \hat{\varphi}_e = \kappa_e$$

cannot be excluded. An example will be given below.

As the system $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ is redundant (summing up over the first m equations yields the flow φ_ℓ on the artificial edge), one w_v, $v = 1,\ldots,n$, can be set to an arbitrary value. Thus, we set for the sink node

$$\hat{w}_n = 1.$$

Assuming that the max flow problem has a positive solution $\hat{\varphi}_\ell > 0$, the corresponding value of the source node turns out to be $\hat{w}_1 = 0$. In contrast to (3.2b) the artificial link is characterized by $\hat{w}_n = 1$, $\hat{w}_1 = 0$, $\hat{\varphi}_\ell > 0$, and $\hat{u}_\ell = 0$.

The next step is not an immediate consequence of the above reasoning. By intuition, however, zero flows in (3.2a) and flows below the capacity in (3.2b) tend to be of minor importance for the primal solution. Thus, we draw our attention to (3.2c) with positive limited flows $\hat{\varphi}_e = \kappa_e > 0$ and positive shadow prices $\hat{u}_e > 0$. Recall that the terms $\kappa_e \hat{u}_e$ determine the value of the dual objective function. At an optimal flow any flow unit starts at the source with $\hat{w}_1 = 0$, uses a sequence of arcs with at least one congested link $\hat{u}_e$ due to at least one bottleneck, and ends up at the sink with $\hat{w}_n = 1$. On its way the flow unit "observes" a non-decreasing sequence of node values $\hat{w}_e$. The decisive observation is that the values can only jump when the corresponding two nodes are connected by a congested link. When every route leaves the region around the starting node following such a congested link and enters the region around the end node, it is natural to define cuts separating the source node $v_1 \in \mathcal{S}$ with $\hat{w}_1 = 0$ and the sink node $v_n \in \mathcal{T}$ with $\hat{w}_n = 1$ by

$$\mathcal{S} = \left\{ v \in \mathcal{V} \,|\, w_v \leqq \hat{w}_1 = 0 \right\} \quad \text{and} \quad \mathcal{T} = \left\{ v \in \mathcal{V} \,|\, w_v > \hat{w}_1 = 0 \right\}. \tag{3.3}$$

Observe that no such cut permits the case of (3.2b): the flow is either zero or at its capacity limit.

Consider an arc e ($e \neq \ell$) pointing from $\mathcal{S}$ to $\mathcal{T}$, then

$$\left.\begin{array}{l} \alpha(e) \in \mathcal{S} \\ \omega(e) \in \mathcal{T} \end{array}\right\} \Longrightarrow \hat{w}_{\alpha(e)} \leqq \hat{w}_1 = 0 < \hat{w}_{\omega(e)} \overset{(3.2c)}{\Longrightarrow} \hat{\varphi}_e = \kappa_e, \hat{u}_e > 0. \tag{3.4}$$

In the opposite case the arc $e \neq \ell$ points from $\mathcal{T}$ to $\mathcal{S}$ such that

$$\left.\begin{array}{l} \alpha(e) \in \mathcal{T} \\ \omega(e) \in \mathcal{S} \end{array}\right\} \Longrightarrow \hat{w}_{\omega(e)} \leqq \hat{w}_1 = 0 < \hat{w}_{\alpha(e)} \overset{(3.2a)}{\Longrightarrow} \hat{\varphi}_e = 0, \hat{u}_e = 0. \tag{3.5}$$

The flow conservation of Theorem 2.22 together with $\omega(\ell) = v_s \in \mathcal{S}$ and $\alpha(\ell) = v_t \in \mathcal{T}$ states for the cut (3.3)

$$0 = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \hat{\varphi}_e - \sum_{\substack{e:\ \alpha(e)\in\mathcal{T} \\ \omega(e)\in\mathcal{S}}} \hat{\varphi}_e \iff \hat{\varphi}_\ell = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \hat{\varphi}_e - \sum_{\substack{e:\ \alpha(e)\in\mathcal{T} \\ \omega(e)\in\mathcal{S} \\ e\neq\ell}} \hat{\varphi}_e.$$

From (3.5) we know that each term of the last sum equals zero. Observing (3.4) we gain

$$\hat{\varphi}_\ell = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \hat{\varphi}_e = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S} \\ \omega(e)\in\mathcal{T}}} \kappa_e = \kappa(\mathcal{S}, \mathcal{T}).$$

Thus, $\hat{\varphi}_\ell$ corresponds to a cut value $\kappa(\mathcal{S}, \mathcal{T})$. To be more concrete, it is the minimum cut value.

3.2 Theorem (Max Flow–Min Cut Theorem) *The maximum flow $\hat{\varphi}_\ell$ equals the minimum capacity of all cuts,*

$$\hat{\varphi}_\ell = \sum_{\substack{e\in\mathcal{E}\\ e\neq\ell}} \hat{u}_e\kappa_e = \sum_{\substack{e:\ \alpha(e)\in\mathcal{S}\\ \omega(e)\in\mathcal{T}}} \kappa_e = \kappa(\mathcal{S},\mathcal{T}).$$

Proof: The Strong Duality Theorem 2.4 states $\hat{\varphi}_\ell = \boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}}$. Every cut can be described by an appropriate (0,1)-vector $\mathbf{u}$ being dually feasible. As has been shown above the rule

$$\hat{u}_e = \begin{cases} 0, & \text{if } \hat{\varphi}_e < \kappa_e \\ 1, & \text{if } \hat{\varphi}_e = \kappa_e \end{cases}$$

for all $e \in \mathcal{E}'$ implies $\boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}}$ being a cut value. As $\boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}} \leqq \boldsymbol{\kappa}^\mathsf{T}\mathbf{u}$ for all feasible $\mathbf{u}$ and particularly all cuts, $\boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}}$ is the minimum cut value. The corresponding node values of the optimal cut $(\mathcal{S},\mathcal{T})$ are $\hat{w}_v = 0$ for all $v \in \mathcal{S}$ and $\hat{w}_v = 1$ for all $v \in \mathcal{T}$. ■

Suppose that the optimal value of the maximum flow problem (3.1.P) is finite and let $\hat{\varphi}_\ell(\boldsymbol{\delta})$ be the optimal value when the capacities $\boldsymbol{\kappa}$ are disturbed by $\boldsymbol{\delta}$. In accordance with (2.20) the program has a unique Karush-Kuhn-Tucker (KKT) vector $\hat{\mathbf{u}}$ if and only if $\hat{\varphi}_\ell$ is differentiable at $\boldsymbol{\delta} = \mathbf{0}$, in which case $\hat{\mathbf{u}}$ follows from

$$\frac{\partial\hat{\varphi}_\ell(\mathbf{0})}{\partial\delta_e} = \hat{u}_e \qquad \forall\, e \in \mathcal{E}.$$

This sensitivity analysis gives rise to the suspicion that upgrading the capacity of a link of the minimum cut increases the optimum value by the same amount, that is $\hat{u}_e = 1$. Raising the capacity of a link with idle capacities has no effect on $\hat{\varphi}_\ell(\mathbf{0})$, that is $\hat{u}_e = 0$.

3.3 Example The above perturbation effect does not apply in general as shown in the following example of Figure 3.2, where $\hat{\mathbf{u}}$ is not unique. The numbers assigned to the arcs denote the given capacities. The circle segments indicate cuts. There is more than one basic feasible solution – namely, two vertices – representing the minimum cut or the maximum flow $\hat{\varphi}_\ell = 4 = \boldsymbol{\kappa}^\mathsf{T}\hat{\mathbf{u}}$. The primal solution is $\hat{\boldsymbol{\varphi}} = (4, 2, 2, 2, 2, 4)^\mathsf{T}$ for both cases. The dual solutions can be characterized as

First cut: $\hat{\varphi}_\ell = 4 = \hat{\varphi}_1 = \kappa_1, \hat{\mathbf{w}} = (0, 1, 1, 1, 1)^\mathsf{T}, \hat{\mathbf{u}} = (1, 0, 0, 0, 0, 0)^\mathsf{T}$
The degeneration is reflected by $\hat{u}_2 = 0 = \hat{\varphi}_2 - \kappa_2$ and $\hat{u}_3 = 0 = \hat{\varphi}_3 - \kappa_3$. The vector $\hat{\mathbf{w}}$ is represented by the cut $(\mathcal{S},\mathcal{T})$ in Figure 3.2, that is $\mathcal{S} = \{v_1\}$ and $\mathcal{T} = \{v_2, v_3, v_4, v_5\}$. There is one link pointing from $\mathcal{S}$ to $\mathcal{T}$, namely, e_1 with $\hat{u}_1 = 1$; cf. (3.4).

Second cut: $\hat{\varphi}_\ell = 4 = \hat{\varphi}_2 + \hat{\varphi}_3 = \kappa_2 + \kappa_3, \hat{\mathbf{w}} = (0, 0, 1, 1, 1)^\mathsf{T}, \hat{\mathbf{u}} = (0, 1, 1, 0, 0, 0)^\mathsf{T}$
The degeneration is reflected by $\hat{u}_1 = 0 = \hat{\varphi}_1 - \kappa_1$. The right cut in Figure 3.2, that is $\mathcal{S} = \{v_1, v_2\}$ and $\mathcal{T} = \{v_3, v_4, v_5\}$, corresponds to the new vector $\hat{\mathbf{w}}$. Two links point from $\mathcal{S}$ to $\mathcal{T}$, where $\hat{u}_2 = 1$ and $\hat{u}_3 = 1$.

□

From an economic point of view the maximum flow problem describes a technical problem. The aim is to produce the maximum output φ_ℓ at given inputs $\boldsymbol{\kappa}$. In standard

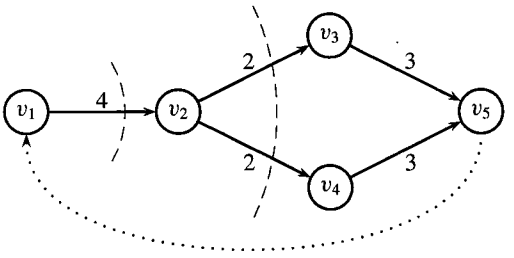


Fig. 3.2
Minimum cut problem
The given capacities imply two optimal cuts, which are indicated by the dashed circle segments.

microeconomic theory this relationship is represented by the *production function* f. The function f determines the *maximum output* producible at given amounts of the factors of production. This fact is referred to as technical (output) efficiency. Hence, the maximum flow problem (3.1.P)

$$f(\boldsymbol{\kappa}) = \max\left\{\varphi_\ell \mid \mathbf{A}\boldsymbol{\varphi} = \mathbf{0},\ \mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}\right\}, \tag{3.6}$$

often abbreviated to $\hat{\varphi}_\ell = f(\boldsymbol{\kappa})$, corresponds to finding an output efficient activity, where

- φ_ℓ is the (final) output of the network flow;
- κ_e, $e \in \mathcal{E}$, are fixed inputs;
- φ_e, $e \in \mathcal{E}$, may be seen as intermediate goods[1] needed to produce the final output φ_ℓ;
- the flow conservation $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ reflects the production technology of the network.

As pointed out in Economides (1996), it is inherent in the structure of a network that many components of a network are required for the provision of a typical service. Thus, network components are *complementary* to each other. This case is shown on the left hand side of Figure 3.3. If one component $(v_1, v_2, v_3, e_1, e_2)$ fails, no output φ_ℓ is producible. Given the link capacities κ_1 and κ_2 and ignoring the nodes, the isoquant has the typical right angled shape as it is known from a linear limitational production function with the maximum output given by $\hat{\varphi}_\ell = \min\{\kappa_1, \kappa_2\}$.

The right hand part of Figure 3.3 shows, however, that the complementarity of network components is not the only conceivable relationsship. For example, two parallel links may act like perfect *substitutes*. A similar reasoning holds good for alternative routes connecting the same OD-pair; see, e.g., Figure 3.2. Both cases will be important with respect to the *reliability of networks*. In view of failing network components one task of network design is to ensure that traffic can reach a given destination at all. This aim of connectivity in abnormal network states can be realized up to a certain degree by constructing parallel routes, which can be substituted by each other. Some aspects of the survivability of networks and especially the costs of building "oversized" networks will be discussed in Section 5.1.

[1] Technically, the production function of each link may be seen as $\varphi_e = \min\{\varphi(\alpha(e)), \kappa_e\}$, where $\varphi(\alpha(e))$ is the amount of flow units initiated at the start node of the link e. As long as the fixed input κ_e exceeds $\varphi(\alpha(e))$ we have $\varphi_e = \varphi(\alpha(e))$.

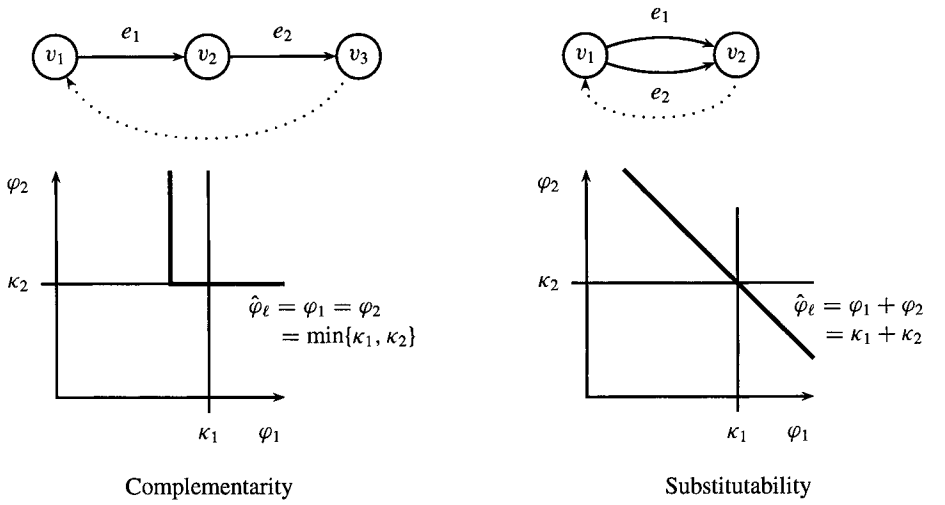


Fig. 3.3 Relationship between link flows

The maximum flow problem has been shown to represent the technical problem of producing the maximum output φ_ℓ at given inputs $\boldsymbol{\kappa}$. As long as we refer to one output this statement is commonly captured in microeconomics by the production function. It will be generalized to many commodities at a later stage of the analysis.

3.1.2 The Minimum Cost Problem for a Single Good

3.1.2.1 One Source and One Sink

The next step is to ask for the minimum cost of producing at least a given output $\bar{\varphi}_\ell$. At this point it is important to note that the link capacities $\boldsymbol{\kappa}$ are supposed to be given. No investment is made to adjust capacities. Therefore investment costs do not occur. This additional aspect will be taken into account in Section 5.1.4. Moreover, we ignore fixed costs which occur, for example, when the carrier rents the capacities. The reason is that fixed costs do not change an optimal solution for the short-termed cost minimization problem at hand. As a consequence the intermediate goods $\boldsymbol{\varphi}$ which are needed to produce the final output φ_ℓ are the only cause of production costs. Let c_e be the constant cost of shipping one unit of φ_e along the arc e ($e \in \mathcal{E}'$) and note that $c_\ell = 0$ by Assumption 3.1. Then $\mathbf{c}^\mathsf{T}\boldsymbol{\varphi}$ is the total transportation cost where $\varphi_1, \ldots, \varphi_m$ are used to produce φ_ℓ. One result of cost minimization should be clear in advance: while a solution $\hat{\boldsymbol{\varphi}}$ to the maximum flow problem can include positive flows on cycles excluding the artificial link, this cannot be cost minimal for $\mathbf{c} > \mathbf{0}$. As an example take the two links connecting the nodes v_4 and v_5 in Figure 3.1. A cost minimal flow will never use both links at the same time. This is a special problem of assignment procedures because avoiding cyclic flows in large networks requires large computational requirement; cf. Janson, Zozaya-Gorostiza (1987).

The aim is to produce at least a given output $\bar{\varphi}_\ell \geqq 0$ at minimum transportation cost, where the link capacities κ are fixed.

$$\begin{aligned} \min\ & \mathbf{c}^\mathsf{T}\boldsymbol{\varphi} \\ \text{s.t. } & \varphi_v^+ = \varphi_v^- && \forall v \in \mathcal{V} \\ & 0 \leqq \varphi_e \leqq \kappa_e && \forall e \in \mathcal{E}' \\ & \varphi_\ell \geqq \bar{\varphi}_\ell \end{aligned}$$

Notice that some of the subsequent programs additionally allow for minimum flows on each link. The corresponding multicommodity flow problem will be discussed at a later stage. Another problem is that no feasible solution exists when the capacities do not suffice to serve for the output $\bar{\varphi}_\ell$. In this case either $\bar{\varphi}_\ell$ must be reduced or the capacities have to be increased. This is the starting point of Section 5.1.4.

The first restriction of the preceding problem is again the flow conservation (2.27), which corresponds to $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ in linear programming. The second restriction determines the set of supply feasible flows $\Phi^S = \{\boldsymbol{\varphi} | \ \mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}\}$. Hence, the primal program (3.7.P) is the cost minimization problem to be solved. The dual program shows particularly the dual variables $\mathbf{q}$ and p, which will be used to value the corresponding primal restrictions. In order to interpret $\mathbf{q}$ as vector of non-negative shadow prices, it is convenient to use the negative primal restriction $-\boldsymbol{\varphi} \geqq -\boldsymbol{\kappa}$. As a matter of fact we know that the primal problem has a feasible solution only if the prespecified output $\bar{\varphi}_\ell$ does not exceed the maximum flow $\hat{\varphi}_\ell$.

(3.7)

$$\begin{array}{llll} \text{(P)} & \min\ \mathbf{c}^\mathsf{T}\boldsymbol{\varphi} & \text{(D)} & \max\ -\boldsymbol{\kappa}^\mathsf{T}\mathbf{q} + \bar{\varphi}_\ell \mathrm{p} \\ & \text{s.t. } \mathbf{A}\boldsymbol{\varphi} = \mathbf{0} & & \text{s.t. } \mathbf{A}^\mathsf{T}\mathbf{v} - \mathbf{q} + \mathbf{e}_\ell \mathrm{p} \leqq \mathbf{c} \\ & \boldsymbol{\varphi} \leqq \boldsymbol{\kappa} & & \mathbf{v} \quad \text{unrestr.} \\ & \varphi_\ell \geqq \bar{\varphi}_\ell & & \mathbf{q} \geqq \mathbf{0} \\ & \boldsymbol{\varphi} \geqq \mathbf{0} & & \mathrm{p} \geqq 0 \end{array}$$

In contrast to the minimum cut problem the dual variables $\mathbf{q}$, p, and $\mathbf{v}$ now have the dimensions

$$[\mathrm{v}_v] = \frac{€}{\text{flow unit in node } v}, \quad [\mathrm{q}_e] = \frac{€}{\text{flow unit on edge } e}, \quad [\mathrm{p}] = \frac{€}{\text{flow unit on edge } \ell}.$$

The difference compared to the former dual variables is simply that the value of a flow unit depends here on the link where they are observed. As has been mentioned before, the matrix $\mathbf{A}$ transforms node flows into link flows. Thus its elements a_{ve} must have the dimension *flow units in node v per flow unit on edge e*. An appealing interpretation of the dual variables is proposed in Ahuja et al. (1993). Assume that the dual restriction assigned to the flow on link $e \in \mathcal{E}$ is active. Then,

$$\mathrm{v}_{\alpha(e)} - c_e - \mathrm{q}_e = \mathrm{v}_{\omega(e)}$$

The *potential value* of a flow unit at the start node is reduced by the real transportation cost along the arc and by the value of the capacity unit consumed by the flow

unit. The result is a reduced *potential value* of the flow unit at the end node. This observation is reflected by Samuelson's (1954) construction of the *iceberg model*, which ignores capacity restrictions. Similar to an iceberg, a commodity looses some of its value using up resources in proportion to the transportation distance. In other words, the price of a commodity must rise with the distance between the place of production and the place of consumption.

Suppose that both problems (3.7.P) and (3.7.D) have feasible solutions. Then an optimal solution to (3.7.P), say $\hat{\boldsymbol{\varphi}}$, and an optimal solution to (3.7.D), say $(\hat{\mathbf{v}}, \hat{\mathbf{q}}, \hat{\mathrm{p}})$, satisfy

$$\mathbf{c}^{\mathsf{T}}\hat{\boldsymbol{\varphi}} = \hat{\mathrm{p}}\bar{\varphi}_\ell - \hat{\mathbf{q}}^{\mathsf{T}}\boldsymbol{\kappa} \tag{3.8}$$

by the strong duality theorem. An optimal solution, therefore, requires *zero profits*, that is the output $\bar{\varphi}_\ell$ evaluated at the shadow price $\hat{\mathrm{p}}$ equals the operating (or transport) cost $\mathbf{c}^{\mathsf{T}}\hat{\boldsymbol{\varphi}}$ plus the rental value of the capital stock $\hat{\mathbf{q}}^{\mathsf{T}}\boldsymbol{\kappa}$. Notice that this argument is based on the shadow prices $(\hat{\mathbf{q}}, \hat{\mathrm{p}})$, which may be seen as internal prices of the network.

Let $C_P(\boldsymbol{\kappa}, \bar{\varphi}_\ell)$ and $C_D(\mathbf{c})$ be the optimal values of the primal program (3.7.P) and the dual program (3.7.D), respectively. Suppose that both values are finite. Then the primal program has a unique KKT vector $(\hat{\mathbf{q}}, \hat{\mathrm{p}})$ if and only if C_P is differentiable at $(\boldsymbol{\kappa}, \bar{\varphi}_\ell)$ and

$$\frac{\partial C_P(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \bar{\varphi}_\ell} = \hat{\mathrm{p}} \geqq 0 \quad \text{and} \quad \frac{\partial C_P(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \kappa_e} = -\hat{\mathrm{q}}_e \leqq 0 \qquad \forall e \in \mathcal{E}.$$

Increasing the amount to be transported raises transport costs. Starting with higher capacities (at zero investment cost) reduces transport cost. The dual program has a unique KKT vector $\hat{\boldsymbol{\varphi}}$ if and only if C_D is differentiable at $\mathbf{c}$ and

$$\frac{\partial C_D(\mathbf{c})}{\partial c_e} = \hat{\varphi}_e \geqq 0 \qquad \forall e \in \mathcal{E}.$$

A higher link cost per unit increases the transport costs.

The complementary slackness conditions of Theorem 2.5 are

$$\begin{aligned}
(\hat{\varphi}_e - \kappa_e)\,\hat{\mathrm{q}}_e &= 0 \qquad e \in \mathcal{E}', \\
(\hat{\varphi}_\ell - \bar{\varphi}_\ell)\,\hat{\mathrm{p}} &= 0, \\
(\hat{\mathrm{v}}_{\alpha(e)} - \hat{\mathrm{v}}_{\omega(e)} - \hat{\mathrm{q}}_e - c_e)\,\hat{\varphi}_e &= 0 \qquad e \in \mathcal{E}, \\
(\hat{\mathrm{v}}_{\alpha(\ell)} - \hat{\mathrm{v}}_{\omega(\ell)} - \hat{\mathrm{q}}_\ell + \hat{\mathrm{p}} - c_\ell)\,\hat{\varphi}_\ell &= 0.
\end{aligned}$$

The second condition says that it cannot be cost minimal to produce more than needed, $\varphi_\ell > \bar{\varphi}_\ell$, if the output has a positive shadow price, $\hat{\mathrm{p}} > 0$. By Assumption 3.1 ($\hat{\varphi}_\ell < \kappa_\ell$ and $c_\ell = 0$), the first condition implies $\hat{\mathrm{q}}_\ell = 0$. At the same time this condition can be used to express the rental value of the capital stock $\hat{\mathbf{q}}^{\mathsf{T}}\boldsymbol{\kappa}$ by the flow value $\hat{\mathbf{q}}^{\mathsf{T}}\hat{\boldsymbol{\varphi}}$ so that (3.8) becomes

$$\hat{\mathrm{p}}\bar{\varphi}_\ell = (\mathbf{c} + \hat{\mathbf{q}})^{\mathsf{T}}\hat{\boldsymbol{\varphi}}.$$

On every link e the direct transport cost per unit c_e is increased by the shadow price $\hat{q}_e$. This correction, however, is zero if there are excess capacities, $\hat{\varphi}_e < \kappa_e$.

The same reasoning can be applied to the third complementary condition for all $e \neq \ell$.

$$\begin{aligned} \kappa_e > \hat{\varphi}_e > 0 &\Longrightarrow c_e &&= \hat{v}_{\alpha(e)} - \hat{v}_{\omega(e)} > 0 \\ \hat{q}_e > 0 \Longrightarrow \kappa_e = \hat{\varphi}_e > 0 &\Longrightarrow c_e + \hat{q}_e &&= \hat{v}_{\alpha(e)} - \hat{v}_{\omega(e)} > 0 \end{aligned}$$

Regarding Figure 3.3 we have to distinguish between complementary and substitutable network components. The case of two parallel links e_1 and e_2 connecting the same pair of nodes v_1 and v_2 yields

$$c_{e_1} + \hat{q}_{e_1} = c_{e_2} + \hat{q}_{e_2} = \hat{v}_{v_1} - \hat{v}_{v_2}.$$

In the second case of two links in a line we have

$$(c_{e_1} + \hat{q}_{e_1}) + (c_{e_2} + \hat{q}_{e_2}) = \hat{v}_{v_1} - \hat{v}_{v_3}.$$

These facts are similar (although not the same) to Ohm's law of electrical networks consisting of parallel resistors or series of resistors.

For the remaining artificial link ℓ with $c_\ell = 0$, $\kappa_\ell > \varphi_\ell > 0$, and $\hat{q}_\ell = 0$ the above complementary slackness condition gives

$$\hat{v}_{\alpha(\ell)} - \hat{v}_{\omega(\ell)} + \hat{p} = 0, \tag{3.9}$$

where $\alpha(\ell)$ and $\omega(\ell)$ are the sink v_t and the source v_s, respectively. This is an important result in the following sense. Every unit traveling from v_s to v_t uses a route ρ, which may be seen as a sequence of arcs or, equivalently, as a sequence of nodes. At every node the item passes, it has lost some of its value corresponding to the preceding link cost. But on the last artificial link when returning to its starting point (and its initial value) the value of every item is increased by the same amount $\hat{p}$ independently of what route it takes. In other words, every transport from v_s to v_t has the same value per flow unit.

$$\sum_{e \in \rho} (c_e + \hat{q}_e) = \sum_{e \in \rho} (\hat{v}_{\alpha(e)} - \hat{v}_{\omega(e)}) = \hat{v}_{\alpha(\rho)} - \hat{v}_{\omega(\rho)} = \hat{v}_{\omega(\ell)} - \hat{v}_{\alpha(\ell)} = \hat{p}$$

This observation is sometimes referred to as[2]

3.4 Proposition (Kirchhoff's Voltage Law) *Regarding all links of the network with a positive flow, the sum of all* potential differences $\hat{v}_{\alpha(e)} - \hat{v}_{\omega(e)}$ *(that is the voltage over the branch e) along a closed path is zero.*

Kirchhoff's Voltage Law has an immediate consequence, because a cost minimal flow $\hat{\varphi}$ requires that there is no route from v_s to v_t with lower transport cost than any used route. Thus, all flow units move along a cost minimal route with transport

[2] Further analogies to electric circuits including Ohm's law are discussed in Sasaki, Inouye (1974).

costs $\hat{\mathrm{p}}$ given in (3.9). Note that $\hat{\mathrm{p}}$ is sometimes referred to as *shortest path distance*[3], where the distance increases when the frictions along that route become heavier.

An easy network example to illustrate the above derived results is given in Figure 3.4. For $\mathbf{c} = (1, 3, 2, 3, 2, 0)^\mathsf{T}$, $\boldsymbol{\kappa} = (4, 2, 2, 3, 3, \infty)^\mathsf{T}$, and $\bar{\varphi}_\ell = 3$ an optimal primal solution is $\hat{\boldsymbol{\varphi}} = (3, 1, 2, 1, 2, 3)^\mathsf{T}$. To check the optimality use the dual solution $\hat{\mathbf{q}} = (0, 0, 2, 0, 0, 0)^\mathsf{T}$, $\hat{\mathbf{v}} = (9, 8, 5, 4, 2)^\mathsf{T}$, $\hat{\mathrm{p}} = 7$ and compare the values of the objective functions (17). The figure should be read in the following sense: $\hat{\mathrm{v}}_1 \to c_1 \to \hat{\mathrm{v}}_2 \to c_2 \to \hat{\mathrm{v}}_3 \to c_4 \to \hat{\mathrm{v}}_5 \to \hat{\mathrm{p}} \to \hat{\mathrm{v}}_1$. The second route has a bottleneck at κ_3, so that $\hat{\mathrm{v}}_1 \to c_1 \to \hat{\mathrm{v}}_2 \to c_3 + \hat{\mathrm{q}}_3 \to \hat{\mathrm{v}}_4 \to c_5 \to \hat{\mathrm{v}}_5 \to \hat{\mathrm{p}} \to \hat{\mathrm{v}}_1$.

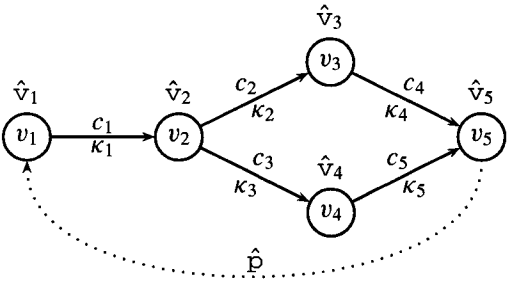


Fig. 3.4
Minimum cost problem
Only the capacity κ_3 has a positive shadow price $\hat{\mathrm{q}}_3$.

3.5 Remark Having introduced the production function f by (3.6), the corresponding *cost function* C results from (3.7.P) by definition, that is

$$C(\mathbf{c}, \boldsymbol{\kappa}, \bar{\varphi}_\ell) = \inf\left\{\mathbf{c}^\mathsf{T}\boldsymbol{\varphi} \mid f(\boldsymbol{\kappa}) \geqq \bar{\varphi}_\ell,\ \boldsymbol{\varphi} \geqq \mathbf{0}\right\}. \tag{3.10}$$

The infimum in this equation declares that, on the one hand, $C(\mathbf{c}, \boldsymbol{\kappa}, \bar{\varphi}_\ell) = +\infty$ if $f(\boldsymbol{\kappa}) < \bar{\varphi}_\ell$. If, on the other hand, $f(\boldsymbol{\kappa}) > \bar{\varphi}_\ell$, the cost minimal flow is not output efficient. The reason is simply that the fixed capacities cannot be reduced. Both cases are problems of network design. The first case requires to install additional capacity units on certain links. As the installation of capacity units incurs costs, network design should avoid cases with excess capacities, that is $\hat{\varphi}_e < \kappa_e$. These aspects will be discussed in Section 3.2.3 with respect to the multicommodity flow problem. □

3.6 Notational Remark The minimum cost problem (3.7) is sometimes discussed in a slightly different form: a firm producing its commodity at v_s wants to accommodate the demand y_m^D at the consumer destination v_m. In order to make this task more explicit, let us rewrite the flow conservation rule $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ with the last column of $\mathbf{A}$ and the last element of the link flow $\boldsymbol{\varphi}$ referring to the fictitious link ℓ. Using the notation of (2.32) $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ can be replaced by $\mathbf{B}\boldsymbol{\varphi} = \mathbf{b}$, where φ_ℓ has been eliminated from $\boldsymbol{\varphi}$. The primal problem of (3.7) now takes the form

$$\min\left\{\mathbf{c}^\mathsf{T}\boldsymbol{\varphi} \mid \mathbf{B}\boldsymbol{\varphi} = \mathbf{b},\ \boldsymbol{\varphi} \leqq \boldsymbol{\kappa},\ -b_m \geqq y_m^D,\ \boldsymbol{\varphi} \geqq \mathbf{0}\right\},$$

with $\mathbf{b} = (b_1, \ldots, b_m)^\mathsf{T}$ and $b_1 = -b_m$. The outflow $b_1 \geqq 0$ of the source v_1 equals the inflow $b_m \leqq 0$ of the sink v_m. Finally, the inflow must cover at least the demand

[3] The contradiction with regard to dimensions can be dissolved by assuming constant costs of shipping one flow unit along one distance unit.

y_m^D at the sink (consumer destination). Of course such a renaming of variables leaves the cost function untouched. □

3.1.2.2 Many Sources and Many Sinks

Up until now the output is supplied by one source as shown in Figure 3.2, p. 83. A first step towards a more general setting is to suppose several places of production. By analogy, the commodity is needed at different consumer destinations. Regarding this good there is no difference where it is produced or consumed. In particular, consumers do not prefer commodities with respect to their place of production (homogeneous good). The case of many distinguishable commodities will be discussed below.

The following models are variations of the minimum cost problem (3.7). They demonstrate the power of the approach as several requirements can be implemented at relatively little effort.

(1) Transportation problem. Find a flow of least cost that ships from supply sources to consumer destinations.[4] Each node is either a source, a transshipment node, or a sink. With regard to Notational Remark 3.6, we start with a decomposition of the basic flow relationship $\mathbf{B}\boldsymbol{\varphi} = \mathbf{b}$.

1. Supply nodes are collected in the set of origins $\mathcal{O}$ having no inflows: $\mathbf{B}_1\boldsymbol{\varphi} = \mathbf{b}_1$ (the elements of the matrix $\mathbf{B}_1$ are either 0 or 1; cf. Table 3.1 below).
   In addition each source $v \in \mathcal{O}$ has the capacity y_v^S, that is $\mathbf{b}_1 \leqq \mathbf{y}^S$.
2. The transshipment nodes $v \in \mathcal{I}$ satisfy $\mathbf{B}_2\boldsymbol{\varphi} = \mathbf{b}_2$ with $\mathbf{b}_2 = \mathbf{0}$.
3. The set of destinations $\mathcal{D}$ includes all demand nodes having no outflows: $\mathbf{B}_3\boldsymbol{\varphi} = \mathbf{b}_3$ (the elements of the matrix $\mathbf{B}_3$ are either 0 or -1).
   In addition the demand y_v^D of each sink $v \in \mathcal{D}$ must be served, $-\mathbf{b}_3 \geqq \mathbf{y}^D$.

The conservation of flows requires that all flows generated at some source end in some node, i.e., $\mathbf{1}^\top\mathbf{b}_1 + \mathbf{1}^\top\mathbf{b}_3 = 0$. The advanced cost minimization problem can now be stated as

$$
\begin{array}{lrll}
(3.11) & \min & \mathbf{c}^\top\boldsymbol{\varphi} & \text{(transport cost)} \\
& \text{s.t. } (\mathbf{1}^\top\mathbf{B}_1 + \mathbf{1}^\top\mathbf{B}_3)\boldsymbol{\varphi} = 0 & & \text{(flow conservation)} \\
& & \mathbf{B}_1\boldsymbol{\varphi} \leqq \mathbf{y}^S & \text{(supply)} \\
& & \mathbf{B}_2\boldsymbol{\varphi} = \mathbf{0} & \text{(transshipment)} \\
& & -\mathbf{B}_3\boldsymbol{\varphi} \geqq \mathbf{y}^D & \text{(demand)} \\
& & \boldsymbol{\varphi} \leqq \boldsymbol{\kappa} & \text{(link capacities)} \\
& & \boldsymbol{\varphi} \geqq \mathbf{0} &
\end{array}
$$

To see whether this problem has a feasible flow $\boldsymbol{\varphi}$, all sources and sinks may be connected to a super-origin v_s or super-destination v_t, respectively. This trick has

[4] When every source node is connected to every demand node and when there are no transshipment nodes and no capacity constraints the problem (3.11) is termed *Hitchcock's Transportation Problem*; see Sheffi (1985).

been illustrated in Figure 2.11, it is useful and admissible as long as we stick to the assumption of a homogeneous good. The capacities of the new links are set to $\kappa(v_s, v) = y_v^S$ for all $v \in \mathcal{O}$ and $\kappa(v, v_t) = y_v^D$ for all $v \in \mathcal{D}$. Every feasible flow in this augmented network requires that $\sum_{v \in \mathcal{D}} y_v^D$ determines a maximum flow. And this is the case when the capacity of every cut in the augmented network is at least as large as the maximum flow; see Theorem 3.2. Of course, a first check for plausibility is $\sum_{v \in \mathcal{O}} y_v^S \geqq \sum_{v \in \mathcal{D}} y_v^D$, for otherwise the cumulated production capacity does not suffice to match the cumulated demand. The corresponding supply-demand-theorem is proved in Gale (1957).

(2) Spatial Price Equilibrium. Suppose there are no transshipment points such that $\mathbf{B}$ consists of $\mathbf{B}_1$ and $\mathbf{B}_3$. Every node v is either a place of production (see v_1 and v_6 in Figure 3.5) with supply y_v^S or a destination with given demand y_v^D (see v_2–v_5 in Figure 3.5); such a graph is called *bipartite*. Moreover, each production node is connected to all consumer destinations. There is one homogeneous good. If a consumer buys one unit of the commodity produced at v he has to pay $c_e + p_v^S$, where e is the link that leads from plant v to the consumer place and p_v^S is the price of the commodity at v. Problems of this type are discussed in detail in Takayama, Judge (1971).

It is useful to go through one provisional example. Indeed, this case goes back

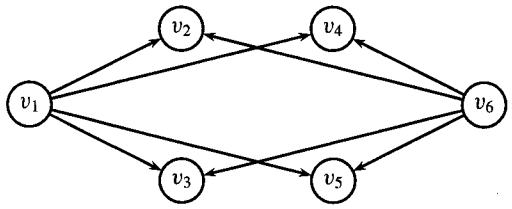


Fig. 3.5
Spatial price equilibrium
The length of the arrows do not necessarily indicate the distance between nodes.

to Samuelson (1952) who discusses a trade model of any number of regions each of which is characterized by an excess supply function. Omitting balanced regions he shows that net import regions never export to any region and, inversely, that a net export region never imports from any region. Hence, the above classification of demand and supply nodes is not an assumption but rather a result of Samuelson's approach.

The corresponding arc/path incidence matrix $\mathbf{B}$ is given in the following table. Observe that the first and the sixth row represent places of production with outflows only; they determine $\mathbf{B}_1$. The remaining rows belong to consumer destinations with inflows only; they constitute $\mathbf{B}_3$. Regarding the pair of linear programs below it is particularly worthwhile to draw our attention to the signs of the elements of the two matrices.

All consumers seek to satisfy their demand $\mathbf{y}^D$ at minimum cost which include transportation cost as well as direct expenditures for the commodity. The general problem of cost minimization (probably after renaming the variables) and its dual program take the form

Table 3.1 The node/arc incidence matrix of the network in Figure 3.5.

B		links								
	a_{ve}	1	2	3	4	5	6	7	8	
nodes	1	1	1	1	1	0	0	0	0	y_1^S, w_1
	2	−1	0	0	0	−1	0	0	0	p_2^D
	3	0	−1	0	0	0	−1	0	0	p_3^D
	4	0	0	−1	0	0	0	−1	0	p_4^D
	5	0	0	0	−1	0	0	0	−1	p_5^D
	6	0	0	0	0	1	1	1	1	y_6^S, w_6
		φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8	

(3.12)

$$\begin{array}{llll} \text{(P)} \min & \mathbf{c}^{\mathsf{T}}\boldsymbol{\varphi} + \mathbf{p}^{S\mathsf{T}}\mathbf{y}^S & & \\ \text{s.t.} & -\mathbf{B}_1\boldsymbol{\varphi} + \mathbf{y}^S = \mathbf{0} & & \\ & -\mathbf{B}_3\boldsymbol{\varphi} \geqq \mathbf{y}^D & & \\ & \boldsymbol{\varphi} \geqq \mathbf{0} & & \\ & \mathbf{y}^S \geqq \mathbf{0} & & \end{array} \qquad \begin{array}{llll} \text{(D)} \max & \mathbf{y}^{D\mathsf{T}}\mathbf{p}^D & & \\ \text{s.t.} & -\mathbf{B}_1{}^{\mathsf{T}}\mathbf{w} - \mathbf{B}_3{}^{\mathsf{T}}\mathbf{p}^D \leqq \mathbf{c} & & \\ & \mathbf{w} \leqq \mathbf{p}^S & & \\ & \mathbf{w} \quad \text{unrestr.} & & \\ & \mathbf{p}^D \geqq \mathbf{0} & & \end{array}$$

Let $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{y}}^S)$ and $(\hat{\mathbf{w}}, \hat{\mathbf{p}}^D)$ be optimal solutions for the primal and the dual program, respectively. As shown below the dual variables $\hat{\mathbf{p}}^D$ have the meaning of prices at consumer destinations. If all nodes of production have a positive supply $\mathbf{y}^S > \mathbf{0}$, then the optimal values $\hat{\mathbf{w}}$ coincide with the prices $\mathbf{p}^S$ at the supply nodes.

The evaluation of the complementary slackness conditions generates several statements. A positive supply at node v requires the optimal shadow price $\hat{w}_e$ to be equal to the given price p_v^S.

$$\hat{y}_v^S > 0 \implies \hat{w}_v = p_v^S$$
$$\hat{w}_v < p_v^S \implies \hat{y}_v^S = 0$$

If all supply nodes with $\hat{y}_v^S = 0$ are removed from the network, there is no need to distinguish between $\hat{\mathbf{w}}$ and $\mathbf{p}^S$. In this case the conditions on optimal link flows

$$\hat{\varphi}_e > 0 \implies -\hat{w}_{\alpha(e)} + \hat{p}_{\omega(e)}^D = c_e$$
$$-\hat{w}_{\alpha(e)} + \hat{p}_{\omega(e)}^D < c_e \implies \hat{\varphi}_e = 0$$

imply for consumer prices

$$\hat{p}_{\omega(e)}^D = c_e + p_{\alpha(e)}^S \quad \text{if} \quad \hat{\varphi}_e > 0.$$

This statement requires especially that there are no capacity restrictions regarding links or supply. As a positive demand $y_v^D > 0$ needs $\hat{\varphi}_e > 0$ for at least one link ending at v, the price at that node can be calculated by

$$\hat{p}_v^D = \min_{e:\, \omega(e)=v} c_e + p_{\alpha(e)}^S.$$

Because this price is always positive, the consumers will never buy more of the good than needed.

$$\hat{p}_v^D > 0 \implies \sum_{e:\, \omega(e)=v} \hat{\varphi}_e = y_v^D$$

$$\sum_{e:\, \omega(e)=v} \hat{\varphi}_e > y_v^D \implies \hat{p}_v^D = 0$$

For positive prices $\hat{\mathbf{p}}^D > \mathbf{0}$ the primal restrictions imply flow conservation via

$$\mathbf{B}_1 \hat{\boldsymbol{\varphi}} + \mathbf{B}_3 \hat{\boldsymbol{\varphi}} = \mathbf{B}\hat{\boldsymbol{\varphi}} = \mathbf{y}^S - \mathbf{y}^D \quad \text{where} \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_3 \end{pmatrix}.$$

Recall again the example of Table 3.1: the node/arc incidence matrix $\mathbf{B}$ includes the matrix $\mathbf{B}_3$ the elements of which are either 0 or -1.

The collection $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{y}}^S, \hat{\mathbf{p}}^D)$ constitutes a *spatial price equilibrium* in the sense that no customer can take advantage of switching to another supplier. If c_e is an affine function of distance and the price p_v^S is the same for all producers, then the solution is what one might expect immediately: each customer is served by his closest supplier. On some graphic interpretations of this result see Hall (1989).

A spatial price equilibrium suggests some further reaching problems which go beyond the scope of this book.

1. Optimal location of plants; see Koopmans, Beckmann (1957) on the location (of fractional parts) of n plants to n locations.
   One trick to solve the location problem is suggested by Figure 2.11, p. 64. Let $\mathcal{O}$ be the set of potential sources with the production capacity y_v^S and introduce a super-origin v_s. The fictitious link e from v_s to some production place $v \in \mathcal{O}$ has the capacity y_v^S and no transportation cost ($c_e = 0$) but fixed construction cost q_e. A binary choice variable ψ_e indicates whether the facility is located at v ($\psi_e = 1$) or not ($\psi_e = 0$). The objective is then to minimize the sum of transportation costs between sources and sinks (i.e., $\mathbf{c}^\top \boldsymbol{\varphi}$) and all construction costs including especially $\psi_e q_e$. Finally, an optimum binary solution $\hat{\boldsymbol{\psi}}$ determines where to locate the plants.
2. Optimal location of a newcomer
3. Spatial price discrimination etc.

(3) Transportation problem. Another variant of the transportation problem refers to an adequate assignment of trips. Given the number of trips generated at any node as well as the number of trips ending at any node, the problem is to find a trip assignment of least cost that brings the trip attraction $\boldsymbol{\varphi}^-$ into line with the trip production $\boldsymbol{\varphi}^+$. By intuition, the resulting traffic flow may be seen as the most likely trip table to arise in reality.

In order to keep the example as simple as possible, capacity restrictions are ignored. Using $\boldsymbol{\varphi} = \boldsymbol{\Delta}\mathbf{h}$ the above linear program can be expressed in route flows $\mathbf{h}$. According to (2.30) the route costs are $\tilde{\mathbf{c}} := \boldsymbol{\Delta}^\top \mathbf{c}$ so that the objective function becomes $\tilde{\mathbf{c}}^\top \mathbf{h}$. Kirchhoff's law requires that the given vectors $\boldsymbol{\varphi}^+$ and $\boldsymbol{\varphi}^-$ satisfy

$\mathbf{1}^{\mathsf{T}}\boldsymbol{\varphi}^{+} = \mathbf{1}^{\mathsf{T}}\boldsymbol{\varphi}^{-}$, where $\mathbf{b} = \boldsymbol{\varphi}^{+} - \boldsymbol{\varphi}^{-}$. The restriction $\mathbf{B}\boldsymbol{\varphi} = \mathbf{b}$ now takes the form $\mathbf{B}\boldsymbol{\Delta}\mathbf{h} = \boldsymbol{\varphi}^{+} - \boldsymbol{\varphi}^{-}$. For the sake of simplicity suppose that every OD-pair is connected by one and only one route, that is $\boldsymbol{\Delta} = \mathbf{I}$. In this case the arc/node incidence matrix $\mathbf{B}$ coincides with $\mathbf{O} - \mathbf{D}$, where $\mathbf{O}$ and $\mathbf{D}$ are the origin/path incidence matrix and the destination path incidence matrix, respectively. In accordance with (2.31) the new transportation problem and its dual can be stated as:

(3.13)

$$\begin{array}{lll} \text{(P)} \ \min \ \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h} & \qquad \text{(D)} \ \max \ \boldsymbol{\varphi}^{+\mathsf{T}}\mathbf{u} + \boldsymbol{\varphi}^{-\mathsf{T}}\mathbf{w} & \\ \text{s.t.} \ \mathbf{O}\,\mathbf{h} = \boldsymbol{\varphi}^{+} & \qquad \text{s.t.} \ \mathbf{O}^{\mathsf{T}}\mathbf{u} + \mathbf{D}^{\mathsf{T}}\mathbf{w} \leqq \tilde{\mathbf{c}} & \\ \mathbf{D}\,\mathbf{h} = \boldsymbol{\varphi}^{-} & \qquad \mathbf{u} & \text{unrestr.} \\ \mathbf{h} \geqq \mathbf{0} & & \mathbf{w} \text{ unrestr.} \end{array}$$

Regarding a primal solution $\hat{\mathbf{h}}$ and a dual solution $(\hat{\mathbf{u}}, \hat{\mathbf{w}})$, one complementary slackness condition is $(\hat{\mathrm{u}}_i + \hat{\mathrm{w}}_j - \tilde{c}_\rho)\hat{h}_\rho = 0$. The dual variables u_i and w_j are assigned to the outflow φ_i^{+} and the inflow φ_j^{-}, respectively, and ρ indicates the path connecting the origin v_i and the destination v_j. A positive flow from v_i to v_j requires that the potential value $\hat{\mathrm{u}}_i$ of the origin exceeds the potential value $\hat{\mathrm{w}}_j$ of the destination by $\tilde{c}_\rho$. That corresponds to the statement of Kirchhoff's Voltage law (Proposition 3.4). At the same time one restriction is again redundant because of Kirchhoff's Current Law.

(4) Transshipment problem. Hillier, Lieberman (1995, p. 187) call (3.13.P), which has no intermediate transfer points, the transport problem. The additional possibility of transshipment, where each transshipment node can also be a sink or a source, requires different methods of assignment because this case introduces the choice between alternative routes having different costs per unit. This extension of the *transport problem* to include the routing decisions is referred to as the *transshipment problem.*

3.1.3 Multicommodity Flows

3.1.3.1 Cost Minimization

In the flow problem (3.11) the origin and the destination of a particular flow unit have no importance. Take for example Figure 2.11, p. 64. A consumer in the place v_4 is indifferent between units coming from v_1 or v_2. Regarding a homogeneous good, the introduction of v_s and v_t is therefore plausible. This section introduces differentiated flow units, the origins and destinations of which are given. As an example we can think of sources each of which produces a distinct good. In this setting a case where each firm provides its commodity only to its immediate environs cannot be a meaningful solution. Another example for such a *multicommodity network* is that of trips from places of residence to working-places.

Note for a strongly connected network that the output of an arbitrary source can be sent to an arbitrary sink and conversely. Regarding multicommodity flows we make the restriction that every flow unit from a certain origin must be sent to a certain destination. This generates a new class of problems. For explanatory purposes let us

start with the cost minimal production of a given multicommodity flow. Afterwards we hand in the necessary extensions for the underlying production technology. The reason is that the former production function with one output must be substituted by the more general concept of an output correspondence with many commodities. In this case the notion of a maximum output is far more difficult to define.

The next step is to generalize the analysis with respect to more than one output of the network. A first example for such a multicommodity flow problem can be constructed as follows. Suppose the network of Figure 3.1, p. 78, has more than one source having no inflows and more than one sink without any outflows. The minimum flow from a source v_s to a sink v_t may then be encoded by a corresponding artificial link $e_k = (v_t, v_s)$ with the restriction $\varphi_k \geqq y_k$. The flows on these feedback

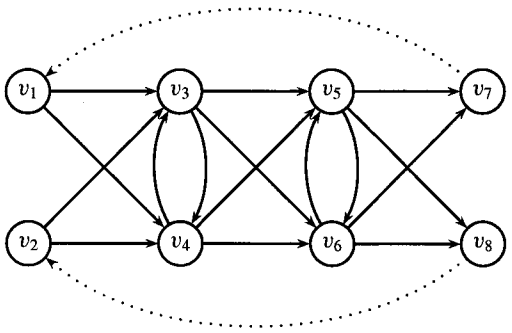


Fig. 3.6
Cost minimization
The example represents a multicommodity flow problem with two sources and two destinations. The objective is to realize at least given flows on the artificial links at minimum cost. The two artificial links are indicated by the dotted curves.

links are now interpreted as the outputs $\mathbf{y}$ of the network, where $y_e = 0$ is valid on all real links of the network; cf. Figure 3.6. The minimum cost problem (3.7) and its dual may now be stated as[5]

$$
\begin{array}{llll}
(3.14) & \text{(P)} \ \min \ \mathbf{c}^{\mathrm{T}}\boldsymbol{\varphi} & \text{(D)} \ \max & -\boldsymbol{\kappa}^{\mathrm{T}}\mathbf{q} + \mathbf{y}^{\mathrm{T}}\mathbf{p} \\
& \text{s.t.} \ \mathbf{A}\boldsymbol{\varphi} = \mathbf{0} & \text{s.t.} \ \mathbf{A}^{\mathrm{T}}\mathbf{w} - \mathbf{q} + \mathbf{p} \leqq \mathbf{c} & \\
& \boldsymbol{\varphi} \leqq \boldsymbol{\kappa} & \mathbf{w} & \text{unrest.} \\
& \boldsymbol{\varphi} \geqq \mathbf{y} & \mathbf{q} & \geqq \mathbf{0} \\
& \boldsymbol{\varphi} \geqq \mathbf{0} & & \mathbf{p} \geqq \mathbf{0}
\end{array}
$$

Bryan, O'Kelly (1999, p. 287) use also the restriction $\boldsymbol{\varphi} \geqq \underline{\boldsymbol{\kappa}}$ to prevent links from being underutilized. Heuristically, this is helpful when the installation of some connection incurs fixed cost or cost of maintenance requiring a minimum utilization of that link.[6] Section 5.2.2 comes back to this problem in a slightly modified setting.

[5] An alternative formulation of the feasible set would be $\mathbf{B}\boldsymbol{\varphi}^j = \mathbf{b}^j$ and $\sum_j \boldsymbol{\varphi}^j = \boldsymbol{\varphi}$, where $\boldsymbol{\varphi}^j \geqq \mathbf{0}$ denotes the link flows of commodity j and $\boldsymbol{\varphi}$ is the aggregate link flow; see Bergendorff et al. (1997).

[6] The idea is similar to the workings of a brain: rarely used neural connections are broken up.

The two programs of (3.14) are closely related to road traffic networks with the exception that traffic networks rest upon the assumption that all nodes can be sources, transit points, and sinks with respect to different commodities at the same same time. Problems of this type are in the center of interest of Section 3.2. Hence, let us abstain from discussing features of solutions to (3.14). Instead, the attention is drawn to the production technology of a multicommodity network which constitutes the basis of the cost minimization problem (3.14.P). Recall that the production function (3.6) represents the production technology for the case of a unique network output.

3.1.3.2 Production Technology

Regarding the artificial links, the example of Figure 3.6 suggests technologies with more than one output. The production *function* f is, therefore, substituted by a production *correspondence* P. This output correspondence describes what *output vectors* $\mathbf{y}$ can be produced given the input vector $\boldsymbol{\kappa}$. Due to the special construction of the vector $\mathbf{y}$, only its components referring to artificial links may be interpreted as final outputs. All other components are set to zero. In accordance with the preceding example, the *production possibility set* is given by

$$P(\boldsymbol{\kappa}) = \{\mathbf{y} |\ \mathbf{y} \text{ is producible by } \boldsymbol{\kappa}\}\,, \tag{3.15}$$

where $P(\boldsymbol{\kappa})$ is a subset of some abstract output space $\mathcal{Y}$. The statement "$\mathbf{y}$ is producible by $\boldsymbol{\kappa}$" means that a non-negative link flow $\boldsymbol{\varphi}$ exists which satisfies, on the one hand, $\boldsymbol{\varphi} \geqq \mathbf{y}$. On the other hand, this flow must not exceed the capacities, $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$, and the flow conservation must be fulfilled, $\mathbf{A}\,\boldsymbol{\varphi} = \mathbf{0}$. Notice that the flow vector $\boldsymbol{\varphi}$ is ambiguous. Some of its components refer to fictitious links and are interpreted as *outputs* of the network. The remaining components correspond to real links and indicate *intermediate goods*. An activity $(\mathbf{y}, \boldsymbol{\kappa})$ is called output efficient if no $\bar{\mathbf{y}}$ with $\bar{\mathbf{y}} \geq \mathbf{y}$ can be produced at the given capacities $\boldsymbol{\kappa}$. On networks with a single output all efficient activities must realize the maximum flow. For multicommodity networks, output efficient activities $(\mathbf{y}, \boldsymbol{\kappa})$ are described by the (positive) frontier of the production possibility set $P(\boldsymbol{\kappa})$.

Reasonable production correspondences possess some regularity properties[7], which are discussed in detail in Bobzin (1998). For the current purpose it is enough to identify the output space $\mathcal{Y}$ by the non-negative real space of proper dimension.

[P1a] Possibility of Inaction. Each capacity vector can be used to produce a "zero output".

$$\mathbf{0} \in P(\boldsymbol{\kappa}) \qquad \forall\, \boldsymbol{\kappa} \geqq \mathbf{0}$$

[P1b] No Land of Cockaigne. No positive output can be produced at zero capacities.

$$\mathbf{y} \notin P(\mathbf{0}) \qquad \forall\, \mathbf{y} \geq \mathbf{0}$$

[7] Similar properties with respect to a highway production function can be found in Färe et al. (1982). Be aware that the authors make use of a so-called generalized congestion function that has been introduced by Inman (1978). They treat link flows and capacities (i.e., highway scale) as inputs, while the level of service (i.e., speed) is interpreted as output.

[P2] **Attainability of Each Production.** For each output vector $\mathbf{y} \geqq \mathbf{0}$ a capacity vector $\boldsymbol{\kappa} \geqq \mathbf{0}$ exists such that the activity $(\boldsymbol{\kappa}, \mathbf{y})$ is a possible (but not necessarily obtainable) process.

$$\forall\, \mathbf{y} \geqq \mathbf{0},\ \exists\, \boldsymbol{\kappa} \geqq \mathbf{0}\colon \quad \mathbf{y} \in P(\boldsymbol{\kappa})$$

[P3] **Disposability of Inputs.** If the output vector $\mathbf{y}$ can be served by a capacity vector $\boldsymbol{\kappa}$, then increased capacities $\tilde{\boldsymbol{\kappa}}$ are also sufficient for the production of $\mathbf{y}$, i.e., $P(\boldsymbol{\kappa}) \subset P(\tilde{\boldsymbol{\kappa}})$.

$$\forall\, \boldsymbol{\kappa}, \tilde{\boldsymbol{\kappa}} \geqq \mathbf{0},\ \ \tilde{\boldsymbol{\kappa}} \geqq \boldsymbol{\kappa}\colon \quad \mathbf{y} \in P(\boldsymbol{\kappa}) \implies \mathbf{y} \in P(\tilde{\boldsymbol{\kappa}})$$

[P4] **Disposability of Outputs.** If an output vector $\mathbf{y}$ can be produced by the capacities $\boldsymbol{\kappa}$, then each vector $\tilde{\mathbf{y}} \geqq \mathbf{0}$, which is not larger than $\mathbf{y}$ for any component, can be produced by making use of $\boldsymbol{\kappa}$.

$$\forall\, \mathbf{y}, \tilde{\mathbf{y}} \geqq \mathbf{0},\ \ \mathbf{y} \geqq \tilde{\mathbf{y}}\colon \quad \mathbf{y} \in P(\boldsymbol{\kappa}) \implies \tilde{\mathbf{y}} \in P(\boldsymbol{\kappa})$$

[P5] **Boundedness** or impossibility of arbitrary abundance. The outputs $\mathbf{y}$ cannot be increased arbitrarily at constant inputs $\boldsymbol{\kappa} \geqq \mathbf{0}$.

$$\forall\, \boldsymbol{\kappa} \geqq \mathbf{0}\colon \quad P(\boldsymbol{\kappa}) \text{ is bounded.}$$

Similar to the production function (3.6) the output correspondence P is given by

$$\mathbf{y} \in P(\boldsymbol{\kappa}) \iff \exists\, \boldsymbol{\varphi} \geqq \mathbf{0}\colon \mathbf{A}\boldsymbol{\varphi} = \mathbf{0},\ \mathbf{y} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$$

and it is not too hard to prove that P satisfies [P1a]–[P5]. Although all of the above properties are plausible there is one caveat concerning congestion which constitutes one of the most important problems in network economics. Following Färe, Grosskopf (1983), congestion occurs when it is possible to reduce inputs and increase outputs. This effect would contradict [P3]. It is implausible with respect to link capacities, but reasonable for the intermediate inputs $\boldsymbol{\varphi}$. When too many flow units on a link disturb each other, the throughput may be reduced. As the current setting is not able to cover congestion phenomena we shall come back to this point at a later stage.

The next assumption reflects a useful mathematical property. As "favorable" activities frequently lie on the border of production possibility sets, we suppose them to be feasible. Although closed level sets are difficult to justify in economic terms, there is hardly any economic reason objecting it.

[P6] **Closedness.** When a sequence of possible activities $(\boldsymbol{\kappa}^\nu, \mathbf{y}^\nu)$ converges to the limit $(\boldsymbol{\kappa}^0, \mathbf{y}^0)$, then this limit is possible, too. When applied to the output correspondence, this means

$$P\colon \mathbb{R}^m_+ \to \mathfrak{P}(\mathbb{R}^m_+) \text{ is a closed mapping.}$$

Another desirable property refers to convex level sets. This simplification ignores that flows consist of indivisible flow units. As long as the size of a flow unit is small compared to the total flow, the mathematical advantage exceeds the actual misspecification by far.

[P7] **Convexity.** For perfectly divisible flows the production possibility set $P(\kappa)$ is convex for all $\kappa \geqq \mathbf{0}$.

$$\forall \tilde{\mathbf{y}}, \mathbf{y} \in P(\kappa),\ \forall \lambda \in [0,1]: \quad \lambda\tilde{\mathbf{y}} + (1-\lambda)\mathbf{y} \in P(\kappa)$$

Again one can show that the output correspondence satisfies [P6] and [P7]. Axiom [P7] has a mathematical analogue with respect to perfectly divisible inputs κ. We state [P8] just for the sake of completeness, being aware that one of the major problems in network design are lumpy or indivisible investments.

[P8] **Quasi-concavity.** The output correspondence P is quasi-concave for perfectly divisible capacities.

$$\forall \tilde{\kappa}, \kappa \geqq \mathbf{0},\ \forall \lambda \in [0,1]: \quad P(\tilde{\kappa}) \cap P(\kappa) \subset P\big(\lambda\tilde{\kappa} + (1-\lambda)\kappa\big)$$

Based upon the properties of the production technology, the next step is to define an adequate cost function as has been done in (3.10). The primal program (3.14.P) then takes the more general form

$$C(\mathbf{c}, \kappa, \mathbf{y}) = \inf\left\{\mathbf{c}^{\mathrm{T}}\varphi \mid \mathbf{y} \in P(\kappa),\ \varphi \geqq \mathbf{0}\right\}. \tag{3.16}$$

Recall that any transport along an artificial arc can be done at zero cost by Assumption 3.1, p. 78. The disadvantage with this notation of a cost function results from the complex structure of the network at hand. First, link capacities are assumed to be fixed and fixed cost of infrastructure provision are ignored. Second, the only cost of interest results from transport of intermediate goods φ which are implicitly defined by $\mathbf{y} \in P(\kappa)$. Nevertheless, (3.14.P) is a relatively easy problem of cost minimization. The next section overcomes the notational problem by an alternative representation of flow conservation.

Without going into the details of cost minimal flows at this point, one may ask furthermore for the *capacity (or power) of a multicommodity network*. The former maximum flow denotes the capacity of a network with respect to a single output. For multicommodity networks one may apply the construct of the *(output) distance function*[8] defined by

$$t_O(\mathbf{y}, \kappa) := \inf\left\{\lambda \geqq 0 \mid \mathbf{y} \in \lambda P(\kappa)\right\}. \tag{3.17}$$

Given any feasible output $\mathbf{y} \in P(\kappa)$ such that $t_O(\mathbf{y}, \kappa) \leqq 1$, one can stretch $\mathbf{y}$ along a ray through the origin, so that $\mathbf{y}/t_O(\mathbf{x}, \kappa)$ determines the maximum feasible

[8] On the properties of gauges or distance functions implied by [P1a]–[P8] see Bobzin (1998).

commodity bundle in the direction of that ray. Of course, any ray[9] may lead to a different network capacity, where the collection of all outputs $\mathbf{y}$ with $t_O(\mathbf{y}, \kappa) = 1$ is referred to as production possibility frontier. Weighing all outputs $\mathbf{y}$ by prices $\mathbf{p}$ would then lead to a revenue maximization problem (see Section 3.2.3.4) of the form

$$R(\mathbf{p}, \kappa) = \max \left\{ \mathbf{p}^{\mathrm{T}} \mathbf{y} |\ t_O(\mathbf{y}, \kappa) = 1 \right\}.$$

Having equilibrium prices in mind, Yang et al. (2000) propose a similar approach with special regard to equilibrium flows on urban transportation networks. Network equilibria will be in the center of interest in the next section.

The examples in Figure 3.1 and Figure 3.6 show three types of nodes – namely, sources, transshipment nodes, and sinks. In accordance with the flow conservation the transport from a source to a sink is represented by a corresponding flow on an artificial edge leading back from the respective sink to the respective source. Regarding traffic networks this approach is more of a hindrance than a help; nevertheless Hall (1978) uses that construction for the characterization of network equilibria. The next section shows how to avoid artificial links without neglecting the flow conservation. Artificial flows are substituted by a trip table. Section 3.2 generalizes the approach also by the aspect that each node may be a source and a sink at the same time. Take for example a traffic network. Some trips start at a particular node while other trips end there. Moreover, there may be further trips leading through that node.

To sum up, Section 3.1 has introduced the necessary concepts that determine the production technology of multicommodity networks including the special case of one homogeneous good. The preliminary examples have shown how to develop corresponding cost functions such as (3.16). The subsequent sections make use of the same reasoning, however, in a modified form of representation which is more adequate for transport networks.

3.2 Cost Minimal Trip Assignment

3.2.1 Production Technology

3.2.1.1 Feasible Network Activities

Having introduced some basic concepts of transport networks, we now analyze capacitated multicommodity networks where all nodes can be origins, transit nodes, or destinations at the same time. The aim is to characterize equilibrium flows when taking congestion effects into account. The analysis is based upon a given network topology with fixed link capacities. Network design will be in the center of interest of the next chapter.

[9] Jarvis, Martinez (1977) just sum up all components of $\mathbf{y}$ and maximize that term with regard to a capacitated network. That approach is only valid if all commodities can be substituted on a one-to-one basis. It is then consequent to define the system capacity by the maximum number of flow units (or customers) allowed in the network.

In what follows we assume a *strongly connected network*, that is each ordered pair of nodes is joined by a route. Take, for instance, the example of Figure 3.6. If the dotted links do exist, the example represents a strongly connected network. In this case it is not useful to describe the flows from v_1 to v_7 or from v_8 to v_2 by y_ℓ and $y_{\ell+1}$, respectively. For that reason we switch to a new notation. The number of units traveling from an origin v_i to a destination v_j is denoted $x_{(i,j)}$. The trip table $\mathbf{x}$ is in fact a vector that includes the traffic for all OD-pairs. Notice that every driver may have the opportunity to choose different routes connecting the same OD-pair. All drivers using route ρ constitute the path flow h_ρ, where $\mathbf{h}$ is the flow vector of all routes.

The advanced production technology of a network includes the following items.

- The *generated traffic* $\mathbf{x}$ (formerly given by $\mathbf{y}$) corresponds to the multicommodity output which is supplied by the network. Bear in mind that every flow unit has a determined origin and a determined destination so that different flows cannot be treated as one homogeneous good. Contrasted with that, in the max flow problem (3.1.P) the only element of the trip table was $\bar{\varphi}_\ell = x_{(1,n)}$.
- The route flows $\mathbf{h}$ are connected to the output of the network by the OD/path incidence matrix $\mathbf{\Lambda}$, that is $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}$. At the same time each route flow vector determines a link flow vector by (2.28), i.e., $\mathbf{\Delta}\,\mathbf{h} = \boldsymbol{\varphi}$. The incidence matrices $\mathbf{\Delta}$ and $\mathbf{\Lambda}$ describe the network topology. In other words, the *topology of the network* is given and encoded by the two incidence matrices. Aspects of (discrete) network design are ignored for the moment.
  The extension of the *transport problem* to include routing decisions is referred to as the *transshipment problem*. According to the multicommodity approach all nodes may be origins, destinations, and intermediate transfer points at the same time.
- The components of the capacity vector $\boldsymbol{\kappa}$ are treated like *fixed inputs* to the network. In fact $\boldsymbol{\kappa}$ reflects a feature of the fixed infrastructure such as highways, streets, railways, airports, canals, and so on. Capacities are transformed by the network on a one-to-one basis into link flows $\boldsymbol{\varphi}$ which are interpreted as intermediate goods needed to produce the final output $\mathbf{x}$. Thus, link flows are restricted to the respective capacity, $\mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$. Discrete or continuous capacity enlargements are an investment problem to be solved by the network provider.
  The infrastructure capital is no subject to wear and tear, by assumption. Consequently, we omit to consider maintenance work.
- For explanatory purposes it is useful to distinguish vehicles from the above fixed infrastructure. Clearly vehicles are fixed inputs in view of service providers such as taxi drivers or households owing private cars. Against that the number of vehicles is much easier to vary than the topology of a network. As an example, think of a railroad company which has to decide on the number of trains using its railroad network.
- There are *variable inputs* (fuel, energy, time, etc.) needed to make a trip along an arc or path. Thus, the amount of inputs depends, on the one hand, on the number of trips made. On the other hand, the inputs needed per trip may be constant

or they depend on the total number of trips. For instance, one needs more time per trip when there is a traffic jam caused by peak demand. These congestion effects reduce the level of service, which may be measured by speed, risk, or other indicators. One way of capturing congestion is to define so-called link performance functions. For example, the function $\tau_e(\varphi, \kappa)$ measures the travel time on link e depending on all traffic volumes and the respective capacities. In most cases we shall use its monetary equivalent, namely, the unit cost function $c_e(\varphi, \kappa)$.

The last item concentrates on the production of one particular trip, where each trip follows a route ρ being a sequence of arcs. The consumed resources reflect the behavior of the driver (e.g., choice of speed) as well as the properties of these links, that is the number of links used, the length of each link, the utilization of each link and so on. The network technology abstracts from these detailed aspects of trip production. Instead it draws the attention to aggregate flows and how to accommodate traffic demand by the fixed infrastructure possessing certain features (e.g., capacities).

The first three items describe the production of a trip matrix $\mathbf{x}$ being composed of route flows $\mathbf{h}$ or link flows φ. Given the network topology and the capacities of the links the advanced *production possibility set* becomes

$$P(\kappa) = \left\{ \mathbf{x} \mid \mathbf{x} \geqq \mathbf{0},\ \Lambda\,\mathbf{h} = \mathbf{x},\ \Delta\,\mathbf{h} = \varphi,\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{0} \leqq \varphi \leqq \kappa \right\}.$$

Note that the output correspondence P again satisfies the properties [P1a]–[P6] and particularly [P7] (convexity[10]). If we allow for transportation from any node to any other node, then [P2] requires a strongly connected network.[11] With that one can always assign $x_{(i,j)}$ to some existing route and the resulting link flows φ are feasible for sufficiently large κ. The only difference between the two output correspondences is that the former balance equation $\mathbf{A}\,\varphi = \mathbf{0}$ with φ including artificial links is now substituted by $\Delta\,\mathbf{h} = \varphi$ and $\Lambda\,\mathbf{h} = \mathbf{x}$. Because $x_{(i,j)}$ describes flow units leaving the node v_i and entering the node v_j at the same time, the condition on flow conservation implicitly holds true: no flow unit vanishes accidentally at any link or node.

The problem with the above production possibility set is that it depends on a capacity vector κ representing the fixed topology of the network. Contrasted with that, we are interested here in transport cost so that the analysis has to focus on the intermediate goods φ. In doing so let us rearrange the notation somewhat. Regarding the production possibility set, every traffic vector $\mathbf{x} \in P(\kappa)$ can be represented by an appropriately chosen, but not necessarily unique link flow vector φ. This vector has two essential features. The first property refers to the network topology which is determined by the set of nodes and links. Hence, any flow φ which is feasible with respect to $\mathbf{x}$ must be an element of the so-called *input requirement set*

$$L(\mathbf{x}) = \left\{ \varphi \mid \varphi = \Delta\,\mathbf{h},\ \Lambda\,\mathbf{h} = \mathbf{x},\ \varphi \geqq \mathbf{0},\ \mathbf{h} \geqq \mathbf{0} \right\}.$$

[10] As long as $P(\kappa)$ is detemined by linear equations and linear inequalities, convexity is not difficult to prove but awkwardly to prove.

[11] Otherwise we have to define an output space $\mathcal{X}$ such that $P(\kappa) = \emptyset$ whenever $\mathbf{x} \notin \mathcal{X}$.

As long as κ can be adjusted so that $\varphi = \kappa$, the output correspondence P and the input correspondence L are inverse to each other.

$$\mathbf{x} \in P(\varphi) \iff \varphi \in L(\mathbf{x}) \tag{3.18}$$

This is the point where the demand for traffic comes in. Let the given trip table $\mathbf{x}^D$ be equal to the generated traffic $\mathbf{x}$. Then the link flow vector φ is said to be *demand feasible* if $\varphi \in L(\mathbf{x}^D)$.

The second property of a feasible flow vector takes into account that the link flows are restricted to fixed capacities. To stress this aspect, an element of the set

$$\Phi^S(\kappa) := \left\{ \varphi \mid \mathbf{0} \leqq \varphi \leqq \kappa \right\}$$

is referred to as *supply feasible*.

To simplify notation in unambiguous cases we use Φ^S instead of $\Phi^S(\kappa)$ and Φ^D instead of $L(\mathbf{x}^D)$. Now any producible trip matrix can be expressed by link flows and (3.18) turns out to be

$$\mathbf{x}^D \in P(\kappa) \iff \varphi \in \Phi^D \cap \Phi^S. \tag{3.19}$$

The inclusion on the right hand side will be discussed at a later stage in more detail with respect to route flows. As pointed out in Beckmann et al. (1956, p. 59), it is important to note that demand refers to trips and capacity concerns link flows. The connecting link is found in the distribution of trips over the network. Every trip matrix $\mathbf{x} \in P(\kappa)$ is supplied by the network. Thus, any *fixed* trip table $\mathbf{x}^D$ which is contained in the production possibility set $P(\kappa)$ constitutes a "network equilibrium" in the sense that supply matches demand. Bear in mind that this equilibrium ignores aspects of quality of service. Although the network accommodates the demand for traffic, it may do so at a very low level of service. Furthermore, this notion of an equilibrium even permits the existence of another trip matrix $\bar{\mathbf{x}} \in P(\kappa)$ with $\bar{x}_{(i,j)} > x^D_{(i,j)}$ for all OD-pairs. The corresponding excess supply does not contradict a network equilibrium as defined above.

The corresponding *equilibrium flow* results from $\varphi \in \Phi^D \cap \Phi^S$, where the relationship between $\mathbf{x}^D$ and φ follows from $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x} = \mathbf{x}^D$ and $\varphi = \mathbf{\Delta}\,\mathbf{h}$. Together with the capacity restrictions the problem of finding some route flow $\mathbf{h}$ such that $\mathbf{x}^D \in P(\kappa)$ or, equivalently,

$$\begin{aligned} \mathbf{\Lambda}\,\mathbf{h} &= \mathbf{x}^D \\ \mathbf{\Delta}\,\mathbf{h} &\leqq \kappa \\ \mathbf{h} &\geqq \mathbf{0} \end{aligned} \tag{3.20}$$

is referred to as *multicommodity flow problem*. As has been mentioned above all flow variables φ and $\mathbf{h}$ are assumed to be perfectly divisible. On the maximum flow or the capacity of such a multicommodity network see the corresponding remarks on p. 97.

In the case of a fixed trip table it is relatively easy to find network equilibria. In a more general setting (see Section 4.2) the demand for trips $\mathbf{x}^D$ depends on

the traffic conditions which in turn are determined by the generated traffic $\mathbf{x}$. An existing OD-pattern $\mathbf{x}$ is an equilibrium if it gives rise to traffic conditions that will maintain the same demand, i.e., $\mathbf{x}^D = \mathbf{x}$. These traffic conditions will be measured by a corresponding price system.

3.2.1.2 Special Features of Data Networks

The multicommodity flow problem can also be applied to telecommunications systems or data networks. The consumption of network resources, however, depends on the method how traffic is handled by the hardware.

Packet-switched describes the type of network in which relatively small units of data, say packets, are routed through a network based upon the destination address contained within each packet (store-and-forward network). Breaking communication down into packets allows the same data path to be shared among many users in the network. This type of communication between sender and receiver is known as *connectionless* (rather than dedicated). Most traffic over the Internet uses packet switching and the Internet is basically a connectionless network.

The opposite of packet-switched is *circuit-switched* characterizing a type of network such as the regular voice telephone network in which the communication circuit (path) for the call is set up and dedicated to the participants in that call. For the duration of the connection, all resources on that circuit are unavailable for other users and new sessions must be rejected. The same concept is valid for networks where each communication line allocates a given portion of its total transmission capacity for the session. Again these resources are blocked for other users.

Circuit switching is rarely used for data networks. The reason is not the complexity of packet switching but the inefficient use of links. Typical data sessions simply consist of very short bursts of high activity followed by long periods of inactivity. Although store-and-forward networks are hard to control, they have the advantage to use the resources more efficiently.

3.2.1.3 Aspects of Trip Assignment

The following sections are in general concerned with the costs of producing a trip matrix $\mathbf{x}$. To ensure that $\mathbf{x}$ is producible the first approach assumes sufficiently large link capacities so that $\boldsymbol{\varphi} < \boldsymbol{\kappa}$ holds always true, see [P2]. In this case the capacities $\boldsymbol{\kappa}$ can be ignored and the restriction turns out to be $\boldsymbol{\varphi} \in L(\mathbf{x})$. The second approach in Section 3.2.3 is concerned with the case $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$. It is implicitly assumed that a feasible flow exists, i.e., $L(\mathbf{x}) \cap \Phi^S(\boldsymbol{\kappa}) \neq \emptyset$. If the intersection is empty, the capacities must be increased such that $\boldsymbol{\kappa} \in L(\mathbf{x})$. The corresponding investment problems of network provision are discussed at a later stage, see Section 5.1.

Let us start with a service provider, or carrier as synonym, who wants to produce a trip matrix $\mathbf{x}$ at minimum cost. The cost function is then determined by

$$C(\mathbf{x}) = \min \left\{ \mathbf{c}^{\mathsf{T}} \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in L(\mathbf{x}) \right\}, \tag{3.21}$$

where c_e indicates the cost of traveling along link e. The idea is to assign the trips $x_{(i,j)}$ to routes, and thus to links, such that the total travel costs attain a minimum. The first procedure is usually called *routing*, while the second step is sometimes referred to as *loading*. At first the cost per trip along a road e is supposed to be a constant c_e and the total travel cost for that road is $c_e\varphi_e$. An advanced approach allows for congestion effects. Now the cost per car c_e itself is an increasing function of φ_e so that the total cost per link becomes $c_e(\varphi_e)\,\varphi_e$. In the most general case, c_e may be a function of all link flows which takes into account that edge flows are disturbed by other link flows, for instance, at junctions. The total cost is then $c_e(\boldsymbol{\varphi})\,\varphi_e$; see Section 3.2.5. If a profit maximizing monopolist knows the demand function $\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p})$ or its inverse $\boldsymbol{\xi}^{-1}(\mathbf{x}^D)$, he seeks to maximize $\boldsymbol{\xi}^{-1}(\mathbf{x}^D)^\top\mathbf{x}^D - C(\mathbf{x}^D)$; see Section 4.2. Notice that a solution to this problem requires to make all trips simultaneously or at least within a restricted period of time. It is not allowed to perform trips one after the other. Moreover, a profit maximal solution is a network equilibrium. The trips supplied by the service provider coincide with the demand for trips. In the case of a fixed trip table $\mathbf{x}^D$ it suffices to solve (3.21).

A similar approach applies to traffic networks. The trip table $\mathbf{x}^D$ incorporates the demand for traffic. At first this demand is assumed to be fixed. In a more general approach $\mathbf{x}^D$ is a function of prices per trip. The crucial point is now who decides on the route choice. A fictitious central decision-maker minimizes the total travel cost of the system as given in (3.21). This solution corresponds to a *normative traffic assignment* and is referred to as *system optimality*. Another aspect is that of *user optimality*, where every driver chooses the cheapest route by himself. The corresponding *descriptive traffic assignment* is characterized by *Wardrop's conditions*. They ensure for any traffic equilibrium that no driver has an incentive to change his route choice. The notion of user and system optimality was coined by Dafermos (1968). Bear in mind that both solutions determine traffic equilibria because the demand for traffic is always identical to the generated traffic. Nevertheless, when the users' behavior violates system optimality one has to think of congestion or road pricing to have a favorable influence on the individual route choice.

At present one prominent example of road pricing can be watched in Great Britain. Since February 17, 2002 the City of London has imposed a daily £5 "congestion charge" on most motorists entering an eight-square-mile downtown zone. The charge has reduced traffic by 20%. Following Pierre Lemieux[12], the economic argument is straight forward. *Motorists imposing congestion on others should pay for the space they use, just as they do when they pay parking charges. Those for whom the benefits of driving downtown is lower than the congestion charge will choose not to go, or to use other means of transportation than their private cars; the others will reimburse the cost they impose, instead of being subsidized by the general taxpayers. The level of congestion will decrease to its optimal level.* That opinion is not too optimistic. A number of cities around the world have already led

[12] Pierre Lemieux is co-director of the Economics and Liberty Research Group at the Université du Québec en Outaouais. Web-site: `www.pierrelemieux.org`.

the way. Trondheim, Oslo, Singapore — all three have excellent electronic road-pricing schemes. They have reduced traffic and raised the average speed of vehicles within the cities.

3.2.2 Fixed Link Costs per Unit

3.2.2.1 System Optimality

This section continues with the problem of minimizing transportation cost on uncapacitated networks. The task is to introduce the ideas of system optimality and user optimality. A first result will be that both concepts coincide in the current simple setting. This outcome becomes invalid at the latest when congestion takes place.

The first problem is to realize a given trip table $\mathbf{x}^D$ at minimum total traveling cost.

$$\min \left\{ \mathbf{c}^T \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in \Phi^D \right\} \tag{3.22}$$

As has been mentioned above this provisional problem is based on the following simplifications: we seek a demand feasible flow $\boldsymbol{\varphi} \in \Phi^D$. There are no capacity constraints $\boldsymbol{\kappa}$ so that supply feasibility has no importance. Neglecting capacity constraints has the technical advantage that traffic congestion or even queues do not occur by assumption. The trip table $\mathbf{x}^D$ is fixed and represents the demand for travel to be served. For this reason, $L(\mathbf{x})$ has been substituted in (3.21) by Φ^D so that a solution to (3.22) yields an equilibrium. The link costs per flow unit (or average costs) $\mathbf{c}$ are assumed to be constant, particularly they do not increase when traffic becomes heavier. The objective function of the cost minimization problem represents the total travel costs of the system. An optimal traffic assignment, therefore, corresponds to *system optimality*. Again the cost minimization problem (3.22) results in a linear program.[13]

$$\begin{aligned} \min\ & \mathbf{c}^T \boldsymbol{\varphi} \\ \text{s.t.}\ & \boldsymbol{\varphi} - \boldsymbol{\Delta}\,\mathbf{h} = \mathbf{0} \\ & \boldsymbol{\Lambda}\,\mathbf{h} - \mathbf{x} = \mathbf{0} \\ & \mathbf{x} = \mathbf{x}^D \\ & \boldsymbol{\varphi} \geqq \mathbf{0} \\ & \mathbf{h} \geqq \mathbf{0} \\ & \mathbf{x} \geqq \mathbf{0} \end{aligned} \tag{3.23}$$

This problem is given in terms of link flows $\boldsymbol{\varphi}$, route flows $\mathbf{h}$, and a trip matrix $\mathbf{x}$. The decisive variable, however, is $\mathbf{h}$. The equilibrium condition requires $\mathbf{x} = \mathbf{x}^D$. The link flows $\boldsymbol{\varphi}$ can be eliminated using the arc/path incidence matrix in the first restriction.

[13] The restrictions can equivalently be expressed as $\mathbf{x}^D \in P(\boldsymbol{\kappa})$ for sufficiently large capacities.

Thereby the restrictions reduce to the multicommodity flow problem (3.20). Defining path costs by summing up over all link costs, that is (2.30) $\tilde{\mathbf{c}} := \mathbf{\Delta}^{\mathsf{T}}\mathbf{c}$, results in

$$\mathbf{c}^{\mathsf{T}}\boldsymbol{\varphi} = \mathbf{c}^{\mathsf{T}}\mathbf{\Delta}\,\mathbf{h} = \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h}\,.$$

In doing so the problem (3.22) shows only route flows. For subsequent purposes it is worthwhile to distinguish explicitly feasible arc flows and feasible route flows,

$$(3.24)\quad \Phi^D = \left\{\boldsymbol{\varphi} \mid \boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h},\ \mathbf{h} \in H^D\right\} \quad \text{and} \quad H^D = \left\{\mathbf{h} \mid \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0}\right\}.$$

Bear in mind that every $\mathbf{h} \in H^D$ is assigned to precisely one $\boldsymbol{\varphi} \in \Phi^D$, but the inversion is not true. Given a $\boldsymbol{\varphi} \in \Phi^D$ there may be more than one $\mathbf{h} \in H^D$ so that $\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h}$. The reason is simply that h_ρ reveals more information on a flow unit of that route than φ_e, which refers to flow units on merely one particular link.

3.7 Remark Eliminating the link flow vector $\boldsymbol{\varphi}$ is more convenient regarding the subsequent problem of user optimality, where every traveler has to choose one admissible route. The difficulty with this formulation is that in general the number of paths exceeds the number of links by far. The elimination of the route flow variable $\mathbf{h}$ can be done at the cost of a less comfortable problem formulation, see page 66. □

The linear program (3.23) now can be rewritten as (3.25.P) with (3.25.D) being its dual program. Note that the primal *traffic assignment problem* requests system optimal route flows.[14]

$$(3.25)\qquad \begin{array}{lll} \text{(P)}\ \min\ \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h} & & \text{(D)}\ \max\ \mathbf{x}^{D\mathsf{T}}\mathbf{p} \\ \text{s.t.}\ \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D & & \text{s.t.}\ \mathbf{\Lambda}^{\mathsf{T}}\mathbf{p} \leqq \tilde{\mathbf{c}} \\ \mathbf{h} \geqq \mathbf{0} & & \mathbf{p}\ \text{unrestr.} \end{array}$$

Note that the ρ^{th} column of $\mathbf{\Lambda}$, say $\mathbf{\Lambda}_\rho$, has one and only one component equal to 1, that is

$$\mathbf{\Lambda}_{(s,t),\rho} = 1 \iff \left[\alpha(\rho) = v_s \text{ and } \omega(\rho) = v_t\right].$$

Hence, ρ denotes a route connecting the origin v_s with the destination v_t. The dual restrictions thus state

$$(3.26)\qquad \mathbf{\Lambda}_\rho{}^{\mathsf{T}}\hat{\mathbf{p}} = \hat{\mathrm{p}}_{(s,t)} \leqq \tilde{c}_\rho \quad \text{whenever } \rho \in \mathcal{R}(s,t).$$

The dimension of the dual variables

$$[\mathrm{p}_{(s,t)}] = \frac{€}{\text{trip from } v_s \text{ to } v_t}$$

suggests $\mathrm{p}_{(s,t)}$ to be the (shadow) price of a trip from v_s to v_t. Let $C_P(\mathbf{x}^D)$ and $C_D(\tilde{\mathbf{c}})$ be the optimal values of the primal program (3.25.P) and the dual program (3.25.D),

[14] The programs (3.25) of finding a cost minimal trip assignment should be compared with the problem of finding a cost minimal trip distribution (3.13), which is based upon different information about traffic, i.e., nodal trip production and trip attraction instead of a trip table. On the state of the art applying primal simplex algorithms to minimum cost network flows see Helgason, Kennington (1995).

respectively. Suppose furthermore that both optimal values are finite. Then the primal program has a unique[15] KKT vector $\hat{\mathbf{p}}$ if and only if C_P is differentiable at $\mathbf{x}^D$ and

$$\frac{\partial C_P(\mathbf{x}^D)}{\partial x^D_{(s,t)}} = \hat{\mathrm{p}}_{(s,t)} \qquad \forall\,(s,t). \tag{3.27}$$

Although $\mathbf{p}$ is unrestricted in sign the marginal costs indicate $\mathbf{p}$ to be non-negative. The dual program indeed will never attain a maximum if one $\mathrm{p}_{(s,t)}$ is negative provided $x^D_{(s,t)} > 0$ and $\tilde{c}_{(s,t)} \geqq 0$. Hence, it is plausible to interpret $\mathbf{p}$ as a vector of non-negative shadow prices.

The partial derivatives in (3.27) depend also on the parameters $\tilde{\mathbf{c}}$. Thus under the premises of (3.27), the shadow price vector may be seen as a function of the parameters $\tilde{\mathbf{c}}$ and $\mathbf{x}^D$.

$$\hat{\mathbf{p}} = \mathbf{p}(\tilde{\mathbf{c}}, \mathbf{x}^D) \equiv \nabla_{\mathbf{x}} C_P(\mathbf{x}^D; \tilde{\mathbf{c}})$$

The dual program has a unique KKT vector $\hat{\mathbf{h}}$ if and only if C_D is differentiable at $\tilde{\mathbf{c}}$ and

$$\frac{\partial C_D(\tilde{\mathbf{c}})}{\partial \tilde{c}_\rho} = \hat{h}_\rho \qquad \forall\,\rho. \tag{3.28}$$

The function C_D depends also on the parameters $\mathbf{x}^D$. Similar to $\hat{\mathbf{p}}$ the optimum solution $\hat{\mathbf{h}}$ is a function of the parameters $\tilde{\mathbf{c}}$ and $\mathbf{x}^D$.

$$\hat{\mathbf{h}} = \mathbf{h}(\tilde{\mathbf{c}}, \mathbf{x}^D) \equiv \nabla_{\tilde{\mathbf{c}}} C_D(\tilde{\mathbf{c}}; \mathbf{x}^D)$$

The complementary slackness conditions of the two linear programs at hand are

$$\begin{aligned} \boldsymbol{\Lambda}_\rho \hat{\mathbf{h}} - x^D_{(s,t)} &= 0 \qquad \forall\,(s,t)\,, \\ \hat{h}_\rho \left(\boldsymbol{\Lambda}_\rho{}^{\mathsf{T}} \hat{\mathbf{p}} - \tilde{c}_\rho\right) &= 0 \qquad \forall\,\rho\,. \end{aligned}$$

The first condition simply states that the generated traffic in terms of route flows matches the demand for traffic. Making use of the definition $\boldsymbol{\Lambda}_\rho{}^{\mathsf{T}} \hat{\mathbf{p}} = \hat{\mathrm{p}}_{(s,t)}$, cf. (3.26), the second complementary slackness condition can be rewritten as

$$\hat{h}_\rho \left(\hat{\mathrm{p}}_{(s,t)} - \tilde{c}_\rho\right) = 0 \qquad \forall\,\rho \in \mathcal{R}(s,t). \tag{3.29}$$

A positive route flow $\hat{h}_\rho$ requires $\hat{\mathrm{p}}_{(s,t)} = \tilde{c}_\rho$. Each route incurring higher costs ($\tilde{c}_\rho > \hat{\mathrm{p}}_{(s,t)}$) must be avoided ($\hat{h}_\rho = 0$). A route with lower costs ($\tilde{c}_\rho < \hat{\mathrm{p}}_{(s,t)}$) contradicts $\hat{\mathbf{p}}$ to be optimal for the dual program. This result is not surprising: cost minimization in this context simply asks for using always the cheapest route available.

3.8 Summary The optimum value $\hat{\mathrm{p}}_{(s,t)}$ of the dual variable $\mathrm{p}_{(s,t)}$ denotes the *minimum travel cost* per trip from v_s to v_t. This is frequently referred to as *shortest path distance*, where the distance is measured in monetary units. As an abbreviation, the *distance* between two nodes signifies the length of the shortest path between

[15] A unique shadow price vector is plausible in view of Summary 3.8 below.

them. For the following analysis it turns out to be useful to distinguish explicitly between shadow prices $\mathbf{p}$ and minimum travel cost $\mathbf{p}$ so that

$$\hat{\mathbf{p}} = \mathbf{p} \quad \text{with} \quad p_{(s,t)} := \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho. \tag{3.30}$$

Moreover, under appropriate assumptions $\hat{\mathbf{p}}$ is the marginal cost with respect to a parametric variation of $x^D_{(s,t)}$. □

3.9 Remark One of the best known algorithms to solve the *shortest path problem* in (3.30) has been presented in Dijkstra (1959).[16] Given a node v_s this algorithm calculates the shortest paths to any other node of the network. One important observation is the following: each shortest path from v_s to v_t has a last link, say (v_{t-1}, v_t). Dropping this link generates the shortest path from v_s to v_{t-1}. This property is reflected by *Bellman's equations*

$$\hat{\mathrm{p}}_{(s,s)} = 0 \quad \text{and} \quad \hat{\mathrm{p}}_{(s,i)} = \min \left\{ \hat{\mathrm{p}}_{(s,k)} + c_{ki} \middle| \, i \neq k \right\} \quad (i = 2, \ldots, t),$$

where $c_{ki} = \infty$ if there is no link from v_k to v_i. For an extended version of this approach see Evans (1976). From an economic point of view this principle implies that trips are not "evenly distributed" over the network, but they use some roads extensively.

In networks where all edges show the same link cost per unit the shortest paths are characterized by the minimum number of links used. In this case the shortest path problem is sometimes called *min-hop routing*. Another associated problem is to find the *most reliable path*. In this case the weight or reliability of a link γ_e measures the probability that a line is operational. If failures of distinct arcs are statistically independent, the reliability of a path ρ is given by the product (rather than the sum) of all link weights involved. Taking logarithms and substituting $c_e = -\ln \gamma_e$ results in a shortest path problem of the above type. □

Finally, Theorem 2.4 states that if $\hat{\mathbf{h}}$ is a primal solution and if $\hat{\mathbf{p}}$ is a dual solution, then

$$C_P(\mathbf{x}^D; \tilde{\mathbf{c}}) = \tilde{\mathbf{c}}^\mathsf{T} \hat{\mathbf{h}} = \mathbf{x}^{D\mathsf{T}} \hat{\mathbf{p}} = C_D(\tilde{\mathbf{c}}; \mathbf{x}^D).$$

Again the network output evaluated at shadow prices equals the minimum transportation cost. This observation is also true for any OD-pair.

$$\hat{\mathrm{p}}_{(s,t)} x^D_{(s,t)} = \sum_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho \hat{h}_\rho, \quad \text{where } x^D_{(s,t)} = \sum_{\rho \in \mathcal{R}(s,t)} \hat{h}_\rho \text{ and } \hat{\mathrm{p}}_{(s,t)} = \tilde{c}_\rho \text{ for } \hat{h}_\rho > 0.$$

Notice that at least one $\hat{h}_\rho$ must be positive when the demand for traffic $x^D_{(s,t)}$ is positive. As long as $\tilde{\mathbf{c}}$ is independent of the traffic assignment, so is $\hat{\mathbf{p}}$.

3.2.2.2 User Optimality

Up until now the cost minimal traffic assignment $\hat{\mathbf{h}}$ corresponds to system optimality. This normative assignment is optimal for the network as a whole, but is it the

[16] For a review of alternative algorithms solving the shortest path problem see, e.g., Bertsekas (1991).

best individual route choice? This point of view refers to user optimality. Wardrop (1952) published a paper in which he proposed two different optimality principles for assigning traffic to a network. Note that in this context the journey time is measured in terms of transportation cost.

Wardrop's first principle: The journey times on all the routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.

Wardrop's second principle: The average journey time is a minimum.

As can be found in Larsson, Patriksson (1999), the first performance criterion is based upon the assumption of rational traveler behavior with respect to each user of the congested traffic network seeking to minimize his/her own travel time, and it is therefore also known as the *user optimum*, or equilibrium *principle*. Wardrop's second performance criterion is the minimization of the average travel time (or, equivalently, the total travel time for a given number of trips[17]), and it is therefore referred to as the *system optimum principle*. A technical statement of the first principle follows from (3.31a) and (3.31b). Being in an equilibrium situation just means that no trip-maker has an incentive to change his behavior. Note that the individual choice of routes is by no means guaranteed to satisfy the second principle, and in most cases it does not.

3.10 Definition *A path flow* $\mathbf{h} \in H^D$ *is called a* Wardrop equilibrium *(or* user equilibrium*) when no driver has a less costly alternative route.*

$$\rho_1 \neq \rho_2,\ \rho_1 \sim \rho_2 \colon h_{\rho_1} > 0 \implies \tilde{c}_{\rho_1} \leqq \tilde{c}_{\rho_2}$$

As no unused path has lower cost the Wardrop equilibrium may be restated as follows: let $p_{(s,t)}$ *be the minimum travel cost on any route from* v_s *to* v_t *as defined in* (3.30). *For each* $(v_s, v_t) \in \mathcal{V} \times \mathcal{V}$ *and each* $\rho \in \mathcal{R}(s,t)$ *the equilibrium flow satisfies*

$$\hat{h}_\rho > 0 \implies \tilde{c}_\rho = p_{(s,t)}, \tag{3.31a}$$

$$\hat{h}_\rho = 0 \implies \tilde{c}_\rho \geqq p_{(s,t)}. \tag{3.31b}$$

The relation (3.31b) is not really an implication because the right hand side holds always true by the definition of $p_{(s,t)}$. The preceding characterization of a user equilibrium represents the view point of an observer monitoring the behavior of network users. If $\hat{h}_\rho > 0$, he knows that there is no alternative preferred route. In this sense users reveal their preferences by choosing the cheapest routes. Smith (1979a, p. 299) deals with an alternative formulation of the Wardrop equilibrium stressing the users' point of view. The negation of (3.31) yields

$$\tilde{c}_\rho > p_{(s,t)} \implies \hat{h}_\rho = 0, \tag{3.32a}$$

$$\tilde{c}_\rho = p_{(s,t)} \implies \hat{h}_\rho \geqq 0. \tag{3.32b}$$

[17] Holding the trip table $\mathbf{x}^D$ fixed determines the number of trips so that minimizing the average cost per trip is the same as minimizing total transportation cost.

If ρ is not the cheapest available route, the network users ignore it. Again (3.32b) provides no real additional information because route flows are always non-negative by definition. This is not an immediate consequence of a route being the cheapest alternative. The two implications (3.31a) and (3.32a) are reflected by (3.29) provided the vector of shadow prices $\mathbf{p}$ coincides with $\mathbf{p}$; see (3.30).

3.11 Example The following figure represents a simple example of two parallel routes connecting the same OD-pair (v_s, v_t) and serves to illustrate Wardrop's first principle. Both routes consist of a unique arc. If $\tilde{c}_1 < \tilde{c}_2$, we know $\tilde{c}_1 = p_{(s,t)}$ by

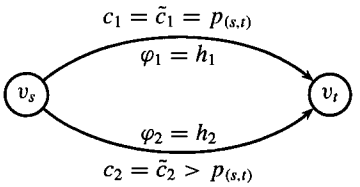


Fig. 3.7
Wardrop's first principle
In a user equilibrium the flow on the second route is always zero while the first (uncapacitated) route shows a non-negative flow.

(3.30). Moreover, no trip-maker will use the second route, $\hat{h}_2 = 0$. □

The following proposition states for the current deterministic[18] settings with constant transportation costs per trip and a fixed trip table that there is no difference between system and user optimality.

3.12 Proposition *A path flow* $\hat{\mathbf{h}}$ *is a cost minimizing solution to* (3.25.P) *if and only if it is a Wardrop equilibrium.*

Proof: The necessary part of the proof assumes $\hat{\mathbf{h}}$ to be a cost minimal path flow, but *no* Wardrop equilibrium. From the definition of a Wardrop equilibrium there must be at least one case of two alternative routes, say ρ_1 and ρ_2, connecting the same OD-pair (v_s, v_t) with $\hat{h}_{\rho_1} > 0$ and $\tilde{c}_{\rho_1} > \tilde{c}_{\rho_2}$, see (3.31a). At the same time (3.29) implies $\tilde{c}_{\rho_1} = \hat{p}_{(s,t)} > \tilde{c}_{\rho_2} \geqq p_{(s,t)}$. But this inequality contradicts the dual restrictions of (3.25.D) such that $\hat{\mathbf{h}}$ cannot be a cost minimal solution to (3.25.P).

Sufficiency of the proof: the two statements of (3.31) imply

$$(\tilde{c}_\rho - p_{(s,t)})(h_\rho - \hat{h}_\rho) \geqq 0 \quad \forall h_\rho \geqq 0.$$

This is obvious for $\hat{h}_\rho > 0$, because then $\tilde{c}_\rho - p_{(s,t)} = 0$. In the other case with $\hat{h}_\rho = 0$, one has $\tilde{c}_\rho \geqq p_{(s,t)}$ which is multiplied by some $h_\rho \geqq 0$. Summing up over all routes and noting that $\hat{\mathbf{h}} \in H^D$ is feasible for (3.25.P) gives in vector notation

$$(\tilde{\mathbf{c}} - \mathbf{\Lambda}^\mathsf{T}\mathbf{p})^\mathsf{T}(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \quad \forall \mathbf{h} \in H^D.$$

Notice that every feasible $\mathbf{h} \in H^D$ and particularly $\hat{\mathbf{h}}$ requires $\mathbf{\Lambda h} = \mathbf{x}^D$. Thus the above inequality becomes

$$\tilde{\mathbf{c}}^\mathsf{T}\mathbf{h} \geqq \tilde{\mathbf{c}}^\mathsf{T}\hat{\mathbf{h}} \quad \forall \mathbf{h} \in H^D.$$

[18] On some aspects of stochastic route choice see Section 4.1.1.

If we assume $\hat{\mathbf{h}}$ not to be cost minimal then an $\mathbf{h} \in H^D$ exists with $\tilde{\mathbf{c}}^T\mathbf{h} < \tilde{\mathbf{c}}^T\hat{\mathbf{h}}$. This contradicts the preceding inequality. ■

3.13 Remark (Game Theoretic Aspects) A user equilibrium state can be interpreted as a Nash equilibrium in a non-cooperative game among flows between different OD-pairs; cf. Dafermos, Sparrow (1969), Friesz (1985) or Larsson, Patriksson (1999). In this game, each group of travelers between an OD-pair observes the traffic conditions that result from the decisions made by the other network participants, and then distributes its travel demand among the OD-routes in such a way that the routes utilized in the pair are among the least costly ones, at the costs obtained when the OD-travel demand has been distributed. Hence, in such a *Nash equilibrium* no actor can unilaterally increase his payoff by changing his strategy, given that all other agents are playing their Nash equilibrium strategies. This approach grows in importance when we give up the assumptions of an uncapacitated network with fixed costs per trip and of a fixed trip table because then the interrelationship between individual decisions become more evident.

Contrasted with the non-cooperative behavior in a user equilibrium which is realistic for passenger traffic on urban networks, modeling of intercity freight movements by a single network carrier should rest upon a cooperative reasoning as it is supposed for the derivation of system equilibria. In reality where vehicle drivers and haulage firms utilize one network at the same time it would be valuable to have models where some movements are organized in a cooperative manner, while other network processes are performed without cooperation. Such a model has been proposed in Harker (1988) on the basis of a variational inequality approach that will be introduced at a later stage.

Finally, the notion of a user equilibrium should be thought of as a steady state evolving after a transient (disequilibrium) phase in which the travelers successively adjust their route-choices, in order to minimize travel costs under the prevailing traffic conditions, until a situation with stable route travel costs and route flows has been reached. See, e.g., Friesz et al. (1994) and Section 3.2.5.3 on the issue of an adjustment process leading to a Wardrop equilibrium. □

3.2.3 Capacity Constraints

3.2.3.1 Fixed Capacities

(a) On the Definition of Link Capacities

Link capacities are taken into account mainly for two reasons:

1. determination of a theoretical maximum upper bound on flows through links per period;
2. modelling of congestion effects when the load of traffic becomes heavier.

The first item refers to a fixed physical limit that restricts the flow on that link from above. For example, the transmission capacity or *bandwidth* of a computer channel or communication line indicates the theoretical maximum capacity of a

connection.[19] But when the theoretical bandwidth is approached, negative factors such as transmission delay can cause deterioration in quality. The second item is usually referred to as *reliability* of a network which addresses the need of error-free communication. In data networks, for example, each signal produces noise in its surroundings that can derange other signals. These errors may be corrected up to a certain amount, but beyond this point messages cannot be guaranteed to be delivered without errors. The capacity of such a channel therefore depends on the bandwidth as well as the signal-to-noise ratio; see Bertsekas, Gallager (1992, p. 51) for further details.

The same observations apply to traffic networks. Figure 3.8 illustrates two fundamental relationships between traffic flow and speed as well as traffic density. Diagrams of that type have a long history and are derived, e.g., in Daganzo (1997). The basic observations are (a) that the period a vehicle occupies parts of the road decreases in speed and (b) that each driver disturbs the traffic in his immediate surroundings. This "noise" increases with speed and is reflected by a growing safety distance. If the congestion effect outweighs the occupation effect, we obtain the falling branch of the speed-flow relationship. As a direct contrast we gain the increasing part of the density-flow relationship. The more safety distance each driver needs, the less vehicles can utilize the road at the same time and the lower is the traffic concentration. By contrast, an increasing traffic density forces all drivers to slow down.

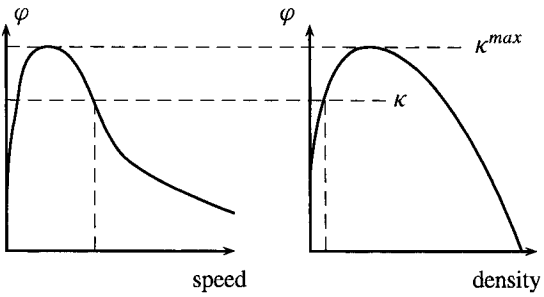


Fig. 3.8
Traffic flow relationships
The flow φ, the speed, and the density are measured in vehicles per hour, kilometers per hour, and vehicles per kilometer, respectively.

The relevant outcome here is that the flow φ cannot exceed the upper bound κ^{max}. This notion corresponds to the "bandwidth of a road". Unfortunately, this limit in general restricts the road users to drive much slower than wanted. Introducing a limited capacity κ may be seen as the traffic load to which the road works at a reliable level to guarantee a certain speed at a maximum traffic density. Such a level of service may also include aspects of risk, security, etc. Road traffic theory, however, emphasizes the opposite aspect. Here limited capacities of transportation channels have been introduced as a means for modeling non-linear congestion effects from increasing loads; cf., e.g., Larsson, Patriksson (1999). A flow exceeding this capacity incurs additional cost in the form of a lower level of service. The increased

[19] Bandwidth indicates the speed at which information can be transferred. In analog systems, it is measured in cycles per second (Hertz, Hz=1/s) and in digital systems in binary bits per second bit/s or bps.

density raises either the risk of accidents or forces the drivers to slow down. Apart from possible queuing delays, network users experience longer trip times. Other definitions of capacity take also dynamic aspects into account. The highway capacity, as defined by the Highway Capacity Manual (TRB, 1994), is the maximum flow rate achievable at a specific location on a roadway under prevailing roadway, traffic, and control conditions.[20] The term "prevailing traffic conditions" refers to a stationary (equilibrium) traffic flow as opposed to acceleration or deceleration. See Zhang (2001) for further details on highway capacity.

3.14 Remark A severe problem ignored by Figure 3.8 is that the capacity of a link cannot be analyzed independently of other parts of the network. For example, taking junctions into account the capacity of a link is only valid if the flow can leave the road by entering the relevant consecutive road. If the outgoing channel is occupied, the flow has to wait. To overcome this objection, detailed theory on the behavior of queues is needed. Another approach is suggested by the maximum flow-minimum cut theorem. In this setting the maximum traffic load on a network is restricted to the smallest cut value. Hence, every flow unit finds always a serving facility. Unfortunately, this formulation is flawed when applied to road traffic networks because there is no controlling agency that prevents trip-makers from starting their trip when the network has reached its capacity. □

(b) Implementation of Capacity Restrictions

In order to keep the analysis as simple as possible, suppose first that each link e has a fixed maximum link flow $\kappa_e^{max} = \kappa_e$ or in vector notation $\boldsymbol{\kappa}$. Here, κ_e corresponds to the maximum of the respective capacity curve as given in Figure 3.8. The commodities are homogeneous in the sense that every flow unit of each good uses one unit of capacity of each arc. Thus all *supply feasible* flows $\boldsymbol{\varphi}$ on the network are restricted to

$$\Phi^S = \left\{\boldsymbol{\varphi} \middle| \, \mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}\right\} \quad \text{or} \quad H^S = \left\{\mathbf{h} \middle| \, \mathbf{0} \leqq \boldsymbol{\Delta}\,\mathbf{h} \leqq \boldsymbol{\kappa}\right\}.$$

The problem becomes clear when imposing capacity restrictions on both arcs in Figure 3.7. The former analysis of Section 3.2.2 ignores these restrictions so that any optimal link flow satisfies $\mathbf{0} \leqq \hat{\boldsymbol{\varphi}} < \boldsymbol{\kappa}$ by assumption. In this case one can choose always an all-or-nothing strategy to assign trips to routes. This procedure needs to solve the shortest path problems (3.30) and then it assigns all of the demand $x^D_{(s,t)}$ to one of the cheapest available routes. The corresponding link flows tend to require huge capacities on a few number of links which is not realistic. Supply feasible sets are implemented in the subsequent analysis where we lay special emphasis on the appropriate pricing of capacities.

At a later stage capacity restrictions will be handled implicitly through non-linear cost functions to model congestion effects. Having the above reasoning in mind, the time τ_e needed to traverse arc e increases when congestion takes place. The *link*

[20] Branston (1976) distinguishes similarly between the terms "capacity", "practical capacity" and "service volumes".

performance function $\tau_e(\varphi_e, \kappa_e)$ describes these effects: the closer a link flow φ_e is to the respective link capacity κ_e the longer is the delay due to congestion effects. Note in accordance with Figure 3.8 that the flow can exceed a given capacity at the cost of lower speed. Furthermore, a capacity extension has the effect of shifting the capacity curve upwards. That is, more cars can be served per period at a constant level of service. The corresponding link cost function is given by $c_e = p_t \tau_e$, where p_t is a conversion factor between time and money, $[p_t] = €/h$. To give some examples, the U.S. Bureau of Public Roads (BPR, 1964) as well as Drissi-Kaitouni (1993) suggest the following functional forms of time-congestion relationships, respectively. Notice that the BPR function is homogeneous of degree zero in φ_e and κ_e so that the level of service, say τ_e, is constant as long as the intensity of road utilization φ_e/κ_e remains unchanged.

$$\tau_e = t_e \left[1 + \beta_0 \left(\frac{\varphi_e}{\kappa_e} \right)^{\beta_1} \right] \quad or \tag{3.33a}$$

$$\tau_e = t_e + \beta_0 (\max \{0, \varphi_e - \kappa_e\})^{\beta_1} \tag{3.33b}$$

Due to different properties, every link has its own free flow travel time t_e. The positive constants β_0 and β_1 take care for rapidly increasing link costs if $\varphi_e > \kappa_e$. The idea behind this construction is that a fixed capacity represents the maximal throughput at a given level of service. When the flow exceeds this capacity the level of service drops down.

Similar functions are developed for data networks with queuing which require, however, $\varphi_e < \kappa_e^{max}$. A frequently used formula (see Bertsekas, Gallager (1992, p. 434)) describes the mean delay of a message on link e by[21]

$$\tau_e = t_e + \frac{\alpha_e}{\kappa_e^{max} - \varphi_e},$$

where t_e is the processing and propagation delay and α_e is a positive parameter not explained here. Notice that this function has a pole at $\kappa_e = \varphi_e$. An application of queuing theory to road traffic is proposed in Boyce et al. (1981) stating Davidson's travel time function as

$$\tau_e = t_e \left[1 + \beta_2 \frac{\varphi_e}{\kappa_e^{max} - \varphi_e} \right].$$

Again β_2 is a calibration parameter and t_e denotes the free flow travel time. The fraction can be shown to represent the ratio of average delay and free flow travel time.

Another way to consider the physical process of congestion is to introduce an explicit congestion function $\kappa_e(\varphi_e)$. Such a function is used in Merchant, Nemhauser (1978), where $\kappa_e(\varphi_e)$ denotes the amount of traffic that exits arc e. For small traffic loads this function satisfies $\kappa_e(\varphi_e) = \varphi_e$, but when the traffic becomes heavier one has $\kappa_e(\varphi_e) < \varphi_e$ where the drivers not served have to wait in a queue. In contrast to Figure 3.8, the authors assume a non-decreasing capacity function. This is not

[21] Due to queuing properties this function is inhomogeneous in κ_e^{max} and φ_e.

necessarily a contradiction; for example, Bertsekas, Gallager (1992, p. 51) refer to *Shannon's theorem* by which the capacity of a communication channel κ_e follows from

$$\kappa_e = \kappa_e^{max} \log_2 \left[1 + \frac{\varphi_e}{\beta_0 \, \kappa_e^{max}} \right], \qquad [\kappa_e] = \text{bits/s}$$

where κ_e^{max} is the bandwidth, φ_e is the signal power, and β_0 is the noise power per unit of κ_e^{max}. Of course, implementing non-linear congestion effects prevents us from applying linear programming methods.

(c) System Optimality

In what follows the uncapacitated problem of cost minimization (3.22) is augmented by maximum flows on all arcs. With regard to *system optimality*, any solution to the advanced problem must now be demand feasible as well as supply feasible.

$$\min \left\{ \mathbf{c}^\mathsf{T} \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in \Phi^D \cap \Phi^S \right\} \quad \text{or} \quad \min \left\{ \tilde{\mathbf{c}}^\mathsf{T} \mathbf{h} \mid \mathbf{h} \in H^D \cap H^S \right\}$$

In accordance with (3.25) the route flow oriented linear program of cost minimization and its dual can be rewritten as

(3.34)

$$\begin{array}{lll} \text{(P)} \ \min \ \tilde{\mathbf{c}}^\mathsf{T} \mathbf{h} & & \text{(D)} \ \max \ \mathbf{x}^{D\mathsf{T}} \mathbf{p} - \boldsymbol{\kappa}^\mathsf{T} \mathbf{q} \\ \text{s.t.} \ \boldsymbol{\Lambda} \mathbf{h} = \mathbf{x}^D & & \text{s.t.} \ \boldsymbol{\Lambda}^\mathsf{T} \mathbf{p} - \boldsymbol{\Delta}^\mathsf{T} \mathbf{q} \leqq \tilde{\mathbf{c}} \\ \boldsymbol{\Delta} \mathbf{h} \leqq \boldsymbol{\kappa} & & \mathbf{p} \quad \text{unrestr.} \\ \mathbf{h} \geqq \mathbf{0} & & \mathbf{q} \geqq \mathbf{0} \end{array}$$

where the trip table $\mathbf{x}^D$ represents the fixed demand. The maximum supply of the network is limited via the capacity vector $\boldsymbol{\kappa}$. Notice that the non-positive dual variables which are assigned to the capacity constraints have been substituted by non-negative shadow prices $\mathbf{q}$.

$$[p_{(s,t)}] = \frac{€}{\text{trip from } v_s \text{ to } v_t} \quad \text{and} \quad [q_e] = \frac{€}{\text{flow unit on edge } e}$$

Let $\hat{\mathbf{h}}$ be a minimum solution to (3.34.P) and $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ be a maximum solution to (3.34.D). Then, similar to (3.8), the output $\mathbf{x}^D$ evaluated at the shadow price $\hat{\mathbf{p}}$ equals the operating (or transport) cost $\mathbf{c}^\mathsf{T} \hat{\boldsymbol{\varphi}}$ plus the rental value of the capital stock $\hat{\mathbf{q}}^\mathsf{T} \boldsymbol{\kappa}$.

$$\mathbf{x}^{D\mathsf{T}} \hat{\mathbf{p}} = \tilde{\mathbf{c}}^\mathsf{T} \hat{\mathbf{h}} + \boldsymbol{\kappa}^\mathsf{T} \hat{\mathbf{q}} = \mathbf{c}^\mathsf{T} \hat{\boldsymbol{\varphi}} + \boldsymbol{\kappa}^\mathsf{T} \hat{\mathbf{q}}.$$

According to the pair of programs in (3.25) the optimal values of the primal program is now denoted $C_P(\mathbf{x}^D, \boldsymbol{\kappa})$. Suppose again that the optimal values of the primal program (3.34.P) and the dual program (3.34.D) are finite. Then the primal program has a unique KKT vector $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ if and only if C_P is differentiable at $(\mathbf{x}^D, \boldsymbol{\kappa})$. In this case the statements of (3.27) and (3.28) are to be completed by

$$\frac{\partial C_P(\mathbf{x}^D, \boldsymbol{\kappa})}{\partial \kappa_e} = \hat{q}_e \geqq 0 \qquad \forall e \in \mathcal{E}.$$

As C_P depends also on the parameters $\tilde{\mathbf{c}}$, this relationship is frequently written in the form $\hat{\mathbf{q}} = \mathbf{q}(\mathbf{x}^D, \boldsymbol{\kappa}, \tilde{\mathbf{c}})$, where the function $\mathbf{q}$ denotes the gradient $\nabla_{\kappa} C_P(\mathbf{x}^D, \boldsymbol{\kappa}, \tilde{\mathbf{c}})$. The additional complementary slackness condition calls for

$$(3.35) \qquad \Big(\sum_{\rho:\, e\in\rho} \hat{h}_\rho - \kappa_e\Big)\,\hat{q}_e = (\hat{\varphi}_e - \kappa_e)\,\hat{q}_e = 0 \qquad \forall e \in \mathcal{E}.$$

In the easy case where all links are unsaturated, that is $\hat{\boldsymbol{\varphi}} = \boldsymbol{\Delta}\hat{\mathbf{h}} < \boldsymbol{\kappa}$, the optimum shadow prices of the capacities equal zero, $\hat{\mathbf{q}} = \mathbf{0}$. The programs in (3.34) reduce to (3.25). If $\hat{q}_e > 0$ the road e is filled to its capacity κ_e and we say that link e is *saturated*. The minimum travel cost per trip on every route using this particular link is increased by the cost term $\hat{q}_e$. To be more concrete we have for every route ρ leading from v_s to v_t

$$(3.36) \qquad \hat{h}_\rho \Big(\hat{p}_{(s,t)} - \sum_{e\in\rho} \hat{q}_e - \tilde{c}_\rho\Big) = \hat{h}_\rho\Big(\hat{p}_{(s,t)} - \sum_{e\in\rho}(\hat{q}_e + c_e)\Big) = 0$$

which corresponds to (3.29). A positive route flow $\hat{h}_\rho > 0$ indicates a pricing rule

$$(3.37) \qquad \hat{p}_{(s,t)} = \tilde{c}_\rho + \sum_{e\in\rho} \hat{q}_e = \sum_{e\in\rho}(c_e + \hat{q}_e) \qquad (\text{or } \hat{\mathbf{p}} = \tilde{\mathbf{c}} + \boldsymbol{\Delta}^{\mathsf{T}}\hat{\mathbf{q}}),$$

which is sometimes called element based charging[22] by which a route specific penalty

$$\hat{\tilde{q}}_\rho := \sum_{e\in\rho} \hat{q}_e \qquad (\text{or } \hat{\tilde{\mathbf{q}}} = \boldsymbol{\Delta}^{\mathsf{T}}\hat{\mathbf{q}})$$

is imposed on each route. Larsson, Patriksson (1999) refer to $\hat{\mathbf{p}} = \tilde{\mathbf{c}} + \hat{\tilde{\mathbf{q}}}$ as generalized route travel costs. Each link of the route working at its limit increases the route cost by the corresponding shadow price. This phenomenon has already been described in the example of Figure 3.4, p. 88. Now even less comfortable routes with a high $\tilde{c}_\rho$ may be used if the cheaper routes are congested. The optimal values of the link specific shadow prices $\hat{\mathbf{q}}$ determine route specific shadow prices $\hat{\tilde{\mathbf{q}}}$ that drivers on saturated routes ($\hat{\tilde{q}}_\rho \geqq 0$) are willing to pay for letting them continue to use routes that are cheaper than the non-saturated ones ($\hat{\tilde{q}}_\rho = 0$). McBride, Mamer (2001) solve (3.34.P) in terms of link flows. They derive the same pricing rule and present a "shortest path-based pricing algorithm" (a modified simplex algorithm) to solve the problem more efficiently.

[22] Element or capacity based charging as defined by the UK regulator of telecommunications industry Oftel (`http://www.oftel.gov.uk/`) means charging on the basis of the costs of the capacity in the network and switches needed to provide a service.

(d) External Pecuniary Costs

In accordance with Summary 3.8, the generalized route travel costs $\hat{\mathbf{p}}$ can be interpreted again as the vector of minimum route costs (or shortest path distances).

$$\hat{p}_{(s,t)} = p_{(s,t)} \quad \text{with} \quad p_{(s,t)} := \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho + \hat{\tilde{q}}_\rho \quad \text{and} \quad \hat{\tilde{q}}_\rho = \sum_{e \in \rho} \hat{q}_e \tag{3.38}$$

The price increments $\hat{\mathbf{q}}$ imposed by all network users in common are termed *pecuniary external economies* by Viner (1931). A price $\hat{q}_e$ represents an increment to (producers') rent on intramarginal capacity units of κ_e and is not part of the real costs of increased output. The real marginal costs are only those which have to be paid for the required marginal inputs. To clarify this statement think of an optimal primal solution $\hat{\mathbf{h}}$ with an initial trip table $\mathbf{x}_0^D$. Let ρ_1 be the uniquely determined cheapest route from v_s to v_t with free capacities on the corresponding links, $\hat{q}_e^0 = 0$ for all $e \in \rho_1$. A marginal increase of $x_{(s,t)}^D$ so that $\mathbf{x}_0^D \to \mathbf{x}_1^D$ is then assigned to $\hat{h}_{\rho_1}$ and the real marginal costs of that trip are

$$\frac{\partial C_P(\mathbf{x}_0^D, \boldsymbol{\kappa})}{\partial x_{(s,t)}^D} = \hat{p}_{(s,t)}^0 = \sum_{e \in \rho_1} c_e = \tilde{c}_{\rho_1}.$$

Suppose now that another additional trip cannot use the particular route ρ_1 because of *one* active capacity restriction on link e. When other travelers cannot be displaced from using that link, the trip-maker chooses the cheapest of the remaining routes, say $\rho_2 \in \mathcal{R}(s,t) \setminus \{\rho_1\}$. Again, if this route has free capacities on all arcs then

$$\frac{\partial C_P(\mathbf{x}_1^D, \boldsymbol{\kappa})}{\partial x_{(s,t)}^D} = \hat{p}_{(s,t)}^1 = \sum_{e \in \rho_2} c_e = \tilde{c}_{\rho_2} > \tilde{c}_{\rho_1}.$$

The real marginal costs of the last trip are $\tilde{c}_{\rho_2}$. The crucial point is that all trips from v_s to v_t describe a homogeneous good and, therefore, have the same value, namely, $\tilde{c}_{\rho_2}$. The real marginal costs of all intramarginal trips are increased by $\hat{q}_e^1 = \tilde{c}_{\rho_2} - \tilde{c}_{\rho_1}$ representing the pecuniary external economies. An alternative solution is to increase κ_e marginally so that the last trip-maker can use the cheaper route ρ_1 instead of ρ_2. The value of the hypothetical capacity extension follows from

$$\frac{\partial C_P(\mathbf{x}_1^D, \boldsymbol{\kappa})}{\partial \kappa_e} = \hat{q}_e^1 = \tilde{c}_{\rho_2} - \tilde{c}_{\rho_1}.$$

Again all capacity units of κ_e have the same value. In the example the value of the intramarginal capacity units is increased from $\hat{q}_e^0 = 0$ to $\hat{q}_e^1$. The real marginal costs of the last trip are $\tilde{c}_{\rho_1}$, where $\hat{q}_e$ represents the external pecuniary costs. Both concepts together indicate

$$\frac{\partial C_P(\mathbf{x}_1^D, \boldsymbol{\kappa})}{\partial x_{(s,t)}^D} = \hat{p}_{(s,t)}^1 = \tilde{c}_{\rho_2} = \tilde{c}_{\rho_1} + \hat{q}_e^1 = \tilde{c}_{\rho_1} + \frac{\partial C_P(\mathbf{x}_1^D, \boldsymbol{\kappa})}{\partial \kappa_e}$$

Note that these results are apparent in all markets with excess demands. The price effects detected above must not be mixed up with external effects due to congestion.

Summarizing, side constraints may be used for calculating proper link tolls to be imposed on the travelers in order to limit some volumes of traffic to levels that are acceptable. The validity of using prescriptive side constraints in traffic models rests on the assumption that their effects are transferable to the perception of travel costs among the trip-makers, for example as queuing delays or link tolls.

3.15 Remark (External Costs on Congested Networks) Some authors refer to the extra cost $\mathbf{q}$ as *delay*, which is incurred when the demand for a link exceeds its capacity. At this point the term "delay" is as misleading as the term "congestion" because it suggests effects of external diseconomies. But the observation that an excess demand causes congestion and forces all drivers to slow down their speed is *not* described by the shadow prices. Recall that fixed capacities have been justified by the assumption of a constant level of service which is measured in terms of travel speed. Delays, however, suggest that the speed slows down so that $\boldsymbol{\kappa}$ varies. The price vector $\hat{\mathbf{q}}$ merely ensures that trip-makers become indifferent between certain routes having different route costs. The effect is that network users are willing to choose a "longer" route when the corresponding "delay" with respect to a "shorter" route is reflected by the shadow prices for capacitated links on the initially more favorable route. Hence, $\hat{\mathbf{q}}$ represents the link tolls that drivers on saturated routes are willing to pay for being allowed to continue to use routes with lower trip times than non-saturated ones. Bear in mind that this argument refers to system optimality, where the travelers have to pay system optimal prices $\hat{\mathrm{p}}_{(s,t)}$ instead of $\hat{c}_\rho$ with $\rho \in \mathcal{R}(s,t)$. On the difference between pecuniary and technological externalities see Remark 3.35, p. 143.

In other settings with the focus on reliability instead of bandwidth link capacities are used to model congestion effects. Capacities then represent the so-called saturation link flows. When a link is saturated, congestion effects result in *queuing* and any excess flow will accumulate in the queue.[23] In an equilibrium state, the saturated links may therefore carry stationary queues. Link capacities also arise naturally when links are signal-controlled. In networks where saturated links have queues at their exits the multipliers $\mathbf{q}$ may be interpreted as the equilibrium time delays caused by the waiting time in steady state link queues. The link queuing interpretation also provides a queue equilibrium characterization of solutions to the model.[24] A feasible link flow $\boldsymbol{\varphi}$ for the capacitated traffic assignment problem together with a vector of link queuing delays is defined to be a queue equilibrium if the links unsaturated at $\boldsymbol{\varphi}$ carry no queues; see Larsson, Patriksson (1999).

In recent years it is common practice for modelers to work directly with equilibrium conditions of the form (3.36). The trick is to omit $\boldsymbol{\Lambda}\,\mathbf{h} \leqq \boldsymbol{\kappa}$ and to model

[23] In telephone networks there is no queuing and the congestion is measured by the probability that all lines are occupied.

[24] In case the traffic flow is not in a steady state, the link flows are unstable, the generalized route travel costs and the route flows vary, and the queue at each exit of a link is building up if the link is over-saturated, and dissolves if it is de-saturating.

equivalent of a link delay due to increased flows. The link specific version of (3.36) is then rewritten as

$$\hat{\varphi}_e\left(\hat{\mathrm{p}}_e - c_e - d_e\hat{\varphi}_e\right) = 0, \quad \hat{\varphi}_e \geqq 0, \quad \hat{\mathrm{p}}_e - c_e - d_e\hat{\varphi}_e \geqq 0.$$

Such a flow pattern is not necessarily optimal to the cost minimization problem at hand but to some other quadratic assignment problems. For example, Jones, Theise (1989) use this method to introduce congestion effects into the problem of finding a spatial price equilibrium (3.12) as given on p. 91. □

(e) User Optimum

The interpretation of the above system optimal solution as a Wardrop equilibrium rests on the assumption that the effects of the capacity constraints can be transferred to the trip-makers' perception of travel cost. Without a network provider (or some other institution) forcing the travelers to pay appropriate trip prices $\hat{\mathbf{p}}$, the travelers tend to ignore the shadow prices (or fines) of the capacitated links. In fact their route choice depends only on the route costs $\tilde{\mathbf{c}}$, see Definition 3.10. At the beginning nobody is willing to choose any route which is not among the cheapest routes. Notice that this behavior may rule out a feasible solution because some travelers cannot finish their trips on the chosen route due to capacity restrictions; these travelers are blocked by the network. The only way to overcome this trap is to introduce new aspects into users' route choice. That may be the risk not to finish a trip on the route chosen or other disadvantages depending on the excess demand per link. The corresponding "market solution" is presented in the subsequent section.

Another problem of system optimality with regard to (3.38) and (3.37) is quite obvious. As can be seen in (3.35), the shadow prices $\hat{\mathbf{q}}$, $\hat{\tilde{\mathbf{q}}}$, and thus $\mathbf{p}$ reflect optimal traffic assignments, which are unknown to the network participants in advance. This lack of information makes it more difficult for travelers to behave in a system optimal sense by making appropriate route choices. All network users may know the complete vector of travel costs $\tilde{\mathbf{c}}$ but they cannot observe $\hat{\mathbf{q}}$, $\hat{\tilde{\mathbf{q}}}$, and $\mathbf{p}$ outside an equilibrium flow $\hat{\mathbf{h}}$. Being in a system optimum $\hat{\mathbf{h}}$, however, reveals the value of $\hat{\mathbf{q}}$ and, thus, $\hat{\tilde{\mathbf{q}}} = \boldsymbol{\Delta}^{\mathsf{T}}\hat{\mathbf{q}}$. In this case there is no less costly alternative route choice for any of the drivers.

$$\rho_1 \neq \rho_2,\ \rho_1 \sim \rho_2 \colon h_{\rho_1} > 0 \implies \tilde{c}_{\rho_1} + \hat{\tilde{\mathrm{q}}}_{\rho_1} \leqq \tilde{c}_{\rho_2} + \hat{\tilde{\mathrm{q}}}_{\rho_2}$$

How do cost minimizing network users behave offside a network optimum? To give an idea, let $\tilde{\mathbf{q}}(\mathbf{h})$ denote the true supplementary charges prevailing at any flow $\mathbf{h}$ such that $\tilde{\mathbf{q}}(\hat{\mathbf{h}}) = \hat{\tilde{\mathbf{q}}}$. For every disequilibrium $\mathbf{h} \neq \hat{\mathbf{h}}$ some drivers can find a less costly route so that $\mathbf{h}$ changes permanently and $\tilde{\mathbf{q}}(\mathbf{h})$ cannot be observed directly. Hence, individuals tend to base their route choice on expected route costs $\tilde{\mathbf{c}}^E$ which are computed of $\tilde{\mathbf{c}}$ and the expected supplementary charges $\tilde{\mathbf{q}}^E$. Now let $\mathbf{p}^E$ be the expected minimum trip costs,

$$p^E_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}^E_\rho = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho + \tilde{\mathrm{q}}^E_\rho.$$

A user equilibrium $\hat{\mathbf{h}}$ requires all expectations to be fulfilled, $\tilde{\mathbf{q}}^E = \tilde{\mathbf{q}}(\hat{\mathbf{h}})$, and

$$\begin{aligned} p^E_{(s,t)} < \tilde{c}^E_\rho &\implies \hat{h}_\rho = 0, \\ \hat{h}_\rho > 0 &\implies p^E_{(s,t)} = \tilde{c}^E_\rho. \end{aligned}$$

Of course, trip-makers have to revise their expectations if the real network conditions disprove them, i.e., $\tilde{\mathbf{q}}^E \neq \tilde{\mathbf{q}}(\mathbf{h})$. One such mechanism follows from adaptive expectations, where the presumed prices of step $k+1$ are given by the expected prices of step k plus a weighted forecasting error which results from the realized minimum cost vector $\mathbf{p}(\mathbf{h}_k)$ at step k.

$$\mathbf{p}^E_{k+1} = \mathbf{p}^E_k + \delta\left(\mathbf{p}(\mathbf{h}_k) - \mathbf{p}^E_k\right), \qquad 0 < \delta < 1$$

This behavior implies a repeated sequence of the form

$$\mathbf{p}^E_k \rightarrow \mathbf{h}_k \rightarrow \mathbf{p}(\mathbf{h}_k) \rightarrow \mathbf{p}^E_{k+1}$$

which may start at $\tilde{\mathbf{q}}^E = \mathbf{0}$ or $\mathbf{p}^E_0 = \tilde{\mathbf{c}}$. The procedure would stop if expectations coincide with reality so that $\mathbf{p}^E_{k+1} = \mathbf{p}^E_k = \mathbf{p}(\hat{\mathbf{h}})$. Bear in mind that it cannot be guaranteed in advance that such a mechanism converges to a steady state where users' choice is consistent with their expectations even if a system optimum exists. On two tâtonnement models for calculating user equilibria see Friesz et al. (1994). The authors present approaches where the network users must estimate the trip costs with full or incomplete information on the travel cost of the past and the present day. Furthermore, they investigate how traveler information systems can help to improve the adjustment process after the failure of some network components.

To sum up, we have learned for the current linear case as given in (3.23) that the system optimum coincides with a Wardrop equilibrium povided the travelers anticipate correct costs of travel. Expectation errors, however, require further mechanisms such that the price tâtonnement leads to a system optimum where prevailing prices acknowledge anticipations of them.

(f) Game Theoretic Aspects

Yang, Lam (1996) derive optimal road tolls within a bilevel programming model. Without going into the details of bilevel programming at this point, the basic idea is to split one problem into two smaller but interdependent problems. For an illustration look at the capacitated traffic assignment problem (3.34.P) in link flow notation. The dual problem is to find an appropriate valuation of the trip matrix $\mathbf{x}^D$ and of the capacities κ. This suggests to take apart capacity related questions from finding optimal flows. In doing so an additional actor – namely, the market[25] – is introduced who maximizes the value of the excess demand on all links by choosing a price vector $\mathbf{q}$ and taking the flow φ^* as given. The traffic assignment φ^* in turn represents the optimal reaction on a given price vector $\mathbf{q}^*$ selected by the market.

[25] The behavior of the market reflects the observation that a price increases in the case of an excess demand and decreases in the case of an excess supply.

$$\begin{array}{lll} \max\ (\boldsymbol{\varphi}^* - \boldsymbol{\kappa})^{\mathsf{T}}\mathbf{q} & & \min\ (\mathbf{c} + \mathbf{q}^*)^{\mathsf{T}}\boldsymbol{\varphi} \\ \text{s.t.} \quad \mathbf{q} \leqq \bar{\mathbf{q}} & & \text{s.t.} \quad \boldsymbol{\varphi} - \boldsymbol{\Delta}\,\mathbf{h} = \mathbf{0} \\ \mathbf{q} \geqq \mathbf{0} & & \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D \\ & & \mathbf{h} \geqq \mathbf{0} \end{array}$$

The vector $\bar{\mathbf{q}}$ includes prohibitive prices such that $\varphi_e = 0$ if $q_e = \bar{q}_e$. This setting presupposes that the travelers can choose a feasible link flow which implicitly satisfies $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$, $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$, and $\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D$. The complementary slackness conditions of the market problem

$$\begin{array}{llll} \varphi_e^* - \kappa_e - \hat{\mu}_e \leqq 0, & \hat{q}_e \geqq 0, & (\varphi_e^* - \kappa_e - \hat{\mu}_e)\hat{q}_e = 0, & \forall\, e \in \mathcal{E} \\ \bar{q}_e - \hat{q}_e \geqq 0, & \hat{\mu}_e \geqq 0, & (\bar{q}_e - \hat{q}_e)\hat{\mu}_e = 0 & \forall\, e \in \mathcal{E} \end{array}$$

show that the dual variable $\hat{\mu}_e$ is always zero. The fact that $\hat{\mu}_e > 0$ implies $\bar{q}_e = \hat{q}_e > 0$ which gives in turn $\varphi_e^* - \kappa_e = \hat{\mu}_e > 0$ reveals a contradiction: either $\bar{q}_e = \hat{q}_e$ yields $\varphi_e^* = 0$ ($\bar{q}_e$ is defined this way) or we have $\varphi_e^* \leqq \kappa_e$. With $\hat{\mu}_e = 0$ the market sets $\hat{q}_e = 0$ for every unsaturated link with $\varphi_e^* < \kappa_e$. Any positive toll $\hat{q}_e > 0$ implies a saturated link $\varphi_e^* = \kappa_e$.

In the case where the two problems are solved simultaneously, a solution $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{q}})$ has been found if $\hat{\boldsymbol{\varphi}}$ solves the cost minimization problem given $\hat{\mathbf{q}}$, and at the same time $\hat{\mathbf{q}}$ maximizes the value of excess demand given $\hat{\boldsymbol{\varphi}}$. This solution is referred to as a *Nash equilibrium*, where each player's choice is a best response to the strategy actually chosen by his rival.

Bilevel programming, however, assumes that the *upper level (maximum) problem* is solved first and that the reaction which is determined by the *lower level (minimum) problem* is known in advance. To be precise, bilevel programming[26] supposes that the variable $\boldsymbol{\varphi}$ in the upper level problem is constrained to be an optimal solution of the lower level problem. The outcome is usually called the *von Stackelberg equilibrium*, where here the market plays the part of the indepent actor and the dependent travelers have to follow the leader's decision. Bear in mind that the two types of equilibria can differ in general. Furthermore, this example of bilevel programming is of more importance when the tolls $\sum_{e \in \rho} q_e$ along the cheapest route ρ connecting v_s and v_t reduce the demand for that OD-pair. These aspects will be discussed in Section 4.3.2.

3.2.3.2 Other Resource Constraints

(a) Modified Capacity Needs

The linear program (3.34) and particularly the set of restrictions is frequently used in alternative contexts, cf. for instance Gallego, van Ryzin (1997). Suppose for a multicommodity framework that each good is produced at a distinct node $v \in \mathcal{V}$. That is, node v_s produces good s and the demand $x^D_{(s,t)}$ at consumer destination v_t is covered by the shipment $\sum_{\rho \in \mathcal{R}(s,t)} h_\rho$ such that $\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D$. Assume furthermore

[26] For an up-to-date bibliography review see Vicente, Calamai (1994).

that each unit of h_ρ needs $\Xi_{e\rho}$ units of resource e having the price c_e. In rail transport, for example, κ_e denotes the capacity supplied by one train on arc e and the units of all commodities moving along this link can be of different size. Intertemporal aspects such as connecting trains are ignored for the moment; for details on synchronization of network processes see Chapter 6. Omitting these effects, however, has no importance when there is only one good to be distributed and all trips to the consumer destinations start at the only production place at the same time. Given the capacities κ_e, $e \in \mathcal{E}$, the resource usage is restricted to $\Xi\,\mathbf{h} \leqq \boldsymbol{\kappa}$. The task is to find a feasible vector $\mathbf{h} \geqq \mathbf{0}$ that minimizes transportation costs. The primal linear program (3.34.P) becomes now

$$
\begin{aligned}
(3.39) \qquad \min \quad & \mathbf{c}^{\mathsf{T}}\Xi\,\mathbf{h} \\
\text{s.t.} \quad & \Xi\,\mathbf{h} \leqq \boldsymbol{\kappa} \\
& \Lambda\,\mathbf{h} = \mathbf{x}^D \\
& \mathbf{h} \geqq \mathbf{0}.
\end{aligned}
$$

With minor modifications the solution technique that has been discussed in the sequel of (3.34) applies to this more general model as well.

(b) Capacitated Trips

Another case of resource constraints follows from minimizing transport costs when each trip along arc e has a capacity $\tilde{\kappa}_e$, where at the most κ_e trips along this arc are available. For example, compare passenger cars and busses or think of transportation by trucks of different size. It is assumed here for explanatory purposes that the units of all commodities are equal-sized and that each shipment along an arc has the same capacity. This idea can be justified if we think of one vehicle per arc which commutes between the two adjacent nodes. Given the demand for transportation $\mathbf{x}^D$ each freight $x^D_{(s,t)}$ is decomposed into packets which are sent along different routes. Let h_ρ designate the number of packets using route ρ. As before, the OD/path incidence matrix Λ yields $\Lambda\,\mathbf{h} = \mathbf{x}^D$. The next step is to calculate the shipment along each arc, i.e., $\Delta\,\mathbf{h}$ by making use of the arc/path incidence matrix Δ. Hence, we need at least φ_e trips along arc e satisfying $\tilde{\kappa}_e\varphi_e \geqq \sum_{\rho:\, e\in\rho} h_\rho$ or in vector notation $\mathbf{K}\,\boldsymbol{\varphi} \geqq \Delta\,\mathbf{h}$, where $\mathbf{K}$ is a diagonal matrix with the elements $\tilde{\kappa}_e$ on the leading diagonal. When each trip along arc e incurs fixed cost c_e, the minimization of transport costs can be stated as

$$
\begin{aligned}
\min \quad & \mathbf{c}^{\mathsf{T}}\boldsymbol{\varphi} \\
\text{s.t.} \quad & \Lambda\,\mathbf{h} = \mathbf{x}^D \\
& \mathbf{K}\,\boldsymbol{\varphi} - \Delta\,\mathbf{h} \geqq \mathbf{0} \\
& \boldsymbol{\varphi} \leqq \boldsymbol{\kappa} \\
& \boldsymbol{\varphi} \in \mathbb{Z}^m_+ \\
& \mathbf{h} \geqq \mathbf{0}.
\end{aligned}
$$

Although the matrix $\mathbf{K}$ should be invertible, the variable $\boldsymbol{\varphi}$ cannot be eliminated as in the linear program (3.23). This is due to the facts that an optimal solution does not require the equality in the second restriction especially when the link flow vector must be integer, $\boldsymbol{\varphi} \in \mathbb{Z}_+^m$.

Apart from postulating an integer solution there is no substantial difference in comparison to the linear program (3.34). The corresponding analysis is left to the reader.

(c) Capacitated Nodes

With special regard to aviation, network nodes such as airports are also subject to congestion. Nodal capacity restrictions are a particular problem of so-called hub-and-spoke networks, where traffic is redirected through special upgraded nodes, namely, hubs; see Section 5.2.3.1 for further details. Beckmann (1999, p. 157) calls attention to capacitated hub-and-spoke systems, where each hub is able to serve a maximum number of flow units per period. Let b_v denote the capacity of the hub located at node $v \in \mathcal{V}$. Then, the cost minimization problem (3.34.P) must be supplemented with some restriction of the form

$$\sum_v \varphi_{iv} + \sum_t x_{vt}^D \leqq b_v \qquad \forall v \in \mathcal{V}.$$

The inflow of each node plus the generated traffic at that node is limited to an upper bound b_v.

(d) Generalized (Capacity) Restrictions

Larsson, Patriksson (1999) discuss a traffic assignment problem similar to (3.34.P), but with a more general set of convex restrictions, $\mathbf{g}(\boldsymbol{\varphi}) \leqq \mathbf{0}$. Capacity restrictions of the form $\boldsymbol{\varphi} = \boldsymbol{\Delta}\mathbf{h} \leqq \boldsymbol{\kappa}$ are just a special case of these side constraints. The Lagrangean function of the modified problem (3.34.P)

$$\mathcal{L}(\mathbf{h}, \mathbf{p}, \mathbf{q}) = \tilde{\mathbf{c}}^\mathsf{T}\mathbf{h} + \mathbf{p}^\mathsf{T}(\boldsymbol{\Lambda}\,\mathbf{h} - \mathbf{x}^D) + \mathbf{q}^\mathsf{T}\mathbf{g}(\boldsymbol{\Delta}\,\mathbf{h})$$

suggests to define

$$\bar{\mathbf{c}} = \tilde{\mathbf{c}} + \boldsymbol{\Delta}^\mathsf{T}\nabla\mathbf{g}(\boldsymbol{\Delta}\,\mathbf{h})\mathbf{q}$$

as *generalized route travel costs*. According to Definition 3.10 the *generalized Wardrop equilibrium* satisfies the adjusted statements of (3.31a) and (3.31b)

$$\hat{h}_\rho > 0 \implies \bar{c}_\rho = \hat{\mathrm{p}}_{(s,t)},$$
$$\hat{h}_\rho = 0 \implies \bar{c}_\rho \geqq \hat{\mathrm{p}}_{(s,t)},$$

where the optimal Lagrange multipliers $\hat{\mathbf{p}}$ have the meaning of minimal generalized route costs; see Larsson, Patriksson (1999, Theorem 2.1).

3.2.3.3 Existence of Feasible Solutions

Considering fixed link capacities, the primal problem (3.34.P) of specifying a cost minimal trip assignment needs not have a feasible solution. This observation is not new and has been discussed in terms of the maximum flow problem. Of course, the same result holds true for multicommodity networks: *fixed link capacities determine an upper bound on feasible flows and thus on the traffic demand that can be served.* In situations where no feasible flow exists several questions arise.

1. What is the minimum investment cost I to increase κ, so that the cost minimization problem (3.34) has a technically feasible solution with $\mathbf{h} \in H^D$ and $\mathbf{h} \in H^S$? Problems of that type will be discussed in Section 5.2.
2. How to adjust κ, when investment cost I *and* travel cost $\tilde{\mathbf{c}}^T\mathbf{h}$ are to be minimized over a certain time horizon? Bear in mind that travel costs are measured per period while the capacities are determined by capital stocks which indicate durable inputs.
3. When a feasible solution exists, it is sometimes important to know whether a feasible solution still exists if one edge or node fails. This aspect refers to the *survivability* of networks and will be treated in Section 5.3.

All of the aforementioned problems seek a demand feasible, least cost flow pattern that can be realized within given capacity restricitons. In alternative settings, however, it is more natural to ask for revenue maximizing activities that can be realized on a capacitated network. This inverse issue presumes that parts of the demand can be rejected particularly when their costs exceed the corresponding revenues.

3.2.3.4 Revenue Maximization

From the carrier's point of view the cost minimal production of a trip matrix has a revenue maximizing counterpart which takes link capacities and output prices as given. This section gives a brief introduction of what is usually referred to as *network revenue management.* Talluri, van Ryzin (1999), for example, discuss such a problem where the task is to determine decision rules for sequentially accepting or denying (probabilistic) requests for trips. If different customers are willing to pay different prices to receive essentially the same service, the carrier has an incentive to implement price discrimination. For example, airline business travelers usually pay a higher fare than leisure travelers, although both trips may be seen as perfect substitutes. Kraft et al. (2000) discuss *discount allocation*[27] as one technique to protect seats for high valued future demand by limiting the current availability of low priced fares. Fare restrictions, such as Saturday night overstays as well as limitations

[27] Two other strategies are *over-booking* and *traffic management.* Following Kraft et al. (2000) the purpose of over-booking is to compensate for reservations made but not used at a statistical "no show" rate. Traffic management addresses the situation where different customers receive different services requiring the use of different resources. The purpose is to determine an optimal traffic mix with respect to the fact that short distance passengers pay a higher fare per mile.

on the total amount of low priced capacity offered, attempt to prevent travelers with a high willingness to pay from being able to utilize discounted fares.

The general problem of (deterministic) revenue maximization is as follows. Let $\mathbf{h}^D$ be the given demand for itineraries which cannot be substituted by each other even if $h^D_{\rho'}$ and $h^D_{\rho''}$ refer to the same OD-pair, i.e., $\rho' \sim \rho''$. Furthermore, all prices p_ρ per trip on route ρ are supposed to be given by the market. Revenue maximization and its dual then have the form[28]

$$
\begin{array}{lll}
(3.40) & \text{(P)} \max \ \mathbf{p}^\mathsf{T}\mathbf{h} & \text{(D)} \min \ \boldsymbol{\kappa}^\mathsf{T}\mathbf{q} + \mathbf{h}^{D\mathsf{T}}\mathbf{p} \\
& \text{s.t. } \boldsymbol{\Delta}\,\mathbf{h} \leqq \boldsymbol{\kappa} & \text{s.t. } \boldsymbol{\Delta}^\mathsf{T}\mathbf{q} + \mathbf{p} \geqq \mathbf{p} \\
& \mathbf{h} \leqq \mathbf{h}^D & \mathbf{q} \geqq \mathbf{0} \\
& \mathbf{h} \geqq \mathbf{0} & \mathbf{p} \geqq \mathbf{0}
\end{array}
$$

where the shadow prices $\mathbf{q}$ of the capacities are identified as *bid prices*. The complementary slackness (Theorem 2.5) has three major implications.

$$(\alpha_1)\ \hat{\mathrm{p}}_\rho > 0 \Rightarrow \hat{h}_\rho = h^D_\rho > 0 \Rightarrow \sum_{e\in\rho} \hat{\mathrm{q}}_e + \hat{\mathrm{p}}_\rho = p_\rho \Rightarrow \sum_{e\in\rho} \hat{\mathrm{q}}_e < p_\rho \Rightarrow (\alpha_1)$$

$$(\alpha_2)\ \sum_{e\in\rho} \hat{\mathrm{q}}_e + \hat{\mathrm{p}}_\rho > p_\rho \Rightarrow \hat{h}_\rho = 0 < h^D_\rho \Rightarrow \hat{\mathrm{p}}_\rho = 0 \Rightarrow \sum_{e\in\rho} \hat{\mathrm{q}}_e > p_\rho \Rightarrow (\alpha_2)$$

$$0 < \hat{h}_\rho < h^D_\rho \Rightarrow \hat{\mathrm{p}}_\rho = 0 \quad \text{and} \quad \sum_{e\in\rho} \hat{\mathrm{q}}_e = p_\rho$$

The second and the third case say that there is at least one link $e \in \rho$ with $\hat{\mathrm{q}}_e > 0$ working at its capacity limit κ_e. This corresponds to $\hat{h}_\rho < h^D_\rho$. Now suppose an optimal primal solution $\hat{\mathbf{h}}$ as well as a dual solution $(\hat{\mathbf{q}}, \hat{\mathbf{p}})$ such that $\mathbf{p}^\mathsf{T}\hat{\mathbf{h}} = \hat{\mathbf{q}}^\mathsf{T}\boldsymbol{\kappa} + \hat{\mathbf{p}}^\mathsf{T}\mathbf{h}^D$. Due to capacity restrictions, an additional request for a trip on route ρ is rejected in the second and the third case. The request is accepted only in the first case provided all links involved show free capacities, $\sum_{e\in\rho} \hat{\mathrm{q}}_e = 0$, such that $\hat{\mathrm{p}}_\rho = p_\rho$. The same result follows from the assumption that the primal problem has a unique KKT vector $(\hat{\mathbf{q}}, \hat{\mathbf{p}})$. In this case the optimal value of the primal program $R(\mathbf{p}, \boldsymbol{\kappa}, \mathbf{h}^D)$ is differentiable at $\mathbf{h}^D$ and

$$\frac{\partial R(\mathbf{p}, \boldsymbol{\kappa}, \mathbf{h}^D)}{\partial h^D_\rho} = \hat{\mathrm{p}}_\rho.$$

Increasing h^D_ρ, i.e., accepting a request, improves the revenue only if $\hat{\mathrm{p}}_\rho > 0$ (see the first case) and if there is sufficient capacity left on each link with respect to the increment of demand.

The idea in Talluri, van Ryzin (1999) for incorporating more stochastic information into the linear programming approach is to replace the demand vector $\mathbf{h}^D$ by a random vector. The authors suggest to simulate N independent samples of $\mathbf{h}^D_i$ and

[28] Revenue maximization is in fact an advanced *Knapsack problem* which denotes an integer program of the form $\max\{\mathbf{p}^\mathsf{T}\mathbf{h} |\ \mathbf{h} \in \mathbb{Z}^n_+,\ \mathbf{a}^\mathsf{T}\mathbf{h} \leqq \kappa\}$ with $\mathbf{a} > \mathbf{0}$ representing the resource consumption per flow unit.

calculate the corresponding vectors $\hat{\mathbf{q}}_i(\kappa, \mathbf{h}_i^D)$. The arithmetic mean $\mathbf{q}^E(\kappa, \mathbf{h}^D) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{q}}_i(\kappa, \mathbf{h}_i^D)$ is then used for the bid prices.[29] Hertz Rent-a-Car and Scandinavian Airlines System use corresponding decision rules in their revenue management systems: accept a request for itinerary ρ if and only if the fare exceeds the sum of the bid prices for all the units of capacity used by the itinerary

$$\sum_{e \in \rho} \mathrm{q}_e^E(\kappa, \mathbf{h}^D) < p_\rho$$

and if there is sufficient capacity left on each link. Similar results with respect to dynamic programming are discussed in Talluri, van Ryzin (1998).

Regarding road networks it is useful to think of trips between OD-pairs rather than of demand for distinct itineraries. The reason is that alternative routes connecting the same OD-pair indicate perfect substitutes. Thus, $\mathbf{h}^D$ is replaced by $\mathbf{x}^D$ and the corresponding restriction becomes $\mathbf{\Lambda}\,\mathbf{h} \leqq \mathbf{x}^D$. Furthermore, we require that all route prices p_ρ with respect to the same OD-pair are equal, $p_\rho = p_{(s,t)}$ for all $\rho \in \mathcal{R}(s,t)$, because no customer would accept a commodity if a cheaper perfect substitute exists. Contrasted with the cost minimization problem (3.34), p. 114, the adjustment of (3.40) then yields

(3.41)

$$\begin{array}{lrl} \text{(P)} & \max & \mathbf{p}^\mathsf{T}\mathbf{h} \\ & \text{s.t.} & \mathbf{\Lambda}\,\mathbf{h} \leqq \boldsymbol{\kappa} \\ & & \mathbf{\Lambda}\,\mathbf{h} \leqq \mathbf{x}^D \\ & & \mathbf{h} \geqq \mathbf{0} \end{array} \qquad \begin{array}{lrl} \text{(D)} & \min & \boldsymbol{\kappa}^\mathsf{T}\mathbf{q} + \mathbf{x}^{D\mathsf{T}}\mathbf{p} \\ & \text{s.t.} & \mathbf{\Lambda}^\mathsf{T}\mathbf{q} + \mathbf{\Lambda}^\mathsf{T}\mathbf{p} \geqq \mathbf{p} \\ & & \mathbf{q} \qquad\quad \geqq \mathbf{0} \\ & & \mathbf{p} \geqq \mathbf{0} \end{array}$$

where the components $\mathrm{p}_{(s,t)}$ of the dual variable $\mathbf{p}$ are now assigned to OD-pairs. For excess capacities on all links ($\hat{\mathbf{q}} = \mathbf{0}$) the carrier would accommodate the complete demand $\mathbf{x}^D$ implying $\hat{\mathrm{p}}_{(s,t)} = p_{(s,t)}$. In all other cases at least some link capacities are used up with $\hat{\mathrm{q}}_e > 0$ for that link. The optimal transport $\hat{\mathbf{h}}$ supplied by the carrier is then characterized by similar complementarity conditions that have been discussed in the sequel of (3.40).

It might be interesting to analyze under what conditions cost minimization according to (3.34) and revenue maximization given in (3.41) have the same optimal flow pattern $\check{\mathbf{h}}$. This coincidence requires at least $\mathbf{\Lambda}\,\check{\mathbf{h}} = \mathbf{x}^D$ and implies the same pattern of capacity utilization. Unfortunately there is no need that two coinciding KKT vectors $(\mathbf{q}^1, \mathbf{p}^1)$ and $(\mathbf{q}^2, \mathbf{p}^2)$ exist, which solve (3.34) and (3.41), respectively. To see this, pick out any positive route flow $\check{h}_\rho > 0$ with $\rho \in \mathcal{R}(s,t)$ and compare the dual constraints with opposite signs of the two sums.

$$\mathrm{p}^1_{(s,t)} - \sum_{e \in \rho} \mathrm{q}^1_e = \tilde{c}_\rho \quad \text{and} \quad \mathrm{p}^2_{(s,t)} + \sum_{e \in \rho} \mathrm{q}^2_e = p_\rho$$

[29] Gallego, van Ryzin (1997) consider in their dynamic pricing system that the pricing decisions induce demand which cannot be separated from traditional, capacity-oriented yield management decisions.

Even when the two sums are positive we cannot conclude a non-negative profit per flow unit, i.e., $p_\rho \geqq \tilde{c}_\rho$, as long as $\mathrm{p}^1_{(s,t)} \neq \mathrm{p}^2_{(s,t)}$. Conversely, $\mathrm{p}^1_{(s,t)} = \mathrm{p}^2_{(s,t)}$ is insufficient to deduce the same route flow pattern. Such a statement would need more information about the difference between the price p_ρ and the cost $\tilde{c}_\rho$.

3.2.4 Alternative Mathematical Formulations

3.2.4.1 Ordinary Convex Programs

The invalidity of a traffic equilibrium model of the form (3.25.P) may be due to its structural limitations, that is, its inherent simplicity which makes it inapplicable to more complex traffic problems. Larsson, Patriksson (1999) give some examples of such limitations: (a) the travel demand is presumed to be independent of travel times, (b) there is no discrimination between different types of vehicles, and (c) the model allows every road to carry arbitrarily large volumes of traffic. The latter case has been discussed in Section 3.2.3 and it was argued that the link performance may go down the closer the link flow is to the link capacity. This congestion effect can be modeled by increasing cost functions $\mathbf{c}(\boldsymbol{\varphi})$. As a first consequence the objective function is no longer linear in the former cost minimization problems. The same observation, however, applies to the feasible sets if, for instance, a price sensitive demand is taken into account. At first glance economists try to overcome these new aspects by transforming the former linear programs into convex problems with corresponding optimum conditions. Nevertheless, not all equilibrium conditions fit into a plausible convex program. For this reason, in recent years it is common practice for modelers to work directly with equilibrium conditions. Before proceeding with the economic analysis some of these approaches[30] are introduced. Moreover, their relations to convex programming are stated.

The determination of market equilibria may be a good example to justify the need for more general concepts in order to solve irregular problems. A market equilibrium is generated by the interaction of separate economic units that pursue their own interests (possibly in a non-cooperative manner). Nevertheless, in regular cases the equilibrium conditions are equivalent to the solution of a virtual single optimization problem. This problem is based on aggregate demand functions $\mathbf{x}^D(\mathbf{p})$ as well as aggregate supply functions $\mathbf{x}^S(\mathbf{p})$, and in accordance with Carey (1977), market clearing follows from the maximization of $f(\mathbf{p}) = \oint_{\mathbf{0}}^{\mathbf{p}} \mathbf{x}^D(\mathbf{z})\mathrm{d}\mathbf{z} - \oint_{\mathbf{0}}^{\mathbf{p}} \mathbf{x}^S(\mathbf{z})\mathrm{d}\mathbf{z}$ possibly with respect to some feasible set (the symbol $\oint$ denotes line integrals). The concept of variational inequality problems introduced below helps to solve such problems, when f does not exist, e.g., when the integrability of $\mathbf{x}^S$ or $\mathbf{x}^D$ does not apply. Without an objective function f the maximization of f becomes void, although the concept of market clearing remains valid.

Up until now the objective function of cost minimization, that is $\mathbf{c}^\mathsf{T}\boldsymbol{\varphi}$, depends on fixed link costs per unit $\mathbf{c}$ and the feasible set consists of linear restrictions.

[30] According to Florian, Hearn (1995) the network equilibrium model will be equivalently formulated as a variational inequality problem, a nonlinear complementarity problem, and a fixed point problem.

Linear programming is therefore the adequate tool to solve the previous problems. In general, however, link costs may change with respect to link flows $\boldsymbol{\varphi}$ due to congestion phenomena. Thus the question arises how to solve the above cost minimization problems when the objective function is substituted by $\mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi}$ so that

$$\min\left\{\mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi}\middle|\ \boldsymbol{\varphi}\in\Phi^{D}\right\}$$

denotes an *ordinary convex program* with linear restrictions. One standard approach to solve ordinary convex programs of the more general form

$$\min\{f(\mathbf{x})|\ \mathbf{x}\in\Omega\}, \tag{3.42}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\Omega \subset \mathbb{R}^n$ is a closed convex set, has been introduced in Section 2.2.2. It deals with Lagrangean functions and their respective saddle points, which are characterized by Karush-Kuhn-Tucker (KKT) conditions.

3.2.4.2 Variational Inequalities

Some new insights follow from solving so-called variational inequality problems. Given fixed link costs we know from Theorem 2.16 and Theorem 2.17, p. 47, that (3.22) has the same solution as the variational inequality

$$\mathbf{c}^{\mathsf{T}}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}}) \geqq 0 \qquad \forall\,\boldsymbol{\varphi}\in\Phi^{D}.$$

Similarly the primal problem (3.25.P) corresponds to the variational problem

$$\tilde{\mathbf{c}}^{\mathsf{T}}(\mathbf{h}-\hat{\mathbf{h}}) \geqq 0 \qquad \forall\,\mathbf{h}\in H^{D}.$$

The two problems may be seen as simple cases of a broad class of variational inequality problems. In fact, Theorem 3.23 below notes under what conditions an ordinary convex program of the type (3.42) is equivalent to a variational inequality problem defined as follows.

3.16 Definition (Variational Inequality Problem)[31] *The finite-dimensional variational inequality problem is to determine a vector* $\hat{\mathbf{x}}\in\Omega\subset\mathbb{R}^n$, *such that*

$$\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x}-\hat{\mathbf{x}}) \geqq 0 \qquad \forall\,\mathbf{x}\in\Omega, \tag{3.43}$$

where $\boldsymbol{\gamma}$ *is a continuous function from* Ω *to* $\mathbb{R}^n$ *and* Ω *is a closed convex set.*

By intuition, this inequality is similar to the subgradient inequality (2.9), p. 45, for a convex function f. Indeed, in the case of $\boldsymbol{\gamma}(\hat{\mathbf{x}})\in\partial f(\hat{\mathbf{x}})$ we have

$$f(\mathbf{x}) \geqq f(\hat{\mathbf{x}}) + \boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x}-\hat{\mathbf{x}}) \qquad \forall\,\mathbf{x}\in\Omega.$$

Recall that $\mathbf{0}\in\partial f(\hat{\mathbf{x}})$ is necessary and sufficient for $\hat{\mathbf{x}}$ to be in the minimum set of f, i.e., $f(\mathbf{x}) \geqq f(\hat{\mathbf{x}})$ for all $\mathbf{x}\in\mathbb{R}^n$. The corresponding result is now given by Nagurney (1999, Prop. 1.1): a vector $\hat{\mathbf{x}}$ solves (3.43) with $\Omega=\mathbb{R}^n$ if and only if $\boldsymbol{\gamma}(\hat{\mathbf{x}})=\mathbf{0}$.

[31] Cf. Nagurney (1999, p. 4).

The preceding subgradient inequality is closely related to convex programming as has been shown in Theorem 2.13, p. 46. This theorem requires that f and $\boldsymbol{\gamma}$ are related by $\boldsymbol{\gamma} = \nabla f$. When no such relationship exists, and when $\boldsymbol{\gamma}$ but not f is known, it turns out to be useful to analyze the remaining part, namely, the variational inequality problem (3.43).

An algorithm that can be applied to solve any variational inequality problem of the form (3.43) is presented in Nagurney (2000b). This algorithm guarantees convergence provided the function $\boldsymbol{\gamma}$ is monotone and Lipschitz continuous (and that a solution exists).

Geometric interpretation: The *normal cone*, say $\text{cone}(\hat{\mathbf{x}})$, to a set Ω at $\hat{\mathbf{x}}$ consists of all vectors $\mathbf{y}$ which do not make an acute angle with any line segment in Ω with $\hat{\mathbf{x}}$ as endpoint.

$$\text{cone}(\hat{\mathbf{x}}) := \left\{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y}^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) \leqq 0 \quad \forall \mathbf{x} \in \Omega\right\}$$

Thus a sufficient condition for $\hat{\mathbf{x}}$ to be a solution to the variational inequality (3.43) is that $-\boldsymbol{\gamma}(\hat{\mathbf{x}})$ is orthogonal to the feasible set Ω at point $\hat{\mathbf{x}}$; see Figure 3.9.[32]

$$-\boldsymbol{\gamma}(\hat{\mathbf{x}}) \in \text{cone}(\hat{\mathbf{x}})$$

The right hand side of this relation refers to the feasible set Ω. The left hand side is related to the objective function f if we assume that $\boldsymbol{\gamma}(\hat{\mathbf{x}})$ is a subgradient of f at $\hat{\mathbf{x}}$, that is $\boldsymbol{\gamma}(\hat{\mathbf{x}}) \in \partial f(\hat{\mathbf{x}})$ or $\boldsymbol{\gamma}(\hat{\mathbf{x}}) = \nabla f(\hat{\mathbf{x}})$ in the special case of a differentiable function f.

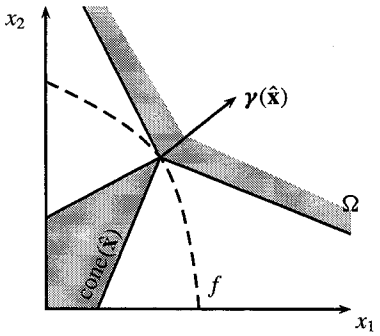


Fig. 3.9
Variational inequality problem
For $\hat{\mathbf{x}}$ being an inner point of Ω the cone reduces to the singleton $\mathbf{0}$.

Kinderlehrer, Stampacchia (1980) prove by virtue of Brouwer's fixed point theorem that the variational inequality problem (3.43) has a solution if $\boldsymbol{\gamma}$ is continuous

[32] An example where the normal cone reduces to just one vector is to realize at least a utility level U at minimum expenditures,

$$e(\mathbf{p}, U) = \min\left\{\mathbf{p}^{\mathrm{T}}\mathbf{x} \mid U \leqq u(\mathbf{x}),\ \mathbf{x} \geqq \mathbf{0}\right\}.$$

Assuming the utility function u to be differentiable, the gradient of the objective function is $\mathbf{p}$, while the normal cone is the (unique) ray having the direction of $-\mathbf{p}$. For a graphic representation it is useful to originate both the gradient and the normal cone at the optimal consumption bundle $\hat{\mathbf{x}}$.

and Ω is compact and convex; see Theorem 3.28. A similar result can be established for an unbounded feasible set Ω. The trick is to define a closed ball $B_r(\mathbf{0})$ of radius r and center $\mathbf{0}$. One can then show that $|\hat{\mathbf{x}}_r| < r$, where $\hat{\mathbf{x}}_r$ is a solution to

$$\boldsymbol{\gamma}(\hat{\mathbf{x}}_r)^\top(\mathbf{x} - \hat{\mathbf{x}}_r) \geqq 0 \qquad \forall\, \mathbf{x} \in \Omega \cap B_r(\mathbf{0})$$

for some $r > 0$, is a necessary and sufficient condition for (3.43) having a solution. Another criterion can be found in the literature on variational inequalities under the keyword "coerciveness".

In almost all economic problems of this book the feasible set Ω is of the type

$$\Omega = \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \leqq \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0}\right\},$$

where $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable and $\mathbf{h}\colon \mathbb{R}^n \to \mathbb{R}^p$ is linear affine. The following result is taken from Tobin (1986);[33] it relates the KKT conditions to a variational inequality problem.

3.17 Theorem ***Necessary conditions for a solution:*** *Let $\hat{\mathbf{x}} \in \Omega$ be a solution to the variational inequality problem* (3.43) *and suppose that the gradients $\nabla g_i(\hat{\mathbf{x}})$ of the active restrictions (that is $g_i(\hat{\mathbf{x}}) = 0$) and $\nabla h_k(\hat{\mathbf{x}})$, $k = 1,\ldots, p$, are linearly independent. Then multipliers $\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2) \in \mathbb{R}^{m+p}$ exist such that*

$$\begin{aligned}
&\boldsymbol{\gamma}(\hat{\mathbf{x}}) + \nabla\mathbf{g}(\hat{\mathbf{x}})\,\hat{\boldsymbol{\lambda}}_1 + \nabla\mathbf{h}(\hat{\mathbf{x}})\,\hat{\boldsymbol{\lambda}}_2 = \mathbf{0},\\
&\mathbf{g}(\hat{\mathbf{x}}) \leqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}_1 \geqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}_1{}^\top\mathbf{g}(\hat{\mathbf{x}}) = 0,\\
&\mathbf{h}(\hat{\mathbf{x}}) = \mathbf{0}.
\end{aligned}$$

Sufficient conditions for a solution: *If g_i, $i = 1,\ldots, m$, are convex and $\hat{\mathbf{x}} \in \Omega$, $\hat{\boldsymbol{\lambda}}_1 \in \mathbb{R}^m$ and $\hat{\boldsymbol{\lambda}}_2 \in \mathbb{R}^p$ satisfy the preceding conditions, then $\hat{\mathbf{x}}$ is a solution to the variational inequality problem* (3.43).

Be aware that the variables $\mathbf{x}$ in Theorem 3.17 are unrestricted in sign. Non-negativity constraints must be taken into account explicitly. If $\mathbf{x} \geqq \mathbf{0}$, the first condition can be rewritten as

$$(3.44)\quad \hat{\boldsymbol{\lambda}}_3 = \boldsymbol{\gamma}(\hat{\mathbf{x}}) + \nabla\mathbf{g}(\hat{\mathbf{x}})\,\hat{\boldsymbol{\lambda}}_1 + \nabla\mathbf{h}(\hat{\mathbf{x}})\,\hat{\boldsymbol{\lambda}}_2 : \qquad \hat{\boldsymbol{\lambda}}_3 \geqq \mathbf{0}, \quad \hat{\mathbf{x}} \geqq \mathbf{0}, \quad \hat{\boldsymbol{\lambda}}_3^\top\hat{\mathbf{x}} = 0.$$

The next property of monotonicity has the same importance as a convex objective function in convex programming.

3.18 Definition (Monotonicity) *The function $\boldsymbol{\gamma}\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is said to be monotone (or antimonotone in the case of $\leqq$ instead of $\geqq$) if*

$$(3.45)\qquad [\boldsymbol{\gamma}(\mathbf{x}_1) - \boldsymbol{\gamma}(\mathbf{x}_2)]^\top(\mathbf{x}_1 - \mathbf{x}_2) \geqq 0 \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \Omega.$$

When the strict inequality holds true for all $\mathbf{x}_1 \neq \mathbf{x}_2$, we say that $\boldsymbol{\gamma}$ is strictly monotone.

The correspondence $\boldsymbol{\Gamma}\colon \Omega \subset \mathbb{R}^n \to \mathfrak{P}(\mathbb{R}^n)$ is said to be monotone *if*

[33] The linear independence constraint qualification can be found in Bazaraa et al. (1993, Chap. 5.3.1).

$$(3.46) \qquad (\mathbf{y}_1 - \mathbf{y}_2)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) \geqq 0 \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \Omega,\ \mathbf{y}_1 \in \Gamma(\mathbf{x}_1),\ \mathbf{y}_2 \in \Gamma(\mathbf{x}_2).$$

Strict monotonicity is sufficient to show that there is at most one solution.

3.19 Theorem (Uniqueness) *If $\hat{\mathbf{x}}$ solves (3.43) and $\boldsymbol{\gamma}$ is strictly monotone, then $\hat{\mathbf{x}}$ is the only solution.*

Proof: Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two distinct solutions, then

$$\left.\begin{aligned} \boldsymbol{\gamma}(\mathbf{x}_1)^\mathsf{T}(\mathbf{x}_2 - \mathbf{x}_1) \geqq 0 \\ \boldsymbol{\gamma}(\mathbf{x}_2)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) \geqq 0 \end{aligned}\right\} \implies [\boldsymbol{\gamma}(\mathbf{x}_1) - \boldsymbol{\gamma}(\mathbf{x}_2)]^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) \leqq 0,$$

which contradicts (3.45). ■

A criterion to detect strict monotonicity can be derived by making use of the Jacobian matrix.

3.20 Theorem *Let $\boldsymbol{\gamma}$ be continuously differentiable on Ω. Then, $\boldsymbol{\gamma}$ is strictly monotone if the Jacobian matrix*[34] *$\mathbf{J}_{\boldsymbol{\gamma}}$ is positive definite at every $\mathbf{x} \in \Omega$.*

$$\mathbf{y}^\mathsf{T}\mathbf{J}_{\boldsymbol{\gamma}}(\mathbf{x})\mathbf{y} > 0 \qquad \forall \mathbf{y} \neq \mathbf{0}$$

Proof: For every pair $\mathbf{x}_1$ and $\mathbf{x}_2$ with $\mathbf{x}_1 \neq \mathbf{x}_2$ we define

$$g(t) := \boldsymbol{\gamma}\big(\mathbf{x}_2 + t(\mathbf{x}_1 - \mathbf{x}_2)\big)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2)$$

$$\text{with} \quad g'(t) = \big[\mathbf{J}_{\boldsymbol{\gamma}}\big(\mathbf{x}_2 + t(\mathbf{x}_1 - \mathbf{x}_2)\big)(\mathbf{x}_1 - \mathbf{x}_2)\big]^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2).$$

By the mean value theorem there is a $0 < \lambda < 1$ such that

$$g(t_1) - g(t_2) = g'\big(\lambda t_1 + (1 - \lambda)t_2\big)\,(t_1 - t_2).$$

Using $t_1 = 1$ and $t_2 = 0$ results in $g(1) - g(0) = g'(\lambda)$ or

$$\boldsymbol{\gamma}(\mathbf{x}_1)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) - \boldsymbol{\gamma}(\mathbf{x}_2)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^\mathsf{T}\mathbf{J}_{\boldsymbol{\gamma}}\big(\mathbf{x}_2 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)\big)(\mathbf{x}_1 - \mathbf{x}_2)$$

When the Jacobian matrix on the right hand side $\mathbf{J}_{\boldsymbol{\gamma}}$ is positive definite, then $\boldsymbol{\gamma}$ is strictly monotone. ■

The inverse result requires that every $\mathbf{x} \in \Omega$ and $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ can be expressed by two distinct $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ and an appropriate $0 < \lambda < 1$. This is always possible for $\Omega = \mathbb{R}^n$. In this case the word "if" in the theorem can be substituted by "if and only if".

The subsequent results reveal the relationship between the objective function f in the convex program (3.42) and the function $\boldsymbol{\gamma}$ in the variational inequality (3.43). The feasible sets coincide. In the case of $\nabla f = \boldsymbol{\gamma}$ we have $\mathbf{H}_f = \mathbf{J}_{\boldsymbol{\gamma}}$ for the Hessean matrix of f. This requires f to be twice differentiable. Using the mean value theorem there is always an $\tilde{\mathbf{x}} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$ for some $0 < \lambda < 1$ such that

$$f(\mathbf{x}_1) = f(\mathbf{x}_2) + \nabla f(\mathbf{x}_2)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)^\mathsf{T}\mathbf{H}_f(\tilde{\mathbf{x}})(\mathbf{x}_1 - \mathbf{x}_2).$$

[34] Remember $\mathbf{J}_{\boldsymbol{\gamma}} = (\nabla\boldsymbol{\gamma})^\mathsf{T}$.

Hence, the convexity of a twice differentiable function f, i.e.,

$$f(\mathbf{x}_1) \geqq f(\mathbf{x}_2) + \nabla f(\mathbf{x}_2)^{\mathsf{T}}(\mathbf{x}_1 - \mathbf{x}_2)$$

by Theorem 2.13, is equivalent to having a positive semidefinite Hessean matrix $\mathbf{H}_f$. A more general result, which does not need twice differentiability but merely monotonicity, is

3.21 Theorem[35] *Let $f\colon \Omega \to \mathbb{R}$ be differentiable on the nonempty open convex set Ω. Then, f is (strictly) convex if and only if $\nabla f = \boldsymbol{\gamma}$ is (strictly) monotone.*

Without differentiability the gradient mapping ∇f is to be substituted by the more general subdifferential mapping ∂f. Rockafellar (1972) shows in the sequel of Theorem 24.9 that the subdifferential ∂f of a closed proper convex function $f\colon \mathbb{R}^n \to]-\infty, +\infty]$ is monotone, that is (3.46) holds good for ∂f.

The following two theorems deal with the case, where f is deduced from $\boldsymbol{\gamma}$. The variational inequality problem (3.43) is then equivalent to a convex program with the convex objective function f. Corresponding results with non-differentiable and even discontinuous correspondences, i.e., set valued mappings, instead of continuous functions are given in de Palma, Nesterov (1998), Theorem 9 and 10.

3.22 Theorem (Integrability) *Let the vector mapping $\boldsymbol{\gamma}\colon \Omega \to \mathbb{R}^n$ be continuously differentiable on an open convex set $\Omega_0 \subset \Omega$. Then $\boldsymbol{\gamma}$ is a gradient mapping on Ω_0 if and only if the Jacobian Matrix $\mathbf{J}_{\boldsymbol{\gamma}} = (\nabla \boldsymbol{\gamma})^{\mathsf{T}}$ is symmetric for all $\mathbf{x} \in \Omega_0$.*

Under this symmetry condition the *line integral*[36]

$$f(\mathbf{x}) := \oint_{\mathbf{a}}^{\mathbf{x}} \boldsymbol{\gamma}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} = \sum_{j=1}^{n} \int_a^b \gamma_j(\boldsymbol{\sigma}(t))\, \mathrm{d}\sigma_j(t)$$

is independent of the path starting at $\boldsymbol{\sigma}(a) = \mathbf{a}$ and ending at $\boldsymbol{\sigma}(b) = \mathbf{x}$. The function $\boldsymbol{\gamma}$ is said to be *integrable* with $\boldsymbol{\gamma} = \nabla f$. (Note $f(\mathbf{a}) = 0$ by the definition of f.) An equivalent formulation results from a continuously differentiable path $\boldsymbol{\sigma}(t)$ with $a \leqq t \leqq b$; see Heuser (1992, Satz 180.3).

$$f(\mathbf{x}) = \int_a^b \boldsymbol{\gamma}\big(\boldsymbol{\sigma}(t)\big)^{\mathsf{T}} \dot{\boldsymbol{\sigma}}(t)\, \mathrm{d}t = \sum_{j=1}^{n} \int_a^b \gamma_j\big(\boldsymbol{\sigma}(t)\big)\, \dot{\sigma}_j(t)\, \mathrm{d}t.$$

---

[35] For the complete proof see Bazaraa et al. (1993, Theorem 3.3.4).

[36] Every function $\boldsymbol{\gamma}\colon X \to \mathbb{R}^n$ is called a *vector mapping* and every function $f\colon X \to \mathbb{R}$ is called a *scalar mapping*. The vector mapping $\boldsymbol{\gamma}$ is said to be a *gradient mapping* on an open set G if there is a scalar mapping f on G so that $\boldsymbol{\gamma}(\mathbf{x}) = \nabla f(\mathbf{x})$ for all $\mathbf{x} \in G$. The idea is to choose a path $\boldsymbol{\sigma}$ with the starting point $\mathbf{a}$ and the end point $\mathbf{b}$ lying in G. Then, similar to the fundamental theorem of calculus

$$\oint_{\boldsymbol{\sigma}} \nabla f(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} = \oint_{\mathbf{a}}^{\mathbf{b}} \boldsymbol{\gamma}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} = f(\mathbf{b}) - f(\mathbf{a}).$$

This statement means that the line integral of a continuous gradient mapping on a set G is *path independent*. Now the question arises of what conditions make a vector mapping $\boldsymbol{\gamma}$ being a gradient mapping. The exact answer is given in Theorem 3.22, but one might memorize only that the antiderivative of $\boldsymbol{\gamma}$ must be path independent. Hence, $\boldsymbol{\gamma} \to \nabla f \to f$ is possible. Cf. Heuser (1992, Satz 182.2).

Choosing the special path $\sigma(t) = t\mathbf{x} + (1-t)\mathbf{a}$ for $0 \leqq t \leqq 1$ such that $\dot{\sigma} = \mathbf{x} - \mathbf{a}$, the above integral can be rewritten as

$$f(\mathbf{x}) = \int_0^1 \boldsymbol{\gamma}\big(t\mathbf{x} + (1-t)\mathbf{a}\big)^{\mathsf{T}}(\mathbf{x} - \mathbf{a})\,\mathrm{d}t.$$

This form of the *potential function* f is frequently used in the literature on traffic assignment; cf., e.g., de Palma, Nesterov (1998). Of course, when each γ_j depends solely on σ_j with $\sigma_j(a) = a_j$ and $\sigma_j(b) = x_j$ the line integral corresponds to the sum of usual integrals.

$$f(\mathbf{x}) = \sum_{j=1}^{n} \int_{a_j}^{x_j} \gamma_j(z_j)\,\mathrm{d}z_j$$

This type of function is widely used in transportation theory since Beckmann et al. (1956).

The fundamental result in relating convex programs to variational inequalities is

3.23 Theorem *Let $\boldsymbol{\gamma}$ be integrable with $\boldsymbol{\gamma} = \nabla f$ and let the antiderivative f be convex. Then the variational inequality problem* (3.43) *is equivalent to the ordinary convex program* (3.42).

Proof: On the convexity of f and the integrability condition of $\boldsymbol{\gamma}$ step back to the Theorems 3.22 and 3.21, respectively.

Necessity: the gradient $\nabla f(\hat{\mathbf{x}})$ of a convex function f satisfies

$$f(\mathbf{x}) \geqq f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \qquad \forall \mathbf{x} \in \Omega.$$

Thus any solution of the variational inequality problem (3.43) solves (3.42) such that

$$f(\mathbf{x}) \geqq f(\hat{\mathbf{x}}) \qquad \forall \mathbf{x} \in \Omega.$$

The sufficient part does not need the convexity of f. The condition $\mathbf{0} \in \nabla f(\hat{\mathbf{x}}) + \partial\delta(\hat{\mathbf{x}}|\Omega)$ is necessary and sufficient for $\hat{\mathbf{x}}$ to be in the minimum set of $f(\hat{\mathbf{x}}) + \delta(\hat{\mathbf{x}}|\Omega)$, where δ is the indicator function of Ω. As can be found in Rockafellar (1972, p. 215), the subdifferential of δ satisfies $\mathrm{cone}(\hat{\mathbf{x}}) \equiv \partial\delta(\hat{\mathbf{x}}|\Omega)$ with $\partial\delta(\hat{\mathbf{x}}|\Omega)$ being empty if $\hat{\mathbf{x}} \notin \Omega$. Hence, $-\nabla f(\hat{\mathbf{x}}) \in \mathrm{cone}(\hat{\mathbf{x}})$ which implies (3.43). The corresponding graphic representation has been given in Figure 3.9. ■

Rearranging the above subgradient inequality to

$$f(\mathbf{x}) - f(\hat{\mathbf{x}}) \geqq \nabla f(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \qquad \forall \mathbf{x} \in \Omega,$$

gives an idea of the following analysis. It relates the change of f, when jumping from $\hat{\mathbf{x}}$ to $\mathbf{x}$, to the corresponding value when moving on a hyperplane which is tangent to the graph of f.

3.2.4.3 Algorithmic Remark

Most optimization methods for convex problems focus on the reaction of f at $\mathbf{x}$ when going into a *direction* $\mathbf{d} := \mathbf{z} - \mathbf{x}$, where $\mathbf{x} \neq \mathbf{z}$ and $\mathbf{z}, \mathbf{x} \in \Omega$. For minimization

problems the directions of interest yield $f(\mathbf{x}) > f(\mathbf{x}+\mathbf{d})$, where every point of the path $\{\mathbf{x}+\lambda\mathbf{d}|\ 0 \leqq \lambda \leqq 1\}$ is feasible as Ω is closed and convex. A more general approach starts with the opposite observation.

3.24 Definition (Ascent and Descent Directions) *A vector* $\mathbf{d}$ *is called an ascent direction if an* $\varepsilon > 0$ *exists such that*

$$f(\mathbf{x}+\lambda\mathbf{d}) - f(\mathbf{x}) > 0 \quad \forall \lambda \in]0, \varepsilon[.$$

If the opposite inequality holds true, $\mathbf{d}$ *is a descent direction.*

The same reasoning with respect to any subgradient $\mathbf{y} \in \partial f(\mathbf{x})$ states

$$f(\mathbf{x}+\mathbf{d}) - f(\mathbf{x}) \geqq \mathbf{y}^{\mathrm{T}}\mathbf{d} \qquad \forall \mathbf{x}+\mathbf{d} \in \Omega.$$

Thus a convex function f attains a minimum at $\mathbf{x} = \hat{\mathbf{x}}$ when no feasible point $\hat{\mathbf{x}} + \mathbf{d} \in \Omega$ determines a descent direction $\mathbf{d}$ with $\mathbf{y}^{\mathrm{T}}\mathbf{d} \leqq f(\mathbf{x}+\mathbf{d}) - f(\mathbf{x}) < 0$. The inverse argument says for $\mathbf{y} = \gamma(\hat{\mathbf{x}})$ being a subgradient of f at $\hat{\mathbf{x}}$ that the variational inequality (3.43) holds true in the form

$$\mathbf{y}^{\mathrm{T}}\mathbf{d} \geqq 0 \qquad \forall \hat{\mathbf{x}}+\mathbf{d} \in \Omega. \tag{3.47}$$

As long as a descent direction exists, the objective value can be improved by moving from $\mathbf{x}$ to $\mathbf{x}+\mathbf{d}$. When there is more than one descent direction it is heuristically useful to go into the direction along which f decreases at the fastest local rate. In order to compute that direction of steepest descent it is useful to define the one-sided directional derivative of f at $\mathbf{x}$ with respect to the direction $\mathbf{d}$ by

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{\lambda \downarrow 0} \frac{f(\mathbf{x}+\lambda\mathbf{d}) - f(\mathbf{x})}{\lambda}.$$

This definition can be interpreted best by using the approximation $f(\mathbf{x}+\mathbf{d}) \approx f(\mathbf{x}) + f'(\mathbf{x}, \mathbf{d})$. A statement on the largest change of f follows from Rockafellar (1972, Theorem 23.4): if f is a proper convex function, then

$$f'(\mathbf{x}; \mathbf{d}) = \sup\{\mathbf{y}^{\mathrm{T}}\mathbf{d}|\ \mathbf{y} \in \partial f(\mathbf{x})\}$$

and $f'(\mathbf{x}, \mathbf{d})$ is finite for every $\mathbf{d}$ if $\mathbf{x} \in \mathrm{int}(\mathrm{Dom}\, f)$. Furthermore, if f is differentiable at $\mathbf{x}$, one has $\hat{\mathbf{y}} = \nabla f(\mathbf{x})$ and all directional derivatives are finite and two-sided, $f'(\mathbf{x}; \mathbf{d}) = -f'(\mathbf{x}; -\mathbf{d})$. Furthermore,

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d}. \tag{3.48}$$

Without differentiability we only know $f'(\mathbf{x}; \hat{\mathbf{d}}) \geqq -f'(\mathbf{x}; -\hat{\mathbf{d}})$. The next step is to define the *direction of steepest ascent* $\hat{\mathbf{d}}$ of f at $\mathbf{x}$ by

$$f'(\mathbf{x}; \hat{\mathbf{d}}) = \max\{f'(\mathbf{x}, \mathbf{d})|\ |\mathbf{d}| \leqq 1\},$$

where $|\mathbf{d}| := (\mathbf{d}^{\mathrm{T}}\mathbf{d})^{1/2}$ is the Euclidean norm. Under the suppositions of (3.48) the objective function satisfies

$$|f'(\mathbf{x};\mathbf{d})| = |\hat{\mathbf{y}}^{\mathsf{T}}\mathbf{d}| \leqq |\hat{\mathbf{y}}|\,|\mathbf{d}| \leqq |\hat{\mathbf{y}}| \quad \text{with} \quad \hat{\mathbf{y}} = \nabla f(\mathbf{x})$$

by making use of Schwarz' inequality. Letting

$$\hat{\mathbf{d}} = \begin{cases} \mathbf{0} & \text{if } \hat{\mathbf{y}} = \mathbf{0}, \\ \hat{\mathbf{y}}/|\hat{\mathbf{y}}| & \text{if } \hat{\mathbf{y}} \neq \mathbf{0}. \end{cases}$$

we compute $|\hat{\mathbf{d}}| = 1$ and

$$f'(\mathbf{x};\hat{\mathbf{d}}) = \frac{\nabla f(\mathbf{x})^{\mathsf{T}}\nabla f(\mathbf{x})}{|\nabla f(\mathbf{x})|} = |\nabla f(\mathbf{x})| = |f'(\mathbf{x};\hat{\mathbf{d}})|.$$

In summary, *the gradient $\nabla f(\mathbf{x})$, if it exists, has the direction of the steepest ascent. Consequently, $-\nabla f(\mathbf{x})$ has the direction of the steepest descent.* A similar result without differentiability can be found in Bazaraa et al. (1993, Theorem 6.3.11).

3.25 Remark (Frank-Wolfe Algorithm) Consider the convex program

$$\min\left\{f(\mathbf{x})\,|\,\mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geqq \mathbf{0}\right\},$$

where f is a differentiable convex function. In order to find an optimal $\hat{\mathbf{x}}$ the following algorithm solves a sequence of linear programs. In doing so it makes use of feasible directions.

Step 1: Find an initially feasible $\mathbf{x}_0$.

$$\mathbf{x}_0 \leftarrow \arg\min\left\{\nabla f(\mathbf{0})^{\mathsf{T}}\mathbf{x}\,|\,\mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geqq \mathbf{0}\right\}$$

Step 2: Find another feasible point trying to go into the direction of steepest descent.

$$\mathbf{x}_1 \leftarrow \arg\min\left\{\nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{x}\,|\,\mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geqq \mathbf{0}\right\}$$

Step 3: Calculate the optimal length of the step to go.

$$\mathbf{x}_0 \leftarrow (1-\hat{\lambda})\mathbf{x}_0 + \hat{\lambda}\mathbf{x}_1 \quad \text{where} \quad \hat{\lambda} = \arg\min_{0 \leqq \lambda \leqq 1} f((1-\lambda)\mathbf{x}_0 + \lambda\mathbf{x}_1)$$

Stop if $\mathbf{x}_0$ changes sightly and $\nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{x}_1$ is close to $\nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{x}_0$. Otherwise return to Step 2.

Step 3 makes use of $\mathbf{x}_0 + \lambda\mathbf{d}$ with $\mathbf{d} = \mathbf{x}_1 - \mathbf{x}_0$, where Step 2 implies $\nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{x}_1 \leqq \nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{x}_0$ or, equivalently, $\nabla f(\mathbf{x}_0)^{\mathsf{T}}\mathbf{d} \leqq 0$. Theorem 2.13 states for a convex function f

$$f(\mathbf{x}_0 + \lambda\mathbf{d}) - f(\mathbf{x}_0) \geqq \nabla f(\mathbf{x}_0)^{\mathsf{T}}(\lambda\mathbf{d}) = \lambda f'(\mathbf{x}_0;\mathbf{d}).$$

Without convergence the right hand side is negative for every $\lambda > 0$. For $\lambda \to 0$ the limits of both sides are equal. Therefore, choosing λ small enough assures $f(\mathbf{x}_0 + \lambda\mathbf{d}) - f(\mathbf{x}_0) < 0$, so that $\mathbf{d}$ is a descent direction. An application of the Frank-Wolfe method with respect to a user optimal traffic assignment is proposed in Boyce et al. (1981). □

3.2.4.4 Nonlinear Complementary Problems

One special case of the variational inequality problem (3.43) follows immediately from the rearrangement

$$\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{x} \geqq \boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{x}} \qquad \forall\, \mathbf{x} \in \Omega.$$

Note that there is no need for $\boldsymbol{\gamma}$ to be integrable. Minimizing the left hand side of the inequality results in

3.26 Theorem and Definition[37] *For $\Omega = \mathbb{R}^n_+$ the variational inequality problem* (3.43) *and the* nonlinear complementary problem

(3.49) *find an $\hat{\mathbf{x}} \geqq \mathbf{0}$ such that $\boldsymbol{\gamma}(\hat{\mathbf{x}}) \geqq \mathbf{0}$ and $\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{x}} = 0$*

have the same solution, if any.

Proof: For the necessary part substitute $\mathbf{x} = 2\hat{\mathbf{x}}$ and $\mathbf{x} = \mathbf{0}$ into (3.43). The sufficient part results from $\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{x} \geqq 0$ and the equation in (3.49). ■

If Ω simply consists of non-negativity constraints $\mathbf{x} \geqq \mathbf{0}$, the variational inequality reduces to

$$\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{x}} = 0 \quad \text{and} \quad \boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{x} \geqq 0 \quad \forall\, \mathbf{x} \in \mathbb{R}^n_+.$$

For $\boldsymbol{\gamma} = \nabla f$ the second restriction implies $\nabla f(\hat{\mathbf{x}}) \geqq \mathbf{0}$. Minimizing f over Ω reveals the same KKT conditions $\hat{\mathbf{x}} \geqq \mathbf{0}$, $\nabla f(\hat{\mathbf{x}}) \geqq \mathbf{0}$, and $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{x}} = 0$ as has been given in (3.44).

3.2.4.5 Fixed Point Problems

Fixed point problems will be used to prove the existence of at least one traffic equilibrium. Furthermore, in many cases proving the existence of a fixed point suggests at the same time algorithms to compute them numerically. Although not in the center of interest here, the projection operator defined below can be used to pursue exactly this task.

Let $\mathbf{k}\colon \Omega \to \mathbb{R}^n$ be continuous. The *fixed point problem* is to

(3.50) find an $\hat{\mathbf{x}} \in \Omega$ such that $\hat{\mathbf{x}} = \mathbf{k}(\hat{\mathbf{x}})$.

In what follows the function $\mathbf{k}$ is substituted by the orthogonal projection of $\mathbf{z}$ on Ω defined by

$$P_\Omega(\mathbf{z}) := \arg\min\{|\mathbf{z} - \mathbf{x}| \mid \mathbf{x} \in \Omega\}.$$

The relation $\hat{\mathbf{x}} = P_\Omega(\mathbf{z})$ states that $\hat{\mathbf{x}}$ is a point in Ω closest to $\mathbf{z}$; see Figure 3.10.

$$|\mathbf{z} - \hat{\mathbf{x}}| = \inf\{\, |\mathbf{z} - \mathbf{x}| \mid \mathbf{x} \in \Omega\}$$

[37] This result is from Kinderlehrer, Stampacchia (1980). It is restated in Nagurney (1999, p. 9), Prop. 1.4.

Notice that $\mathbf{x} = P_\Omega(\mathbf{x})$ for all $\mathbf{x} \in \Omega$. One can show that the solution $\hat{\mathbf{x}}$ is unique if Ω is a non-empty closed convex set.

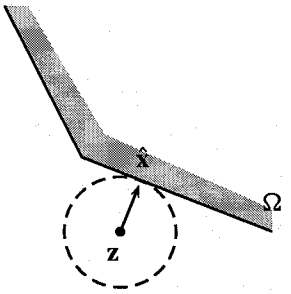


Fig. 3.10
The projection operator P_Ω

3.27 Theorem[38] *A point $\hat{\mathbf{x}}$ is an optimal solution to the variational inequality problem* (3.43) *if and only if*

$$\hat{\mathbf{x}} = P_\Omega\big(\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}})\big). \tag{3.51}$$

Proof: The theorem claims for some $\hat{\mathbf{x}} \in \Omega$ the validity of the following equivalence:

$$\boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \qquad \forall \mathbf{x} \in \Omega \tag{3.52a}$$

$$\iff \quad |\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}}) - \hat{\mathbf{x}}| \leqq |\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}}) - \mathbf{x}| \quad \forall \mathbf{x} \in \Omega \tag{3.52b}$$

The second inequality expands to

$$|\boldsymbol{\gamma}(\hat{\mathbf{x}})| \leqq |\boldsymbol{\gamma}(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})| \qquad \forall \mathbf{x} \in \Omega$$

$$\iff \quad \boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}\boldsymbol{\gamma}(\hat{\mathbf{x}}) \leqq \boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}\boldsymbol{\gamma}(\hat{\mathbf{x}}) + 2\boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}(\mathbf{x} - \hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})^\mathsf{T}(\mathbf{x} - \hat{\mathbf{x}}) \quad \forall \mathbf{x} \in \Omega$$

(3.52a) $\implies$ (3.52b): This inequality is true whenever $\hat{\mathbf{x}}$ is a solution to (3.52a).

(3.52b) $\implies$ (3.52a): Given the last inequality, the minimum of the right hand side is attained at $\hat{\mathbf{x}}$ or at $t = 0$ if we substitute $\mathbf{x}$ by $(1 - t)\hat{\mathbf{x}} + t\bar{\mathbf{x}}$; this is always possible because Ω is convex. Observing $\mathbf{x} - \hat{\mathbf{x}} = t(\bar{\mathbf{x}} - \hat{\mathbf{x}})$, we define the auxiliary function

$$\Gamma(t) := 2t\,\boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}(\bar{\mathbf{x}} - \hat{\mathbf{x}}) + t^2\,(\bar{\mathbf{x}} - \hat{\mathbf{x}})^\mathsf{T}(\bar{\mathbf{x}} - \hat{\mathbf{x}}).$$

The minimum of $\Gamma(t)$ at $t = 0$ requires $\Gamma'(0) \geqq 0$ for any $\bar{\mathbf{x}} \in \Omega$ or

$$\Gamma'(0) = 2\boldsymbol{\gamma}(\hat{\mathbf{x}})^\mathsf{T}(\bar{\mathbf{x}} - \hat{\mathbf{x}}) \geqq 0 \quad \forall \bar{\mathbf{x}} \in \Omega$$

and this is (3.52a). ■

To get an idea of the statement in Theorem 3.27, suppose $\Omega = \mathbb{R}^n$ and $\nabla f = \boldsymbol{\gamma}$. That is, we modify $\hat{\mathbf{x}}$ in the interior of Ω by $\mathbf{d} = -\boldsymbol{\gamma}(\hat{\mathbf{x}})$, where $-\boldsymbol{\gamma}(\hat{\mathbf{x}})$ has the direction of the steepest descent at $\hat{\mathbf{x}}$: $f(\hat{\mathbf{x}}) \geqq f(\hat{\mathbf{x}} + \mathbf{d})$. The minimum has been found when $\boldsymbol{\gamma}(\hat{\mathbf{x}}) = \mathbf{0}$ so that no further change of $\hat{\mathbf{x}}$ implies a decrease of f.

[38] This theorem is a special case of Patriksson (1994), Theorem 3.8.

When the feasible set is restricted to $\Omega = \mathbb{R}^n_+$ (note that $\boldsymbol{\gamma}(\mathbf{x}) = \mathbf{0}$ may be inadmissible), the fixed point problem (3.50) with $\mathbf{z} = \hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}})$ is to

$$\text{find an } \hat{\mathbf{x}} \geqq \mathbf{0} \text{ such that } \hat{\mathbf{x}} = \arg\min \left\{ |\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}}) - \mathbf{x}| \mid \mathbf{x} \geqq \mathbf{0} \right\}. \tag{3.53}$$

The explicit form of the objective function

$$|\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}}) - \mathbf{x}| = \left[\sum_{j=1}^{n} (\hat{x}_j - \gamma_j(\hat{\mathbf{x}}) - x_j)^2 \right]^{1/2}$$

requires for an optimal choice of x_j

$$\hat{x}_j = \begin{cases} \hat{x}_j - \gamma_j(\hat{\mathbf{x}}) & \text{if } \hat{x}_j - \gamma_j(\hat{\mathbf{x}}) > 0, \\ 0 & \text{if } \hat{x}_j - \gamma_j(\hat{\mathbf{x}}) \leqq 0. \end{cases}$$

The first case is a contradiction if $\gamma_j(\hat{\mathbf{x}}) \neq 0$ so that $\hat{x}_j > 0 \implies \gamma_j(\hat{\mathbf{x}}) = 0$. The second case needs $\gamma_j(\hat{\mathbf{x}}) \geqq \hat{x}_j = 0$ so that $\gamma_j(\hat{\mathbf{x}}) > 0 \implies \hat{x}_j = 0$. Both arguments yield the nonlinear complementary problem (3.49) and this is equivalent to the variational inequality problem (3.43) by the assumption $\Omega = \mathbb{R}^n_+$. In terms of the fixed point problem (3.50) the two cases can be summarized by

$$\hat{x}_j = k_j(\hat{\mathbf{x}}) = \max \left\{ \hat{x}_j - \gamma_j(\hat{\mathbf{x}}), 0 \right\} \qquad j = 1, \ldots, n.$$

Last but not least the mapping $P_\Omega \colon \Omega \to \Omega$ allows a fairly simple statement on the existence of a fixed point; see Kinderlehrer, Stampacchia (1980, Theorem 3.1):

3.28 Theorem (Existence) *Let Ω be a non-empty compact convex subset in $\mathbb{R}^n$. If the function $\boldsymbol{\gamma}$ is continuous, then the variational inequality problem* (3.43) *has a solution $\hat{\mathbf{x}}$.*

Proof: From the definition of $P_\Omega \colon \Omega \to \Omega$ the continuity of P_Ω follows from the assumption that $\boldsymbol{\gamma}$ is continuous and that the minimum function preserves that property. Then Brouwer's fixed point theorem guarantees the existence of a fixed point $\hat{\mathbf{x}}$ with respect to the compact set Ω. Theorem 3.27 states in turn that this fixed point solves (3.43). ■

While (3.49) refers to a vector valued mapping $\mathbf{k} \colon \Omega \to \mathbb{R}^n$ the next characterization of optimal solutions to (3.43) refers to a more general correspondence or set valued mapping $\widehat{X} \colon \Omega \to \mathfrak{P}(\mathbb{R}^n)$. A fixed point is now given by $\hat{\mathbf{x}} \in \widehat{X}(\hat{\mathbf{x}})$ rather than $\hat{\mathbf{x}} = \mathbf{k}(\hat{\mathbf{x}})$.

3.29 Theorem[39] *The point $\hat{\mathbf{x}}$ is an optimal solution to the variational inequality problem* (3.43) *if and only if $\hat{\mathbf{x}} \in \widehat{X}(\hat{\mathbf{x}})$, where $\widehat{X}(\hat{\mathbf{x}})$ is the minimum set of*

$$\min \left\{ \boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}} (\mathbf{x} - \hat{\mathbf{x}}) \mid \mathbf{x} \in \Omega \right\}. \tag{3.54}$$

3.30 Summary The preceding sections introduced the following list of programs:

[39] See Patriksson (1994), Theorem 3.9.

(3.42) Find an $\hat{\mathbf{x}} \in \Omega$ such that $\hat{\mathbf{x}} = \arg\min\{f(\mathbf{x}) | \mathbf{x} \in \Omega\}$.
(2.11) Find an $\hat{\mathbf{x}} \in \Omega$ such that $\mathbf{y}^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0$ for all $\mathbf{x} \in \Omega$, where $\mathbf{y} \in \partial f(\hat{\mathbf{x}})$.
(2.12) Find an $\hat{\mathbf{x}} \in \Omega$ such that $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0$ for all $\mathbf{x} \in \Omega$.
(3.43) Find an $\hat{\mathbf{x}} \in \Omega$ such that $\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0$ for all $\mathbf{x} \in \Omega$.
(3.49) Find an $\hat{\mathbf{x}} \geqq \mathbf{0}$ such that $\boldsymbol{\gamma}(\hat{\mathbf{x}}) \geqq \mathbf{0}$ and $\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}\mathbf{x}$ for all $\mathbf{x} \geqq \mathbf{0}$.
(3.50) Find an $\hat{\mathbf{x}} \geqq \mathbf{0}$ such that $\hat{\mathbf{x}} = \arg\min\{|\hat{\mathbf{x}} - \boldsymbol{\gamma}(\hat{\mathbf{x}}) - \mathbf{x}| \, | \, \mathbf{x} \geqq \mathbf{0}\}$.
(3.54) Find an $\hat{\mathbf{x}} \in \Omega$ such that $\hat{\mathbf{x}} = \arg\min\{\boldsymbol{\gamma}(\hat{\mathbf{x}})^{\mathsf{T}}(\mathbf{x} - \hat{\mathbf{x}}) | \mathbf{x} \in \Omega\}$.

Figure 3.11 on the next page gives a brief overview on some of the relationships between the above mentioned problems. □

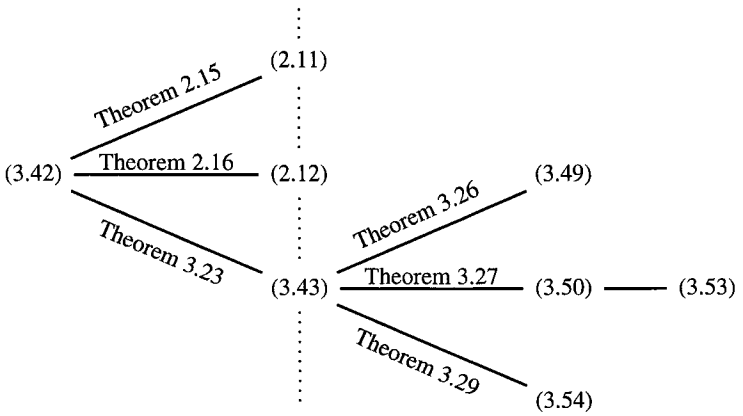


Fig. 3.11 Overview of introduced problems

3.2.5 Congested Networks

3.2.5.1 System Optimality

(a) The Convex Programming Approach

Before the results of the previous section can be applied to finding a user optimal trip assignment with flow dependent link costs, we go one step back to discuss the corresponding problem of system optimality. Here we focus on the simple case where the convex function $f(\boldsymbol{\varphi}) = \mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi}$ has the gradient $\nabla f = \mathbf{c}$ such that the vector of marginal link costs $\mathbf{c}$ is integrable and Theorem 3.23 applies. The (strict) convexity of f is then equivalent to the (strict) monotonicity of $\mathbf{c}$ so that we can give a statement on the uniqueness of a solution; see Theorem 3.19. Regarding user optimality we shall deal with link costs $\mathbf{c}$ instead of marginal link costs $\mathbf{c}$. In this case $\mathbf{c}$ needs not be integrable so that no equivalent convex program exists. Nevertheless an appropriately

defined variational inequality program is capable of recording the conditions of a Wardrop equilibrium.

Apart from these mathematical arguments, system optimality constitutes a reference case which can be compared with all alternative traffic assignments. When there is no ruling instance particularly user optimal solutions occur. They differ from system optimal traffic assignments because trip-makers ignore the external effects they impose on the rest of the network. Again the reason can be reduced to the fact that travelers select their routes on the basis of link costs $\mathbf{c}$ instead of marginal link costs $\mathbf{c}$, which would be system optimal; cf. Dafermos, Sparrow (1969). In the tradition of Pigou (1920) we are then interested in appropriate incentives which make network users behave in a system optimal way.

In contrast to the linear programming case, the travel time or travel cost is now allowed to depend on traffic flows. Because of traffic congestion the link travel times increase when the load of traffic is getting heavier. Depending on the characteristics of the real-world traffic situation modeled and the purpose of the model, they may also include a variety of other model components (Larsson, Patriksson (1999)). Although explicit capacity constraints of the form $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$ are ignored, we allow for implicit capacity restrictions in the sense of Section 3.2.3.1.

Since 1952 when Wardrop presented optimality principles for assigning traffic to a road network numerous authors have reexamined the above problem with special regard to capacitated networks. Among the first were Beckmann, McGuire (1953) whose concepts can be found in the pioneering work of Beckmann et al. (1956). For a comprehensive list of references see Patriksson (1994). The approach presented here is similar to that of Bell, Iida (1997). We analyze the problem of cost minimization (3.22) under the additional aspect of flow dependent link costs per unit or per user. It is important to notice that $\mathbf{c}(\boldsymbol{\varphi})$ indicates that every link cost function c_e may be influenced by all link flows $\boldsymbol{\varphi}$. One reason for this assumption comes from the existence of intersections where the flow of one link disturbs the flows of other links. Another reason for interaction between traffic on different links is that of two-way streets which are modeled as two anti-parallel arcs.

The new cost minimization problem[40] requests to realize a given trip table $\mathbf{x}^D$ at minimum transportation cost, where capacity constraints $\boldsymbol{\kappa}$ are ignored.

$$C(\mathbf{x}^D) = \min \left\{ \mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}} \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in \Phi^D \right\} \tag{3.55}$$

Similar to the linear program (3.23), this problem is given in terms of arc flows $\boldsymbol{\varphi}$. The same steps as before can be used to eliminate $\boldsymbol{\varphi}$, that is

[40] This problem corresponds to the cost minimization in Beckmann et al. (1956, p. 91). Notice that the authors use a different notation to describe the restrictions of the network (no route flows $\mathbf{h}$) and that their network is an undirected graph (all road can be used in both directions).

$$\varphi_{ij} = \sum_t [\varphi_{ij}^{(t)} - \varphi_{ji}^{(t)}], \quad x_{(s,t)}^D = \sum_j [\varphi_{sj}^{(t)} - \varphi_{js}^{(t)}],$$

where $\varphi_{ij}^{(t)}$ is the link flow from node i to j with destination t.

$$\begin{aligned} \boldsymbol{\varphi} &= \boldsymbol{\Delta}\,\mathbf{h} && \boldsymbol{\Delta} = \text{arc/path incidence matrix,} \\ \tilde{\mathbf{c}}(\mathbf{h}) &= \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h}) && \tilde{\mathbf{c}} = \text{path costs.} \end{aligned}$$

3.31 Remark An example with explicit channel capacities in the link cost functions $c_e(\varphi_e, \kappa_e)$ is discussed in Bertsekas, Gallager (1992, p. 436). The problem is to find a cost minimal routing in data networks.

$$\min\left\{\mathbf{c}(\boldsymbol{\varphi}, \boldsymbol{\kappa})^{\mathsf{T}}\boldsymbol{\varphi} \mid \boldsymbol{\varphi} \in \Phi^D\right\}$$

The feasible set Φ^D includes the demand for traffic $\mathbf{x}^D$, where $x^D_{(s,t)}$ is the arrival rate of traffic entering the network at node v_s and destined for node v_t. □

In accordance with (3.24) the problem (3.55) takes the form

$$C(\mathbf{x}^D) = \min\left\{\tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\mathbf{h} \mid \mathbf{h} \in H^D\right\}. \tag{3.56}$$

The remarks on (3.24), p. 105, suggest that a unique solution to (3.56) implies a unique solution for (3.55), namely, $\hat{\mathbf{h}} = \boldsymbol{\Delta}\,\hat{\boldsymbol{\varphi}}$. The inverse, however, is not true. Indeed, any solution to (3.55) can be compatible with multiple solutions of (3.56).

The analysis starts with some remarks on the new elements in the above given problems. Let f and $\tilde{f}$ be the objective functions of (3.55) and (3.56) with the effective domains Φ^D and H^D, respectively. Similarly, $\mathbf{c}$ and $\tilde{\mathbf{c}}$ denote the *marginal link costs* and the *marginal route costs*, that is

$$\begin{aligned} \nabla f(\boldsymbol{\varphi}) &= \mathbf{c}(\boldsymbol{\varphi}) = \mathbf{c}(\boldsymbol{\varphi}) + \nabla\mathbf{c}(\boldsymbol{\varphi})\boldsymbol{\varphi} && \forall\,\boldsymbol{\varphi} \\ \text{and}\quad \nabla\tilde{f}(\mathbf{h}) &= \tilde{\mathbf{c}}(\mathbf{h}) = \tilde{\mathbf{c}}(\mathbf{h}) + \nabla\tilde{\mathbf{c}}(\mathbf{h})\mathbf{h} && \forall\,\mathbf{h}, \end{aligned}$$

where $\nabla\mathbf{c}$ (and analogously $\nabla\tilde{\mathbf{c}}$) is the collection of gradients

$$\nabla\mathbf{c} = (\nabla c_1, \ldots, \nabla c_n) = \begin{pmatrix} \frac{\partial c_1}{\partial \varphi_1} & \cdots & \frac{\partial c_n}{\partial \varphi_1} \\ \vdots & & \vdots \\ \frac{\partial c_1}{\partial \varphi_n} & \cdots & \frac{\partial c_n}{\partial \varphi_n} \end{pmatrix},$$

which corresponds to the transposed Jacobian matrix $\mathbf{J}_{\mathbf{c}}$. It is useful to bear the following relationships in mind when comparing results in terms of route flows with those expressed by link flows.

$$\begin{aligned} \boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}: \quad f(\boldsymbol{\Delta}\,\mathbf{h}) &= \mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h})^{\mathsf{T}}\boldsymbol{\Delta}\,\mathbf{h} = [\boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h})]^{\mathsf{T}}\mathbf{h} = \tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\mathbf{h} =: \tilde{f}(\mathbf{h}) \\ \nabla\tilde{f}(\mathbf{h}) &= \boldsymbol{\Delta}^{\mathsf{T}}\nabla f(\boldsymbol{\Delta}\,\mathbf{h}) \quad \text{or} \quad \tilde{\mathbf{c}}(\mathbf{h}) = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h}) \\ \nabla\tilde{\mathbf{c}}(\mathbf{h}) &= \boldsymbol{\Delta}^{\mathsf{T}}\nabla\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h})\boldsymbol{\Delta} \end{aligned}$$

By Theorem 2.14, the two functions f and $\tilde{f}$ are convex if and only if the corresponding Hessean matrices $\mathbf{H}_f = \nabla\mathbf{c}^{\mathsf{T}}$ and $\mathbf{H}_{\tilde{f}} = \nabla\tilde{\mathbf{c}}^{\mathsf{T}}$ are positive semidefinite. That is, $\mathbf{y}^{\mathsf{T}}\,\mathbf{H}_f(\boldsymbol{\varphi})\,\mathbf{y} \geqq 0$ for all $\mathbf{y}$ at each $\boldsymbol{\varphi} \in \Phi^D$ with respect to the Hessean matrix of f. One special case results from

3.32 Assumption (Separable Cost Functions) *The link cost functions* $\mathbf{c}$ *are separable such that every component* c_e, $e \in \mathcal{E}$, *is independent of the flows on any other link in the network save* φ_e.

This assumption corresponds to the technical time-congestion relationsships that have been introduced on page 113. The travel time increases as the link flow φ_e converges to the respective capacity κ_e. Regarding the marginal link costs, we now gain

$$\mathfrak{c}_e(\varphi_e) = c_e(\varphi_e) + \frac{\mathrm{d}c_e(\varphi_e)}{\mathrm{d}\varphi_e}\,\varphi_e$$

so that

$$\mathbf{H}_f = \mathbf{J}_{\mathfrak{c}} = \nabla \mathbf{c}^{\mathsf{T}} = \begin{pmatrix} \frac{\partial \mathfrak{c}_1}{\partial \varphi_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial \mathfrak{c}_n}{\partial \varphi_n} \end{pmatrix}$$

is positive semidefinite if each link cost c_e, $e \in \mathcal{E}$, is a non-decreasing function of φ_e. This ensures f to be convex so that (3.55) is a convex program. From the remarks on Theorem 3.20 we know that if $\mathbf{H}_f$ is positive definite, then f is strictly convex. This case will be used in Proposition 3.36 to verify a unique optimal link flow vector.

Assumption 3.32 is unreasonable when applied to route costs because different routes may have links in common. With that $\mathbf{H}_{\tilde{f}}$ is in general not positive semidefinite and, therefore, $\tilde{f}$ needs not be convex. Fortunately, we can prove the convexity of $\tilde{f}$ when f is convex.

3.33 Proposition (Convexity) *For each convex function* f, *the function* $\tilde{f}$ *defined by* $\tilde{f}(\mathbf{h}) := f(\mathbf{\Delta}\,\mathbf{h})$ *is convex. Furthermore, for each convex function* $\tilde{f}$, *the function* C *defined by* $C(\mathbf{x}^D) := \inf\{\tilde{f}(\mathbf{h}) |\ \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D\}$ *is convex, see* (3.56).

This proposition is an immediate consequence of Rockafellar (1972, Theorem 5.7). The problem (3.56) can now be solved in terms of a Lagrangean function with the Lagrange multipliers $\boldsymbol{\pi}$.

$$\mathcal{L}(\mathbf{h}, \boldsymbol{\pi}) = \tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\mathbf{h} + \boldsymbol{\pi}^{\mathsf{T}}(\mathbf{x}^D - \mathbf{\Lambda}\,\mathbf{h}) \tag{3.57}$$

Let $\mathcal{L}$ be convex in $\mathbf{h}$ (Proposition 3.33), concave in $\boldsymbol{\pi}$, and differentiable in both $\mathbf{h}$ and $\boldsymbol{\pi}$. Then, the Karush-Kuhn-Tucker (KKT) conditions for $(\hat{\mathbf{h}}, \hat{\boldsymbol{\pi}})$ to be a saddle point of $\mathcal{L}$ are

$$\begin{aligned} &\nabla_{\mathbf{h}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\boldsymbol{\pi}}) \geqq \mathbf{0}, \quad \hat{\mathbf{h}}^{\mathsf{T}}\nabla_{\mathbf{h}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\boldsymbol{\pi}}) = 0, \quad \hat{\mathbf{h}} \geqq \mathbf{0}, \\ &\nabla_{\boldsymbol{\pi}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\boldsymbol{\pi}}) = \mathbf{0}, \qquad\qquad\qquad\qquad \hat{\boldsymbol{\pi}} \text{ unrestricted in sign.} \end{aligned}$$

The partial derivative with respect to $\mathbf{h}$ follows from

$$\nabla_{\mathbf{h}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\boldsymbol{\pi}}) = \tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \nabla\tilde{\mathbf{c}}(\hat{\mathbf{h}})\hat{\mathbf{h}} - \mathbf{\Lambda}^{\mathsf{T}}\hat{\boldsymbol{\pi}}.$$

The optimal value in (3.56) is assumed to be finite. Thus the cost minimization program has a unique KKT vector $\hat{\boldsymbol{\pi}}$ if and only if the function C is differentiable at $\mathbf{x}^D$, in which case $\hat{\boldsymbol{\pi}}$ is given by the envelope theorem,

$$\hat{\pi}_{(s,t)} = \frac{\partial C(\mathbf{x}^D)}{\partial x^D_{(s,t)}} \equiv \pi_{(s,t)}(\mathbf{x}^D). \tag{3.58}$$

Notice that each component $\hat{\pi}_{(s,t)}$ of the KKT vector is a function of the trip table $\mathbf{x}^D$. It has the dimension

$$[\pi_{(s,t)}] = \frac{€}{\text{trip from } v_s \text{ to } v_t}.$$

Choose a route $\rho \in \mathcal{R}(s,t)$ and note again $\hat{\pi}_{(s,t)} = \mathbf{\Lambda}_\rho{}^\top \hat{\boldsymbol{\pi}}$ by (3.26). Also let $\mathcal{R}$ be the set of all routes available in the network. Then, the KKT conditions yield

$$\begin{aligned} & \hat{h}_\rho \left[\tilde{c}_\rho(\hat{\mathbf{h}}) + \sum_{r \in \mathcal{R}} \frac{\partial \tilde{c}_r(\hat{\mathbf{h}})}{\partial h_\rho} \hat{h}_r - \hat{\pi}_{(s,t)} \right] = 0 \qquad \forall \rho \in \mathcal{R}(s,t) \\ \iff \quad & \hat{h}_\rho \left[\tilde{\mathfrak{c}}_\rho(\hat{\mathbf{h}}) - \hat{\pi}_{(s,t)} \right] = 0 \qquad \forall \rho \in \mathcal{R}(s,t). \end{aligned} \tag{3.59}$$

Although at first glance the Lagrange multipliers $\hat{\boldsymbol{\pi}}$ are unrestricted in sign, we know that $x^D_{(s,t)} > 0$ requires $\hat{h}_\rho > 0$ for at least one route $\rho \in \mathcal{R}(s,t)$. Thus $\hat{\pi}_{(s,t)}$ is non-negative for any optimal solution with non-negative and non-decreasing route costs. Wardrop's second principle can now be stated as follows: (a) all routes in use between an OD-pair have equal marginal costs and (b) no unused route between that OD-pair has lower marginal cost.

$$\hat{h}_\rho > 0 \implies \tilde{\mathfrak{c}}_\rho(\hat{\mathbf{h}}) = \hat{\pi}_{(s,t)}, \tag{3.60a}$$

$$\hat{h}_\rho = 0 \implies \tilde{\mathfrak{c}}_\rho(\hat{\mathbf{h}}) \geqq \hat{\pi}_{(s,t)}. \tag{3.60b}$$

In equation (3.59) the sum results from the fact that the total costs on all routes are calculated by $h_\rho \tilde{c}_\rho(\mathbf{h})$. If we assume that the partial derivatives vanish, (3.59) equals (3.29). In general, however, the sum is expected to be positive. The reason is that increasing one route flow, say h_{ρ_1}, increases not only the route costs $\tilde{c}_{\rho_1}$ but also other route costs. Take, for instance, a second route ρ_2 which shares one arc e with route ρ_1 and recall that link costs c_e are increasing functions of link flows φ_e. The bracketed terms can now be interpreted as follows. The Lagrange multiplier $\hat{\pi}_{(s,t)}$ represents the *social costs* caused by the last trip from v_s to v_t to the network as a whole. The *private costs* $\tilde{c}_\rho(\hat{\mathbf{h}})$ are paid by the trip-maker. Finally, the monetary equivalent of the *external effects* incurred to the rest of the network is given by the remaining sum. A positive value is conveniently taken as an expression for *congestion*. Analogous statements result for link flows. The addition of another vehicle on the road will slow down all existing vehicles and so increase the costs of all other road users, cf. Walters (1987). The corresponding observation with respect to physical terms goes back at least to Pigou (1920): *The source of the general divergences between the values of marginal social and marginal private net product [...] is the fact that [...] a part of the product of a unit of resources consists of something, which, instead of coming in the first instance to the person who invests the unit, comes instead, in the first instance [...], as a positive or negative item, to other people.*

An important observation on marginal route costs follows from the fact that routes are sequences of links. Restating marginal route costs yields

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}) = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\hat{\mathbf{h}}) \iff \tilde{c}_\rho(\hat{\mathbf{h}}) = \sum_{e\in\rho} c_e(\boldsymbol{\Delta}\,\hat{\mathbf{h}}) \quad \forall\rho.$$

The marginal cost of an additional vehicle using route ρ corresponds to the sum of marginal link costs included by the route. Any pricing mechanism can take advantage of this result because there is no need to have full information on route choice. The outcomes of road pricing will show that it is enough to alter travel costs on the link level in order to redirect traffic flows in a favorable way.

Regarding a user optimum which will be discussed in detail in the subsequent sections, the typical behavior of trip-makers can be described as follows. Each individual assumes that its trip has no effect on the route costs of its own route and that the costs of all other routes remain unchanged. As all drivers behave in this way, they ignore not only the immediate effect on their own route, that is $\frac{\partial\tilde{c}_\rho(\hat{\mathbf{h}})}{\partial h_\rho}\,\hat{h}_\rho$, but also the indirect effects imposed on all other routes. This behavior suggests that system optima differ from user optima. The subsequent sections confirm this conjecture.

3.34 Remark (Time varying demand) The traffic assignment problem (3.56) was formulated in terms of a fixed trip table $\mathbf{x}^D$ which is valid, for instance, during a busy hour. In off-peak periods reduced trip tables appear and, therefore, other traffic assignments would be optimal. In telecommunications networks, for instance, an enormous effort would be needed to modify the routing whenever the traffic demand changes so that it is useful to think of the following heuristic. Suppose $\hat{\mathbf{h}}$ is an optimal solution regarding the peak demand $\mathbf{x}^D$. Then, it is natural to work with *fractions of flows*

$$\tilde{\theta}_\rho = \frac{\hat{h}_\rho}{x^D_{(s,t)}} \qquad \forall\rho \in \mathcal{R}(s,t)$$

implying in vector notation $\boldsymbol{\Lambda}\,\tilde{\boldsymbol{\theta}} = \mathbf{1}$. In off-peak periods with presumably $\mathbf{x} < \mathbf{x}^D$, we then realize a path flow $\mathbf{h} < \hat{\mathbf{h}}$ according to $\tilde{\theta}_\rho\, x_{(s,t)} = h_\rho$. Of course, this scaling strategy mistreats networks with asymmetric congestion effects, for instance, at some particularly sensitive junctions. The advantage, however, is that the routing process is easy to implement. □

3.35 Remark (Capacity Restrictions) The above interpretations contain a subtle point that may be a source of some confusion. An externality is present whenever the production of trips by some agent is "directly" affected by the actions of another trip-maker in the network, see, e.g., Mas-Colell et al. (1995). This is the case when $\partial\tilde{c}_\rho(\hat{\mathbf{h}})/\partial h_\rho$ reflects, for example, delays caused by mutual obstructions. When higher route flows induce increasing prices of capacity usage so that $\tilde{c}_\rho$ goes up, the effect is referred to as *pecuniary externality* by Viner (1931). Thus, following Scitovsky's (1954) classification, when increasing trip prices result from a traffic demand exceeding scarce capacities "at a constant level of service" (see Section 3.2.3) we have pecuniary externalities. When drivers are forced to slow down their speed because of increasing traffic loads we have a pure or *technological externality*. Today pecuniary external effects are synonymous with general market interdependence and with price taking behavior. Mas-Colell et al. (1995) show that the market is precisely the mechanism that internalizes these effects correctly.

The difference between the two phenomena becomes clear if explicit capacity restrictions $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$ with $\boldsymbol{\Delta}\,\mathbf{h} = \boldsymbol{\varphi}$ are introduced into the cost minimal trip assignment problem (3.56).[41] This changes the optimal solution (3.59) as follows. Firstly, the term $+\,\mathbf{q}^{\mathsf{T}}(\boldsymbol{\Delta}\,\mathbf{h} - \boldsymbol{\kappa})$ with the multipliers $\mathbf{q}$ must be added to the Lagrangean function (3.57). Secondly, the partial derivative with respect to $\mathbf{h}$ is to be augmented by $\boldsymbol{\Delta}^{\mathsf{T}}\hat{\mathbf{q}}$ or equivalently $\hat{\tilde{\mathbf{q}}}$, where $\hat{\mathbf{q}}$ is a cost minimal vector of shadow prices of the capacities and $\hat{\tilde{\mathbf{q}}}$ is the corresponding vector of shadow prices assigned to routes. Eventually, (3.59) turns out to be

$$\hat{h}_\rho \left[\tilde{c}_\rho(\hat{\mathbf{h}}) + \sum_{r \in \mathcal{R}} \frac{\partial \tilde{c}_r(\hat{\mathbf{h}})}{\partial h_\rho}\, \hat{h}_r + \sum_{e \in \rho} \hat{\tilde{q}}_e - \hat{\pi}_{(s,t)} \right] = 0 \qquad \forall\, \rho \in \mathcal{R}(s,t)$$

real marginal cost — marginal technical external cost — marginal pecuniary external cost — marginal social cost

For every road $\rho \in \mathcal{R}(s,t)$ in use the marginal social cost $\hat{\pi}_{(s,t)}$ is increased by the sum of shadow prices of all links involved. As usual an unsaturated link $\hat{\varphi}_e < \kappa_e$ implies $\hat{\tilde{q}}_e = 0$. If $\hat{\tilde{q}}_e > 0$, the road is working at its limit $\hat{\varphi}_e = \kappa_e$. □

(b) The Variational Inequality Approach

Regarding the objective function $\tilde{f}$ of the cost minimization problem (3.56), the marginal route costs $\tilde{\mathbf{c}} := \nabla \tilde{f}$ suggest a variational inequality problem

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall\, \mathbf{h} \in H^D. \tag{3.61}$$

By Theorem 3.23, this problem is equivalent to (3.56) having the form

$$\min \left\{ \tilde{f}(\mathbf{h}) \mid \mathbf{h} \in H^D \right\}$$

provided that $\tilde{f}$ is convex (Proposition 3.33). The Lagrangean function (3.57) with $\tilde{f}(\mathbf{h})$ instead of $\tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\mathbf{h}$ yields the same KKT conditions as before. If the marginal route cost $\tilde{c}_\rho(\hat{\mathbf{h}})$ exceeds the marginal social cost $\hat{\pi}_{(s,t)}$, system optimality requires that no flow is assigned to that route, $\hat{h}_\rho = 0$. A positive route flow $\hat{h}_\rho > 0$ implies $\tilde{c}_\rho(\hat{\mathbf{h}}) = \hat{\pi}_{(s,t)}$. This case yields also

$$\frac{\partial \tilde{f}(\hat{\mathbf{h}})}{\partial h_\rho} \equiv \tilde{c}_\rho(\hat{\mathbf{h}}) = \hat{\pi}_{(s,t)} = \frac{\partial C(\mathbf{x}^D)}{\partial x^D_{(s,t)}}$$

under the preconditions of (3.58). A marginal change of one route flow has the same effect as a marginal change of the demand for traffic on the respective OD-pair. Bear in mind that these arguments are based on marginal route costs $\tilde{c}_\rho(\hat{\mathbf{h}})$ which in general differ from route costs per trip, that is $\tilde{c}_\rho \neq \tilde{c}_\rho$.

[41] It is assumed that the set of feasible assignments is not empty despite capacity restrictions.

Before discussing (3.61) with respect to a user optimum it is useful to restate the variational inequality problem in terms of link flows. Again let f denote the (convex) objective function of (3.55). If there is an $\mathbf{h}$ with $\boldsymbol{\varphi} = \boldsymbol{\Delta}\mathbf{h}$ then $f(\boldsymbol{\varphi}) = \tilde{f}(\mathbf{h})$ by the definition of $\tilde{\mathbf{c}}$. The marginal link costs are $\mathbf{c} := \nabla f$, where the symmetry of the second-order partial derivatives implies $\partial \mathfrak{c}_i / \partial \varphi_j = \partial \mathfrak{c}_j / \partial \varphi_i$. Note again that link costs differ in general from marginal link costs, $\mathbf{c} \neq \mathbf{c}$. Using $\sigma(t) = \boldsymbol{\varphi} t$ and observing $f(\mathbf{0}) = 0$ yields

$$f(\boldsymbol{\varphi}) = \oint_{\mathbf{0}}^{\boldsymbol{\varphi}} \mathbf{c}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} = \int_0^1 \mathbf{c}(\sigma(t))^{\mathsf{T}} \dot{\sigma}(t) \mathrm{d}\, t$$

$$= \sum_{e \in \mathcal{E}} \int_0^1 \mathfrak{c}_e(\sigma(t)) \dot{\sigma}_e(t) \mathrm{d}\, t = \sum_{e \in \mathcal{E}} \int_0^1 \mathfrak{c}_e(\boldsymbol{\varphi} t) \varphi_e \mathrm{d}\, t.$$

On the Assumption 3.32 of separable link costs this function simplifies to

$$f(\boldsymbol{\varphi}) = \sum_{e \in \mathcal{E}} \int_0^1 \mathfrak{c}_e(\varphi_e t)\, \varphi_e \, \mathrm{d}\, t.$$

Resubstituting $z = \varphi_e t$ with $\mathrm{d}\, z = \varphi_e \mathrm{d}\, t$ gives a functional form as used in Bell, Iida (1997).

$$f(\boldsymbol{\varphi}) = \sum_{e \in \mathcal{E}} \int_0^{\varphi_e} \mathfrak{c}_e(z) \mathrm{d}z. \tag{3.62}$$

By Theorem 2.14, this function is convex if we assume that the marginal link cost functions are increasing functions of the respective link flow. As mentioned above, this is the case of a positive semidefinite Hessean matrix $\mathbf{H}_f$ such that $\mathbf{y}^{\mathsf{T}} \mathbf{H}_f(\boldsymbol{\varphi})\, \mathbf{y} \geqq 0$ for all $\mathbf{y}$ at every $\boldsymbol{\varphi} \in \Phi^D$.

Another equivalent criterion for the convexity of f results from the monotonicity property of the marginal link cost functions $\mathbf{c}$. By the mean value theorem, there is an $\tilde{\boldsymbol{\varphi}} = \lambda \boldsymbol{\varphi} + (1 - \lambda) \bar{\boldsymbol{\varphi}}$ for some $0 < \lambda < 1$ such that

$$f(\boldsymbol{\varphi}) = f(\bar{\boldsymbol{\varphi}}) + \nabla f(\bar{\boldsymbol{\varphi}})^{\mathsf{T}} (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}}) + \frac{1}{2} (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}})^{\mathsf{T}} \mathbf{H}_f(\tilde{\boldsymbol{\varphi}}) (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}}).$$

If $\mathbf{H}_f$ is positive semidefinite, we gain the convexity criterion of Theorem 2.13.

$$f(\boldsymbol{\varphi}) \geqq f(\bar{\boldsymbol{\varphi}}) + \nabla f(\bar{\boldsymbol{\varphi}})^{\mathsf{T}} (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}})$$

The same inequality holds good for the pair $(\bar{\boldsymbol{\varphi}}, \boldsymbol{\varphi})$ instead of $(\boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}})$. Adding the two inequalities yields $0 \geqq \nabla f(\bar{\boldsymbol{\varphi}})^{\mathsf{T}} (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}}) + \nabla f(\boldsymbol{\varphi})^{\mathsf{T}} (\bar{\boldsymbol{\varphi}} - \boldsymbol{\varphi})$. Rearranging this inequality and using $\nabla f = \mathbf{c}$ results in the *monotonicity* of $\mathbf{c}$,

$$[\mathbf{c}(\boldsymbol{\varphi}) - \mathbf{c}(\bar{\boldsymbol{\varphi}})]^{\mathsf{T}} (\boldsymbol{\varphi} - \bar{\boldsymbol{\varphi}}) \geqq 0 \qquad \forall \boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}} \tag{3.63}$$

which indicates f to be convex. Proposition 3.36 comes back to the special case of strict monotonicity.

Again the crucial step is to apply Theorem 3.23 which states that the cost minimization problem

$$\min\left\{f(\boldsymbol{\varphi})\mid \boldsymbol{\varphi}\in\Phi^D\right\} \tag{3.55}$$

and the variational inequality problem

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^{\mathsf{T}}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}})\geqq 0 \qquad \forall\,\boldsymbol{\varphi}\in\Phi^D \tag{3.64}$$

are equivalent. A similar variational inequality problem is the starting point of the next section when discussing the problem of user optimal traffic assignments. Before going on with this question let us emphasize one further result.

3.36 Proposition (Uniqueness)[42] *If $\hat{\boldsymbol{\varphi}}$ solves* (3.64) *and* $\mathbf{c}$ *is strictly monotone, then $\hat{\boldsymbol{\varphi}}$ is the only equilibrium. An optimal route flow $\hat{\mathbf{h}}$, however, is in general not unique.*

Proof: The strict monotonicity requires that the strict inequality in (3.63) holds true whenever $\boldsymbol{\varphi}\neq\bar{\boldsymbol{\varphi}}$. The rest is an immediate consequence of Theorem 3.19. The remark in the sequel of (3.24) states that there is hardly ever a unique optimal route flow $\hat{\mathbf{h}}$. Remember Theorem 3.21: f is strictly convex if and only if $\mathbf{c}$ is strictly monotone. ■

3.2.5.2 User Optimality

(a) Analogies to System Optimality

The preceding section points out that a link flow vector $\hat{\boldsymbol{\varphi}}$ minimizes total travel cost if and only if it is a solution to (3.64). This problem of system optimality is now confronted with the variational inequality problem

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^{\mathsf{T}}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}})\geqq 0 \qquad \forall\,\boldsymbol{\varphi}\in\Phi^D. \tag{3.65}$$

In accordance with Smith (1979a) it will be shown that a solution to (3.65)[43] represents a *user optimum*. The reason to describe the users' behavior by average link costs instead of marginal link costs is that network users are only concerned with the cost they must bear themselves. Each network user is simply unaware of (or at least unresponsive to) the congestion costs he imposes on others. Thus system optimality and user optimality coincide only if

$$\nabla f\equiv\mathbf{c}=\mathbf{c}. \tag{3.66}$$

Unfortunately, this relationship is not true in the presence of congestion which is commonly ignored by the drivers. Indeed, for $f(\boldsymbol{\varphi})=\mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi}$ we have

$$\nabla f(\boldsymbol{\varphi})=\mathbf{c}(\boldsymbol{\varphi})=\mathbf{c}(\boldsymbol{\varphi})+\nabla\mathbf{c}(\boldsymbol{\varphi})\boldsymbol{\varphi},$$

[42] See Smith (1979a, p. 302).

[43] A solution to (3.65) is equivalent to saying that $-\mathbf{c}(\hat{\boldsymbol{\varphi}})$ is normal to Φ^D at $\hat{\boldsymbol{\varphi}}$.

where the link costs per unit $\mathbf{c}$ are non-decreasing functions of $\boldsymbol{\varphi}$, i.e., the matrix $\nabla\mathbf{c}(\boldsymbol{\varphi})$ has only non-negative elements. Regarding $\mathbf{c}$ the next step is to derive a function which is connected to $\mathbf{c}$ just as f and $\mathbf{c}$ are related to each other. Supposing the matrix $\nabla\mathbf{c}(\boldsymbol{\varphi})$ to be symmetric the line integral

$$g(\boldsymbol{\varphi}) := \oint_{\mathbf{0}}^{\boldsymbol{\varphi}} \mathbf{c}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} = \sum_{e\in\mathcal{E}} \int_0^1 c_e(\boldsymbol{\sigma}(t))\, \mathrm{d}\sigma_e(t) \quad \text{with} \quad \nabla g = \mathbf{c}. \tag{3.67}$$

is path independent, where the paths start at $\boldsymbol{\sigma}(0) = \mathbf{0}$ and end at $\boldsymbol{\sigma}(1) = \boldsymbol{\varphi}$ (some authors prefer the equivalent formulation $g(\boldsymbol{\varphi}) = \int_0^1 \mathbf{c}(t\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi}\,\mathrm{d}t$) with $\boldsymbol{\sigma}(t) = \boldsymbol{\varphi}t$). On the Assumption 3.32 that the link cost functions $\mathbf{c}$ are separable, the above formula simplifies to a sum of usual integrals; this function has been introduced by Beckmann et al. (1956) and is widely used since then.

$$g(\boldsymbol{\varphi}) = \sum_{e\in\mathcal{E}} \int_0^{\varphi_e} c_e(z_e)\, \mathrm{d}z_e \tag{3.68}$$

Furthermore, the Hessean matrix $\mathbf{H}_g$ is positive semidefinite so that g is convex. The objective function g is frequently used with respect to travel time (or link performance) functions τ_e instead of cost functions c_e, see for instance Patriksson (1994, p. 35).

Now Theorem 3.23 states with respect to the convex function g that (3.65) is equivalent to the *traffic assignment problem*

$$G(\mathbf{x}^D) = \min\left\{g(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \Phi^D\right\}. \tag{3.69}$$

The travelers behave together as if they minimize the function g. Although in general $f \neq g$, the value of a system optimum $\boldsymbol{\varphi}^1$ never exceeds the value of a user optimum $\boldsymbol{\varphi}^2$, that is $f(\boldsymbol{\varphi}^1) \leqq f(\boldsymbol{\varphi}^2)$, by the definition of f.

3.37 Remark The difference between total transport costs f in (3.62) and the term g in (3.68) can be clarified best if the analysis is restricted to a single OD-pair which is connected by one link e. The fixed demand for traffic, the route flow, and the link flow then satisfy $x^D_{(s,t)} = \hat{h}_\rho = \hat{\varphi}_e$. The travelers regard the cost per trip $c_e(\varphi_e)$ as their marginal contribution to total cost so that *they behave together* as if they minimize $g(\hat{\varphi}_e) = \int_0^{\hat{\varphi}_e} c_e(z)\mathrm{d}z$. In doing so they ignore their individual contribution to congestion as a part of marginal social cost $\mathfrak{c}_e(\varphi_e) = c_e + \varphi_e\, \mathrm{d}c_e/\mathrm{d}\varphi_e$. That is $f(\hat{\varphi}_e) = \hat{\varphi}_e c_e(\hat{\varphi}_e) = \int_0^{\hat{\varphi}_e} \mathfrak{c}_e(z)\mathrm{d}z$. The underestimation $\hat{\varphi}_e c_e(\hat{\varphi}_e) - g(\hat{\varphi}_e)$ is indicated by the shaded area in Figure 3.12 or, equivalently, by the area between the two curves, i.e., $f(\hat{\varphi}_e) - g(\hat{\varphi}_e)$.

The example is misleading in the sense that travelers have no choice between alternative routes. This unsatisfactory setting will be corrected in Remark 3.41. Oberserve also that the cost curves are meant to be suggestive. A time-congestion relationship according to (3.33a), p. 113, would imply a flat region for $0 \leqq \varphi_e \ll \kappa_e$ such that travel costs are approximately constant in the situation of free flow. □

Before going on with the analysis let us restate the definition of a Wardrop equilibrium (see Definition 3.10) for the case where route costs depend on route

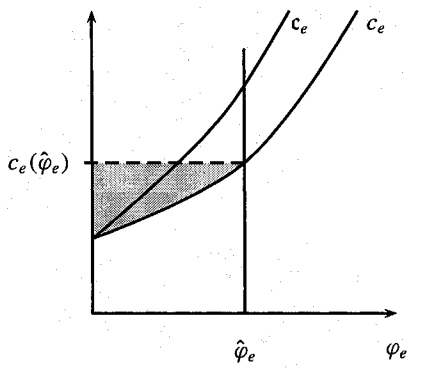


Fig. 3.12
Total cost of transportation
The network users behave as if they need not pay the shaded area, although they do in reality.

flows. The user equilibrium $\hat{\mathbf{h}}$ will be related to two optimum problems for the symmetric case (Proposition 3.39) and for the asymmetric case (Proposition 3.43)

3.38 Definition (Wardrop equilibrium) *A path flow $\hat{\mathbf{h}} \in H^D$ is called a Wardrop equilibrium when, once established, no driver has a less costly alternative route.*

$$\rho_1 \neq \rho_2,\ \rho_1 \sim \rho_2 \colon \hat{h}_{\rho_1} > 0 \implies \tilde{c}_{\rho_1}(\hat{\mathbf{h}}) \leqq \tilde{c}_{\rho_2}(\hat{\mathbf{h}})$$

As no unused path has lower cost, the Wardrop equilibrium may be restated as follows: let $p_{(s,t)}$ be the minimum travel cost on any route from v_s to v_t given the route flow $\hat{\mathbf{h}}$. For each $(v_s, v_t) \in \mathcal{V} \times \mathcal{V}$ and each $\rho \in \mathcal{R}(s,t)$ the equilibrium flow $\hat{\mathbf{h}}$ satisfies

$$\hat{h}_\rho > 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) = p_{(s,t)}, \tag{3.70a}$$

$$\hat{h}_\rho = 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) \geqq p_{(s,t)}. \tag{3.70b}$$

Bear in mind that this definition refers to a (known) minimum travel cost vector $\mathbf{p}$ which prevails at the optimal route flow vector $\hat{\mathbf{h}}$. Similar to Summary 3.8 the vector $\mathbf{p}$ follows from

$$p_{(s,t)} := \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho(\hat{\mathbf{h}}) \qquad \forall\,(s,t)\,. \tag{3.71}$$

The crucial point is that the flow vector $\hat{\mathbf{h}}$ is optimal with respect to the given trip table $\mathbf{x}^D$. Thus $\mathbf{p}$ varies when dealing with a price sensitive demand for traffic. Here the trip table $\mathbf{x}^D$ is assumed to be fixed. In the subsequent Lagrangean approach we seek for a saddle point $(\hat{\mathbf{h}}, \hat{\mathbf{p}})$ so that $\hat{\mathbf{h}}$ and $\mathbf{p}$ fit in well with each other if $\mathbf{p} = \hat{\mathbf{p}}$.

(b) Symmetric Equilibrium Problems

In what follows we are concerned with the vector of minimum travel costs $\mathbf{p}$. In Section 4.2 this vector determines the trip table $\mathbf{x}^D$ which is assumed to be given here. Instead, $\mathbf{p}$ may be seen as a vector of Lagrange multipliers $\mathbf{p}$ of an appropriate cost minimization problem. This shows in turn that $\mathbf{p}$ is a function of parameters $\mathbf{x}^D$.

For every link flow $\boldsymbol{\varphi} \in \Phi^D$ there is a route flow $\mathbf{h}$ with $\boldsymbol{\varphi} = \boldsymbol{\Delta}\mathbf{h}$. Thus it is plausible to define $\tilde{g}(\mathbf{h}) := g(\boldsymbol{\Delta}\mathbf{h})$ and, by analogy to Proposition 3.33, we know that $\tilde{g}$ is convex if g is convex. Then, (3.69) or equivalently (3.65) can be rewritten as[44]

$$G(\mathbf{x}^D) = \min\left\{\tilde{g}(\mathbf{h}) \mid \mathbf{h} \in H^D\right\}. \tag{3.72}$$

The same reasoning as with regard to system optimality results in the Lagrangean function

$$\mathcal{L}(\mathbf{h}, \mathfrak{p}) = \tilde{g}(\mathbf{h}) + \mathfrak{p}^{\mathsf{T}}(\mathbf{x}^D - \boldsymbol{\Lambda}\,\mathbf{h})$$

having similar KKT conditions as before; see (3.57). Regarding the derivatives of $\mathcal{L}$ it is useful to observe

$$\nabla\tilde{g}(\mathbf{h}) = \nabla\boldsymbol{\varphi}\nabla g(\boldsymbol{\varphi}) = \boldsymbol{\Delta}^{\mathsf{T}}\nabla g(\boldsymbol{\varphi}) = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\mathbf{h}) = \tilde{\mathbf{c}}(\mathbf{h}). \tag{3.73}$$

Using (3.73) the condition (3.59) of a system optimum has its user optimal counterpart of the form

$$\hat{h}_\rho\left[\tilde{c}_\rho(\hat{\mathbf{h}}) - \hat{\mathrm{p}}_{(s,t)}\right] = 0 \qquad \forall\,\rho \in \mathcal{R}(s,t). \tag{3.74}$$

A positive route flow $\hat{h}_\rho$ requires $\tilde{c}_\rho(\hat{\mathbf{h}}) = \hat{\mathrm{p}}_{(s,t)}$. Moreover, for $x^D_{(s,t)} > 0$ we know that $\hat{h}_\rho > 0$ for at least one route $\rho \in \mathcal{R}(s,t)$. Thus $\hat{\mathrm{p}}_{(s,t)} \geqq 0$ for any optimal solution with non-negative route costs. Every route incurring higher costs than $\hat{\mathrm{p}}_{(s,t)}$ is not used and, inversely, there is no route $\rho \in \mathcal{R}(s,t)$ available with lower cost. Both statements are reflected by the properties (3.70a) and (3.70b) of a user equilibrium, where the optimal shadow prices have the meaning of minimum travel costs as defined in (3.71).

$$\hat{\mathfrak{p}} = \mathfrak{p} \tag{3.75}$$

These results go back to Beckmann et al. (1956) and are summarized by

3.39 Proposition *An optimal traffic assignment* $\hat{\mathbf{h}}$ *for the convex program* (3.72) *determines a Wardrop equilibrium (Definition 3.38).*

Although the result (3.75) seems to be independent of the given trip table $\mathbf{x}^D$, the values of the shadow prices may change with a variation of the demand for traffic. To see this, suppose that the optimal value $G(\mathbf{x}^D)$ is finite. Then, the problem (3.72) has a unique KKT vector $\hat{\mathfrak{p}}$ if and only if the perturbation function G is differentiable at $\mathbf{x}^D$ and[45]

$$\hat{\mathrm{p}}_{(s,t)} = \frac{\partial G(\mathbf{x}^D)}{\partial x^D_{(s,t)}} \equiv \mathrm{p}_{(s,t)}(\mathbf{x}^D). \tag{3.76}$$

The two vectors of a saddle-point $(\hat{\mathbf{h}}, \hat{\mathfrak{p}})$ of $\mathcal{L}$ are mainly coordinated by the given trip table $\mathbf{x}^D$, namely, $\boldsymbol{\Lambda}\,\hat{\mathbf{h}} = \mathbf{x}^D$ and $\hat{\mathfrak{p}} = \mathfrak{p}(\mathbf{x}^D)$.

[44] On an algorithm solving the dual of that problem see Okutani (1984).

[45] Recall that, by Theorem 3.21, the gradient of a convex function is monotone, i.e., $[\mathfrak{p}(\mathbf{x}_1) - \mathfrak{p}(\mathbf{x}_2)]^{\mathsf{T}}[\mathbf{x}_1 - \mathbf{x}_2] \geqq 0$ for all $\mathbf{x}_1, \mathbf{x}_2 \geqq \mathbf{0}$.

3.40 Remark (Nonlinear Complementarity Problem) The KKT conditions of (3.72) suggest to look at the user equilibrium in view of a nonlinear complementarity problem (see Definition 3.26).

$$\begin{pmatrix}\hat{\mathbf{h}}\\ \hat{\mathbf{p}}\end{pmatrix} \geqq \mathbf{0},\ \begin{pmatrix}\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}}\\ \mathbf{\Lambda}\,\hat{\mathbf{h}} - \mathbf{x}^{D}\end{pmatrix} \geqq \mathbf{0},\ \begin{pmatrix}\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}}\\ \mathbf{\Lambda}\,\hat{\mathbf{h}} - \mathbf{x}^{D}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\hat{\mathbf{h}}\\ \hat{\mathbf{p}}\end{pmatrix} = 0,$$

where $\hat{\mathbf{p}}$ denotes the minimum route costs in accordance with (3.71). Aashtiani (1979) proves that this nonlinear complementarity problem is equivalent to the Wardrop equilibrium if and only if $\tilde{\mathbf{c}} > \mathbf{0}$ and $\mathbf{x}^D \geqq \mathbf{0}$ is fixed. These conditions exclude the case $\sum_{\rho\in\mathcal{R}(s,t)}\hat{h}_\rho > x^D_{(s,t)}$ because then $\hat{p}_{(s,t)} = 0$. With that $\tilde{c}_\rho(\hat{\mathbf{h}}) - \hat{p}_{(s,t)} > 0$ implies $\hat{h}_\rho = 0$ for every $\rho \in \mathcal{R}(s,t)$ contradicting the assumption that the route flows exceed the demand for some OD-pair. Hence, $\mathbf{\Lambda}\,\hat{\mathbf{h}} = \mathbf{x}^D$ in any optimum. The rest follows from Theorem 3.26. Furthermore, Lo, Chen (2000) introduce a so-called non-negative gap function which attains the value 0 if and only its argument solves the preceding nonlinear complementarity problem. In doing so they transform the traffic assignment problem into a continuously differentiable unconstrained minimum problem (Proposition 5) permitting a large number of optimization algorithms for its solution. For example, a gradient method is applied, which searches the minimum in the direction of the negative gradient of the gap function. Furthermore, Lo, Szeto (2002) discuss the nonlinear complementarity problem in a dynamic context. The idea is to restate the problem for each period, that is each departure time, and then to model the overlapping flows. The paper is based upon the conditions of an ideal dynamic user optimum taken from Ran, Boyce (1996): for each OD-pair at each instant of time the actual travel times experienced by the travelers departing at the same time are equal and minimal. □

3.41 Remark (Road Pricing) Pigou (1920) suggested that *the deficiency of the private, as compared with the social, net product [...] can be mitigated by compensation schemes.* This statement can be illustrated by the example in Remark 3.37. At first we need a more realistic setting with at least two links which can be substituted against each other. In this case one can think of link tolls for instance $\mathtt{t}_e = \hat{\varphi}_e \mathrm{d}c_e(\hat{\varphi}_e)/\mathrm{d}\varphi_e$ which induce the travelers to adjust their behavior. This last effect is not represented by Figure 3.12 because there is no route choice so that tripmakers cannot change their behavior. Introducing a second parallel link[46] yields a situation which is shown in the following picture; see Akcelik (1979), Boyce (1980), or Gartner (1980). Notice that the first road is overused (congested) in the user equilibrium UE because the marginal link cost exceeds that of the second link.

The reader may verify that the gray triangle indicates the travel costs the users can avoid in common if they switch from user equilibrium to system optimum. The disadvantage is that in the system optimum the costs per trip c_1 and c_2 fall apart.

[46] Pigou (1920) introduced the idea of a congestion toll by virtue of the famous two-road example. He postulates that one road is wide enough to accommodate the whole traffic without congestion, but is also rough and slow whereas the other is narrow and likely to be congested, but smooth and fast.

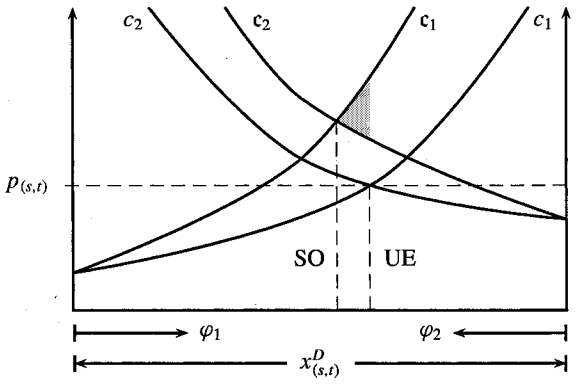


Fig. 3.13
Traffic assignments
The system optimum SO requires $c_1 = c_2$ for two nodes connected by two links. In contrast, the user equilibrium UE implies $p_{(s,t)} = c_1 = c_2$.

That is, not all users profit from this transition, what is evident because the former solution was a user equilibrium. As it cannot be expected that users of e_1 compensate travelers on link e_2 the only way to overcome this problem is to levy an appropriate toll system. In cases where private compensation schemes are not plausible Pigou (1920, Part II Chap. XI) postulated: *[...] for every industry in which the value of the marginal social net product is less than that of the marginal private net product, there will be certain rates of tax, the imposition of which by the State would increase [...] the national welfare; and one rate of tax, which would have the optimum effect.*

In the example at hand, the trip table is fixed by assumption so that any toll system induces a reassignment of traffic where φ_2 increases by the same amount as φ_1 decreases. This effect is referred to as traffic diversion. To arrive at a system optimum, the toll for a trip on link e_1 must exceed that of e_2 by $c_2 - c_1$. Figure 3.13 suggests at least two different solutions.[47]

$$c_1 + \mathfrak{t}_1^a = \mathfrak{c}_1 = \mathfrak{c}_2 = c_2 + \mathfrak{t}_2^a \qquad \mathfrak{t}_1^a, \mathfrak{t}_2^a > 0$$
$$c_1 + \mathfrak{t}_1^b = c_2 \qquad \mathfrak{t}_1^b > 0,\ \mathfrak{t}_2^b = 0$$

This observation of a non-unique toll system goes back to Dafermos, Sparrow (1971). As the number of trips $x^D_{(s,t)}$ is fixed, the users pay $\mathfrak{c}_1\varphi_1 + \mathfrak{c}_2\varphi_2$ in the first case of *marginal cost pricing*.[48] This amount exceeds the initial travel cost by $\mathfrak{t}_1^a\varphi_1 + \mathfrak{t}_2^a\varphi_2$, which may be seen as toll revenue of the levying network provider. The revenue is much smaller but nevertheless positive in the second case, i.e., $\mathfrak{t}_1^b\varphi_1$. Bergendorff et al. (1997) define a non-negative vector $\mathbf{t}$ to be "valid" if the set of tolled user equilibrium solutions is a subset of untolled system optimal solutions. In their study of valid toll vectors the authors demand that travelers only have to pay the true total travel costs. This requires a *toll/subsidy schedule* ($\mathfrak{t}_1^c > 0$, $\mathfrak{t}_2^c < 0$) such

[47] The toll systems $(\mathfrak{t}_1^a, \mathfrak{t}_2^a)$ and $(\mathfrak{t}_2^a, \mathfrak{t}_2^b)$ are not indicated in Figure 3.13 but they can be found immediately on the dashed line representing the system optimum SO. All we have to do is to adjust c_1 and c_2 such that travelers become indifferent between both links.

[48] Walters (1961) discusses marginal cost pricing on the basis of the explicit capacity curve in Figure 3.8, p. 111, instead of expressing congestion phenomena by a link cost function.

that $\mathfrak{t}_1^c\varphi_1 + \mathfrak{t}_2^c\varphi_2 = 0$. Constraining net tolls collected to be zero or minimizing the total toll revenue are tasks discussed by Hearn, Ramana (1998). Larsson, Patriksson (1998) emphasize that the non-uniqueness of the link tolls is a positive feature of the model, since it facilitates the fulfillment of traffic management goals while also satisfying exogenous restrictions on the toll scheme. Furthermore, one could choose a toll vector which simultaneously optimizes a secondary performance function. Bear in mind that network providers need revenues to finance network infrastructure. We shall come back to this issue in the context of network design, see Chapter 5.

The results on pricing schemes in order to attain system optima change when an exogenous capacity restriction κ_2 is also imposed on the second road; see Figure 3.14. As an example think of Bonn, the former capital of Germany, where only a few bridges cross the river Rhine. Shorter routes are only available through ferries with a very limited capacity. Drivers are now faced with a completely different reason for congestion.

Adding exogenous capacity restrictions κ_1 and κ_2 to the Lagrangean function (3.57) implies for the shadow price $\mathfrak{q}_1$ of the first unsaturated road

$$\hat{\varphi}_1 < \kappa_1 \implies \hat{\mathfrak{q}}_1 = 0.$$

As both roads are needed to accommodate total demand, one has furthermore

$$\begin{aligned} \hat{\varphi}_1 = \hat{h}_1 > 0 &\implies \mathfrak{c}_1 + \hat{\mathfrak{q}}_1 = \hat{\pi}_{(s,t)} \implies \mathfrak{c}_1 = \hat{\pi}_{(s,t)} \\ \kappa_2 = \hat{\varphi}_2 = \hat{h}_2 > 0 &\implies \mathfrak{c}_2 + \hat{\mathfrak{q}}_2 = \hat{\pi}_{(s,t)} \end{aligned}$$

and, therefore, $\hat{\mathfrak{q}}_2 > 0$. The marginal cost pricing rule (3.58) with respect to a system optimum proposes for this case

$$\frac{\partial C(\mathbf{x}^D)}{\partial x^D_{(s,t)}} = \pi_{(s,t)} = \mathfrak{c}_1 = \mathfrak{c}_2 + \hat{\mathfrak{q}}_2.$$

That market solution is indicated in Figure 3.14 by the two circles on the dashed capacity curve for κ_2, where $\hat{\mathfrak{q}}_2 = \mathfrak{c}_1 - \mathfrak{c}_2$. From the users' point of view that supplementary charge on the second road is too high because drivers compare the perceived cost per trip c_1 and c_2. Every fare that exceeds $c_1 - c_2$ would be prohibitive for the ferry. Without further tolls by the government a charge $\mathfrak{t}_2 = c_1 - c_2$ makes travelers indifferent between the two roads.

Evans (1992) discusses governments as monopoly suppliers of roads and derives in fact similar results as pointed out by Knight (1924) who favors a private ownership of the resources. The traditional arguments for road congestion pricing assume that the government behaves like a benevolent dictator who does not exploit his monopoly power. Nevertheless, the above results show that any monopolist has strong incentives to raise a substantial amount of revenue. Although congestion pricing provides the correct incentives for road users, it does so at the cost of setting up perverse incentives for governments. Another argument is of similar importance, see Evans (1992) and Small (1992): suppose the government raises a positive revenue by levying tolls on road users under congested conditions. This makes the travelers

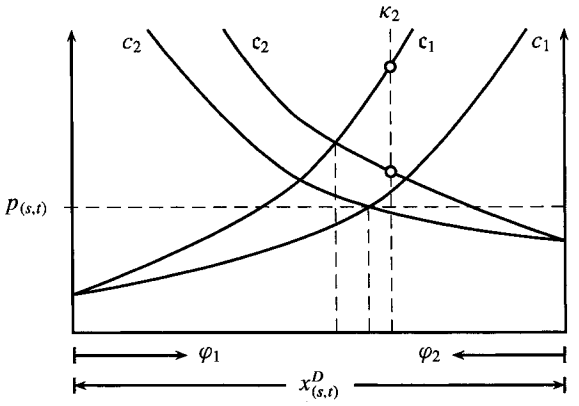


Fig. 3.14
Traffic assignments with an external capacity restriction

Neither the former system optimum nor the user equilibrium can be realized due to a limited throughput of the second road.

even worse off, perhaps to the benefit of other people. The argument is different if we assume that the revenue is used to expand the road system so that those who pay the fees benefit from reduced congestion and, therefore, from reduced tolls; see Chapter 5. Notice that it might be necessary to have disaggregate travel data in the case of different vehicle modes (cars, trucks, buses, etc.) or user classes (residents, commuters, commercial traffic, etc.) □

(c) Asymmetric Equilibrium Problems

Up until now the average travel cost vector $\mathbf{c}(\boldsymbol{\varphi})$ is a gradient of the objective function g at $\boldsymbol{\varphi}$ by definition, see (3.69) and (3.67), so that the Jacobian matrix $\mathbf{J_c} = (\nabla \mathbf{c})^\mathsf{T}$ is symmetric. It is obvious that convex programming solution techniques are no longer applicable if the Jacobian matrix is detected to be asymmetric, see Dafermos (1971). Smith (1983) gives a quite easy example for the asymmetric case so that (3.67) becomes void. Hence, let us go one step back to the variational inequality problem

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^\mathsf{T}(\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}) \geqq 0 \qquad \forall \boldsymbol{\varphi} \in \Phi^D, \tag{3.65}$$

which can be analyzed independently of the fact whether the link costs $\mathbf{c}$ are integrable. Thus, g may exist (symmetric case) or not (asymmetric case).

As network users decide on their route choice, i.e., a distinct collection of links, user optimality is usually defined in terms of route flows instead of link flows. Recalling that arc flows $\boldsymbol{\varphi}$ are related to path flows $\mathbf{h}$ by the arc/path incidence matrix $\boldsymbol{\Delta}$, the problem (3.65) can be restated by the variational inequality

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^\mathsf{T}(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall \mathbf{h} \in H^D, \tag{3.77}$$

where the demand feasible set is based on a fixed trip table $\mathbf{x}^D$. Smith (1983, and 1984a) proposes an algorithm to solve this variational inequality problem. In doing so he exploits $-\tilde{\mathbf{c}}(\mathbf{h})$ as a descent direction in the sense of (3.47). Some other algorithms are compared in Fisk, Nguyen (1982) with respect to convergence properties as well as their numerical performance.

The two preceding problems (3.65) and (3.77) are equivalent just as (3.61) and (3.64) which use marginal cost terms instead of average cost terms. Despite flow dependent link costs the variational inequality (3.77) is closely related to the linear program (3.25), p. 105, where both problems particularly refer to the same feasible set with a fixed trip table. The concrete relationship is given by

3.42 Lemma *The following two conditions are equivalent for a feasible traffic assignment* $\hat{\mathbf{h}} \in H^D$.

(i) The assignment $\hat{\mathbf{h}}$ *solves* (3.77).
(ii) $\min\{\tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \mathbf{h} \mid \mathbf{h} \in H^D\} = \tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \hat{\mathbf{h}}$

Notice that the objective function in (ii) is based on $\tilde{\mathbf{c}}(\hat{\mathbf{h}})$.

Proof: Rewriting the variational inequality (3.77) yields

$$\begin{aligned} & \tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \mathbf{h} \geqq \tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \hat{\mathbf{h}} \qquad \forall \mathbf{h} \in H^D \\ \Longleftrightarrow \quad & \min\{\tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \mathbf{h} \mid \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0}\} \geqq \tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top \hat{\mathbf{h}}. \end{aligned}$$

As $\hat{\mathbf{h}}$ is feasible by assumption, the equivalence of (i) and (ii) follows. ■

This lemma permits us to show that the traffic assignment $\hat{\mathbf{h}}$ considered is also a user equilibrium. The corresponding statement of the following theorem with respect to a price sensitive demand $\mathbf{x}^D$ will be given in Proposition 4.14.

3.43 Proposition[49] *The path flow* $\hat{\mathbf{h}}$ *is a Wardrop equilibrium if and only if* $\hat{\mathbf{h}}$ *solves* (3.77) *or, equivalently,* (3.65).

Proof: *(a) Necessity.* Let $\hat{\mathbf{h}}$ be a user equilibrium flow. Then, Definition 3.38 of a Wardrop equilibrium implies[50]

$$(3.78) \qquad (\tilde{c}_\rho(\hat{\mathbf{h}}) - p_{(s,t)})(h_\rho - \hat{h}_\rho) \geqq 0 \quad \forall \rho \in \mathcal{R}(s,t),\ \forall h_\rho \geqq 0$$

Summing up over all routes and restricting the route flow to $\mathbf{h} \in H^D$ (indeed this weakens the preceding relationship) yields

$$\begin{aligned} & \sum_\rho \big(\tilde{c}_\rho(\hat{\mathbf{h}}) - p_{(s,t)}\big)(h_\rho - \hat{h}_\rho) \geqq 0 \qquad && \forall \mathbf{h} \in H^D \\ \Longleftrightarrow \quad & \big(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^\top \mathbf{p}\big)^\top (\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 && \forall \mathbf{h} \in H^D. \end{aligned}$$

Notice that every feasible $\mathbf{h}$ and especially $\hat{\mathbf{h}}$ satisfies $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$ so that $\mathbf{p}^\top(\mathbf{\Lambda}\,\mathbf{h} - \mathbf{\Lambda}\,\hat{\mathbf{h}}) = 0$. The remaining part of the inequality is (3.77) or, equivalently, (3.65).

(b) Sufficiency. If $\hat{\mathbf{h}} \in H^D$ solves (3.77), then Lemma 3.42 applies. The minimization problem is in fact the same as the primal program (3.25.P) because the objective function includes *constant* route costs $\tilde{\mathbf{c}}(\hat{\mathbf{h}})$. Thus we can apply the

[49] This result has been presented in Dafermos (1971) and Smith (1979a) who identified the equilibrium network flow problem as a variational inequality.

[50] The same reasoning has been used in the proof of Proposition 3.12, p. 109.

necessary part in the proof of Proposition 3.12 showing that $\hat{\mathbf{h}}$ must be a Wardrop equilibrium.[51] ■

The variational inequality problem (3.78) has an equivalent formulation by Theorem 3.26, namely, the nonlinear complementary problem

$$\text{find a non-negative flow } \hat{\mathbf{h}} \geqq \mathbf{0} \text{ such that } \tilde{\mathbf{c}}(\hat{\mathbf{h}}) \geqq \mathbf{p} \text{ and } \big(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{p}\big)^{\mathsf{T}} \hat{\mathbf{h}} \geqq 0.$$

Provided the Lagrange multipliers of the symmetric equilibrium problem satisfy $\hat{\mathbf{p}} = \mathbf{p}$, the preceding complementary problem complies with the system given in (3.74). To sum up, the asymmetric equilibrium problem has no objective function as detected for the symmetric formulation. Nevertheless, optimal flows for both classes of problems can be characterized by the same system of optimum conditions.

3.44 Remark (Multiclass-User Transportation Networks) The variational inequality formulation is a powerful tool which can easily be generalized with respect to further aspects of traffic assignment. For example, Dafermos (1972) presents a network with distinct classes ($j \in \mathcal{J}$) of network users, e.g., trucks and passenger cars. All classes of travelers use the same network encoded by the incidence matrices $\mathbf{\Delta}$ and $\mathbf{\Lambda}$. Let $\boldsymbol{\varphi}_j$, $\mathbf{h}_j$, and $\mathbf{x}_j^D$ be the class specific link flow, route flow, and traffic demand[52]. Then all feasible flow patterns are restricted to $\boldsymbol{\varphi}_j = \mathbf{\Delta}\,\mathbf{h}_j$ and $\mathbf{x}_j^D = \mathbf{\Lambda}\,\mathbf{h}_j$ for all $j \in \mathcal{J}$. Assuming that each class perceives different link costs, we need class specific cost functions $\mathbf{c}_j$ which depend on link flows. In the simplest case these link flows result from adding up class specific link flows, i.e., $\boldsymbol{\varphi} = \sum_{j \in \mathcal{J}} \boldsymbol{\varphi}_j$ in which case $\mathbf{c}_j = \mathbf{c}_j(\boldsymbol{\varphi})$ as before. Summing up class specific link flows may be an oversimplification, for example, when trucks are compared to passenger cars. In this case we need cost functions of the form $\mathbf{c}_j = \mathbf{c}_j(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{|\mathcal{J}|})$, where $|\mathcal{J}|$ is the number of classes. In fact the problem (3.65) remains the same but with a lot of more notational work.

$$\sum_{j \in \mathcal{J}} \mathbf{c}_j(\hat{\boldsymbol{\varphi}}_1, \ldots, \hat{\boldsymbol{\varphi}}_{|\mathcal{J}|})^{\mathsf{T}} (\boldsymbol{\varphi}_j - \hat{\boldsymbol{\varphi}}_j) \geqq 0 \qquad \forall\, (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{|\mathcal{J}|}) \in \Phi_{\mathcal{J}}^D$$

$$\text{where} \quad \Phi_{\mathcal{J}}^D := \big\{ (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{|\mathcal{J}|}) \,|\, \boldsymbol{\varphi}_j = \mathbf{\Delta}\,\mathbf{h}_j,\ \mathbf{x}_j^D = \mathbf{\Lambda}\,\mathbf{h}_j,\ \mathbf{h}_j \geqq \mathbf{0}\ \forall\, j \in \mathcal{J} \big\}$$

□

The next step deals with conditions which ensure the user optimal traffic assignment to be unique. Similar to (3.63) it is assumed that the link costs are strictly monotone (see Definition 3.18).[53]

$$[\mathbf{c}(\boldsymbol{\varphi}_1) - \mathbf{c}(\boldsymbol{\varphi}_2)]^{\mathsf{T}} (\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2) > 0 \qquad \forall \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \quad \text{with} \quad \boldsymbol{\varphi}_1 \neq \boldsymbol{\varphi}_2$$

[51] A similar statement is given in the proof of Theorem 5 of Tobin, Friesz (1988).

[52] The traffic demand of each class of users is fixed and changes of demand between user classes are ruled out. This assumption is misleading if the aim is to model the interaction, for example, between private cars and public vehicles. See, e.g., Florian (1977) on this issue of modal split.

[53] The monotonicity condition prompts Hearn et al. (1984) to discuss the following pair of programs

If $\mathbf{c}$ is continuously differentiable, the preceding inequality is satisfied if and only if the Jacobian matrix $\mathbf{J_c}$ is positive definite at every $\varphi \in \Phi^D$ (but not necessarily symmetric).

$$\mathbf{y}^\top \mathbf{J_c}(\varphi)\,\mathbf{y} > 0 \qquad \forall \mathbf{y} \neq \mathbf{0}$$

This assumption is plausible if $\partial c_e/\partial \varphi_e \gg \partial c_e/\partial \varphi_b$ for all $b \neq e$; but uncontrolled intersections may constitute an important counterexample.

3.45 Proposition (Uniqueness)[54] *If $\hat{\varphi}$ solves* (3.65) *and* $\mathbf{c}$ *is strictly monotone, then the link flow $\hat{\varphi}$ is the only equilibrium. In general, however, there is no unique optimal route flow $\hat{\mathbf{h}}$.*

Proof: see Theorem 3.19. ■

For unique link flows on a subset of links see Fisk (1990).

3.46 Remark (Road Pricing) Let $\hat{\varphi}_s$ and $\hat{\mathbf{h}}_s$ denote a system optimal traffic assignment. A comparison with the corresponding user optimal assignment $\hat{\varphi}_u$ and $\hat{\mathbf{h}}_u$ gives

$$\mathbf{c}(\hat{\varphi}_s)^\top(\varphi - \hat{\varphi}_s) \geqq 0 \quad \forall \varphi \in \Phi^D \quad \longleftrightarrow \quad \mathbf{c}(\hat{\varphi}_u)^\top(\varphi - \hat{\varphi}_u) \geqq 0 \quad \forall \varphi \in \Phi^D;$$
$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}_s)^\top(\mathbf{h} - \hat{\mathbf{h}}_s) \geqq 0 \quad \forall \mathbf{h} \in H^D \quad \longleftrightarrow \quad \tilde{\mathbf{c}}(\hat{\mathbf{h}}_u)^\top(\mathbf{h} - \hat{\mathbf{h}}_u) \geqq 0 \quad \forall \mathbf{h} \in H^D.$$

A toll price system $\mathbf{t}$ with respect to links or $\tilde{\mathbf{t}} = \boldsymbol{\Delta}^\top \mathbf{t}$ with respect to routes modifies individual costs such that drivers behave in a system optimal way if

$$\mathbf{c}(\hat{\varphi}_s) = \mathbf{c}(\hat{\varphi}_s) + \mathbf{t} \quad \text{and} \quad \tilde{\mathbf{c}}(\hat{\mathbf{h}}_s) = \tilde{\mathbf{c}}(\hat{\mathbf{h}}_s) + \tilde{\mathbf{t}}.$$

If the network provider knows the system optimum in advance he computes $\mathbf{t} = \nabla \mathbf{c}(\hat{\varphi}_s)\hat{\varphi}_s$ or $\tilde{\mathbf{t}} = \nabla \tilde{\mathbf{c}}(\hat{\mathbf{h}}_s)\hat{\mathbf{h}}_s$ so that the toll system reflects marginal cost pricing. This provides the correct incentives for network users to realize a system optimum. As has been discussed in Remark 3.41, marginal cost pricing cannot be expected to be the only appropriate pricing scheme. Hearn, Ramana (1998) show that any toll vector $\mathbf{t}$ is adequate in a system optimizing sense if

$$\boldsymbol{\Delta}^\top\big(\mathbf{c}(\hat{\varphi}) + \mathbf{t}\big) \geqq \boldsymbol{\Lambda}^\top \hat{\mathbf{p}},$$
$$\big(\mathbf{c}(\hat{\varphi}) + \mathbf{t}\big)^\top \hat{\varphi} = \hat{\mathbf{p}}^\top \mathbf{x}^D.$$

$$\Gamma_0(\varphi_2) = \min\left\{\mathbf{c}(\varphi_1)^\top(\varphi_1 - \varphi_2) \mid \varphi_1 \in \Phi^D\right\}$$
and $$\Gamma_1(\varphi_2) = \max\left\{\mathbf{c}(\varphi_2)^\top(\varphi_2 - \varphi_1) \mid \varphi_1 \in \Phi^D\right\},$$

where they refer to Γ_1 as "gap function". Be aware that their analysis is based upon the assumption of *concave* link cost functions. The main result is $\Gamma_1(\varphi) \geqq 0 \geqq \Gamma_2(\varphi)$ for all $\varphi \in \Phi^D$ and $\min\left\{\Gamma_1(\varphi) \mid \varphi \in \Phi^D\right\} = \max\left\{\Gamma_2(\varphi) \mid \varphi \in \Phi^D\right\}$. The gap function is useful as a monitor of various algorithms, especially as a stopping criterion.

[54] See Smith (1979a, p. 302) or Dafermos (1980). Dafermos deals with the special case $\mathbf{c}(\varphi) = \mathbf{G}\varphi + \mathbf{c}^f$, where $\mathbf{G}$ is a positive definite $n \times n$ matrix and $\mathbf{c}^f$ is a vector of fixed costs per trip. The line integral then becomes quadratic $g(\varphi) = \frac{1}{2}\varphi^\top \mathbf{G}\varphi + \mathbf{c}^{f\top}\varphi$.

Both conditions may be seen as modifications of the nonlinear complementary problem in Remark 3.40 which applies also to the asymmetric equilibrium problem. Just observe

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \tilde{\mathbf{t}} - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}} \geqq \mathbf{0},$$
$$\left(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \tilde{\mathbf{t}}\right)^{\mathsf{T}}\hat{\mathbf{h}} = \hat{\mathbf{p}}^{\mathsf{T}}\mathbf{\Lambda}\,\hat{\mathbf{h}} = \hat{\mathbf{p}}^{\mathsf{T}}\mathbf{x}^{D}.$$

If needed one may restrict the toll revenue to be zero, i.e., $\mathbf{t}^{\mathsf{T}}\hat{\boldsymbol{\varphi}} = \tilde{\mathbf{t}}^{\mathsf{T}}\hat{\mathbf{h}} = 0$, so that the transport costs $\mathbf{c}(\hat{\boldsymbol{\varphi}})^{\mathsf{T}}\hat{\boldsymbol{\varphi}} = \tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\hat{\mathbf{h}}$ equal the network users' expenditures $\hat{\mathbf{p}}^{\mathsf{T}}\mathbf{x}^{D}$. □

(d) Capacity Restrictions

Remarks 3.35 and 3.41 tell us that it is worthwhile to take a look at explicit capacity restrictions. Considering limited link flows, the supply feasible set $H^S = \left\{\mathbf{h} \mid \mathbf{0} \leqq \mathbf{\Delta h} \leqq \boldsymbol{\kappa}\right\}$ must be added to the variational inequality problem (3.77).

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall \mathbf{h} \in H^D \cap H^S \tag{3.79}$$

Following Lemma 3.42, this problem can be restated as a linear program

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}\hat{\mathbf{h}} = \min\left\{\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}\mathbf{h} \mid \mathbf{h} \in H^S \cap H^D\right\},$$

which has been discussed in the sequel of (3.34), p. 114. It has the same optimum conditions as the augmented convex program (3.72). By Remark 3.35, we know that the optimum condition (3.74) must be corrected as follows

$$\hat{h}_\rho\left[\tilde{c}_\rho(\hat{\mathbf{h}}) - \hat{\mathrm{p}}_{(s,t)} + \hat{\tilde{\mathrm{q}}}_\rho\right] = 0 \qquad \forall \rho \in \mathcal{R}(s,t),$$

where $\hat{\tilde{\mathbf{q}}} = \mathbf{\Delta}^{\mathsf{T}}\hat{\mathbf{q}}$ and $\hat{\mathbf{q}}$ is the non-negative, optimal shadow price vector assigned to the capacities $\boldsymbol{\kappa}$. On the one hand, it takes no wonder that this rule of user optimality includes no external costs. They are just ignored by the users. On the other hand, as a matter of course the Wardrop conditions (3.70) can only be reproduced if travelers take into account the respective term $\hat{\tilde{\mathrm{q}}}_\rho = \sum_{e\in\rho}\hat{\mathrm{q}}_e$ as part of their trip costs. Otherwise – as criticized in Ferrari (1997) – the variational inequality problem is only equivalent to the user equilibrium conditions if capacity constraints are not active. The implementation of correct shadow prices for saturated links is a question of how to allot property rights to network participants. A network provider could force the drivers to pay $\tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \hat{\tilde{\mathbf{q}}}$ but that would not be consistent with the analysis of user optima. A more realistic allotment could be in the order of arrival, i.e., first come first served. The price of consuming capacity units at a distinct period of time may then be interpreted as the risk not to be served; it must be weighed against the cost of less comfortable routes and the cost of driving earlier or later. Such a behavior, however, cannot be discussed on the basis of the current theoretical framework.

Larsson, Patriksson (1995) consider capacity restrictions with respect to the convex program (3.69) (symmetric case) rather than the variational inequality program

(3.79) (asymmetric case). They discuss different solution methods and present an augmented Lagrangean scheme to compute the cost minimal traffic assignment as well as the generalized route costs $\hat{\mathbf{p}} = \mathbf{\Delta}^T\mathbf{c} + \mathbf{\Delta}^T\hat{\mathbf{q}} = \tilde{\mathbf{c}} + \hat{\tilde{\mathbf{q}}}$.

(e) Alternative Definitions of User Equilibria

One problem in Definition 3.38 of a Wardrop equilibrium is the comparison of route costs possibly with respect to an unknown flow pattern $\mathbf{h}^1$, that is $\tilde{c}_{\rho_1}(\mathbf{h}^1) \leqq \tilde{c}_{\rho_2}(\mathbf{h}^1)$. Without perfect knowledge of the network all travelers who want to check this condition need two trips, one on route ρ_1 and one on route ρ_2. This behavior is no problem as long as the route costs $\tilde{c}_{\rho_1}$ and $\tilde{c}_{\rho_2}$ are constant as in Definition 3.10. But for flow dependent route costs a driver on route ρ_1 cannot observe $\tilde{c}_{\rho_2}(\mathbf{h}^1)$ directly because switching to route ρ_2 changes $\mathbf{h}^1$ into $\mathbf{h}^2$. Although it is realistic to assume

$$\tilde{c}_{\rho_1}(\mathbf{h}^2) < \tilde{c}_{\rho_1}(\mathbf{h}^1) < \tilde{c}_{\rho_2}(\mathbf{h}^2) \quad \text{and} \quad \tilde{c}_{\rho_2}(\mathbf{h}^1) < \tilde{c}_{\rho_2}(\mathbf{h}^2)$$

so that the driver tends to choose route ρ_1, we cannot rule out the case $\tilde{c}_{\rho_1}(\mathbf{h}^1) > \tilde{c}_{\rho_2}(\mathbf{h}^1)$ contradicting $\mathbf{h}^1$ to be a Wardrop equilibrium. The only way to find out $\tilde{c}_{\rho_2}(\mathbf{h}^1)$ would be to ask other drivers. As this approach is not very realistic, it is helpful to introduce alternative equilibrium definitions.[55]

Let $\mathrm{d}\mathbf{h}_{\rho_1\rho_2}$ be a vector with -1 in the ρ_1^{th} place, $+1$ in the ρ_2^{th} place, and 0 elsewhere; this vector is particularly useful to describe drivers who switch from route ρ_1 to ρ_2.

3.47 Definition (User Optimized Flows) *The route flow* $\mathbf{h} \in H^D$ *is user optimized if*

$$\rho_1 \neq \rho_2,\ \rho_1 \sim \rho_2 \colon h_{\rho_1} > 0 \implies \tilde{c}_{\rho_1}(\mathbf{h}) \leqq \tilde{c}_{\rho_2}(\mathbf{h} + \epsilon\, \mathrm{d}\mathbf{h}_{\rho_1\rho_2}),$$

whenever $0 < \epsilon \leqq h_{\rho_1}$ *and* $\mathbf{h} + \epsilon\, \mathrm{d}\mathbf{h}_{\rho_1\rho_2} \in H^D$.

Given a user-optimized flow, the definition states that any driver who changes to an alternative route will experience a route cost which is at least as high as the *old* cost on his former route. Equivalently, no driver discovers a cheaper route after he has switched to a new route. Thus no traveler can improve his situation by changing his behavior. This definition is utilized in Heydecker (1986) who considers also capacity constraints in which case the demand feasible set H^D has to be substituted by $H^D \cap H^S$. Smith (1984c) shows that a user-optimized flow (Definition 3.47) is also a Wardrop equilibrium (Definition 3.38), provided the route cost functions are continuous. Moreover, he gives an example with continuous route costs where a Wardrop equilibrium route flow but no user-optimized flow exists. Some other counterexamples for the case of discontinuous link cost functions can be found in de Palma, Nesterov (1998).

An alternative concept follows from

3.48 Definition (Equilibrated Flows) *The route flow* $\mathbf{h} \in H^D$ *is equilibrated if*

$$\rho_1 \neq \rho_2,\ \rho_1 \sim \rho_2 \colon h_{\rho_1} > 0 \implies \tilde{c}_{\rho_1}(\mathbf{h} + \epsilon\, \mathrm{d}\mathbf{h}_{\rho_1\rho_2}) \leqq \tilde{c}_{\rho_2}(\mathbf{h} + \epsilon\, \mathrm{d}\mathbf{h}_{\rho_1\rho_2}),$$

[55] Cf. Heydecker (1986, p. 435 f.) or Patriksson (1994, p. 73) for more references.

whenever $0 < \epsilon \leqq h_{\rho_1}$ *and* $\mathbf{h} + \epsilon\, \mathrm{d}\mathbf{h}_{\rho_1\rho_2} \in H^D$.

Any driver who switches to an alternative route will experience a cost which is at least as great as the *new* cost on his old route. This last concept requires perfect foresight for all network users. Given the route flow $\mathbf{h}^1$, the traveler anticipates the effects of his behavior. Choosing an alternative route generates a new route flow $\mathbf{h}^2$ and this is advantageous if $\tilde{c}_{\rho_1}(\mathbf{h}^1) \geqq \tilde{c}_{\rho_1}(\mathbf{h}^2) > \tilde{c}_{\rho_2}(\mathbf{h}^2)$. Intuitively the definition describes a route flow where no such case exists, i.e., $\mathbf{h}^1$ would be optimal.

All of the above reasoning suggests that the route cost c_ρ decreases if some drivers switch to an alternative route. Unfortunately, one can construct simple network examples where $\tilde{c}_{\rho_1}(\mathbf{h}^1) < \tilde{c}_{\rho_1}(\mathbf{h}^2)$ although the flow on route ρ_1 decreases. One example can be found in Nagurney (1999, p. 164) referring to Braess' paradox; see also Section 3.2.5.3. Irregular cases of that type are omitted here for the sake of brevity.

Another approach to overcome the drawbacks of a deterministic user equilibrium is suggested by Daganzo, Sheffi (1977). If users have imperfect knowledge of the network or if they make other errors in their route choice process, it is natural to modify Wardrop's criterion of a user equilibrium as follows. In a *stochastic user equilibrium* no user *believes* that he can improve his travel time by unilaterally changing routes. We shall come back to this point at a later section.

3.2.5.3 Further Properties of User Optima

(a) Existence

In order to show that a user optimal traffic flow does exist, let us start with the symmetric case as described in (3.72). This is only for the sake of simplicity and Proposition 3.49 below includes also the asymmetric case. Regarding (3.72) the value of the objective function $\tilde{g}$ at some feasible point $\mathbf{h} \in H^D$ can be improved if we are allowed to change $\mathbf{h}$ by $\mathbf{d}$ so that $\mathbf{h} + \mathbf{d} \in H^D$ and $\tilde{g}(\mathbf{h}) > \tilde{g}(\mathbf{h} + \mathbf{d})$. The vector $\mathbf{d}$ has been introduced as a descent direction of $\tilde{g}$. To improve $\tilde{g}$ most rapidly, let us go into the direction of the steepest descent of $\tilde{g}$ at $\mathbf{h}$. That direction can be computed immediately because $\tilde{g}$ is differentiable with $\nabla\tilde{g} = \tilde{\mathbf{c}}$ so that $\mathbf{d} = -\tilde{\mathbf{c}}$. The idea is now to find a sequence of steps where $\mathbf{h}$ is substituted by $\mathbf{h} + \mathbf{d}(\mathbf{h})$ until $\tilde{g}$ cannot be reduced any more. In this case we have either $\mathbf{d}(\hat{\mathbf{h}}) = \mathbf{0}$ or $\hat{\mathbf{h}}$ lies on the border of H^D such that $\hat{\mathbf{h}} + \mathbf{d}(\hat{\mathbf{h}})$ is infeasible.

In doing so we must be aware that the step $\mathbf{d}(\mathbf{h}^1)$ can be too far and the new point would be infeasible, i.e., $\mathbf{h}^1 + \mathbf{d}(\mathbf{h}^1) \notin H^D$. In this case the new point is substituted by a new one, say $\mathbf{h}^2 \in H^D$, which is nearest to $\mathbf{h}^1 + \mathbf{d}(\mathbf{h}^1)$, where the distance is measured by the Euclidean norm. In accordance with the fixed point problem (3.53) the new point is calculated by making use of the orthogonal projection $\mathbf{k}$, that is $\mathbf{h}^{t+1} = \mathbf{k}(\mathbf{h}^t)$ with

$$\mathbf{k}(\mathbf{h}^t) := \arg\min \left\{ |\mathbf{h}^t + \mathbf{d}(\mathbf{h}^t) - \mathbf{h}| \mid \mathbf{h} \in H^D \right\}. \tag{3.80}$$

In the particular case of $\mathbf{h}^t + \mathbf{d}(\mathbf{h}^t) \in H^D$ one has $\mathbf{h}^{t+1} = \mathbf{k}(\mathbf{h}^t) = \mathbf{h}^t + \mathbf{d}(\mathbf{h}^t)$ as postulated above. A graphic representation of the process $\mathbf{h}^0 \to \mathbf{h}^1 \to \mathbf{h}^2 \to \cdots \to \hat{\mathbf{h}}$ is given in Figure 3.15.

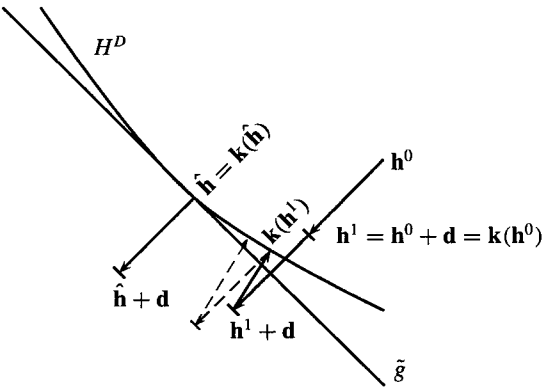


Fig. 3.15
Illustration of the fixed point problem

Assume $-\nabla\tilde{g} = -\tilde{\mathbf{c}} = \mathbf{d} = \text{const.}$ Starting at $\mathbf{h}^0$ yields $\mathbf{k}(\mathbf{h}^0) = \mathbf{h}^0 + \mathbf{d}$. This point in turn leads to an infeasible point $\mathbf{h}^1 + \mathbf{d}$. The nearest point in H^D is $\mathbf{k}(\mathbf{h}^1) = \mathbf{h}^2$. The point where $\tilde{g}$ is tangent to H^D is the only point with $\hat{\mathbf{h}} = \mathbf{k}(\hat{\mathbf{h}})$.

The following proposition is an immediate consequence of Theorem 3.27. It goes back to Smith (1979a) and allows for asymmetric equilibrium problems.

3.49 Proposition (Fixed Point Representation) *A route flow* $\hat{\mathbf{h}}$ *is a fixed point*

$$\hat{\mathbf{h}} = \mathbf{k}(\hat{\mathbf{h}}) \tag{3.81}$$

if and only if it is a user optimum which solves the variational inequality problem (3.77).

Furthermore, (3.80) together with Figure 3.15 suggests a method of the form $\mathbf{h}^t = \mathbf{k}(\mathbf{h}^{t+1})$ to compute a fixed point $\hat{\mathbf{h}}$. Indeed, Dafermos (1980) provided a slightly modified algorithm of that type which converges to a Wardrop equilibrium $\hat{\mathbf{h}}$.

The explicit minimum problem to be solved is

$$\min\left\{\left[\sum_{\rho}\left(h_\rho^t + d_\rho(\mathbf{h}^t) - h_\rho\right)^2\right]^{1/2} \;\middle|\; h_\rho \geqq 0 \;\forall\rho,\; \sum_{\rho\in\mathcal{R}(s,t)} h_\rho = x^D_{(s,t)} \;\forall(v_s, v_t)\right\}.$$

Without the second restriction the optimal h_ρ^{t+1} would be

$$h_\rho^{t+1} = k_\rho(\mathbf{h}^t) = \max\left\{h_\rho^t + d_\rho(\mathbf{h}^t), 0\right\}. \tag{3.82}$$

Given the trip table, this choice is only feasible if $\sum_{\rho\in\mathcal{R}(s,t)} d_\rho(\mathbf{h}^t) = 0$ for every OD-pair. Similar to the construction of user optimized flows in Definition 3.47, this equation requires that drivers using one route ρ_1 can switch to other routes, say ρ_2, only if both routes connect the same OD-pair, $\rho_1 \sim \rho_2$. Every $d_{\rho_1} < 0$ requires some other components of $\mathbf{d}$ to be positive. Unfortunately, going into the direction of steepest descent $\mathbf{d} = -\tilde{\mathbf{c}} < \mathbf{0}$ offends always against this rule. Thus, with respect to Figure 3.15, the mapping from $\mathbf{h}^2$ to $\mathbf{k}(\mathbf{h}^2)$ including the correction of $\mathbf{h}^2 + \mathbf{d}$ represents the general case to be computed.

The rule (3.82) has a more intuitive interpretation when going back to the Definition 3.38 of a Wardrop equilibrium. It is then useful to redefine

$$d_\rho(\mathbf{h}) := -\tilde{c}_\rho(\mathbf{h}) + p_{(s,t)}(\mathbf{h}) \qquad \forall\,(s,t),\ \forall \rho \in \mathcal{R}(s,t),$$

where $p_{(s,t)}(\mathbf{h})$ is the minimum travel cost from v_s to v_t prevailing at the route flow $\mathbf{h}$. Thus h_ρ is reduced if $\tilde{c}_\rho(\mathbf{h}) > p_{(s,t)}(\mathbf{h})$. Although again $\mathbf{d} \leqq \mathbf{0}$, the rule (3.82) now reflects the conditions (3.70) of a user optimum. On the one hand, a positive route flow $\hat{h}_\rho = k_\rho(\hat{\mathbf{h}}) > 0$ requires $d_\rho(\hat{\mathbf{h}}) = 0$ or $\tilde{c}_\rho(\hat{\mathbf{h}}) = p_{(s,t)}(\hat{\mathbf{h}})$. On the other hand, $\hat{h}_\rho = k_\rho(\hat{\mathbf{h}}) = 0$ implies $d_\rho(\hat{\mathbf{h}}) \leqq 0$. Nevertheless the same drawback is valid as before. For almost all feasible flows $\mathbf{h} \in H^D$ the reassignment $\mathbf{h} + \mathbf{d}(\mathbf{h})$ is infeasible with respect to H^D and we are asked to compute a feasible flow which is nearest to that reassignment.

To prove the existence of a Wardrop equilibrium $\hat{\mathbf{h}}$ we have to show that a fixed point of the form (3.81) exists. In doing so Brouwer's fixed point theorem gives an answer: *Let C be a non-empty compact convex subset in $\mathbb{R}^n$. If the function $\mathbf{f}\colon C \to C$ is continuous, then $\mathbf{f}$ has a fixed point $\hat{\mathbf{x}}$, i.e., $\hat{\mathbf{x}} = \mathbf{f}(\hat{\mathbf{x}})$.* From the definition of $\mathbf{k}\colon H^D \to H^D$ in (3.80) its continuity follows from the observation that $\mathbf{d} = -\tilde{\mathbf{c}}$ is continuous and that the minimum function preserves continuity. Without capacity constraints there is at least one $\mathbf{h} \in H^D$. This observation is based upon the assumption of a strongly connected network so that every driver can find a feasible route. As the set H^D of demand feasible route flows is also convex and compact we gain as an application of Theorem 3.28:

3.50 Proposition (Existence) *Let $\tilde{\mathbf{c}}$ be continuous on H^D, and let H^D be a non-empty compact convex set. Then $\mathbf{k}\colon H^D \to H^D$, defined by* (3.82)*, is continuous and has a fixed point $\hat{\mathbf{h}} = \mathbf{k}(\hat{\mathbf{h}})$ being a Wardrop equilibrium.*

A similar result with respect to link flows $\boldsymbol{\varphi} \in \Phi^D$ and additional capacity constraints ($\boldsymbol{\varphi} \in \Phi^S$) is given in Smith (1979a, p. 302). An alternative proof based on Helly's theorem can be found in de Palma, Nesterov (1998).

(b) Stability

Having shown the existence of a Wardrop equilibrium on the basis of the orthogonal projection (3.80), one might think of the process described by Figure 3.15 as a procedure for detecting such an equilibrium. One disadvantage with this approach is that the flow $\mathbf{h}$ is modified by a sequence of discrete steps, the length of which hopefully converges to zero. An alternative dynamic model of traffic assignment in which drivers smoothly change their route choice to take advantage of cheaper routes is proposed by Smith (1984, 1984c). This approach represents dynamics in traffic assignment even when the total demand for trips does not vary with time. The idea is to define shares $\theta_{\rho_1\rho}(\mathbf{h})$ of a route flow h_{ρ_1} switching from route ρ_1 to an alternative route ρ if the first route is more expensive than the second one. Consequently, one needs $\theta_{\rho_1\rho}(\mathbf{h}) = -\theta_{\rho\rho_1}(\mathbf{h})$. Furthermore, the number of drivers leaving a costly route should increase with the cost differential. An appropriate *instantaneous* swapping rate follows from

$$\theta_{\rho_1\rho}(\mathbf{h}) = \begin{cases} \alpha[\tilde{c}_{\rho_1}(\mathbf{h}) - \tilde{c}_\rho(\mathbf{h})] & \text{if } \rho \sim \rho_1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ is a conversion factor for dimensional consistency that can be calibrated in the particular application. The term "instantaneous" emphasizes that this rate refers to a point of time. Clearly $\theta_{\rho_1\rho}(\mathbf{h})$ can numerically exceed h_{ρ_1}, but as no flow can be negative, the rate of leaving drivers must come to an end after some period of time. At the latest for $h_{\rho_1} = 0$ the swapping rate cannot be positive. Redefining the above leaving rate by the corresponding entering rate suggests

$$\Theta_{\rho_1\rho}(\mathbf{h}) = \begin{cases} -\theta_{\rho_1\rho}(\mathbf{h}) & \text{if } h_{\rho_1} > 0, \\ \max\{0, -\theta_{\rho_1\rho}(\mathbf{h})\} & \text{if } h_{\rho_1} = 0. \end{cases} \tag{3.83}$$

Before proceeding with dynamics observe that a traffic assignment $\hat{\mathbf{h}}$ is a Wardrop equilibrium at any instant if and only if no traveler has an incentive to alter his route choice.

$$\hat{h}_{\rho_1} \cdot \max\{0, \theta_{\rho_1\rho}(\hat{\mathbf{h}})\} = 0 \qquad \forall \rho \sim \rho_1$$

In fact, this is equivalent to the optimum condition (3.74): all positive route flows $\hat{h}_{\rho_1}$ require that there is no alternative route with lower cost, that is $\theta_{\rho_1\rho}(\hat{\mathbf{h}}) = 0$. If ρ_1 is not among the cheapest routes, we have $\theta_{\rho_1\rho}(\hat{\mathbf{h}}) > 0$ and at the same time $\hat{h}_{\rho_1} = 0$ for any equilibrium flow. For every disequilibrium the instantaneous change of h_{ρ_1} at time t is the sum of all switching rates entering ρ_1 or leaving it:

$$\dot{h}_{\rho_1}(t) = \sum_{\rho \sim \rho_1} \Theta_{\rho_1\rho}(\mathbf{h}(t))$$

with $\dot{\mathbf{h}} = \mathbf{0}$ indicating a Wardrop equilibrium. Smith (1984b) shows by means of a Lyapunov theorem that a similar dynamic system converges to a non-empty set of equilibria when route costs $\tilde{\mathbf{c}}$ are continuously differentiable and monotone[56]. This result addresses the *global stability* by which a network equilibrium is said to be stable if convergence of the flow pattern to some equilibrium is guaranteed from arbitrary initial conditions; see also Friesz (1985).

Zhang, Nagurney (1996) and Nagurney (1999) discuss both local and global stability issues. Their dynamic model for the travel route choice adjustment process was proposed in Dupuit, Nagurney (1993) and in fact they detect that (3.83) is closely related to the orthogonal projection (3.80). Without going into the details of that analysis, the authors describe an adjustment process similar to (3.83) and they show that this process is stable if the link cost functions $\mathbf{c}$ are monotone in the link load pattern $\boldsymbol{\varphi}$.

3.51 Remark (Braess' Paradox) Stability of user equilibria is of particular interest when changing the topology of a transport network, for example, when the network designer opens a new link. Some drivers may now take advantage of that link so that the corresponding adjustment process transforms the former user equilibrium into a new one. In doing so the network participants ignore their impacts on the rest

[56] Dafermos, Nagurney (1984) show by virtue of a strong monotonicity property that small changes in the transportation cost functions induce small changes in the equilibrium pattern, that is, jumps as can be expected for instable equilibria are ruled out.

of the network. Indeed, Braess (1968) gave an example where the total travel costs as well as the travel costs for every user *increase* although the restrictions on the network are relaxed. Today that observation is referred to as *Braess' paradox* and it is recognized by almost all traffic engineers.

The following simple example is taken from Braess (1968). For $c_5 = \infty$ the reader may verify that six travelers driving from v_1 to v_4 should split into two groups of three cars. One group uses the route (v_1, v_2, v_4) and the other one travels along (v_1, v_3, v_4) that is $\boldsymbol{\varphi} = (3, 3, 3, 3, 0)$. The cost per trip is 83 on both paths and no driver can unilaterally benefit from switching to the other route. (By the way this solution is also system optimal.)

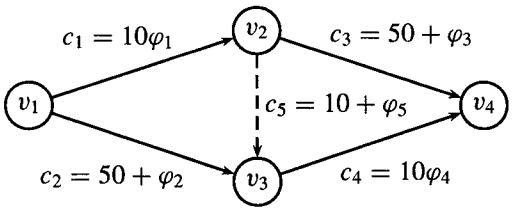


Fig. 3.16
Braess' Paradox
Introducing the fifth link is the same as reducing travel cost c_5 from infinity to $c_5 = 10 + \varphi_5$.

In the next step the network designer opens the fifth link to traffic at a link cost of $c_5 = 10 + \varphi_5$. Now one driver of the first group discovers that he can reduce his own trip cost by 2 (compute 30+11+40=81) if he alters his route choice into (v_1, v_2, v_3, v_4). The additional costs for the rest of the network are irrelevant for this decision. By similar reasoning, one driver of the second group with new costs of $53 + 40 = 93$ can reduce his costs by 1 if he also switches to that route (40+12+30=92). The new flow $\boldsymbol{\varphi} = (4, 2, 2, 4, 2)$ constitutes a new Wardrop equilibrium but it implies $c_\rho = 92$ for every route so that all drivers are worse off due to considerable cost increments on the first and fourth link.

The described paradox is no sleight of hand. Similar physical phenomena are described in Pöppe (1995) with respect to springs or resistors in parallel or in line. As can be found in Patriksson (1994, p. 91), however, under strong assumptions on the travel costs, Braess' Paradox may be guaranteed not to occur. □

3.2.5.4 Routing in Data Networks

Regarding road networks, travelers are assumed to choose their route in advance. In other networks, where the detailed path descriptions (that is the OD/path incidence matrix $\boldsymbol{\Lambda}$) are not known at the origins, it may be necessary to maintain a routing table at each node. These tables consist of routing variables $\theta_{ij}^{(t)}$ defined as the fraction of packets arriving at node i with destination t which are routed directly to node j. Each packet arriving at node i carries merely its destination address t but it has no information about the route it uses. A facility (e.g., a switch) located at node i then collects all incoming flows, classifies them with respect to their destination t, and assigns each class to outgoing channels in accordance with the routing table.

To point out the relationship with the former assignment problem, suppose that the OD/path incidence matrix $\boldsymbol{\Lambda}$ is known. Let $\hat{\boldsymbol{\varphi}}$ be a *system optimal traffic*

assignment solving[57]

$$\min\left\{\mathbf{c}(\boldsymbol{\varphi})^{\mathsf{T}}\boldsymbol{\varphi} \mid \boldsymbol{\Delta}\,\mathbf{h} = \boldsymbol{\varphi},\ \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^{D},\ \mathbf{h} \geqq \mathbf{0}\right\}. \tag{3.55}$$

Hence, there is some $\hat{\mathbf{h}}$ satisfying $\boldsymbol{\Delta}\,\hat{\mathbf{h}} = \hat{\boldsymbol{\varphi}}$ or more specifically

$$\hat{\varphi}_e = \sum_{\rho} \Delta_{e\rho}\hat{h}_\rho = \sum_{t\in\mathcal{V}}\sum_{s\in\mathcal{V}}\sum_{\rho\in\mathcal{R}(s,t)} \Delta_{e\rho}\hat{h}_\rho.$$

With regard to data networks, the first new element is that each link flow can be decomposed into destination specific flows

$$\hat{\varphi}_e = \sum_{t\in\mathcal{V}} \hat{\varphi}_e^{(t)} \quad \text{and} \quad \hat{\varphi}_e^{(t)} = \sum_{s\in\mathcal{V}}\sum_{\rho\in\mathcal{R}(s,t)} \Delta_{e\rho}\hat{h}_\rho.$$

The second additional element follows from the fact that routing is performed by some node specific facility.[58] It has the task to assign the outflow $\zeta_v^{(t)}$ of node v with destination t to outgoing links in accordance with some routing variables which are to be determined in a favorable manner. On a directed network outflows are given by

$$\zeta_v^{(t)} = \sum_{e:\ \alpha(e)=v} \hat{\varphi}_e^{(t)} \qquad \forall v \in \mathcal{V},\ \forall t \in \mathcal{V},$$

where every link e is uniquely defined by its tail $\alpha(e)$ and its head $\omega(e)$. For a network with n nodes that suggests to define routing shares

$$\theta_{ij}^{(t)} := \hat{\varphi}_{ij}^{(t)} / \zeta_i^{(t)} \qquad i = 1,\ldots,n,\ j = 1,\ldots,n,\ t = 1,\ldots,n, \tag{3.84}$$

where $\theta_{ij}^{(t)} = 0$ if there is no arc e from i to j. In contrast, $\theta_{ij}^{(t)} > 0$ implies an existing link from $\alpha(e) = i$ and $\omega(e) = j$. As usual the nodal flow vector is denoted by $\boldsymbol{\zeta}^{(t)} = (\zeta_1^{(t)},\ldots,\zeta_n^{(t)})^{\mathsf{T}}$ and the *routing patterns* are

$$\boldsymbol{\Theta}^{(t)} = \left[\theta_{ij}^{(t)}\right] \qquad \forall t \in \mathcal{V}.$$

Apart from (3.84), all admissible routing patterns require

$$\sum_{j\in\mathcal{V}} \theta_{ij}^{(t)} = 1,\ \theta_{ij}^{(t)} \geqq 0,\ \theta_{ii}^{(t)} = 0,\ \theta_{tj}^{(t)} = 0. \tag{3.85}$$

Thus the only fractions we can choose are the $\theta_{ij}^{(t)}$ with $i \neq j$ and $i \neq t$. Finally, the conservation of flows must be satisfied for every destination node $t \in \mathcal{V}$, that is

[57] Apart from the objective function such a routing problem is discussed in Bertsekas, Gallager (1992, p. 451 ff.). Be aware that we ignore here all problems of queuing or any other stochastic processes.

[58] There are two distinct approaches to routing traffic, namely, connection-oriented and connectionless; see Corne et al. (2000). Connectionless networks use information contained within the data packet itself. Connection-oriented networks establish the route first and subsequently use it for all traffic associated with the connection until it is torn down.

$$\begin{pmatrix} \zeta_1^{(t)} \\ \vdots \\ \zeta_n^{(t)} \end{pmatrix} = \begin{pmatrix} x^D_{(1,t)} \\ \vdots \\ x^D_{(n,t)} \end{pmatrix} + \begin{pmatrix} \theta_{11}^{(t)}\zeta_1^{(t)} + \cdots + \theta_{n1}^{(t)}\zeta_n^{(t)} \\ \vdots \\ \theta_{1n}^{(t)}\zeta_1^{(t)} + \cdots + \theta_{nn}^{(t)}\zeta_n^{(t)} \end{pmatrix}$$

or, equivalently, in vector notation

$$\boldsymbol{\zeta}^{(t)} = \mathbf{x}^D_{(t)} + \boldsymbol{\Theta}^{(t)\mathsf{T}}\boldsymbol{\zeta}^{(t)} \qquad \forall t \in \mathcal{V}.$$

The outflow $\boldsymbol{\zeta}^{(t)}$ is composed of the initiated flows $\mathbf{x}^D_{(t)}$ plus the inflow $\boldsymbol{\Theta}^{(t)\mathsf{T}}\boldsymbol{\zeta}^{(t)}$ from all predecessor (or upstream) nodes. In particular one has $\zeta_t^{(t)} = x^D_{(t,t)}$ which may be negative if the corresponding amount of traffic leaves the network at that node. The collection of $\mathbf{x}^D_{(t)}$ without exiting traffic determines the former trip matrix $\mathbf{x}^D$. Suppose that all of the quadratic routing patterns $\mathbf{I} - \boldsymbol{\Theta}^{(t)}$ can be inverted, then the flow conservation conditions take the form

$$\boldsymbol{\zeta}^{(t)} = \left[\mathbf{I} - \boldsymbol{\Theta}^{(t)\mathsf{T}}\right]^{-1}\mathbf{x}^D_{(t)} \qquad \forall t \in \mathcal{V}. \tag{3.86}$$

The results obtained so far suggest that all we have to do is to substitute the link flow concept in (3.55) by a corresponding node flow concept. However, if we examine a capacitated version of this traffic assignment problem with respect to optimal routing patterns, total link flows are restricted by link capacity constraints, see, e.g., Mason (1985).

$$0 \leqq \varphi_e = \sum_{t \in \mathcal{V}} \varphi_e^{(t)} \leqq \kappa_e \qquad \forall e \in \mathcal{E} \tag{3.87}$$

The last step towards an appropriate adjustment of (3.55) is to check for the objective function. Up until now the transport costs per link are given by $c_e(\boldsymbol{\varphi})\,\varphi_e$. Such a cost function is of minor importance for telecommunications networks. Instead a frequently used link cost or delay function is $C_e(\varphi_e) = \frac{\alpha}{\kappa_e - \varphi_e}$ with $0 \leqq \varphi_e < \kappa_e$, where α indicates the mean packet length; for further details see Section 5.2.3.3. Using this or even more general cost functions $C_e(\boldsymbol{\varphi})$, the cost minimizing routing problem turns out to be

$$\min_{\boldsymbol{\Theta}^{(t)},\, t \in \mathcal{V}} \left\{ \sum_{e \in \mathcal{E}} C_e(\boldsymbol{\varphi}) \mid (3.84), (3.85), (3.86), (3.87) \right\}.$$

A feasible solution consists of n routing patterns $\hat{\boldsymbol{\Theta}}^{(t)}$, $t = 1,\ldots,n$. For each destination $t \in \mathcal{V}$ the routing pattern $\hat{\boldsymbol{\Theta}}^{(t)}$ determines together with the given demand $\mathbf{x}^D_{(t)}$ the nodal flow $\hat{\boldsymbol{\zeta}}^{(t)}$. These nodal flows and the routing patterns suffice to compute feasible total link flows $\hat{\boldsymbol{\varphi}}$, i.e., $\mathbf{0} \leqq \hat{\boldsymbol{\varphi}} \leqq \boldsymbol{\kappa}$. All what remains to be done is to configure nodes with corresponding routing tables, that is, a switch located at node v needs to know the v^{th} vector of every routing pattern $\hat{\boldsymbol{\Theta}}^{(t)}$.

The main computational problems are at hand.

(a) Even in well behaved cases the KKT conditions result in a large coupled nonlinear system in the routing variables.

(b) The matrices $[\mathbf{I} - \boldsymbol{\Theta}^{(t)}]$ must be non-singular, cf. (3.86). If one matrix is singular for some t, the routing is said to be *deadlocked.*

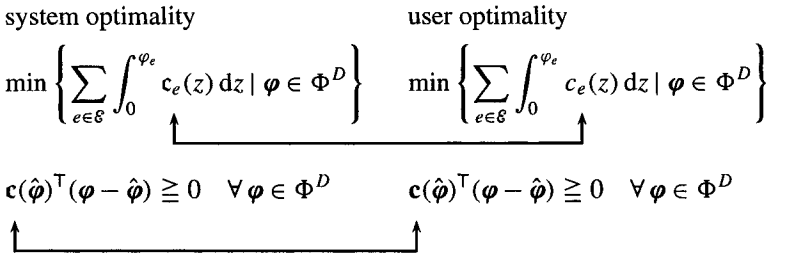
(c) The most important problem is to avoid that messages are routed along cycles. Mason (1985) shows that the routing is *acyclic* (or loop-free) with respect to destination t if $[\boldsymbol{\Theta}^{(t)\mathsf{T}}]^n = 0$, where $n = t$ is the number of nodes. Similar observations hold true for problems of synchronization of network processes; see Chapter 6.

3.3 Summary

Chapter 3 has started with a capacitated network for one commodity. We have derived the maximum flow which corresponds to a technically efficient activity. This is the same as having a production function that determines the maximum output for given amounts of inputs, say link capacities. The next step leads to an optimization problem with capacitated link flows as inputs such that a given OD-flow is produced at minimum transport costs. It is not too hard to generalize this single commodity problem to many places of production and many consumer destinations. Nevertheless, more realistic transport problems need a reformulation with regard to multicommodity flows by which each OD-flow represents a distinct good. In this case the concept of a production function must be replaced by more general production correspondences so that the former notion of a maximum output or a "network capacity" becomes void. Any feasible activity may now be described by an output distance function so that all technically efficient flow patterns form a so-called production possibility frontier (or transformation curve). Irrespective of these problems on flow maximization, the formulation of a cost minimal production of overlapping multicommodity flows remains valid. Given the network structure with fixed capacities, the superordinate task is to realize a prespecified trip table at minimum transport costs. One major problem, however, results from the fact that individual network users do not necessarily cooperate in minimizing total transport costs but they pursue their own aim of minimizing private costs. In doing so they ignore external costs caused by themselves and this is the core problem of a cost minimal trip assignment.

The analysis of system optimality and user optimality suggests that the solutions differ in general. Regarding system optimality the traffic assignment is oriented at the KKT conditions (3.59). They require that the route flows are controlled by *marginal route costs per trip* $\tilde{\mathbf{c}}$ which include the effects that individuals impose on the rest of the network. Against that Dafermos, Sparrow (1969) exposed that user optima are characterized by similar KKT conditions, namely, (3.74): the traffic assignment is now regulated by *route costs per trip* $\tilde{\mathbf{c}}$. The two approaches differ because individual network participants ignore their own disturbing effects on the other network flows. They observe the minimum route costs $\mathbf{p}$ at the prevailing network conditions and believe that these prices are constant independent of their own behavior.

Technically, the difference may be seen best by comparing the two minimization problems. For the symmetric and the asymmetric equilibrium problem we have



These problems can equivalently be expressed in route flows $\mathbf{h}$ instead of link flows $\boldsymbol{\varphi}$. The corresponding rearrangement is meaningful because network users decide on their route choice on the basis of route costs rather than link costs. Each additional trip on route ρ, $r = 1,\ldots, n$, increases total travel costs $\tilde{f}(\mathbf{h}) = \tilde{\mathbf{c}}(\mathbf{h})^{\mathsf{T}}\mathbf{h}$ of a traffic flow $\mathbf{h}$ approximately by

$$\tilde{\mathbf{c}}_\rho(\mathbf{h}) = \tilde{c}_\rho(\mathbf{h}) + \sum_{r=1}^{n} \frac{\partial \tilde{c}_r(\mathbf{h})}{\partial h_\rho}\, h_r.$$

In networks where users' costs are not affected by traffic flows and thus when there is no congestion, the last term is zero so that $\tilde{\mathbf{c}}_\rho = \tilde{c}_\rho$. In this case the normative and the descriptive traffic assignment have the same result; cf. Steenbrink (1974). In all other cases marginal and average costs differ. As long as there is no force that assigns the last cost term to the party causing it, user decisions consider only the immediate effect $\tilde{c}_\rho(\mathbf{h})$. All travelers ignore the indirect effects, that is to say the external costs they impose on the rest of the network. The reason is plausible because each trip-maker experiences himself to be of no influence on the total network process. As a consequence all users together tend to neglect remarkable holdups in traffic.

To correct the users' behavior one may think of a network provider who knows the given trip table $\mathbf{x}^D$ and the corresponding system optimal traffic assignment $\hat{\mathbf{h}}$. Such a provider is in principle able to calculate adequate tolls per road (cf. Yang, Huang (1998)) or route so that travelers internalize their true costs per trip. The problems of this idea are obvious. Every route needs in general a different fee and the costs to levy the tolls are expected to be huge. Moreover, the trip table and, therefore, the pricing scheme may change in time (e.g., rush hours, growing networks, technical progress, etc.). Another problem is that of a price dependent demand for traffic which is discussed in the next section. All these reasons require a more or less complete revision of the price system.

The following review summarizes the most important aspects of the above analysis of traffic assignment. Most of the arguments can also be found in Boyce, Daskin (1997).

1. The cost minimizing drivers are assumed to have *perfect information* concerning all details of the network. The consequence is a deterministic user equilibrium which rules out any variations in the perception of the network by its participants. In particular network users make no mistakes due to imperfect knowledge of the network and of the processes performed on it. The next section additionally considers some stochastic elements in the users' route choice.

2. User decisions have no *time-of-day dimension*, especially they cannot choose departure times; under these assumptions travelers respond to congestion primarily through route choice decisions. An important dimension of choice available to network participants for avoiding congestion (rush hour) is thus left out, namely, the decision regarding departure time, see Ben-Akiva (1985) or Arnott et al. (1990). Several approaches implementing a time-of-day dimension do exist, for example de Palma et al. (1983) and discuss the departure time choice with respect to queue delays during a peak period. Their model consists of relations describing (1) waiting time, (2) queue length, (3) arrival model, and (4) utility function which depends on travel time and arrival time. Mahmassani, Herman (1984) propose a model which combines the users' choice of departure time and the route choice. A plausible extension for the definition of a Wardrop equilibrium is then that no traveler can unilaterally improve his travel cost by changing departure time and/or route. The demand for a dynamic network equilibrium research has also been picked up by Friesz et al. (1989). They show in the context of a continuous time optimal control problem that there are dynamic traffic assignment problems which preserve the static results above. Furthermore, Friesz et al. (1993) propose a variational inequality formulation of the dynamic user equilibrium when certain regularity conditions hold good. Observe that dynamic models are restricted not only to flow conservation for links and nodes but also to flow propagation through space and time. An extensive analysis of dynamic route choice is given in Ran, Boyce (1996). On a variational inequality formulation of the discrete-time dynamic user-optimal route choice problem see Chen (1999).

3. The processes on the network are restricted to a *given network*. The nodes, their connections, the capacities, and the functionality, all these network elements remain unchanged although the construction of the network may be suboptimal with regard to the task it has to perform. Network design considers possible adjustments that will improve the performance of a network, see Chapter 5. In what we have learned so far Braess' paradox indicates that there may be awkward properties of a network by which the user optimal link costs can increase after the addition (or the improvement) of one or more links. That problem affects also traffic light control.

4. Examining network flows requests to collect flow data. One provisional approach is to open check points and to count link flows per period and probably to measure speed as an indicator for the quality of service. Even if this method describes a user optimum, it cannot reveal all information needed: optimal link flows $\hat{\boldsymbol{\varphi}}$ may be unique but in general route flows $\hat{\mathbf{h}}$ are not. We know where flow units are but the flow conservation rule does not tell completely where they come from and where they go to.

5. Network users tend to ignore parts of their travel cost by shifting it to the rest of network participants. This behavior can lead to a severe increase in total travel cost. Basically the analysis has shown indeed that there is an adequate instrument capable of correcting this defect in accordance with the origination principle. Such a *pricing scheme*, i.e., a system of link or route tolls, changes the user behavior so as to achieve a system optimum.

4

Utilization of Transport Networks: Extensions

4.1 Stochastic Elements of Traffic Assignment

4.1.1 Stochastic Route Choice

(a) Wardrop's Principles Revisited

This chapter pursues the core problem of traffic assignment by introducing several additional aspects. In the first instance we extend the analysis with regard to stochastic route choice. After that the analysis is expanded with respect to a price sensitive demand for traffic. The chapter ends with selected extensions, such as dynamic traffic assigment, that indicate further directions of research.

Stochastic choice of routes, modes, departure times and so on requires deep knowledge of stochastic theory. Even though we are not going into detail, it is nevertheless useful to elaborate some basic principles when path choice follows certain probabilistic rules. A *deterministic* optimal traffic assignment $\hat{\boldsymbol{\varphi}}$ or $\hat{\mathbf{h}}$ according to the Wardrop conditions (3.70) requires perfect knowledge on the part of the drivers and that each driver perceives costs in the same way. Allowing for variations in user perception and knowledge of the network suggests to substitute the deterministic equilibrium by a stochastic approach.

The network users do not necessarily realize the user optimal assignment. One reason may be that travelers make mistakes in their choice process due to their imperfect information on the network. Their errors, however, are not systematic so that the mean error is zero. In accordance with Daganzo, Sheffi (1977), it is natural to modify Wardrop's deterministic principle of route choice as follows.

4.1 Definition *In a **stochastic user equilibrium** no user* believes *that he can improve his travel time by unilaterally changing his route.*

Given the demand for traffic $\mathbf{x}^D$, a stochastic user equilibrium describes the trip assignment $\mathbf{h}$ most likely to arise in reality. As all travelers have their own expectations on route costs which are the basis for their path choice, it is suggestive to discuss route flows as random variables. The next section derives the most likely

traffic pattern on the basis of a multinomial distribution. The trip table and the probabilities for choosing a certain route are assumed to be known.

Observe that imperfect information can be reduced by electronic route guidance. On a combined route guidance and road pricing system in congested traffic networks with stochastic route choice see Yang (1999). Against that, Mahmassani, Chang (1987) tackle the problem of imperfect information by a different approach based on boundedly rational decision-making. In accordance with Simon (1955), this method seeks an acceptable solution for traffic assignment rather than a necessarily optimal one.

(b) Multinomial Distribution

For explanatory purposes, pick out one OD-pair (v_s, v_t) and renumber the available routes $\rho \in \mathcal{R}(s,t)$ by $\rho = 1,\ldots,r$. Let the flows on all these routes be random variables such that h_ρ indicates that the route ρ has been chosen by h_ρ travelers. The joint probability mass function is supposed to be given by

$$\text{(4.1a)} \qquad \text{Prob}\{h_1,\ldots,h_r|\ \pi_1,\ldots,\pi_r|\ x^D_{(s,t)}\} = \frac{x^D_{(s,t)}!}{h_1!\cdots h_r!}\,{\pi_1}^{h_1}\cdots{\pi_r}^{h_r},$$

$$\text{(4.1b)} \qquad \text{where} \quad x^D_{(s,t)} = \sum_{\rho=1}^{r} h_r \quad \text{and} \quad \sum_{\rho=1}^{r} \pi_r = 1.$$

A vector of random variables showing this probability mass function is said to have a *multinomial distribution* (see, e.g., Rohatgi (1976)) and the inclosed fraction is called the *multinomial coefficient*. Furthermore, π_ρ is the probability of selecting route ρ. The expected value of trips on route ρ is

$$E(h_\rho) = \pi_\rho\, x^D_{(s,t)}$$

and its variance is $\pi_\rho\,(1-\pi_\rho)x^D_{(s,t)}$. Moreover, the relative frequencies $h_\rho/x^D_{(s,t)}$ serve as maximum-likelihood estimators of π_ρ.

This approach has severe drawbacks, especially when distinct routes overlap at some links. Daganzo, Sheffi (1977) avoid this problem by using link specific random variables which are mutually independent. The corresponding route specific variables are then assumed to have a multivariate normal distribution (see Rohatgi (1976)). The cost of more accuracy is that we need much more knowledge on the basic properties of probability distributions. Disregarding that objection, the most likely traffic assignment in the multinomial approach results from maximizing $\text{Prob}\{h_1,\ldots,h_r|\cdot\}$ over the route flows. In doing so we take the logarithm of (4.1a) and make use of Stirling's approximation[1], namely, $\ln(x!) \approx x\ln x - x$.

[1] Stirling's approximation states $n! \to n^n\,\mathrm{e}^{-n}\sqrt{2\pi n}$ as $n \to \infty$. Taking logarithms implies $\ln(n!) \approx (n+\tfrac{1}{2})\ln n - n + \tfrac{1}{2}\ln(2\pi)$. That approximation is suitable even for small numbers. For $n = 50$ we compute $\ln(n!) \approx 148.478$, $(n+\tfrac{1}{2})\ln n - n + \tfrac{1}{2}\ln(2\pi) \approx 148.476$, and $n\ln n - n \approx 145.601$.

$$\ln \text{Prob}\{\cdot\} \approx x^D_{(s,t)} \ln x^D_{(s,t)} - x^D_{(s,t)} - \sum_{\rho=1}^{r} (h_\rho \ln h_\rho - h_\rho) + \sum_{\rho=1}^{r} h_\rho \ln \pi_\rho$$

It is not too hard to show[2] that the maximum of $\ln \text{Prob}\{h_1, \ldots, h_r | \cdot\}$ subject to $x^D_{(s,t)} = \sum_\rho h_\rho$ is attained at

$$h_\rho = \pi_\rho x^D_{(s,t)} \qquad \rho = 1, \ldots, r,$$

that is at the expected trip values. Indeed, substituting these values into the above approximation shows $\ln \text{Prob}\{\cdot\} \approx 0$ which indicates that the probability of this structure of traffic assignment is far greater than any other state. Recall that $x^D_{(s,t)}$ is a known constant and let us assume for the moment that all path choice probabilities are equal, $\pi_\rho = 1/r$, so that the last sum is also constant. Then the above approximation can be reduced to either of the following forms.

$$\ln \text{Prob}\{\cdot\} \approx \text{const.} - x^D_{(s,t)} \sum_{\rho=1}^{r} \frac{h_\rho}{x^D_{(s,t)}} \ln \frac{h_\rho}{x^D_{(s,t)}} \tag{4.2a}$$

$$\ln \text{Prob}\{\cdot\} \approx \text{const.} - \sum_{\rho=1}^{r} h_\rho \ln h_\rho \tag{4.2b}$$

In transportation economics it is common practice to focus on the last sum; some examples will be given below. This sum corresponds to the multinomial coefficient which indicates the number of permutations of $x^D_{(s,t)}$ with respect to the available routes $\rho = 1, \ldots, r$. *Maximizing the multinomial coefficient for equally likely path choices is the same as searching for the most likely route choice between the OD-pair* (v_s, v_t) *to arise in reality.* Note that the maximization of $\ln \text{Prob}\{\cdot\}$ is the same as minimizing $\sum_\rho h_\rho \ln h_\rho$. For computational convenience the last formula is often substituted by $\sum_\rho (h_\rho \ln h_\rho - h_\rho)$. The difference is a constant $\sum_\rho h_\rho = x^D_{(s,t)}$ which has no effect on some optimization procedure.

4.2 Remark Stirling's approximation is valid for very large numbers. Nevertheless, the following computations make use of this approximation even for small non-negative numbers. In particular, for $x = 0$ the logarithm is not defined. In this case we put $x \ln x - x = 0$. Fortunately, $0! = 1$ or $\ln(0!) = 0$ confirm this correction. □

(c) Entropy

The entropy describes the probability of a certain system state. It measures the degree of disorder and/or irreversibility. Irreversible processes continue on their own as long as the system reaches a situation of maximum likelihood. The entropy then attains its maximum. An example is the complete mixture of two gases. Here entropy indicates the largest number of associated microstates.[3] A growing entropy means that a

[2] Just use the Lagrangean approach and recall $\sum_{\rho=1}^{r} \pi_r = 1$.

[3] On the corresponding Lagrangean procedure of entropy maximization see also Buhr, Pauck (1981, p. 411–416).

system goes over from one situation to another situation with higher probability. This concept is frequently applied to network systems where the adjustment procedure of pressing the traffic assignment into the direction of a user equilibrium is interpreted as a process of growing entropy. Similar concepts can also be applied to location theory. For an equilibrium model describing the spatial distribution of households and producers by means of entropy maximization see for instance Batten, Roy (1982).

An appropriate measure of the dispersion or spread of a discrete probability distribution is the following entropy function, see Theil (1967).

4.3 Definition (Entropy) *The entropy function* $\chi(\mathcal{J})$ *of a discrete*[4] *probability distribution defined on the set* $\mathcal{J}$ *is*

$$\chi(\mathcal{J}) = -\sum_{j\in\mathcal{J}} \text{Prob}\{A_j\} \ln \text{Prob}\{A_j\},$$

where $\text{Prob}\{A_j\}$ *is the probability of the event* A_j *and* $\mathcal{J}$ *is the set of mutually exclusive and exhaustive events.*

If we substitute $\text{Prob}\{A_j\}$ by the route choice probabilities, then

$$\chi(\mathcal{R}(s,t)) = -\sum_{\rho\in\mathcal{R}(s,t)} \pi_\rho \ln \pi_\rho. \tag{4.3}$$

Using the maximum likelihood estimators $\pi_\rho = h_\rho / x^D_{(s,t)}$, we obtain the sum in (4.2a). The results so far obtained are twofold. On the one hand, given π_ρ we can expect to observe h_ρ. On the other hand, observing h_ρ gives rise to estimate the path choice probability at π_ρ.

4.4 Remark (Entropy of a probability distribution) Further results can be derived if we have more information on the network. For example, assume that the expectation value $E[\tilde{c}_\rho(h_\rho)]$ of the route cost $\tilde{c}_\rho(h_\rho)$ is known, where h_ρ can take the values $h_{\rho 1},\ldots, h_{\rho z}$ with probabilities $\pi_{\rho 1},\ldots, \pi_{\rho z}$. Wilson (1970) shows that maximizing the (path related) entropy

$$\max\left\{\sum_{j=1}^{z} \pi_{\rho j} \ln \pi_{\rho j} \,\middle|\, \sum_{j=1}^{z} \pi_{\rho j} = 1,\ \sum_{j=1}^{z} \pi_{\rho j} \tilde{c}_\rho(h_{\rho j}) = E[\tilde{c}_\rho(h_\rho)]\right\}$$

yields the following maximum entropy estimates ($j = 1,\ldots, z$)

$$\pi_{\rho j} = \frac{\exp[-\theta \tilde{c}_\rho(h_{\rho j})]}{\sum_{k=1}^{z} \exp[-\theta \tilde{c}_\rho(h_{\rho k})]}.$$

Here the parameter θ is the Lagrange multiplier of the second restriction. Functional forms of the preceding type are characteristic for logit assignment models and will frequently return below. □

[4] A corresponding formula with respect to a continuous distribution defines the entropy for any thermodynamic system; see Georgescu-Roegen (1987).

The next step is to generalize the result for one OD-pair to the whole network. The number of permutations across all OD-pairs $W(\mathbf{x}^D)$ is given by

$$W(\mathbf{x}^D) = \prod_{(s,t)} W\big(x^D_{(s,t)}\big) \quad \text{with} \quad W\big(x^D_{(s,t)}\big) = \frac{x^D_{(s,t)}!}{\prod_{\rho\in\mathcal{R}(s,t)} h_\rho!}$$

where the last term indicates the number of permutations of $x^D_{(s,t)}$ with respect to the available routes $\rho \in \mathcal{R}(s,t)$, i.e., the multinomial coefficient in (4.1a). Approximating each factor in accordance with (4.2b) yields an alternative expression for the entropy

$$\chi(\mathbf{h}) = -\sum_{(s,t)} \sum_{\rho\in\mathcal{R}(s,t)} h_\rho \ln h_\rho - h_\rho = -\mathbf{h}^\mathsf{T}(\ln \mathbf{h} - \mathbf{1}),$$

where, as mentioned above, $\sum_{(s,t)} \sum_{\rho\in\mathcal{R}(s,t)} h_\rho = \mathbf{h}^\mathsf{T}\mathbf{1}$ is the constant total number of trips observed. The function $\chi(\mathbf{h})$ is meant to be suggestive as it expresses the same idea as (4.3). In computing this function it is implicitly assumed that the joint probability distribution of $\mathbf{h}$ is the product of OD-specific, independent multinomial distributions. Although this supposition cannot be expected to hold true (recall the condition of flow conservation $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$ and remember the objection against a multinomial distribution), it is a widely used approach due to its relative simplicity. To sum up, *entropy maximization is one method of finding the most likely trip assignment.*

4.1.2 Estimates from Trip Tables

(a) Logit Assignment

The attractiveness of the entropy function $\chi(\mathbf{h})$ can be demonstrated by an optimization problem proposed in Fisk (1980). Bear in mind that the approach is justified only by the optimum conditions it produces. There are no objections to the feasible set: for any given trip table $\mathbf{x}^D$ the not necessarily unique demand feasible route flows must satisfy the conservation of flows $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$. The objective function, however, is composed of the the sum of the total travel costs $\tilde{g}(\mathbf{h})$, see (3.72) on p. 149, and the negative entropy $\chi(\mathbf{h})$ adjusted by some dispersion parameter θ to be calibrated. There is no immediate economic reason why the network users should try to minimize such a function. Nevertheless, suppose that all travelers together behave as if they do so. The next section hands in some justifications for this approach. Now[5]

$$\min\left\{\tilde{g}(\mathbf{h}) + \frac{1}{\theta}\mathbf{h}^\mathsf{T}\big(\ln \mathbf{h} - \mathbf{1}\big)\,\middle|\, \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0}\right\}, \tag{4.4}$$

for which the resulting KKT conditions give rise to an equilibrated assignment in which flows are shared among sets of paths $\mathcal{R}(s,t)$ between each OD-pair according

[5] Chen, Alfa (1991) apply a method of successive averages (cf. Powell, Sheffi (1982)) as well as improved algorithms to solve this problem.

to a *multinomial logit distribution.*[6] Setting up an appropriate Lagrangean function with Lagrange multipliers $\boldsymbol{\mu}$, the KKT conditions for an optimal traffic assignment $\hat{\mathbf{h}}$ turn out to be

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \frac{1}{\theta} \ln \hat{\mathbf{h}} - \boldsymbol{\Lambda}^{\mathrm{T}} \hat{\boldsymbol{\mu}} \geqq \mathbf{0}, \quad \hat{\mathbf{h}} \geqq \mathbf{0}, \quad \hat{\mathbf{h}}^{\mathrm{T}} \left[\tilde{\mathbf{c}}(\hat{\mathbf{h}}) + \frac{1}{\theta} \ln \hat{\mathbf{h}} - \boldsymbol{\Lambda}^{\mathrm{T}} \hat{\boldsymbol{\mu}} \right] = 0,$$
$$\boldsymbol{\Lambda} \hat{\mathbf{h}} = \mathbf{x}^{D}.$$

For any positive route flow $\hat{h}_\rho > 0$ with $\rho \in \mathcal{R}(s,t)$ one has $\tilde{c}_\rho(\hat{\mathbf{h}}) + \frac{1}{\theta} \ln \hat{h}_\rho = \hat{\mu}_{(s,t)}$. Together with the condition of flow conservation $\sum_{\rho \in \mathcal{R}(s,t)} \hat{h}_\rho = x^D_{(s,t)}$ we obtain similar *path demand functions* for any $\rho \in \mathcal{R}(s,t)$ as before.

$$\hat{h}_\rho = \pi_\rho \, x^D_{(s,t)} \qquad \text{where} \quad \pi_\rho := \frac{\exp[-\theta \tilde{c}_\rho(\hat{\mathbf{h}})]}{\sum_{r \in \mathcal{R}(s,t)} \exp[-\theta \tilde{c}_r(\hat{\mathbf{h}})]} \tag{4.5}$$

Of course, all route flows between the OD-pair (v_s, v_t) sum up to the demand $x^D_{(s,t)}$ such as the path choice probabilities sum up to one $\sum_{r \in \mathcal{R}(s,t)} \pi_\rho = 1$ or in vector notation $\boldsymbol{\Lambda}\, \boldsymbol{\pi} = \mathbf{1}$. The larger $\tilde{c}_\rho$ relative to alternative routes, the smaller is the flow $\hat{h}_\rho$ on that route. The interpretation of θ as *dispersion factor* follows from rewriting (4.5) as

$$\ln(\hat{h}_{\rho_1} / \hat{h}_{\rho_2}) = -\theta \left[\tilde{c}_{\rho_1}(\hat{\mathbf{h}}) - \tilde{c}_{\rho_2}(\hat{\mathbf{h}}) \right] \tag{4.6}$$

for alternative routes $\rho_1 \sim \rho_2$. Increasing θ means that an increasing part of traffic is allocated to the more expensive route.

Similar results are derived in Boyce, Daskin (1997) from a stochastic route choice model with a more appropriate treatment of (conditional) route choice probabilities. In this approach, however, the entropy is introduced as a constraint to the above minimization approach. The authors wish to model a pattern of route choices the dispersion of which must exceed a certain minimum.

The *logit model* (4.5) depends on unknown path choice probability π_ρ. On the estimation of choice probability functions see Domencich, McFadden (1975, Chap. 5). The most important fact is that (4.6) can be completed by

$$\ln(\pi_{\rho_1} / \pi_{\rho_2}) = \ln(h_{\rho_1} / h_{\rho_2})$$

so that all path choice probabilities can be estimated from empirical route cost functions $\tilde{\mathbf{c}}$.

[6] In general the multinomial logit formula is unreasonable for route choice analysis when routes overlap so that too much traffic is allocated to these routes. Instead Daganzo (1979, Chap. 1 and 1982) recommends multinomial probit models to represent stochastic route choice. Unfortunately, the choice function of a multinomial probit model cannot be easily written in closed form, except for very simple examples.

(b) Further Results from Stochastic User Equilibria

The minimum problem (4.4) depends among others on the cost term $\tilde{g}$ having the property $\nabla\tilde{g} = \tilde{\mathbf{c}}$. Such a function does not exist if the Jacobian matrix $\mathbf{J_c} = (\nabla\mathbf{c})^\top$ is asymmetric. Section 3.2.5.2 has shown how to substitute the minimization of $\tilde{g}$ by a corresponding variational inequality. This observation is also valid in the present context, where the stochastic approach of entropy maximization is substituted by a corresponding route choice behavior of trip-makers; see, e.g., Bell, Iida (1997, Chap. 6).

The starting point is a given trip table $\mathbf{x}^D$ and a user equilibrium $\hat{\mathbf{h}}$ at which network users reveal that they are willing to pay at least $\tilde{c}_\rho(\hat{\mathbf{h}})$ for any route ρ with a positive flow. In contrast, travelers expose that they are not willing to pay the corresponding cost for any of the unused routes. Let $\tilde{p}^w_\rho(\mathbf{h})$ denote the willingness to pay for route ρ at the prevailing traffic pattern $\mathbf{h}$. Then, the behavior of network users can be described by

$$\tilde{c}_\rho(\hat{\mathbf{h}}) > \tilde{p}^w_\rho(\hat{\mathbf{h}}) \implies \hat{h}_\rho = 0, \tag{4.7a}$$

$$\tilde{c}_\rho(\hat{\mathbf{h}}) \leqq \tilde{p}^w_\rho(\hat{\mathbf{h}}) \implies \hat{h}_\rho \geqq 0, \tag{4.7b}$$

$$\hat{h}_\rho > 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) \leqq \tilde{p}^w_\rho(\hat{\mathbf{h}}). \tag{4.7c}$$

An immediate consequence is that a positive route flow $\hat{h}_\rho > 0$ is possible, even if ρ is not among the cheapest routes. That is, some drivers are willing to pay more for route ρ than for the cheapest alternative route, $\tilde{p}^w_\rho > p_{(s,t)}$ with $\rho \in \mathcal{R}(s,t)$. The subsequent discussion shows that the following "symmetric" implication *does not hold true*

$$\hat{h}_\rho = 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) \geqq \tilde{p}^w_\rho(\hat{\mathbf{h}}).$$

According to the path demand functions (4.5) of the logit model with positive flows on all routes, the behavior of trip-makers suggests to approximate $\tilde{\mathbf{c}}(\hat{\mathbf{h}})$ by $\tilde{\mathbf{p}}^w(\hat{\mathbf{h}})$ so that the path demand functions take the form $\mathbf{h} = \mathbf{h}^D(\tilde{\mathbf{p}}^w)$. On the assumption that the inverse of these functions exists we obtain functions that describe the willingness to pay for a trip on route ρ, that is in vector notation $\tilde{\mathbf{p}}^w = \tilde{\mathbf{p}}^w(\mathbf{h})$ satisfying $\mathbf{h}^D(\tilde{\mathbf{p}}^w)^\top\tilde{\mathbf{p}}^w = \mathbf{h}^\top\tilde{\mathbf{p}}^w(\mathbf{h})$. The use of path demand functions following from an entropy maximizing approach is the reason to call a path flow $\hat{\mathbf{h}}$ a *stochastic user equilibrium*; it reflects the stochastic aggregate behavior of all travelers.

Regarding the path demand function (4.5), the minimum problem (4.4) can now be expressed by a variational inequality problem

$$\left[\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \tilde{\mathbf{p}}^w(\hat{\mathbf{h}})\right]^\top(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall \mathbf{h} \in H^D. \tag{4.8}$$

This problem will be analyzed with respect to a price sensitive traffic demand in Proposition 4.14 below. For a fixed trip table $\mathbf{x}^D$ the behavior of the trip-makers is reflected by the conditions given in Theorem 3.17 on p. 129, i.e.,

$$\hat{\boldsymbol{\mu}} := \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \tilde{\mathbf{p}}^w(\hat{\mathbf{h}}) + \boldsymbol{\Lambda}^\top\hat{\boldsymbol{\lambda}}: \qquad \hat{\boldsymbol{\mu}} \geqq \mathbf{0}, \quad \hat{\mathbf{h}} \geqq \mathbf{0}, \quad \hat{\boldsymbol{\mu}}^\top\hat{\mathbf{h}} = 0,$$
$$\boldsymbol{\Lambda}\hat{\mathbf{h}} - \mathbf{x}^D = \mathbf{0},$$

where the components of $\hat{\mu}$ and $\hat{\lambda}$ are assigned to the routes and the OD-pairs of the network, respectively. For a positive demand $x^D_{(s,t)}$ there must be at least one positive route flow $\hat{h}_{\rho'}$ with $\rho' \in \mathcal{R}(s,t)$ having the immediate consequence

$$\hat{h}_{\rho'} > 0 \implies \tilde{c}_{\rho'}(\hat{\mathbf{h}}) - \tilde{p}^w_{\rho'}(\hat{\mathbf{h}}) + \hat{\lambda}_{(s,t)} = 0$$

Thus we can expect $\hat{\lambda}_{(s,t)}$ to be non-negative because otherwise

$$\hat{\lambda}_{(s,t)} < 0 \implies \tilde{c}_{\rho}(\hat{\mathbf{h}}) > \tilde{p}^w_{\rho}(\hat{\mathbf{h}}) \qquad \forall\, \rho \in \mathcal{R}(s,t),$$

in which case the drivers are forced to pay more than their willingness to pay due to a fixed demand $x^D_{(s,t)}$. Taking the regularity condition $\hat{\boldsymbol{\lambda}} \geqq \mathbf{0}$ into account, the implications (4.7) follow from

$$\hat{\mu}_\rho := \tilde{c}_\rho(\hat{\mathbf{h}}) - \tilde{p}^w_\rho(\hat{\mathbf{h}}) + \hat{\lambda}_{(s,t)}: \qquad \hat{\mu}_\rho \geqq 0, \quad \hat{h}_\rho \geqq 0, \quad \hat{\mu}_\rho \hat{h}_\rho = 0$$

4.5 Remark Readers who are familiar with linear programming may assume fixed route costs and a fixed willingness to pay for all routes to obtain similar results from the following dual pair of linear programs.

$$\begin{array}{lll} \text{(P)} & \min\ (\tilde{\mathbf{c}} - \tilde{\mathbf{p}}^w)^\mathsf{T}\mathbf{h} & \\ & \text{s.t.} \quad \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D & \\ & \qquad\quad \mathbf{h} \geqq \mathbf{0} & \end{array} \qquad \begin{array}{lll} \text{(D)} & \max\ \mathbf{x}^{D\mathsf{T}}\boldsymbol{\lambda} & \\ & \text{s.t.} \quad \boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\lambda} \leqq \tilde{\mathbf{c}} - \tilde{\mathbf{p}}^w & \\ & \qquad\quad \boldsymbol{\lambda} \text{ unrestricted in sign} & \end{array}$$

Again, it is plausible to suppose that there is at least one route for every OD-pair on which the willingness to pay exceeds the corresponding route cost. The primal objective function then maximizes the positive difference between the aggregate willingness to pay $\tilde{\mathbf{p}}^{w\mathsf{T}}\hat{\mathbf{h}}$ and the total transport cost $\tilde{\mathbf{c}}^\mathsf{T}\hat{\mathbf{h}}$ provided a feasible path flow exists. □

With a stochastic user equilibrium solving (4.8) we know a third traffic assignment besides a deterministic user equilibrium and a deterministic system optimum. Prashker, Bekhor (2000) compare these results with respect to different levels of congestion. Their numerical examples reflect what is (at least partially) expected by intuition. For uncongested networks, the user equilibrium and the system optimum are similar. This observation can be explained by negligible external effects. In this case a stochastic assignment performs worse than a deterministic assignment because the traffic is distributed among more expensive routes although not needed. For moderately congested networks the stochastic assignment performs better than the deterministic counterparts. In this range the stochastic effects "balance" in some way the congestion effects. For highly congested situations, the deterministic and stochastic user equilibrium become very similar because both solutions show a highly dispersed traffic flow pattern.

4.6 Remark Stochastic traffic assignment has been justified by the fact that travelers have imperfect information on the network. Another justification for entropy maximization is that deterministic user equilibria (with full information for all travelers) are not necessarily unique. In this case the set $\widehat{H}$ of all route flows solving

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}(\mathbf{h}-\hat{\mathbf{h}}) \geqq 0 \qquad \forall \mathbf{h} \in H^D$$

includes more than one element. Then, the entropy maximization

$$\min\left\{\frac{1}{\theta}\mathbf{h}^{\mathsf{T}}\big(\ln \mathbf{h}-\mathbf{1}\big)\,\Big|\,\mathbf{h}\in \widehat{H}\right\} \tag{4.9}$$

seeks for the most likely route flow in the class of all user equilibria. A similar problem will be discussed in Remark 4.7. □

4.1.3 Estimates from Link Flows

(a) Most Likely Path Flows

In the preceding section the trip table $\mathbf{x}^D$ was assumed to be known. From the viewpoint of data collection an even more realistic situation is that we observe some arc flow $\hat{\boldsymbol{\varphi}}$ which gives implicitly some deterministic user equilibrium for an unknown trip table $\mathbf{x}^D$.

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^{\mathsf{T}}(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}}) \geqq 0 \qquad \forall \boldsymbol{\varphi} \in \Phi^D$$

That is, $\hat{\boldsymbol{\varphi}}$ satisfies particularly the two constraints $\hat{\boldsymbol{\varphi}} = \boldsymbol{\Delta}\,\mathbf{h}$ and $\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D$. As a path flow $\mathbf{h}$ compatible with $\hat{\boldsymbol{\varphi}}$ is not necessarily unique, Bell, Iida (1997, p. 100) ask for the *most likely path flows* by solving the entropy maximizing problem[7]

$$\min\big\{\mathbf{h}^{\mathsf{T}}\big(\ln(\mathbf{h})-\mathbf{1}\big)\,|\,\hat{\boldsymbol{\varphi}} = \boldsymbol{\Delta}\,\mathbf{h},\ \mathbf{h}\geqq \mathbf{0}\big\}. \tag{4.10}$$

Let us assume for the sake of simplicity that the most likely path flow $\hat{\mathbf{h}}$ is positive with regard to the given flow observation $\hat{\boldsymbol{\varphi}}$. This is plausible in the light of Stirling's approximation of the entropy. With that the KKT conditions for an optimal solution yield

$$\ln \hat{\mathbf{h}} = \boldsymbol{\Delta}^{\mathsf{T}}\hat{\boldsymbol{\mu}} \quad \text{or} \quad \ln \hat{h}_\rho = \sum_{e\in\rho}\hat{\mu}_e \quad \forall \rho,$$

and

$$\boldsymbol{\Delta}\,\hat{\mathbf{h}} = \hat{\boldsymbol{\varphi}},$$

where $\hat{\mu}_e$, $e \in \mathcal{E}$, is the optimal value of the Lagrange multiplier assigned to link e with an observed flow $\hat{\varphi}_e$. Having solved the equivalent system

$$\sum_{\rho:\, e'\in\rho} \exp\left(\sum_{e\in\rho}\hat{\mu}_e\right) = \hat{\varphi}_{e'} \qquad \forall e' \in \mathcal{E}$$

[7] A similar formulation can be found in Bell (1995). The constraint, however, is based on a direct relationship between link flows and the OD-trip matrix, i.e., $\hat{\boldsymbol{\varphi}} = \boldsymbol{\Xi}\,\mathbf{x}^D$, where each element of the matrix $\boldsymbol{\Xi}$ denotes the proportion of the corresponding OD-flow using the link considered.

we know also an entropy maximizing path flow $\hat{\mathbf{h}}$ which in turn reveals a trip table $\hat{\mathbf{x}}^D = \mathbf{\Lambda}\hat{\mathbf{h}}$.

A comparison of (4.9) and (4.10) indicates that an entropy maximizing route flow $\hat{\mathbf{h}}$ with respect to (4.10) ignores any information on travel costs. Although $\hat{\mathbf{h}}$ is one among possibly many route flows satisfying $\hat{\boldsymbol{\varphi}} = \mathbf{\Delta}\,\mathbf{h}$, there is no guarantee that $\hat{\mathbf{h}}$ is also user optimal with respect to $\hat{\mathbf{x}}^D$ as required in (4.9). This observation is not trivial because $\mathbf{c}(\hat{\boldsymbol{\varphi}})^\top\hat{\boldsymbol{\varphi}} = \tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top\hat{\mathbf{h}}$ holds always true by definition. The decisive fact is that $\hat{\boldsymbol{\varphi}}$ is assumed to be a user equilibrium for an unknown trip table $\mathbf{x}^D$, but there is no need that the traffic demand $\hat{\mathbf{x}}^D$ determined by $\hat{\mathbf{h}}$ coincides with that vector. The next section overcomes this problem under the supposition that we can observe not only the flow pattern $\hat{\boldsymbol{\varphi}}$, but also the cost structure $\mathbf{c}(\hat{\boldsymbol{\varphi}})$.

4.7 Remark (Most Likely Trip Matrix) Suppose that we observe the link flow $\hat{\boldsymbol{\varphi}}$ and that the route cost function $\tilde{\mathbf{c}}$ is given in advance. If $\hat{\boldsymbol{\varphi}}$ is a user optimized flow pattern, what is the most likely trip matrix? Following Fisk (1988), let us restrict the maximum entropy *trip matrix estimation* to user optimal traffic assignments. Then, the problem to be solved is

(4.11)
$$\min_{\mathbf{h},\mathbf{x}^D}\left\{\mathbf{x}^{D\top}(\ln\mathbf{x}^D - \mathbf{1}) \mid \mathbf{\Delta}\,\mathbf{h} = \hat{\boldsymbol{\varphi}},\ \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \tilde{\mathbf{c}}(\mathbf{h})^\top(\mathbf{z} - \mathbf{h}) \geqq 0\ \forall\mathbf{z} \in H^D\right\},$$

where the last constraint indicates user optimality. Notice that this problem assumes all link flows being observable. Fisk (1988), however, refers to a situation where only a subset of link flows is known. Recall that a unique link flow $\hat{\boldsymbol{\varphi}}$ can be compatible with many route flows $\mathbf{h}$ each of which can determine another trip matrix $\mathbf{x}^D$. For further approaches on the estimation of trip matrices from link traffic count data see Fisk, Boyce (1983b). □

(b) Cost Preserving Trip Tables

Given the link flow $\hat{\boldsymbol{\varphi}}$ and the link costs $\mathbf{c}(\hat{\boldsymbol{\varphi}})$, what do we know about the underlying trip table $\mathbf{x}^D$ being consistent with a Wardrop equilibrium? This problem has been introduced in Nguyen (1977) and later it was refined by Sherali et al. (1994). In order to answer this question let us abstain from finding an entropy maximizing path flow pattern, see (4.10), or trip matrix, see (4.11). Again all feasible route flows $\mathbf{h}$ must satisfy $\mathbf{\Delta}\,\mathbf{h} = \hat{\boldsymbol{\varphi}}$ and every feasible $\mathbf{h}$ determines its own trip table by $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$. Furthermore, any Wardrop equilibrium reflects the behavior of the users to choose one among the cheapest available routes. Computing route costs $\tilde{\mathbf{c}} = \mathbf{\Delta}^\top\mathbf{c}(\hat{\boldsymbol{\varphi}})$, the minimum route cost for every OD-pair follows from $p_{(s,t)} = \min_{\rho\in\mathcal{R}(s,t)}\tilde{c}_\rho$ so that all routes of potential usage are known. For computational purposes define a vector $\tilde{\mathbf{c}}^*$ having the components $\tilde{c}^*_\rho = \tilde{c}_\rho$ if ρ is one of the cost minimal routes and otherwise $\tilde{c}^*_\rho = \mu\tilde{c}_\rho$, where μ is a sufficiently large constant. In doing so the linear program

$$\min\left\{\tilde{\mathbf{c}}^{*\top}\mathbf{h} \mid \hat{\boldsymbol{\varphi}} = \mathbf{\Delta}\,\mathbf{h},\ \mathbf{h} \geqq \mathbf{0}\right\}$$

provides a not necessarily unique optimal solution $\hat{\mathbf{h}}$ that has several interesting properties. First of all it reproduces the observed link flows $\hat{\boldsymbol{\varphi}} = \mathbf{\Delta}\,\hat{\mathbf{h}}$ and it estimates

a trip matrix $\mathbf{x}^D = \mathbf{\Lambda}\,\hat{\mathbf{h}}$. Due to the previously defined constant μ, a route flow $\hat{h}_\rho$ is only positive when ρ is a cost minimal connection of the respective OD-pair. All other routes are not used so that $\hat{\mathbf{h}}$ constitutes a Wardrop equilibrium. The same reasoning shows $\tilde{\mathbf{c}}^{*\top}\hat{\mathbf{h}} = \tilde{\mathbf{c}}^{\top}\hat{\mathbf{h}} = \mathbf{c}(\hat{\boldsymbol{\varphi}})^{\top}\hat{\boldsymbol{\varphi}}$. That is, the flow $\hat{\mathbf{h}}$ reproduces the observed total travel cost. Recall again that this approach is not able to discriminate between alternative solutions to the minimum problem. One way out would be to solve (4.9) that is to pick the most likely path flow which is consistent with a user equilibrium. For further modifications needed to accommodate inconsistent flow data see Sherali et al. (1994).

Finally, note that the estimates for the path demand as well as the trip table are also valid for a price sensitive traffic demand. Computing $\mathbf{x}^D = \mathbf{\Lambda}\hat{\mathbf{h}}$ then constitutes one point of the demand function $\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p})$ that will be introduced in Section 4.2, where $\mathbf{p}$ is as usual the vector of minimum travel cost with respect to all OD-pairs.

4.1.4 Doubly Constrained Models

(a) Symmetric Cost Structure

In what follows link flows, route flows as well as the trip table are unknown. Instead the origin totals $\boldsymbol{\varphi}^+$ and destination totals $\boldsymbol{\varphi}^-$ can be observed, but the matching of the trip destinations with the trip origins, i.e., the trip distribution $\mathbf{x}^D$ is allowed to be influenced by network conditions. This is the case of *combined trip distribution and traffic assignment.* The two most important restrictions for that setting have been noted in

$$\varphi_s^+ = \sum_j x^D_{(s,j)} \quad \forall s \in \mathcal{V} \quad \text{and} \quad \varphi_t^- = \sum_i x^D_{(i,t)} \quad \forall t \in \mathcal{V} \tag{2.26}$$

or, equivalently, in terms of route flows

$$\boldsymbol{\varphi}^+ = \mathbf{O}\,\mathbf{h} \quad \text{and} \quad \boldsymbol{\varphi}^- = \mathbf{D}\,\mathbf{h}. \tag{2.31}$$

These two conditions are the reason to refer to models of that type as doubly constrained; cf., e.g., Sheffi (1985). As the set Φ^D of feasible link flows $\boldsymbol{\varphi}$ depends on $\mathbf{x}^D$, it is useful to introduce

$$\Phi' := \left\{(\boldsymbol{\varphi}, \mathbf{x}^D) \mid \mathbf{\Delta}\mathbf{h} = \boldsymbol{\varphi},\ \mathbf{\Lambda}\mathbf{h} = \mathbf{x}^D,\ \mathbf{O}\,\mathbf{h} = \boldsymbol{\varphi}^+,\ \mathbf{D}\,\mathbf{h} = \boldsymbol{\varphi}^-,\ \mathbf{h} \geqq \mathbf{0}\right\}$$

including $\boldsymbol{\varphi}^+$ and $\boldsymbol{\varphi}^-$ as given parameters. The task is now to find the most likely trip distribution $\mathbf{x}^D$ together with a compatible traffic assignment $\boldsymbol{\varphi}$ or $\mathbf{h}$ subject to fixed origin totals $\boldsymbol{\varphi}^+$ and destination totals $\boldsymbol{\varphi}^-$.

Again entropy maximization starts with the number of permutations of trips across all OD-pairs

$$W(\mathbf{x}^D) = \frac{\varphi^{\#}!}{\prod_{(s,t)} x^D_{(s,t)}!},$$

where $\varphi^{\#} = \sum_s \varphi_s^+ = \sum_t \varphi_t^- = \sum_s \sum_t x_{(s,t)}^D$ is the total number of trips. Hence, granting that every $x_{(s,t)}^D$ is sufficiently great, Stirling's approximation yields by the same reasoning as before

$$\chi(\mathbf{x}^D) \approx -\mathbf{x}^{D\mathsf{T}}\big(\ln(\mathbf{x}^D) - \mathbf{1}\big), \tag{4.12}$$

cf., e.g., Wilson, Senior (1974). Bear in mind that the gradient of this entropy function turns out to be $-\ln(\mathbf{x}^D)$. In accordance with (4.4) the most likely trip table together with a stochastic user optimal link flow solves[8]

$$\min\left\{g(\boldsymbol{\varphi}) + \mathbf{x}^{D\mathsf{T}}\big(\ln(\mathbf{x}^D) - \mathbf{1}\big) \mid (\boldsymbol{\varphi}, \mathbf{x}^D) \in \Phi'\right\}, \tag{4.13}$$

where g denotes the total travel cost as defined in (3.67) having the link costs $\mathbf{c}(\boldsymbol{\varphi})$ as gradient. This integrability conditions requires a symmetric matrix $\nabla\mathbf{c}$. Problems of that type go back at least to Florian et al. (1975) and Evans (1976, p. 43)[9] and can equivalently be expressed in terms of path flows eliminating link flows by $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$. To be more concrete the corresponding Lagrangean function is

$$\mathcal{L}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \tilde{g}(\mathbf{h}) + \mathbf{x}^{D\mathsf{T}}\big(\ln(\mathbf{x}^D) - \mathbf{1}\big) + \boldsymbol{\lambda}^{\mathsf{T}}\left(\boldsymbol{\varphi}^+ - \mathbf{O}\,\mathbf{h}\right) + \boldsymbol{\mu}^{\mathsf{T}}\left(\boldsymbol{\varphi}^- - \mathbf{D}\,\mathbf{h}\right),$$

where the trip table must hold $\mathbf{x}^D = \boldsymbol{\Lambda}\,\mathbf{h}$. Regarding the subsequent KKT conditions we need the gradient of the objective function with respect to $\mathbf{h}$, that is

$$\tilde{\mathbf{c}}(\mathbf{h}) + \boldsymbol{\Lambda}^{\mathsf{T}} \underbrace{\nabla\big(\mathbf{x}^{D\mathsf{T}}\big(\ln(\mathbf{x}^D) - \mathbf{1}\big)\big)}_{=\ln(\mathbf{x}^D)=\ln(\boldsymbol{\Lambda}\mathbf{h})} = \tilde{\mathbf{c}}(\mathbf{h}) + \boldsymbol{\Lambda}^{\mathsf{T}} \ln(\mathbf{x}^D)$$

with the ρ^{th}-element

$$\tilde{c}_\rho(\mathbf{h}) + \ln x_{(s,t)}^D \quad \text{for} \quad \rho \in \mathcal{R}(s,t).$$

The corresponding KKT conditions require for every route $\rho \in \mathcal{R}(s,t)$ and every OD-pair

$$\tilde{c}_\rho(\hat{\mathbf{h}}) + \ln \hat{x}_{(s,t)}^D - \hat{\lambda}_s - \hat{\mu}_t \geqq 0, \quad \hat{h}_\rho \geqq 0, \quad \hat{h}_\rho\left[\tilde{c}_\rho(\hat{\mathbf{h}}) + \ln \hat{x}_{(s,t)}^D - \hat{\lambda}_s - \hat{\mu}_t\right] = 0.$$

On the one hand, a positive route flow requires

$$\hat{h}_\rho > 0 \implies \hat{x}_{(s,t)}^D = \exp(-\tilde{c}_\rho(\hat{\mathbf{h}}) + \hat{\lambda}_s + \hat{\mu}_t). \tag{4.14}$$

On the other hand, all traffic is assigned to cheapest routes.

$$\rho_1 \sim \rho_2,\ \tilde{c}_{\rho_1}(\hat{\mathbf{h}}) < \tilde{c}_{\rho_2}(\hat{\mathbf{h}}) \implies \hat{h}_{\rho_2} = 0$$

[8] Erlander (1980, Chap. 6) discusses a corresponding entropy constrained problem as linear program. Moreover, LeBlanc, Abdulaal (1982) suggest a similar model taking also mode choice into account.

[9] Evans (1976) presented a converging algorithm that solves this problem. Florian et al. (1975) suggested another method which is an adaption of the Frank and Wolfe algorithm.

For the sake of notational convenience let $p_{(s,t)}(\hat{\mathbf{h}})$ denote the minimum travel cost between a source v_s and a destination v_t prevailing at the optimal route flow $\hat{\mathbf{h}}$. For a positive traffic demand $\hat{x}^D_{(s,t)}$ there must be at least one positive route flow joining the OD-pair at hand such that

$$-p_{(s,t)}(\hat{\mathbf{h}}) = \ln \hat{x}^D_{(s,t)} - \hat{\lambda}_s - \hat{\mu}_t \tag{4.15}$$

and every least cost route $\rho \in \mathcal{R}(s,t)$ with $\tilde{c}_\rho(\hat{\mathbf{h}}) = p_{(s,t)}(\hat{\mathbf{h}})$ shows a non-negative flow $\hat{h}_\rho$. In other words we have for any $\rho \in \mathcal{R}(s,t)$

$$\tilde{c}_\rho(\hat{\mathbf{h}}) \geqq p_{(s,t)}(\hat{\mathbf{h}}), \qquad \hat{h}_\rho \geqq 0, \qquad \hat{h}_\rho\big(\tilde{c}_\rho(\hat{\mathbf{h}}) - p_{(s,t)}(\hat{\mathbf{h}})\big) = 0.$$

Contrasted with the path demand functions of Section 4.1.2, rearrange (4.15) to obtain *OD-demand functions*

$$\hat{x}^D_{(s,t)} = R_s\, S_t \exp(-p_{(s,t)}(\hat{\mathbf{h}})), \tag{4.16}$$

where $R_s := \exp \hat{\lambda}_s$ and $S_i := \exp \hat{\mu}_i$ are the so-called balancing factors. The last step is to compute the Lagrange multipliers $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\mu}}$. For that reason the system (4.16) can be substituted into (2.26) such that

$$\begin{aligned} \varphi^+_s &= R_s \sum_{t \in \mathcal{V}} S_t \exp(-p_{(s,t)}(\hat{\mathbf{h}})) \quad \forall s \in \mathcal{V} \qquad \text{(outflows)}, \\ \varphi^-_t &= S_t \sum_{s \in \mathcal{V}} R_s \exp(-p_{(s,t)}(\hat{\mathbf{h}})) \quad \forall t \in \mathcal{V} \qquad \text{(inflows)}. \end{aligned} \tag{4.17}$$

Notice that one of these equations is redundant due to Kirchhoff's law, $\sum_s \varphi^+_s = \sum_t \varphi^-_t = \varphi^\#$.

4.8 Remark The entropy in (4.13) is usually divided by a dispersion parameter θ to be estimated. The system (4.16) then takes the form

$$\hat{x}^D_{(s,t)} = R_s\, S_t \exp(-\theta p_{(s,t)}) \quad \text{with} \quad R_s := \exp(\theta \hat{\lambda}_s) \text{ and } S_t := \exp(\theta \hat{\mu}_t).$$

Redefining R_s and S_t yields the classical solution

$$\hat{x}^D_{(s,t)} = A_s B_t\, \varphi^+_s\, \varphi^-_t \exp(-\theta p_{(s,t)}) \quad \text{with} \quad A_s := R_s/\varphi^+_s \text{ and } B_t := S_t/\varphi^-_t.$$

Demand functions of that type have been derived in Wilson (1967, 1970) by maximizing the entropy of the flow pattern subject to appropriate constraints. The exponential term indicates a deterrence function for the OD-pair (v_s, v_t). The "potentials" of the originating node and the terminating node are represented by φ^+_s and φ^-_t, respectively. Eventually, A_s may be interpreted as a measure for the importance of the origin v_s relative to all receiving nodes, i.e.,

$$A_s = \Big[\sum_{t \in \mathcal{V}} S_t \exp(-p_{(s,t)}(\hat{\mathbf{h}}))\Big]^{-1}.$$

A similar reasoning may be applied to B_t as a measure for the relative importance of the destination node v_t. □

(b) Asymmetric Cost Structure

The results so far obtained are based upon the assumption that the cost term g does exist, which requires the matrix $\nabla \mathbf{c}$ to be symmetric. In what follows we relax this supposition and provide a variational inequality formulation of the convex minimum problem (4.13). A similar approach can be found in Florian et al. (1999), however, their multi-class multi-purpose approach is ignored here for the sake of clarity. In accordance with Theorem 3.23, we make use of the gradient of the negative entropy, i.e., $\ln(\hat{\mathbf{x}}^D)$, to obtain

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^\top(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}})+\frac{1}{\theta}(\ln \hat{\mathbf{x}}^D)^\top(\mathbf{x}^D-\hat{\mathbf{x}}^D)\geqq 0 \quad \forall\,(\boldsymbol{\varphi},\mathbf{x}^D)\in\Phi',$$

where θ is the dispersion parameter mentioned in the previous remark. In order to verify that the corresponding optimum conditions reveal the same demand functions as have been derived in (4.16) with respect to the symmetric case, we eliminate $\boldsymbol{\varphi}$ and $\mathbf{x}^D$ to obtain

$$\mathbf{c}(\boldsymbol{\Delta}\hat{\mathbf{h}})^\top\boldsymbol{\Delta}(\mathbf{h}-\hat{\mathbf{h}})+\frac{1}{\theta}(\ln \boldsymbol{\Lambda}\hat{\mathbf{h}})^\top\boldsymbol{\Lambda}(\mathbf{h}-\hat{\mathbf{h}})\geqq 0 \qquad \forall \mathbf{h}\colon \mathbf{O}\mathbf{h}=\boldsymbol{\varphi}^+,\ \mathbf{D}\mathbf{h}=\boldsymbol{\varphi}^-,\ \mathbf{h}\geqq\mathbf{0}.$$

Rewriting the inequality as

$$\Big[\tilde{\mathbf{c}}(\hat{\mathbf{h}})+\frac{1}{\theta}\boldsymbol{\Lambda}^\top(\ln \boldsymbol{\Lambda}\,\hat{\mathbf{h}})\Big]^\top(\mathbf{h}-\hat{\mathbf{h}})\geqq 0,$$

Theorem 3.17 implies the following conditions for an optimal path flow $\hat{\mathbf{h}}$:

$$\hat{\boldsymbol{\alpha}}:=\tilde{\mathbf{c}}(\hat{\mathbf{h}})+\frac{1}{\theta}\boldsymbol{\Lambda}^\top(\ln \boldsymbol{\Lambda}\hat{\mathbf{h}})-\mathbf{O}^\top\hat{\boldsymbol{\lambda}}-\mathbf{D}^\top\hat{\boldsymbol{\mu}}\colon \qquad \hat{\boldsymbol{\alpha}}\geqq\mathbf{0},\quad \hat{\mathbf{h}}\geqq\mathbf{0},\quad \hat{\boldsymbol{\alpha}}^\top\hat{\mathbf{h}}=0$$
$$\mathbf{O}\hat{\mathbf{h}}=\boldsymbol{\varphi}^+,\quad \mathbf{D}\hat{\mathbf{h}}=\boldsymbol{\varphi}^-$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are multipliers associated with sources and sinks, respectively. A positive flow on some route $\rho\in\mathcal{R}(s,t)$ therefore implies

$$\hat{h}_\rho>0 \implies \hat{\alpha}_\rho=\tilde{c}_\rho+\frac{1}{\theta}x^D_{(s,t)}-\hat{\lambda}_s-\hat{\mu}_t=0$$

and that corresponds to the OD-demand function of (4.16).

An extended version of the above concept is provided in Friesz et al. (1986). The relevant idea is to analyze an overall network consisting of subnetworks with each being provided by a different carrier. The total network is then used by several shippers. For the sake of clarity we assume here that there is only one transport mode and every shipper s produces the same commodity at its production site v_s, where p^S_s denotes the respective producer price, i.e., $\mathbf{p}^S$ in vector notation. Given the demand $\boldsymbol{\varphi}^-$ for all consumer destinations, the demand for traffic results mainly from a production vector $\boldsymbol{\varphi}^+$ to be determined. Note that every shipper pursues the task of minimizing the final delivered price (p^S_s plus transport cost) by an appropriate route

choice. The stochastic behavior of competing shippers can now be derived from a variational inequality of the form

$$\underbrace{\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}(\mathbf{h}-\hat{\mathbf{h}})}_{\text{transport cost}}+\underbrace{\mathbf{p}^{S\mathsf{T}}(\boldsymbol{\varphi}^{+}-\hat{\boldsymbol{\varphi}}^{+})}_{\text{cost of production}}+\frac{1}{\theta}\underbrace{(\ln\hat{\mathbf{x}}^{D})^{\mathsf{T}}(\mathbf{x}^{D}-\hat{\mathbf{x}}^{D})}_{\text{entropy}}\geqq 0\quad\forall(\mathbf{h},\mathbf{x}^{D},\boldsymbol{\varphi}^{+})\in H',$$

where H' is derived from Φ' as follows

$$H':=\left\{(\mathbf{h},\mathbf{x}^{D},\boldsymbol{\varphi}^{+})|\ \boldsymbol{\Lambda}\mathbf{h}=\mathbf{x}^{D},\ \mathbf{O}\mathbf{h}=\boldsymbol{\varphi}^{+},\ \mathbf{D}\mathbf{h}=\boldsymbol{\varphi}^{-},\ \mathbf{h}\geqq\mathbf{0}\right\}.$$

The same steps as before yield now

$$\left[\tilde{\mathbf{c}}(\hat{\mathbf{h}})+\mathbf{O}^{\mathsf{T}}\mathbf{p}^{S}+\frac{1}{\theta}\boldsymbol{\Lambda}^{\mathsf{T}}(\ln\boldsymbol{\Lambda}\hat{\mathbf{h}})\right]^{\mathsf{T}}(\mathbf{h}-\hat{\mathbf{h}})\geqq 0,\quad\forall\mathbf{h}\colon\mathbf{O}\mathbf{h}=\boldsymbol{\varphi}^{+},\ \mathbf{D}\mathbf{h}=\boldsymbol{\varphi}^{-},\ \mathbf{h}\geqq\mathbf{0}$$

with the following conditions for an optimal path flow $\hat{\mathbf{h}}$.

$$\hat{\boldsymbol{\alpha}}:=\tilde{\mathbf{c}}(\hat{\mathbf{h}})+\mathbf{O}^{\mathsf{T}}\mathbf{p}^{S}+\frac{1}{\theta}\boldsymbol{\Lambda}^{\mathsf{T}}(\ln\boldsymbol{\Lambda}\hat{\mathbf{h}})-\mathbf{O}^{\mathsf{T}}\hat{\boldsymbol{\lambda}}-\mathbf{D}^{\mathsf{T}}\hat{\boldsymbol{\mu}}\colon\quad\hat{\boldsymbol{\alpha}}\geqq\mathbf{0},\quad\hat{\mathbf{h}}\geqq\mathbf{0},\quad\hat{\boldsymbol{\alpha}}^{\mathsf{T}}\hat{\mathbf{h}}=0$$
$$\mathbf{O}\hat{\mathbf{h}}=\hat{\boldsymbol{\varphi}}^{+},\quad\mathbf{D}\hat{\mathbf{h}}=\boldsymbol{\varphi}^{-}$$

Observe that $\hat{\mathbf{h}}$ determines a production vector $\hat{\boldsymbol{\varphi}}^{+}$, a trip matrix $\hat{\mathbf{x}}^{D}$, as well as a link flow $\hat{\boldsymbol{\varphi}}$. A positive flow on route $\rho\in\mathcal{R}(s,t)$ requires particularly

$$\hat{h}_{\rho}>0\implies\hat{\alpha}_{\rho}=\tilde{c}_{\rho}(\hat{\mathbf{h}})+p_{s}^{S}+\frac{1}{\theta}x_{(s,t)}^{D}-\hat{\lambda}_{s}-\hat{\mu}_{t}=0.$$

Transportation demand functions thus take the form

$$x_{(s,t)}^{D}=\exp(\theta\hat{\lambda}_{s})\,\exp(\theta\hat{\mu}_{t})\,\exp[-\theta\,(\tilde{c}_{\rho}(\hat{\mathbf{h}})+p_{s}^{S})]\quad\text{with}\quad\rho\in\mathcal{R}(s,t),$$

where $\tilde{c}_{\rho}(\hat{\mathbf{h}})+p_{s}^{S}$ is the final delivered price at consumer destination v_t if the commodity is produced at site v_s and shipped along route ρ. For further details with respect to many commodities and many transport modes see Friesz et al. (1986).

Having found an optimal path flow $\hat{\mathbf{h}}$, which is interpreted as optimal routing of *shippers*, a second model is used to describe competing *carriers*, each of which controls a specific subnetwork. All carriers have the aim to serve the given demand which is assigned by shippers to the specific subnet at minimum transport cost; to avoid complicated notations recall (3.56) on p. 140 and think of it as shipper specific problem to be solved. This is an appealing approach because carriers are interested in a system optimal usage of their own network while the shippers' demand is user optimized.

4.9 Remark The basic idea of shipper-carrier networks is also valid for (capacitated) telecommunications networks. For example, Faragó et al. (1995) describe the physical infrastructure of a carrier on which several *logical* or *virtual subnetworks* (for instance, one for each type of traffic) can coexist, sharing the same physical transmission and switching capacities. As the carrier incurs only fixed network costs

but no transport cost, he pursues the task of revenue maximization by (a) assigning physical capacities to virtual subnetworks and (b) by sharing the load of each virtual network among routes that connect the same endpoints. These decisions together with a stochastic arrival rate to routes and a stochastic holding time of calls determine end-to-end blocking probabilities for traffic, which must be taken into account when computing the total network revenue. □

(c) Gravity Models

In models of competitive retail facility location, customers can choose between several available facilities. Hotelling (1929) proposed that customers select the closest store. However, in reality a proportion of customers will not necessarily select the nearest shop. Huff (1966) proposed that the probability for selecting a store by some customer is proportional to the store's area (or potential) and inversely proportional to a power of the distance to that store. This rule is the basis of gravity models; they combine the attractiveness of retail facilities with some distance decay functions in order to explain spatial interaction patterns. For further details see, e.g., Alonso (1987).

The basic gravity hypothesis asserts a multiplicative relationship between mean interaction frequencies and the effects of origin, destination, and separation attributes. In the present framework the stochastic user traffic assignment generates a gravity model consisting of OD-demand functions as given in Remark 4.8. Picking out one OD-pair (v_s, v_t) the interaction between the respective nodes (or regions) is described by

$$\hat{x}^D_{(s,t)} = R_s\, S_t\, \exp(-\theta p_{(s,t)}).$$

Here the *exponential deterrence function* $\exp(-\theta p_{(s,t)})$ includes a distance sensitivity parameter θ that can also be interpreted as dispersion factor; see (4.6). Of course, the demand for traffic is a decreasing function of travel costs $p_{(s,t)}$, where travel costs indicate a measure for the spatial separation between two regions, say s and t. The meaning of R_s and S_t can be explained best by expressing inflows and outflows in accordance with (4.17). We can then compute the fractional parts[10]

$$\frac{\hat{x}^D_{(s,t)}}{\varphi^+_s} = \frac{S_t \exp(-p_{(s,t)})}{\sum_{j\in\mathcal{V}} S_j \exp(-p_{(s,j)})} \quad \text{and} \quad \frac{\hat{x}^D_{(s,t)}}{\varphi^-_t} = \frac{R_s \exp(-p_{(s,t)})}{\sum_{i\in\mathcal{V}} R_i \exp(-p_{(i,t)})}.$$

For two destinations t_1 and t_2 with the same evaluated distance from source s, that is $p_{(s,t_1)} = p_{(s,t_2)}$, one has

$$\frac{\hat{x}^D_{(s,t_1)}}{\hat{x}^D_{(s,t_2)}} = \frac{S_{t_1}}{S_{t_2}}.$$

In this sense S_t indicates the *attraction* of a target area t *relative* to other regions with the same distance. Similarly, two origins s_1 and s_2 with the same distance to a destination t, that is $p_{(s_1,t)} = p_{(s_2,t)}$, satisfy

[10] In the case where $x^D_{(s,t)}$ denotes an interaction probability, $\hat{x}^D_{(s,t)}/\varphi^+_s$ and $\hat{x}^D_{(s,t)}/\varphi^-_t$ represent the conditional origin and destination probabilities, respectively.

$$\frac{\hat{x}^D_{(s_1,t)}}{\hat{x}^D_{(s_2,t)}} = \frac{R_{s_1}}{R_{s_2}}.$$

In this sense R_s measures the *relative importance* of a source node s for a target area t. On further properties of gravity models and methods for estimating the parameters (R_v, S_v, θ for all $v \in \mathcal{V}$) see Sen, Smith (1995).

4.10 Summary Section 3.2 deals with the cost minimal assignment of a fixed trip table. In doing so we have started with an uncapacitated network that has constant link costs per flow unit. In this setting there is no difference between system and user optima because marginal route costs coincide with route costs per trip. This is distinct for capacitated networks so that congestion takes place. As motorists tend to ignore their disturbing impacts onto the rest of the network, the realized trip assignment in accordance with Wardrop's first principle deviates from a system optimal assignment that takes external costs into consideration. In the tradition of Pigou it has been shown that road pricing is in principle able to overcome this deficiency. One fundamental observation is to distinguish between link and path flows. While demand refers to trips, the supply side is characterized by capacities that are related to flows on roads.

The preceding results refer to the case where all trip-makers behave together as if they would minimize some superordinated artificial objective function. Without the implicit integrability condition there is no need that such a function exists and further analysis is required to describe user optima. Variational inequalities reveal not only the behavior of network users but can also be used to examine existence and uniqueness of equilibria. Such variational inequality formulations are a powerful tool that can also be used to capture more general traffic assignment problems. The subsequent analysis supplements the deterministic traffic assignment with elements of stochastic route choice. Based on different traffic counts, entropy maximizing problems then yield the most likely path flows. Again more general results can be obtained by making use of variational inequality formulations. □

4.2 Traffic Assignment with Price Sensitive Demand

4.2.1 Price Dependent Demand for Traffic

Following Beckmann, McGuire (1953) we have to distinguish two aspects, namely, that of *traffic diversion* and *traffic diminution* (or *creation*). The first problem is concerned with rerouting a fixed volume of traffic. The second issue takes into account that tolls reduce the demand for traffic but it essentially ignores the traffic assignment problem. Recall the example of the City of London. The congestion charge has reduced traffic volumes by 20% within the eight-square-mile downtown zone. The remaining traffic now can be reorganized under less restrictive traffic conditions.[11] With regard to our analytical framework the formerly fixed traffic demand, i.e., the trip table $\mathbf{x}^D$, will now be described by price sensitive demand

[11] The primary benefit of congestion pricing may not be to reduce traffic flows. In many situations in which such pricing is urgently needed, the effect is intended to increase

functions that influence three decisions at the same time, cf. Patriksson (1994): the number of trips generated at the zones (travel choice), the destination of generated trips (trip distribution) and the route choice (user optimal trip assignment). Models of this type are characterised by the fact that network congestion effects are taken implicitly into consideration in the travel choice model: when travel costs increase due to congestion, the demand for the traffic concerned is typically reduced. At equilibrium, the travel costs are those produced by a user optimal allocation of the travel demand onto the network; cf., e.g., Fisk, Boyce (1983a).

The elements of the enhanced approach may be seen best when going back to the user optimal traffic assignment problems (3.69), p. 147, or (3.72), p. 149. Congestion effects are incorporated by travel cost functions $\mathbf{c}$ or $\tilde{\mathbf{c}}$ and the conservation of flow constraint is the same as before. All that must be added to both of the problems are appropriate OD-demand functions relating the trip table $\mathbf{x}^D$ to trip prices $\mathbf{p}$ the travelers have to pay for journeys from any origin to any destination. To sum up we obtain the following list of ingredients.

1. The *conservation of flows* follows from the network model which is encoded in the incidence matrices $\mathbf{\Delta}$ and $\mathbf{\Lambda}$. As before link flows $\boldsymbol{\varphi}$, route flows $\mathbf{h}$, and the generated traffic $\mathbf{x}^S$ must satisfy

$$\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h} \quad \text{and} \quad \mathbf{x}^S = \mathbf{\Lambda}\,\mathbf{h}.$$

2. The *link performance* or *link cost functions* $\mathbf{c}$ reflect congestion effects depending on the prevailing traffic load $\boldsymbol{\varphi}$. As trip-makers decide on the basis of route cost $\tilde{\mathbf{c}}$, we use again

$$\tilde{\mathbf{c}}(\mathbf{h}) = \mathbf{\Delta}^{\mathrm{T}}\mathbf{c}(\mathbf{\Delta}\,\boldsymbol{\varphi}).$$

   If time varying arrival rates are taken into account, these cost functions need to be adjusted in order to capture queuing effects.

3. The *path choice criterion* follows from Wadrop's first principle. The minimum travel costs $\mathbf{p}$ prevailing at a traffic assignment $\mathbf{h}$ are computed by

$$p_{(s,t)} = p_{(s,t)}(\mathbf{h}) = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho(\mathbf{h})$$

   and all drivers make only use of least cost routes.

4. The *origin to destination demand model* introduces as new element the travel choice function

$$\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p})$$

   with elements $x^D_{(s,t)} = \xi_{(s,t)}(\mathbf{p})$ denoting the aggregate demand of all active drivers whose willingness to pay per trip from v_s to v_t is at least as high as $p_{(s,t)}$ at the prevailing price vector $\mathbf{p}$. Of course there are potential drivers who decide to stay at home (or to use different trip modes) at the given fare $\mathbf{p}$.

---

total traffic flow over the day by increasing the efficiency with which facilities are used. Moreover, revenues from congestion pricing may be used for financing investments in infrastructure.

In a dynamic context this model should capture furthermore the distribution of departure times. Some advice for such a dynamic network equilibrium approach is given in Ben-Akiva (1985).

5. A *user equilibrium* has been found when the generated traffic $\hat{\mathbf{x}}^S$ and a corresponding route choice or user optimized traffic assignment $\hat{\mathbf{h}}$ imply route costs $\tilde{\mathbf{c}}(\hat{\mathbf{h}})$ and thus "market prices" $\hat{\mathbf{p}}$ such that the demand for traffic $\hat{\mathbf{x}}^D = \boldsymbol{\xi}(\hat{\mathbf{p}})$ satisfies market clearing

$$\hat{\mathbf{x}}^S = \hat{\mathbf{x}}^D.$$

An equivalent statement with regard to some equilibrium path flow $\hat{\mathbf{h}}$ is

$$\boldsymbol{\Lambda}\,\hat{\mathbf{h}} = \boldsymbol{\xi}\big(\mathbf{p}(\hat{\mathbf{h}})\big).$$

Alternatively we shall see that these conditions can be expressed in terms of route costs $\tilde{\mathbf{c}}$ and prices $\mathbf{p}$ the travelers have to pay.

The crucial point of the last remark is the twofold meaning of $p_{(s,t)}$. The supply side of the network determines $p_{(s,t)}$ as the minimum of all route costs $\tilde{c}_\rho$ with $\rho \in \mathcal{R}(s,t)$, which increases with growing congestion. On the demand side, $p_{(s,t)}$ is related to the (marginal) willingness to pay for trips from v_s to v_t; it decreases when congestion lowers the comfort of network utilization or, equivalently, less people decide to make a trip if growing prices indicate a diminishing level of service. As travelers decide on their trip and route choice by "comparing" what they are willing to pay for a trip with route costs at prevailing traffic conditions, we make for the sake of convenience the following

4.11 Assumption (Invertible Demand Functions) *The demand functions $\boldsymbol{\xi}$ are invertible such that*[12]

$$\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p}) \iff \mathbf{p} = \boldsymbol{\xi}^{-1}(\mathbf{x}^D).$$

*for all $\mathbf{p} \geqq \mathbf{0}$ and all $\mathbf{x}^D \geqq \mathbf{0}$. The inverse demand functions are referred to as **marginal willingness to pay**. Accordingly the **willingness to pay** is defined by $k(\mathbf{x}^D) := \int_{\mathbf{0}}^{\mathbf{x}^D} \boldsymbol{\xi}^{-1}(\mathbf{z})^\mathsf{T} \mathrm{d}\mathbf{z}$ provided that this line integral exists.*

One immediate implication is

$$\mathbf{p}^\mathsf{T}\boldsymbol{\xi}(\mathbf{p}) = \mathbf{p}^\mathsf{T}\mathbf{x}^D = \mathbf{x}^{D\mathsf{T}}\boldsymbol{\xi}^{-1}(\mathbf{x}^D).$$

Another implication with respect to the Jacobian matrices of $\boldsymbol{\xi}$ and its inverse follows from Heuser (1992, Umkehrsatz 171.1) when the demand function $\boldsymbol{\xi}$ is continuously differentiable:

$$\mathbf{J}_{\boldsymbol{\xi}^{-1}}(\mathbf{x}^D) = [\mathbf{J}_{\boldsymbol{\xi}}(\mathbf{p})]^{-1} \quad \text{with} \quad \mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p}).$$

Although dealing with aggregate market demand functions which are assumed to be independent of any income distribution, it will be useful to have in mind some results obtained from the theory of the household. For further implications see Remark 4.16.

4.12 Remark (Theory of the Household) Under mild assumptions on the differentiability of utility functions, invertible demand functions follow from duality

[12] Nagurney (1999, p. 154–155) identifies $\xi^{-1}_{(s,t)}(\mathbf{x}^D)$ as travel disutility functions. This notion is misleading in all respects and is discarded in our research.

theory. Without going into the details of the theory of consumer behavior let us compare a few standard results derived from two transposed problems, namely, utility maximization[13] and expenditure minimization, as well as their respective dual problems. A household endowed with the budget y maximizes its utility $U = U(\mathbf{x})$ at given commodity prices $\mathbf{p}$ if the commodity bundle $\hat{\mathbf{x}}$ solves the problem

$$V(\mathbf{p}, y) = \max \left\{ U(\mathbf{x}) \mid \mathbf{p}^{\mathsf{T}}\mathbf{x} \leqq y,\ \mathbf{x} \geqq \mathbf{0} \right\}, \tag{4.18}$$

where V is referred to as *indirect utility function*. As V is homogeneous of degree zero in $\mathbf{p}$ and y, we can rewrite $V^*(\mathbf{p}^*) := V(\mathbf{p}^*, 1)$ with $\mathbf{p}^* := \mathbf{p}/y$ to state the dual problem

$$\breve{U}(\mathbf{x}) = \min \left\{ V^*(\mathbf{p}^*) \mid \mathbf{p}^{*\mathsf{T}}\mathbf{x} \leqq 1,\ \mathbf{p}^* \geqq \mathbf{0} \right\}. \tag{4.19}$$

On the regularity conditions to obtain $U = \breve{U}$ see Blackorby et al. (1978). The decisive observation regarding invertible demand functions follows from *Roy's theorem* and its dual, i.e., *Wold's theorem*. While Roy's theorem generates Marshallian demand functions $\mathbf{x}^M$ from the indirect utility function V^*, Wold's theorem does the same with respect to inverse demand functions $\mathbf{p}^W$ and the direct utility function U. With regard to Assumption 4.11 suppose $\hat{\mathbf{x}}$ and $\hat{\mathbf{p}}^*$ to solve the primal problem (4.18) and the dual problem (4.19), respectively; then

$$\hat{\mathbf{x}} = \mathbf{x}^M(\hat{\mathbf{p}}^*) := \frac{\nabla V^*(\hat{\mathbf{p}}^*)}{\hat{\mathbf{p}}^{*\mathsf{T}}\nabla V^*(\hat{\mathbf{p}}^*)} \quad \Longleftrightarrow \quad \hat{\mathbf{p}}^* = \mathbf{p}^W(\hat{\mathbf{x}}) := \frac{\nabla U(\hat{\mathbf{x}})}{\hat{\mathbf{x}}^{\mathsf{T}}\nabla U(\hat{\mathbf{x}})}$$

under appropriate regularity conditions.[14] An even simpler version follows from quasi-linear preferences $U(x_0, \mathbf{x}) = x_0 + \breve{U}(\mathbf{x})$ with $p_0 > 0$ and the corresponding indirect utility $V(p_0, \mathbf{p}, y)$. Letting $V^{\circledast}(\mathbf{p}^{\circledast}, y^{\circledast}) := V(1, \mathbf{p}^{\circledast}, y^{\circledast})$ with $\mathbf{p}^{\circledast} := \mathbf{p}/p_0$ and $y^{\circledast} := y/p_0$, the reader may verify[15]

$$\hat{\mathbf{x}} = -\nabla_{\mathbf{p}^{\circledast}} V^{\circledast}(\hat{\mathbf{p}}^{\circledast}, y^{\circledast}) \quad \Longleftrightarrow \quad \hat{\mathbf{p}}^{\circledast} = \nabla\breve{U}(\hat{\mathbf{x}}).$$

For any $p_0 > 0$ the demand $\hat{x}_0$ follows residually from $\hat{x}_0 = y^{\circledast} - \hat{\mathbf{p}}^{\circledast\mathsf{T}}\hat{\mathbf{x}}$.

Similar results are valid, if we draw the attention to the transposed problem of (4.18), that is, the minimum expenditure problem which is restricted to a given utility level $\bar{U}$. In the notation of Caputo (2000) the expenditure function e is defined by

$$e(\mathbf{p}, U) = \min \left\{ \mathbf{p}^{\mathsf{T}}\mathbf{x} \mid U(\mathbf{x}) \geqq \bar{U},\ \mathbf{x} \geqq \mathbf{0} \right\}, \tag{4.20}$$

---

[13] In accordance with Beckmann, Golob (1974), the utility approach serves to relate flow theory to welfare economics and to point out the various restrictive assumptions underlying more conventional network models.

[14] A detailed proof may start with the definition $V^*(\hat{\mathbf{p}}^*) = U(\mathbf{x}^M(\mathbf{p}^*))$ such that $\nabla V^*(\mathbf{p}^*) = \mathbf{A}^{\mathsf{T}}\nabla U(\hat{\mathbf{x}})$, where $\mathbf{A}$ denotes the Jacobian matrix of the Marshallian demand functions at $\mathbf{p}^*$, e.g., $\mathbf{A} = \mathbf{J}_{\mathbf{x}^M}(\mathbf{p}^*)$. On the assumption that $\mathbf{A}$ can be inverted, we find also $(\mathbf{A}^{-1})^{\mathsf{T}}\nabla V^*(\mathbf{p}^*) = \nabla U(\hat{\mathbf{x}})$. The rest of the proof follows from computing the Lagrange multipliers of (4.18) and (4.19) both for an inner solution. A comparison then shows the inverse relationship of $\hat{\mathbf{x}} = \mathbf{x}^M(\hat{\mathbf{p}}^*)$ and $\hat{\mathbf{p}}^* = \mathbf{p}^W(\hat{\mathbf{x}})$.

[15] Observe $p_0 = [x_0 + \hat{\mathbf{x}}^{\mathsf{T}}\nabla\breve{U}(\mathbf{x})]^{-1}$.

where an optimal solution $\hat{\mathbf{x}} = \mathbf{x}^H(\mathbf{p}, \bar{U})$ is referred to as Hicksian or compensated demand. The problem of expenditure minimization also has a dual problem the solution of which is the called the transformation or distance function t_U. It is related to the indirect utility V^* by a program similar to the construction principle of (4.20); see Blackorby et al. (1978, p. 30). An alternative approach[16] derives t_U from the utility function U:

$$t_U(\mathbf{x}, \bar{U}) = \max \left\{\lambda > 0 |\ U(\mathbf{x}/\lambda) \geqq \bar{U}\right\} \tag{4.21}$$

The given commodity bundle $\mathbf{x}$ is moved along a ray through the origin such that $\hat{\mathbf{x}} = \mathbf{x}/\hat{\lambda}$ is the smallest vector on that ray which can generate the utility level $\bar{U}$. The duality of the pair of functions (e, t_U) is reflected by Mahler's inequality, i.e., $t_U(\mathbf{x}, \bar{U})\, e(\mathbf{p}^*, \bar{U}) \leqq \mathbf{p}^{*\mathsf{T}}\mathbf{x}$ for all $\mathbf{p}^*$ and all $\mathbf{x}$. For economists the importance of this relationship follows from *Shephard's theorem*: for any pair of so-called polar points $(\hat{\mathbf{p}}^*, \hat{\mathbf{x}})$ satisfying this inequality as an equation it is necessary and sufficient that $\hat{\mathbf{x}}$ solves the problem of expenditure minimization. Under appropriate regularity conditions Shephard (1953) provides furthermore two systems of equations, namely,[17]

$$\hat{\mathbf{x}} = \mathbf{x}^H(\hat{\mathbf{p}}^*, \bar{U}) := \nabla_{\mathbf{p}^*} e(\hat{\mathbf{p}}^*, \bar{U}) \quad \Longleftrightarrow \quad \hat{\mathbf{p}}^* = \nabla_{\mathbf{x}} t_U(\hat{\mathbf{x}}, \bar{U}),$$

which indicate Hicksian demand functions and their reverse for the case of $\hat{\mathbf{p}}^{*\mathsf{T}}\hat{\mathbf{x}} = 1$. This corresponds again to what has been supposed in Assumption 4.11. □

4.2.2 Traffic Equilibria

(a) System Optimality

In accordance with Beckmann et al. (1956, p. 60) we begin with the problem of generating some given traffic vector $\mathbf{x}^S$ at minimum travel cost. This is (3.57), p. 141, with the restriction $\mathbf{x}^S = \mathbf{\Lambda}\,\mathbf{h}$. From (3.58) it is known that the traffic conditions prevailing at $\mathbf{x}^S$ determine the shadow prices of the OD-pairs, that is in general

$$\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\mathbf{x}^S).$$

Furthermore, a traffic equilibrium must take into account that the generated amount of traffic $\mathbf{x}^S$ has to match the demand[18] at the prevailing price system, $\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p})$, such that

$$\mathbf{x}^D = \mathbf{x}^S.$$

[16] This way of computing the distance function t_U corresponds to the derivation of the output distance function on p. 97. Indeed, one can show that the expenditure function e corresponds to a distance function derived from the indirect utility function V^*.

[17] The equation on the left hand side is usually referred to as Shephard's lemma.

[18] The aggregate demand functions are derived from the individual Marshallian demand functions. For the sake of simplicity we drop the aggregate income and its distribution from our notation; see Remark 4.16 below on the corresponding implications.

Letting $\hat{\mathbf{x}}$ denote market clearing, the preceding equation suggests a solution $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ of the form

$$\hat{\mathbf{p}} = \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}}) = \boldsymbol{\pi}(\hat{\mathbf{x}}).$$

Unfortunately, this relationship is almost always not true as for any $\hat{h}_\rho > 0$ the costs per trip differ from the marginal route costs.

$$\hat{p}_{(s,t)} = \tilde{c}_\rho(\hat{\mathbf{h}}) \leqq \tilde{\mathbf{c}}_\rho(\hat{\mathbf{h}}) = \hat{\pi}_{(s,t)}$$

This problem disappears only when the users orient their demand to $\tilde{\mathbf{c}}(\mathbf{h})$ instead of $\tilde{\mathbf{c}}(\mathbf{h})$. Every solution minimizing total cost must increase $\hat{\mathbf{p}}$ by appropriate amounts reflecting the external costs per trip incurred to the network as a whole. Nevertheless the above reasoning helps to find user optima.

(b) User Optimality

The remarks on condition (3.74), p. 149, for a user optimum point out that the optimal Lagrange multipliers $\hat{\mathbf{p}}$ have the meaning of minimum route costs $\mathbf{p}$ with respect to the prevailing traffic conditions. The route costs $\hat{\mathbf{p}} = \mathbf{p}$ are internally generated by the network and depend on some traffic matrix $\mathbf{x}^S = \boldsymbol{\Lambda}\,\mathbf{h}$ as can be seen in (3.71) regarding $\mathbf{p}$ as a function of $\mathbf{h}$ or in (3.76) concerning shadow prices. That is particularly $\hat{\mathbf{p}} = \mathbf{p}(\mathbf{x}^S)$. Again an equilibrium[19] requires market clearing denoted by $\hat{\mathbf{x}}$ at which the generated traffic $\mathbf{x}^S$ matches the demand for traffic $\mathbf{x}^D$.

Having determined the relevant relationship between minimum travel costs and generated traffic $\mathbf{p} = \mathbf{p}(\mathbf{x}^S)$ on the basis of the above mentioned conditions or

$$\text{(4.22a)} \qquad \mathbf{x}^S = \boldsymbol{\Lambda}\,\hat{\mathbf{h}},$$

$$\text{(4.22b)} \qquad \hat{h}_\rho\left[\tilde{c}_\rho(\hat{\mathbf{h}}) - p_{(s,t)}\right] = 0, \quad \hat{h}_\rho \geqq 0, \quad \tilde{c}_\rho(\hat{\mathbf{h}}) \geqq p_{(s,t)}, \qquad \forall\,\rho \in \mathcal{R}(s,t),\ \forall\,(s,t),$$

we are able to define user optima with regard to price sensitive traffic demand. A pair $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ is called a user optimum if

$$\text{(4.23)} \qquad \hat{\mathbf{p}} = \mathbf{p}(\mathbf{x}^S), \quad \hat{\mathbf{x}} = \mathbf{x}^S = \mathbf{x}^D, \quad \text{and} \quad \mathbf{x}^D = \boldsymbol{\xi}(\hat{\mathbf{p}}).$$

Accounting for path flows by means of $\hat{\mathbf{x}} = \boldsymbol{\Lambda}\,\hat{\mathbf{h}}$, where $\hat{\mathbf{h}}$ is not necessarily unique, the pair $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ satisfies $\hat{\mathbf{p}}^{\mathsf{T}}\hat{\mathbf{x}} = (\boldsymbol{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}})^{\mathsf{T}}\hat{\mathbf{h}} = \tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}\hat{\mathbf{h}}$ due to the definition of $\hat{p}_{(s,t)} = \boldsymbol{\Lambda}_\rho{}^{\mathsf{T}}\hat{\mathbf{p}}$ (see (3.26) on p. 105) as minimum travel cost between the OD-pair (v_s, v_t). The system (4.23) together with (4.22) gives rise to describe a Wardrop equilibrium by a trip table $\hat{\mathbf{x}}$ solving the fixed point problem

$$\hat{\mathbf{x}} = \boldsymbol{\xi}\big(\mathbf{p}(\hat{\mathbf{x}})\big).$$

[19] Given the trip table, Fernandez, Friesz (1983) refer to the equilibration of flows over the network during the traffic assignment step with respect to Wardrop's principle as *traffic equilibration*. In contrast *market equilibration* means the equilibration of demand and supply where both are sensitive with respect to minimum route costs incorporating congestion effects.

On Assumption 4.11 this is the same as

$$\text{find an } \hat{\mathbf{x}} \text{ such that } \mathbf{p}(\hat{\mathbf{x}}) = \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}}).$$

The generated traffic on the network must cover the demand for traffic such that the prices generated by the network equal the minimum route costs. An equivalent fixed point problem is

$$\hat{\mathbf{p}} = \mathbf{p}\big(\boldsymbol{\xi}(\hat{\mathbf{p}})\big).$$

Here we seek for a price vector $\hat{\mathbf{p}}$ such that $\boldsymbol{\xi}(\hat{\mathbf{p}}) = \boldsymbol{\chi}(\hat{\mathbf{p}})$, where $\mathbf{x}^S = \boldsymbol{\chi}(\hat{\mathbf{p}})$ is the (hopefully existing) inverse of $\hat{\mathbf{p}} = \mathbf{p}(\mathbf{x}^S)$.

The corresponding formulation as a nonlinear complementarity problem is ignored at this point. Following Friesz (1985), all we have to do is to substitute $\mathbf{x}^D$ by $\boldsymbol{\xi}(\mathbf{p})$ in the corresponding statement of Remark 3.40.

4.2.3 User Optimality in Terms of a Variational Inequality Problem

Regarding the trip table $\hat{\mathbf{x}}$ we need an appropriate traffic assignment $\hat{\mathbf{h}}$ such that $\boldsymbol{\Lambda}\,\hat{\mathbf{h}} = \hat{\mathbf{x}}$. In order to emphasize that the two feasible sets Φ^D and H^D depend on the *generated amount of traffic* $\mathbf{x}^S$ both sets are redefined as follows.

$$\begin{aligned} \Phi_X^S &= \big\{(\boldsymbol{\varphi}, \mathbf{x}^S) |\ \boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h},\ (\mathbf{h}, \mathbf{x}^S) \in H_X^S\big\} \\ \text{and}\quad H_X^S &= \big\{(\mathbf{h}, \mathbf{x}^S) |\ \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^S,\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{x}^S \geqq \mathbf{0}\big\} \end{aligned}$$

The superscript S refers to the supply side of the network because the demand functions have been recorded separately. In accordance with Definition 3.38 the advanced definition of a user optimum with price sensitive demand for trips reflects (4.22).

4.13 Definition (Wardrop equilibrium) *Based on Assumption 4.11, a feasible flow pattern* $(\hat{\mathbf{h}}, \hat{\mathbf{x}}) \in H_X^S$ *is a Wardrop equilibrium if*

$$\begin{aligned} \hat{h}_\rho > 0 &\implies \tilde{c}_\rho(\hat{\mathbf{h}}) = \xi_{(s,t)}^{-1}(\hat{\mathbf{x}}), \\ \hat{h}_\rho = 0 &\implies \tilde{c}_\rho(\hat{\mathbf{h}}) \geqq \xi_{(s,t)}^{-1}(\hat{\mathbf{x}}). \end{aligned} \tag{4.24}$$

Notice that the supply side is incorporated in the set H_X^S. This is an important condition because $\hat{\mathbf{h}}$ and $\hat{\mathbf{x}}$ are linked to each other so that not all non-negative pairs $(\mathbf{h}, \mathbf{x})$ are relevant. The demand side is taken into account by $\mathbf{p} = \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}})$. A component $\xi_{(s,t)}^{-1}$ denotes the travelers' marginal willingness to pay for a trip on any route from v_s to v_t. Note on the one hand that trip-makers will not accept any price exceeding the minimum travel cost $p_{(s,t)}$ from v_s to v_t, but most of them are willing to pay prices higher than $p_{(s,t)}$. On the other hand, they are forced to pay at least $p_{(s,t)}$. If the route cost $\tilde{c}_\rho$ exceeds this price, the demand for trips on route ρ is zero.

Figure 4.1 gives a graphic representation of these observations with respect to a network consisting of two roads ρ_1 and ρ_2 joining the OD-pair (v_s, v_t). The route

cost $\tilde{c}_{\rho_2}$ exceeds that of the first route and it is assumed in particular that $\tilde{c}_{\rho_2}(0) > \tilde{c}_{\rho_1}(\hat{h}_{\rho_1}) = \hat{p}_{(s,t)}$ to assure an all-or-nothing assignment. Thus all travelers use the first route, where $x_{(s,t)} = \hat{h}_{\rho_1}$ and $\hat{p}_{(s,t)}$ indicate the user optimal traffic assignment. No user accepts the higher price of the second route, i.e., $\hat{h}_{\rho_2} = 0$. However, if the first route is suspended all trip-makers are forced to use the remaining road so that $x_{(s,t)} = \check{h}_{\rho_2}$. Consequently, the marginal willingness to pay increases from $\hat{p}_{(s,t)}$ to $\check{p}_{(s,t)}$.

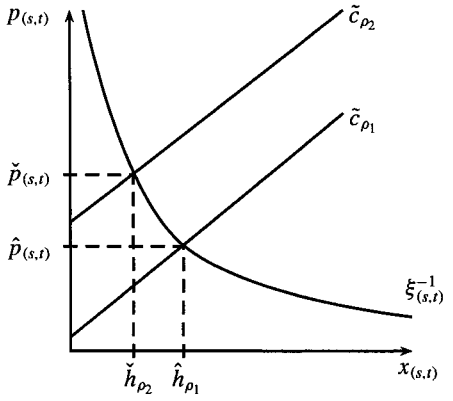


Fig. 4.1
Wardrop equilibrium
The inverse demand function $\xi^{-1}_{(s,t)}$ indicates the marginal willingness to pay for a trip from v_s to v_t. Two alternative routes ρ_1 and ρ_2 can be chosen, but the cost of the second route is prohibitively high when compared to the first route.

Outside a Wardrop equilibrium we have to distinguish two cases.

1. Some travelers do not choose the cheapest route,

$$h_{\rho'} > 0 \text{ and } \tilde{c}_{\rho'}(\mathbf{h}) > \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_{\rho}(\mathbf{h}).$$

   This is the question of a proper traffic assignment.
2. Even if all network users choose one cost minimal route, there may be some trip-makers who are willing to pay more than the price for the cheapest available route,[20]

$$h_{\rho'} > 0 \text{ and } \xi^{-1}_{(s,t)}(\mathbf{x}) > \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_{\rho}(\mathbf{h}).$$

   An easy example of this situation will be discussed in Figure 4.3. This case is important with respect to system optima which refer to marginal route cost $\tilde{c}_\rho$ with $\tilde{c}_\rho \geqq \tilde{c}_\rho$. Suppose we increase $\tilde{c}_\rho$ by an appropriate amount reflecting the external costs of a trip. This reduces h_ρ to those travelers who are willing to compensate the rest of the network for the disadvantage of another trip.

4.14 Proposition[21] *On the Assumption 4.11 of invertible demand functions, the conditions of a Wardrop equilibrium under price dependent demand are equivalent to the following variational inequality problems:*

[20] This is the case of Figure 4.1 if $\check{h}_{\rho_2}$ would be assigned to the first route and nothing to the second route.

[21] See Dafermos (1982) or Patriksson (1994, p. 85), Theorem 3.15.

1. *Find a pair* $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{x}}) \in \Phi_X^S$ *such that*

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^\top(\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}) - \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}})^\top(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \qquad \forall\, (\boldsymbol{\varphi}, \mathbf{x}) \in \Phi_X^S. \tag{4.25}$$

2. *Find a pair* $(\hat{\mathbf{h}}, \hat{\mathbf{x}}) \in H_X^S$ *such that*

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top(\mathbf{h} - \hat{\mathbf{h}}) - \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}})^\top(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \qquad \forall\, (\mathbf{h}, \mathbf{x}) \in H_X^S. \tag{4.26}$$

Proof: Remember $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$ and $\tilde{\mathbf{c}}(\hat{\mathbf{h}}) = \boldsymbol{\Delta}^\top\mathbf{c}(\boldsymbol{\Delta}\mathbf{h})$. Eliminating $\boldsymbol{\varphi}$ in (4.25) and taking into account the definition of Φ_X^S yields (4.26). The only difference between the two inequalities is that each $\mathbf{h}$ in (4.26) determines a unique $\boldsymbol{\varphi}$ in (4.25), but the inverse is not true.

Necessity: For any $\mathbf{h} \geqq \mathbf{0}$ the conditions (4.24) for a user equilibrium imply[22]

$$\left(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^\top\boldsymbol{\xi}^{-1}(\hat{\mathbf{x}})\right)^\top(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall\, \mathbf{h} \geqq \mathbf{0}. \tag{4.27}$$

This is (4.26) by using $\mathbf{x} = \boldsymbol{\Lambda}\,\mathbf{h}$ such that $(\mathbf{h}, \mathbf{x}) \in H_X^S$. Notice that (4.27) must be satisfied only for all non-negative route flows and *not* for all $\mathbf{h} \geqq \mathbf{0}$ and $\mathbf{x} \geqq \mathbf{0}$. Indeed, Theorem 3.26 would then imply $\hat{\mathbf{x}} \geqq \mathbf{0}$, $\hat{\mathbf{p}} \leqq \mathbf{0}$, and $\hat{\mathbf{p}}^\top\hat{\mathbf{x}} = 0$ which is senseless from an economic point of view.

Sufficiency: Suppose a pair $(\hat{\mathbf{h}}, \hat{\mathbf{x}}) \in H_X^S$ satisfies the variational inequality problem (4.26). Then a vector $\hat{\mathbf{h}} \geqq \mathbf{0}$ exists with $\hat{\mathbf{x}} = \boldsymbol{\Lambda}\,\hat{\mathbf{h}}$. Choose one route ρ connecting v_s and v_t and change $\hat{h}_\rho$ by $\varepsilon \neq 0$. This yields $h_\rho = \hat{h}_\rho + \varepsilon$. With that $\hat{x}_{(s,t)}$ becomes $x_{(s,t)} = \hat{x}_{(s,t)} + \varepsilon$. Substituting the newly constructed vectors $\mathbf{h}$ and $\mathbf{x}$ into the inequality (4.26) yields

$$\tilde{c}_\rho(\hat{\mathbf{h}})\,\varepsilon - \xi_{(s,t)}^{-1}(\hat{\mathbf{x}})\,\varepsilon \geqq 0.$$

(a) If $\hat{h}_\rho > 0$, the perturbation ε may be positive or negative. This requires $\tilde{c}_\rho(\hat{\mathbf{h}}) - \xi_{(s,t)}^{-1}(\hat{\mathbf{x}}) = 0$.
(b) If $\hat{h}_\rho = 0$, the perturbation ε must be positive so that $\tilde{c}_\rho(\hat{\mathbf{h}}) - \xi_{(s,t)}^{-1}(\hat{\mathbf{x}}) \geqq 0$. ■

The two variational inequality problems in Proposition 4.14 need some explanations. The first part of (4.25) has been discussed in (3.65) where Φ^D was based on a fixed trip table $\mathbf{x}^D$. The same is true for (4.26) in comparison to (3.77) where H^D depends on the given traffic demand $\mathbf{x}^D$. Furthermore, we know that the formulation in (4.25) or (4.26) is appropriate whenever the Jacobian matrices with respect to $\mathbf{c}$, $\tilde{\mathbf{c}}$ or $\boldsymbol{\xi}^{-1}$ are asymmetric. Nevertheless, whenever the link cost functions $\mathbf{c}$ are integrable they imply a convex transport cost term

$$g(\boldsymbol{\varphi}) = \int_{\mathbf{0}}^{\boldsymbol{\varphi}} \mathbf{c}(\mathbf{z})^\top \mathrm{d}\mathbf{z},$$

which was minimized in (3.69) with respect to $\boldsymbol{\varphi} \in \Phi^D$. Any solution to that problem was then shown to satisfy the Wardrop conditions of a user optimal traffic

[22] We make use of the same reasoning as in the proof of Proposition 3.43, p. 154.

assignment. Now the question arises as to how to interpret (4.25) if we assume for explanatory purposes that the Jacobian matrix $\mathbf{J}_{\xi^{-1}}$ is symmetric, too. Theorem 3.22 then states that inverse demand functions are also integrable with

$$k(\mathbf{x}^D) := \oint_{\mathbf{0}}^{\mathbf{x}^D} \boldsymbol{\xi}^{-1}(\mathbf{z})^{\top} \mathrm{d}\mathbf{z} = \sum_{(s,t)\in\mathcal{B}} \int_0^1 \xi_{(s,t)}^{-1}\big(\sigma(z)\big)\, \mathrm{d}\sigma_{(s,t)}(z),$$

where the path σ starts at $\sigma(0) = \mathbf{0}$ and ends at $\sigma(1) = \mathbf{x}^D$. Furthermore, if $\xi_{(s,t)}^{-1}$ is independent of the demand for traffic regarding any other OD-pair $(s, t) \in \mathcal{B}$ in the network, then[23]

$$k(\mathbf{x}^D) = \sum_{(s,t)\in\mathcal{B}} \int_0^{x_{(s,t)}^D} \xi_{(s,t)}^{-1}(z)\, \mathrm{d}z. \tag{4.28}$$

Each integral indicates the area under the respective inverse demand function. Consequently, $k(\mathbf{x}^D)$ measures the aggregate *willingness to pay* for traffic represented by the trip table $\mathbf{x}^D$ and the *marginal willingness to pay* follows from

$$\frac{\partial k(\mathbf{x}^D)}{\partial x_{(s,t)}^D} = \xi_{(s,t)}^{-1}(x_{(s,t)}^D).$$

Before proceeding with the minimization problem (4.29) below, let us hand in two remarks on the measurement of how consumers are affected by changes in the economic environment. Recall for the current setting that network users decide on their trip and route choice on the basis of prevailing route costs per trip. If some authority modifies these prices in order to induce a system optimal traffic assignment, many travelers are concerned with a new situation where at least some consumer prices and possibly their income (cf. commuters) have changed. In the ideal case every network user should be made better off by a system optimal pricing scheme, where the source of that benefit results from the reduction of total transport cost including congestion. This task is by far not trivial, as has been indicated in Remark 3.41 on road pricing.

4.15 Remark (Welfare measurement) The willingness to pay is closely related to the consumer's surplus. With regard to a household the consumer's surplus addresses the question of how his welfare changes when some occurrence changes the consumer prices from $\mathbf{p}^0$ to $\mathbf{p}^1$ and the household's income from y^0 to y^1. Examples of such occurrences regarding networks may be the opening of a new bridge or airport or the introduction of a congestion charge as can be observed in the City of London. Given the indirect utility function V of some household concerned we are interested in the utility difference $U^1 - U^0$ or more precisely

[23] In the framework of discrete choice models the integrals in (4.28) are usually substituted by terms of the form $S_{(s,t)}(\tilde{c}_\rho | \rho \in \mathcal{R}(s,t))\, x_{(s,t)}^D$, where $S_{(s,t)}$ is a so-called *satisfaction measure*. In accordance with Williams (1977), it is closely related to the consumers' surplus introduced below and Daganzo (1982) provides a corresponding equivalence theorem by which the modified approach yields also a network equilibrium.

$$V(\mathbf{p}^1, y^1) - V(\mathbf{p}^0, y^0) = \oint_{\mathbf{p}^0, y^0}^{\mathbf{p}^1, y^1} \left[\nabla_{\mathbf{p}} V(\mathbf{p}, y)^{\mathsf{T}} \mathrm{d}\mathbf{p} + \frac{\partial V(\mathbf{p}, y)}{\partial y} \mathrm{d}y \right],$$

where V is defined in (4.18). Using Roy's identity the line integral can be expressed as

$$V(\mathbf{p}^1, y^1) - V(\mathbf{p}^0, y^0) = \oint_{\mathbf{p}^0, y^0}^{\mathbf{p}^1, y^1} \left[- \left(\mathbf{x}^M(\mathbf{p}, y) \frac{\partial V(\mathbf{p}, y)}{\partial y} \right)^{\mathsf{T}} \mathrm{d}\mathbf{p} + \frac{\partial V(\mathbf{p}, y)}{\partial y} \mathrm{d}y \right],$$

where $\mathbf{x}^M(\mathbf{p}, y)$ designate Marshallian demand functions. As the marginal utility of income $\partial V / \partial y$ is not observable, let us assume that it is a positive constant λ. Caution: this "trick" may lead to a severe mistake as discussed in the subsequent Remark 4.16. Disregarding this objection the welfare change or (negative) *consumer's surplus* (CS) may be rewritten for a constant income $y^0 = y^1$ as follows[24]

$$\frac{1}{\lambda} \left[V(\mathbf{p}^1, y^0) - V(\mathbf{p}^0, y^0) \right] = - \oint_{\mathbf{p}^0}^{\mathbf{p}^1} \mathbf{x}^M(\mathbf{p}, y^0)^{\mathsf{T}} \mathrm{d}\mathbf{p} =: -CS.$$

Notice that the welfare cannot increase when $\mathbf{p}^1 \geqq \mathbf{p}^0$. In order to derive a range for this welfare change with unknown λ, observe the minimum expenditure e to realize a given utility level U at given commodity prices $\mathbf{p}$; see (4.20). For the household's optimal choice one obtains $e(\mathbf{p}^0, U^0) = y^0$ and $e(\mathbf{p}^1, U^1) = y^1 = y^0$. Using Shephard's lemma to derive the Hicksian (or compensated) demand functions, that is $\mathbf{x}^H(\mathbf{p}, U) = \nabla_{\mathbf{p}} e(\mathbf{p}, U)$, it follows for the *compensating variation* CV and the *equivalent variation* EV

$$CV = e(\mathbf{p}^1, U^0) - e(\mathbf{p}^0, U^0) = \oint_{\mathbf{p}^0}^{\mathbf{p}^1} \mathbf{x}^H(\mathbf{p}, U^0)^{\mathsf{T}} \mathrm{d}\mathbf{p},$$

$$EV = e(\mathbf{p}^1, U^1) - e(\mathbf{p}^0, U^1) = \oint_{\mathbf{p}^0}^{\mathbf{p}^1} \mathbf{x}^H(\mathbf{p}, U^1)^{\mathsf{T}} \mathrm{d}\mathbf{p}.$$

The compensating variation CV corresponds to the amount of money[25] the household needs to be as well off as before after the increase in prices; $\mathbf{p}^1 \geqq \mathbf{p}^0$. The equivalent variation EV measures the amount of money the household could save today if it realizes merely the lower utility level U^1 that can be realized at increased prices.

---

[24] See Williams (1977) for an application to transport networks with special regard to the framework of random utility theory. The author computes, among other things, the consumers' surplus on the basis of the path demand functions (4.5) with respect to a multinomial logit distribution. Observe, however, that these functions have been derived on the basis of a fixed trip table so that the consumers' surplus omits any effects of traffic creation or diminution.

[25] An equivalent expression for the compensating variation is $CV = e\big(\mathbf{p}^1, V(\mathbf{p}^0, y^0)\big) - y^0$, where e, with the utility level U expressed by the indirect utility function V, is referred to as *money metric utility function*. A similar result holds true for the equivalent variation $EV = y^1 - e\big(\mathbf{p}^0, V(\mathbf{p}^1, y^1)\big)$.

For the initial household's optimum with $\mathbf{x}^H(\mathbf{p}^0, U^0) = \mathbf{x}^M(\mathbf{p}^0, y^0)$ and $y^0 = e(\mathbf{p}^0, U^0)$ the Slutzky equation tells us for every commodity $j, k = 1,\ldots, n$

$$\frac{\partial x_j^H(\mathbf{p}^0, U^0)}{\partial p_k} = \frac{\partial x_j^M(\mathbf{p}^0, y^0)}{\partial p_k} + \frac{\partial x_j^M(\mathbf{p}^0, y^0)}{\partial y} \frac{\partial e(\mathbf{p}^0, U^0)}{\partial p_k}$$

$$\text{with} \quad \frac{\partial e(\mathbf{p}^0, U^0)}{\partial p_k} = x_k^H(\mathbf{p}^0, U^0) = x_k^M(\mathbf{p}^0, y^0).$$

If all goods are normal, that is $\partial x_j^M / \partial y > 0$ for all $j = 1,\ldots, n$, then $x_j^H(\mathbf{p}, U^0) > x_j^M(\mathbf{p}, y^0)$ whenever $\mathbf{p} \geq \mathbf{p}^0$.

For $\mathbf{x}^H(\mathbf{p}^1, U^1) = \mathbf{x}^M(\mathbf{p}^1, y^0)$ and $y^0 = e(\mathbf{p}^1, U^1)$, the same argument yields the inequality $x_j^H(\mathbf{p}, U^1) < x_j^M(\mathbf{p}, y^0)$ whenever $\mathbf{p} \leq \mathbf{p}^1$. Thus, for increasing prices $\mathbf{p}^0 \to \mathbf{p}^1$ one obtains

$$EV < CS < CV.$$

This relationship may be explained best by a figure representing the case of one good j, the price of which increases from p_j^0 to p_j^1. Given the income y^0, the utility level U^0 falls down to U^1. The shaded area in Figure 4.2 indicates the consumer's surplus lost by the rising price. Be aware that the price change does not influence the income y^0. If a rising price p_j increases the income ($y^0 < y^1$), the measures for differences in welfare positions cannot be compared to each other.

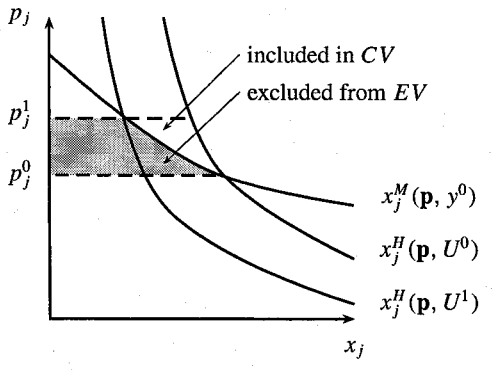


Fig. 4.2
Consumer's surplus

The consumer loses the surplus CS indicated by the shaded area. The equivalent variation EV as well as the compen- sating variation CV are not indicated explicitly.

To get an idea of the willingness to pay observe the compensated demand function $x_j^H(\cdot, U^0)$. Given p_j^0 the household has to pay $p_j^0 x_j^H(p_j^0, \mathbf{p}^{-j}, U^0)$, where $\mathbf{p}^{-j}$ is the price vector without its j^{th} component. If the price goes up to p_j^1 and other things remain equal, the household has to pay the additional amount CV to preserve his utility level U^0. One says that the household is willing to pay the higher amount if someone compensates it for the increase in expenditures. The disadvantage with this argument is that in reality no such compensating instance exists and that neither the utility level U nor the income y can be expected to be constant. Nevertheless, it is common practice to calculate the willingness to pay on the basis of Marshallian demand functions. The resulting term is then associated

with the area under the inverse demand function, i.e., the marginal willingness to pay, with respect to the interval $[0, x_j]$. A welfare change due to price alterations is approximated by the consumer's surplus CS ignoring any income shifts. A lower bound and an upper bound is then given by the equivalent variation EV and the compensating variation CV, respectively. The corresponding aggregate term for all network users with respect to all OD-pairs is given in (4.28). The analog reasoning applies also to firms. See Diewert (1986) for an analysis of changes in infrastructure services utilized by firms. □

4.16 Remark (Path independence) When the market demand functions ξ for trips are derived in the context of a general equilibrium analysis, path independence needs severe attention. Let $x^M_{j\hbar}(\mathbf{p}, y_\hbar)$ be the Marshallian demand function of household $\hbar$ for good j ($j = 1,\ldots, n$) given the household's income $y_\hbar$ and the commodity prices $\mathbf{p}$. The market demand for that good is then denoted by

$$x_j(\mathbf{p}, \mathbf{y}) = \sum_\hbar x^M_{j\hbar}(\mathbf{p}, y_\hbar) \qquad j = 1,\ldots, n.$$

Using Roy's identity and setting the marginal values of incomes to one, it can be shown (see, e.g., Buhr, Wagner (1995)) that a change in prices and income has the following monetary measure of aggregate welfare change:

$$\sum_\hbar \oint_{\mathbf{p}^0,\mathbf{y}^0}^{\mathbf{p}^1,\mathbf{y}^1} \left[-\mathbf{x}^M_\hbar(\mathbf{p}, y_\hbar)^\mathsf{T} \mathrm{d}\mathbf{p} + \mathrm{d}y_\hbar\right].$$

Without normalization of some variables involved, Chipman, Moore (1976, 1980) (for one consumer) as well as Blackorby, Donaldson (1999) (for many consumers) have proved that there are no "well behaved" preferences such that this line integral would be path independent. One way to escape this trap is to normalize the aggregate income $y = \sum_\hbar y_\hbar$ to one. Substituting in the line integral $\mathbf{p}$ by $\mathbf{p}^* = \mathbf{p}/y$ and $y_\hbar$ by $y^*_\hbar = y_\hbar/y$, Blackorby, Donaldson (1999) show that the line integral is path independent if and only if there is an "aggregate consumer" with homothetic preferences.

Regarding an urban traffic network the assumption $\sum_\hbar \mathrm{d}y^*_\hbar = 0$ is hard to justify. Just think of a new road that connects a residential area with an industrial park. It is not too hard to imagine that some commuters profit ($\mathrm{d}y^*_\hbar > 0$), but no other traveler is worse off. Problems of this type are ignored throughout the book. □

Having in mind the results on how to measure aggregate welfare changes, the problems of Proposition 4.14 of finding Wardrop equilibria can be rewritten as convex programs provided that some regularity conditions hold good. On the integrability assumptions, Theorem 3.23 states that the variational inequality problem (4.25) is equivalent to the following problem, which has been introduced by Beckmann et al. (1956, p. 63).

$$\min\left\{g(\varphi) - k(\mathbf{x}) \mid (\varphi, \mathbf{x}) \in \Phi^S_X\right\} \tag{4.29}$$

This problem requires g and k to be convex and concave, respectively, which holds true when the Jacobian matrix $\mathbf{J}_c(\varphi)$ is positive semidefinite and $\mathbf{J}_{\xi^{-1}}$ is negative

semidefinite.[26] Every solution maximizes a term $k(\mathbf{x})$ for the willingness to pay for $\mathbf{x}$ minus a travel cost term $g(\boldsymbol{\varphi})$ under the condition of flow conservation $\mathbf{x} = \boldsymbol{\Lambda}\,\mathbf{h}$ and $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$. As g is based on average cost $\mathbf{c}$ rather than marginal cost $\mathbf{c}$, the optimal value of (4.29) is *not the aggregate travelers' surplus*; see Remark 3.37. However, we shall use it as a rough approximation, having in mind that this requires small external (congestion) effects in which case average and marginal costs differ only slightly. On marginal cost pricing to internalize congestion effects see Section 4.2.4.

Even without integrability the existence of a user equilibrium solving the variational inequality problem (4.26) is guaranteed by Theorem 3.28 provided that the feasible set H_X^S in (4.26) can be shown to be a non-empty convex and compact set; see Dafermos (1982). The crucial point is that H_X^S is not bounded from above without capacity restrictions. Thus it is worthwhile to have an alternative existence theorem which has been developed by Aashtiani, Magnanti (1981). The following proposition goes back to the nonlinear complementarity problem in Remark 3.40, where the fixed trip table $\mathbf{x}^D$ is to be substituted by the demand function $\boldsymbol{\xi}(\mathbf{p})$.

4.17 Proposition (Existence) *Suppose the route cost functions $\tilde{c}_p$ are non-negative and continuous for all routes. Also suppose that the demand functions $\xi_{(s,t)}$ are continuous and bounded from above for all OD-pairs. Then the nonlinear complementarity problem in Remark 3.40 with $\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p})$ has a solution $(\hat{\mathbf{h}}, \hat{\mathbf{p}})$.*

Again the proof is based on Brouwer's fixed point theorem and can be found in Aashtiani, Magnanti (1981). Similar to Remark 3.40, the authors prove also that the solution $(\hat{\mathbf{h}}, \hat{\mathbf{p}})$ is a user equilibrium if the route cost functions are positive and the demand functions are non-negative. The uniqueness of a solution to (4.26) is an immediate consequence of Theorem 3.19.

4.18 Proposition (Uniqueness) *Let the route cost function $\tilde{\mathbf{c}}$ and the negative inverse demand function $-\boldsymbol{\xi}^{-1}$ be strictly monotone. Then* (4.26) *has at most one solution.*

Another approach similar to (4.26) is discussed in Bell, Iida (1997, p. 102f.) who completely eliminate the trip table $\hat{\mathbf{x}}$. Here the traffic demand is a function of *expected* costs per trip $p^E_{(s,t)}$ from any origin v_s to any destination v_t

$$\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p}^E).$$

Some frequently used functional forms are $x^D_{(s,t)} = x^0_{(s,t)} \exp(-\beta\, p^E_{(s,t)})$ for logit choice models or just $x^D_{(s,t)} = x^0_{(s,t)} - \beta\, p^E_{(s,t)}$. With regard to Assumption 4.11 on invertible demand functions and making use of market clearing $\mathbf{x}^D = \mathbf{x}^S = \boldsymbol{\Lambda}\,\mathbf{h}$, we may define

$$\mathbf{p}^E = \boldsymbol{\zeta}(\mathbf{h}) := \boldsymbol{\xi}^{-1}(\boldsymbol{\Lambda}\,\mathbf{h}).$$

With that the variational inequality (4.27) takes the form known from (3.78), p. 154:

$$\left(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^\top \boldsymbol{\zeta}(\hat{\mathbf{h}})\right)^\top (\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall\, \mathbf{h} \geqq \mathbf{0}. \tag{4.30}$$

[26] It may be worth to apply Theorem 3.23 with respect to (4.29). Notice that $\nabla g = \mathbf{c}$ and $\nabla k = \boldsymbol{\xi}^{-1}$, which results in the variational inequality problem (4.25).
Furthermore, Florian, Nguyen (1974) proposed a method which is based on *Generalized Benders Decomposition* for computing such equilibrium flows.

Again by Theorem 3.23, the conditions for an optimal path flow $\hat{\mathbf{h}}$ are

$$\hat{\mathbf{h}}^{\mathsf{T}}\big(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\boldsymbol{\zeta}(\hat{\mathbf{h}})\big) = 0, \quad \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\boldsymbol{\zeta}(\hat{\mathbf{h}}) \geqq \mathbf{0}, \quad \hat{\mathbf{h}} \geqq \mathbf{0}.$$

As usual, a path flow $\hat{h}_\rho$ joining the OD-pair (v_s, v_t) is zero if the corresponding route cost $\tilde{c}_\rho(\hat{\mathbf{h}})$ is greater than the expected cost $\zeta_{(s,t)}(\hat{\mathbf{h}})$ for the OD-pair concerned. In contrast any positive flow requires $\tilde{c}_\rho(\hat{\mathbf{h}}) \leqq \zeta_{(s,t)}(\hat{\mathbf{h}})$ at the prevailing traffic conditions.

Regarding simple networks let us assume for explanatory purposes that $\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\boldsymbol{\zeta}(\hat{\mathbf{h}})$ is integrable. This supposition is hard to justify because Theorem 3.22, p. 131, requires a symmetric Jacobian matrix with respect to $\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\boldsymbol{\zeta}(\hat{\mathbf{h}})$ which is satisfied at the most if the number of OD-pairs equals the number of routes. Using the integrability assumption in conjunction with Theorem 3.23, p. 132, the variational inequality problem (4.30) is equivalent to

$$\min\big\{Z(\mathbf{h}) \mid \mathbf{h} \geqq \mathbf{0}\big\} \quad \text{with} \quad Z(\mathbf{h}) := \int_{\mathbf{0}}^{\mathbf{h}} \big(\tilde{\mathbf{c}}(\mathbf{z}) - \mathbf{\Lambda}^{\mathsf{T}}\boldsymbol{\zeta}(\mathbf{z})\big)^{\mathsf{T}} \mathrm{d}\mathbf{z}.$$

This problem is in fact the same as (4.29) when substituting

$$\tilde{g}(\mathbf{h}) = \int_{\mathbf{0}}^{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z} \quad \text{and} \quad k(\mathbf{\Lambda}\,\mathbf{h}) = \int_{\mathbf{0}}^{\mathbf{h}} \boldsymbol{\zeta}(\mathbf{z})^{\mathsf{T}} \mathbf{\Lambda}\, \mathrm{d}\mathbf{z}.$$

Figure 4.3 picks out one OD-pair and *the* route connecting it. The gray area indicates what one might identify as net benefit with respect to that OD-pair resulting from flow $\hat{h}_\rho$. But this is only true if the upward curve would represent marginal route cost, which corresponds to the same caveat as has been made in the sequel of (4.29). Nevertheless, when the trip choice of all network participants is determined by the preceding minimum problem, they behave together as if they maximize the difference between their willingness to pay and a cost term which corresponds to the area below the curve representing average route costs.

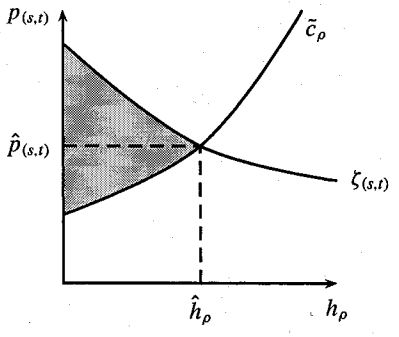


Fig. 4.3
Measuring net benefit
The consumers behave as if they maximize the shaded area with respect to all OD-pairs.

A similar figure with constant $\tilde{c}_\rho$ is given in Steenbrink (1974, p. 45). This is the case where average and marginal route costs coincide, $\mathfrak{c}_\rho = \tilde{c}_\rho$. Consequently, as indicated by Figure 3.12 on p. 148, $\mathfrak{c}_\rho > \tilde{c}_\rho$ results in an overuse of that road.

4.2.4 Marginal Cost Pricing

In what follows marginal cost pricing is used to correct individual traffic decisions in order to obtain system optimal traffic flows. System optimality is achieved when the net benefit of traffic is maximized, where net benefit of traffic means the difference between the aggregate willingness to pay and total traffic costs. Bear in mind that marginal cost pricing implies travelers to pay more than transport cost, i.e., $[\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \tilde{\mathbf{c}}(\hat{\mathbf{h}})]^{\mathsf{T}}\hat{\mathbf{h}} > 0$. This difference is needed to cover fixed cost of carrying on the network. With regard to a non-profit network provider it may be necessary to substitute marginal cost pricing by a toll/subsidy scheme constraining net tolls collected to be zero; cf. Remark 3.41.

The starting point is that of finding user optimal path flows for the combined travel choice/trip assignment problem (4.29). After eliminating link flows this problem takes the form

$$\min_{\mathbf{h},\mathbf{x}^D}\left\{\tilde{g}(\mathbf{h}) - k(\mathbf{x}^D)\middle|\ (\mathbf{h},\mathbf{x}^D) \in H_X^S\right\} \tag{4.31}$$

$$= \min_{\mathbf{h},\mathbf{x}^D}\left\{\oint_{\mathbf{0}}^{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{z})^{\mathsf{T}}\mathrm{d}\mathbf{z} - \oint_{\mathbf{0}}^{\mathbf{x}^D} \boldsymbol{\xi}^{-1}(\mathbf{z})^{\mathsf{T}}\mathrm{d}\mathbf{z}\middle|\ \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{x}^D \geqq \mathbf{0}\right\}.$$

In fact, the condition for a user equilibrium (4.27) corresponds to the set of KKT conditions

$$\left[\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\xi}^{-1}(\boldsymbol{\Lambda}\,\hat{\mathbf{h}})\right]^{\mathsf{T}}\hat{\mathbf{h}} = 0, \qquad \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\xi}^{-1}(\boldsymbol{\Lambda}\,\hat{\mathbf{h}}) \geqq \mathbf{0}, \qquad \hat{\mathbf{h}} \geqq \mathbf{0}.$$

The problems (4.29) or, equivalently, (4.31) are discussed in Yang, Huang (1998) with respect to system optimality. In this case $\tilde{g}$ and $\tilde{\mathbf{c}}$ have to be substituted by the total transport cost $\tilde{f}$ and the marginal route costs $\tilde{\mathbf{c}}$, respectively, so that (4.32) yields the correct expression for the *net benefit of traffic*.

$$\min\left\{f(\boldsymbol{\varphi}) - k(\mathbf{x}^d)\middle|\ (\boldsymbol{\varphi},\mathbf{x}^D) \in \Phi_X^S\right\} = \min\left\{\tilde{f}(\mathbf{h}) - k(\boldsymbol{\Lambda}\,\mathbf{h})\middle|\ \mathbf{h} \geqq \mathbf{0}\right\} \tag{4.32}$$

The KKT conditions are the same as before but with marginal route costs,

$$\left[\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\xi}^{-1}(\boldsymbol{\Lambda}\,\hat{\mathbf{h}})\right]^{\mathsf{T}}\hat{\mathbf{h}} = 0, \qquad \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\xi}^{-1}(\boldsymbol{\Lambda}\,\hat{\mathbf{h}}) \geqq \mathbf{0}, \qquad \hat{\mathbf{h}} \geqq \mathbf{0},$$

where

$$\tilde{\mathbf{c}}(\mathbf{h}) = \tilde{\mathbf{c}}(\mathbf{h}) + \nabla\tilde{\mathbf{c}}(\mathbf{h})\mathbf{h}. \tag{4.33}$$

If a controlling agency charges a route specific toll system $\tilde{\mathbf{t}} = \nabla\tilde{\mathbf{c}}(\hat{\mathbf{h}})\hat{\mathbf{h}}$ so that users perceive route costs $\tilde{\mathbf{c}} + \tilde{\mathbf{t}}$, then a new traffic equilibrium with price sensitive demand is established. In view of the network provider it is more advantageous to express this toll system in terms of link flows. One reason is that there are fewer links than routes and congestion results from link flows. Another reason refers to a more realistic implementation of tollgates to raise tolls. In doing so remember the

definition of route costs $\tilde{\mathbf{c}}(\mathbf{h}) = \mathbf{\Delta}^{\mathsf{T}}\mathbf{c}(\mathbf{\Delta}\,\mathbf{h})$ and compute marginal link costs on the basis of $\mathbf{c}(\boldsymbol{\varphi})$. Then[27]

$$\mathbf{c}(\boldsymbol{\varphi}) = \mathbf{c}(\boldsymbol{\varphi}) + \nabla\mathbf{c}(\boldsymbol{\varphi})\boldsymbol{\varphi} \tag{4.34}$$

and premultiplying by the transpose of $\mathbf{\Delta}$ yields (4.33). Even the optimal link specific toll system $\mathbf{t} = \nabla\mathbf{c}(\hat{\boldsymbol{\varphi}})\hat{\boldsymbol{\varphi}}$ inducing $\tilde{\mathbf{t}} = \mathbf{\Delta}^{\mathsf{T}}\mathbf{t}$ may be very difficult to compute if link flows interact. In the easy case of separable cost functions (4.34) reduces to

$$c_e(\varphi_e) = c_e(\varphi_e) + t_e \quad \text{with} \quad t_e = \frac{\mathrm{d}c_e(\varphi_e)}{\mathrm{d}\varphi_e}\,\varphi_e \qquad \forall\, e \in \mathcal{E}.$$

As has been discussed in the sequel of Figure 3.13 this pricing rule implies again that the travelers pay more than the costs caused by the realized trip table. Substituting $\tilde{\mathbf{t}}$ into the KKT conditions of (4.32) reveals the total toll revenue

$$\begin{aligned}\tilde{\mathbf{t}}^{\mathsf{T}}\hat{\mathbf{h}} &= \left(\boldsymbol{\xi}^{-1}(\mathbf{\Lambda}\,\hat{\mathbf{h}})\right)^{\mathsf{T}}\mathbf{\Lambda}\,\hat{\mathbf{h}} - \tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}\hat{\mathbf{h}} \\ &= \mathbf{p}^{\mathsf{T}}\boldsymbol{\xi}(\mathbf{p}) - \tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}\hat{\mathbf{h}}\end{aligned}$$

for the case of marginal cost pricing; cf. Larsson, Patriksson (1998). Ignoring fixed cost of network provision, Hearn, Ramana (1998) now impose an additional non-profit constraint on the network carrier. That requires a toll/subsidy schedule where net tolls collected are zero. Notice that in the current model a non-zero toll system changes not only the routing decisions but also the trip choice. This fact must be taken into account if some toll/subsidy schedule should generate the same path flows as marginal cost pricing.

4.2.5 Non-invertible Demand Functions

(a) Conjugate Functions

Regarding route flows, the combined travel choice/trip assignment problems (4.29) or (4.31) are substantially based upon the assumption that the demand functions $\boldsymbol{\xi}$ are invertible. In what follows equivalent programs are derived which can be used, for instance, when the demand functions cannot be inverted, cf. Dafermos, Nagurney (1984). Another important result expresses (4.31) solely in terms of price variables instead of the quantity variables $\mathbf{h}$ and $\mathbf{x}^D$.

The following reasoning on duality is heavily based on conjugate convex and conjugate concave functions introduced below. Observe that this theory deals with pairs of functions, say (f, f^*), and pairs of points, say $(\mathbf{x}, \mathbf{x}^*)$. It may be helpful to identify variables indicated by a star as *price variables facing quantities* which are denoted by ordinary variables.

[27] It may be helpful to observe $\nabla\tilde{\mathbf{c}}(\mathbf{h}) = \mathbf{\Delta}^{\mathsf{T}}\nabla\mathbf{c}(\boldsymbol{\varphi})\mathbf{\Delta}$, where $\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h}$.
The corresponding result for a network with two routes can be found in Smith (1979b).

The subsequent definitions and results on conjugate functions are taken from Rockafellar (1972). Let f be a proper[28] convex function. The *conjugate convex function* f^* of f is defined by

$$f^*(\mathbf{x}^*) = \sup_{\mathbf{x}} \left\{\mathbf{x}^\mathsf{T}\mathbf{x}^* - f(\mathbf{x})\right\}$$

and one can show that f^* is in turn convex; see Figure 4.4 for a graphic representation. Whenever f is a proper convex function, the pair (f, f^*) satisfies *Fenchel's inequality* $f(\mathbf{x}) + f^*(\mathbf{x}^*) \geqq \mathbf{x}^\mathsf{T}\mathbf{x}^*$ for all $\mathbf{x}$ and $\mathbf{x}^*$. Furthermore, if f is closed[29], the biconjugate f^{**} of f^* satisfies $f = f^{**}$ (Rockafellar, Theorem 12.2). For any closed proper convex function f the following conditions for a pair $(\mathbf{x}, \mathbf{x}^*)$ are equivalent to each other (Rockafellar, Theorem 23.5): (a) $f(\mathbf{x}) + f^*(\mathbf{x}^*) = \mathbf{x}^\mathsf{T}\mathbf{x}^*$, (b) $\mathbf{x}^* = \partial f(\mathbf{x})$ and (c) $\mathbf{x} = \partial f^*(\mathbf{x}^*)$. Furthermore, by Corollary 23.5.1, the subdifferentials ∂f and ∂f^* are inverse mappings, that is $\mathbf{x}^* = \partial f(\mathbf{x})$ if and only if $\mathbf{x} = \partial f^*(\mathbf{x}^*)$. When f is differentiable and ∇f is invertible[30], one has $\partial f^*(\mathbf{x}^*) = (\nabla f)^{-1}(\mathbf{x}^*)$. A pair $(\mathbf{x}, \mathbf{x}^*)$ satisfying the preceding conditions (a)–(c) is called a *pair of dual points*. Now assume that both f and f^* are differentiable and let $\mathbf{a}$ and $\mathbf{a}^*$ be two given points. Observe $f(\mathbf{x}) - f(\mathbf{a}) = \oint_{\mathbf{a}}^{\mathbf{x}} \nabla f(\mathbf{z})^\mathsf{T} d\mathbf{z}$ and $f^*(\mathbf{x}^*) - f^*(\mathbf{a}^*) \coloneqq \oint_{\mathbf{a}^*}^{\mathbf{x}^*} \nabla f^*(\mathbf{z})^\mathsf{T} d\mathbf{z}$. Then

$$(4.35)\quad f(\mathbf{x}) + f^*(\mathbf{x}^*) = \mathbf{x}^\mathsf{T}\mathbf{x}^* = \oint_{\mathbf{a}}^{\mathbf{x}} \nabla f(\mathbf{z})^\mathsf{T} d\mathbf{z} + \oint_{\mathbf{a}^*}^{\mathbf{x}^*} \nabla f^*(\mathbf{z})^\mathsf{T} d\mathbf{z} + f(\mathbf{a}) + f^*(\mathbf{a}^*)$$

can be added to the above list of conditions (a)–(c). Typically we require also $f(\mathbf{a}) + f^*(\mathbf{a}^*) = \mathbf{a}^\mathsf{T}\mathbf{a}^*$. Given $\mathbf{a}$ induces $\mathbf{a}^* = \nabla f(\mathbf{a})$ or, conversely, given $\mathbf{a}^*$ induces $\mathbf{a} = \nabla f^*(\mathbf{a}^*)$.

4.19 Remark (On the Concept of Conjugate Functions) Figure 4.4 depicts a pair (f, f^*) by making use of the derivative f' of f; it includes the above cited properties (a)–(c) and furthermore (4.35). Observe $a^* = f'(a)$ as well as $x^* = f'(x)$. The dual statements follow from $a = (f^*)'(a^*)$ and $x = (f^*)'(x^*)$, where $(f^*)'$ is the inverse of f', i.e., $(f^*)' = (f')^{-1}$.

Comparing the shaded areas in Figure 4.4 to those of Figure 4.3 gives an idea of how to benefit from conjugate functions with respect to economic analysis. □

The same concept applies to n-proper concave functions with the exception that the conjugate concave $g^\star$ of g is defined by $g^\star(\mathbf{x}^*) = \inf_{\mathbf{x}} \left\{\mathbf{x}^\mathsf{T}\mathbf{x}^* - g(\mathbf{x})\right\}$. Caution: in general $g^\star \neq -(-g)^*$. For the convex function $f = -g$ one has $g^\star(\mathbf{x}^*) = -f^*(-\mathbf{x}^*)$. With that f and f^* can be substituted by g and $g^\star$ in (4.35). A similar reasoning can be found in Evans (1976) in her appendix on conjugate functions.

---

[28] A proper function is finite for at least one point of its domain and attains nowhere the value $-\infty$.

[29] On the definition of a closed function see Rockafellar (1972, p. 52). All we need to know about closed functions is that, for a proper convex function, closedness is the same as lower semi-continuity, in which case the epigraph of the function is closed, too.

[30] When ∇f is invertible, the function $h(\mathbf{x}^*) \coloneqq \mathbf{x}^{*\mathsf{T}}(\nabla f)^{-1}(\mathbf{x}^*) - f\big((\nabla f)^{-1}(\mathbf{x}^*)\big)$ is called the *Legendre transform* relative to f. The close relation of f^* and h becomes clear if we are allowed to substitute $\mathbf{x} = (\nabla f)^{-1}(\mathbf{x}^*)$.

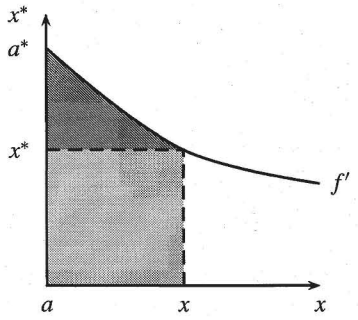


Fig. 4.4
Conjugate convex f^* of f
The shaded area under the curve f' is $f(x)$. The light gray rectangle is given by xx^*. Calculating the dark gray area by $xx^* - f(x)$ yields $f^*(x^*)$ being negative in our example because $a^* > x^*$.

For those with eyes for the finer details notice the difference between the symbols $*$ and $\star$ denoting conjugate convex and conjugate concave functions, respectively. Many authors do not distinguish the two cases having in mind that essentially the former convex supremum problem is substituted by a concave infimum problem. The "only" effect is that the direction of corresponding inequalities must be reversed.

(b) Notational Conventions

Before going on with the analysis of traffic equilibria we need two additional variables: a route specific price vector $\tilde{\mathbf{w}}$ as well as a link specific price vector $\mathbf{w}$. The pair of points $(\mathbf{x}, \mathbf{x}^*)$ introduced in the previous section then corresponds to one of the following pairing of price variables and quantity variables.

$$(\mathbf{p}, \mathbf{x}), \quad (\tilde{\mathbf{c}}, \mathbf{h}) \quad \text{or} \quad (\tilde{\mathbf{w}}, \mathbf{h}), \quad (\mathbf{c}, \boldsymbol{\varphi}) \quad \text{or} \quad (\mathbf{w}, \boldsymbol{\varphi})$$

Furthermore, remember the definition of the minimum travel cost vector $\mathbf{p}$.

$$p_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho(\mathbf{h}) \qquad \forall\, (s,t) \in \mathcal{B} \tag{4.36}$$

That vector $\mathbf{p}$ assigns a price to each OD-pair. However, in what follows it is useful to have a price for each route. As there are in general more routes than OD-pairs we need an appropriate number of "copies" of the components of $\mathbf{p}$. They are created by letting $\tilde{\mathbf{w}} := \mathbf{\Lambda}^\top \mathbf{p}$ such that the route specific prices satisfy[31] $\tilde{w}_\rho = \mathbf{\Lambda}_\rho{}^\top \mathbf{p} = p_{(s,t)}$ if $\rho \in \mathcal{R}(s,t)$. Restating Wardrop's first principle now yields

$$\tilde{w}_\rho < \tilde{c}_\rho \implies \hat{h}_\rho = 0 \quad \text{and} \quad \hat{h}_\rho > 0 \implies \tilde{w}_\rho = \tilde{c}_\rho.$$

In particular, the price $\tilde{w}_\rho$ for every route in use equals $\tilde{c}_\rho$. For later purposes also a link specific price vector $\mathbf{w}$ is needed satisfying

$$w_e < c_e \implies \hat{\varphi}_e = 0 \quad \text{and} \quad \hat{\varphi}_e > 0 \implies w_e = c_e.$$

An immediate consequence for all utilized routes is

[31] The same reasoning has been introduced in (3.26) on p. 105.

$$\hat{h}_\rho > 0 \implies \hat{\varphi}_e > 0 \ \forall e \in \rho \implies w_e = c_e \ \forall e \in \rho, \implies \tilde{c}_\rho = \tilde{w}_\rho = \sum_{e \in \rho} w_e.$$

A link cost c_e exceeding w_e indicates that the link is out of service.

$$\begin{aligned} w_e < c_e &\overset{(1)}{\implies} \hat{\varphi}_e = 0 \implies \hat{h}_\rho = 0 \quad \text{for all routes } \rho \text{ containing link } e \\ &\overset{(2)}{\implies} \tilde{w}_\rho < \tilde{c}_\rho \qquad\qquad\quad \text{for all routes } \rho \text{ containing link } e \end{aligned}$$

To keep track of the analysis observe the following summary of results holding true by definition.

$$\begin{array}{llll} \boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h} & \tilde{\mathbf{c}} = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c} \geqq \tilde{\mathbf{w}} & \tilde{\mathbf{w}} = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{w} & \mathbf{w}^{\mathsf{T}}\boldsymbol{\varphi} = \mathbf{c}^{\mathsf{T}}\boldsymbol{\varphi} = \mathbf{c}^{\mathsf{T}}\boldsymbol{\Delta}\,\mathbf{h} = \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h} \\ \mathbf{x} = \boldsymbol{\Lambda}\,\mathbf{h} & \tilde{\mathbf{w}} = \boldsymbol{\Lambda}^{\mathsf{T}}\mathbf{p} \leqq \tilde{\mathbf{c}} & & \mathbf{p}^{\mathsf{T}}\mathbf{x} = \mathbf{p}^{\mathsf{T}}\boldsymbol{\Lambda}\,\mathbf{h} = \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} \end{array}$$

At last, many results for pairs of conjugate functions will be applied to functions, say the cost term g, the argument of which is a linear transformation of another variable, that is $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$ in the case at hand. Up until now the naming convention was $g(\boldsymbol{\varphi}) = g(\boldsymbol{\Delta}\,\mathbf{h}) \equiv \tilde{g}(\mathbf{h})$. With regard to duality theory, however, it turns out to be useful to give $\tilde{g}$ (and all corresponding functions) a new name, that is to say $(g\boldsymbol{\Delta})$ with

$$g(\boldsymbol{\varphi}) = g(\boldsymbol{\Delta}\,\mathbf{h}) \equiv \tilde{g}(\mathbf{h}) \equiv (g\boldsymbol{\Delta})(\mathbf{h}).$$

This convention will be applied to all subsequent results. For example, on the assumption that $\nabla g = \mathbf{c}$, equation (3.73) on p. 149 becomes $\tilde{\mathbf{c}} = \nabla\tilde{g} = \nabla(g\boldsymbol{\Delta})$. Recall also that the convexity (or concavity) of g is preserved when going over from g to $(g\boldsymbol{\Delta})$; see Proposition 3.33 on p. 141.

(c) Marginal Willingness to Pay Pricing

Going back to the travel choice/trip assignment problem (4.31), p. 200, the first intuitive step is to eliminate the trip matrix $\mathbf{x}^D$ in the expression for the aggregate willingness to pay

$$k(\mathbf{x}^D) = k(\boldsymbol{\Lambda}\mathbf{h}) =: (k\boldsymbol{\Lambda})(\mathbf{h}), \tag{4.37}$$

where the concavity of k passes to $(k\boldsymbol{\Lambda})$ by Proposition 3.33. The conjugate concave functions of k and $(k\boldsymbol{\Lambda})$ are again concave and denoted by $k^\star$ and $(k\boldsymbol{\Lambda})^\star$, respectively. Before giving these functions an economic interpretation, we restate the definition of the conjugate function $(k\boldsymbol{\Lambda})^\star$ as well as its biconjugate function $(k\boldsymbol{\Lambda})^{\star\star}$ and point out some of its properties.

$$(k\boldsymbol{\Lambda})^\star(\tilde{\mathbf{w}}) = \inf_{\mathbf{h}} \left\{ \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} - (k\boldsymbol{\Lambda})(\mathbf{h}) \right\} \tag{4.38a}$$

$$(k\boldsymbol{\Lambda})^{\star\star}(\mathbf{h}) = \inf_{\tilde{\mathbf{w}}} \left\{ \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} - (k\boldsymbol{\Lambda})^\star(\tilde{\mathbf{w}}) \right\} \tag{4.38b}$$

From an economic point of view it is important to notice that both definitions refer to the *route specific price vector* $\tilde{\mathbf{w}}$ and not to $\tilde{\mathbf{c}}$. The reason will be explained in context

with (4.41).[32] The notion of the biconjugate function $(k\mathbf{\Lambda})^{**}$ may be dropped if $(k\mathbf{\Lambda})$ is assumed to be closed. Then

$$(k\mathbf{\Lambda}) = (k\mathbf{\Lambda})^{**}. \tag{4.39}$$

It is assumed that the willingness to pay $(k\mathbf{\Lambda})$ is finite for any non-negative route flow $\mathbf{h} \geqq \mathbf{0}$ and $-\infty$ otherwise, hence it is n-proper. The same property is transferred to $(k\mathbf{\Lambda})^*$ with respect to $\tilde{\mathbf{w}}$ by virtue of (4.38a) (and turns back to $(k\mathbf{\Lambda})^{**}$ via (4.38b)). As has been emphasized with respect to $(g\Delta)$, a differentiable function $(k\mathbf{\Lambda})$ satisfies

(4.40)
$$\nabla k(\mathbf{x}) = \boldsymbol{\xi}^{-1}(\mathbf{x}) = \mathbf{p} \quad \text{and} \quad \nabla(k\mathbf{\Lambda})(\mathbf{h}) = \mathbf{\Lambda}^{\mathsf{T}} \nabla k(\mathbf{\Lambda}\mathbf{h}) = \mathbf{\Lambda}^{\mathsf{T}} \boldsymbol{\xi}^{-1}(\mathbf{\Lambda}\mathbf{h}) = \tilde{\mathbf{w}}$$

by the definition of k. While the gradient of k reveals the OD-specific marginal willingness to pay $\mathbf{p}$, the gradient of $(k\mathbf{\Lambda})$ yields the path specific price vector $\tilde{\mathbf{w}}$, both of which are related by $\tilde{\mathbf{w}} = \mathbf{\Lambda}^{\mathsf{T}}\mathbf{p}$.

What is the economic meaning of the conjugate concave function $(k\mathbf{\Lambda})^*$? The definition (4.38a) states that it is the negative difference between the aggregate willingness to pay and the expenditures to realize the trip table which is decomposed into route flows. Thus, $-(k\mathbf{\Lambda})^*$ denotes the maximum *consumers' surplus*. A graphic representation of this observation is given in Figure 4.4, p. 203. Just imagine a network with one route ρ connecting the OD-pair (v_s, v_t) and associate (x, x^*, f') with $(h_\rho, \tilde{w}_\rho, \xi^{-1}_{(s,t)})$.

The result of (4.39) suggests to substitute the definition of $(k\mathbf{\Lambda})^{**}$ into (4.31). To avoid a mixture of infimum and minimum problems, assume that the infima in the above definitions of conjugate concave functions are attained for every $\tilde{\mathbf{w}} \geqq \mathbf{0}$ and for every $\mathbf{h} \geqq \mathbf{0}$. Then (4.31) goes over into a *minimax problem.*

$$\begin{aligned} &\min_{\mathbf{h}} \left\{ (g\Delta)(\mathbf{h}) - \min_{\tilde{\mathbf{w}}} \left\{ \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} - (k\mathbf{\Lambda})^*(\tilde{\mathbf{w}}) \right\} \middle| \, \mathbf{h} \geqq \mathbf{0},\ \tilde{\mathbf{w}} \geqq \mathbf{0} \right\} \\ &\quad = \min_{\mathbf{h}} \max_{\tilde{\mathbf{w}}} \left\{ (g\Delta)(\mathbf{h}) - \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} + (k\mathbf{\Lambda})^*(\tilde{\mathbf{w}}) \middle| \, \mathbf{h} \geqq \mathbf{0},\ \tilde{\mathbf{w}} \geqq \mathbf{0} \right\} \end{aligned}$$

The structure of this problem is similar to a bilevel programming problem. Given the path flow $\mathbf{h}$, we seek a route price vector $\tilde{\mathbf{w}}$ that minimizes $\tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} - (k\mathbf{\Lambda})^*(\tilde{\mathbf{w}})$. The result $(k\mathbf{\Lambda})(\mathbf{h})$ indicates that the travelers are not prepared to pay more than their willingness to pay. The opposite problem takes prices $\tilde{\mathbf{w}}$ as being fixed. It then maximizes $\tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} - (g\Delta)(\mathbf{h})$ with respect to $\mathbf{h}$. This reminds us of maximizing the producers' surplus provided the travelers would base their traffic demand on marginal route costs rather than average costs.

At first sight the minimax problem is unusual in the sense that the minimization is performed with respect to quantity variables while the maximization refers to price variables. This procedure, however, is common practice when applying the Lagrangean method. The minimax problem has the appealing property that it is based on two variables, namely, $\mathbf{h}$ and $\tilde{\mathbf{w}}$, having the same size. Nevertheless, a price

[32] Nevertheless, Beckmann, Kapur (1972) show that the introduction of $\tilde{\mathbf{w}}$ can be omitted at the disadvantage of augmenting the problems by the constraint $\mathbf{\Lambda}^{\mathsf{T}}\mathbf{p} \leqq \tilde{\mathbf{c}}$.

vector $\tilde{\mathbf{w}}$ veils the fact that travelers choose only cheapest routes. In other words, all routes ρ joining the same OD-pair (v_s, v_t) are valued by the trip-makers by the same price $p_{(s,t)} = \tilde{w}_\rho$, which notably reduces the number of independent price variables. The next step pursues the task of eliminating $\tilde{\mathbf{w}}$ in the minimax problem by an appropriate expression depending on $\mathbf{p}$.

In order to realize this aim we refer to Rockafellar (1972, Theorem 16.3). This result relates the infimum problem (4.38a) to a new supremum problem, which is by far not trivial, but relatively simple from a technical point of view. On the assumption that a path flow $\mathbf{h}$ exists such that $\mathbf{x} = \mathbf{\Lambda}\,\mathbf{h}$ is in the relative interior of the effective domain of k, the theorem states

$$(k\mathbf{\Lambda})^\star(\tilde{\mathbf{w}}) = \sup_{\mathbf{p}} \left\{k^\star(\mathbf{p}) \middle|\ \mathbf{\Lambda}^\mathsf{T}\mathbf{p} = \tilde{\mathbf{w}}\right\}. \tag{4.41}$$

While definition (4.37) connects the two functions k and $(k\mathbf{\Lambda})$ expressing the willingness to pay, equation (4.41) reveals a similar relationship between $k^\star$ and $(k\mathbf{\Lambda})^\star$, but with an optimization step in between. Furthermore, draw your attention to the constraint of (4.41). It relates the variables $\mathbf{p}$ and $\tilde{\mathbf{w}}$ by an equation. This is true by definition and has been utilized in the definition (4.38a) of $(k\mathbf{\Lambda})^\star$. Choosing the route cost vector $\tilde{\mathbf{c}}$ instead of $\tilde{\mathbf{w}}$ would result almost always in a contradiction because $\tilde{\mathbf{w}} \leqq \tilde{\mathbf{c}}$.

Disregarding the intrinsic importance of (4.41), it can be used to eliminate $\tilde{\mathbf{w}}$ from the above minimax problem. If the supremum in (4.41) is attained for every $\mathbf{p} \geqq \mathbf{0}$, the minimax problem can be rewritten as

$$\min_{\mathbf{h}} \max_{\mathbf{p}} \left\{(g\mathbf{\Delta})(\mathbf{h}) - \mathbf{p}^\mathsf{T}\mathbf{\Lambda}\,\mathbf{h} + k^\star(\mathbf{p}) \middle|\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{p} \geqq \mathbf{0}\right\}, \tag{4.42}$$

where

$$\tilde{\mathbf{w}}^\mathsf{T}\mathbf{h} = \mathbf{p}^\mathsf{T}\mathbf{\Lambda}\,\mathbf{h} = \mathbf{p}^\mathsf{T}\mathbf{x}^D.$$

Problems of that type are discussed in Rockafellar (1972, Part VII) and we shall come back to it in the next section, where corresponding optimum conditions are introduced.

Before doing so it is worthwhile to note some properties of the two pairs of functions $(k, k^\star)$ and $((k\mathbf{\Lambda}), (k\mathbf{\Lambda})^\star)$. As has been mentioned above, the two functions k and $(k\mathbf{\Lambda})$ indicate the travelers' willingness to pay for a trip table $\mathbf{x}^D$. Similarly, the economic interpretation of $k^\star$ and $(k\mathbf{\Lambda})^\star$ is the same. Both of them represent the negative consumers' surplus. Nevertheless it is useful to go back to the fundamental relationship (4.35). If k and $k^\star$ are differentiable, then $\nabla k = \boldsymbol{\xi}^{-1}$, in accordance with (4.40), and $\nabla k^\star = (\nabla k)^{-1} = \boldsymbol{\xi}$. Furthermore, substitute $\mathbf{a}^*$ by $\mathbf{0}$ and $\mathbf{a}$ by $\mathbf{p}^0 = \boldsymbol{\xi}^{-1}(\mathbf{0})$ or, equivalently, $\boldsymbol{\xi}(\mathbf{p}^0) = \mathbf{0}$ into (4.35). Then the line integral for k in (4.31) may be rewritten as

$$k(\mathbf{x}^D) = \oint_{\mathbf{0}}^{\mathbf{x}^D} \boldsymbol{\xi}^{-1}(\mathbf{z})^\mathsf{T}\mathrm{d}\mathbf{z} = \mathbf{p}^\mathsf{T}\mathbf{x}^D - \oint_{\mathbf{p}^0}^{\mathbf{p}} \boldsymbol{\xi}(\mathbf{z})^\mathsf{T}\mathrm{d}\mathbf{z} - \mathbf{0}^\mathsf{T}\mathbf{p}^0 = \mathbf{p}^\mathsf{T}\mathbf{x}^D - k^\star(\mathbf{p}). \tag{4.43}$$

The second line integral denotes the conjugate concave $k^\star$ of k. Expression (4.43) states that the area below the demand curves or the willingness to pay is equal to

the realized expenditures *plus* the consumers' surplus if we exchange the upper and the lower limit of the integral. Be aware that $\mathbf{p}^0$ is a prohibitive price vector where the complete demand for trips vanishes simultaneously, which implies $k^*(\mathbf{p}^0) = 0$. In order to avoid a new restriction of the form $\mathbf{p}^0 \geqq \mathbf{p}$ we set $\xi(\mathbf{p}) = \mathbf{0}$ whenever $\mathbf{p} \geqq \mathbf{p}^0$. Nevertheless some demand functions can achieve zero already for $\mathbf{p} \leqq \mathbf{p}^0$.

The equality in (4.43) requires some attention as it is only valid if the pair $(\mathbf{x}^D, \mathbf{p})$ is chosen appropriately. Notice that the following proposition makes use of superdifferentials Δk and Δk^* due to the concavity of k. The conditions (a)–(c) are the same as those which have been introduced on p. 202.

4.20 Proposition *If k is n-proper, one has $k(\mathbf{x}) + k^*(\mathbf{p}) \leqq \mathbf{p}^\mathsf{T}\mathbf{x}$ for all $\mathbf{p}$ and all $\mathbf{x}$ (Fenchel's inequality). If k is also concave, then for any vector $\hat{\mathbf{x}}$ the following two conditions on a vector $\hat{\mathbf{p}}$ are equivalent to each other:*

(a) $k(\hat{\mathbf{x}}) + k^*(\hat{\mathbf{p}}) = \hat{\mathbf{p}}^\mathsf{T}\hat{\mathbf{x}}$;
(b) $\hat{\mathbf{p}} \in \Delta k(\hat{\mathbf{x}})$ (or for differentiability $\hat{\mathbf{p}} = \nabla k(\hat{\mathbf{x}}) = \xi^{-1}(\hat{\mathbf{x}})$).

Furthermore, if k is closed, another condition can be added to the list:

(c) $\hat{\mathbf{x}} \in \Delta k^*(\hat{\mathbf{p}})$ (or $\hat{\mathbf{x}} = \nabla k^*(\hat{\mathbf{p}}) = \xi(\hat{\mathbf{p}})$ provided $\nabla k = \xi^{-1}$ is invertible),

in which case the superdifferential Δk^* is the inverse of Δk in the sense of set valued mappings. A pair $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ satisfying the conditions (a)–(c) is called a pair of dual points.

The proof is omitted because the result is merely a collection of Theorem 23.5 and Corollary 23.5.1 in Rockafellar (1972). The first condition has been explained graphically with respect to Figure 4.4 and was used to demonstrate the relationship between the willingness to pay k and the negative consumers' surplus k^*. Given the integrability of the inverse OD-demand functions ξ^{-1}, condition (a) is equivalent to having $\hat{\mathbf{p}} = \xi^{-1}(\hat{\mathbf{x}})$. Moreover, under the regularity condition of k being closed, we gain one special case of Proposition 4.20

$$\hat{\mathbf{p}} = \xi^{-1}(\hat{\mathbf{x}}) = \nabla k(\hat{\mathbf{x}}) \iff \hat{\mathbf{x}} = \xi(\hat{\mathbf{p}}) = \nabla k^*(\hat{\mathbf{p}})$$

because the demand function ξ is the inverse of ξ^{-1}. Without differentiability the analogous statement remains true, however, in a more general form. Condition (b) is most commonly used in the characterization of an optimal solution to the combined travel choice/trip assignment problem (4.31), while condition (c) is usually employed with respect to the equivalent problem (4.42). The latter will be shown in the discussion of (4.44) below.

Observing $\hat{\mathbf{x}} = \xi(\hat{\mathbf{p}})$ one may ask whether the willingness to pay k affects only trip choice, say $\hat{\mathbf{x}}$, or even route choice, say $\hat{\mathbf{h}}$. In order to answer this question, recall that a path flow is not necessarily unique given the trip table, i.e., $\hat{\mathbf{x}} = \mathbf{\Lambda}\hat{\mathbf{h}}$. Nevertheless, Corollary 4.21 shows that the conditions in Proposition 4.20 on a pair $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ pass on to a corresponding pair $(\hat{\mathbf{h}}, \hat{\tilde{\mathbf{w}}})$. All that must be done is to apply the preceding proposition to $(k\mathbf{\Lambda})$.

4.21 Corollary *If $(k\mathbf{\Lambda})$ is an n-proper closed concave function, then for any vector $\hat{\mathbf{h}}$ the following three conditions on a vector $\hat{\tilde{\mathbf{w}}}$ are equivalent to each other:*

(a) $(k\mathbf{\Lambda})(\hat{\mathbf{h}}) + (k\mathbf{\Lambda})^*(\hat{\tilde{\mathbf{w}}}) = \hat{\tilde{\mathbf{w}}}^\top \hat{\mathbf{h}}$;
(b) $\hat{\tilde{\mathbf{w}}} \in \Delta(k\mathbf{\Lambda})(\hat{\mathbf{h}})$ *(or for differentiability* $\hat{\tilde{\mathbf{w}}} = \nabla(k\mathbf{\Lambda})(\hat{\mathbf{h}}) = \mathbf{\Lambda}^\top \boldsymbol{\xi}^{-1}(\mathbf{\Lambda}\,\hat{\mathbf{h}})$*)*;
(c) $\hat{\mathbf{h}} \in \Delta(k\mathbf{\Lambda})^*(\hat{\tilde{\mathbf{w}}})$ *(or* $\hat{\mathbf{h}} = (\nabla(k\mathbf{\Lambda}))^{-1}(\hat{\tilde{\mathbf{w}}})$ *provided* $\nabla(k\mathbf{\Lambda})$ *is invertible).*

Disregarding that any trip table $\hat{\mathbf{x}}$ may be compatible with many path flows $\hat{\mathbf{h}}$, Proposition 4.20 lays emphasis on trip choice which is contrasted by Corollary 4.21 stressing route choice. The most important relationships may be visualized best by the following commutative diagram with quantity variables on the left hand side and price variables on the right hand side.

$$\begin{array}{ccccc} & \hat{\mathbf{x}} \to & k(\hat{\mathbf{x}}) & \to & \nabla k(\hat{\mathbf{x}}) = \hat{\mathbf{p}} \\ \hat{\mathbf{x}} = \mathbf{\Lambda}\,\mathbf{h} \;\Big|\; & & \| & & \Big|\; \mathbf{\Lambda}^\top \hat{\mathbf{p}} = \hat{\tilde{\mathbf{w}}} \\ & \hat{\mathbf{h}} \to & (k\mathbf{\Lambda})(\hat{\mathbf{h}}) & \to & \nabla(k\mathbf{\Lambda})(\hat{\mathbf{h}}) = \hat{\tilde{\mathbf{w}}} \end{array}$$

We may complete this diagram by adding corresponding terms with respect to k^* and $(k\mathbf{\Lambda})^*$, where the above cited conditions on pairs of dual points must be fulfilled.

(d) Average Cost Pricing

The analytical results obtained so far all serve to determine a Wardrop equilibrium under price sensitive demand. Assuming invertible demand functions the corresponding variational inequality problems have been introduced in (4.25) and (4.26). Provided the cost functions as well as the inverse demand functions are integrable, these problems were substituted by (4.29) or (4.31) which has been shown to be equivalent to (4.42). In accordance with (4.43) that minimax problem can be transformed again into[33]

$$\min_{\mathbf{h}} \max_{\mathbf{p}} \left\{ \oint_{\mathbf{0}}^{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{z})^\top d\mathbf{z} - \mathbf{p}^\top \mathbf{\Lambda}\,\mathbf{h} + \oint_{\mathbf{p}^0}^{\mathbf{p}} \boldsymbol{\xi}(\mathbf{z})^\top d\mathbf{z} \,\Big|\, \mathbf{h} \geqq \mathbf{0},\ \mathbf{p} \geqq \mathbf{0} \right\}. \tag{4.44}$$

This formulation has the advantage of avoiding the assumption of invertible demand functions, but for the moment we stick to the supposition of integrability. Carey (1987) confronts the preceding "quantity-price" formulation with the initial "quantity-quantity" formulation in (4.31) in a pure minimum problem. Nevertheless, the thorough analysis of conjugate functions confirms that the objective function is to be maximized over prices. Observe furthermore that the objective function in (4.44), say $\mathcal{L}$, is convex in $\mathbf{h}$ and concave in $\mathbf{p}$ by assumption. If $\mathcal{L}$ is also differentiable in both $\mathbf{h}$ and $\mathbf{p}$, Theorem 2.4.1 in Walk (1989) applies. The following result on a user optimal trip choice/traffic assignment generalizes the statement of Remark 3.40 on p. 150 with respect to a price sensitive demand for traffic.

The pair $(\hat{\mathbf{h}}, \hat{\mathbf{p}})$ is a saddle point of $\mathcal{L}$

$$\mathcal{L}(\hat{\mathbf{h}}, \mathbf{p}) \leqq \mathcal{L}(\hat{\mathbf{h}}, \hat{\mathbf{p}}) \leqq \mathcal{L}(\mathbf{h}, \hat{\mathbf{p}}) \qquad \forall \mathbf{h},\ \forall \mathbf{p}$$

[33] Be aware that this problem is *not* the frequently used version of Aashtiani (1979) as can be found in Fernandez, Friesz (1983, problem (16)).

if and only if

$$\nabla_{\mathbf{h}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\mathbf{p}}) = \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}} \geqq \mathbf{0}, \quad \big(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}}\big)^{\mathsf{T}}\hat{\mathbf{h}} = 0, \quad \hat{\mathbf{h}} \geqq \mathbf{0}, \tag{4.45a}$$

$$\nabla_{\mathbf{p}}\mathcal{L}(\hat{\mathbf{h}}, \hat{\mathbf{p}}) = \boldsymbol{\xi}(\hat{\mathbf{p}}) - \mathbf{\Lambda}\,\hat{\mathbf{h}} \leqq \mathbf{0}, \quad \big(\boldsymbol{\xi}(\hat{\mathbf{p}}) - \mathbf{\Lambda}\,\hat{\mathbf{h}}\big)^{\mathsf{T}}\hat{\mathbf{p}} = 0, \quad \hat{\mathbf{p}} \geqq \mathbf{0}. \tag{4.45b}$$

The first line of this nonlinear complementarity problem simply restates Wardrop's first principle, cf. Definition 4.13. That is, for any route $\rho \in \mathcal{R}(s,t)$,

$$\tilde{c}_\rho(\hat{\mathbf{h}}) > \hat{p}_{(s,t)} \implies \hat{h}_\rho = 0, \tag{4.46a}$$

$$\hat{h}_\rho > 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) = \hat{p}_{(s,t)}. \tag{4.46b}$$

The second line requires particularly that the generated traffic $x^S_{(s,t)}$ must match the demand $x^D_{(s,t)}$ for $\hat{p}_{(s,t)} > 0$. In the case of a *positive excess supply* one has $\hat{p}_{(s,t)} = 0$.

$$\xi_{(s,t)}(\hat{\mathbf{p}}) < \sum_{\rho \in \mathcal{R}(s,t)} \hat{h}_\rho \implies \hat{p}_{(s,t)} = 0 \tag{4.47a}$$

$$\hat{p}_{(s,t)} > 0 \implies \xi_{(s,t)}(\hat{\mathbf{p}}) = \sum_{\rho \in \mathcal{R}(s,t)} \hat{h}_\rho \tag{4.47b}$$

Patriksson (1994, p. 85) refers to the case where always market clearing prevails at some feasible minimum cost vector $\hat{\mathbf{p}}$, so that particularly (4.45b) reduces to $\mathbf{x}^D = \boldsymbol{\xi}(\hat{\mathbf{p}}) = \mathbf{\Lambda}\,\mathbf{h}$. Applying Theorem 3.26 with respect to (4.45a), we gain a restatement of the variational inequality problem (4.27) at which we have started. Without presuming market clearing, (4.45a) induces together with (4.45b) a new variational inequality problem of the form

$$\big(\tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \mathbf{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}}\big)^{\mathsf{T}}(\mathbf{h} - \hat{\mathbf{h}}) + \big(\mathbf{\Lambda}\hat{\mathbf{h}} - \boldsymbol{\xi}(\hat{\mathbf{p}})\big)^{\mathsf{T}}(\mathbf{p} - \hat{\mathbf{p}}) \geqq 0 \quad \forall\, (\mathbf{h}, \mathbf{p}) \geqq \mathbf{0}. \tag{4.48}$$

Such a formulation can be found in Fisk, Boyce (1983a) and it is discussed in further detail in the subsequent section. It has the advantage that neither integrability nor invertibility is needed.

4.22 Remark (On the meaning of p) Up until now $\mathbf{p}$ has been interpreted as minimum route cost vector defined by (4.36), p. 203. This is in almost all cases correct, but in view of (4.45b) there is an exception which can be explained in light of Figure 4.5. The demand for trips on route ρ_1 is zero if a cheaper alternative route ρ_2 exists: $\tilde{c}_1 > \tilde{c}_2$ with $\rho_1 \sim \rho_2$. In Figure 4.5 all points on the demand curve h_1^D above $\tilde{c}_2$ are mapped to the point $(h_1 = 0, \tilde{c}_2)$. In the special case of $\tilde{c}_2 = \tilde{c}_1$ the demand is distributed between both routes so that all points on the horizontal line are admissible. When $\tilde{c}_1 < \tilde{c}_2$ all points on h_1^D are possible. For instance, look at the point $(\tilde{c}_1, h_1) = (p_{(s,t)}, x^{D0}_{(s,t)})$, where all of the demand is routed along ρ_1 implying $h_2 = 0$. The last case refers to a *positive excess supply*, where more trips are generated than needed and the marginal willingness to pay drops to zero. Given $h_1 > x^{D1}_{(s,t)}$ the value of a trip is set to $p_{(s,t)} = 0$. Hence, all points on h_1^D beyond $x^{D1}_{(s,t)}$ are mapped to the horizontal axis.

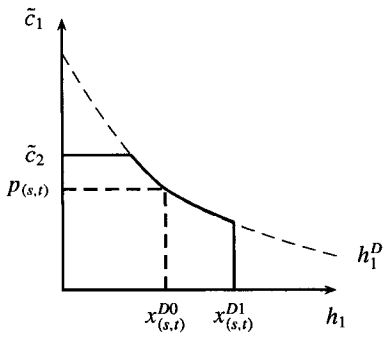


Fig. 4.5
Equilibrium flows on route ρ_1
The example refers to two alternative routes, namely, ρ_1 and ρ_2, connecting the nodes v_s and v_t with the demand $x_{(s,t)}^D = h_1 + h_2$.

At least when $\tilde{\mathbf{c}} > \mathbf{0}$ is independent of $\mathbf{h}$, the last case seems to contradict the above definition of $\mathbf{p}$ as minimum route cost vector because no $p_{(s,t)}$ can be zero and positive at the same time. However, (4.47a) requires $\hat{h}_\rho > 0$ for at least one route $\rho \in \mathcal{R}(s,t)$ contradicting (4.46b) so that $(\hat{\mathbf{h}}, \hat{\mathbf{p}})$ cannot be optimal. The only case in which $\hat{p}_{(s,t)}$ falls short of minimum travel costs is $\xi_{(s,t)}(\hat{\mathbf{p}}) = \sum_{\rho \in \mathcal{R}(s,t)} \hat{h}_\rho = 0$ and $\hat{p}_{(s,t)} = 0$. We shall keep this case of minor importance in mind and stick to the above interpretation of $\mathbf{p}$. □

The rest of this section completes the analysis with regard to the initial "quantity-quantity" formulation in (4.31) and the "quantity-price" formulation in (4.42). Consequently one may ask for a pure "price-price" expression depending on path cost $\tilde{\mathbf{w}} = \tilde{\mathbf{c}}(\mathbf{h})$ and the marginal willingness to pay $\mathbf{p} = \boldsymbol{\xi}^{-1}(\mathbf{x})$. In fact all that must be done is suggested by (4.42): compute the conjugate convex function

$$(g\boldsymbol{\Delta})^*(\tilde{\mathbf{w}}) := \sup\left\{\tilde{\mathbf{w}}^\mathsf{T}\mathbf{h} - (g\boldsymbol{\Delta})(\mathbf{h})\right\} \quad \text{with} \quad \tilde{\mathbf{w}} = \boldsymbol{\Lambda}^\mathsf{T}\mathbf{p}$$

and substitute the result into (4.42) to obtain a maximum problem over the variables $\mathbf{p}$. The discussion of that problem goes beyond the scope of this book and can be found in slightly different form in Carey (1985). Hence, we just report the relevant properties of the pairs (g, g^*) and $\left((g\boldsymbol{\Delta}), (g\boldsymbol{\Delta})^*\right)$ in the subsequent two corollaries. In doing so observe that Proposition 4.20 has a "convex" counterpart with respect to the function g and its subdifferential ∂g.

4.23 Corollary *If g is a proper closed convex function, then for any link flow $\hat{\boldsymbol{\varphi}}$ the following three conditions on a* ***link specific price vector*** *$\hat{\mathbf{w}}$ are equivalent to each other:*

(a) $g(\hat{\boldsymbol{\varphi}}) + g^*(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^\mathsf{T}\hat{\boldsymbol{\varphi}}$;
(b) $\hat{\mathbf{w}} \in \partial g(\hat{\boldsymbol{\varphi}})$;
(c) $\hat{\boldsymbol{\varphi}} \in \partial g^*(\hat{\mathbf{w}})$.

If g is differentiable with $\hat{\mathbf{w}} = \nabla g(\hat{\boldsymbol{\varphi}}) = \mathbf{c}(\hat{\boldsymbol{\varphi}})$ and if the gradient ∇g is invertible such that $\hat{\boldsymbol{\varphi}} = \mathbf{c}^{-1}(\hat{\mathbf{w}})$, then (b) and (c) take the form

$$\hat{\mathbf{w}} = \mathbf{c}(\hat{\boldsymbol{\varphi}}) \iff \hat{\boldsymbol{\varphi}} = \mathbf{c}^{-1}(\hat{\mathbf{w}}). \tag{4.49}$$

The equivalent condition (a) suggests that the newly defined term g^* is something like the producers' surplus provided g would be based on marginal link cost rather

than average link cost. The reader may verify that statement for the special case of separable link cost functions,

$$g(\boldsymbol{\varphi}) = \sum_{e \in \mathcal{E}} g_e(\varphi_e) \quad \text{with} \quad g_e(\varphi_e) := \int_0^{\varphi_e} c_e(z_e)\, \mathrm{d}z_e.$$

Indeed one can show that the concept of conjugate functions even applies to sums of functions so that condition (a) turns out to be

$$\sum_{e \in \mathcal{E}} g_e(\hat{\varphi}_e) + \sum_{e \in \mathcal{E}} g_e^*(\hat{w}_e) = \sum_{e \in \mathcal{E}} \hat{w}_e\, \hat{\varphi}_e.$$

Figure 4.6 depicts a graphic representation of g_e^* with respect to one link e included in the vector pair $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{w}})$. In accordance with Figure 4.4, the graph of c_e illustrates $g_e' = c_e$. The difference between the rectangle $\hat{w}_e\hat{\varphi}_e$ and the area below the cost curve $g_e(\hat{\varphi}_e)$ corresponds to $g_e^*(\hat{w}_e)$.

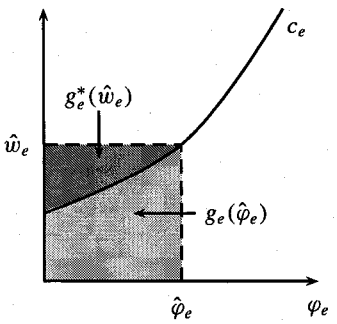


Fig. 4.6
Conjugate convex g_e^* of g_e
The figure represents the same case as in Figure 4.4 with the exception of an increasing function $g_e' = c_e$.

The link specific formulation in Corollary 4.23 has a route specific counterpart which is more appropriate with regard to travelers' path choice. The facts needed for that rearrangement are $\boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}$ and $g(\boldsymbol{\Delta}\,\mathbf{h}) \equiv (g\boldsymbol{\Delta})(\mathbf{h})$. Taking (4.49) into account, link costs $\mathbf{w} = \mathbf{c}(\boldsymbol{\varphi})$ induce route cost by $\tilde{\mathbf{w}} = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{w} = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h}) = \tilde{\mathbf{c}}(\mathbf{h})$. With that Corollary 4.23 can be applied to $(g\boldsymbol{\Delta})$ rather than g.

4.24 Corollary *If $(g\boldsymbol{\Delta})$ is a proper closed convex function, then for any vector $\hat{\mathbf{h}}$ the following three conditions on a* ***route specific price vector*** *$\hat{\tilde{\mathbf{w}}}$ are equivalent to each other:*

(a) $(g\boldsymbol{\Delta})(\hat{\mathbf{h}}) + (g\boldsymbol{\Delta})^*(\hat{\tilde{\mathbf{w}}}) = \hat{\tilde{\mathbf{w}}}^{\mathsf{T}}\hat{\mathbf{h}}$;
(b) $\hat{\tilde{\mathbf{w}}} \in \partial(g\boldsymbol{\Delta})(\hat{\mathbf{h}})$;
(c) $\hat{\mathbf{h}} \in \partial(g\boldsymbol{\Delta})^*(\hat{\tilde{\mathbf{w}}})$.

Although route costs are almost always not separable, condition (a) has the same interpretation as provided in Figure 4.6. Assuming $(g\boldsymbol{\Delta})$ to be differentiable, condition (b) confirms the equality of route prices $\hat{\tilde{\mathbf{w}}}$ and route cost $\tilde{\mathbf{c}}$:

$$\hat{\tilde{\mathbf{w}}} = \nabla(g\boldsymbol{\Delta})(\hat{\mathbf{h}}) = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\hat{\mathbf{h}}) = \tilde{\mathbf{c}}(\hat{\mathbf{h}})$$

Furthermore, if $\nabla(g\boldsymbol{\Delta})$ is invertible, one has

$$\hat{\tilde{\mathbf{w}}} = \tilde{\mathbf{c}}(\hat{\mathbf{h}}) \iff \hat{\mathbf{h}} = \tilde{\mathbf{c}}^{-1}(\hat{\tilde{\mathbf{w}}}).$$

by the equivalence of (b) and (c).

4.25 Remark (Summary) Under certain regularity conditions, a series of equivalent programs exists that characterize a Wardrop equilibrium for the combined trip choice/traffic assignment problem. Leaving the "well behaved" case, some appealing programs must be dropped from the list of programs. For example, if the demand functions ξ cannot be inverted or if ξ^{-1} is not integrable, the "quantity-quantity" problem (4.31) becomes void and all we can do is to stick to the "quantity-price" problem in (4.44). Similarly, a pure "price-price" formulation requires invertible cost functions. Differentiability, however, is of minor importance. It assures merely that the subdifferential, e.g., ∂g, or the superdifferential, e.g., Δk, consists of one and only one element at the respective point of interest. For numerical purposes this gradient has the favorable property that it indicates the direction of steepest ascent. □

(e) Fenchel Duality

This and the subsequent section give a brief overview on duality theory with special regard to the trip choice/traffic assignment problem. The results restate indeed what we have obtained so far, but lay emphasis on different technical properties. Readers who are not interested in these details may jump to Section 4.3.2.

Almost all of the above reasoning focuses on the conjugate concave of k or $(k\boldsymbol{\Lambda})$ because the aim was to find a substitute for k in the case of non-invertible demand functions. In fact, the starting point (4.31) recommends to analyze another relationship which follows from $(g\boldsymbol{\Delta})$ and $(k\boldsymbol{\Lambda})$ both of which are functions of route flows $\mathbf{h}$. Let $(g\boldsymbol{\Delta})$ be proper convex and $(k\boldsymbol{\Lambda})$ be n-proper concave and confront the infimum problem (4.31) with a supremum problem as follows.

$$\begin{aligned} &\inf_{\mathbf{h}} \{(g\boldsymbol{\Delta})(\mathbf{h}) - (k\boldsymbol{\Lambda})(\mathbf{h})\} \\ &\sup_{\tilde{\mathbf{w}}} \left\{(k\boldsymbol{\Lambda})^*(\tilde{\mathbf{w}}) - (g\boldsymbol{\Delta})^*(\tilde{\mathbf{w}})\right\} \end{aligned} \tag{4.50}$$

Both programs are attractive because the infimum problem depends solely on the quantity variable $\mathbf{h}$ and the supremum problem shows only the corresponding price variable $\tilde{\mathbf{w}}$. Hence, a user optimal trip assignment can be described completely by a vector of appropriate route prices. On the aforementioned assumptions, *Fenchel's Duality Theorem* (cf. Rockafellar (1972, Theorem 31.1)) proposes that the infimum and the supremum in (4.50) are equal if either of the following conditions is satisfied:[34]

[34] The effective domain Dom g of a proper function g denotes the set of points at which g is finite. The same applies to n-Dom k of an n-proper function k. The relative interior rint C of a convex set C is defined by the interior of C regarding its affine hull. For example, the relative interior of a line in $\mathbb{R}^3$ is the line without its endpoints. See Rockafellar (1972) for further details.

1. $\text{rint}(\text{Dom}(g\mathbf{\Delta})) \cap \text{rint}(\text{n-Dom}(k\mathbf{\Lambda})) \neq \emptyset$
2. $(g\mathbf{\Delta})$ and $(k\mathbf{\Lambda})$ are closed and $\text{rint}(\text{Dom}(g\mathbf{\Delta})^*) \cap \text{rint}(\text{n-Dom}((k\mathbf{\Lambda})^\star)) \neq \emptyset$

If both conditions hold good, the infimum and supremum are necessarily finite. The two conditions require by intuition that the corresponding effective domains overlap sufficiently. Some further applications of Fenchel's Duality Theorem as well as further mathematical details can be found in Ekeland, Témam (1999). Their Proposition 4.1 combines the Corollaries 4.21 and 4.24.

4.26 Proposition *The following two conditions are equivalent to each other:*

(i) The infimum and the supremum in (4.50) *are equal and* $\hat{\mathbf{h}}$ *and* $\hat{\tilde{\mathbf{w}}}$ *are solutions to the infimum and the supremum problem, respectively.*

(ii) $\hat{\mathbf{h}}$ *and* $\hat{\tilde{\mathbf{w}}}$ *satisfy the extremality condition*

$$(g\mathbf{\Delta})(\hat{\mathbf{h}}) + (g\mathbf{\Delta})^*(\hat{\tilde{\mathbf{w}}}) = (k\mathbf{\Lambda})(\hat{\mathbf{h}}) + (k\mathbf{\Lambda})^\star(\hat{\tilde{\mathbf{w}}})$$

which is equivalent to

$$\hat{\tilde{\mathbf{w}}} \in \partial(g\mathbf{\Delta})(\hat{\mathbf{h}}) \quad \textit{and} \quad \hat{\tilde{\mathbf{w}}} \in \Delta(k\mathbf{\Lambda})(\hat{\mathbf{h}})$$

Notice that the equivalence in (ii) is an immediate consequence of Corollaries 4.21 and 4.24. Subtracting $\hat{\mathbf{h}}^\mathsf{T}\hat{\tilde{\mathbf{w}}}$ on both sides, shows that the left hand side is non-negative while the right hand side in non-positive. Hence, both sides are equal to zero and the two corollaries apply.

Just for the sake of completeness, the duality of the pair (4.50) can be transferred immediately to the initial problem (4.31). The adjusted pair of dual programs reads

$$\begin{aligned} &\inf_{\mathbf{h},\mathbf{x}^D} \left\{(g\mathbf{\Delta})(\mathbf{h}) - k(\mathbf{x}^D) |\ \mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D\right\}, \\ &\sup_{\tilde{\mathbf{w}},\mathbf{p}} \left\{k^\star(\mathbf{p}) - (g\mathbf{\Delta})^*(\tilde{\mathbf{w}}) |\ \mathbf{\Lambda}^\mathsf{T}\mathbf{p} = \tilde{\mathbf{w}}\right\}, \end{aligned} \tag{4.51}$$

where, by analogy to (4.41), $(g\mathbf{\Delta})^*$ satisfies

$$(g\mathbf{\Delta})^*(\tilde{\mathbf{w}}) = \inf_{\mathbf{w}} \left\{g^*(\mathbf{w}) |\ \mathbf{\Delta}^\mathsf{T}\mathbf{w} = \tilde{\mathbf{w}}\right\}.$$

In relation to (4.29) and (4.50) this expression gives rise to state another pair of dual programs

$$\begin{aligned} &\inf_{\mathbf{h},\boldsymbol{\varphi}} \{g(\boldsymbol{\varphi}) - (k\mathbf{\Lambda})(\mathbf{h}) |\ \mathbf{\Delta}\,\mathbf{h} = \boldsymbol{\varphi}\} \\ &\sup_{\tilde{\mathbf{w}},\mathbf{w}} \left\{(k\mathbf{\Lambda})^\star(\tilde{\mathbf{w}}) - g^*(\mathbf{w}) |\ \mathbf{\Delta}^\mathsf{T}\mathbf{w} = \tilde{\mathbf{w}}\right\}, \end{aligned} \tag{4.52}$$

a variation of which can be found in Fukushima (1984). He solves the supremum problem for a special functional form of separable link cost functions $c_e(\varphi_e)$ and a fixed traffic demand by virtue of a subgradient method.

(f) Generalized Duality

As the conditions of Fenchel's Duality Theorem in the sequel of (4.50) are not very comfortable, it is worthwhile to have alternative conditions. Following Walk (1989, Section 2.3), let the cost term $(g\Delta)$ and the willingness to pay k be defined on H and K, respectively. Consider the *negative* objective function in (4.44) and define the Lagrangean form $\mathcal{K} := -\mathcal{L}$ by

$$\mathcal{K}(\mathbf{h}, \mathbf{p}) = \mathbf{p}^{\mathsf{T}} \Lambda\, \mathbf{h} - (g\Delta)(\mathbf{h}) - k^{\star}(\mathbf{p}).$$

Furthermore, define the *(generalized) conjugate functions*

$$(g\Delta)^{\circledast}(\mathbf{p}) = \sup_{\mathbf{h}} \left\{\mathbf{p}^{\mathsf{T}} \Lambda\, \mathbf{h} - (g\Delta)(\mathbf{h})\right\},$$
$$k^{\star\circledast}(\mathbf{h}) = \inf_{\mathbf{p}} \left\{\mathbf{p}^{\mathsf{T}} \Lambda\, \mathbf{h} - k^{\star}(\mathbf{p})\right\}.$$

Caution: $(g\Delta)^{*}$ is a function of route specific prices $\tilde{\mathbf{w}}$, while $(g\Delta)^{\circledast}$ is a function of OD-specific prices $\mathbf{p}$. Nevertheless, one has $(g\Delta)^{*}(\tilde{\mathbf{w}}) = (g\Delta)^{\circledast}(\mathbf{p})$ whenever $\tilde{\mathbf{w}} = \Lambda^{\mathsf{T}}\mathbf{p}$. A similar observation holds true for $k = k^{\star\star}$ and $k^{\star\circledast}$. Let the supremum exist for all $\mathbf{p} \in \hat{K} \subset K$ and the infima for all $\mathbf{h} \in \hat{H} \subset H$. Then, similar to Fenchel's Duality Theorem,

$$\sup\left\{k^{\star\circledast}(\mathbf{h}) - (g\Delta)(\mathbf{h})\right\} \quad \text{and} \quad \inf\left\{(g\Delta)^{\circledast}(\mathbf{p}) - k^{\star}(\mathbf{p})\right\}$$

constitute a pair of dual problems and Theorem 2.3.1 in Walk (1989) states

4.27 Theorem *The following three statements are equivalent:*

1. *There is a pair* $(\hat{\mathbf{h}}, \hat{\mathbf{p}}) \in H \times K$ *such that for any* $(\mathbf{h}, \mathbf{p}) \in H \times K$ *one has*

$$\mathcal{K}(\mathbf{h}, \hat{\mathbf{p}}) \leqq \mathcal{K}(\hat{\mathbf{h}}, \hat{\mathbf{p}}) \leqq \mathcal{K}(\hat{\mathbf{h}}, \mathbf{p}).$$

2. *The following relationships hold true*

$$\inf_{\mathbf{p}\in K} \mathcal{K}(\hat{\mathbf{h}}, \mathbf{p}) = \min_{\mathbf{p}\in K} \mathcal{K}(\hat{\mathbf{h}}, \mathbf{p}) = k^{\star\circledast}(\hat{\mathbf{h}}) - (g\Delta)(\hat{\mathbf{h}}) = \mathcal{K}(\hat{\mathbf{h}}, \hat{\mathbf{p}}),$$
$$\sup_{\mathbf{h}\in H} \mathcal{K}(\mathbf{h}, \hat{\mathbf{p}}) = \max_{\mathbf{h}\in H} \mathcal{K}(\mathbf{h}, \hat{\mathbf{p}}) = (g\Delta)^{\circledast}(\hat{\mathbf{p}}) - k^{\star}(\hat{\mathbf{p}}) = \mathcal{K}(\hat{\mathbf{h}}, \hat{\mathbf{p}}).$$

3. *A pair* $(\hat{\mathbf{h}}, \hat{\mathbf{p}}) \in \hat{K} \times \hat{H}$ *exists such that*

$$\hat{\mathbf{p}}^{\mathsf{T}} \Lambda\, \hat{\mathbf{h}} = (g\Delta)(\hat{\mathbf{h}}) + (g\Delta)^{\circledast}(\hat{\mathbf{p}}),$$
$$\hat{\mathbf{p}}^{\mathsf{T}} \Lambda\, \hat{\mathbf{h}} = k^{\star}(\hat{\mathbf{p}}) + k^{\star\circledast}(\hat{\mathbf{h}}).$$

To give an intuition on the functions involved look at Figure 4.7; it picks up the measurement of the net benefit in Figure 4.3, p. 199. The decreasing curve indicates the (inverse) demand for trips between the OD-pair (v_s, v_t). The increasing curve gives the cost per trip on the only route ρ connecting (v_s, v_t). Each of the functions $(g\Delta)$, $(g\Delta)^{\circledast}$, $k^{\star}$, and $k^{\star\circledast}$ is characterized by a distinct area. The particular regions

$-k^*(\hat{\mathbf{p}})$ and $(g\boldsymbol{\Delta})^*(\hat{\mathbf{p}})$ have been identified as consumers' surplus and producers' surplus, respectively, were we have to remind the caveats given in the text due to the average cost curve $\tilde{c}_\rho$. The value $\hat{\mathbf{p}}^\mathsf{T}\boldsymbol{\Lambda}\hat{\mathbf{h}}$ is represented by the dashed rectangle.

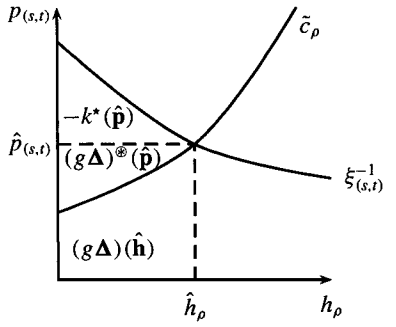


Fig. 4.7
Duality
The area $k^{*\circledast}(\hat{\mathbf{h}})$ is not indicated but may be calculated by
$(g\boldsymbol{\Delta})(\hat{\mathbf{h}}) + (g\boldsymbol{\Delta})^{\circledast}(\hat{\mathbf{p}}) - k^*(\hat{\mathbf{p}})$

(g) Summary

The analysis of user equilibrium with price sensitive demand for traffic rests upon programs the optimum conditions of which describe the behavior of network users. Route choice then follows in accordance with Wardrop's first principle as given in Definition 4.13. On the basis of the two regularity assumptions of invertible demand functions and integrability, Beckmann et al. (1956) introduced the artificial, convex problem (4.29) where the objective function describes the difference between the aggregate willingness to pay k and a term g that denotes the line integral of link cost functions. Thus g would indicate total travel cost if we would use marginal link cost functions rather than average link costs. In this case (4.29) goes over into a problem of maximizing the net benefits of a traffic flow pattern. Without integrability the convex program is to be substituted by one of the two variational inequalities to be found in Proposition 4.14. Any solution corresponds to a Wardrop equilibrium, but an overall objective function as described above does not exist any longer. When the demand functions cannot be inverted, alternative formulations are needed in order to find user equilibria. One first approach transforms the quantity-quantity formulation in (4.29) into an equivalent price-quantity minimax problem, cf. (4.42). The same steps lead to a pure price-price statement as given in (4.51). We have shown by virtue of duality theory that these problems are equivalent under certain regularity conditions. If these conditions fail to hold true, one cannot expect that duality results prevail. Nevertheless, we can now choose that formulation which fits best to the information that can be observed in reality.

4.3 Special Issues

4.3.1 Preliminary Remarks

Chapter 3 will now be closed by several approaches that extend the results so far obtained. Starting with the variational inequality problem (4.48), the following

Section 4.3.2 suggests an approach – namely, *bilevel programming* – how to simplify this problem by separating it into two interdependent but smaller problems. This attempt is contrasted in Section 4.3.3 with a much more extensive approach of considering several distinct classes of network users sharing one network. The formulation demonstrates the power of the variational inequality approach in the sense that merely a renaming of variables suffices to capture the new problem.

Section 4.3.4 ignores the aspect of a price sensitive demand for trips. Instead, we hand in some basic features of traffic assignment that must be taken into account when all movements on the network are characterized by features of space and time. Queuing is presumably the most important cause for delays so that an appropriate traffic assignment should not ignore this particular problem of route choice.

The last extension deals with a completely different issue of network theory, namely, spatial price equilibria. The understanding of this problem is crucial for the explanation of traffic demand. Households, for example, make trips to shopping centers with the cheapest available commodity supply, where their expenditures include the costs of a trip. Correspondingly, firms must take into account costs of delivery between places of production and consumer destinations. As a consequence, firms as well as households have strong incentives to find optimal locations what in turn has a lasting effect on future trip tables. The described problem of finding spatial price equilibria will be solved again by virtue of a variational inequality formulation.

4.3.2 Bilevel Programming

The preceding Section 4.2 has characterized user optimal solutions to the combined problem of travel choice and trip assignment. Bilevel programming serves for the current setting as one approach to separate the two interdependent subproblems. The minimum route costs $\mathbf{p}$ can then be interpreted as market prices that determine the demand for traffic $\mathbf{x}^D$. Increasing or decreasing the price $p_{(s,t)}$ of a trip from v_s to v_t consequently leads to a corresponding traffic diminution or creation. This is the first subproblem, namely, travel choice. It is contrasted with route choice which is represented by a path flow $\mathbf{h}$. In accordance with Wardrop's first principle this flow results from travelers' behavior by which always the cheapest among alternative routes are selected. Regarding route choice a change of $p_{(s,t)}$ redirects path flows, what is referred to as traffic diversion. Observe also that each traffic assignment results in a distinct route cost vector $\tilde{\mathbf{c}}$ which in turn determines $\mathbf{p}$.

4.28 Remark (Principle of Bilevel Programming) Bilevel programming has two sets of variables as has been suggested in Section 3.2.3.1 (f), p. 119. The two sets are represented by x and y in the following example. Let us start with the general minimum problem

$$\min\{f(x, y) |\ x \in X,\ y \in Y,\ h(x, y) = 0\}.$$

Bilevel programming deals with the case in which one set of variables, say y, is constrained to be a solution of an inner optimization problem,

$$\min\left\{f(x, y) |\ x \in X,\ h(x, y) = 0,\ y \in \arg\max\{F(x, y) |\ y \in S(x)\}\right\},$$

where $S(x)$ is some subset of Y; see, e.g., Boyce, Daskin (1997) whose example of designing a network will be reproduced in the next chapter. The inner maximum problem is referred to as *lower level problem*, while the outer minimum problem is called the *upper level problem*. This separation principle is particularly appealing when the decision-maker of the minimum problem is able to compute the reaction of his opponent, who decides on the maximum problem. Notice, however, that the definition of $S(x)$ may be crucial when describing the conjectural behavior of the second decision-maker. For example, he can take the choice of x as being fixed (von Stackelberg follower) or he computes in turn some conjectural reaction of the first agent before making his own decision. □

This principle of splitting one problem into two smaller but interdependent problems can now be applied to our combined travel choice/trip assignment problem, cf. Bell, Iida (1997, p. 156 ff.). For the current setting the variational inequality (4.48) suggests two interdependent problems, which are quite similar to the nonlinear complementarity problem of Remark 3.40, p. 150, with the exception that the fixed trip table $\mathbf{x}^D$ is now substituted by a price sensitive demand for traffic.

$$\begin{pmatrix} \tilde{\mathbf{c}}(\hat{\mathbf{h}}) - \boldsymbol{\Lambda}^{\mathsf{T}}\hat{\mathbf{p}} \\ \boldsymbol{\Lambda}\,\hat{\mathbf{h}} - \boldsymbol{\xi}(\hat{\mathbf{p}}) \end{pmatrix}^{\mathsf{T}} \left(\begin{pmatrix} \mathbf{h} \\ \mathbf{p} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{h}} \\ \hat{\mathbf{p}} \end{pmatrix} \right) \geqq 0 \qquad \forall \begin{pmatrix} \mathbf{h} \\ \mathbf{p} \end{pmatrix} \geqq \mathbf{0}.$$

The upper problem is that of finding a user optimal trip assignment $\hat{\mathbf{h}}$ given a price vector $\hat{\mathbf{p}}$ which solves the lower problem. This second problem in turn is to find an optimal price vector $\hat{\mathbf{p}}$ such that the given route flow $\hat{\mathbf{h}}$ which solves the first problem determines a market equilibrium where the generated traffic $\mathbf{x}^S = \boldsymbol{\Lambda}\,\hat{\mathbf{h}}$ matches the demand for traffic $\mathbf{x}^D = \boldsymbol{\xi}(\hat{\mathbf{p}})$. The optimum conditions have been presented in (4.46) and (4.47). They show explicitly the interdependence of prices $\hat{\mathbf{p}}$ and path flows $\hat{\mathbf{h}}$ so that there is no need to go into further technical details at this point. Taking the two problems apart, it is common practice to interpret the upper one as a problem of price taking network users choosing among routes. In comparison the market may be identified with the decision-maker of the lower problem. It adjusts prices in a manner as if he seeks to minimize the value of any excess supply.

Such an endogenous price mechanism has the widely discussed disadvantage that travelers perceive route cost rather than marginal cost which they impose on the network. Hence, one might think of an exogenous authority in order to correct this problem. Yang, Lam (1996, 1997) describe a corresponding problem of bilevel programming where the upper level program determines a link specific toll pattern $\mathbf{t}$ for a capacitated network subject to a user optimal travel choice and trip assignment. On the assumption of invertible traffic demand functions, a user optimum is described by a pair $(\hat{\mathbf{x}}, \hat{\boldsymbol{\varphi}})$ which depends on the toll pattern $\mathbf{t}$ by virtue of travelers' perceived link cost $\check{\mathbf{c}}(\boldsymbol{\varphi}, \mathbf{t}) = \mathbf{c}(\boldsymbol{\varphi}) + \mathbf{t}$. The objective of the upper level program may be the maximization of the net benefit of all network participants or just the maximization of total toll revenue, which is the same as profit maximization because there are only fixed costs of operating the network.

4.3.3 Multiclass-User Traffic Assignment

The combined travel choice/trip assignment model is extended in Nagurney, Dong (2002a) with respect to a multiclass-user traffic network equilibrium. Ignoring their class specific (generalized) travel cost functions (see Remark 3.44) the idea can be grasped as follows. Let j denote a class of travelers out of a set of classes $\mathcal{J}$. Each class is characterized by its own invertible demand function $\mathbf{x}_j^D = \boldsymbol{\xi}_j(\mathbf{p})$, where $\mathbf{p}$ is the vector of cheapest route costs paid by all trip-makers irrespective of their class. The collection of demand vectors is denoted $\boldsymbol{\varkappa}^D = (\mathbf{x}_1^D, \mathbf{x}_2^D, \ldots)$. In terms of class specific route flows the traffic demand is given by $\mathbf{x}_j^D = \boldsymbol{\Lambda}\,\mathbf{h}_j$, where total route flows are $\mathbf{h} = \sum_{j\in\mathcal{J}} \mathbf{h}_j$. The collection of route flows is denoted $\mathfrak{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots)$. The same is supposed for link flows, $\boldsymbol{\varphi}_j = \boldsymbol{\Delta}\,\mathbf{h}_j$ and $\boldsymbol{\varphi} = \sum_{j\in\mathcal{J}} \boldsymbol{\varphi}_j$. On this basis (4.26) becomes

$$\sum_{j\in\mathcal{J}} \left[\tilde{\mathbf{c}}(\hat{\mathbf{h}})^\top(\mathbf{h}_j - \hat{\mathbf{h}}_j) - \boldsymbol{\xi}_j^{-1}(\hat{\mathbf{x}}_j)^\top(\mathbf{x}_j - \hat{\mathbf{x}}_j)\right] \geqq 0 \qquad \forall(\mathfrak{h}, \boldsymbol{\varkappa}) \in H_{\boldsymbol{\varkappa}}^S$$

$$\text{with} \quad H_{\boldsymbol{\varkappa}}^S := \left\{(\mathfrak{h}, \boldsymbol{\varkappa}) \mid \mathbf{x}_j = \boldsymbol{\Lambda}\mathbf{h}_j,\ \mathbf{x}_j \geqq \mathbf{0},\ \mathbf{h}_j \geqq \mathbf{0}, \mathbf{h} = \textstyle\sum_j \mathbf{h}_j,\ j \in \mathcal{J}\right\}$$

The corresponding conditions of a generalized Wardrop equilibrium are

$$\hat{h}_{\rho j} > 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) = \xi_{(s,t)j}^{-1}(\hat{\mathbf{x}}_j),$$
$$\hat{h}_{\rho j} = 0 \implies \tilde{c}_\rho(\hat{\mathbf{h}}) \geqq \xi_{(s,t)j}^{-1}(\hat{\mathbf{x}}_j).$$

If travelers of class j choose route ρ, the minimum route cost for the OD-pair at hand, say $\tilde{c}_\rho(\hat{\mathbf{h}})$, equals the j^{th} class' marginal willingness to pay for that trip, say $\xi_{(s,t)j}^{-1}(\hat{\mathbf{x}}_j)$. Nagurney, Dong (2002a) consider also some sort of price differentiation, where each class bears route cost $\tilde{c}_{\rho j}$ according to their willingness to pay. Then each class has its own price system. Irrespective of this modification, the existence[35] of a user optimal traffic assignment and its uniqueness can be proved by properties of $\tilde{\mathbf{c}}$ as given in Theorem 3.28 and Theorem 3.19, respectively. Nagurney et al. (2002b) lay special emphasis on class specific multi-criteria cost functions with variable weights of the criteria such as trip time, cost per trip and safety. A further extension with respect to stochastic traffic assignment (entropy maximization) is proposed in Lam, Huang (1992).

Ferrari (1995) discusses the preceding Wardrop equilibrium together with the variational inequality problem in view of *mode choice* instead of user classes. Here the inverse demand functions for trip on mode j stem from a logit model. An additional feature follows from capacitated roads. A toll pattern is introduced when the above solution does not satisfy capacity constraints. A multimode multiclass-user

[35] An alternative proof of existence where user classes, say trucks and passenger cars, perceive asymmetrical route costs is given in Braess, Koch (1979). This proof is based upon Kakutani's fixed point theorem and requires also continuous and monotone route costs for every user class. Moreover, it is shown that convex programming methods are not applicable.

model with demand functions derived from a random utility approach is proposed in Cantarella (1997), where the author abstains from inverting the traffic demand functions. In its most general form the model is analyzed as a fixed point problem with respect to link flows, that is $\hat{\varphi} = \varphi(\mathbf{c}(\hat{\varphi}))$.

4.3.4 Dynamic Traffic Assignment

Up until now one of the most important determinants of travel cost has been travel time, which can be reduced within limits by speeding up or which is increased due to congestion. Nevertheless, dynamic aspects such as the favorable choice of departure time in order to bypass congestion have not been taken into account. Another objection refers to the definition of link capacities as maximum flow per period, i.e., usually the busy hour, probably with respect to a certain level of service. The whole traffic load is then assigned at once to the capacitated network and it seems as if all travelers use every link included in their chosen route all over the period concerned. That may be true for circuit switched traffic on telecommunications networks but it is hard to justify for road traffic. One way out might be to interpret a traffic assignment, say $\mathbf{h}$, as an average network state, where all travelers leaving a link (or the network at their destination) are substituted immediately by following trip-makers. But in this case the path flow $\mathbf{h}$ would refer to a point of time while the trip table $\mathbf{x}^D = \mathbf{\Lambda}\,\mathbf{h}$ is related to the rush hour.[36] All these objections indicate that there is a strong need for a consistent reformulation with respect to simultaneous movements through space *and* time. On space-time networks for the the dynamic traffic assignment problem see, e.g., Powell et al. (1995, p. 195 ff.) and Ran, Boyce (1996).

Movements of flow units in the presence of queuing can be described best by picking out one particular link $e = (v_a, v_b)$ of the network; see Figure 4.8. The travel time across the road is assumed to be constant $\tau_e = 2$ so that all drivers are forced to choose the same constant speed which results from the given distance of the road. The specified capacity κ_e indicates the maximum number of cars that can pass the road with respect to the period τ_e at the corresponding level of service, say speed. The new element is that a flow unit can enter v_b only if it is free. Otherwise the traveler has to wait in a queue, which is represented by a new node v_e. He can be served after all other cars in front of him have left that queue and v_b becomes free. This queue discipline is termed FIFO (first in first out)[37] or FCFS (first come first served).

The disadvantage with the representation of queuing in Figure 4.8 is that we sometimes need much more detailed information on the functioning of queues.

[36] A consistent interpretation would be to take a photo of the network and to count all drivers on certain routes for that point of time so that $\mathbf{x}^D = \mathbf{\Lambda}\,\mathbf{h}$ denotes the instantaneous demand for traffic. Correspondingly, the capacity of a road is the maximum number of cars that can be served by the road at the same time and at a given level of service. Observe, however, that path flows are scattered through space but link flows occupy single roads. Hence, the relationship $\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h}$ must be broken up.

[37] Carey (1992) points out that FIFO leads to several problems in a dynamic traffic assignment, as it yields a nonconvex feasible set.

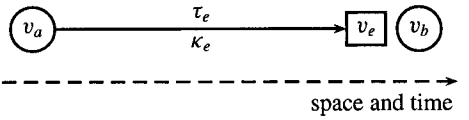


Fig. 4.8
Introducing queuing

Regarding the queue v_e it is in most cases useful to distinguish between a so-called "waiting room" and a "server". The maximum length of the queue κ_e^w refers to the maximum number of cars that can stay together in the waiting room, e.g., the place in front of a traffic light. In each period at the most κ_e^s travelers can be served, these customers leave the queue and enter v_b, where the service time τ_s is set to 0 for convenience. All remaining and incoming cars have to wait one more period $\tau_w = 1$ until the server becomes free.

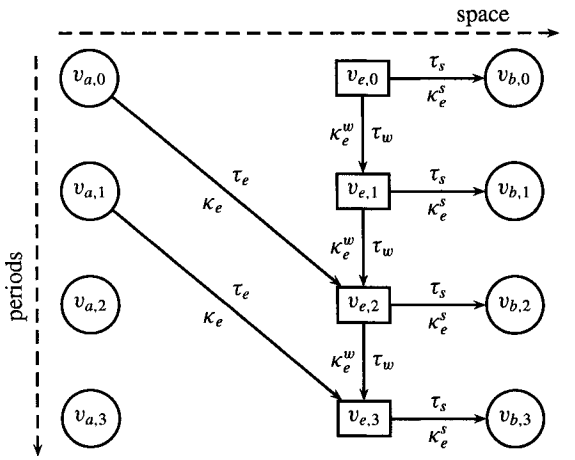


Figure 4.9
Dynamic traffic flows with queuing

The described steps are illustrated in Figure 4.9 with respect to three periods. Each node of the network and especially v_a, v_b are expanded to four nodes, $v_{i,t}$, $i \in \mathcal{V}$ and $t = 0, 1, 2, 3$, representing additionally the point of time. The initial movement $v_a \rightarrow v_b$ becomes now a motion in space and time, for example

$$v_{a,0} \xrightarrow{\tau_e=2} v_{e,2} \xrightarrow{\tau_w=1} v_{e,3} \xrightarrow{\tau_s=0} v_{b,3}$$

if the driver has to wait for one period in the *queue of link e*. Such a formulation can be found in Drissi-Kaitouni (1993), who deals with a dynamic traffic assignment problem on the basis of a capacitated network with queuing, where capacity constraints are handled implicitly through route cost functions.

At first glance Figure 4.9 suggests that the representation of dynamic traffic flows would merely be a problem of enumeration. It seems as if all we have to do is to encode the topology of the *expanded network*[38] by an augmented arc/path incidence

[38] The expanded network is no longer strongly connected as supposed by the prior analysis. For example, the node $v_{b,0}$ in Figure 4.9 cannot be reached by a driver in $v_{a,0}$.

matrix $\mathbf{\Delta}$ and a new OD/path incidence matrix $\mathbf{\Lambda}$. Notice, however, that several further problems must be solved.

Queue discipline. Other types of networks may need a different formulation, for example, when all incoming link flows are collected by only one queue. In that case the queue discipline FIFO may be unsuitable and prioritization of some parts of traffic is often used to improve efficiency; cf., for example, bus lanes, roundabout traffic, or just different rules on the right of way.

Queue dynamics. The length of a queue grows as long as the number of arriving cars exceeds the number of vehicles served.[39] With that a favorable route choice and consequently route pricing changes in time, which is in the center of interest in peak load pricing. Observe furthermore that a stochastic arrival process may be superimposed by traffic guidance systems.

Blocking. When a queue is working at its capacity, that is the waiting room of a queue is full, incoming cars must be rejected. As a consequence no vehicle can leave the predecessor node and a backlock takes place in that node although it has free capacities. Furthermore, a node can initiate traffic only if its queue has idle capacities, otherwise that traffic is blocked, it cannot enter the network.

Flow conservation. The new nodes representing queues must also satisfy the conservation of flow. No flow unit is lost accidentally within the network. Regarding path flows that condition implies $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^D$, where the trip table $\mathbf{x}^D$ refers, for instance, to a busy hour and the route flow $\mathbf{h}$ denotes all flows through space and time that are finished within the busy hour. In fact each path flow is counted during a given period at its point of destination. Consequently, $\boldsymbol{\varphi} = \mathbf{\Delta}\,\mathbf{h}$ determines the link flows for the same period,[40] but link capacities $\boldsymbol{\kappa}$ may be defined for shorter intervals. Alternatively, link flows may be computed by summing up all route flows (starting somewhere at any point of time) which use the link at hand at a certain point of time.

Cost of travel. Each link cost function must take into account a proper measure of trip time including the pure time τ_e needed to overcome space plus a sojourn time which is composed of the waiting time (multiples of τ_w) and the service time τ_s. Without choice of speed each link cost function c_e depends particularly on the respective traffic load φ_e with a corresponding length of the queue as well as on all other link flows by means of the queuing discipline and backlacks. Computing route cost by $\tilde{\mathbf{c}}(\mathbf{h}) = \mathbf{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\varphi})$ would ignore the effects of changing network states by which drivers on the same route may be confronted with different delays.

Further aspects will be discussed in Chapter 5 with special regard to telecommunications and data networks; see p. 275. We shall discuss in particular the minimization of the mean delay by virtue of a capacity assignment procedure that determines

[39] On the computation of the queue length experienced by commuters entering a respective link within the variational inequality framework see, e.g., Lam, Huang (2002).

[40] It is assumed that the network state at any point of time does not change within the period of observation. In particular, the number of initiated trips equals the number of finished trips.

the dimensioning of server capacities κ^s. This network design problem refers to the "mean state" of a network flow pattern.

Apart from all these caveats the true difficulty may be the enumeration of feasible paths. Just imagine an OD-pair (v_i, v_j) and let $\mathcal{R}_t(i, j)$ be the set of available interconnecting routes through space and time with respect to a period $t = 0, 1, 2, \ldots$ On the one hand, most nodes are irrelevant for this set as they exist only for a certain time interval. On the other hand, the set of feasible routes shrinks with the traffic load, because queuing delays prevent travelers from choosing certain options. Thus the remaining routes need more time and drivers have to depart earlier if they want to arrive at v_j at some appointed time.

All of the above objections can be tackled by more or less restrictive regularity assumptions, where some details of this work can be found in Drissi-Kaitouni (1993). Indeed, he shows that the new problem of finding a user optimal trip assignment $\hat{\mathbf{h}}$ can be reduced to a variational inequality problem of the familiar type

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}})^{\mathsf{T}}(\mathbf{h} - \hat{\mathbf{h}}) \geqq 0 \qquad \forall \mathbf{h} \in H^D.$$

On further details of dynamic traffic assignment see Ran, Boyce (1996), particularly Section 5.3.2.

4.3.5 Spatial Price Equilibrium

The knowledge of a spatial price equilibrium gives further insights into the origins of demand for traffic or trade as it was primarily considered in Samuelson (1952). A standard reference for further readings on spatial price equilibria is Takayama, Judge (1971). The subsequently presented approach deals with a single homogeneous commodity that is produced at different places and must be shipped to consumer destinations. The corresponding multicommodity spatial price equilibrium needs some more notational work and can be found in Fisk (1987). The importance of deriving equilibria of that type is not only to compute the implied traffic flows. It rather has far reaching consequences on the location of economic activities and the specialization pattern of regions. For example, firms entering a market have to decide where to locate their production plants and households have to choose some place of residence. Of course, this is an interdependent problem, because firms have to distribute their commodities to customers and households presumably settle down in the environment of their working place.

A provisional approach of a spatial price equilibrium has been introduced in (3.12) on page 91. Friesz et al. (1984) and later Miller et al. (1996) demonstrate that this problem has a more general variational inequality formulation. It considers

1. flow dependent link cost functions $\mathbf{c}(\boldsymbol{\varphi})$, where φ_e is the total amount of the good passing link e and c_e denotes the respective transport cost per unit of the commodity;
2. price sensitive demand functions $\mathbf{y}^D(\mathbf{p}^D)$, where $y_v^D(\mathbf{p}^D)$ is the demand at consumer destination v and p_v^D is the commodity price consumers are willing to pay at that node;

3. price sensitive supply functions $\mathbf{y}^S(\mathbf{p}^S)$, where $y_v^S(\mathbf{p}^S)$ is the supply at production place v and p_v^S is the price demanded by the supplier at that node; and
4. the condition of flow conservation $\mathbf{B}\,\boldsymbol{\varphi} = \mathbf{y}^S - \mathbf{y}^D$ with the node/arc incidence matrix $\mathbf{B}$ of the directed network at hand; it especially allows for the case that the commodity is produced and consumed locally.

Notice that every node can be a production place and a consumer place at the same time having the advantage that transport cost is zero within the node. Furthermore, flow conservation implies market clearing of the form $\sum_{v\in\mathcal{V}}(y_v^S - y_v^D) = 0$ for any equilibrium transport pattern $\hat{\boldsymbol{\varphi}}$, which reallocates nodal supply $\mathbf{y}^S$ in accordance with nodal demand $\mathbf{y}^D$. Consequently, a *spatial price equilibrium* requires that no household and no firm can improve its payoff at the given spatial (or nodal) price system, say $\boldsymbol{\lambda}$, by moving some amount of the good to a different place.

A solution to the following abstract variational inequality will be shown to satisfy all conditions required by a spatial price equilibrium. Remark 4.29 below hands in some plausibility checks for the case of integrability. On the assumption that the demand functions as well as the supply functions are invertible[41], say $\mathbf{p}^D = \mathbf{p}^D(\mathbf{y}^D)$ and $\mathbf{p}^S = \mathbf{p}^S(\mathbf{y}^S)$, the collection $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{y}}^D, \hat{\mathbf{y}}^S)$ constitutes a spatial price equilibrium if and only if

$$\mathbf{c}(\hat{\boldsymbol{\varphi}})^\top(\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}}) - \mathbf{p}^D(\hat{\mathbf{y}}^D)^\top(\mathbf{y}^D-\hat{\mathbf{y}}^D) + \mathbf{p}^S(\hat{\mathbf{y}}^S)^\top(\mathbf{y}^S-\hat{\mathbf{y}}^S) \geqq 0 \qquad \forall\,(\boldsymbol{\varphi}, \mathbf{y}^D, \mathbf{y}^S) \in \Omega,$$

where the feasible set Ω is defined by

$$\Omega = \left\{(\boldsymbol{\varphi}, \mathbf{y}^D, \mathbf{y}^S) \middle|\, \mathbf{B}\,\boldsymbol{\varphi} = \mathbf{y}^S - \mathbf{y}^D,\ \boldsymbol{\varphi} \geqq \mathbf{0},\ \mathbf{y}^D \geqq \mathbf{0},\ \mathbf{y}^S \geqq \mathbf{0}\right\}.$$

4.29 Remark (Integrability) Chao, Friesz (1984) discuss the problem of finding a spatial price equilibrium on the basis of corresponding line integrals with separable supply, demand, and transport cost functions.[42] Assuming integrability our variational inequality problem is equivalent to

$$\max\left\{\oint_{\mathbf{0}}^{\mathbf{y}^D} \mathbf{p}^D(\mathbf{z})\mathrm{d}\mathbf{z} - \oint_{\mathbf{0}}^{\mathbf{y}^S} \mathbf{p}^S(\mathbf{z})\mathrm{d}\mathbf{z} - \oint_{\mathbf{0}}^{\boldsymbol{\varphi}} \mathbf{c}(\mathbf{z})\mathrm{d}\mathbf{z} \middle|\, (\boldsymbol{\varphi}, \mathbf{y}^D, \mathbf{y}^S) \in \Omega\right\},$$

where the first line integral corresponds to consumers' willingness to pay and the second term describes production cost in the case of marginal cost pricing. The third line integral differs from total transport cost and describes the user optimal traffic assignment on the basis of perceived link cost $\mathbf{c}$ instead of marginal link cost $\mathbf{c}$. Substituting $\mathbf{c}$ for $\mathbf{c}$ would result in a welfare maximizing approach, where the sum of consumers' surplus and producers' surplus is reduced by total transport cost. □

Without going into further details with respect to the existence and the uniqueness of a solution to the preceding variational inequality problem, we have still to

[41] In the case of non-invertible supply and demand functions alternative formulations are needed as has been discussed in Section 4.2.5.

[42] In the light of variational inequalities almost all results of their sensitivity analysis remain valid when dropping the integrability assumption; see Tobin (1987).

prove that a solution $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{y}}^D, \hat{\mathbf{y}}^S)$ represents a spatial price equilibrium. In doing so let $\boldsymbol{\lambda}$ denote the Lagrange multiplier of the equality constraint with node specific components λ_v. Then, by Theorem 3.17, an optimal solution $(\hat{\boldsymbol{\varphi}}, \hat{\mathbf{y}}^D, \hat{\mathbf{y}}^S)$ is characterized by the KKT conditions

$$\hat{\boldsymbol{\mu}} := \begin{pmatrix} \mathbf{c}(\hat{\boldsymbol{\varphi}}) \\ -\mathbf{p}^D(\hat{\mathbf{y}}^D) \\ \mathbf{p}^S(\hat{\mathbf{y}}^S) \end{pmatrix} - \begin{pmatrix} \mathbf{B}^\top \\ -\mathbf{I} \\ \mathbf{I} \end{pmatrix} \hat{\boldsymbol{\lambda}} \quad \text{so that} \quad \hat{\boldsymbol{\mu}} \geqq \mathbf{0}, \quad \begin{pmatrix} \hat{\boldsymbol{\varphi}} \\ \hat{\mathbf{y}}^D \\ \hat{\mathbf{y}}^S \end{pmatrix} \geqq \mathbf{0} \quad \hat{\boldsymbol{\mu}}^\top \begin{pmatrix} \hat{\boldsymbol{\varphi}} \\ \hat{\mathbf{y}}^D \\ \hat{\mathbf{y}}^S \end{pmatrix} = 0$$

$$\text{and} \quad \mathbf{B}\,\hat{\boldsymbol{\varphi}} - \hat{\mathbf{y}}^S + \hat{\mathbf{y}}^D = \mathbf{0}.$$

Regarding demand and supply at some node v the complementarity conditions say

$$\hat{y}_v^D > 0 \implies p_v^D(\hat{\mathbf{y}}^D) = \hat{\lambda}_v \quad \text{and} \quad p_v^D(\hat{\mathbf{y}}^D) < \hat{\lambda}_v \implies \hat{y}_v^D = 0,$$
$$\hat{y}_v^S > 0 \implies p_v^S(\hat{\mathbf{y}}^S) = \hat{\lambda}_v \quad \text{and} \quad p_v^S(\hat{\mathbf{y}}^S) > \hat{\lambda}_v \implies \hat{y}_v^S = 0.$$

Four major cases must be distinguished with respect to the spatial price system $\hat{\boldsymbol{\lambda}}$ in networks:

(a) $\hat{\lambda}_v = p_v^S(\hat{\mathbf{y}}^S)$ for pure sources with positive demand $\hat{y}_v^D > 0$ and no supply $\hat{y}_v^S = 0$;
(b) $\hat{\lambda}_v = p_v^D(\hat{\mathbf{y}}^D)$ for pure sinks with positive supply $\hat{y}_v^S > 0$ and no demand $\hat{y}_v^D = 0$;
(c) $\hat{\lambda}_v = p_v^S(\hat{\mathbf{y}}^S) = p_v^D(\hat{\mathbf{y}}^D)$ for consumer places with $\hat{y}_v^D > 0$ being also places of production, i.e., $\hat{y}_v^S > 0$;
(d) $p_v^D(\hat{\mathbf{y}}^D) \leqq \hat{\lambda}_v \leqq p_v^S(\hat{\mathbf{y}}^S)$ for transshipment points without supply and demand.

The last case seems to be ambiguous. However, the price $\hat{\lambda}_v$ can be calculated recursively from observing optimal arc flows.

$$\hat{\varphi}_e > 0 \implies c_e(\hat{\boldsymbol{\varphi}}) = \hat{\lambda}_{\omega(e)} - \hat{\lambda}_{\alpha(e)} \quad \text{and} \quad c_e(\hat{\boldsymbol{\varphi}}) > \hat{\lambda}_{\omega(e)} - \hat{\lambda}_{\alpha(e)} \implies \hat{\varphi}_e = 0$$

Every node imports the commodity only from one of the "relatively" cheapest predecessors, where the term "relatively" emphasizes that transport cost must be taken into consideration. All other nodes with $\hat{\lambda}_{\alpha(e)} + c_e(\hat{\boldsymbol{\varphi}}) > \hat{\lambda}_{\omega(e)}$ are irrelevant for the importing node $\omega(e)$. Conversely, all exporting nodes supply the commodity only to those nodes with the highest increase in price. All other nodes with $\hat{\lambda}_{\alpha(e)} > \hat{\lambda}_{\omega(e)} - c_e(\hat{\boldsymbol{\varphi}})$ are irrelevant for the emitting node $\alpha(e)$. As a consequence, following any utilized path $\rho \in \mathcal{R}(s, t)$ of a commodity shipment, its price starts at $p_s^S(\mathbf{y}^S)$ and is increased successively by $c_e(\hat{\boldsymbol{\varphi}})$ until it reaches the destination v_t with

$$p_s^S(\hat{\mathbf{y}}^S) + \sum_{e \in \rho} c_e(\hat{\boldsymbol{\varphi}}) = p_t^D(\hat{\mathbf{y}}^D).$$

No alternative path with higher transport cost will be used and, conversely, no cheaper route does exist. Given a positive path flow on route ρ, all of the quantities involved in the preceding price equation are positive, that is $\hat{y}_s^S$, $\hat{y}_t^D$, and $\hat{\varphi}_e$ for all $e \in \rho$ with $\rho \in \mathcal{R}(s, t)$. Hence, the equation can be rewritten completely in terms of

the spatial price system $\hat{\boldsymbol{\lambda}}$. All nodes $v \in \mathcal{V}$ receive the value $\hat{\lambda}_v$ and the difference between any two adjacent node is $c_e(\hat{\boldsymbol{\varphi}})$.

Eventually, the KKT conditions yield a zero profit condition.

$$0 = \hat{\boldsymbol{\mu}}^{\mathsf{T}} \begin{pmatrix} \hat{\boldsymbol{\varphi}} \\ \hat{\mathbf{y}}^D \\ \hat{\mathbf{y}}^S \end{pmatrix} = \mathbf{c}^{\mathsf{T}}\hat{\boldsymbol{\varphi}} - \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{B}\hat{\boldsymbol{\varphi}} - \mathbf{p}^{D\mathsf{T}}\hat{\mathbf{y}}^D + \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{I}\hat{\mathbf{y}}^D + \mathbf{p}^{S\mathsf{T}}\hat{\mathbf{y}}^S - \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{I}\hat{\mathbf{y}}^S$$
$$= \mathbf{c}^{\mathsf{T}}\hat{\boldsymbol{\varphi}} - \mathbf{p}^{D\mathsf{T}}\hat{\mathbf{y}}^D + \mathbf{p}^{S\mathsf{T}}\hat{\mathbf{y}}^S$$

Taking the complementary slackness conditions into account the same statement can be expressed solely in terms of the spatial price system $\hat{\boldsymbol{\lambda}}$.

$$-\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{B}\,\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\hat{\mathbf{y}}^D - \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\hat{\mathbf{y}}^S = \mathbf{c}^{\mathsf{T}}\hat{\boldsymbol{\varphi}}.$$

Dafermos, Nagurney (1989) discuss a similar model with constant link cost and with supply and demand functions which are linear in node specific prices. Each supply node is directly connected to every demand node so that transshipment nodes can be omitted. The complementarity slackness conditions on link flows then become

$$\hat{\varphi}_e > 0 \implies c_e = p^D_{\omega(e)}(\hat{\mathbf{y}}^D) - p^S_{\alpha(e)}(\hat{\mathbf{y}}^S)$$
$$\hat{\varphi}_e = 0 \implies c_e \geqq p^D_{\omega(e)}(\hat{\mathbf{y}}^D) - p^S_{\alpha(e)}(\hat{\mathbf{y}}^S)$$

where the variables $\hat{\mathbf{y}}^D$ or $\hat{\mathbf{y}}^S$ may be substituted by $\mathbf{B}\,\hat{\boldsymbol{\varphi}} = \hat{\mathbf{y}}^S - \hat{\mathbf{y}}^D$. The suggested solution method for the consumers' point of view is to eliminate $\mathbf{y}^S$. The algorithm then computes an increasing sequence of consumer prices (with decreasing demand and increasing supply) until an equilibrium occurs. The inverse steps may be applied to the supply side of the market.

4.30 Summary Section 3.2 deals with the traffic assignment of a fixed trip table so that any form of road pricing merely results in traffic diversion. This effect is now superimposed by a price sensitive demand so that additional effects of traffic diminution or traffic creation must be taken into account. With these features several questions arise as to how to measure welfare. This is a complex problem particularly when path independence and invertibility of demand functions are not satisfied, in which case we make use of variational inequality formulations. Nevertheless, it has been shown for the regular case that marginal cost pricing maximizes the net benefit of traffic. This solution differs, however, from a user optimum, where travelers are oriented towards average route costs rather than marginal costs. In doing so they ignore the external costs imposed on the rest of the network, which might be corrected again by an appropriate road pricing scheme. Several alternative formulations indicate how to solve the same problem in alternative settings. They have been summarized on p. 215. The extensions indicate how to apply variational inequality problems to further aspects that have been neglected so far. □

4.4 Summary

The traffic assignment problem has been introduced in Chapter 3 as the core problem of transport networks. It is based on several simplifying assumptions that have been

relaxed in this chapter. Section 4.1 widens the analysis of the former deterministic traffic assignment by a stochastic route choice. This leads to the notion of a stochastic user equilibrium with path demand functions that describe the most likely traffic flow pattern. Depending on what data can be observed – that is link flows, path flows, trip tables, or nodal trip generation and trip attraction – different approaches where presented in order to find the most likely traffic assignment.

The following Section 4.2 abandons the assumption of a fixed trip table by taking a price sensitive demand for traffic into account. In this case we can expect that any change of the prevailing traffic conditions induces not only a reassignment of traffic flows (traffic diversion), but it also affects traffic volumes. An improvement of traffic conditions usually increases traffic flows. This traffic creation is to be contrasted with traffic diminution, for example, when the costs of travel are increased by raising tolls. On the assumption that the traffic demand functions are invertible, we are able to describe user optima which are characterized by Wardrop's first principle. Nevertheless, welfare measurement becomes difficult because travelers take "asymmetric" average costs rather than "symmetric" marginal costs as the basis for their traffic demand. Afterwards we have presented several ways to analyze the combined travel choice/trip assignment problem. All these approaches are equivalent under certain regularity assumptions, but they refer to distinct sets of variables. We can now choose that formulation which fits best to the information on traffic processes that can be observed in reality.

The chapter ends with further issues that extend the traffic assignment problem discussed so far. Bilevel programming is one approach to simplify the complex travel choice/trip assignment problem by separating it into two interdependent but smaller problems. The multiclass-user traffic assignment extends the analysis with respect to distinct classes of travelers such as passenger cars and trucks. Dynamic traffic assignment distinguishes explicitly between the two dimensions that characterize movements through space and time. Finally, a spatial price equilibrium describes the impact of transport costs on local consumer prices.

5

Design of Transport Networks

5.1 Introductory Aspects of Network Optimization

5.1.1 Construction of Networks

The results obtained so far refer to a network with given features. The two sets of nodes and edges as well as their combinations such as the sets of feasible routes are fixed and known.[1] Furthermore, capacity restrictions, congestion effects, and transport costs qualify the level of services produced by the network. On the basis of these properties multicommodity flow problems including routing problems have been discussed. Moreover, some rudimentary aspects of location theory were mentioned. The analysis is now to be continued with respect to modifications of the network and of its properties.

The network design problem is to choose facilities which should be added to a transportation network (network widening) or to determine capacity enhancements of existing network facilities (network deepening) which are in some respect optimal.[2] In both cases we are interested in cost reducing or quality improving investments. Following this characterization network design suggests to classify two types of problems: *discrete* and *continuous* network design problems; cf. Friesz (1985), for example.

Discrete network design problems are concerned with the topology of the network. The basic ingredients are a set of potential nodes and a set of potential edges. Typical tasks are explained best by giving some examples.

- Establish a new transport network from the scratch (greenfield approach). For instance, think of a newcomer in the telecommunications industry. Another case is

[1] Cellular mobile telephone networks constitute an important exception where the topology of a network can vary at least to some extent with the movements of customers considered as end nodes. Such effects make network design considerably more difficult.

[2] In most cases we lay emphasis on the network operator's point of view. A more general approach refers to *cost-benefit analysis* which is a widely used technique of applied welfare economics. Cost-benefit analysis throws light on the *social* desirability of undertaking an economic project, cf. Chakravarty (1987).

concerned with the reconstruction of networks after disasters like earthquakes or floods (scorched earth approaches). A less drastic example is that of snow removal to put a network back into service as quick as possible.

- Extend a given network by the construction of new nodes or links. Additional nodes may refer to new warehouses, industrial parks, airports, ports, etc. The connection of nodes to a network requires the construction of new roads, rail links, tunnels, bypasses, shipping canals etc. Recall that parts of a network are related to each other as complements, other elements are substitutes. In this sense, certain links are necessary to reach a particular node; other interconnections just improve its accessibility.
- Growing networks naturally reach a state where improvements not only incur costs of construction but also demolition costs, which have deep impact, for example, on urban growth; see Hochman, Pines (1980). The problem is here to weigh up strategies of pure network extensions and strategies of reconstruction.
- Of course, there are mixed cases which need elements of discrete and continuous network design. An example is the provision of new transport services on an urban subway network which requires new tunnels (and stations) as well as an upgrade of old tracks.
- An advanced problem addresses the question of optimal location of points of interconnection for traffic exchange between autonomous networks. Awduche et al. (1998) analyze such a problem in the context of Internet Service Providers with the objective to minimize the cost of *peering*, which improves the efficiency of Internet-domain traffic. The same problem is valid for national and international telecommunications networks as well as for road traffic with respect to different travel modes or for commercial road systems.

One important approach in constructing networks is the computation of *minimum spanning trees* (MST). Although not in the center of interest here it is useful to have an idea of such trees, because networks with minimal construction costs and huge link capacities tend to have a "tree-like" topology. Such networks have some important disadvantages. Blockage can occur with no alternative routes available and tree networks are very vulnerable to network failures such as temporal maintenance works or a durable destruction of network elements. The failure of an arc of a tree just results in decomposing the tree into two disconnected trees; see, e.g., Dantzig, Thapa (1997, Lemma 9.17). Traffic that is generated in one tree cannot be terminated in the other tree any longer. From an economic point of view it is therefore worth to install *redundant facilities* to avoid traffic breakdowns due to a malfunction of some network elements. The *reliability* of a graph indicates the probability that the "surviving" network without its failing elements remains connected.[3] Interrupted routes can then be substituted by alternative working routes. For instance, if one link in the left hand part of Figure 5.1 below fails, one node is isolated from the rest of

[3] Ferrari (1988, 1991) defines the reliability of a motorway transport system as the probability that, starting from the instant when reliability is measured within a certain period of time, no drops in speed to an extent deemed hazardous will occur. Hence, reliability can be increased if the conditions of instability are known and if there are control instruments to intervene before traffic break downs take place.

the network. The situation is even worse when the upper node is out of service. The (surviving) network is immune against these cases if the lower nodes are connected by some link so that the chain goes over into a circuit. Another side effect is that the comfort of network usage increases with the set of alternative routes induced by traverse links or bypasses.

What is a minimum spanning tree? Without going into the details of graph theory, let $\mathcal{G}$ be a graph on n vertices. This graph $\mathcal{G}$ is said to be a *tree* if it satisfies three conditions: (1) it has $n-1$ edges, (2) it is connected, and (3) it contains no circuits. Jungnickel (1994) shows that any two of these conditions imply the third. Every graph $\mathcal{G}$ having n vertices but more than $n-1$ edges generates some subgraph by removing all but $n-1$ edges. If such a subgraph is a tree, it is referred to as *spanning tree*. The term "spanning" becomes apparent taking note of the fact that trees are connected graphs where each pair of nodes is joined by a consecutive chain of edges. At last a *MST*[4] is a spanning tree, but has weights or lengths associated with the edges, and the sum of the weights of its edges is at a minimum. When the edge weights indicate costs one speaks also of a minimum cost spanning tree. Notice that it is not always optimal for a network carrier to connect all nodes scattered through space to the network; nevertheless some of them, e.g., postal service providers, are obliged by regulators to provide universal services. The term *coverage* addresses the question of how many nodes, say points of demand, have access to the network and how many are isolated from it.

Two real-world examples may illustrate the importance of MSTs for network design.

- A group of individuals, who are separated by varying distances, wish to be connected together in a telephone network. Although an MST cannot do anything about the distance from one connection to another, it can be used to determine the least costly paths with no cycles in this network, thereby connecting everyone at minimum cost.
- Another useful application of MST would be finding airline routes. The vertices of the graph represent cities, and the edges indicate routes between the cities. Obviously, the further one has to travel, the more it will cost, so MST can be applied to optimize airline routes by finding the least costly paths with no cycles.

An easy example of a tree is a graph with three nodes and two edges. Given the link weights c_1, c_2 and c_3 the tree shown in the left hand part of Figure 5.1 indicates an MST. The second graph represents a so-called *Steiner tree*: given the

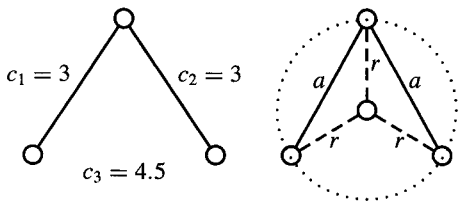


Fig. 5.1
Special trees

[4] *Kruskal's* and *Prim's algorithm* differ in their methodology, but both eventually end up with the MST. Kruskal's algorithm uses edges, and Prim's algorithm uses vertex connections in determining the MST.

set of vertices of $\mathcal{G}$ a solution to the MST may be improved[5] if we are allowed to add (arbitrarily many) vertices which are referred to as Steiner points. Figure 5.1 shows the example of the vertices of an equilateral triangle with side length a and an MST of length $2a$. Adding the center of the dotted circumcircle (with radius $r = \sqrt{3}\,a/3$) as Steiner point, the length of the MST reduces to $3r = \sqrt{3}\,a$, i.e., the length of the dashed Steiner tree. Recall again that the Steiner point is decisive for the reliability of the network. In particular, the center nodes of star networks (or concentrators in hierarchical networks) must be protected against demolition. To limit the disadvantages from a failing concentrator one can require that no more than m nodes should be disconnected from the network.

Many features of networks depend on their topological structure. The following list includes the most important network topologies, where possible advantages or disadvantages are at hand without going into further details.

- A *star network* consists of a center node to which all other peripheral nodes are connected by one physical link.
- In a *tree network* each pair of nodes is connected by a unique route.
- A *mesh network* consists of nodes which are connected with at least two other nodes. In a fully connected mesh network there is a direct branch between any two nodes.
- A *bus network* has a common transmission line and all stations are connected to the single bus.
- In *ring networks* every node has a unique predecessor (or upstream node) as well as a unique successor (or downstream node). The transmission follows a predetermined direction.

Of course, all of the above topologies can be combined in more or less hierarchical structures. For example, one might think of a failsafe backbone network with the expensive topology of a ring or fully connected mesh network. Each node could then serve as a concentrator for a subordinated star or tree network, which may be less reliable but cheaper.

Continuous network design problems have the task of improving the functionality of networks or of network elements so that the level of service (comfort, risk, travel time, etc.) delivered to the customers increases. These problems take the network topology as given. They are concerned with the parameterization of a network (link or node capacity, user charges, etc.). An important problem in the management of urban transportation systems is determining improvement levels to the roads and freeways of the network. Improving streets reduces travel time and travel cost and thus lessens congestion in the city. The determination of improvement levels is known as the (continuous) network design model. A useful classification can be found in Magnanti, Wong (1984):

- *Fixed charge network design problems* deal with the aspect of fixed arc design

[5] In accordance with Bern, Graham (1996) the Steiner problem of finding the shortest network cannot be solved exactly within reasonable time. Nevertheless, suboptimal solutions are successfully employed, for instance, in the design of electronic circuits.

cost. Ignoring capacity restrictions and other side constraints this case may be subsumed to discrete design models. However, we shall follow an extremely simplified approach with perfectly divisible capacity increments so that the existing capacities can be varied continuously. Another extreme position is pursued in Wessäly (2000). He specifies all sequences of potential capacities and corresponding sequences of investments for every link. This procedure requires thorough information on the network to be designed and is reduced in almost all cases to a few alternative investment programs to be compared.

- *Budget design problems* introduce an upper bound on investment costs. A fixed budget represents an exogenous inflow of resources, the origin of which is not explained within the model. This approach may either be seen as a preliminary stage of determining the optimum budget size or as a matter of fact. Take, for instance, a fixed public amount of expenditures on material infrastructure.
- *Network design/traffic equilibrium problems* are augmented by equilibrium conditions as, for example, Wardrop's first principle of route choice. The problem is to find an optimal network design in terms of additional facilities or capacity enhancements, when the network flow pattern is restricted to be an equilibrium. Solution methods require theory indicating how solutions to the equilibrium conditions respond to changes in design decisions. Two aspects of travelers' behavior must be distinguished: traffic diversion and traffic generation. On the one hand, the improvement of roads may cause the network users to choose different routes (reassignment) or to select new destinations (redistribution). On the other hand, an upgraded network may induce a rising number of realized trips, that is, some travelers increase their frequencies of trips and other potential network users opt in. At least in the long term location decisions must also be taken into account. Improving the accessibility of an industrial park or the expansion of an airport have in general far reaching consequences. If locational decisions ignore the total cost of providing and servicing roads, wrong incentives harden or even intensify the inefficient design and use of networks, cf., e.g., Walters (1968, p. 95 ff.).

Of course, most network design problems include aspects of both, discrete and continuous network improvements. As an example, Section 5.2.3.1 introduces the concept of so-called hub-and-spoke networks. The idea is to upgrade distinct points in the set of nodes which will be called hubs. All other point are connected to one of these hubs and the remaining links are broken up. As a consequence, all hubs play the role of a concentrator in the respective star network, which looks like a wheel with *spokes* connecting the minor nodes to the respective *hub*. Finally, all hubs are interconnected possibly by a mesh network. This is the discrete network design problem. The last step, however, is decisive to take advantage of the hub-and-spoke topology: as most of the traffic is concentrated on interhub links, the carrier has an incentive to increase capacities on these links to realize economies of scale. The later reasoning is a matter of continuous network design.

One fundamental property of network design is that solitary network "improvements" must be evaluated on the basis of their response to the rest of the network. Network designers, however, must account for several problems which are not obvious at first sight.

- Traffic usually causes pollution external to the network. For road specific emissions an improvement in the cost structure may result in an increase in total emission even for a constant traffic demand. For example, noise increases with speed. Another reason is that traffic may be rerouted along roads with higher emissions. In fact, this problem is always apparent where decision-makers ignore external effects. A similar outcome might be observed for generated traffic. An increase in travel demand results in a decrease in emission if the new user optimal flow pattern implies a more intensive usage of roads with lower emission; see Nagurney (2000a). Of course, the converse case can also be constructed, where traffic diminution induces higher emissions.
- Another astonishing effect has been introduced as the so-called Braess' paradox; see Remark 3.51, p. 162. The addition of a link can result in an increase of travel cost for every user. On circumstances that avoid this counterintuitive phenomenon see Pas, Principio (1997).
- Projects intended to improve the functionality of networks may produce traffic congestion for a long time prior to their completion. The short-termed disadvantages must be evaluated against future benefits. Moreover, the transition towards a new equilibrium may consist of series of situations with traffic disequilibria. A control theoretic formulation to model such a disequilibrium network design problem is proposed in Friesz et al. (1998).

A final remark draws the attention to a fact that is frequently ignored by pure network design problems, namely, *maintenance work* that occurs at regular intervals. Although cost of maintenance amounts to a substantial part of network operating cost, there are several more or less questionable arguments in doing so. Maintenance cost are of minor importance as long as the network design problem merely seeks the technical realization of demand feasible network flows. At least for the short run – most of the subsequent models refer to one period – investment cost exceed cost of maintenance by far. A weaker argument is that we need far more difficult mathematical tools such as dynamic programming to implement repetitive operating cost. The weakest point is that construction projects generally are more (politically) appealing than maintenance projects. Consequently, most of the results below must be reviewed with regard to financing network infrastructure in the long run. On a detailed survey on methods of cost allocation in several European countries see Link et al. (1999). They distinguish between road functions, road types, and vehicle types and gather costs of construction, reconstruction including resurfacing, maintenance (grass cutting, sweeping, road marking, street lighting, etc.), and administration (police, traffic wardens).

5.1.2 Network Improvements

In what follows it is almost always assumed for the sake of simplicity that increments of any link capacity are perfectly divisible. This assumption helps to avoid integer programming and other technical difficulties. Furthermore, discrete network design will be treated in some way as a borderline case of continuous network design. The

addition of a new link is just interpreted as all other capacity increments with the exception that it starts at zero capacity.

The *continuous network design* problem of the network provider takes the network topology as given. The problem is concerned with the parameterization of the network, where the attention is drawn to link capacities $\boldsymbol{\kappa}$ which restrict the non-negative link flows to $\mathbf{0} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$. Consequently, the output of the network is bounded from above. Increasing this upper bound therefore needs appropriate capacity extensions $\boldsymbol{\delta} \geqq \mathbf{0}$ so that the flow restrictions become

$$\boldsymbol{\varphi} \leqq \boldsymbol{\kappa} + \boldsymbol{\delta}.$$

For perfectly divisible capacity increments, capacities can be adjusted continuously. Towards a more realistic approach many authors suppose $\boldsymbol{\delta}$ to be integer. In other models δ_e indicates some indivisible capital unit which is installed ($\psi_e = 1$) or not ($\psi_e = 0$). Such aspects will be introduced to some extent at a later stage. Another detail concerns the network properties before reconstruction. Observe that the capacity κ_e of some link e can be zero before the investment. In this case $\delta_e > 0$ has the meaning of constructing a new link.

The investment costs of the capacity extension are given by $\mathbf{q}^\top \boldsymbol{\delta}$, where the price vector $\mathbf{q}$ is assumed to be constant. The component q_e denotes the "cost per capacity unit" regarding link e.

5.1 Remark (Lack of consistency) The components of $\mathbf{q}$ are frequently termed *market prices*. Recall that capacity is measured in flow units per period and that capacity increments are assumed to be perfectly divisible.

$$[\varphi_e] = [\kappa_e] = [\delta_e] = \text{flow units on edge } e \text{ (per period)}$$

In the present context one unit of a capacity increment represents one additional flow unit (per period) and q_e is the corresponding price of the investment. Such an approach abstains at least from two facts: (1) road investments are in general utilized over a period which is indeed longer than one hour[6] and (2) road investments constitute an infrastructure capital stock that must be converted into flow capacities. Models, which ignore such objections, are far from being realistic but they simplify some provisional calculations greatly.

A relatively simple method to overcome the above problems is given in Keeler, Small (1977) who define *rental costs of road e* with capacity κ_e by

[6] Anderson (1972) provides static and dynamic models for determining least cost investments in the special case of electricity supply. He considers (integer) capacity restricitions as well as transmission losses. Furthermore, the supply side is characterized by various technologies using different energy sources such as gas, fossile fuels, nuclear power, or hydroelectric power. As electricity storage is a particular problem, hydroelectric and nuclear power stations which have limited flexibility in adjusting the magnitudes of power supply are used for the continuous base loads. By contrast, gas turbines with high generation costs accommodate especially peak loads. The author analyzes among others a vintage model that distinguishes between capital costs, maintenance and replacement costs, and operating costs, in order to find an optimum mode of system operation.

$$I_e(\kappa_e) = R_e(\kappa_e) + M_e(\kappa_e) + rA_e(\kappa_e) \quad \text{with} \quad R_e(\kappa_e) := \frac{rK_e(\kappa_e)}{1-(1+r)^{-T}},$$

where r is the interest rate, T is the lifetime of the road, R_e is the annuity of the construction cost K_e, M_e is the maintenance cost, and A_e is the land acquisition cost. Thus

$$q_e := I_e(\kappa_e)/\kappa_e$$

represents the rental price of one capacity unit per year. The annuity K_e describes the cost of purchasing durable assets by spreading over the economic life of the asset rather than treating such costs as immediate expenses when they occur. The rationale for doing so is that the benefits of owning such assets are spread over their economic life, e.g., savings in transport costs, and hence the costs should be distributed over the same period. In the case of physical plant and equipment, these costs are called *depreciation*. On a combined machine replacement and capacity expansion problem as an infinite horizon linear program see Jones et al. (1988). They solve that problem by decomposing it into a sequence of single period linear programs. □

Notice also that the functional form of investment costs may be too restrictive. Dantzig et al. (1979) use, for example, more general convex investment cost functions I_e, $e \in \mathcal{E}$, where $I_e(\delta_e)$ denotes the investment cost associated with the investment decision δ_e. A very general form of dealing with discrete structures of available capacities is used in Wessäly (2000). He specifies sequences of potential capacities and the corresponding sequences of costs for incremental capacities. In this case q_e depends on both the actual capacity κ_e and the increment δ_e. Although this procedure is useful in many practical applications as telecommunications networks, it loads a lot of detailed information onto the formulation of problems. In order to keep the analysis as simple as possible we release it from that burden. The easiest case refers to perfectly divisible capacity units which can be installed at constant cost per unit.

Several further important problems are ignored by the above approaches to record investment costs. For example, most road works incur fixed cost (e.g., laying cables or pipes into ditches) so that it makes sense to bunch several steps of capacity extensions. Moreover, in many practical applications the links show the property of increasing returns to scale. Take for example a thin wire with a given capacity. Certain multiples of this capacity can be obtained either by installing an appropriate number of additional thin wires or by just one thick wire. The second choice may be much cheaper than the first. Both cases, fixed costs as well as increasing returns to scale, contradict usually the assumption of convex investment cost functions at least locally. Eventually, it is worthwhile to mention that economies of scale have already far reaching implication at the level of constructing a network. Hub-and-spoke networks indicate one example which will be discussed in Section 5.2.3.1.

Another problem is that of defining the capacity, see Figure 3.8 on page 111. When modeling congestion effects it is assumed that the capacity determines the maximum number of homogeneous flow units (cars, bits, etc.) per unit of time *at a given level of service* (velocity, reliability, etc.). The models below keep the quality of service unchanged so that a capacity increment just raises the maximum number of cars that can be served. This simplification ignores that many investments pursue

the task of improving the quality of service, in which case the network operates more reliably or its participants drive faster or safer. Although these measures reduce consequences of mutual obstructions, the maximum number of vehicles passing through the network may be constant, they just perceive higher comfort. In all cases of ambiguity, the exact meaning of "capacity" will be stressed.

5.1.3 The Maximum Flow Problem Revisited

The first problem of interest is that of a network carrier who wants to maximize the flow from a unique source to a unique sink given the capacities $\boldsymbol{\kappa}$. Furthermore, capacities can be enlarged, but investment costs $\mathbf{q}^\top\boldsymbol{\delta}$ must not exceed a given budget $\bar{I}$. This budget represents an exogenous amount of resources the origin of which is not explained here. Similar to the maximum flow problem (3.1.P) on page 78 the aim is to maximize the flow $\varphi_\ell (= \mathbf{e}_\ell^\top\boldsymbol{\varphi})$ on the artificial arc ℓ which leads back from a specified sink to a distinct source. Given the network topology each flow has to satisfy the flow conservation $\mathbf{A}\boldsymbol{\varphi} = \mathbf{0}$ with the node/arc incidence matrix $\mathbf{A}$. The modified capacity restrictions are $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa} + \boldsymbol{\delta}$, where the additional budget constraint is $\mathbf{q}^\top\boldsymbol{\delta} \leqq \bar{I}$. The complete linear program is given in (5.1.P), where (5.1.D) is the corresponding dual program including the cut value $\boldsymbol{\kappa}^\top\mathbf{u}$. In contrast to the former maximum flow problem (3.1.P) the new variable $\boldsymbol{\delta}$ indicates that the new program is mainly concerned with the optimal design of the network.

(5.1)

$$\begin{array}{llrl} \text{(P)} & \max & \varphi_\ell & \\ & \text{s.t.} & \mathbf{A}\boldsymbol{\varphi} & = \mathbf{0} \\ & & \boldsymbol{\varphi} - \boldsymbol{\delta} & \leqq \boldsymbol{\kappa} \\ & & \mathbf{q}^\top\boldsymbol{\delta} & \leqq \bar{I} \\ & & \boldsymbol{\varphi} & \geqq \mathbf{0} \\ & & \boldsymbol{\delta} & \geqq \mathbf{0} \end{array} \qquad \begin{array}{llrll} \text{(D)} & \min & \boldsymbol{\kappa}^\top\mathbf{u} + \bar{I}\mathrm{b} & & \\ & \text{s.t.} & \mathbf{A}^\top\mathbf{w} + \mathbf{u} & \geqq \mathbf{e}_\ell & \\ & & -\mathbf{u} + \mathbf{q}\mathrm{b} & \geqq \mathbf{0} & \\ & & \mathbf{w} & & \text{unrestr.} \\ & & \mathbf{u} & \geqq \mathbf{0} & \\ & & \mathrm{b} & \geqq 0 & \end{array}$$

By Assumption 3.1, the capacity κ_ℓ of the artificial arc can be increased at zero cost ($q_\ell = 0$). Thus, the maximum flow $\hat{\varphi}_\ell$ is always smaller than κ_ℓ so that the complementary dual variable of the following dual problem satisfies $\hat{\mathrm{u}}_\ell = 0$.

Let $(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\delta}})$ and $(\hat{\mathbf{w}}, \hat{\mathbf{u}}, \hat{\mathrm{b}})$ designate optimal primal and dual solutions, respectively. As the increments $\boldsymbol{\delta}$ are perfectly divisible, each unit of the budget can be used to increase the flow φ_ℓ, so that $\mathbf{q}^\top\hat{\boldsymbol{\delta}} = \bar{I}$. At the same time $\mathrm{b} = 0$ cannot be an optimal solution, because then $\mathbf{u} = \mathbf{0}$ and, by the strong duality theorem, $\varphi_\ell = 0 = \boldsymbol{\kappa}^\top\mathbf{u} + \bar{I}\mathrm{b}$. Thus, $\hat{\mathrm{b}} > 0$. Both observations are reflected by the complementary slackness condition

$$\hat{\mathrm{b}}\left[\mathbf{q}^\top\hat{\boldsymbol{\delta}} - \bar{I}\right] = 0.$$

Moreover, we have

$$\hat{\delta}_e\left[q_e\hat{\mathrm{b}} - \hat{\mathrm{u}}_e\right] = 0 \quad \text{and} \quad \hat{\mathrm{u}}_e\left[\hat{\varphi}_e - (\kappa_e + \hat{\delta}_e)\right] = 0 \qquad \forall e \in \mathcal{E}.$$

When $\hat{\delta}_e > 0$ then $q_e\hat{\mathrm{b}} = \hat{\mathrm{u}}_e > 0$ and, therefore, $\hat{\varphi}_e = \kappa_e + \hat{\delta}_e$. *Each investment is used for a link working at its new capacity limit.* Furthermore

$$\hat{b}\,\bar{I} = \hat{b}\,\mathbf{q}^{\mathsf{T}}\hat{\boldsymbol{\delta}} = \hat{\mathbf{u}}^{\mathsf{T}}\hat{\boldsymbol{\delta}}. \tag{5.2}$$

An immediate interpretation of this relationship is suggested by the dimensions of the dual variables

$$[b] = \frac{\text{flow units on edge } \ell}{€} \quad \text{and} \quad [u_e] = \frac{\text{flow units on edge } \ell}{\text{flow unit of edge } e}.$$

Both sides of (5.2) express the additional flow on the artificial arc ℓ induced by an efficient use of the budget. This result is confirmed by Theorem 2.4 (Strong Duality).

$$\hat{\varphi}_\ell = \boldsymbol{\kappa}^{\mathsf{T}}\hat{\mathbf{u}} + \bar{I}\hat{b} = \hat{\mathbf{u}}^{\mathsf{T}}(\boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}) = \hat{\mathbf{u}}^{\mathsf{T}}\hat{\boldsymbol{\varphi}}. \tag{5.3}$$

In accordance with the Max Flow–Min Cut Theorem 3.2, the maximum flow value $\hat{\varphi}_\ell = \hat{\mathbf{u}}^{\mathsf{T}}\hat{\boldsymbol{\varphi}}$ equals the minimum cut value $\hat{\mathbf{u}}^{\mathsf{T}}(\boldsymbol{\kappa} + \hat{\boldsymbol{\delta}})$ of the modified network.

In response to Remark 5.1 the budget constraint $\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} \leqq \bar{I}$ should be interpreted as follows. The network carrier operates on a network with capacities $\boldsymbol{\kappa}$ and $\mathbf{q}$ denotes market prices to rent additional capacity units for one period. Consequently, $\bar{I}$ represents the available budget per period. This treatment of the budget constraint is indeed realistic if we think of several overlapping telecommunications networks, where carriers have the option to rent transshipment volumes from a competitor.

5.1.4 Minimum Investment Cost

Whereas the preceding problem computes the maximum flow φ_ℓ on the artificial arc ℓ given the budget $\bar{I}$, the following problem calculates the minimum investment cost required to reach the flow $\bar{\varphi}_\ell$. Again it is not explained where the resources needed for the investment come from. Both problems are called to be *dual*, as they capture the same aim of efficient use of the network provider's budget. Note that the term *dual* is meant to be suggestive. Economists often apply this term to pairs of problems and concepts that are technically similar except that the role of quantities and prices, and/or maximization and minimization, and/or objective function and constraints, have been reversed (Mas-Colell et al. (1995, p. 58)).

The aim is to determine the minimum investment cost which guarantees at least the flow $\bar{\varphi}_\ell$. The price vector $\mathbf{q}$ of capacity units is positive with the exception of $q_\ell = 0$. Bear in mind that capacity units can differ from link to link due to the link properties such as the length of the connections. With that the minimum problem in its extensive form is given by

$$\begin{array}{lrll}
\min & I & & \\
\text{s.t.} & \mathbf{A}\boldsymbol{\varphi} & = \mathbf{0} & \text{flow conservation} \\
& \boldsymbol{\varphi} - \boldsymbol{\delta} & \leqq \boldsymbol{\kappa} & (\bar{\varphi}_\ell < \kappa_\ell) \\
& \varphi_\ell & \geqq \bar{\varphi}_\ell & (\bar{\varphi}_\ell > 0) \\
& I - \mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} & \leqq 0 & (q_\ell = 0) \\
& I,\ \boldsymbol{\varphi},\ \boldsymbol{\delta} & \geqq \mathbf{0}. &
\end{array}$$

Starting from scratch with $\kappa = \mathbf{0}$, the optimum solution $\hat{\delta}$ represents the capacities of an optimal network with respect to total costs of supplying these capacities. As the minimum investment cost will never exceed $\mathbf{q}^{\mathsf{T}}\delta \geqq 0$, we can rewrite the problem and note its dual.

(5.5)

$$\begin{array}{llll} \text{(P)} & \min & \mathbf{q}^{\mathsf{T}}\delta & \\ & \text{s.t.} & \mathbf{A}\varphi & = \mathbf{0} \\ & & \varphi - & \delta \leqq \kappa \\ & & \mathbf{e}_\ell^{\mathsf{T}}\varphi & \geqq \bar{\varphi}_\ell \\ & & \varphi & \geqq \mathbf{0} \\ & & & \delta \geqq \mathbf{0} \end{array} \qquad \begin{array}{llll} \text{(D)} & \max & -\kappa^{\mathsf{T}}\mathbf{q} + \bar{\varphi}_\ell \mathrm{p} & \\ & \text{s.t.} & \mathbf{A}^{\mathsf{T}}\mathbf{v} - & \mathbf{q} + \mathbf{e}_\ell \mathrm{p} \leqq \mathbf{0} \\ & & & \mathbf{q} \leqq \mathbf{q} \\ & & \mathbf{v} & \text{unrestr.} \\ & & & \mathbf{q} \geqq \mathbf{0} \\ & & & \mathrm{p} \geqq 0 \end{array}$$

In order to interpret the dual variables $\mathbf{q}$ as non-negative shadow prices it is convenient to reverse the sign of the corresponding capacity restrictions, i.e., $-\varphi + \delta \geqq -\kappa$.

Regarding the artificial link it is supposed that κ_ℓ is sufficiently high, so that $\hat{\varphi}_\ell < \kappa_\ell$ and $\hat{\mathrm{q}}_\ell = 0$. When the increment of an ordinary link is positive $\hat{\delta}_e > 0$, then $q_e = \hat{\mathrm{q}}_e > 0$ and, therefore, $\hat{\varphi}_e = \kappa_e + \hat{\delta}_e$. Again all positive investment decisions refer to saturated arcs.

Let $(\hat{\varphi}, \hat{\delta})$ be a minimum solution to (5.5.P). Then, by Theorem 2.4 (Strong Duality), there is a dual solution $(\hat{\mathbf{w}}, \hat{\mathbf{q}}, \hat{\mathrm{p}})$ satisfying

$$\hat{I}(\kappa, \bar{\varphi}_\ell) = \mathbf{q}^{\mathsf{T}}\hat{\delta} = -\kappa^{\mathsf{T}}\hat{\mathbf{q}} + \bar{\varphi}_\ell \hat{\mathrm{p}}. \tag{5.6}$$

Together with the complementary slackness conditions $\hat{\delta}^{\mathsf{T}}(\hat{\mathbf{q}} - \mathbf{q}) = 0$ and $\hat{\mathbf{q}}^{\mathsf{T}}[\hat{\varphi} - (\kappa + \hat{\delta})] = 0$ we gain

$$\hat{\mathrm{p}}\,\bar{\varphi}_\ell = \hat{\mathbf{q}}^{\mathsf{T}}(\kappa + \hat{\delta}) = \hat{\mathbf{q}}^{\mathsf{T}}\hat{\varphi}. \tag{5.7}$$

This result is similar to (5.3) but the difference is immediate from the dimensions of the dual variables

$$[\mathrm{q}_e] = \frac{€}{\text{flow units of edge } e} \quad \text{and} \quad [\mathrm{p}] = \frac{€}{\text{flow units on edge } \ell}.$$

While (5.3) is measured in flow units on edge ℓ, equation (5.7) has the dimension €. The flow value on the artificial link ("revenue") equals the total flow value on all ordinary links and this is the same as the value of all capacities ("fixed costs" plus "investment costs").

The "duality" of (5.1.P) and (5.5.P) results from the following reasoning. Suppose the budget in (5.1.P) is substituted by the minimum value of (5.5.P), that is $\bar{I} = \hat{I} = I(\kappa, \bar{\varphi}_\ell)$, and inversely that the minimum flow in (5.5.P) is substituted by the maximum value of (5.1.P), that is $\bar{\varphi}_\ell = \hat{\varphi}_\ell = \varphi_\ell(\kappa, \bar{I})$. Then, an optimal solution $(\hat{\varphi}, \hat{\delta})$ to the maximum flow problem is feasible for the minimum investment cost problem. Now if we can find a feasible solution to (5.5.D) such that the strong duality result (5.6) holds true, then $(\hat{\varphi}, \hat{\delta})$ is also an optimal solution to (5.5.P).

Recall that we have argued $\hat{\mathrm{b}} > 0$. A comparison of (5.3) and (5.6) in the form

$$\frac{\bar{\varphi}_\ell}{\hat{\mathrm{b}}} = \boldsymbol{\kappa}^{\mathsf{T}} \begin{pmatrix} \hat{\mathbf{u}} \\ \hat{\mathrm{b}} \end{pmatrix} + \bar{I} \quad \text{and} \quad \hat{\mathrm{p}}\, \bar{\varphi}_\ell = \boldsymbol{\kappa}^{\mathsf{T}} \hat{\mathbf{q}} + \bar{I},$$

suggests to choose $\hat{\mathrm{p}} = 1/\hat{\mathrm{b}} > 0$ and $\hat{\mathbf{q}} = \hat{\mathbf{u}}/\hat{\mathrm{b}} = \hat{\mathrm{p}}\hat{\mathbf{u}} \geqq \mathbf{0}$. Moreover, letting $\hat{\mathbf{v}} = -\hat{\mathbf{w}}/\hat{\mathrm{b}} = -\hat{\mathrm{p}}\hat{\mathbf{w}}$, the restrictions in (5.1.D) become the constraints of (5.5.D)

$$\begin{aligned} \mathbf{A}^{\mathsf{T}}\hat{\mathbf{w}} + \hat{\mathbf{u}} \geqq \mathbf{e}_\ell &\iff -\mathbf{A}^{\mathsf{T}}\hat{\mathbf{v}} + \hat{\mathbf{q}} \geqq \mathbf{e}_\ell \hat{\mathrm{p}} \\ -\hat{\mathbf{u}} + \mathbf{q}\hat{\mathrm{b}} \geqq \mathbf{0} &\iff -\hat{\mathbf{q}} + \hat{\mathbf{q}} \geqq \mathbf{0} \end{aligned}$$

To sum up, (5.1.P) and (5.5.P) have the same optimal solutions for $\bar{I} = \hat{I}$ and $\bar{\varphi}_\ell = \hat{\varphi}_\ell$. This fact is what economists frequently refer to as being dual. This observation is reconfirmed by the following assumption. Let $\hat{\varphi}_\ell = \varphi_\ell(\boldsymbol{\kappa}, \bar{I})$ and $\hat{I} = I(\boldsymbol{\kappa}, \bar{\varphi}_\ell)$ be differentiable solutions to (5.1.P) and (5.5.P), respectively. In the case of differentiability one obtains

$$\frac{\partial \varphi_\ell(\boldsymbol{\kappa}, \bar{I})}{\partial \bar{I}} = \hat{\mathrm{b}} \quad \text{and} \quad \frac{\partial I(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \bar{\varphi}_\ell} = \hat{\mathrm{p}} \implies \frac{\partial \varphi_\ell(\boldsymbol{\kappa}, \bar{I})}{\partial \bar{I}} = \hat{\mathrm{b}} = \frac{1}{\hat{\mathrm{p}}} = \frac{1}{\dfrac{\partial I(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \bar{\varphi}_\ell}}.$$

This is the marginal effect of a budget increment on the maximum flow and vice versa. By analogy, the marginal productivity of a capacity increment follows from

$$\frac{\partial \varphi_\ell(\boldsymbol{\kappa}, \bar{I})}{\partial \kappa_e} = \hat{\mathrm{u}}_e \quad \text{and} \quad \frac{\partial I(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \kappa_e} = -\hat{\mathrm{q}}_e$$

$$\implies \frac{\partial \varphi_\ell(\boldsymbol{\kappa}, \bar{I})}{\partial \kappa_e} = \hat{\mathrm{u}}_e = \frac{\hat{\mathrm{q}}_e}{\hat{\mathrm{p}}} = -\frac{1}{\hat{\mathrm{p}}} \frac{\partial I(\boldsymbol{\kappa}, \bar{\varphi}_\ell)}{\partial \kappa_e}.$$

This result may be misleading in the sense that increasing the capacity of one link has no effect on the output when other capacity restrictions are active.

5.2 Summary In Section 5.1 the maximum flow problem for a single commodity has been augmented by a budget constraint which provides resources in order to extend link capacities. The dual linear program can still be interpreted as a minimum cut problem. The max-flow-min-cut theorem remains valid in the sense that it refers now to the extended capacities. With regard to an optimal solution it takes no wonder that all investments are used to widen the bottlenecks of the network. Contrasted with the maximum flow problem is the problem of detecting the minimum investment costs that suffice to ensure a minimum flow. Economists usually refer to these programs a being dual because the role of the flow and the investment cost as objective function or restriction is reversed. This type of duality is confirmed whenever the optimum value of one problem is used as constraint in the other program. In this case the optimum solutions coincide. □

5.2 Multicommodity Flow Problem

5.2.1 Feasibility Conditions

The maximum flow problem (5.1) is now upgraded with respect to a multicommodity flow problem. Remember at first the attempt to determine the maximum producible trip tables of a capacitated network which follows from the output distance function (3.17), p. 97, of an appropriately defined output correspondence.[7] We have seen that this approach does not yield *the* maximum output of a multicommodity network so that it is consequent to refrain from the task of maximizing a multicommodity flow. Instead, the analysis is restricted to find feasibility conditions of a given trip table.

The preceding section refers particularly to one link, namely, the fictitious flow φ_ℓ connecting the unique sink with the unique source. In what follows we return to the multicommodity flow problem with a given traffic demand $\mathbf{x}^D$. Problems of traffic generation are ignored at first. As usual, the trip table $\mathbf{x}^D$ determines the set of demand feasible route flows by virtue of the flow conservation condition

$$H^D = \left\{ \mathbf{h} \mid \mathbf{\Lambda}\, \mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0} \right\} \tag{5.8}$$

where the link flows can be computed by $\boldsymbol{\varphi} = \mathbf{\Delta}\, \mathbf{h}$. The capacity restrictions are the same as before,

$$\mathbf{0} \leqq \boldsymbol{\varphi} = \mathbf{\Delta}\, \mathbf{h} \leqq \boldsymbol{\kappa} + \boldsymbol{\delta}, \tag{5.9}$$

where $\boldsymbol{\delta}$ indicates the capacity increments of interest. In this setting the maximization of multiple path flows by increasing link capacities subject to a given budget is of minor importance because the way of expanding $\mathbf{x}^D$ is not known. However, in accordance with (5.5.P) the "dual" problem remains meaningful: what is the minimum investment cost $I = \mathbf{q}^\mathsf{T} \hat{\boldsymbol{\delta}}$ to realize the trip table $\mathbf{x}^D$? This question will be analyzed in the next section. First, a criterion is derived which indicates whether a capacity vector $\boldsymbol{\kappa}$ suffices to serve the demand $\mathbf{x}^D$ or not. By intuition, this criterion compares the capacity of each cut to the corresponding flow value of the cut.

The starting point lies at the initial capacities $\boldsymbol{\kappa}$ (with $\boldsymbol{\delta} = \mathbf{0}$). If the system (5.8) and (5.9) has no feasible solution, one has to find an (not necessarily unique) investment $\boldsymbol{\delta}$ such that the new capacities $\bar{\boldsymbol{\kappa}} = \boldsymbol{\kappa} + \boldsymbol{\delta}$ suffice to serve the demand. Of course, the criterion whether $\boldsymbol{\kappa}$ or $\bar{\boldsymbol{\kappa}}$ ensure a feasible flow is the same. A feasible traffic assignment $\mathbf{h}$ (or *routing* in the sense of data networks) exists if and only if the capacitated *multicommodity flow problem*[8] (5.10.A) has a solution $(\mathbf{h}, \mathbf{k})$ with non-negative excess capacities $\mathbf{k}$. In this case no capacity κ_e of any link must be increased

[7] The output distance function is defined by $\inf\{\lambda \geqq 0 \mid \mathbf{x} \in \lambda P(\boldsymbol{\kappa})\}$ and searches the largest feasible trip table along a ray through the origin having the direction $\mathbf{x}$; see Bobzin (1998). A positive finite solution $\hat{\lambda}$ yields the maximum feasible trip table $\mathbf{x}/\hat{\lambda}$. Such an approach embedded into a user optimal trip assignment is described in Yang et al. (2000). Of course, the capacities can be stated in a way which guarantees a predefined level of service; see Figure 3.8 on page 111.

[8] Frank et al. (2001) call this problem the edge-capacitated multicommodity flow problem in order to distinguish it from their node-capacitated multicommodity flow problem.

to realize the given trip table $\mathbf{x}^D$. By Theorem 2.6 (Farkas), the system (5.10.A) has a solution if and only if the system (5.10.B) has *no* solution.

$$
\begin{array}{lrlrl}
(5.10) & \text{(A)} & \Lambda\,\mathbf{h} & = \mathbf{x}^D & \\
 & & \Delta\,\mathbf{h} + \mathbf{k} & = \boldsymbol{\kappa} & \\
 & & \mathbf{h} & \geqq \mathbf{0} & \\
 & & \mathbf{k} & \geqq \mathbf{0} &
\end{array}
\qquad
\begin{array}{lrlrl}
 & \text{(B)} & \mathbf{x}^{D\mathsf{T}}\mathbf{u} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{v} & > 0 & \\
 & & \Lambda^{\mathsf{T}}\mathbf{u} - \Delta^{\mathsf{T}}\mathbf{v} & \leqq \mathbf{0} & \\
 & & \mathbf{u} & \geqq \mathbf{0} & \\
 & & \mathbf{v} & \geqq \mathbf{0} &
\end{array}
$$

For a proof of this statement, multiply the first and the second equation of the first system by $\mathbf{u}$ and $\mathbf{v}$, respectively. Every feasible solution $(\mathbf{u}, \mathbf{v})$ for the second system then implies

$$\mathbf{x}^{D\mathsf{T}}\mathbf{u} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{v} = \mathbf{h}^{\mathsf{T}}(\Lambda^{\mathsf{T}}\mathbf{u} - \Delta^{\mathsf{T}}\mathbf{v}) - \mathbf{k}^{\mathsf{T}}\mathbf{v} > 0.$$

Thus the validity of (5.10.B) contradicts (5.10.A) and vice versa:

$$\Lambda^{\mathsf{T}}\mathbf{u} - \Delta^{\mathsf{T}}\mathbf{v} \leqq \mathbf{0},\ \mathbf{v} \geqq \mathbf{0},\ \mathbf{k} \geqq \mathbf{0} \implies h_\rho < 0 \text{ for at least one route } \rho,$$

$$\Lambda^{\mathsf{T}}\mathbf{u} - \Delta^{\mathsf{T}}\mathbf{v} \leqq \mathbf{0},\ \mathbf{v} \geqq \mathbf{0},\ \mathbf{h} \geqq \mathbf{0} \implies \mathrm{k}_e < 0 \text{ for at least one link } e.$$

Although insufficient, a first simple check for a feasible pair $(\mathbf{u}, \mathbf{v})$ is to choose $\mathbf{v}$ as a unit vector with respect to some link e', i.e., $\mathbf{v} = \mathbf{e}_{e'}$ so that $\mathbf{v}^{\mathsf{T}}\boldsymbol{\kappa} = \kappa_{e'}$. As an example one can think of the capacity of a bridge connecting two districts of a town being separated by a river. The second inequality of the second system requires for each route $\rho \in \mathcal{R}(s, t)$

$$\mathrm{u}_{(s,t)} = \Lambda_\rho^{\mathsf{T}}\mathbf{u} \leqq \sum_{e \in \rho} \mathrm{v}_e = \begin{cases} 1 & \text{if } e' \in \rho, \\ 0 & \text{if } e' \notin \rho. \end{cases} \tag{5.11}$$

When every route from v_s (on the left hand side of the river) to v_t (on the right hand side of the river) passes link e', we are allowed to set $\mathrm{u}_{(s,t)} = 1$. The link e' cannot be substituted by any other link in which case $\mathbf{x}^{D\mathsf{T}}\mathbf{u} > \kappa_e$ indicates that there is no feasible solution $(\mathbf{h}, \mathbf{k})$ for the first system. This is the same idea as has been introduced in the minimum cut problem. The reasoning suggests to test all partitions $(\mathcal{O}, \mathcal{D})$ of the set of nodes $\mathcal{V}$. Choosing

$$\mathrm{v}_e = \begin{cases} 1 & \text{if } \alpha(e) \in \mathcal{O} \text{ and } \omega(e) \in \mathcal{D}, \\ 0 & \text{otherwise,} \end{cases}$$

yields the cut value $\boldsymbol{\kappa}^{\mathsf{T}}\mathbf{v}$ which indicates the upper limit of flows initiated in $\mathcal{O}$ and terminating in $\mathcal{D}$. Each flow unit concerned must use at least one link with $\mathrm{v}_e = 1$. Set for these flow units $\mathrm{u}_{(s,t)} = 1$ and $\mathrm{u}_{(s,t)} = 0$ otherwise. The corresponding flow value $\mathbf{x}^{D\mathsf{T}}\mathbf{u}$ is the lower limit to provide a feasible flow. If this flow value exceeds the cut value $\boldsymbol{\kappa}^{\mathsf{T}}\mathbf{v}$, no feasible route flow vector $\mathbf{h}$ exists. Notice again that this test is not sufficient to show that the first system has no feasible solution. The test simply ignores route flows which start and end in one of the two regions but lead through the other area.

A more general test is suggested by (5.11). Given the edge weights $\mathbf{v}$ the corresponding route weights are determined by $\mathbf{w} := \mathbf{\Delta}^{\mathsf{T}}\mathbf{v}$. Similar to (3.30), the largest feasible value of $u_{(s,t)}$ corresponds to the "shortest" route with $\rho \in \mathcal{R}(s,t)$.

$$\bar{u}_{(s,t)} := \min_{\rho \in \mathcal{R}(s,t)} w_\rho$$

Substituting these values into the first inequality of (5.10.B) yields the following proposition of Iri (1971)[9]; it describes technically feasible trip tables by a system of infinitely many inequalities.

5.3 Proposition *A capacity vector $\boldsymbol{\kappa}$ is feasible for* (5.10.A) *if and only if*

$$\boldsymbol{\kappa}^{\mathsf{T}}\mathbf{v} \geqq \mathbf{x}^{D\mathsf{T}}\bar{\mathbf{u}} \qquad \forall \mathbf{v} \geqq \mathbf{0}, \tag{5.12}$$

where $\bar{\mathbf{u}}$ is the vector of shortest path values with respect to edge weights $\mathbf{v}$.
Caution: $\bar{\mathbf{u}}$ must be recalculated for every $\mathbf{v}$.

This type of inequality will return in the subsequent analysis of network failures; cf. Section 5.3. Even if a feasible solution exists with respect to a fully operable network, the network carrier wants to know the consequences of failing network elements and how to bypass them by an appropriate network design. *Survivability* of transportation networks is an important practical problem that addresses the design of networks which are still functional after the failure of certain network components; see Grötschel et al. (1995). In a broader sense the probability of network failures pertains also the *reliability* of networks which refers to the probability that an initiated network process can be finished in a regular manner. On probability measures of network reliability[10] see Ball et al. (1995a).

In general many investments may exist which ensure a feasible solution for the multicommodity flow problem. The task of the following sections is to determine optimal investment decisions or to suggest heuristics that work well. A relatively easy case is to compute the cost minimal investment such that a given traffic pattern can be guaranteed for the normal operating state. An advanced problem account also for budget constraints. In cases where feasible flows exist, this budget can also be used just to improve certain features of a network. Without feasible solutions the budget may be too low to realize any appropriate investment. In this case the network provider has to decide on a *blocking strategy* by which some traffic is rejected. Such a blocking strategy may be useful even when the capacities suffice to serve the ordinary traffic demand (possibly during a busy hour) except for short-term peak loads. The network provider then has to weigh up the cost of investment and the probability to block a certain fraction of traffic. Finally, investment decisions have long term effects. While variable travel costs are determined by inputs such as fuel, investment costs represent durable inputs which can be utilized over a certain time horizon. Thus investments on network improvements can at least

[9] See also Kakusho, Onaga (1971).

[10] Contrasted with reliability of a network is the *availability* of its components, which is defined as the probability that the component is operating. Corresponding measures include the mean time to failure (that includes maintenance work) as well as the time to repair.

partially be used in the future. Furthermore, investments in material infrastructure are often irrevocably fixed. Network design then runs the risk of being completely predetermined by decisions of the past which holds true especially in the case of considerable demolition costs.

One further result of investment in infrastructure follows from savings in transport cost by the extension of presumably congested low cost paths. At least to some extent such an investment strategy may be self-financing. A computable general equilibrium model which is concerned with financing of road infrastructure with respect to savings in congestion cost is presented in Conrad, Heng (2002).

5.2.2 Minimum Investment Cost

5.2.2.1 A Reference Case

The minimum cost traffic assignment problem (3.34.P) and its dual in Chapter 3 are the starting point of the following analysis. They are restated here for convenience.

$$
\begin{array}{lll}
(5.13) & \text{(P)} \ \min \ \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h} & \text{(D)} \ \max \ \mathbf{x}^{D\mathsf{T}}\mathbf{p} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{q} \\
& \text{s.t.} \ \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^{D} & \text{s.t.} \ \boldsymbol{\Lambda}^{\mathsf{T}}\mathbf{p} - \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{q} \leqq \tilde{\mathbf{c}} \\
& \boldsymbol{\Delta}\,\mathbf{h} \leqq \boldsymbol{\kappa} & \mathbf{p} \quad \text{unrestr.} \\
& \mathbf{h} \geqq \mathbf{0} & \mathbf{q} \geqq \mathbf{0}
\end{array}
$$

where the system determining the feasible set of the primal program has been introduced in Chapter 3, (3.20) as *multicommodity flow problem*. It has a solution if and only if (5.12) holds true.

In what follows network improvements are related to capacity adjustments which are reflected by investment costs. The reference case (5.15) below introduces these aspects into (5.13.P). It is based on the following simplifying assumptions which will be relaxed more or less in the subsequent analysis.

- The trip table $\mathbf{x}^D$ is held fixed. A price sensitive demand for traffic will be considered in Section 5.2.2.5.
- The link cost vector $\mathbf{c}$ is constant and implies constant route costs via $\tilde{\mathbf{c}} = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}$. Fixed costs per link, e.g., costs of maintenance, are omitted.[11] For telecommunications networks pure transport costs are usually ignored; nevertheless transmission delays may be seen as opportunity cost of a reduced level of service. The case of zero transport cost is discussed in the next section. In contrast, Section 5.2.2.3 deals with additional fixed operating cost as well as flow dependent cost functions to model congestion effects.
- The link capacities $\boldsymbol{\kappa}$ can be changed by the amount $\boldsymbol{\delta}$, which is at first unrestricted in sign. This is not very realistic and throughout the subsequent analysis we shall assume $\boldsymbol{\delta} \geqq \mathbf{0}$. Irrevocable investments (irreversibility) in prior periods

[11] According to information from the Federal Highway Administration (FHWA, 2000) maintenance cost absorb 15% of U. S. highway revenues.

are then sunk cost. Furthermore, capacity increments are assumed to be perfectly divisible. This is supposed for convenience only, because otherwise we have to go into the details of integer programming. In this case the construction of an arc ($\psi_e = 1$) with a given capacity ($\psi_e \kappa_e$) incurs fixed arc design costs ($q_e \psi_e$); otherwise ψ_e is set to zero (all-or-nothing road, lumpiness). Apart from that many techniques for solving linear integer programs first solve the linear programming relaxation of the model with $0 \leqq \psi_e \leqq 1$ instead of $\psi_e = 0$ or 1.

- There is no budget constraint that determines an upper bound on investment. Budget constraints are introduced in Section 5.2.2.4. As has been mentioned above the budget is explained at first by an exogenous source. A model where investment costs are to be covered by corresponding payments of the travelers will be introduced at a later stage.

Let $\hat{\mathbf{h}}$ be an optimal path flow to (5.13.P) so that total transport cost $\tilde{\mathbf{c}}^T\hat{\mathbf{h}}$ is finite. Then an optimal solution $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ to (5.13.D) exists with $\tilde{\mathbf{c}}^T\hat{\mathbf{h}} = \hat{\mathbf{p}}^T\mathbf{x}^D - \hat{\mathbf{q}}^T\boldsymbol{\kappa}$. We may now think of favorable perturbations $\boldsymbol{\delta}$ of the capacities $\boldsymbol{\kappa}$, where $\mathbf{q}^T\boldsymbol{\delta}$ is the cost of this capacity adjustment. Note that the market price vector $\mathbf{q}$ and the optimal shadow price vector $\hat{\mathbf{q}}$ may be different.

Declare $\hat{\mathbf{h}}(\boldsymbol{\delta})$ to be an optimal solution to the perturbed problem so that $\hat{\mathbf{h}}(\mathbf{0})$ is an solution to the initial unperturbed problem. On the one, hand this perturbation is worth realizing if

$$\tilde{\mathbf{c}}^T\hat{\mathbf{h}}(\mathbf{0}) > \tilde{\mathbf{c}}^T\hat{\mathbf{h}}(\boldsymbol{\delta}) + \mathbf{q}^T\boldsymbol{\delta}.$$

This is the case of a self-financing investment strategy in the sense that savings in transportation cost outweigh investment cost. On the other hand, we gain:

5.4 Proposition *No perturbation*[12] $\boldsymbol{\delta}$ *improves the value of the cost minimal result in* (5.13.P) *if the market price vector* $\mathbf{q}$ *equals the shadow price vector* $\hat{\mathbf{q}}(\mathbf{0})$ *of the unperturbed problem.*

$$\tilde{\mathbf{c}}^T\hat{\mathbf{h}}(\mathbf{0}) \leqq \tilde{\mathbf{c}}^T\hat{\mathbf{h}}(\boldsymbol{\delta}) + \hat{\mathbf{q}}(\mathbf{0})^T\boldsymbol{\delta} \qquad \forall\, \boldsymbol{\delta}. \tag{5.14}$$

The proof is a consequence of optimality conditions for the following pair of dual programs. In general (5.13.D) implies $\mathbf{q} \neq \hat{\mathbf{q}}(\mathbf{0})$. In this case we have to choose an adjustment $\boldsymbol{\delta}$ such that the dual solution of the perturbed problem yields $\hat{\mathbf{q}}(\boldsymbol{\delta}) = \mathbf{q}$. The idea is to solve an augmented problem of cost minimization, where at first the shadow prices $\mathbf{q}$ in (5.15.D) can differ from $\mathbf{q}$ in (5.13.D).

(5.15)

$$\begin{array}{llll} \text{(P)} & \min\ \tilde{\mathbf{c}}^T\mathbf{h} + \mathbf{q}^T\boldsymbol{\delta} & & \\ & \text{s.t.}\ \boldsymbol{\Lambda}\,\mathbf{h} & = \mathbf{x}^D & \\ & \quad \boldsymbol{\Delta}\,\mathbf{h} - & \boldsymbol{\delta} \leqq \boldsymbol{\kappa} & \\ & \quad\ \mathbf{h} & \geqq \mathbf{0} & \\ & & \boldsymbol{\delta}\ \text{unrestr.} & \end{array} \qquad \begin{array}{llll} \text{(D)} & \max\ \mathbf{x}^{DT}\mathbf{p} - \boldsymbol{\kappa}^T\mathbf{q} & & \\ & \text{s.t.}\ \boldsymbol{\Lambda}^T\mathbf{p} - \boldsymbol{\Delta}^T\mathbf{q} \leqq \tilde{\mathbf{c}} & & \\ & \qquad\qquad \mathbf{q} = \mathbf{q} & & \\ & \qquad\ \mathbf{p} & \text{unrestr.} & \\ & \qquad\qquad \mathbf{q} & \geqq \mathbf{0} & \end{array}$$

Given positive market prices $\mathbf{q} > \mathbf{0}$, the dual program requires $\mathbf{q} = \hat{\mathbf{q}} > \mathbf{0}$. With that the complementary slackness yields

[12] Notice that $\boldsymbol{\delta}$ is unrestricted in sign!

$$[\Delta\hat{\mathbf{h}} - \hat{\boldsymbol{\delta}} - \boldsymbol{\kappa}]^{\mathsf{T}}\hat{\mathbf{q}} = 0 \quad \xrightarrow{\mathbf{q}=\hat{\mathbf{q}}>\mathbf{0}} \quad \hat{\boldsymbol{\delta}} = \Delta\hat{\mathbf{h}} - \boldsymbol{\kappa}.$$

This optimum perturbation is now applied to (5.13). After the new capacities $\hat{\boldsymbol{\kappa}} := \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}$ have been substituted into (5.13.P), the optimal shadow prices $\hat{\mathbf{q}}(\mathbf{0})$ in (5.13.D) and $\hat{\mathbf{q}}$ in (5.15.D) coincide. With that the statement of (5.14) is that no further perturbation improves the result of (5.13) at $\mathbf{q} = \hat{\mathbf{q}}$. Furthermore, one has $\hat{\mathbf{q}} > \mathbf{0}$ such that all links are filled up to their respective capacity, $\Delta\hat{\mathbf{h}} = \hat{\boldsymbol{\kappa}}$.

Proposition 5.4 now gets the following meaning: *the capacity vector $\hat{\boldsymbol{\kappa}}$ determines the smallest network capable of producing $\mathbf{x}^D$ at minimum cost, where the word "smallest" refers to the capacities of a network with a given topology.* Note that the cost term includes transportation cost as well as investment cost. At last a primal optimum solution $(\hat{\mathbf{h}}, \hat{\boldsymbol{\delta}})$ together with a dual optimum solution $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ yields

$$\tilde{\mathbf{c}}^{\mathsf{T}}\hat{\mathbf{h}} + \mathbf{q}^{\mathsf{T}}\hat{\boldsymbol{\delta}} + \boldsymbol{\kappa}^{\mathsf{T}}\hat{\mathbf{q}} = \mathbf{x}^{D\mathsf{T}}\hat{\mathbf{p}}, \tag{5.16}$$

by the strong duality theorem. The value of the trip matrix at shadow prices $\hat{\mathbf{p}}$ equals the sum of total transportation cost plus the replacement cost of the new capital stock $\boldsymbol{\kappa}+\hat{\boldsymbol{\delta}}$ evaluated at market prices $\mathbf{q} = \hat{\mathbf{q}}$. For any positive demand $x^D_{(s,t)} > 0$ there must be at least one route $\rho \in \mathcal{R}(s,t)$ with a positive flow, say $\hat{h}_\rho > 0$. In this case the complementary slackness conditions with respect to the first dual restriction result in

$$\hat{p}_{(s,t)} = \tilde{c}_\rho + \hat{\tilde{q}}_\rho \qquad \text{with} \quad \hat{\tilde{q}}_\rho := \sum_{e\in\rho} \hat{q}_e = \sum_{e\in\rho} q_e.$$

In an optimal solution the price for a journey between any OD-pair is composed of the transport cost per trip plus the cost of all capacity units occupied by the trip. Remember Remark 5.1: the interpretation of $\mathbf{q}$ as market prices for capacity units is merely valid within narrow limits.

5.2.2.2 An Application to Telecommunications Networks

Problem (5.15.P) is revisited in the network improvement model (5.17.P) with the exception that installed capacity units cannot be resold, $\boldsymbol{\delta} \geqq \mathbf{0}$. This is realistic for many real-world examples. Moreover, in telecommunications networks the transport cost $\tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h}$ are of minor importance compared to investment costs $\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta}$. The easiest case is that of $\tilde{\mathbf{c}} = \mathbf{0}$.

As there are many capacity enlargements which ensure the existence of a feasible route flow, we are now interested in those increments incurring the minimum investment costs $\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta}$. Such a model is suggested, for example, in Alevras et al. (1997). Given an arc specific *market price* vector $\mathbf{q} > \mathbf{0}$ the minimum investment cost follows from

(5.17)

$$\begin{array}{llll} \text{(P)} & \min & \mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} & \\ & \text{s.t.} & \Lambda\mathbf{h} & = \mathbf{x}^D \\ & & \Delta\mathbf{h} - & \boldsymbol{\delta} \leqq \boldsymbol{\kappa} \\ & & \mathbf{h} & \geqq \mathbf{0} \\ & & & \boldsymbol{\delta} \geqq \mathbf{0} \end{array} \qquad \begin{array}{llll} \text{(D)} & \max & \mathbf{x}^{D\mathsf{T}}\mathbf{p} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{q} & \\ & \text{s.t.} & \Lambda^{\mathsf{T}}\mathbf{p} - \Delta^{\mathsf{T}}\mathbf{q} & \leqq \mathbf{0} \\ & & & \mathbf{q} \leqq \mathbf{q} \\ & & \mathbf{p} & \text{unrestr.} \\ & & & \mathbf{q} \geqq \mathbf{0} \end{array}$$

In fact these two problems have the same target as (5.5). The primal problem is sometimes called *minimum cost capacity installation* (MCCI) for multicommodity network flows. Moreover, the variables $\boldsymbol{\delta}$ are frequently required to be integer so that the MCCI problem becomes a mixed integer problem. The complexity[13] of this problem is discussed in Bienstock et al. (1998).[14] An application to telecommunications networks with sequences of discrete capacity enlargements can be found in Wessäly (2000, p. 39). Magnanti et al. (1995) distinguish explicitly between two types of facilities that can be installed on arcs. One facility has a fixed low capacity and the other facility transmits at a multiple rate. The capacity κ_e then follows from the number of installed facilities on that link.

Let $(\hat{\mathbf{h}}, \hat{\boldsymbol{\delta}})$ and $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ be optimal solutions to the primal and the dual program, respectively. Observe at first, when the initial network holds true $\boldsymbol{\Delta}\hat{\mathbf{h}} \leqq \boldsymbol{\kappa}$, the minimum investment costs are $\hat{I} = \mathbf{q}^{\mathsf{T}}\hat{\boldsymbol{\delta}} = 0$. This implies $\hat{\boldsymbol{\delta}} = \mathbf{0}$ provided that $\mathbf{q} > \mathbf{0}$. In the other case the complementary slackness indicates for positive investments that the shadow price of the respective capacity equals the market price of the capacity increment.

$$\text{(5.18.a)} \qquad \hat{\delta}_e > 0 \implies \hat{\mathrm{q}}_e = q_e > 0 \implies \sum_{\rho:\, e\in\rho} \hat{h}_\rho = \hat{\varphi}_e = \kappa_e + \hat{\delta}_e$$

Bear in mind that the value of a link is zero as long as the road is unsaturated. Of course, the optimal investment is then zero.

$$\text{(5.18.b)} \qquad \sum_{\rho:\, e\in\rho} \hat{h}_\rho = \hat{\varphi}_e < \kappa_e + \hat{\delta}_e \implies \hat{\mathrm{q}}_e = 0 < q_e \implies \hat{\delta}_e = 0$$

Next choose a route ρ connecting the OD-pair (v_s, v_t). Given a positive route flow the shadow price $\hat{\mathrm{p}}_{(s,t)}$ of the demand $x^D_{(s,t)}$ is computed by the sum of the shadow prices of all link capacities involved. This principle is usually referred to as element based charging.

$$\begin{aligned} &\hat{h}_\rho > 0 \implies \hat{\mathrm{p}}_{(s,t)} = \hat{\tilde{\mathrm{q}}}_\rho \\ &\hat{\mathrm{p}}_{(s,t)} < \hat{\tilde{\mathrm{q}}}_\rho \implies \hat{h}_\rho = 0 \end{aligned} \qquad \text{with} \quad \hat{\tilde{\mathrm{q}}}_\rho := \sum_{e\in\rho} \hat{\mathrm{q}}_e \quad \text{and} \quad \hat{\mathrm{p}}_{(s,t)} = \boldsymbol{\Lambda}_\rho{}^{\mathsf{T}}\hat{\mathbf{p}}$$

The price $\hat{\mathrm{p}}_{(s,t)}$ is, therefore, called the *value of the shortest path* $\rho \in \mathcal{R}(s,t)$ with respect to the edge weights $\hat{\mathrm{q}}_e$. As $\mathbf{q} \geqq \mathbf{0}$ (and thus $\tilde{\mathbf{q}} \geqq \mathbf{0}$) it is always possible to choose a non-negative optimal value of $\hat{\mathrm{p}}_{(s,t)}$.

For a differentiable primal optimum value $\hat{I} = I(\boldsymbol{\kappa}, \mathbf{x}^D)$ which requires a unique dual solution we gain for the respective shadow prices

$$\frac{\partial I(\boldsymbol{\kappa}, \mathbf{x}^D)}{\partial \kappa_e} = -\hat{\mathrm{q}}_e \quad \text{and} \quad \frac{\partial I(\boldsymbol{\kappa}, \mathbf{x}^D)}{\partial x^D_{(s,t)}} = \hat{\mathrm{p}}_{(s,t)}.$$

[13] Complexity addresses the question of how much effort is needed to find a solution or to verify a suggested solution; cf. Reith, Vollmer (2001).

[14] For those who are familiar with complexity theory the authors prove particularly that the problem is "NP-hard" for general graphs.

Starting with a larger capacity κ_e reduces the investment cost $\hat{I}$ needed. An increasing demand requires a higher budget.

An interesting version of the primal program starts with the assumption that the multicommodity flow problem $\mathbf{\Delta h} \leqq \boldsymbol{\kappa}$, $\mathbf{\Lambda h} = \mathbf{x}^D$, and $\mathbf{h} \geqq \mathbf{0}$ has a feasible solution with relatively small excess capacities. Now solve the minimum investment problem (5.17.P) subject to

$$\mathbf{\Lambda h} = \mathbf{x}^D + \mathrm{d}\mathbf{x}^D.$$

This is the design problem of a network carrier who is asked to serve the additional demand $\mathrm{d}\mathbf{x}^D$. As an example, one might think of two telecommunications network carriers one of which wants to provide services by leading parts of its traffic volumes through the competitor's network. An agreement on interconnection must then clarify what path flows can be served at what prices. Alternatively, we can think of a provider whose mean demand $\mathbf{x}^D$ is increased by $\mathrm{d}\mathbf{x}^D$ during a peak load period. The idea is the same as before: find the "smallest" network characterized by $\boldsymbol{\kappa} + \boldsymbol{\delta}$ capable of serving the demand for traffic, where the size is measured in terms of investment costs. Be aware that a profit maximizing network provider can decide not to serve all of the demand. When the investment cost exceeds the expected rise in revenue, the network extension is refused and a possible excess demand will be blocked.

Although this section ignores transportation cost, data networks show in general travel costs in form of delays. Section 5.2.3.3 introduces this aspect together with a budget constraint.

5.2.2.3 Traffic Diversion

(a) Fixed Link Costs

The aim of the minimum investment cost problem (5.17.P) is similar to that of the capacitated minimum cost problem (5.15.P) with the exception of having no transport costs, i.e., $\tilde{\mathbf{c}} = \mathbf{0}$. This is realistic for telecommunications networks, but in general transportation costs cannot be neglected. Ragarding positive route or link costs the network design serves a second task. The idea is to increase capacities of certain links so that the traffic can be redirected along cheaper routes. This strategy contrasts investment cost with savings in total transportation cost and has already been mentioned as "self-financing investment program" in the context of Proposition 5.4. One problem, however, that runs the risk to be ignored here will be discussed at a later stage with respect to the survivability of networks: investment strategies that concentrate on a few least cost routes in the end bring out tree-like topologies[15] which are extremely vulnerable against network failures.

The following minimum cost problem has already been discussed in (5.15), where $\boldsymbol{\delta}$ is unrestricted in sign. If installed capacities cannot be resold, this imposes $\boldsymbol{\delta} \geqq \mathbf{0}$ as can be found in (5.17). This program, however, neglects transportation

[15] Small et al. (1992) argue that optimal investments would only lead to thicker roads on major highways.

costs $\tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h}$. Notice furthermore that the trip matrix $\mathbf{x}^D$ is fixed so that only effects of traffic diversion can be analyzed. Considering also traffic generation requires a price sensitive demand. The advanced *network improvement model*[16] as well as its dual program differ only slightly from (5.17), but we restate them here for convenience.

(5.19)

$$\begin{array}{llll} \text{(P)} & \min & \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h}+\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} & \\ & \text{s.t.} & \boldsymbol{\Lambda}\,\mathbf{h} & = \mathbf{x}^D \\ & & \boldsymbol{\Delta}\,\mathbf{h} - & \boldsymbol{\delta} \leqq \boldsymbol{\kappa} \\ & & \mathbf{h} & \geqq \mathbf{0} \\ & & & \boldsymbol{\delta} \geqq \mathbf{0} \end{array} \qquad \begin{array}{llll} \text{(D)} & \max & \mathbf{x}^{D\mathsf{T}}\mathbf{p} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{q} & \\ & \text{s.t.} & \boldsymbol{\Lambda}^{\mathsf{T}}\mathbf{p} - \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{q} & \leqq \tilde{\mathbf{c}} \\ & & \mathbf{q} & \leqq \mathbf{q} \\ & & \mathbf{p} & \text{unrestr.} \\ & & \mathbf{q} & \geqq \mathbf{0} \end{array}$$

Let the demand feasible route flow $\check{\mathbf{h}}$ be a restricted optimum solution for $\boldsymbol{\delta} = \mathbf{0}$. The idea to intensify the use of cheaper roads follows from two distinct routes connecting the same OD-pair, $\rho_1 \sim \rho_2$. If $\tilde{c}_{\rho_1} < \tilde{c}_{\rho_2}$ and $\check{h}_{\rho_1}, \check{h}_{\rho_2} > 0$, there must be a capacity restriction κ_e on the first route forcing some drivers to switch to the second inferior route. When a capacity increment δ_e allows δ_e drivers out of h_{ρ_2} to use the favorable route ρ_1 (that requires all other links on route ρ_1 to have excess capacities capable of serving δ_e), the investment is advantageous if savings in transport cost exceed investment costs,

$$\delta_e(\tilde{c}_{\rho_2} - \tilde{c}_{\rho_1}) > \delta_e q_e.$$

Of course, the new flow needs not be an optimal solution, but it improves the value of the objective function. In any case an optimal solution $(\hat{\mathbf{h}}, \hat{\boldsymbol{\delta}})$ satisfies (5.18.a), (5.18.b), and

(5.20)

$$\begin{array}{l} \hat{h}_\rho > 0 \implies \hat{\mathrm{p}}_{(s,t)} = \tilde{c}_\rho + \hat{\tilde{\mathrm{q}}}_\rho \\ \hat{\mathrm{p}}_{(s,t)} < \tilde{c}_\rho + \hat{\tilde{\mathrm{q}}}_\rho \implies \hat{h}_\rho = 0 \end{array} \qquad \text{with} \quad \hat{\tilde{\mathrm{q}}}_\rho := \sum_{e\in\rho} \hat{\mathrm{q}}_e,$$

which restates Wardrop's first principle of route choice. Given a positive traffic demand $x^D_{(s,t)}$, there must be at least one positive path flow $\hat{h}_\rho > 0$ with $\rho \in \mathcal{R}(s,t)$. Hence, $\hat{\mathrm{p}}_{(s,t)}$ determines again the *minimum route costs* on any route from v_s to v_t including the shadow price $\hat{\tilde{\mathrm{q}}}_\rho$ for that route.

$$\hat{\mathrm{p}}_{(s,t)} = \min_{\rho\in\mathcal{R}(s,t)} \tilde{c}_\rho + \hat{\tilde{\mathrm{q}}}_\rho > 0$$

All routes with higher travel cost are rejected by the travelers. The statement of (5.16) must be qualified in the sense that an excess capacity on link e implies $\hat{\mathrm{q}}_e = 0 < q_e$ as well as $\hat{\delta}_e = 0$. Spare capacity units cannot be resold and the current framework offers no chance to take advantage of their demolition.

5.5 Remark (Fixed Operating Cost per Link) The previous problems can be augmented by taking into consideration that any positive link flow $\varphi_e > 0$ incurs

[16] A general model of that type is given in Magnanti, Wong (1984). The authors assume that all or nothing of κ_e can be installed at fixed arc design cost q_e. Without capacity constraints this mixed integer problem is referred to as *fixed charge design problem*.

link specific fixed cost c_e^f. Moreover, we may assume that any positive flow requires resources[17] k_e in advance for setting up the link; cf. Ng, Rardin (1997). Another example refers to maintenance work which reduces the temporal availability of a link. In any case the corresponding fixed cost can be avoided if the link is out of service. The following minimum cost problem includes the diagonal matrix $\mathbf{K}$ having the elements k_e on the leading diagonal.

$$\begin{array}{lrl}
\min & \mathbf{c}^{\mathrm{T}}\boldsymbol{\varphi}+\mathbf{c}^{f\mathrm{T}}\boldsymbol{\psi} & \\
\text{s.t.} & \boldsymbol{\varphi}+\ \mathbf{K}\boldsymbol{\psi} \leqq \boldsymbol{\kappa} & \\
\mathbf{\Lambda h} & = \mathbf{x}^D & \\
\mathbf{\Delta h} - & \boldsymbol{\varphi} = \mathbf{0} & \\
& \boldsymbol{\varphi} - \ \mu\boldsymbol{\psi} \leqq \mathbf{0} & \\
\mathbf{h} & \geqq \mathbf{0} & \\
& \boldsymbol{\varphi} \geqq \mathbf{0} & \\
& \psi_e \in \{0,1\} \qquad \forall e \in \mathcal{E}, &
\end{array}$$

where μ is some positive constant exceeding every feasible φ_e. The restriction $\varphi_e \leqq \mu\psi_e$ guarantees $\hat{\psi}_e = 1$ for $\hat{\varphi}_e > 0$. In the other case with $\hat{\varphi}_e = 0$ the auxiliary variable ψ_e can take the values 0 or 1. As the objective function is to be minimized, we gain $\hat{\psi}_e = 0$ if $\hat{\varphi}_e = 0$. Summing up, ψ_e is one if link e is operable and working and zero if not. Campbell (1994) records the problem of fixed costs by requiring minimum flows across links, $\boldsymbol{\varphi} \geqq \underline{\boldsymbol{\kappa}}$. This flow threshold prevents links from being underutilized. The idea is the same as for hub-and-spoke networks: shut down links with an insufficient utilization or ensure minimum flows on all constructed links (close-down or continue problem). If no feasible flow $\underline{\boldsymbol{\kappa}} \leqq \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$ exists for the given traffic demand $\mathbf{x}^D$, the network must be redesigned. □

(b) Congested Links

The previous section has started with constant link costs $\mathbf{c}$ and explicit capacity constraints $\mathbf{\Delta h} = \boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$ as it is appropriate for the framework of linear programming. Consequently, traffic assignment follows a pricing scheme based on shadow prices for scarce resources. Nevertheless, Chapter 3 has shown how to apply capacities which are usually given for some predefined level of service in order to model congestion effects by virtue of appropriate link cost functions. The difficulty, however, is that an investment decision δ_e not only expands the link capacity κ_e for a constant level of service, but can also be used to improve that level of service. For example, Dantzig et al. (1979) refer to the effect of adding lanes to a road link. This investment increases, on the one hand, the capacity and, on the other hand, it may also allow to increase the speed limit which reduces the free flow travel time.

[17] Throughout the book k_e has the meaning of an excess capacity. At this point it is the minimum excess supply needed for different purposes that guarantee the operability of the link.

On the assumption of separable link cost functions, where the travel cost on each link is independent of any other flow, link cost may be written as[18]

$$c_e(\varphi_e, \kappa_e + \delta_e) := p_\tau \tau_e(\varphi_e, \kappa_e + \delta_e) \qquad \forall e \in \mathcal{E}.$$

The parameter p_τ indicates the value of one unit of time and τ_e measures the link performance including congestion effects on basis of a realized flow φ_e and the available capacity $\kappa_e + \delta_e$. One such performance function has been suggested by the U.S. Bureau of Public Roads (BPR, 1964), where τ_e measures the travel time on link e.

$$\tau_e(\varphi_e, \kappa_e + \delta_e) = t_e \left[1 + \beta_0 \left(\frac{\varphi_e}{\kappa_e + \delta_e} \right)^{\beta_1} \right]$$

This function is convex whenever the parameters β_0, β_1, and $\kappa_e + \delta_e$ as well as the free flow travel t_e are non-negative. In accordance with the above example, it may be necessary to abstain from a constant free flow travel time t_e, when capacity extension δ_e change the functionality of the link. Observe that the flow φ_e may exceed the capacity $\kappa_e + \delta_e$ at the disadvantage of remarkable transport cost reflecting a drop in quality of service due to congestion. Valuating travel times by a corresponding price, capacity restrictions now can be incorporated in a convex objective function of the form

$$f(\boldsymbol{\varphi}, \boldsymbol{\kappa} + \boldsymbol{\delta}) := \sum_{e \in \mathcal{E}} c_e(\varphi_e, \kappa_e + \delta_e)\varphi_e.$$

Aside from transportation cost it may be also necessary to consider more general non-decreasing convex[19] investment cost functions $I_e(\delta_e)$ instead of $q_e \delta_e$.

$$I(\boldsymbol{\delta}) := \sum_{e \in \mathcal{E}} I_e(\delta_e)$$

Reusing the demand feasible set Φ^D as defined on p. 105, the *system optimum design problem* is then to realize a given trip table at minimum cost including costs of travel and investment.

$$\min_{\boldsymbol{\varphi}, \boldsymbol{\delta}} \left\{ f(\boldsymbol{\varphi}, \boldsymbol{\kappa} + \boldsymbol{\delta}) + I(\boldsymbol{\delta}) \mid \boldsymbol{\varphi} \in \Phi^D,\ \boldsymbol{\delta} \geqq \mathbf{0} \right\} \tag{5.21}$$

A solution $(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\delta}})$ describes a network – including capacities $\hat{\boldsymbol{\kappa}} = \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}$ – that is capable of serving a demand feasible flow $\hat{\boldsymbol{\varphi}}$ which represents at the same time a system optimal traffic assignment. Such a solution to the design model (5.21) has

[18] The design variable δ_e refers here exclusively to capacity increments. Ben-Akiva et al. (1985) provide further design criteria that are based on safety considerations as well as on operation and construction cost.

[19] This is a strong assumption because in realistic problems many investment functions I_e show decreasing marginal cost; cf. LeBlanc, Abdulaal (1984). In this case at the most local optimum solutions can be obtained. Furthermore, Abdulaal, LeBlanc (1979a) give simple extensions of the investment function $I_e(\delta_e)$, when capacity demolitions $\delta_e < 0$ are taken into account.

an important disadvantage for route choice is carried out by network participants. Hence, the solution must be checked by using a separate user optimum assignment run on the augmented network with corresponding revisions of the investment strategy. This is done to make sure that the capacity increases are worthwhile at the user-optimum flows, not just at the system-optimum flows used in the network design model; cf. LeBlanc, Abdulaal (1984).

Revisions of the system optimum investment strategy with respect to a user optimal traffic assignment lead to the question when to stop this process. The answer can be given in the light of bilevel programming. Let $\hat{\varphi}$ designate a user optimum link flow after the capacity increments are chosen to be $\hat{\delta}$. Then, the *user optimum network design problem* is to solve

$$\min_{\delta} \left\{ f(\hat{\varphi}, \kappa + \delta) + I(\delta) \mid \hat{\varphi} \text{ solves (5.22b)}, \ \delta \geqq \mathbf{0} \right\}, \tag{5.22a}$$

$$\mathbf{c}(\hat{\varphi}, \kappa + \hat{\delta})^{\mathsf{T}} (\varphi - \hat{\varphi}) \geqq 0 \qquad \forall \varphi \in \Phi^D \quad (\hat{\delta} \text{ solves (5.22a)}), \tag{5.22b}$$

which is far more difficult than the system optimum design problem. The network design problem (5.22a) is restricted to a user optimal route choice, where the corresponding trip assignment problem (5.22b) has been introduced in Section 3.2.5.2. The group of non-cooperative travelers take the capacities $\kappa + \hat{\delta}$ as well as the route choices of other travelers as given. In doing so the user optimal traffic assignment has been interpreted as a Nash equilibrium, where all trip-makers give mutually best answers to the route choice of the others. Regarding the system designer as additional player, the Nash equilibrium is now embedded into a new game with the designer as the von Stackelberg leader and the group of the trip-makers in the dependent position (von Stackelberg follower). The designer takes into account responses in travelers' route choice to alternative investment programs, while the group of travelers treats capacities as being exogenously given to their decisions. In particular, they do not try to influence investment decisions by altering their route choice.

Notice that both problems of network design consider only the effects of reassigning traffic flows. Holding the trip table (incorporated in Φ^D) fixed and calculating traffic flows $\hat{\varphi}$ or $\hat{\mathbf{h}}$ tells us, what traffic will be *diverted*. In the short-run the only choices left to drivers are those of route (and speed), while the trip table remains unchanged. On this we may superimpose the effects of traffic *generated*. That is, the demand for traffic changes with respect to the new traffic conditions. Section 5.2.2.5 below will pick up this observation and considers explicitly the more realistic case of a price sensitive demand.

Remember also another implication of optimal solutions to (5.22): optimal investment programs concentrate network flows on fewer links with low transportation costs and enormous capacities. If such a road fails, large amounts of traffic cannot be terminated and the reliability of the network diminishes; see Section 5.3.

5.2.2.4 Budget Design Problem

The budget design problem inverts the minimum cost problem by taking an exogenous upper bound on investment costs into consideration. As an example we refer to

a government which designates a fixed amount of tax income to road infrastructure investments. The instructed network designer is then asked to assign these resources to network improvements in the "most desirable" way. In contrast the previous sections pursued the task to determine an optimal investment program irrespective of the height of investment cost. Such a strategy was shown to be self-financing at least in parts by savings in transport cost. An important disadvantage of this aproach is that travelers benefit from reduced transport cost, but the network provider carries investment cost. Consequently, a network owner, who has the right to exclude customers from network utilization, would raise prices for network services in order to finance the investment program. Nevertheless, extensive investments and delayed revenues may exceed the financial capability of the network provider by far, which imposes an additional exogenous budget constraint onto the design problem; see Remark 5.6 on p. 261 for further comments. A similar problem arises with respect to public roads.[20] If it is not possible or at least not wanted to charge people who drive on a road in proportion to their use of the service, investment programs must be financed at least partially by public tax revenue. Again network participants who benefit from infrastructure investments can differ from those who pay the tax. In any case all means of financing investments must be evaluated with respect toi their revenue potential (i.e., their ability to raise the needed funds), their equity (i.e., the fairness of distributing costs to benefits[21]), their efficiency (which pertains to the ability of the government to collect these revenues), and their political acceptability; see Gibson, Pavlak (2002).

When it is not possible to assign investment cost to those who profit from network improvements and when optimal solutions $(\hat{\mathbf{h}}, \hat{\boldsymbol{\delta}})$ to (5.19.P) require considerable investments $I = \mathbf{q}^\top \hat{\boldsymbol{\delta}}$, it seems to be plausible to introduce a new element into the analysis, namely, a budget constraint $I \leqq \bar{I}$ which determines an exogenous upper bound on investment. The advanced design problem may be explained best by the following linear program, where the trip table $\mathbf{x}^D$ as well as the route costs $\tilde{\mathbf{c}}$ are held fixed. Furthermore, the investment cost functions are assumed to be linear, $I(\delta_e) = q_e \delta_e$. The primal program (5.23.P) seeks an investment decision $\hat{\boldsymbol{\delta}}$ and a demand feasible path flow $\hat{\mathbf{h}}$ in order to minimize total travel costs. (For the sake of convenience we use a non-negative dimensionless dual variable $\mathfrak{b}$.)

[20] Public roads are sometimes referred to as public goods, presumably because they are publicly owned. With congestion in mind *non-rivalry* cannot be claimed for roads, as the service available for consumption is reduced with every car. *Non-excludability* may apply to many roads, but today at least for highly congested roads, instruments exist to exclude drivers from consuming road services at acceptable cost.

[21] As an example, observe a motor fuel tax. Equity would require that heavyweight trucks should pay a much greater tax burden due to their multiple damage to the roadway in comparison to other light vehicles. The AASHO (American Association of State Highway Officials) road test carried out in the US in 1958 in Illinois yielded so-called AASH(T)O factors which describe the impacts between axle-load and road deformation. The result was that the road deformation increases with the fourth power of axle-load.

(5.23)

$$\begin{array}{llrl} \text{(P)} & \min & \tilde{\mathbf{c}}^{\mathsf{T}}\mathbf{h} & \\ & \text{s.t.} & \mathbf{\Lambda}\,\mathbf{h} & = \mathbf{x}^{D} \\ & & \mathbf{\Delta}\,\mathbf{h} - & \boldsymbol{\delta} \leqq \boldsymbol{\kappa} \\ & & \mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} & \leqq \bar{I} \\ & & \mathbf{h} & \geqq \mathbf{0} \\ & & \boldsymbol{\delta} & \geqq \mathbf{0} \end{array} \qquad \begin{array}{llrl} \text{(D)} & \max & \mathbf{x}^{D\mathsf{T}}\mathbf{p} - \boldsymbol{\kappa}^{\mathsf{T}}\mathbf{q} - \bar{I}\mathrm{b} & \\ & \text{s.t.} & \mathbf{\Lambda}^{\mathsf{T}}\mathbf{p} - \mathbf{\Delta}^{\mathsf{T}}\mathbf{q} & \leqq \tilde{\mathbf{c}} \\ & & \mathbf{q} - \mathbf{q}\mathrm{b} & \leqq \mathbf{0} \\ & & \mathbf{p} & \text{unrestr.} \\ & & \mathbf{q} & \geqq \mathbf{0} \\ & & \mathrm{b} & \geqq 0 \end{array}$$

Notice again that the capacity restrictions together with the budget constraint determine an upper bound on traffic flows so that the trip table may be infeasible. This case has no importance when the budget $\bar{I}$ exceeds the optimum value of (5.17.P) such that a demand feasible solution with $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^{D}$ is guaranteed. In case of infeasibility the carrier has to decide in a different problem what traffic should be rejected.

Even when the primal solutions to (5.19.P) and (5.23.P) coincide, the dual solutions can differ significantly. Just assume $\mathbf{q}^{\mathsf{T}}\hat{\boldsymbol{\delta}} < \bar{I}$ so that $\hat{\mathrm{b}} = 0$ and, therefore, $\hat{\mathbf{q}} = \mathbf{0}$ with respect to (5.23.D). In contrast a positive increment $\hat{\delta}_e$ implies $\hat{\mathrm{q}}_e = q_e > 0$ in (5.19.D). The difference results from the fact that all of the resources incorporated in $\bar{I}$ are available in advance for the network designer with regard to (5.23). The design problem (5.19), however, assumes that resources are disposable at the moment when they are bought and installed. Nevertheless, if a dual solution includes the optimal value $\hat{\mathrm{b}} = 1$, then (5.19) and (5.23) describe in fact the same network design problem in the sense that $\bar{I} = \mathbf{q}^{\mathsf{T}}\hat{\boldsymbol{\delta}}$. Contrasted with (5.19), where every advantageous capacity increment is realized, we now have to concentrate on those increments with the highest cost reducing effect. In any case, on the assumption that $\hat{\mathrm{b}} = 1$, every optimal solution satisfies (5.18.a) and (5.18.b): all links with a positive investment $\hat{\delta}_e > 0$ work at their capacity limit $\kappa_e + \hat{\delta}_e$ and idle capacities contradict any positive investment. Moreover, the traffic assignment in accordance with (5.20) remains valid and especially the interpretation of $\hat{\mathrm{p}}_{(s,t)}$ as minimum route cost persists.

Observe that a link specific increment $\hat{\delta}_e > 0$ is always reflected by raising at least one route flow h_ρ using that link. As all paths consist of sequences of consecutive links, no link can be separated from its complementary links. Hence, $\hat{\delta}_e$ is almost always one part of a complex investment program, where adjustments in $\hat{\delta}_e$ imply changes in many other investments. Otherwise one runs into the risk of shifting bottlenecks merely from one link to the next; see Figure 5.2 as an example. Two features of an optimal network design reflect this problem.

Demand feasibility The available capacities along all path flows between an OD-pair may restrict the maximum flow between that pair to a value that is insufficient for $x^{D}_{(s,t)}$. In this case we are interested in one or more paths $\rho \in \mathcal{R}(s,t)$ along which all relevant capacities are expanded in order to guarantee a demand feasible path flow $\sum_{\rho\in\mathcal{R}(s,t)} \hat{h}_\rho = x^{D}_{(s,t)}$. This procedure implies path specific investment cost $\sum_{e\in\rho} q_e\hat{\delta}_e$ to be contrasted with route cost $\tilde{c}_\rho$. In fact it is misleading to concentrate on paths with minimum route cost, because this would ignore the

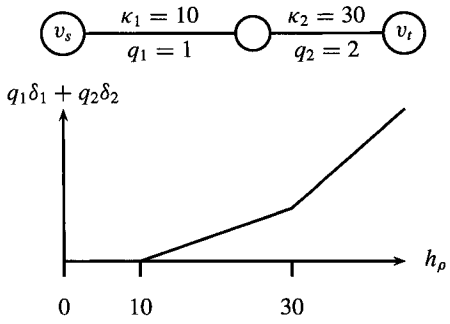


Fig. 5.2
Complementary investments
The maximum flow along route ρ joining the OD-pair (v_s, v_t) is $\kappa_1 = 10$. An investment on the first link alone can expand this capacity up to $\kappa_1 = \kappa_2 = 30$. Further extensions require investments on both links.

budget constraint. Moreover, route specific investment cost per flow unit increase when more and more links along that route become saturated.

Traffic assignment Given a demand feasible path flow, e.g., $\mathbf{\Delta}\check{\mathbf{h}} \leqq \boldsymbol{\kappa}$, one can think of investments which expand the capacity of cost minimal paths. This strategy reduces total travel costs and it is undoubtedly advantageous if savings in transport cost outweigh investment costs. Two objections indicate that the design problem (5.23) must be handled with care. (1) Although an investment program can reduce total cost, the designer is only allowed to take advantage of this effect within the limit of the budget constraint. (2) With increasing investment cost in mind the designer realizes all feasible investment programs that minimize total travel cost, even if the investment costs exceed the savings in transport cost.

Given the system optimum design problem (5.21), it is not too hard to restate the linear program (5.23) with an exogenous budget constraint as a convex program that takes congestion effects by link cost functions into account; cf. Dantzig et al. (1979).

$$\min_{\boldsymbol{\varphi},\boldsymbol{\delta}} \left\{ f(\boldsymbol{\varphi}, \boldsymbol{\kappa} + \boldsymbol{\delta}) \mid I(\boldsymbol{\delta}) \leqq \bar{I},\ \boldsymbol{\varphi} \in \Phi^D,\ \boldsymbol{\delta} \geqq \mathbf{0} \right\} \tag{5.24}$$

This problem can then be refined in the same way as (5.22) to obtain the corresponding user optimum design problem. We come back to the details of this problem in the next section, where additionally a price sensitive demand for traffic is taken into consideration; see (5.39) on p. 263. The effects of diversion are then superimposed by effects of traffic generation.

5.2.2.5 Price Sensitive Demand for Traffic

(a) Network Design with Many Non-cooperative Decision-Makers

The user optimum network design problem (5.22) indicates one of the most important problems in the construction of transport networks that fit to the needs of selfish network users who tend to ignore their individual impacts on the rest of the network. The network carrier[22], on the one hand, provides transport services which

[22] As a matter of fact we stick here to one network provider. In any event, we should *not* take it for granted that the peculiarities of road networks require that highways be and remain a

are produced on a network with features that are determined by the network designer, e.g., topology, capacities, level of service, etc. On the other hand, network users decide on what services they consume (e.g., route and mode choice) and in what quantities (e.g., travel choice). These decisions are superimposed by their choice of departure time, for example, to bypass daily rush hours. Such a temporary congestion has also a spatial dimension, where location decisions may shift the disadvantages of congestion from one area to another. All these aspects must be taken into account by the network designer whose network implements incentives or even instructions that make customers behave in a favorable way. In the end the success of a network provider can be measured in many ways, that is profit, throughput, reliability, safety and so on.

In what follows we emphasize the position of the network carrier, because railroads, telecommunications, and local public utilities all possess to some degree the characteristics of a natural monopoly[23] with corresponding market power. According to Sharkey (1987), an industry is a *natural monopoly* if total costs of production are lower when a single firm produces the entire industry output than when any collection of two or more firms divide the total output among themselves. Whether this criterion of a subadditive cost function is true or not with respect to network industries is not a question of this book. It is worth mentioning, however, that the users' comfort often increases with the extent of a network as suggested by Metcalfe's law.[24] For that reason it is worthwhile to look at the network design problem from an unregulated monopolist's point of view. This is the case of a Cournot monopoly where a *profit maximizing monopolist* supplies transport services, say trips with a certain level of service, on his privately owned network. Contrasted with the standard model, the present approach features many commodities. Afterwards the assumption of profit maximization will be dropped, which is more realistic for publicly owned networks such as a road network. The tasks are then (1) the efficient use of an exogenous budget as well as (2) welfare maximization, where the investment cost must be covered by an appropriate tax revenue.

In order to reduce the complexity of the problem we restrict the analysis to one period. Although investments in infrastructure have a strong impact on location decisions, and vice versa, the analysis of such dynamic effects require further reaching

---

government monopoly, or that the principles of private enterprise cannot be applied to solve the problem of providing road services; see Friedman, Boorstin (1996). As an example think of competitive telecommunications systems. In many countries they are operated as government-owned monopolies, but in a few countries competition is allowed, and the sector has learned how to avoid the extremes of congestion and over-investment.

[23] Although the provision of a public road network may underly the characteristics of a natural monopoly we must be careful with such statements. In many cases they just reflect public opinion that sees the provision of road, streets and highways as the responsibility of government.

[24] A wide range of network industries with characteristics of natural monopolies can be found in Shy (2001), that is, e.g., urban subway networks or several firms supplying standardized products.

approaches.[25] Moreover, it is assumed that the traffic is uniformly distributed over time. Hence, rush hours do not occur or, if the analysis refers to a busy hour, peak loads cannot be bypassed by corresponding choices in departure time. Eventually, we must be careful with the interpretation of capacity increments δ. In the short run the monopolist as a carrier seeks to produce a trip matrix $\mathbf{x}$ being a composition of link flows φ at minimum *transportation cost*, while the capital stock representing the network and especially the capacities κ is fixed. As the corresponding input choices are irrevocably set by prior decisions the physical capital stock has *sunk cost* as monetary equivalent which does not affect any profit maximizing decisions. At least for the long-run, however, the monopolist as network provider has to make decisions on the extent of his network. In the setting at hand the only "hardware" related variables that can be changed are the perfectly divisible link capacities, that is to say $\kappa \rightarrow \hat{\kappa} = \kappa + \delta$. We assume here again that $\mathbf{q}$ is a vector of given prices per capacity unit so that *investment costs* amount to $\mathbf{q}^{\mathrm{T}}\delta$. Recalling Remark 5.1, the preceding objections suggest that it may be more consistent to interpret $\mathbf{q}^{\mathrm{T}}\delta$ as cost of leasing capacities for one period rather than investment cost.

(b) Profit Maximizing Network Monopolist

The first problem to be discussed is that of a network monopoly, where the network owner pursues the task of profit maximization. With regard to a railway or airline company, the monopolist, on the one hand, bears the travel cost[26] as well as the investment cost. A possible budget constraint is not taken into consideration. On the other hand, the service provider obtains *revenues* by selling trips. Regarding these revenues it is supposed that the monopolist has full information on the (OD-specific) demand functions ξ and that he selects a vector of non-negative fares $\boldsymbol{\gamma}$, where γ_e is the fare per unit traversing link e. In doing so the fares of all trips along any route are determined, i.e., $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\gamma}$. At first sight, the trip choice as well as the route choice is left to the customers; they always select one of the cheapest feasible routes. Consequently, the route specific fares schedule the traffic volumes as well as the customers' route choice.[27]

With regard to traffic demand the monopolist has to consider mainly two aspects: (1) congestion and (2) traffic generation or diminution. Improving the performance of a link e by a positive investment $\delta_e > 0$ reduces the respective link cost and increases in turn the demand for trips traversing that link. This last effect may be caused by traffic diversion, i.e., rerouting, or by traffic generation. Similar observations hold true when fares are increased to cover investment costs for network improvements. While network improvements raise the demand for traffic, price

[25] On the special problem of dynamic benefit measures with regard to infrastructure services see Diewert (1986).

[26] On a road network travel cost are immediately born by the car drivers.

[27] MacKie-Mason, Varian (1995) discuss the problem of pricing congestible network resources within the framework of club theory which refers to the provision of shared goods. The results differ because the authors consider the demand for a single network resource as a decreasing function of traffic intensity, which is implemented here via link cost functions.

increments have the contrary effect. In any case a non-negative profit requires that revenues cover transportation cost as well as investment cost. All these details affect the optimal solution of the following profit maximization problem and are summarized in Figure 5.3.

$$\begin{aligned}
\max_{\delta,\gamma}\ & \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\varphi} - \mathbf{c}(\boldsymbol{\varphi}, \boldsymbol{\kappa} + \boldsymbol{\delta})^{\mathsf{T}}\boldsymbol{\varphi} - \mathbf{q}^{\mathsf{T}}\boldsymbol{\delta} \\
\text{s.t. } & \tilde{\boldsymbol{\gamma}} = \boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\gamma}, \quad p_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{\gamma}_\rho, && \text{(route choice)} \\
& \boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h}, \quad \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}, && \text{(flow conservation)} \\
& \mathbf{x} = \boldsymbol{\xi}(\mathbf{p}), && \text{(trip choice)} \\
& \boldsymbol{\delta} \geqq \mathbf{0}, \quad \boldsymbol{\gamma} \geqq \mathbf{0}
\end{aligned}$$

The first two restrictions convert link fares $\boldsymbol{\gamma}$ into route specific fares $\tilde{\boldsymbol{\gamma}}$, where the travelers are assumed to use one of the cheapest routes with OD-specific prices $\mathbf{p}$. The next two conditions incorporate the network topology as well as flow balance equations. The following condition takes the consumers' behavior concerning trip and route choice into account. At last all variables of the problem must be non-negative. Notice that the feasible set permits alternative expressions for the revenue in the objective function.

$$\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\varphi} = \boldsymbol{\gamma}^{\mathsf{T}}\,\boldsymbol{\Delta}\mathbf{h} = \tilde{\boldsymbol{\gamma}}^{\mathsf{T}}\mathbf{h}$$

The lowest price $p_{(s,t)}$ for a trip between the OD-pair (v_s, v_t) can be mapped to all routes joining that OD-pair. The result yields an auxiliary variable $\tilde{\mathbf{w}} := \boldsymbol{\Lambda}^{\mathsf{T}}\mathbf{p} \leqq \tilde{\boldsymbol{\gamma}}$ which reflects travelers' route choice in accordance with Wardrop's first principle: a path flow $\hat{h}_\rho$ is zero whenever $\tilde{w}_\rho < \tilde{\gamma}_\rho$. Furthermore, one has $\tilde{w}_\rho = \tilde{\gamma}_\rho$ for every positive route flow and thus

$$\tilde{\boldsymbol{\gamma}}^{\mathsf{T}}\mathbf{h} = \tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{h} = \mathbf{p}^{\mathsf{T}}\boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{p}^{\mathsf{T}}\mathbf{x} = \mathbf{p}^{\mathsf{T}}\boldsymbol{\xi}(\mathbf{p}).$$

To sum up, revenues can equivalently be expressed in link flows $\boldsymbol{\varphi}$, route flow $\mathbf{h}$, or OD-traffic volumes $\boldsymbol{\xi}$,

$$\mathbf{p}^{\mathsf{T}}\boldsymbol{\xi}(\mathbf{p}) = \tilde{\boldsymbol{\gamma}}^{\mathsf{T}}\mathbf{h} = \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\varphi}. \tag{5.25}$$

The flow chart in Figure 5.3 collects the aforementioned properties of the profit maximization problem. First of all the monopolist sets a pricing scheme $\boldsymbol{\gamma}$ and determines an investment program $\boldsymbol{\delta}$. Following the subsequent arrows, both decisions determine link costs $\mathbf{c}$, which take congestion effects into account. In the end a positive profit π requires $(\boldsymbol{\gamma} - \mathbf{c}(\cdot))^{\mathsf{T}}\boldsymbol{\varphi} > \mathbf{q}^{\mathsf{T}}\boldsymbol{\delta}$.

In general one may expect some sort of markup pricing, for if the monopolist sets $\boldsymbol{\gamma} = \mathbf{c}$ or, equivalently, $\tilde{\boldsymbol{\gamma}} = \tilde{\mathbf{c}}$, the monopoly supply is running at a loss $\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta}$. To avoid this problem the monopolist can either choose $\boldsymbol{\delta} = \mathbf{0}$ to achieve a zero profit or he sets $\gamma_e > c_e$ for at least some links. The latter case is particularly relevant when establishing a new network. Another case demonstrates one source of monopoly power. Starting at $\boldsymbol{\gamma} = \mathbf{c}$ and $\boldsymbol{\delta} = \mathbf{0}$ one can think of investments that reduce travel cost, $c_e(\boldsymbol{\varphi}, \boldsymbol{\kappa}) > c_e(\boldsymbol{\varphi}, \boldsymbol{\kappa} + \boldsymbol{\delta})$. Hence, the carrier runs the network at a profit if he

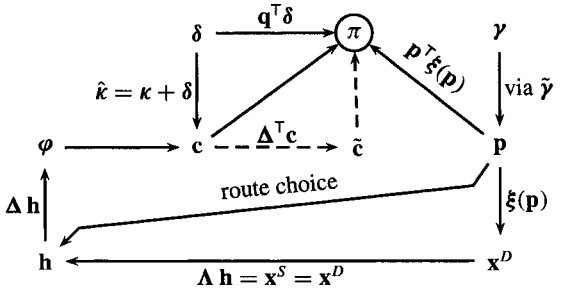


Fig. 5.3
Profit maximization (1)
The diagram represents the profit π and its main determinants.

holds the pricing scheme γ fixed, and if the network improvement δ reduces total transport cost $\mathbf{c}(\varphi, \kappa)^\top\varphi$ more than the investment cost $\mathbf{q}^\top\delta$. The only incentive for a monopolist to reduce his fares γ would be that the induced traffic raises revenue by an amount that outweighs the corresponding increase in transportation cost.

A first step towards a more detailed analysis is to eliminate the variables φ, γ, and $\mathbf{x}$ from the above profit maximization problem.

$$\begin{aligned} &\max_{\delta,\tilde{\gamma}} \ \tilde{\gamma}^\top\mathbf{h} - \mathbf{c}(\Delta\,\mathbf{h}, \kappa + \delta)^\top\Delta\,\mathbf{h} - \mathbf{q}^\top\delta \\ &\text{s.t. } p_{(s,t)} = \min_{\rho\in\mathcal{R}(s,t)} \tilde{\gamma}_\rho, \quad \Lambda\,\mathbf{h} = \xi(\mathbf{p}), \quad \delta \geqq \mathbf{0}, \quad \tilde{\gamma} \geqq \mathbf{0} \end{aligned} \tag{5.26}$$

The minimum constraint is the main problem of this program. The pricing scheme on the right hand side is set by the monopolist, but travelers accept only those routes with the lowest prices. This process seems to result in a user optimal traffic assignment, but in fact the price setting carrier is interested in a system optimal assignment that minimizes total transport cost. Such an assignment could be reached, for example, by marginal cost pricing which, however, cannot be expected to be necessarily profit maximal. Nevertheless, it seems to be reasonable that the monopolist can find a pricing scheme γ that determines a price vector $\mathbf{p}$ which in turn lead to a system optimal traffic assignment. In this case we can eliminate the minimum constraint and profit maximization becomes a problem of finding an optimal investment plan $\hat{\delta}$ together with an optimal price vector $\hat{\mathbf{p}}$. Using (5.25) for an equivalent expression of the revenue results in

$$\begin{aligned} &\max_{\delta,\mathbf{p}} \ \mathbf{p}^\top\xi(\mathbf{p}) - \tilde{\mathbf{c}}(\mathbf{h}, \kappa + \delta)^\top\mathbf{h} - \mathbf{q}^\top\delta \\ &\text{s.t. } \Lambda\,\mathbf{h} = \xi(\mathbf{p}), \quad \delta \geqq \mathbf{0}, \quad \mathbf{p} \geqq \mathbf{0} \end{aligned} \tag{5.27}$$

Observe that any choice of $\mathbf{p}$ determines the traffic demand $\mathbf{x} = \xi(\mathbf{p})$ which must be assigned to routes, i.e., $\mathbf{x} = \Lambda\,\mathbf{h}$. As such an assignment is not necessarily unique, path flows are also to be determined by the problem. The Lagrangean approach

$$\mathcal{L}(\delta, \mathbf{p}, \mathbf{h}, \mu) = \mathbf{p}^\top\xi(\mathbf{p}) - \tilde{\mathbf{c}}(\mathbf{h}, \kappa + \delta)^\top\mathbf{h} - \mathbf{q}^\top\delta + \mu^\top\big(\Lambda\,\mathbf{h} - \xi(\mathbf{p})\big)$$

with OD-specific Lagrange multipliers μ suggests KKT conditions of the form:

$$
\begin{aligned}
\nabla_{\delta}\mathcal{L} &= -\nabla_{\kappa}\tilde{\mathbf{c}}(\cdot)\hat{\mathbf{h}} - \mathbf{q} && \leqq \mathbf{0}, && \hat{\boldsymbol{\delta}} \geqq \mathbf{0}, && \nabla_{\delta}\mathcal{L}^{\mathsf{T}}\hat{\boldsymbol{\delta}} = 0,\\
\nabla_{\mathbf{p}}\mathcal{L} &= \boldsymbol{\xi}(\hat{\mathbf{p}}) + \nabla\boldsymbol{\xi}(\hat{\mathbf{p}})\hat{\mathbf{p}} - \nabla\boldsymbol{\xi}(\hat{\mathbf{p}})\hat{\boldsymbol{\mu}} && \leqq \mathbf{0}, && \hat{\mathbf{p}} \geqq \mathbf{0}, && \nabla_{\mathbf{p}}\mathcal{L}^{\mathsf{T}}\hat{\mathbf{p}} = 0,\\
\nabla_{\mathbf{h}}\mathcal{L} &= -\tilde{\mathbf{c}}(\cdot) - \nabla_{\mathbf{h}}\tilde{\mathbf{c}}(\cdot)\hat{\mathbf{h}} + \boldsymbol{\Lambda}^{\mathsf{T}}\hat{\boldsymbol{\mu}} && \leqq \mathbf{0}, && \hat{\mathbf{h}} \geqq \mathbf{0}, && \nabla_{\mathbf{h}}\mathcal{L}^{\mathsf{T}}\hat{\mathbf{h}} = 0,\\
\nabla_{\mu}\mathcal{L} &= \boldsymbol{\Lambda}\,\hat{\mathbf{h}} - \boldsymbol{\xi}(\hat{\mathbf{p}}) && = \mathbf{0}.
\end{aligned}
$$

Starting with the simplest result, the last line requires that the traffic generated equals the traffic demand. The third line relates $\hat{\mu}_{(s,t)}$ to marginal route cost as has been stated in (3.59), p. 142. With $\mathcal{R}$ as the set of all feasible paths in the network one has

$$
\hat{h}_{\rho} > 0 \implies \hat{\mu}_{(s,t)} = \tilde{c}_{\rho}(\hat{\mathbf{h}}, \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}) + \sum_{r \in \mathcal{R}} \frac{\partial \tilde{c}_r(\hat{\mathbf{h}}, \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}})}{\partial h_{\rho}}\, \hat{h}_r. \tag{5.28}
$$

Similarly, the second line compares $\hat{\boldsymbol{\mu}}$ with the marginal revenue of a price setting monopolist, that is

$$
\hat{p}_{(s,t)} > 0 \implies \sum_{b \in \mathcal{B}} \frac{\partial \xi_b(\hat{\mathbf{p}})}{\partial p_{(s,t)}}\, \hat{\mu}_b = \xi_{(s,t)}(\hat{\mathbf{p}}) + \sum_{b \in \mathcal{B}} \frac{\partial \xi_b(\hat{\mathbf{p}})}{\partial p_{(s,t)}}\, \hat{p}_b \tag{5.29}
$$

with $\mathcal{B}$ as the set of all OD-pairs. If the traffic demand for each OD-pair is supposed to be independent of all trip prices referring to other OD-pairs, then (5.29) reduces to some sort of mark up pricing in accordance with the Amoroso-Robinson relation[28]

$$
\frac{\partial \xi_{(s,t)}(\hat{\mathbf{p}})}{\partial p_{(s,t)}}\, \hat{\mu}_{(s,t)} = \xi_{(s,t)}(\hat{\mathbf{p}}) + \frac{\partial \xi_{(s,t)}(\hat{\mathbf{p}})}{\partial p_{(s,t)}}\, \hat{p}_{(s,t)} \iff \frac{\hat{p}_{(s,t)} - \hat{\mu}_{(s,t)}}{\hat{p}_{(s,t)}} = -\frac{1}{\eta_{(s,t)}} > 0,
$$

where $\eta_{(s,t)}$ denotes the direct price elasticity of traffic demand with respect to the OD-pair (v_s, v_t). The opaque form of equation (5.29) can also be put into a more familiar form if we assume invertible and continuously differentiable demand functions so that

$$
\begin{aligned}
\hat{\mathbf{p}} > \mathbf{0} &\implies \boldsymbol{\xi}(\hat{\mathbf{p}}) + \nabla\boldsymbol{\xi}(\hat{\mathbf{p}})\hat{\mathbf{p}} - \nabla\boldsymbol{\xi}(\hat{\mathbf{p}})\hat{\boldsymbol{\mu}} = \mathbf{0}\\
&\iff [\nabla\boldsymbol{\xi}(\hat{\mathbf{p}})]^{-1}\boldsymbol{\xi}(\hat{\mathbf{p}}) + \hat{\mathbf{p}} = \hat{\boldsymbol{\mu}}\\
&\iff [\nabla\boldsymbol{\xi}^{-1}(\hat{\mathbf{x}})]\,\hat{\mathbf{x}} + \boldsymbol{\xi}^{-1}(\hat{\mathbf{x}}) = \hat{\boldsymbol{\mu}}.
\end{aligned}
$$

This is what economists might have expected for a Cournot monopoly: the marginal cost $\hat{\mu}_{(s,t)}$ of a route flow as given in (5.28) equals the marginal revenue for the corresponding OD-pair. The first line of the KKT conditions evaluates a positive increment $\hat{\delta}_e > 0$ on link e by the effects it imposes onto the network as a whole, where the route flows disperse the perturbation of this investment.

$$
\hat{\delta}_e > 0 \implies q_e = -\sum_{\rho \in \mathcal{R}} \frac{\partial \tilde{c}_{\rho}(\hat{\mathbf{h}}, \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}})}{\partial \kappa_e}\, \hat{h}_{\rho} \tag{5.30}
$$

[28] An alternative formulation of the Amoroso-Robinson relation yields $\hat{\mu}_{(s,t)} = \hat{p}_{(s,t)}\left(1 + 1/\eta_{(s,t)}\right)$ by which $\hat{\mu}_{(s,t)}$ equals the marginal revenue of a trip.

Easier statements can be obtained at the cost of denying basic network features, for example, if we assume separable link cost functions.[29] Then,

$$\frac{\partial \tilde{c}_\rho(\hat{\mathbf{h}}, \kappa + \hat{\boldsymbol{\delta}})}{\partial \kappa_e} = \sum_{j \in \rho} \frac{\partial c_j(\hat{\varphi}_j, \kappa_j + \hat{\delta}_j)}{\partial \kappa_e} = \begin{cases} \partial c_e(\hat{\varphi}_e, \kappa_e + \hat{\delta}_e)/\partial \kappa_e & \text{if } e \in \rho \\ 0 & \text{otherwise} \end{cases}$$

and (5.30) reduces to

$$\hat{\delta}_e > 0 \implies q_e = -\frac{\partial c_e(\hat{\varphi}_e, \kappa_e + \hat{\delta}_e)}{\partial \kappa_e} \hat{\varphi}_e.$$

The marginal savings in cost per trip on link e due to a positive investment $\hat{\delta}_e$ for all trips passing that link, i.e., $\hat{\varphi}_e$, equals the cost q_e for the last investment unit. In other words, the capacity should be expanded to the point where the marginal cost of an extra capacity unit is equal to the marginal value of user savings brought about by that investment.

The above results confirm what has been suggested in Friedman, Boorstin (1996) with regard to a competitive framework of road provision. The ideal system of prices would have the property that the price paid for a unit of highway service would be just equal to the extra cost of providing that unit of service. Of course, there is generally more than one way of providing an additional unit of service: improving the road, new construction, reducing congestion, and so on. The efficient combination of ways of providing highway services has been reached only when a situation prevails in which the extra cost of providing an additional unit of service is the same no matter how it is provided.

(c) Profit Maximizing Network Monopolist: An Alternative Setting

The approach of the preceding section refers to a network monopoly where the monopolist adopts the problem of a system optimal traffic assignment. Indeed, that procedure is realistic for firms such as a railway company. For a privately owned road network, however, it seems reasonable that car drivers decide on route choice especially if travel costs are born by the travelers and not by the network provider. Nevertheless, the monopolist can try to influence the flow pattern in a favorable way by setting appropriate fares $\boldsymbol{\gamma}$. The former profit maximization problem (5.26) must then be revised to take care of the new decision process. As the carrier is mainly interested in a link specific investment plan, we state the new problem in terms of link flows:[30]

$$\max_{\boldsymbol{\delta}, \boldsymbol{\gamma}} \ \boldsymbol{\gamma}^\top \hat{\boldsymbol{\varphi}} - \mathbf{q}^\top \boldsymbol{\delta} \tag{5.31}$$

$$\text{s.t. } \hat{\boldsymbol{\varphi}} \text{ is a given user optimal traffic assignment,}$$

$$\boldsymbol{\delta} \geqq \mathbf{0}, \quad \boldsymbol{\gamma} \geqq \mathbf{0}$$

[29] One has $\tilde{\mathbf{c}}(\mathbf{h}, \kappa) = \boldsymbol{\Delta}^\top \mathbf{c}(\boldsymbol{\Delta}\mathbf{h}, \kappa) = \boldsymbol{\Delta}^\top \mathbf{c}(\boldsymbol{\varphi}, \kappa)$, by definition, and substitutes link cost functions $c_e(\varphi_e, \kappa_e)$ which have no influence on the remaining links.

[30] On similar formulations of bilevel programming problems see Fisk (1986), where the upper level problem is to be solved by some transportation supplier.

Having selected investments δ as well as link specific fares γ, consumer prices $\mathbf{p}$ are determined on the basis of enhanced capacities $\hat{\kappa} = \kappa + \delta$ and a route specific pricing scheme $\tilde{\gamma} = \Delta^{\top}\gamma$.

$$p_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_{\rho}(\mathbf{h}, \hat{\kappa}) + \tilde{\gamma}_{\rho} \tag{5.32}$$

While $\tilde{\gamma}$ and δ (and thus $\hat{\kappa}$) are fixed by the network provider, a traffic pattern $\mathbf{h}$ that constitutes $\hat{\varphi} = \Delta\,\mathbf{h}$ follows from the behavior of network users. In its most general form this problem of determining a user optimal traffic assignment is given by

$$\min_{(\varphi,\mathbf{x})} \left\{ z(\varphi, \mathbf{x}, \hat{\kappa}) |\ (\varphi, \mathbf{x}) \in \Phi_X^S \right\}, \tag{5.33}$$

where z is some aggregate objective function which incorporates individual trip choice and route choice as suggested by (5.32). The set of supply feasible flows

$$\Phi_X^S = \left\{ (\varphi, \mathbf{x}^S) \mid \varphi = \Delta\,\mathbf{h},\ \Lambda\,\mathbf{h} = \mathbf{x}^S,\ \varphi \geqq \mathbf{0},\ \mathbf{h} \geqq \mathbf{0} \right\} \tag{5.34}$$

does not include any explicit upper bounds on link flows, because capacity restrictions are modeled via cost functions. We have seen in Chapter 3 that there is no need for an aggregate objective function z to exist, but alternative formulations are able to describe the optimal behavior of travelers at an equilibrium, cf. Proposition 4.14 on p. 192. In the current framework with a price sensitive demand (trip choice) and route choice according to Wardrop's first principle, the user optimal traffic assignment is to find a pair $(\hat{\mathbf{h}}, \hat{\mathbf{x}}) \in H_X^S$ such that

$$\left(\tilde{\mathbf{c}}(\hat{\mathbf{h}}, \hat{\kappa}) + \hat{\tilde{\gamma}}\right)^{\top}(\mathbf{h} - \hat{\mathbf{h}}) - \xi^{-1}(\hat{\mathbf{x}})^{\top}(\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \quad \forall\, (\mathbf{h}, \mathbf{x}) \in H_X^S. \tag{5.35}$$

where H_X^S assures particularly flow conservation.

$$H_X^S := \left\{ (\mathbf{h}, \mathbf{x}) |\ \Lambda\,\mathbf{h} = \mathbf{x},\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{x} \geqq \mathbf{0} \right\} \tag{5.36}$$

Recall that the variational inequality problem (5.35) assumes the capacities $\hat{\kappa}$ as well as the route pricing scheme $\hat{\tilde{\gamma}}$ as being fixed by the carrier, see (5.31). Problem (5.35) has been extensively discussed in Section 4.2 of Chapter 3 so that we can draw here the attention to the network design problem of the carrier. In comparison to Figure 5.3 the following flow chart represents the most important features of the above bilevel programming problem (5.31) and (5.35).

The upper part shows the sphere of the profit maximizing network provider: the profit π results from revenues $\gamma^{\top}\varphi$ minus investment cost $\mathbf{q}^{\top}\delta$. Both decisions, choosing γ as well as δ, influence minimum travel cost $\mathbf{p}$ and thus enter the lower part of the figure, which refers to the combined travel choice/traffic assignment by the group of network users. Their decision on travel choice $\mathbf{x}^D$ as well as their route choice constitutes a path flow $\mathbf{h}$, which is reflected by the link flow pattern φ. This result shows two responses. It is related to the utilization of link capacities and thus indicates the degree of congestion which is incorporated in link cost functions $\mathbf{c}$. Hence, the flow pattern reenters the determination of minimum route cost $\mathbf{p}$. The

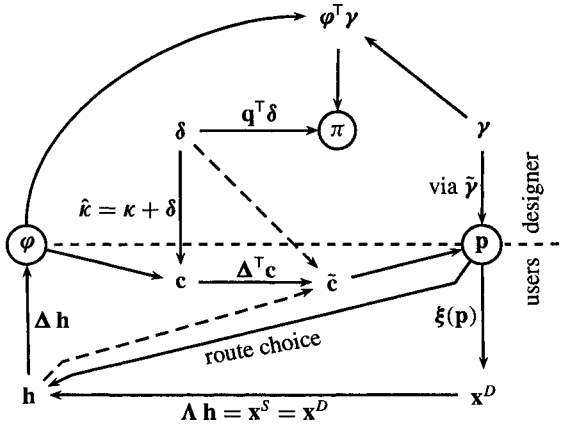


Fig. 5.4
Profit maximization (2)
The bilevel programming problem of a profit maximizing network designer.

second response refers to the network provider whose revenue is an immediate consequence of the realized traffic flow. This bilevel programming problem can hardly be solved analytically due to its complexity. In the ideal, but unrealistic case one would obtain from (5.35) reaction curves that represent link specific demand functions of the form $\varphi^D = \varphi(\kappa + \delta, \gamma)$ which could be used to solve the upper-level program (5.31) of the network designer who is in the independent position of a von Stackelberg leader.

$$\max_{\delta,\gamma} \left\{ \gamma^T \varphi(\kappa + \delta, \gamma) - q^T \delta \mid \delta \geqq 0,\ \gamma \geqq 0 \right\} \tag{5.37}$$

As this solution technique works hardly ever even for extreme simplifications, computer simulation is the only way to evade the problem. The aim is then to successively solve (5.31) and (5.35) with the hope that the corresponding algorithm converges to a situation in which the profit maximizing choice $(\hat{\delta}, \hat{\gamma})$ of the von Stackelberg leader implies a user optimal traffic assignment $\hat{\varphi}$.

Disregarding pure investment decisions let us sum up the most important results on pricing road services. The network provider must choose a pricing scheme $\hat{\tilde{\gamma}} = \Delta^T \hat{\gamma}$ that meets best the following three tasks:

1. The toll system $\hat{\gamma}$ impinges on the carrier's revenue $\hat{\gamma}^T \varphi$.
2. Route specific transport costs $\tilde{c}(\cdot) + \hat{\tilde{\gamma}}$ determine traffic volumes x^D.
3. Link specific tolls $\hat{\gamma}$ influence the user optimal traffic assignment $\hat{\varphi}$.

If profit maximization is not wanted for public roads one might think of a government monopoly providing the network "free of charge", i.e., $\gamma = 0$ as supposed in the next section. But then the network provider loses also control of traffic guidance. These results confirm what can be found in Friedman, Boorstin (1996): *[on] a crowded road ... it would be desirable to discourage traffic*, and that *the people who drive on a road should be charged ... in proportion to their use of the service.*

5.6 Remark (Financing Georgia's Highway Infrastructure) One may ask whether the profit maximization problem of the network provider is a real world problem. Indeed, the largest source of transportation infrastructure revenues for road

construction and maintenance are motor fuel taxes and vehicle taxes, followed by tolls; they account for 52% of Georgia's national average funding sources in the year 2000, see Gibson, Pavlak (2002). This observation suggests to interpret the pricing scheme $\boldsymbol{\gamma}$ in terms of a motor fuel tax. As all roads have a fixed distance, the fuel consumption is approximately constant and γ_e would result from fuel tax per liter times fuel use per kilometer[31] times distance of link e. The implicit "toll scheme" has only limited influence on road guidance for all road services show the same price, even though the traffic diminishing effect may help to reduce congestion.

A second annotation emphasizes that the design problem not necessarily includes a budget constraint. With regard to Georgia's highway infrastructure development Gibson, Pavlak (2002) expose: *The state's highway financing plan uses the federal government's Grant Anticipation Revenue Vehicle (GARVEE) bonding program to accelerate the Governor's Road Improvement Program (GRIP) construction timetable. GRIP targets a network of developmental roads and highways across Georgia. The GARVEE program is a funding mechanism that allows states to raise funds through bond issues that are repaid with future federal transportation "turnback" funds (from federal motor fuel taxes) plus state contributions, if necessary. While this infusion of capital enables Georgia to move forward with much-needed road construction, it will not provide for the costs of future road maintenance. Thus, identifying additional means of increasing state transportation revenues remains an important issue for state policymakers.* In fact, their problem is not too different from that of a profit maximizing network provider. □

(d) Cost Minimization under a Financing Restriction

In the context of a publicly owned road network the carrier is usually forced to operate in the manner of non-profit organizations. Its objective is to design a network that minimizes total transport cost, where the resources to finance investments are fixed in advance, for example, by the government. Road usage is free of charge ($\boldsymbol{\gamma} = \mathbf{0}$) so that the designer has less instruments of traffic guidance. A model that describes the efficient use of state assets requires similar ingredients as before: (1) supply, (2) demand, and (3) market clearing as well as (4) an exogenous budget.

(1) The *supply side* uses link cost functions $\mathbf{c}(\boldsymbol{\varphi}, \boldsymbol{\kappa})$ to capture capacitated links which cause congestion effects when link loads exceed the respective capacities. Moreover, route costs result from $\tilde{\mathbf{c}}(\mathbf{h}, \boldsymbol{\kappa}) = \boldsymbol{\Delta}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\Delta}\,\mathbf{h}, \boldsymbol{\kappa})$ so that total transport cost are equivalently given by

$$f(\boldsymbol{\varphi}, \boldsymbol{\kappa}) := \mathbf{c}(\boldsymbol{\varphi}, \boldsymbol{\kappa})^{\mathsf{T}}\boldsymbol{\varphi} \quad \text{or} \quad \tilde{f}(\mathbf{h}, \boldsymbol{\kappa}) := \tilde{\mathbf{c}}(\mathbf{h}, \boldsymbol{\kappa})^{\mathsf{T}}\mathbf{h}.$$

As capacity restrictions are modeled via cost functions, the set of supply feasible flows Φ_X^S defined by (5.34) does not include any explicit upper bounds on link flows.

[31] Increased fuel efficiency among cars and trucks may considerably reduce tax revenue! If travel would not increase, a perverse incentive holding tax revenue constant might be to cause artificial congestion. Roth (1996) cautions against local governments as the most likely monopolists to wield market power in road space. In doing so they could discover a *cash cow in traffic congestion.*

It just indicates trip tables $\mathbf{x}^S$ that can be realized with respect to flow conservation on the given network.

(2) The *demand side* takes travel choice as well as path choice into account but it ignores here mode choice and location decisions. As before the trip matrix $\mathbf{x}^D$ results from price sensitive demand functions

$$\mathbf{x}^D = \boldsymbol{\xi}(\mathbf{p}),$$

where again $p_{(s,t)}$ indicates the minimum trip cost between the OD-pair (v_s, v_t) prevailing at path flow pattern $\mathbf{h}$ and capacities $\boldsymbol{\kappa}$. That is (5.32) but with $\tilde{\boldsymbol{\gamma}} = \mathbf{0}$:

$$p_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho(\mathbf{h}, \boldsymbol{\kappa}) \tag{5.38}$$

In accordance with Wardrop's first principle, trip-makers choose always routes with the lowest cost, which results in a user optimal trip assignment.

(3) *Market clearing* requires a supply and demand feasible flow pattern $\hat{\mathbf{h}}$ (or $\hat{\boldsymbol{\varphi}}$) as well as a pricing scheme $\hat{\mathbf{p}}$ such that

$$\boldsymbol{\Lambda}\, \hat{\mathbf{h}} = \mathbf{x}^S = \mathbf{x}^D = \boldsymbol{\xi}(\hat{\mathbf{p}}).$$

(4) An exogenous *budget constraint* $\bar{I}$ limits investment costs $I(\boldsymbol{\delta})$ that result from network improvements such as construction of links or expansion of link capacities $\boldsymbol{\delta}$.

$$I(\boldsymbol{\delta}) \leqq \bar{I}$$

Notice that the previous models assume constant cost q_e per capacity unit on link e. The term $I(\boldsymbol{\delta})$ is merely a more general form of $\mathbf{q}^\top \boldsymbol{\delta}$.

The following network design problem is similar to the preceding bilevel programming problem (5.31) and (5.35), but the warrants of the network designer involve several crucial modifications. The objective is to minimize total travel cost by corresponding network improvements, while revenues are dropped from the analysis by providing road services free of tolls, i.e., $\boldsymbol{\gamma} = \mathbf{0}$. Thus traffic volumes as well as traffic assignment cannot be guided through a toll scheme so that the flow pattern would suit to the capabilities of the network. When revenues via toll collects are ruled out, the network designer needs other means to finance investments in infrastructure. These resources are made available in form of a limited budget, for example, by the government which collects a gasoline tax. In this setting the network designer indeed acts like a von Stackelberg leader and the group of network users takes the dependent position of a von Stackelberg follower.

The upper-level problem of the network designer is to minimize total travel cost subject to an exogenous budget constraint $\bar{I}$. In contrast with (5.24), a price sensitive traffic demand superimposes the pure effects of traffic assignment. As the investment decisions are link specific, the advanced user optimum network design problem is given in link flow notation.

$$\min_{\boldsymbol{\delta}} \left\{ f(\hat{\boldsymbol{\varphi}}, \boldsymbol{\kappa} + \boldsymbol{\delta}) \mid I(\boldsymbol{\delta}) \leqq \bar{I},\ \boldsymbol{\delta} \geqq \mathbf{0},\ (\hat{\boldsymbol{\varphi}}, \hat{\mathbf{x}}) \text{ solves (5.40)} \right\} \tag{5.39}$$

The pair $(\hat{\varphi}, \hat{\mathbf{x}})$ is restricted to solve some lower-level problem, which presupposes the decisions of the network designer – namely, the new capacities $\hat{\kappa} = \kappa + \hat{\delta}$ – as fixed. According to (5.35) with $\tilde{\gamma} = \mathbf{0}$, the following variational inequality problem seeks a pair $(\hat{\mathbf{h}}, \hat{\mathbf{x}}) \in H_X^S$ which corresponds to a *user optimum* after $\hat{\delta}$ has been installed.

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}, \hat{\kappa})^\top (\mathbf{h} - \hat{\mathbf{h}}) - \xi^{-1}(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \quad \forall (\mathbf{h}, \mathbf{x}) \in H_X^S. \tag{5.40}$$

This problem is given in route flows $\mathbf{h}$ to emphasize that travelers select between different routes. Remember the problem that a unique link flow vector may be compatible with many route flow vectors, $\varphi = \Delta \mathbf{h}$. The interconnection between the involved variables may be taken from the following figure.

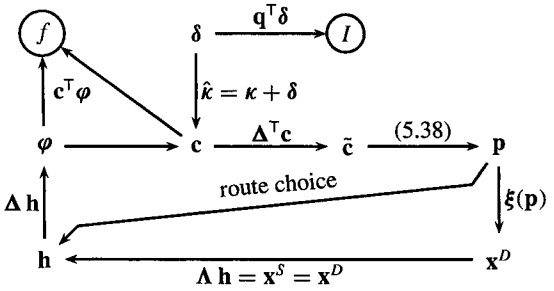


Fig. 5.5
Bilevel programming
The budget design problem seeks optimal capacity increments δ taking the aggregate response of network users into account.

The budget design problem, as described by the bilevel programming problem (5.39) and (5.40), includes several disadvantages that cannot be ignored.

Solutions Although $\gamma = \mathbf{0}$, the problem has the same complexity as solving (5.31) and (5.35). Closed solutions as indicated by (5.37) cannot be assumed to exist so that computer simulations seem to be the only way to find approximate solutions to the large scale network design problem.

Pricing Providing road services free of charge, that is we set $\gamma = \mathbf{0}$ in advance, the government gives away the decisive instrument to reduce congestion effects. Neither total traffic volumes nor route choice (nor time of day utilization nor user types[32]) can be controlled without charging the travelers use of service.

Cost-benefit A solution to the budget design problem implements the efficient use of state assets $\bar{I}$. But do the network improvements following from investments δ justify investment cost $I(\delta) = \bar{I}$? Savings in total transport cost is an insufficient criterion for it ignores traffic creation due to increased comfort of travel. Moreover, the budget $\bar{I}$ may be too low or even too high in which case the designer might realize projects of minor importance.

Power without responsibility Roth (1996) draws the attention to proper incentives to be implemented in the political decision process. He reports that the US federal "Highway Trust Fund" receives the bulk of highway revenues from dedicated

[32] For example, one can think of weight specific tolls that reflect road damage in relation to vehicle weight.

road user charges and distributes them to the states by formula. This results in excessive demand for expensive facilities, because to the states, which are nominally responsible for expenditure decisions, federal funds are costless, and state officials are accountable to their voters only for state funds. Neither road users nor state officials nor federal officials are responsible for paying the costs that arise from their decisions to spend. The road users have no control over their payments into the federal Highway Trust Fund; members of federal Congress have no control over the distribution of the funds between the states; and the states have no control over the amount of federal funds they receive. In this situation, it takes no wonder that the efficient use of state assets seems to be out of reach.

All these objections suggest that publicly owned (road) networks which are provided free of charge tend to be badly designed. Poor services on congested links as well as oversized parts of networks seem to be omnipresent. In this sense Viton (1995) investigates the question under what conditions the private sector will find it profitable to compete with an existing network of public roads by means of building, maintaining, and operating parallel facilities. His results indicate a surprisingly wide range of profitability, where private firms provide improved road services on the basis of profit maximizing pricing schemes. As an example, think of a second causeway connecting the German island Rügen with the main-land. The corresponding so-called Strelasund project was planned though not realized as a toll road parallel to the existing causeway on the basis of a private-public partnership. This case corresponds to the second best congestion pricing in case of an untolled alternative as discussed in Verhoef et al. (1996), where some regulator aims at maximizing total benefits.

(e) Net Benefit Maximization

Maximizing the net benefit for all network participants can be explained best by starting with the system optimum design problem (5.27) of a profit maximizing network provider. All that must be done is to substitute the monopolist's revenue $\mathbf{p}^{\mathsf{T}}\boldsymbol{\xi}(\mathbf{p})$ by the willingness to pay $k(\mathbf{x}^D)$ of the group of travelers. This term has been introduced in Chapter 3, see 200, and is given by

$$k(\mathbf{x}^D) = \int_{\mathbf{0}}^{\mathbf{x}^D} \boldsymbol{\xi}^{-1}(\mathbf{z})^{\mathsf{T}} \mathrm{d}\mathbf{z},$$

where it is assumed that the demand functions $\boldsymbol{\xi}$ are invertible.[33] For consistency, let us express total travel cost by a corresponding line integral of marginal route costs $\tilde{\mathbf{c}}$.

$$\tilde{f}(\mathbf{h}, \boldsymbol{\kappa}) := \tilde{\mathbf{c}}(\mathbf{h}, \boldsymbol{\kappa})^{\mathsf{T}}\mathbf{h} = \int_{\mathbf{0}}^{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{z}, \boldsymbol{\kappa})^{\mathsf{T}} \mathrm{d}\mathbf{z}$$

Since the set of feasible path flows H_X^S as given in (5.36) remains unchanged, we can rewrite (4.32), p. 200, with respect to an optimal network design as follows

[33] The revenue $\mathbf{p}^{\mathsf{T}}\boldsymbol{\xi}(\mathbf{p})$ is the appropriate term, when the monopolist acts as a price setter.

$$\max_{\delta,\mathbf{h},\mathbf{x}^D} \left\{ k(\mathbf{x}^D) - \tilde{f}(\mathbf{h}, \kappa + \delta) - \mathbf{q}^\top \delta \mid \delta \geqq \mathbf{0},\ (\mathbf{h}, \mathbf{x}^D) \in H_X^S \right\}.$$

The notation $(\mathbf{h}, \mathbf{x}^D) \in H_X^S$ stresses market clearing $\mathbf{\Lambda}\,\mathbf{h} = \mathbf{x}^S = \mathbf{x}^D$. The trip table can therefore be eliminated from the preceding maximum problem[34]

$$\max_{\delta,\mathbf{h}} \left\{ k(\mathbf{\Lambda}\,\mathbf{h}) - \tilde{f}(\mathbf{h}, \kappa + \delta) - \mathbf{q}^\top \delta \mid \delta \geqq \mathbf{0},\ \mathbf{h} \geqq \mathbf{0} \right\}. \tag{5.41}$$

The KKT conditions for $(\hat{\delta}, \hat{\mathbf{h}})$ be an optimal solution to this problem turn out to be

$$\left[\mathbf{\Lambda}^\top \xi^{-1}(\mathbf{\Lambda}\,\hat{\mathbf{h}}) - \tilde{\mathbf{c}}(\hat{\mathbf{h}}, \kappa + \hat{\delta})\right]^\top \hat{\mathbf{h}} = 0, \quad \mathbf{\Lambda}^\top \xi^{-1}(\mathbf{\Lambda}\,\hat{\mathbf{h}}) - \tilde{\mathbf{c}}(\hat{\mathbf{h}}, \kappa + \hat{\delta}) \leqq \mathbf{0}, \quad \hat{\mathbf{h}} \geqq \mathbf{0},$$
$$\left[\nabla_\kappa \tilde{f}(\hat{\mathbf{h}}, \kappa + \hat{\delta}) + \mathbf{q}\right]^\top \hat{\delta} = 0, \qquad \nabla_\kappa \tilde{f}(\hat{\mathbf{h}}, \kappa + \hat{\delta}) + \mathbf{q} \geqq \mathbf{0}, \quad \hat{\delta} \geqq \mathbf{0}.$$

The first line requires *marginal cost pricing* for all routes $\rho \in \mathcal{R}(s,t)$ in use to obtain a system optimal traffic assignment,

$$\hat{h}_\rho > 0 \implies p_{(s,t)} = \xi_{(s,t)}^{-1}(\mathbf{\Lambda}\,\hat{\mathbf{h}}) = \tilde{c}_\rho(\hat{\mathbf{h}}, \kappa + \hat{\delta})$$

Whenever marginal route cost $\tilde{c}_\rho$ exceed the marginal willingness to pay $\xi_{(s,t)}^{-1}$ with $\rho \in \mathcal{R}(s,t)$, the route is out of service, i.e., $\hat{h}_\rho = 0$, and thus

$$p_{(s,t)} = \min_{\rho \in \mathcal{R}(s,t)} \tilde{c}_\rho(\mathbf{h}, \cdot).$$

Recalling the computation of marginal route costs $\tilde{\mathbf{c}}$, the route specific fares $\hat{\tilde{\boldsymbol{\gamma}}}$ follow from

$$\tilde{\mathbf{c}}(\hat{\mathbf{h}}, \kappa + \hat{\delta}) = \tilde{\mathbf{c}}(\hat{\mathbf{h}}, \kappa + \hat{\delta}) + \hat{\tilde{\boldsymbol{\gamma}}} \quad \text{with} \quad \hat{\tilde{\boldsymbol{\gamma}}} := \nabla_{\mathbf{h}} \tilde{\mathbf{c}}(\hat{\mathbf{h}}, \kappa + \hat{\delta})\,\hat{\mathbf{h}}.$$

As large networks include much more routes than links, a link specific price scheme $\boldsymbol{\gamma}$ simplifies this representation greatly,[35]

$$\mathbf{c}(\hat{\boldsymbol{\varphi}}, \kappa + \hat{\delta}) = \mathbf{c}(\hat{\boldsymbol{\varphi}}, \kappa + \hat{\delta}) + \hat{\boldsymbol{\gamma}} \quad \text{with} \quad \hat{\boldsymbol{\gamma}} := \nabla_{\boldsymbol{\varphi}} \mathbf{c}(\hat{\boldsymbol{\varphi}}, \kappa + \hat{\delta})\,\hat{\boldsymbol{\varphi}}.$$

Premultiplying this equation by the transpose of $\mathbf{\Delta}$ proves that both pricing schemes coincide because $\tilde{\mathbf{c}} = \mathbf{\Delta}^\top \mathbf{c}$, $\tilde{\mathbf{c}} = \mathbf{\Delta}^\top \mathbf{c}$, and therefore $\tilde{\boldsymbol{\gamma}} = \mathbf{\Delta}^\top \boldsymbol{\gamma}$ (see also p. 140).

The second line of the KKT conditions yields especially

$$\hat{\delta}_e > 0 \implies q_e = -\frac{\partial \tilde{f}(\hat{\mathbf{h}}, \kappa + \hat{\delta})}{\partial \kappa_e} = -\sum_{\rho \in \mathcal{R}} \frac{\partial \tilde{c}_\rho(\hat{\mathbf{h}}, \kappa + \hat{\delta})}{\partial \kappa_e}\,\hat{h}_\rho.$$

[34] Arnott, Kraus (1995) apply this problem to one OD-pair but for two periods in order to derive a first-best congestion charge taking into account commuters' departure time.

[35] Revenues from congestion pricing $\boldsymbol{\gamma}^\top \boldsymbol{\varphi}$ do not necessarily coincide with investment cost $\mathbf{q}^\top \delta$! Provided revenues exceed investment cost, Small (1992) suggests to use the surplus for monetary reimbursements to travelers, possibly in the same way as discussed in Remark 3.41, p. 150, or for substitution of general taxes.

This is the same investment rule as has been discussed for the profit maximizing network provider, see (5.30).

According to the flow charts of Figure 5.3, 5.4, and 5.5 the following diagram represents the interrelationships between the used variables. Due to the numerous relations, the chart is slightly rearranged and shows particularly marginal costs rather than costs per trip. Nevertheless, it is useful to start reading at the investment program δ.

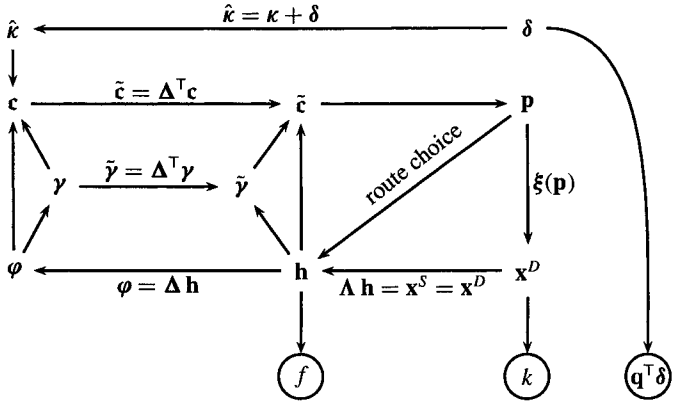


Fig. 5.6 Maximizing the net benefit
The quantities to be determined are δ, $\mathbf{x}^D$, and $\mathbf{h}$. Marginal cost pricing generates γ, $\tilde{\gamma}$, and especially $\mathbf{p}$ endogenously.

When marginal cost pricing is not possible or politically undesirable so that travelers decide on their route choice on the basis of perceived travel cost $\tilde{\mathbf{c}}$, one may ask for a *second best network design*[36] given the corresponding user optimal traffic assignment. Again, the network designer acts like a von Stackelberg leader. He decides on the investment program δ taking into account the response of trip-makers.

$$\max_{\delta} \left\{ k(\hat{\mathbf{x}}) - \tilde{f}(\hat{\mathbf{h}}, \kappa + \delta) - \mathbf{q}^T \delta \mid \delta \geqq \mathbf{0},\ (\hat{\mathbf{h}}, \hat{\mathbf{x}}) \text{ solves (5.40)} \right\}$$

This upper-level program involves a solution to the lower-level program (5.40), which seeks a pair $(\hat{\mathbf{h}}, \hat{\mathbf{x}})$ that solves the combined trip choice/traffic assignment problem given the capacities $\hat{\kappa} = \kappa + \hat{\delta}$. The according variational inequality formulation is restated here for the sake of convenience.

$$\text{(5.40)} \qquad \tilde{\mathbf{c}}(\hat{\mathbf{h}}, \hat{\kappa})^T (\mathbf{h} - \hat{\mathbf{h}}) - \xi^{-1}(\hat{\mathbf{x}})^T (\mathbf{x} - \hat{\mathbf{x}}) \geqq 0 \quad \forall (\mathbf{h}, \mathbf{x}) \in H_X^S.$$

It should be clear from the results obtained so far, that a user optimal traffic pattern $(\hat{\mathbf{h}}, \hat{\mathbf{x}})$ differs from its system optimal counterpart as long as route cost differ from

[36] Wilson (1983) puts this question into a slightly different context. When travel cost to not cover marginal cost of road congestion, does this underpricing raise or lower the optimal level of highway investment? The answer essentially depends on the price elasticity of travel demand.

marginal route cost, that is to say $\tilde{\mathbf{c}}(\mathbf{h}, \hat{\kappa}) - \tilde{\mathbf{c}}(\mathbf{h}, \hat{\kappa}) = \nabla_{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{h}, \hat{\kappa})\, \mathbf{h}$. As a pricing scheme of the form $\tilde{\gamma} = \nabla_{\mathbf{h}} \tilde{\mathbf{c}}(\mathbf{h}, \hat{\kappa})\, \mathbf{h}$ has been ruled out, investments δ are the only instrument to exert a favorable influence. This approach presumably distorts an optimal solution of (5.41) in favor of a network design with smaller side effects between roads or routes and that can be reached by an oversized network. Such a result would at least confirm that the primary benefit of congestion pricing may not be to discourage traffic, but it results in a more efficient utilization of network resources.

The preceding bilevel programming problem includes similar elements as the multi-modal transportation model in Clegg et al. (2001). Although the authors do not target at an optimal network design, they seek an optimal control comprising public transport fares, parking charges, road-use charges, and signal green-times in order to optimize some overall objective function subject to a user-optimal traffic assignment with price sensitive traffic demand. It is argued that system optimality cannot be reached when the control variables underlie further external restrictions that prevent marginal cost pricing. The corresponding second best problem is then tackled by a solution method that solves a sequence of linear programs. It estimates approximately optimal prices for transportation taking account of real-life constraints. In principle it should be possible to apply this method to our network design problem.

In what follows we shall discuss further problems of network construction. The analysis particularly exposes one feature of network design presented so far: cost minimal investment programs have a tendency towards a tree-like network topology with a few broad roads. These kind of networks may have the advantage that *economies of scale* reduce transport cost per unit on the main links as considered in the next section referring to hub-and-spoke networks. The opposite disadvantage will be discussed in Section 5.3: when considerable traffic volumes are concentrated on a few links, *failures of network components* may cause severe losses. Hence, redundancy is an important issue to improve the survivability of networks.

5.7 Summary Having started with the max flow problem for a single commodity, this section takes multicommodity flows into consideration. In this setting it is considerably more difficult to check whether a feasible flow exists or not. We have shown for a reference case where capacity adjustments are unrestricted in sign how to determine the smallest network that is able to support a given flow pattern. One essential result with respect to an optimal solution is that the shadow prices of link capacities coincide with their market prices. Such an outcome cannot be expected to hold good when capacity investments must be non-negative. Another result has been called element based charging and is related to the shadow prices of trips. Such a price denotes, on the one hand, the marginal investment costs of an additional trip. On the other hand, it is the sum of link specific shadow prices, where the sum is taken over all links involved in the corresponding shortest path.

The subsequent analysis takes traffic diversion and traffic generation into consideration. Traffic diversion describes effects on rerouting due to an improved network design, where the concentration of traffic volumes on some appointed links is of particular interest. These implications are in general superimposed by some sort of traffic generation which additionally takes a price sensitive demand for traffic

into account. Aside from some plain settings, the task of an optimal network design becomes now too complex to be solved at once. The corresponding expositions have shown how to split such complicated models into two smaller but interdependent problems. This bilevel programming approach can then be put into the context of game theory. The von Stackelberg approach has the appealing interpretation of a network designer in the leading position and network users in the dependent position. □

5.2.3 Special Cases

5.2.3.1 Hub-and-Spoke Networks

Hub-and-spoke networks are utilized by many distribution systems including airline passenger carriers, express package delivery networks[37], and communications networks. Following O'Kelly (1986) or Bryan, O'Kelly (1999), the idea of hub-and-spoke networks is as appealing as simple. *Hubs* serve as transshipment points and allow the replacement of direct connections between all nodes with fewer indirect connections. Aside from interhub links, the *spokes* connect all minor nodes with one hub. Provided there is only one hub, the term *hub-and-spoke network* is meant to be suggestive as it looks like a wheel. According to Figure 5.7, in a pure hub-and-spoke network all links must begin or end at a hub.

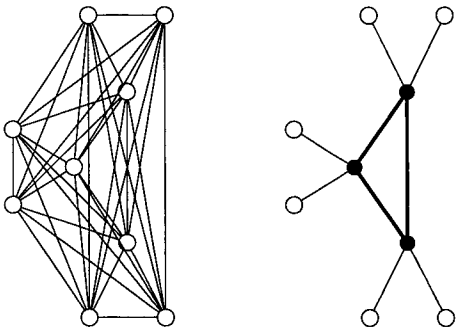


Fig. 5.7
Hub-and-spoke networks
The left hand part shows a fully meshed network. The right hand part represents the same set of nodes including three hubs and an extremely reduced number of connections.

A characteristic feature of such networks is the bundling of flows on the interhub links. Contrasted with what one might expect, these links do not necessarily suffer from congestion if increasing returns to scale prevail. As an example, carriers can use larger aircraft and reduce passenger-mile costs. These *scale economies*, provided they exist, are one of the primary advantages to hubbing. Other advantages refer to reduced maintenance cost due to a smaller number of links and to the sharing of indivisible network facilities. Bailey et al. (1986, Chap. 4) report indeed that many US airlines are focusing on hub-and-spoke operations; since the deregulation act of 1978 they are concentrating their operations at a few large hubs.

[37] On the hub location of several U.S. express package delivery firms see O'Kelly (1986). Most of them have one major hub (typically in or near Ohio) and a supplementary mini hub.

Starting with the latter aspect of joint utilization of indivisible network components, Anderson, Claus (1976) provide a simple example for the idea of sharing fixed costs which at the same time confirms the close-down or continue problem in Remark 5.5. According to Figure 5.8 the undirected network consists of three nodes with a demand for trips given by $(x_{13}^D, x_{21}^D, x_{23}^D) = (1, 1, 1)$. The variable

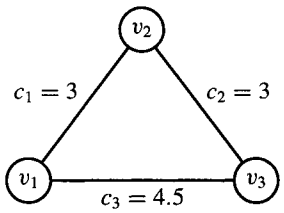


Fig. 5.8
Cost sharing

costs per trip are $(c_1, c_2, c_3) = (3, 3, 4.5)$ and the corresponding fixed link cost follow from $(c_1^f, c_2^f, c_3^f) = (6, 6, 9)$ provided the respective link is in use. Due to fixed cost degression it is advantageous here not to install the third link ($c_3 + c_3^f > c_1 + c_2 + c_1^f/2 + c_2^f/2$) although it is the "shortest" connection between v_1 and v_3. In this case v_2 would play the part of a hub. The situation changes when the demand x_{13}^D jumps from 1 to 2. Now x_{13}^D should be routed along link 3. Of course, this routing may be suboptimal when the technology on some link with heavy traffic can be upgraded so that the respective link costs can be reduced sufficiently.

Observe that the existing nodes of a transport network are not necessarily the only candidates to be upgraded to a hub. Figure 5.1, p. 229, indicates how to improve the network by an additional Steiner point which is for instance centered between v_1, v_2, and v_3. This is a particular problem of location theory where the interconnection with other networks must also be taken into account, cf., e.g., airports that must be accessible by train or passenger car.

Disregarding Steiner points, the coverage problem is to upgrade a subset of all feasible nodes to hubs[38] and then to assign each node to one hub; see Figure 5.7. Thus the set of hubs constitutes a partition of all network nodes $\mathcal{V}$, the so-called *coverage*. An optimal coverage corresponds to a set of hubs such that the transport cost is minimized. Let us assume that all hubs are completely meshed and that travel costs are proportional to the distance between nodes, then it is always suboptimal to pass more than two hubs. In other words, all relevant paths in the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ connecting any OD-pair consist either of one, two or three links. In the latter case the flow starts at v_s passes the hub assigned to it, then uses one interhub link, and enters v_t via the hub assigned to this destination. All other routes with one or two links represent special cases. Alternative routes in particular are ruled out by the network topology. The p-hub median problem as defined in Campbell (1994) goes back at least to Hakimi (1965): it seeks p out of the $|\mathcal{V}|$ nodes such that total transport cost is minimized.[39] The advantage of using interhub connections at reduced travel costs

[38] The coverage problem may be restricted to a maximum number of nodes that can be connected to the same hub or by a maximum number of feasible hubs.

[39] The determination of the optimal number of hubs would be the next step to be solved.

is represented by an exogenous interhub "discount factor" α. Thus all transport costs per unit from origin v_s to destination v_t have the form

$$\tilde{c}_{(s,t)}^{(k,m)} = c_{sk} + \alpha c_{km} + c_{mt},$$

where v_k and v_m are hubs assigned to v_s and v_t, respectively. All special cases with $s = k$, $k = m$, or $m = t$ can be captured by letting $c_{vv} = 0$. Let $\mathcal{V}_p \subset \mathcal{V}$ be a set of p points to be designated as hubs. Then

$$\tilde{c}_{(s,t)}(\mathcal{V}_p) := \min\left\{\tilde{c}_{(s,t)}^{(1,1)}, \tilde{c}_{(s,t)}^{(1,2)}, \ldots, \tilde{c}_{(s,t)}^{(p,p)}\right\}$$

yields the cost minimal assignment of nodes (v_s, v_t) to hubs given by the set $\mathcal{V}_p$. As long as nodes can be assigned to more than one hub, a set of points $\hat{\mathcal{V}}_p$ is a *p-hub median* of $\mathcal{G}$ if for every $\mathcal{V}_p \subset \mathcal{V}$

$$\sum_{s\in\mathcal{V}}\sum_{t\in\mathcal{V}} \tilde{c}_{(s,t)}(\hat{\mathcal{V}}_p)\, x_{(s,t)} \leqq \sum_{s\in\mathcal{V}}\sum_{t\in\mathcal{V}} \tilde{c}_{(s,t)}(\mathcal{V}_p)\, x_{(s,t)},$$

where $x_{(s,t)}$ is the given demand traffic with respect to the OD-pair (v_s, v_t). The necessary modifications to obtain a *p-median problem*,[40] where each node associates to exactly one hub can be found in Campbell (1994). This integer programming formulation goes beyond the scope of this book, however, we can find that it is technically relatively easy to implement fixed link cost or fixed cost of establishing a facility at a hub.[41] The model then consists of the three travel cost categories determining $\tilde{c}_{(s,t)}^{(k,m)}$ and it includes arc construction cost, hub construction cost, and possibly hub congestion cost. A more detailed analysis with regard to piecewise linear interhub link cost functions is proposed in O'Kelly, Bryan (1998). In order to capture economies of scale, these cost functions are decreasing in the respective link flow. The result of hub location with regard to the U.S. airline industry is quite intuitive. As an illustrative example, O'Kelly, Bryan (1998) choose a set of 20 nodes from the set of 25 largest cities (see Figure 5.9) and find especially for $p = 3$ that Chicago, New York City and Dallas are the most valuable locations for hubs.

A different approach for the hub network design problem starts with the traffic assignment model on page 139.

$$\min\left\{\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\varphi}) \mid \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x},\ \boldsymbol{\Delta}\,\mathbf{h} = \boldsymbol{\varphi},\ \boldsymbol{\varphi} \geqq \mathbf{0}\right\} \tag{5.42}$$

If links operate with increasing returns to scale, the link cost per unit decrease with the flow φ_e. This observation induces Horner, O'Kelly (2001) to introduce economies of scale by virtue of a time-congestion relationship that has been introduced as BPR function in (3.33a), p. 113: they just reverse the sign of β_0. As it is hard to justify

[40] The p-hub median problem assigns each OD-pair to the cost minimal pair of hubs, while the p-median problem allocates each node to one hub.

[41] On a slightly modified version of the model with tighter restrictions see Skorin-Karpov et al. (1996) who solve the model numerically for a large data set from the Civil Aeronautics Board which was regulating the U.S. airline industry prior to the deregulation act 1978.

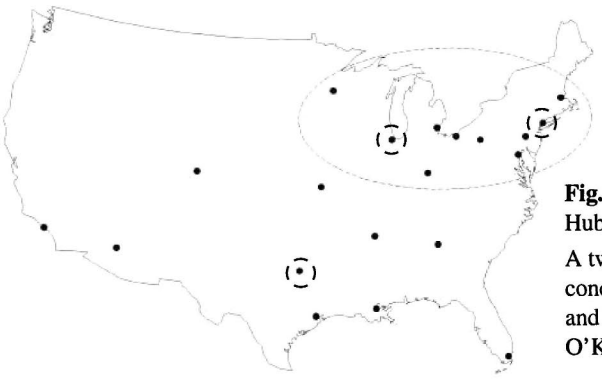


Fig. 5.9
Hub location
A twenty node network showing concentration of cities in the East and Upper Midwest. Source O'Kelly, Bryan (1998).

why a flow exceeding the capacity $\varphi_e > \kappa_e$ should reduce transport cost, it is more adequate to model increasing returns to scale, for instance, by a production function $\varphi_e = {\kappa_e}^{\alpha}$ which is homogeneous of degree $\alpha > 1$. Hence, total link costs

$$f_e(\varphi_e) = \min\left\{q_e\kappa_e \mid \varphi_e \leqq {\kappa_e}^{\alpha}\right\} = q_e{\varphi_e}^{1/\alpha}$$

are homogeneous of degree $1/\alpha < 1$ and transport costs per flow unit diminish.

$$f_e(\varphi_e) = c_e(\varphi_e)\varphi_e \implies c_e(\varphi_e) = q_e{\varphi_e}^{\frac{1-\alpha}{\alpha}} \quad \text{and} \quad \mathfrak{c}_e(\varphi_e) < 0$$

The parameter α could be link specific and determines an endogenous interhub "discount factor" rather than an exogenous discount factor as in the previous approach.

At first sight the "design problem" (5.42) seems to be independent of capacities κ which enter the formulation via the assumed production functions. Horner, O'Kelly (2001) refer to this problem as an evolutionary approach to network design, for the location of hubs is endogenously determined by the minimization of transport cost. Obviously, we cannot expect that this model generates a pure hub-and-spoke network, where all hubs are completely meshed and each node is assigned to exactly one hub. Nevertheless, the numerical results of Horner and O'Kelly confirm what has been suspected earlier: cost minimal transport networks tend to show sparse, even tree-like, network topologies with enormous capacities on a few number of links. Another "evolutionary" approach is proposed by Drezner, Drezner (2001), where passengers choose the hubs that suit them best. The authors apply the gravity rule to that airline hub selection problem with the objective of minimizing the total miles traveled by passengers. They assume that the probability that a customer selects a hub is proportional to the hub's "attractiveness" and inversely proportional to a power of the distance to that hub. This rule is the basis of the gravity model, see p. 184. The numerical result for a three hub configuration as compared to Figure 5.9 recommends to choose one pair of towns in the East and Upper Midwest (e.g., New York City and Chicago) and to add Los Angeles instead of Dallas.

5.8 Remark (Set covering problem) The location of interacting hub facilities is closely related to the so-called *set covering problem*. This is to find the minimum number of facilities, located on nodes, which cover all of the demand. The *maximum*

covering location problem is to find a set of p nodes maximizing the total demand that is covered. As an example one may think of establishing a new network from scratch which starts typically with the coverage of large cities.
Sometimes an arriving demand must be rejected because all facilities are engaged serving other demand. This probability of being blocked can be reduced by multiple coverage (which is a question of network dimensioning), where many facilities are capable of serving the demand. Daskin (1983) proposes a corresponding location problem that maximizes the expected coverage. In communications systems the probability of being blocked is expressed by Erlang's rejection function as will be explained subsequently. □

The benefits from hubbing must be confronted with several disadvantages that limit the applicability of hub-and-spoke networks.

Network topology As has already be mentioned, the traffic is concentrated to a few broad lines, which makes the network vulnerable against failing network components.

Reduced level of service Total network costs are often minimized by forcing some individual interacting pairs to their non-least-cost path. From a customer point of view, passengers prefer fewer stops on the way to their destination. Hence, competitors of a hub-and-spoke network carrier can take advantage by offering certain direct point-to-point connections.

Nodal congestion Hub-and-spoke networks concentrate traffic not only to certain links but also to a small number of hubs. As a consequence, Beckmann (1999, p. 157) calls attention to capacitated nodes, where each hub, e.g., a central airport, is able to serve a maximum number of flow units per period; think, for instance, of the increasing effort of baggage transshipment. The cost minimization problem (5.42) must then be supplemented by

$$\sum_j \varphi_{ij} + \sum_t x_{jt}^D \leqq b_j \qquad \forall\, j,$$

where b_j is the capacity of the hub located at node v_j. In other words, the inflow of a node plus the generated traffic at that node is limited to an upper bound b_j.

Economies of scale Increasing returns to scale cannot be expected to prevail globally. Instead congestion will predominate the ability of networks to ship extremely large flow volumes across the same link.

Irrespective of these objections, one can think of many enhancements to the design of pure hub-and-spoke networks. For example, many package express delivery networks make use of major and mini hub networks. Such hierarchic structures are important features especially for telecommunications networks, some details of which are introduced in the next section. Less directly, hierarchies are evident in the way that some nodes and hubs become the focal points for agglomerating economic activity; cf. O'Kelly (1998).

5.2.3.2 Hierarchic Networks

The idea of hub-and-spoke network design is to upgrade p out of n existing nodes to hubs in order to provide the bundled traffic at reduced total transport cost. Similarly one can think of the location of p new facilities among n existing facilities with a given demand. Elzinga et al. (1976) refer to this question as *multifacility location problem*[42], where they consider all places in space as feasible. As an example the authors consider the location of medical facilities in a battle zone. A more peaceable example can be found in Bertsekas, Gallager (1992, p. 448) who focus on the *local access network design problem*. This problem deals with data networks (with travel cost $\mathbf{c} = \mathbf{0}$) and aims at the design of a network that connects a collection of terminals (nodes) with known demand to a network of "hubs". This problem is often addressed in the context of a *hierarchic strategy*, whereby groups of terminals are connected to various types of *concentrators*, which are in turn connected to higher levels of concentrators, and so on; cf. Figure 5.10 in comparison to the "flat" network of Figure 5.7. If we allow for cross connections of nodes on the same level, such a

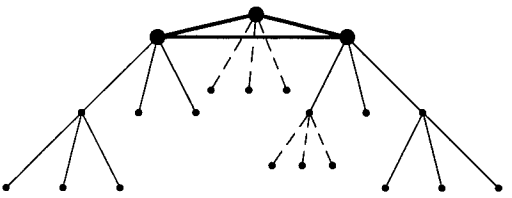


Fig. 5.10
Hierarchic network
All demand nodes are connected to concentrators at the second level which in turn are joined with the bold backbone network at the third level.

hierarchic system corresponds exactly to networks of streets, roads, and highways.

In the *concentrator* or *warehouse location problem* there are n sources ($i = 1,\ldots, n$) located at known geographical points. The task is to connect these sources to p out of m potential concentrator sites ($j = 1,\ldots, m$) that are themselves connected in a superior communications network. The corresponding design problem can be stated by making use of the following two classes of binary variables.

$$\psi_{ij} = \begin{cases} 1 & \text{if link } (i, j) \text{ is constructed} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and} \quad \psi_j^{\circ} = \begin{cases} 1 & \text{if a concentrator is located at } j \\ 0 & \text{otherwise} \end{cases}$$

[42] The usual facility location problem, which is also referred to as (Alfred) Weber problem, is to find a location that minimizes a weighted sum of distances to each of several locations. Originally considered by Fermat in the 17th century, there are now many variations of this problem. One variation is the location of an obnoxious facility such as a garbage dump. In that case the objective is to maximize total (weighted) distance from the given points, where there are constraints, such as cost, that require the location to be within some distance of the points.

Regarding data networks it is common practice to ignore link specific transport cost, i.e., $\mathbf{c} = \mathbf{0}$. Suppose that the construction of some link $e = (i, j)$ incurs fixed cost c_{ij}^f and let c_j^f denote fixed cost of establishing a facility at node j. The costs of interconnecting concentrators are neglected by Bertsekas, Gallager (1992) for convenience, but can be added easily; cf. Lee et al. (1996). The design problem now takes the form

$$\begin{aligned}
\min \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}^f \psi_{ij} + \sum_{j=1}^{m} c_j^f \psi_j^\circ \\
\text{s.t.} \quad & \sum_{j=1}^{m} \psi_{ij} = 1 \quad \forall i, \\
& \sum_{j=1}^{m} \psi_j^\circ = p, \\
& \sum_{i=1}^{n} \psi_{ij} - \kappa_j \psi_j^\circ \leqq 0 \quad \forall j, \\
& \psi_{ij} \in \{0, 1\}, \\
& \psi_j^\circ \in \{0, 1\}
\end{aligned}$$

The first restriction requires that each node is connected to exactly one concentrator, where exactly p out of m concentrators are to be chosen by the second constraint. Each concentrator j is able to handle a maximum number of sources, say κ_j. This problem is a difficult combinatorial problem that usually can be solved only approximately. On a Lagrangean relaxation approach see Lee et al. (1996) who report also numerical results with respect to digital data service networks.

5.2.3.3 Data Networks with Queuing

(a) Elements of Queuing Theory

Up until now it was almost always assumed that transport cost are of minor importance to telecommunications or data networks. In the light of queuing delays that simplification cannot be sustained. In what follows we introduce time needs in a queuing framework that correspond to travel costs. Afterwards the capacity assignment problem will be revisited.

The capacity assignment problem may be seen as a special case of the budget design model. It is supposed that the initial capacities are zero, $\boldsymbol{\kappa}_0 = \mathbf{0}$, so that the investment $\boldsymbol{\delta}$ describes the capacities of the network, $\boldsymbol{\kappa} = \boldsymbol{\delta}$. Given the budget $\bar{I}$ and a unique price q per unit of investment which is valid for all links, the budget corresponds to the total capacity available for distribution among the communication lines, i.e., $\sum_{e \in \mathcal{E}} \kappa_e \leqq \bar{I}/q$. Nevertheless, when links are of different length it is consequent to maintain the budget constraint $\mathbf{q}^\mathsf{T}\boldsymbol{\kappa} \leqq \bar{I}$. Apart from additional stochastic elements, it is assumed that a given routing pattern[43] determines feasible

[43] Routing patterns have been introduced in Section 3.2.5.4 on p. 163.

link loads $\bar{\varphi}$ for each given demand $\mathbf{x}$. This releases the budget design problem (5.24), p. 253, from problems of traffic assignment so that it can be rewritten as

$$\min_{\kappa} \left\{ f(\bar{\varphi}, \kappa) \mid \mathbf{q}^{\top}\kappa \leqq \bar{I},\ \kappa \geqq \mathbf{0} \right\}.$$

One important result will be that it is *not optimal* to choose $\bar{\varphi} = \kappa$ what one might expect at first sight.

Before network design problems of this type can be analyzed, some basic features of queuing models must be introduced. They are collected from Bertsekas, Gallager (1992).

Arrival process of customers Usually it is assumed that the interarrival times are independent and have a common distribution. In many practical situations customers arrive according to a Poisson stream (i.e., exponential interarrival times). Customers may arrive one by one, in batches, or in packets. An example of batch arrivals is the customs office at the border where travel documents of bus passengers have to be checked.

- In *open networks* customers arrive and depart from the network.
- In *closed networks* a fixed number of customers circulate within the network.

Behavior of customers Customers may be patient and willing to wait (for a long time). Or customers may be impatient and leave after a while. For example, in call centers customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while. Sharkey (1995) presents a model where arriving customers are assumed to bark if the observed queue length – and thus the expected waiting costs – exceeds a critical value.

Service times Usually we assume that the service times are independent and identically distributed, and that they are independent of the interarrival times. For example, the service times can be deterministic or exponentially distributed. It can also occur that service times depend on queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large.

Service discipline Customers can be served one by one or in batches. There are many possibilities for the order in which they enter service. We mention: first in first out (FIFO) or first come first served (FCFS), i.e., in order of arrival; service in random order (SIRO); last come first served (LCFS), e.g., in a computer stack or a shunt buffer in a production line; priorities, e.g., rush orders first, shortest processing time first (prioritization is often used to improve efficiency); processor sharing, i.e., in computers that equally divide their processing power over all jobs in the system; preemption, i.e., the job in service is sent back to the buffer if jobs with higher priority arrive.

Service capacity There may be a single server or a group of servers helping the customers.

Waiting room There can be limitations with respect to the number of customers in the system. For example, in a data communication network only finitely many cells can be buffered in a switch. The determination of good buffer sizes is an important issue in the design of these networks.

Kendall introduced a shorthand notation to characterize a range of these queuing models. It is a three-part code a/b/c. The first letter specifies the interarrival time distribution and the second one the service time distribution. For example, for a general distribution the letter G is used, M for the exponential distribution (M stands for memoryless or Markov process), D for deterministic times, and E for the Erlang distribution. The third and last letter specifies the number of servers. Some examples are $M/M/1$, $M/M/m$, $M/G/1$, $G/M/1$ and $M/D/1$. The notation can be extended with an extra letter to cover other queuing models. A forth letter specifies the upper bound on the length of a queue, the queuing capacity. For example, a system with exponential interarrival and service times, one server and a waiting room only for N customers (including the one in service) is abbreviated by the four letter code $M/M/1/N$. A fifth letter indicates the number of objects to be distinguished (population size). The population size and the queuing capacity default to infinity. The sixth position gives the way of selecting objects from the queue (queue discipline), that is FIFO, LIFO, SIRO, etc. The queue discipline is FIFO by default.

(b) Performance Measures

The main metric that Kleinrock (2002) uses for the performance of a computer network is the average time it takes for messages to move across the net. The network consists of sites (nodes) and communication lines (edges) and a switch with infinite queuing space is located at each site to decide how messages are to be processed. It is clear that queues develop at each switch as messages arrive and wait to be served; see Figure 5.11 representing one link e ending at the switch $v_j = \omega(e)$.

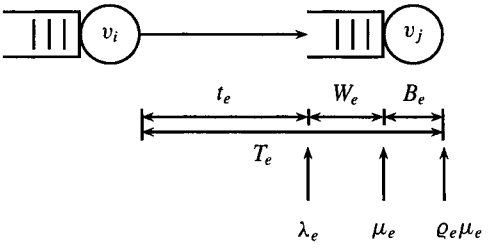


Fig. 5.11
Delays in a waiting system
The delay T_e includes the waiting time W_e in the queue, the service time B_e, and the propagation delay t_e.

It is convenient for the next few paragraphs to drop the subscript v from notation. They explain the functionality of waiting systems associated with each link $e \in \mathcal{E}$.

For the sake of simplicity, the link delay is assumed to consist of two components, namely, the *queuing delay* or waiting time W and the *processing delay* B, which are usually referred to as *sojourn time* or *response time* T.

$$T = W + B \ (+t) \qquad [T] = \text{s/packet}$$

The queuing delay or lifetime of a job is the time an arriving message has to wait in a queue to be transmitted. The processing delay is the service time needed by

the switch to serve the message, that is the time to receive and to send the message correctly. The third delay component – namely, the *propagation delay* t, is omitted here although it can be substantial. This delay refers to the time a bit, and especially the last bit of a message, needs for the physical distance between the transmitter and the receiver. In traffic theory the propagation delay may be the main part of travel time, but here it is a constant of minor importance.[44] The delay measures T, W, and B are continuous random variables, the mean values of which satisfy $E(T) = E(W) + E(B)$.

The utilization of network resources depends also on the distribution of the number of customers K (e.g., orders, jobs, messages, etc.) in the waiting system (including the one or those in service) or on the number of jobs in the queue with Q denoting the length of the queue. While delay variables describe periods of time, the two variables K and Q are discrete random variables regarding one point of time.

The system state at time t is the result of simultaneous arrivals, services and departures in the past. Given an $M/M/1$ queue[45] the stochastic arrival, birth, or input process $\{X(t) | \; t \geqq 0\}$, as well as the departure, death or output process $\{Z(t) | \; t \geqq 0\}$ follow Poisson processes. Here the discrete random variable $X(t)$ denotes the number of arrivals in the period $[0, t]$ and has a probability mass function (pmf)

$$P\{X(t) = x\} = \frac{(\lambda t)^x \, \mathrm{e}^{-\lambda t}}{x!} \quad \text{with} \quad E(X(t)) = \lambda \, t. \tag{5.43}$$

Based on this arrival process the interarrival time A is a continuous random variable with exponential distribution following from

$$P\{A > t\} = P\{X(t) = 0\} = \mathrm{e}^{-\lambda t} \quad \text{or} \quad P\{A \leqq t\} = 1 - \mathrm{e}^{-\lambda t}$$

with an average period between two arrivals or mean arrival time of

$$E(A) = 1/\lambda. \qquad [\lambda] = \text{packets/s}$$

This determines the first M of the $\underline{M}/M/1$ queue, where the decisive parameter λ is referred to as *arrival rate*. Provided the queue is not empty, we may use a service process $\{Y(t) | \; t \geqq 0\}$ that has the same probability mass function as the arrival process but with the *service rate* μ as parameter rather than λ so that particularly $E(Y(t)) = \mu t$; the random variable $Y(t)$ denotes here the number of units accepted by a free server in period $[0, t]$. Consequently, the service time B with

$$P\{B > t\} = P\{Y(t) = 0\} = \mathrm{e}^{-\mu t} \quad \text{or} \quad P\{B \leqq t\} = 1 - \mathrm{e}^{-\mu t}$$

is also exponentially distributed with the mean service time

$$E(B) = 1/\mu, \qquad [\mu] = \text{packets/s}$$

explaining the second M in the $M/\underline{M}/1$ queue.

[44] In fiber optic communication networks the propagation delay can be neglected and storage may be interpreted as transportation through time.

[45] On further details of queuing systems see, e.g., Neumann, Morlock (1993, Sec. 5.3).

5.9 Remark (Memoryless random variables) Exponentially distributed random variables and particularly the interarrival time A as well as the service B have no memory which is referred to as *Markov property* (see Rohatgi (1976, p. 209), Theorem 8). With special regard to the interarrival time A, this property means

$$P\{A > t + \Delta t | \, A > t\} = P\{A > \Delta t\} = e^{-\lambda \Delta t}$$

or, equivalently,

$$P\{A \leqq t + \Delta t | \, A > t\} = P\{A \leqq \Delta t\} = 1 - e^{-\lambda \Delta t}.$$

The probability that a customer arrives in the period $[t, t + \Delta t]$ is independent of the point of time t. This property is characteristic for the busy hour where traffic does not change its "typical" behavior; but it cannot be applied to the period preceding the busy hour (with an arrival rate λ being a stochastic process itself). The stochastic input process $\{X(t) | \, t \geqq 0\}$ that complies with the memoryless interarrival time A corresponds to a stochastic process whose future development depends only on the state arrived at the given instant and not on its past history. It is called a *Markov process*. On a more precise definition and further properties of Markov processes see Kleinrock (1975). □

The utilization rate, traffic intensity or *occupation rate* ϱ represents an expression of system occupation. It indicates the (dimensionless) fractional part of the time a single server $m = 1$ in the $M/M/1$ queue is working:[46]

$$\varrho := \lambda / \mu.$$

When ten customers per minute arrive on average at the queue ($\lambda = 10/60$) and the server needs three seconds to serve a customer ($1/\mu = 3$), the utilization rate of the server is $\varrho = 50\%$. A more thorough analysis as in Neumann, Morlock (1993) proves for $\varrho < 1$ that the stochastic queuing system reaches a steady state with a "constant" or "typical" number of customers when the arrival rate λ matches the *departure rate* $\varrho\mu$; cf. Figure 5.11. The departure process $\{Z(t) | \, t \geqq 0\}$ has therefore the same probability mass function as the arrival process but with $E(Z(t)) = \varrho\mu\, t$, where the random variable $Z(t)$ indicates the number of customers leaving the server in period $[0, t]$.

To sum up, we know how many customers enter and leave the system on average. Moreover, the average service time is $E(B) = 1/\mu$. The next step is to compute the mean waiting time $E(W)$ which in turn depends on the queue length Q. This may be done by observing the last unit that has entered a *deterministic* queue of length Q. It needs W seconds to leave the queue, provided the queuing discipline is FIFO. If the queue length is constant, then Q units must enter the system during the period T at the rate λ. This reasoning yields $Q = \lambda W$. If the server is also taken into account, one has $K = \lambda(W + B)$. The same result also holds true for stochastic systems (see Kleinrock (1975, Chap. 2)), where *Little's Law*

$$K = \lambda T \tag{5.44}$$

represents the fundamental relationship of queuing theory.

[46] For n servers the occupation rate per server would be $\bar{\varrho} = \lambda / (n\,\mu)$, where $n\,\mu$ is the service rate of n servers.

(c) Steady State of an M/M/1 Queue

A waiting system with an arrival process $\{X(t) | \ t \geqq 0\}$ and a simultaneous service process $\{Y(t) | \ t \geqq 0\}$ may be in a *transient state*, where the number of incoming packets, say *births*, differs from the number of packets served, say *deaths*. This situation often occurs as a short-term, start-up condition in which the system is said to be unstable. Contrasted with unstable flows an *equilibrium* refers to a situation where the state of the system, although still probabilistic, is steady. This means, a typical system state with k customers at time t remains unchanged if the switch dispatches on average the same number of customers as arrive in the respective period of time; the number of births equals on average the number of deaths. In other words, the transition probability for a system to go over from a state k to any different state in the next period is zero.

This statement may be explained best by recalling the following probabilities

$$\begin{aligned} P\{X(\Delta t) = 1\} &= \lambda \, \Delta t \, e^{-\lambda \Delta t}, \\ P\{Y(\Delta t) = 1\} &= \mu \, \Delta t \, e^{-\mu \Delta t}, \\ P\{X(\Delta t) = 0\} &= e^{-\lambda \Delta t}, \\ P\{Y(\Delta t) = 0\} &= e^{-\mu \Delta t}. \end{aligned}$$

Letting $\pi_k(t)$ denote the probability of k jobs in the waiting system at time t, we can describe $\pi_k(t + \Delta t)$ by all possible prior states together with the occurrences in the period $[t, t + \Delta t]$. For example, one birth and no death implies $(k - 1) \rightarrow k$, one death and no birth yields $(k + 1) \rightarrow k$, and one birth together with one death (or no birth and no death) results in $k \rightarrow k$. Such a birth-death model is represented by Figure 5.12.

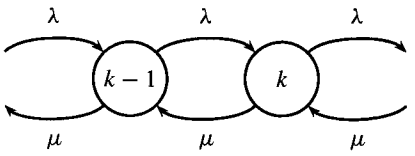


Fig. 5.12
Birth-death model ($M/M/1$ queue)
An equilibrium needs
$\lambda \pi_{k-1} = \mu \pi_k$ for $k = 1, 2, \ldots$.

Hence, when births and deaths are independent and $k \geqq 1$,

$$\begin{aligned} \pi_k(t + \Delta t) = \pi_{k-1}(t) \, \lambda \, \Delta t \, e^{-\lambda \Delta t} e^{-\mu \Delta t} + \pi_k(t) \big(\lambda \, \Delta t \, e^{-\lambda \Delta t} \mu \, \Delta t e^{-\mu \Delta t} \\ + e^{-\lambda \Delta t} e^{-\mu \Delta t} \big) + \pi_{k+1}(t) \mu \Delta t \, e^{-\mu \Delta t} e^{-\lambda \Delta t} + o(\Delta t), \end{aligned}$$

where $o(\Delta t)$ gathers all cases with multiple births and multiple deaths yielding the state k. Rearranging this equation yields

$$\begin{aligned} \frac{\pi_k(t + \Delta t) - \pi_k(t)}{\Delta t} = \Big[\lambda \pi_{k-1}(t) + \big(\lambda \, \mu \, \Delta t + \frac{1}{\Delta t} \big) \pi_k(t) + \mu \pi_{k+1}(t) \Big] e^{-(\lambda + \mu) \Delta t} \\ - \frac{\pi_k(t)}{\Delta t} + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

For $\Delta t \to 0$ it is natural to assume that multiple events do not occur so that the last fraction goes to zero. At the same time we arrive at the formal derivative of $\pi_k(t)$ for $k \geqq 1$.

$$\dot{\pi}_k = \lambda \pi_{k-1} - (\lambda + \mu)\pi_k + \mu \pi_{k+1}$$

A second differential equation is needed for the case $k = 0$ in which the probability of a death is zero and $\pi_{-1} = 0$. The same steps as before then yield

$$\dot{\pi}_0 = -\lambda \pi_0 + \mu \pi_1.$$

This system of infinitely many differential equations reaches a steady state[47] at $\dot{\pi}_k = 0$ for $k = 0, 1, 2 \ldots$ It can be solved by recursive substitution resulting in

$$\lambda \pi_{k-1} = \mu \pi_k \quad \text{or} \quad \pi_k = (\lambda/\mu)^k \pi_0 = \varrho^k \pi_0 \qquad k = 1, 2, \ldots$$

As total probability must sum to unity, one has $\pi_0 = 1/\sum_{k=0}^{\infty} \varrho^k = 1 - \varrho$ provided the stability condition $0 \leqq \varrho < 1$ holds true. Thus $\pi_k = (1 - \varrho)\varrho^k$ and the average number of customers in the system follows from

$$E(K) = \sum_{k=0}^{\infty} k\pi_k = (1-\varrho) \sum_{k=0}^{\infty} k\varrho^k = (1-\varrho)\left[\sum_{k=0}^{\infty} (k+1)\varrho^k - \sum_{k=0}^{\infty} \varrho^k\right]$$
$$= (1-\varrho)\left[\frac{1}{(1-\varrho)^2} - \frac{1}{1-\varrho}\right] = \frac{\varrho}{1-\varrho}.$$

Using the mean number of customers the mean sojourn time results from Little's law.[48]

$$(5.45) \qquad E(K) = \frac{\varrho}{1-\varrho} = \frac{\lambda}{\mu - \lambda} \quad \text{and} \quad E(T) = \frac{1}{\lambda} E(K) = \frac{1/\mu}{1-\varrho} = \frac{1}{\mu - \lambda}$$

For an occupation rate $\varrho = 0$ the customer spends on average no time in the queue and the service time is $1/\mu$. For the sake of completeness, we mention the corresponding results which refer to the queue excluding the server, that is the queue length Q and the waiting time W.

$$E(Q) = \frac{\varrho^2}{1-\varrho} = \frac{\lambda^2/\mu}{\mu - \lambda} \quad \text{and} \quad E(W) = \frac{E(Q)}{\lambda} = \frac{\varrho^2/\lambda}{1-\varrho} = \frac{\lambda/\mu}{\mu - \lambda} = E(T) - \frac{1}{\mu}$$

These and a lot of more general results can be found in Kleinrock (1975). Here we can find also one of the possibly most important results for dimensioning networks: to avoid that the queue eventually grows to infinity, it is required that

$$0 \leqq \varrho = \lambda/\mu < 1.$$

[47] The steady state or statistical equilibrium requires that the upward transition $k \to k+1$ occurs at the same rate as the downward transition $k+1 \to k$.

[48] One version of an $M/M/1$ queue worth to be mentioned can be found in Kleinrock (1975). If the arrivals tend to get discouraged when more and more people are present in the system, $\lambda_k = \alpha/(k+1)$, then $\varrho = 1 - e^{-\alpha/\mu}$, $E(K) = \alpha/\mu$, $\lambda = \mu\varrho$, $E(T) = \alpha/(\mu^2\varrho)$. De Palma et al. (1983) pursue the same idea by introducing a departure time choice during a peak period, where the arrival rates follow from a random utility approach.

Without going into further details of that stability condition, we note that the mean queue length also explodes when $\lambda/\mu = 1$, except in the $D/D/1$ system, i.e., the system with no randomness at all. In a multiserver system $G/G/m$ we need $\varrho < m$. In any case the subsequent sections confirm that it is strongly recommended due to stochastic traffic demand not to operate the network too close to an occupation rate of unity.

(d) Capacity Assignment

The results of queuing theory obtained so far are now set into a more complex framework of networks of queues. Fortunately, Kleinrock's independence approximation shows under fairly general conditions (see Bertsekas, Gallager (1992, p. 211 ff.)) that the analysis is straight forward. For example, if we ignore the propagation delays t_e, the average message delay is given by

$$\tau = \frac{1}{L}\sum_{e\in\mathcal{E}} \lambda_e \tau_e, \qquad [\tau] = \frac{\mathsf{s}}{\mathsf{packet}}$$

where the *network throughput*[49] L with $[L] = \mathsf{packet/s}$ is the total incoming message rate for the network. The mean delay $\tau_e = E(T_e)$ is modeled as an $M/M/1$ queue by which

$$\tau_e = \frac{1/\mu_e}{1-\varrho_e} = \frac{\alpha}{(1-\varrho_e)\kappa_e} = \frac{\alpha}{\kappa_e - \varphi_e}. \qquad [\tau_e] = \frac{\mathsf{s}}{\mathsf{packet}}$$

The new elements are the parameter α denoting the mean message length in bits valid for the entire network, $[\alpha] = \mathsf{bits/packet}$, and the capacity of the channel κ_e as well as the link flow $\varphi_e = \varrho_e \kappa_e$, both of which are measured in bits per period, $[\kappa_e] = [\varphi_e] = \mathsf{bits/s}$. It may be useful to observe that κ_e/α corresponds to the mean packet service rate $1/\mu_e$ as indicated in Figure 5.11. Accordingly, the mean packet arrival rate is given by $\lambda_e = \varphi_e/\alpha$. The total system delay can now be obtained from summing the delays on all links.

$$\sum_{e\in\mathcal{E}} \lambda_e \tau_e = \sum_{e\in\mathcal{E}} \frac{\varphi_e}{\alpha}\tau_e = \sum_{e\in\mathcal{E}} \frac{\varphi_e}{\kappa_e - \varphi_e}$$

Although this term seems to be dimensionless it should be read in the sense of accumulated delay per hour. From the customers' point of view, the mean path delays $\tilde{\tau} = \mathbf{\Delta}^{\mathsf{T}}\tau$ may be seen as one substantial criterion for the assured quality of service.

5.10 Remark (Road networks) Delay functions may also be applied to road traffic networks as there is strong evidence that trip-makers choose shortest routes with respect to time. Fuel consumption which increases with speed seems to be of minor

[49] The network throughput formerly has been defined as $\varphi^{\#} = \sum_{v\in\mathcal{V}} \varphi_v^+$ where all units are about the same size. The only difference is that L consists of packets of different size. As a matter of fact, L is treated here as a constant.

importance. Such an approach can be found in Mason (1985)[50], where the queues represent bottlenecks in the middle of a road. He refers to a given trip table $\mathbf{x}^D$ with the usual demand feasible set

$$\Phi^D = \left\{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi} = \boldsymbol{\Delta}\,\mathbf{h},\ \boldsymbol{\Lambda}\,\mathbf{h} = \mathbf{x}^D,\ \mathbf{h} \geqq \mathbf{0} \right\}.$$

In data networks the efficient use of network resources is a task left to the carrier. Similarly, *system optimal flows* on road networks require to minimize the overall system delay

$$\min \left\{ \sum_{e \in \mathcal{E}} \frac{\varphi_e}{\kappa_e - \varphi_e} \mid \boldsymbol{\varphi} \in \Phi^D \right\}.$$

This problem has been discussed in Section 3.2.5.1 in the context of a cost minimal traffic assignment. All that must be done is to replace the objective function in (3.55), p. 139. Contrasted with system optimality, the potential function

$$\int_0^{\varphi_e} \tau_e(z)\,\mathrm{d}z = \int_0^{\varphi_e} \frac{\alpha}{\kappa_e - z}\,\mathrm{d}z = \alpha \ln \frac{\kappa_e}{\kappa_e - \varphi_e}$$

is the *integral of the average delay* τ_e so that it might be used to compute *a user optimal traffic assignment* by solving

$$\min \left\{ \sum_{e \in \mathcal{E}} \alpha \ln \frac{\kappa_e}{\kappa_e - \varphi_e} \mid \boldsymbol{\varphi} \in \Phi^D \right\}.$$

With regard to the objective function for a user optimum (3.68), p. 147, the discussion can also be omitted. □

The following optimization problem deals with the choice of capacity for each channel such that the mean delay per link in the network is minimized; cf. Kleinrock (2002). The capacities of the queues are infinite by default. Furthermore, the output of the server is limited to the capacity of the outgoing communication lines. In other words, the capacity κ_e of link e is an upper bound on the link flow φ_e and $\sum_{e:\,\alpha(e)=j} \kappa_e$ is the capacity of the server. In the next step, the mean *arrival rate* $\lambda_e = \varphi_e/\alpha$ and the mean *service rate* $\mu_e = \kappa_e/\alpha$ determine the *occupation rate* ϱ_e of the respective link, i.e., $\varrho_e = \lambda_e/\mu_e$. The mean link load is then given by

$$\varphi_e = \alpha\,\lambda_e = \varrho_e\,\kappa_e.$$

This relationship permits us to give alternative expressions for the average link delay, all of which can be found in the literature on communication theory; cf. also (5.45).

$$\lambda_e \tau_e = \frac{\varphi_e}{\kappa_e - \varphi_e} = \frac{\lambda_e}{\mu_e - \lambda_e} = \frac{\varrho_e}{1 - \varrho_e} \tag{5.46}$$

If needed, this delay can be supplemented with the propagation delay t_e in s/bit so that the total delay per link becomes $\lambda_e \tau_e + t_e \varphi_e$.

[50] The functions differ slightly from Mason (1985) because he mixed up the variables φ_e, λ_e, and κ_e in an inappropriate way. Nevertheless, the qualitative results remain valid.

The problem of minimizing the mean message delay in its most general form is

$$\begin{aligned} \min\ & \text{average delay } \tau \\ \text{s.t.}\ & (1)\ \text{channel capacity assignment} \\ & (2)\ \text{routing procedure} \\ & (3)\ \text{topology} \\ & (4)\ \mathbf{q}^{\mathsf{T}}\boldsymbol{\kappa} \leqq \bar{I}. \end{aligned}$$

According to Blessing (2000), this is a problem for which no known algorithm can find the optimum solution apart from brute force, exhaustive approaches. Thus the challenge is to develop algorithms which run in polynomial time and produce designs as near as possible to the optimum. Since lower bounds on the network design problem are only known for simple cases, the method of choice for selecting good algorithms is to compare their results on identical problems. In this sense let us suppose for the sake of simplification that the decisions what sites and communication lines are to be constructed have been made. Thus, the topology (3) of the network is fixed. Furthermore, the routing procedure (2) is assumed to be given. Each switch "knows" how to dispatch arriving packets so that the given demand for traffic $\bar{\mathbf{x}}$ results in a corresponding link flow pattern $\bar{\boldsymbol{\varphi}}$. As long as these topics are not solved one can think of at least three strategies regarding capacity assignment, cf., e.g., Jones, Steele (1989, Chap. 6).

- In an *equal assignment* all channels are constructed with the same capacity. This is realistic, for instance, when thinking of fiber cables with enormous indivisible capacities.
- The *proportional assignment* takes into consideration that there are light and heavy users located at the sources. The budget is then distributed proportionally in accordance with $\kappa_e/\lambda_e = \bar{I}/L$.
- Indeed, the preceding heuristic strategies may be improved by an *optimal assignment* provided the effort to solve this problem is of reasonable extent.

Given the routing procedure as well as the network topology, the above problem of optimizing the network performance takes the form

$$\min_{\boldsymbol{\kappa}} \left\{ \frac{1}{L} \sum_{e \in \mathcal{E}} \lambda_e \tau_e \;\middle|\; \mathbf{q}^{\mathsf{T}}\boldsymbol{\kappa} \leqq \bar{I},\ \boldsymbol{\kappa} \geqq \mathbf{0} \right\}.$$

In the light of the occupation rates ϱ_e, $e \in \mathcal{E}$, an optimal capacity assignment may be interpreted as an *optimal network utilization*. In the preceding optimization problem, $q_e\kappa_e$ denotes the cost to supply κ_e units of capacity for the e^{th} channel, where the total costs must not exceed the budget $\bar{I}$ available for design; $[q_e] = €/(\text{bits/s})$. As has been mentioned above, in communication theory it is frequently assumed that the fixed budget is equivalent to having a fixed total capacity available for distribution among the communication lines. For this reason the above budget design problem is also referred to as *capacity assignment problem*. This is valid, however, only if the prices q_e are the same for all links.

With regard to the objective function and the variables $\boldsymbol{\kappa}$ it turns out to be useful to eliminate $\lambda_e \tau_e$ in accordance with (5.46). This step results in the Lagrangean function

$$\mathcal{L}(\boldsymbol{\kappa}, \nu) = \frac{1}{L} \sum_{e \in \mathcal{E}} \frac{\bar{\varphi}_e}{\kappa_e - \bar{\varphi}_e} + \nu \left[\mathbf{q}^{\mathsf{T}} \boldsymbol{\kappa} - \bar{I} \right] \tag{5.47}$$

having the following optimum conditions for an interior solution $\hat{\boldsymbol{\kappa}} > \mathbf{0}$.

$$\frac{\partial \mathcal{L}}{\partial \kappa_e} = -\frac{\bar{\varphi}_e}{L} \frac{1}{(\hat{\kappa}_e - \bar{\varphi}_e)^2} + \hat{\nu} q_e = 0 \implies \hat{\kappa}_e = \bar{\varphi}_e + \sqrt{\frac{\bar{\varphi}_e}{q_e} \frac{1}{\hat{\nu} L}}$$

$$\frac{\partial \mathcal{L}}{\partial \nu} = \mathbf{q}^{\mathsf{T}} \hat{\boldsymbol{\kappa}} - \bar{I} = 0$$

Eliminating $\hat{\boldsymbol{\kappa}}$ in the latter condition

$$\bar{I} = \mathbf{q}^{\mathsf{T}} \bar{\boldsymbol{\varphi}} + \sum_{j \in \mathcal{E}} \sqrt{q_j \bar{\varphi}_j} \sqrt{\frac{1}{\hat{\nu} L}}$$

and substituting the last square root into the expression for $\hat{\kappa}_e$ yields

$$\hat{\kappa}_e = \bar{\varphi}_e + \frac{\bar{I} - \mathbf{q}^{\mathsf{T}} \bar{\boldsymbol{\varphi}}}{q_e} \frac{\sqrt{q_e \bar{\varphi}_e}}{\sum_{j \in \mathcal{E}} \sqrt{q_j \bar{\varphi}_j}} \qquad \forall e \in \mathcal{E}.$$

Notice that $\bar{I} > \mathbf{q}^{\mathsf{T}} \bar{\boldsymbol{\varphi}}$ and, therefore, $\hat{\kappa}_e > \bar{\varphi}_e$. *The optimal capacity exceeds the mean link load.* Kleinrock (1975, p. 98) explains this phenomenon in the light of the occupation rate ϱ_e. Taking a look at the link delay (5.46) shows that this function has simply a pole at $\varrho_e = 1$. Recalling $\varrho_e = \lambda_e / \mu_e = \varphi_e / \kappa_e$ indicates that the attempt to run a system e near (but below) capacity results in an extreme penalty. With random flow we get occasional bursts of traffic which temporarily overwhelm the server. This type of behavior is characteristic for almost every system one can encounter. In other words, it is strongly recommended to *avoid situations where the network occupation rate approaches 100%.* Eventually the minimal average network delay results as

$$\hat{\tau} = \frac{1}{L} \sum_{e \in \mathcal{E}} \frac{\bar{\varphi}_e}{\hat{\kappa}_e - \bar{\varphi}_e} = \frac{1}{L} \frac{1}{\bar{I} - \mathbf{q}^{\mathsf{T}} \bar{\boldsymbol{\varphi}}} \left[\sum_{e \in \mathcal{E}} \sqrt{q_e \bar{\varphi}_e} \right]^2 .$$

While (5.47) seeks the minimum delay achievable under the given budget $\bar{I}$, Bertsekas, Gallager (1992, p. 439) discuss the inverse problem of finding the minimum budget to guarantee a maximum acceptable delay $\bar{\tau}$.

$$\min \left\{ \mathbf{q}^{\mathsf{T}} \boldsymbol{\kappa} \,\middle|\, \frac{1}{L} \sum_{e \in \mathcal{E}} \frac{\bar{\varphi}_e}{\kappa_e - \bar{\varphi}_e} \leqq \bar{\tau} \right\} \tag{5.48}$$

In the light of the results derived in Section 5.1.4, it takes no wonder that both optimal capacity assignments coincide when $\bar{I} = \mathbf{q}^{\mathsf{T}} \hat{\boldsymbol{\kappa}}$ and $\hat{\tau} = \bar{\tau}$.

The *design procedure* starts with given routing patterns for instance those that have been derived in Section 3.2.5.4. They determine together with the traffic demand the arrival rates λ on all links or, equivalently, the link flows $\bar{\varphi} = \alpha\lambda$. This flow pattern suffices to assign capacities $\hat{\kappa}$ according to the above rule.

In some instances of the network design problem, the goal of achieving minimum cost is realized by a graph of minimum connectivity such that any single node (or edge) failure will disconnect the network. Since single point failures are common in networks, an acceptable design method must provide enough redundancy to survive single node failures. However, any redundancy increases the cost of the network. To balance these conflicting goals, a minimum of disjoint paths must exist between every OD-pair. In the case of a single point network failure, traffic can be sent along an alternative path, albeit at a lower service level. This is the issue of the next section.

5.2.3.4 Telephone Networks without Queuing

(a) Measuring Teletraffic

Regular telephone networks differ from the previous queuing model insofar as an arriving call will be blocked rather than queued if it finds no free communication line. Recall that circuit-switched networks set up a communication circuit for the call and dedicate that path to the participants in that call. For the duration of the connection, all resources on that circuit are unavailable for other users and new sessions must be rejected. Blocked calls get lost in view of the network provider. Agner Krarup Erlang (1878–1929) was the first person to study the problem of telephone networks and is considered to be the founder of traffic theory. By studying a village telephone exchange he worked out a rejection formula, now known as Erlang's formula, to calculate the fraction of callers attempting to call someone outside the village that must wait – provided they are willing to do so – because all of the lines are in use. The corresponding economic network design problem is to determine the optimal number of lines and the network investment cost so that traffic rejection can be reduced to a predefined level.

The framework may be explained best by Erlang's example of a village with s sources or telephones, all of which are connected to one concentrator that connects the village by n lines to the rest of the world. Each line can have two states, it

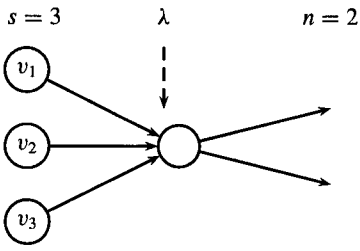


Fig. 5.13
Erlang's example
Three sources in the village ($s = 3$) are connected to the rest of the world via $n = 2$ outgoing channels.

is either free or occupied. Moreover, if there are no calls within the village, the

number of telephone calls equals the number of occupied lines at any instant. This assumption rules out the effect that two telephones, namely, the caller and the called party, are blocked as long as they are joined by an active connection. The possibility that two villagers try to reach the same party in the rest of the world is also ignored. These effects must be taken into account in more general settings, where the network provider is interested in the fraction of arriving calls that can technically be served by the network (free circuit), but are restricted to occupied destinations (called party is blocked).

Using the same notion as before, the input process $\{X(t) | \, t \geqq 0\}$ describes the number of arrivals in the period $[0, t]$ by the random variable $X(t)$.[51] Given a Poisson process (5.43) with the (instantaneous) constant[52] arrival rate λ, one has

$$E(X(t)) = \lambda t \quad \text{or} \quad E(\dot{X}) = \lambda. \qquad [\lambda] = \text{call/s}$$

It is useful to bear in mind two special cases, particularly for the interarrival time $\beta := 1/\lambda$,

$$\begin{aligned} t &= 1 & E(X(1)) &= \lambda, \\ t &= \beta & E(X(\beta)) &= 1. \end{aligned}$$

As long as queuing is impossible, the waiting time W in a queue is always zero. Thus, the mean holding time of a call τ equals the average sojourn time T as well as the average service time B,

$$\tau := E(T) = E(B) = 1/\mu, \qquad [\tau] = \text{s/call}$$

where the service rate μ indicates the number of calls μt accommodated in the period $[0, t]$. If we assume that there are enough lines so that no arriving call must be rejected, then as a matter of principle the output process and the input process are in the steady state, however, all events occur on average at an offset τ. Moreover, if the expected number of starting and terminated calls even out (on average) in each point of time, the expected number of simultaneous calls or *traffic intensity* $E(K)$ at any instant should be constant. It can heuristically be computed by an initial period $[0, \tau]$ such that no started call would be terminated if the holding time is exactly τ.

$$E(K) = E(X(\tau)) = \lambda \tau = \lambda/\mu =: \varrho \qquad [K] = \text{Erl}$$

This equation is in fact a restatement of Little's law (5.44), i.e., $E(K) = \lambda E(T)$, and the former occupation rate ϱ (which was restricted to $0 \leqq \varrho < 1$) now corresponds to the traffic intensity that can exceed unity. It is measured in *Erlang*

[51] In teletraffic theory the amount $X(t)$ is usually referred to as traffic *offered* to the system. Nevertheless, we prefer the economist's point of view by which $X(t)$ denotes the demand for calls because it represents the number of attempts of customers.

[52] In fact, the arrival rate is itself a stochastic process so that λ should be a function of time. Such deliberations go beyond the scope of this book.

(Erl), a dimensionless unit of the average occupancy of a facility during a period of time, usually the busy hour.[53]

The realized *traffic volume* V is the total time of occupied sources during a certain period of observation $[0, t]$. It is measured in Erlang hours (Erl h) or seconds (Erl s),

$$E(V) = E(K)\,t = \varrho\,t = \lambda\,\tau\,t. \qquad [V] = \text{Erl h; Erl s}$$

The last relationship may be interpreted best by the expected arrivals in period $[0, t]$, i.e., $E(X(t)) = \lambda\,t$, times the mean holding time τ which results in the total number of seconds carried by the network in the respective period. For the price p_t per second of a telephone call $p_t\,E(V)$ is the expected revenue of the network provider.

(b) Loss Systems

The fundamental property of loss systems is that incoming calls finding all lines blocked get lost which implies that the calling party makes no reattempts. The *power of a loss system* signifies the maximum load for a prespecified level of losses (guaranteed maximum loss). In order to compute the proportion of blocked traffic we need to know what traffic can be carried by a network with n lines, where the traffic demand results from s sources; cf., e.g., Jung, Warnecke (1998, Sec. 1.7). For the sake of clarity it is useful to distinguish two cases.

Blocking free networks As long as $n \geqq s$, each arriving call finds a free line and will be served.

Networks with blocking If $s > n$ as in Figure 5.13, we can expect that some of the arriving traffic will be rejected by the network.[54]

Apart from stochastic aspects the first case cannot be realized at plausible costs because it would require unreasonable resources for the largest traffic peak that may ever occur. Nevertheless, it is a useful theoretical starting point to analyze the second case of call blocking, where singular peaks are truncated due to cost efficiency.

Blocking free networks A telephone network that accepts all incoming calls requires more lines than sources or more precisely $n \geqq s$. Given a constant arrival rate λ and a mean holding time τ, the mean traffic intensity has been shown to be

$$E(K) = \lambda\tau = \varrho,$$

which corresponds to the mean number of arriving calls in period $[0, \tau]$, i.e., $E(X(\tau)) = E(K)$. Given s sources, the probability b of each source being occupied thus follows from

$$b = \varrho/s. \tag{5.49}$$

[53] The relationship between traffic load and amount of traffic is the same as that for (instantaneous) speed and distance. Given a constant speed v (meters per second), one can compute the distance in meters covered in the period $[0, t]$. Similar to speed, Erlangs describe minutes of active connections during one minute.

[54] The attempt to establish a connection may be unsuccessful due to other reasons not under the responsibility of the network provider, e.g., the called party does not answer. Such aspects go beyond the scope of this analysis.

The probability that k out of s sources are occupied is given by the binomial distribution ($K \sim \mathcal{D}_B(n, b)$) which goes over to a Poisson distribution ($K \sim \mathcal{D}_P(\varrho)$) for a large number of sources $s \to \infty$ and $b \to 0$, so that $sb = \varrho$ remains constant (Rohatgi (1976, p. 200)).

$$P\{K = k\} = \binom{s}{k} b^k (1-b)^{s-k} \xrightarrow[b\to 0]{s\to\infty} \frac{\varrho^k}{k!} e^{-\varrho} = \frac{\varrho^k/k!}{\sum_{v=0}^{\infty} \varrho^v/v!} \tag{5.50}$$

(The last relation makes use of the power series of e^{ϱ}.) Both cases confirm what has been stated above: the expected number of occupied lines is $E(K) = b\,s = \varrho < s \leqq n$. Before going on it is worthwhile to review (5.49) in the form $b = (\lambda/s)\,\tau$. The fraction suggests s sources, all of which independently demand for calls at a constant rate $\lambda_v = \lambda/s$ in accordance with a Poisson process $\{X_v(t) | \, t \geqq 0\}$, $v = 1, \ldots, s$, that is

$$X_v(t) \sim \mathcal{D}_P(\lambda_v t) \quad \text{with} \quad E(X_v(\tau)) = \lambda_v \tau = b. \tag{5.51}$$

Summing up these processes with $\lambda = \sum_{v=1}^{s} \lambda_v$ yields the aggregate demand for calls which follows the common Poisson process

$$X(t) = \sum_{v=1}^{s} X_v(t) \sim \mathcal{D}_P(\lambda t) \quad \text{with} \quad E(X(\tau)) = \lambda \tau = \varrho. \tag{5.52}$$

This result will be modified below in the sense that occupied sources cannot initiate calls.

Networks with blocking Suppose now $n < s$ so that some arriving calls $X(\tau)$ may find all lines blocked, i.e., the *carried load* K satisfies $K = n$ so that the excess demand or *rejected load* R gets lost. In this case, the system states $K = n+1, \ldots, s$ are no longer feasible so that (5.50) must be adjusted. Again two major cases can be distinguished.

First Order Traffic, pure accident traffic, or Poisson traffic assume a constant arrival rate λ of the telephone traffic with a large number of sources in comparison to the number of lines, $s \gg n$, $s \to \infty$. As each line is dedicated to one call, this Erlang model is an $M/M/n/n$ queue in accordance with Kendall's notation.

Second Order Traffic or source traffic supposes also $s > n$, but s is relatively small in comparison to n. Moreover, the arrival rate of telephone traffic depends on the number of free sources. This is the Engset model which corresponds to an $M/M/n/n/s$ queue with s sources being idle or active.

First Order Traffic is based on a telephone network with n lines and an infinite number of sources s. Given a constant arrival rate λ and a mean holding time τ, the probability of a system state with k occupied sources $P\{K = k\}$ is again determined by the traffic intensity $\varrho = \lambda\,\tau$. The computation can be found in Bäßler, Deutsch (1989, p. 240) and results in a discrete Erlang distribution[55] $K \sim \mathcal{D}_{\widetilde{Erl}}(n, \varrho)$ with the probability mass function

[55] Outside of queuing theory, the discrete Erlang distribution is referred to a truncated Poisson distribution.

$$P\{K = k\} = \frac{\varrho^k/k!}{\sum_{v=0}^{n} \varrho^v/v!}.$$

This is the Poisson distribution in (5.50), but the summation is truncated at the last possible state $v = n$. For $n \to \infty$ the probability mass function converges to the Poisson distribution in (5.50) where every arriving call is accepted and we regain $E(K) = \varrho$. For the network designer the most important case is that of n occupied lines so that all incoming calls are blocked. This is done by *Erlang's loss formula* (also referred to as *Erlang B formula*)

$$\pi_n(\varrho) := P\{K = n\} = \frac{\varrho^n/n!}{\sum_{v=0}^{n} \varrho^v/v!} \tag{5.53}$$

which gives the probability for a network with n lines and traffic intensity ϱ that an arriving call gets lost. Thus,

traffic intensity	$X(\tau) = K + R$	$[X(\tau)] = \mathsf{Erl}$
blocked traffic	$R = X(\tau)\,\pi_n(\varrho)$	$[R] = \mathsf{Erl}$
carried traffic	$K = X(\tau)\,\big(1 - \pi_n(\varrho)\big)$	$[K] = \mathsf{Erl}$

Imagine the people of a village want to make calls to the outside world. We don't know when they will want to call or how long their calls will last, but let's suppose that there are on average λ calls starting at each point of time (instantaneous arrival rate), and that the average length of a call is one minute ($\tau = 1$). Erlang wanted to know what fraction of callers would find that all the n lines leading out of the village were already occupied, and so would not be able to make their call until later. Erlang's loss formula gives the answer.

Observe that the addition of further lines reduces the blocking probability $\pi_n(\rho)$. The corresponding investment cost for the $(n+1)^{\text{th}}$ line with a given lifetime is then to be confronted with the respective revenue that follows from the (expected) additional traffic volume ΔV per busy hour[56] ($t = 1$ h), where we have, e.g., 365 busy hours per year.

$$\Delta V = \Delta K\, t = X(\tau)\,\big(\pi_n(\varrho) - \pi_{n+1}(\varrho)\big) \qquad [\Delta V] = \mathsf{Erl\,h}$$

Regarding the decreasing sequence of blocking probabilities it is useful to observe the recursion formula

$$\pi_0(\varrho) = 1, \qquad \pi_1(\varrho) = \frac{\varrho}{1+\varrho}, \qquad \pi_n(\varrho) = \frac{\varrho\,\pi_{n-1}(\varrho)}{n + \varrho\,\pi_{n-1}(\varrho)}.$$

To sum up, the economic criterion for adding new lines is twofold. We can compute, on the one hand, the minimum investment costs to assure a maximum predefined blocking probability in the busy hour. On the other hand, we know that small

[56] Outside the busy hour, most networks show excess capacities so that the construction of new lines would be senseless.

blocking probabilities can only be reached by excessive investments. The optimal blocking probability is then a question whether the present value of the sequence of additional revenues exceeds today's investment cost.

Second Order Traffic refers basically to the same reasoning as before, but the number of sources s is relatively small so that an additional aspect must be taken into account: incoming calls require not only a free line but also a free source to initiate the call. When each occupied line corresponds to one occupied source, the observed aggregate traffic demand $X(\tau)$ must be generated by $s - k$ free sources. Consequently, the parameter $b = \lambda\tau/s$ as defined in (5.49) should be substituted by

$$\tilde{b} = \frac{E(X(\tau))}{s - E(K)} = \frac{\lambda\tau}{s - E(K)}.$$

The question, however, is how to compute λ. With regard to (5.51) we have supposed a source specific arrival rate $\lambda_v = \lambda/s$, the inverse of which gives the interarrival time $\beta_v = 1/\lambda_v$. It is composed of the mean holding time τ and the mean idle time τ_{idle}, i.e., $\beta_v = \tau + \tau_{\text{idle}}$. When occupied sources cannot initiate new calls it is consequent to restrict λ_v to the idle time with the result that (5.52) cannot be applied any longer and

$$\lambda = s\lambda_v = \frac{s}{\tau + \tau_{\text{idle}}} \longrightarrow \lambda = \sum_{k=0}^{n}(s-k)\,P\{K=k\}\,\lambda_v = \frac{s - E(K)}{\tau_{\text{idle}}}.$$

A more rigorous derivation can be found in Cooper (1972, p. 86 ff.). We gain in particular $\tilde{b} = \tau/\tau_{\text{idle}}$ and one can show that the probability of k occupied sources follows from an Engset distribution with the probability mass function

$$P\{K=k\} = \frac{\binom{s}{k}\tilde{b}^k}{\sum_{v=0}^{n}\binom{s}{v}\tilde{b}^v},$$

where the corresponding *Engset's rejection formula* for $P\{K = n\}$ denotes the probability of n occupied lines at one point of time (*time blocking*). This formulation takes into account that all sources whose calls are rejected become immediately free for further attempts (blocked customers cleared). The probability that one arriving call finds a situation where n out of $s - 1$ sources occupy the n existing lines is referred to as *call blocking* and follows from *Engset's loss formula*

$$\pi_n = \frac{\binom{s-1}{n}\tilde{b}^n}{\sum_{v=0}^{n}\binom{s-1}{v}\tilde{b}^v}.$$

The economic results with respect to network design are the same as before, but due to the more accurate blocking probabilities investment decisions may differ. For small numbers of sources s (but $s > n$) the Engset model should be preferred, but the computation of π_n may be difficult because $E(K)$ depends on $\tilde{b}$ which is itself a function of $E(K)$. For $s \to \infty$, however, the Engset distribution converges to the discrete Erlang distribution (see Bäßler, Deutsch (1989, p. 240)) and Erlang's loss formula yields suitable approximations.

5.11 Remark (Erlang C formula) The previous approach operates on a "blocked customer cleared" basis. Consider now the case where blocked customers are delayed in a queue until they are handled by a free server. As an example we refer to the dimensioning server pools where requests for service wait on a FIFO queue until an idle server is available. Given n servers with an unbounded buffer which attend to infinitely many sources ($s \to \infty$), we speak of an $M/M/n$ queue if it has a Poisson input with arrival rate λ and exponentially distributed service times with mean μ. The quotient $\varrho := \lambda/\mu$ denotes the mean carried load because blocking has bee ruled out. Moreover, Cooper (1972) has shown that a steady state requires $\varrho < n$. *Erlang's delay formula* or *Erlang C formula* is used to predict the probability P_W that a call will be delayed

$$P_W = \frac{\dfrac{\varrho^n}{n!}\dfrac{n}{(n-\varrho)}}{\displaystyle\sum_{v=0}^{n-1}\frac{\varrho^v}{v!} + \frac{\varrho^n}{n!}\frac{n}{(n-\varrho)}},$$

and can be used to predict the probability that a call will be delayed more than a certain time. For the $M/M/n$ queue one has particularly for the mean queue length $E(Q)$ and for the mean number of customer $E(K)$ (cf., e.g., Hillier, Lieberman (1995, Chap. 16))

$$E(Q) = P_W\frac{\bar{\varrho}}{1-\bar{\varrho}} \quad \text{and} \quad E(K) = E(Q) + n\bar{\varrho} \quad \text{with} \quad \bar{\varrho} := \frac{\lambda}{n\mu} = n\varrho.$$

On this basis other key queue performance metrics can be calculated, for example, the mean waiting time in the queue $E(W) = E(Q)/\lambda$ by Little's law. □

(c) Quality of Service

Blocking is primarily a problem to obtain the carried traffic load which is the basis for the revenue of the carrier. This viewpoint ignores that blocked customers obtain poor services particularly when they have to wait in long queues or when several reattempts are needed to be served.[57] Two criteria of measuring network accessibility might be

$P\{\text{call blocked}\} < 2\%$	for loss systems,
$P\{\text{waiting time} > t_k\} < 1\%$	for waiting systems.

In this sense, quality of service is basically a problem of network design which pursues the task to deliver services within acceptable time limits; see (5.47) and

[57] Some authors, cf. e.g. Hardy (2001), distinguish explicitly between the two phases of establishing a call and information transfer. In this sense, the *grade of service* (GoS) refers to the ease of a customer to be served. It is contrasted with the *quality of service* (QoS) that defines performance criteria for the pure transfer process.

(5.48) for corresponding optimization problems. Of course, higher levels of service will only be offered to the customers if they are willing to pay for it. The purpose of dimensioning a telecommunications network is in fact to ensure that the expected needs will be met in an economical way both for subscribers and operators.

As an example we refer to a case where the service provider limits the waiting room in order to assure a maximum delay. Consider a queue with buffer size k where there are $k-1$ orders to be processed when a customer request comes in ($M/M/1/k$ queue). Let B_i $(i = 1,\ldots,k)$ be independent, exponentially distributed random variables denoting the processing time of the respective job with $1/\mu$ as mean service time, $B_i \sim \mathcal{D}_{Exp}(1/\mu)$. Then, the conditional waiting time $S_k = \sum_{i=1}^{k} B_i$ for the last customer in a queue of length k (including the service time for himself) is an Erlang distributed[58] random variable, $S_k \sim \mathcal{D}_{Erl}(k, 1/\mu)$, with the following probability density function and cumulative distribution function for $t \geqq 0$:

pdf
$$f_{Erl}(t; k, 1/\mu) = \frac{(\mu t)^{k-1}}{(k-1)!} e^{-\mu t},$$

cdf
$$F_{Erl}(t; k, 1/\mu) = 1 - \sum_{v=0}^{k-1} \frac{(\mu t)^v}{v!} e^{-\mu t}.$$

Let the deadline for the k^{th} job to be completed be t_k. The question is what time should the company promise to deliver the product to the customer so that it meets the deadline with probability 0.99. Therefore, what needs to be determined is a t_k such that

$$P\{S_k \leq t_k\} = F_{Erl}(t_k; k, 1/\mu) = 0.99.$$

The cumulative distribution function $F_{Erl}(t_k; k, 1/\mu)$ shows that decreasing the deadline t_k can be realized by a proportional reduction of the service time $1/\mu$. This can be done, for instance, by increasing the bandwidth of the transmission lines so that more jobs can be served per period (e.g., bits or packets per second, cars per hour, etc.). Holding the service time fixed, the deadline t_k can only be reduced by concentrating on fewer customers, i.e., k must be set to a tighter value.

Alternatively, we can ask for the probability that the k^{th} job grows older than t, i.e., $P\{S_k > t\}$, *and* is finished immediately after t (the job "dies"). In this case $1 - F_{Erl}(t; k, 1/\mu)$ is referred to as survival function and the probability density function f_{Erl} gives an approximation for the probability that the death occurs in the period $[t, t + \Delta t]$.

$$f_{Erl}(t; k, 1/\mu) \cdot \Delta t \approx P\{t < S_k \leq t + \Delta t\}$$

Furthermore, the *hazard (rate) function*, which is no probability density function,

$$h(t) := \lim_{\Delta t \downarrow 0} \frac{P\{t < S_k \leq t + \Delta t\}}{P\{S_k > t\}} = \frac{f_{Erl}(t; k, 1/\mu)}{1 - F_{Erl}(t; k, 1/\mu)} = \frac{(\mu t)^{k-1}/(k-1)!}{\sum_{v=0}^{k-1} (\mu t)^v / v!}$$

[58] Outside from queuing theory, the Erlang distribution with probability density function $f_{Erl}(t; k, 1/\mu)$ is usually referred to as Gamma distribution with probability density function $f_{Gam}(t; k, 1/\mu)$ and the relaxation that $k > 0$ needs not be integer.

describes the instantaneous output rate of the system. The hazard rate gives, to some extent, a more precise characteristic of the dynamic properties of a system. In the light of Erlang's loss formula (5.53) (compare the two formulas for $n = k - 1$) one might think of customers whose access is denied due to an overwhelmed system. While Erlang described that phenomenon with respect to the occupation rate ϱ, the current setting refers to the service volume μt that can be provided in a period $[0, t]$ at a service rate μ. Any excess demand is cut off.

The hazard rate is of particular use in the machinery replacement problem where the equipment has a mean lifetime that may vary with the utilization rate. Failures of network components can be avoided at least to some extent by substituting them in due time. Nevertheless, no network provider can prevent all failures, but fortunately the consequences can be made less severe by an appropriate network design together with an "intelligent" network utilization. This is the issue of the next section.

5.12 Remark (Road traffic flows) Queuing theory can also be applied to road traffic flows to model congestion but with several caveats as stated in Ashton (1966).

- The arrivals which form the input process $\{X(t) | t \geqq 0\}$ for the queue system have been assumed to be independent of any other arrival. This assumption cannot be expected to hold true for road networks with interacting vehicles. Without overtaking, for example, travelers arrive typically in batches at traffic lights where they queue up and await their turn. The same argument implies that equal intervals of time are not equally likely to contain equal numbers of arrivals.
- The propagation delay, i.e., the time need to pass a road, is by far not constant due to variability in speed. Slow network participants such as trucks may have stronger impacts on traffic delays than bottlenecks of the network.
- Furthermore, road junctions as one of the most important causes of congestion differ widely in their working. Crossroads usually include more than one queue rather than collecting all arrivals in one queue. Hence, the service discipline is in general not FIFO across all participants, but priorization is widely used to improve efficiency. Even the flow that arrives from one direction is treated in different ways, for example, when they turn left or right.

Apart from these objections, Ashton (1966, Chap. 7) shows what results of queuing theory pass on to road traffic. He computes, among others, the distribution of waiting times as well as the number of vehicles in the system at any instant. Erlang's loss formula in particular is applied to depict the degree of congestion when the traffic load is ρ and k denotes the number of lanes. □

5.13 Summary The results on detecting minimum investment costs given a flow pattern and its "dual" budget design problem which minimizes transport costs subject to a budget constraint have been recapitulated in Summary 5.7, p. 268. The subsequent expositions focus on special features of transport networks that have far reaching consequences for the nature of networks. Hub-and-spoke networks, and similarly hierarchic networks, proceed with the former result to concentrate traffic volumes on certain links. The reason is that carriers may take advantage of link specific scale economies at least to some extent. As a consequence we observe networks that bundle flows on a few fully developped interhub connections.

Queuing aims at a different form of sharing network resources. It may be seen as one instrument that coordinates individual, non-cooperating network processes. The basic problem is that (stochastic) arrivals of flows claim scarce network resources that can serve only a limited number of jobs at a time. This fact results in delays caused by queues, the length of which vary with the arrivals and departures. For at least one case we are able to detect (probabilistic) steady states of the system and to compute the overall system delay. Minimizing this network performance criterion is similar to the maximum flow problem of a multicommodity network. The corresponding budget design problem shows that it is strongly recommended to avoid situations with occupation rates close to one. In networks without queuing attention is drawn to carried and lost traffic. Loss and rejection formulas give hints for an improved network design. The corresponding problem of finding minimum investment costs in order to realize a given network improvement is an important issue which is left for future research. □

5.3 Survivability Models

5.3.1 Normal Operating State

The preceding approaches of network design have shown two major results. (1) Economies of scale indicate that it may be advantageous to concentrate traffic on inexpensive connections and (2) it is strongly recommended not to operate the network at full load even if it is in working order. In what follows we consider the additional aspect of unreliable network components. Such deficiencies suggest to spread traffic over the network and to lay out surplus capacities in order to make the consequences of temperamental network components less severe.

The task of this section is to work out the relationship between two classes of networks. The first class includes networks the construction cost of which are minimal so that they are usually characterized by low *connectivity* with large capacities on a few links. The other class encloses reliable networks, where all links and nodes are subject to random failure. Here, *survivability* aims at the functionality of the network if one or more network components break down. For low connectivity the trade-off between construction cost and survivability becomes particularly apparent because such a network runs into the risk that many services must be canceled if they are bundled up into one failing component; take, for example, a hub-and-spoke network and eliminate one hub or interhub connection. Thus the question arises as to what does it cost to increase the reliability of networks. Possible measures which reduce the probability of breakdowns are fireprotection, emergency stops on highways or the replacement of worn out equipment in due time; they increase the safety levels for trunk lines and nodes. Nevertheless, not all failures can be precluded at reasonable cost so that a sensible network design may help to reduce expensive safety measures, for instance, by the construction of redundant network elements that serve as substitutes for failing equipments. Aside from assuring network operability these network expansions help also to distribute traffic among alternative paths and

they may serve to improve the overall level of service. As an example we refer to Figure 5.1, p. 229. The left part represents a network with minimal connectivity and the construction of the third link would improve not only reliability but also the comfort by the corresponding shortcut. Contrasted with such a redundant link, the Steiner point in the right hand part of Figure 5.1 may be the optimal solution when taking cost of construction, maintenance, transportation, and reliability into account. In this case safeguard measures are needed to protect the Steiner point from damage. Due to the complexity of network design we present here only a few heuristics to tackle the problem of survivability. Details on further heuristics and computational results for small cases can be found in Grötschel et al. (1995). A rigorous mathematical treatment is presented in Stoer (1992).

The following sections refer to telecommunications networks where transmission costs are of minor importance, i.e., $\mathbf{c} \approx \mathbf{0}$. Although we have seen that delays per link or node may constitute important counterexamples, the only costs taken into account are those associated with the network topology such as the costs of digging trenches and placing fiber cables into service. Besides construction cost, we ignore maintenance cost that accrue to the network provider from every additional communication line. This is merely done for the sake of simplicity as the task is to work out the relationship between cost minimal networks that are characterized by low connectivity and survivability of the network. Let $(\hat{\mathbf{h}}, \hat{\boldsymbol{\delta}})$ be an optimal solution to the minimum investment cost problem (5.17.P), p. 244. Then, $\hat{\mathbf{h}}$ is a demand feasible route flow with respect to the communication demand $\mathbf{x}^D$. Having installed the increments $\hat{\boldsymbol{\delta}}$ such that $\bar{\boldsymbol{\kappa}} = \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}$, a realization $(\hat{\mathbf{h}}, \mathbf{x}^D, \bar{\boldsymbol{\kappa}})$ is called the *normal operating state* which requires in particular that all links and nodes are operational. All variables involved are measured in Mbit/s, where $\mathbf{x}^D$ usually refers to the stochastic but approximately constant demand during a busy hour. Be aware that capacity increments can be used over a long economic life so that minimum investment cost provide a feasible flow for that lifetime. Consequently, (5.17.P) can only be extended to a problem of profit maximization, when revenues and investment cost refer to the same period. One way to overcome this problem of consistency has been discussed in Remark 5.1, p. 233.

The conventional design problem (5.17.P) determines the minimum investment cost to assure that the expanded network is able to accommodate a given demand. In doing so, it abstains unconsciously from reliability issues possibly for the following reason. Redundancy and survivability were not important issues in the design of traditional copper cable networks, because the limited capacity of copper cables resulted in highly diverse routing between nodes in a mesh-like network topology. With the approach of fiber-optics technology and its very high capacities per fiber cable, however, hierarchical routing and bundling of traffic became attractive, resulting in sparse, even tree-like, network topologies with huge capacities on a few number of links. Hence, the failure of a network component may cause severe losses in revenue. Some reasons for failing components are the maintenance of the equipment, human errors such as a cable cut by a digger, or just natural disasters such as an earth quake or fire. One measure to protect networks against failing components is to implement safeguards which take care of the surviving network. The idea is to install

extra capacity units (this includes modifications of the network topology) such that any *failure state* is capable of serving a specified minimum fraction of the demand. The following heuristics calculate the minimum investment costs needed to ensure the survivability of a network. Of course, these costs increase with the minimum standards on survivability. The choice between the models depends on the costs and the ability of the network provider to reroute traffic flows immediately.

5.14 Remark (Real world examples) The need to think about the nation's critical infrastructure has been highlighted by the events of the year 2001, especially the destruction of an essential telephone switch under the World Trade Center, and the fire in the Baltimore tunnel which destroyed two backbone fibers, disrupting Internet traffic. Nevertheless, its destruction hardly ever affected the operativeness of the Internet. In fact, the only reason the Internet was slow at all during September 11, 2001 was in part due to a mass of people rushing to their favorite news sites. Another example is that of KPNQwest or Worldcom, two of the main Internet data carriers of Europe and of the world. If such parts of the Internet (as a network of networks) are shut down, the surviving carriers have to operate the data flows.

Observe that network breakdowns of large extent need further strategies of disaster recovery. For example, the flooding in eastern and southern Germany in 2002 hit no major company harder than Deutsche Bahn, the national railway. The company saw the tracks of some of its main lines near Dresden get washed away, train stations flooded and the collapse of crucial bridges, including one on a high-speed rail line connecting northern and southern Germany. Hartmut Mehdorn, the chief executive officer of Germany's national railway, estimates it will take two and a half years before 400 kilometers of destroyed track, 94 unusable bridges, and more than 200 damaged train stations can be put back into full service. Deutsche Bahn issued a preliminary estimate of over one billion Euro in flood damages and suffered revenue losses. Another much smaller damage results from idle trains and staff which cannot be utilized at least for some time. Similar to snow removal plans, Deutsche Bahn now has to decide what lines are to be reconstructed first. New priorities, however, can lead to delays for other investment projects. □

In order to specify the minimum requirements of routing in the normal operating state as well as some failure state, Alevras et al. (1997) assign three parameters to the communication demand $x^D_{(s,t)}$ between the two nodes v_s and v_t. The *reservation parameter* $\sigma_{(s,t)}$ is a lower bound on the fraction of the demand that must be served when a single node other than the source v_s or the destination v_t (site failure) or a single edge (line failure) is out of service. The *diversification parameter* $\theta_{(s,t)}$ determines an upper bound on the fraction of the demand that can be routed through a single node or edge in case all network components are operating (normal operating state). This spreads traffic across alternative paths. As an immediate consequence of reservation and diversification, the network requires parallel routes to ensure the connectivity of OD-pairs in cases of abnormal network states, where Stoer (1992) distinguishes explicitly between node and edge survivability conditions. Finally, a *path-length parameter* indicates an upper bound on the number of edges or *hops* contained in a path that routes (part of) the demand in the normal operating state. This procedure reduces the probability of a path failure and the qualitative losses

per network component; the latter means that excessive cumulative work rendered by the network along long paths is not lost short before finishing. Bell, Iida (1997) refer to the same aspect as travel time reliability which is the probability that traffic can reach its destination within a given time. Path-length restrictions, however, are ignored here for the sake of simplicity.

5.3.2 Failure States of the Network

Regarding traffic assignment survivability can be introduced in two steps: in the normal operating state with constraints imposed on routing, and in the case of a single node or edge failure through (partial) rerouting of surviving demands. Of course, the immunity of a network can be defined on broader sets of failures.

Diversification This model imposes restrictions on the routing in the normal operating state. The traffic assignment is diversified on several routes for every demand such that the failure of a network component does not affect more than a specified percentage of the demand. This approach manages failure states *without rerouting*.

Reservation Here are no restrictions on the routing in the normal operating state. The approach does a *complete rerouting* even of those parts of the demand which are not immediately affected by the failure. The idea is to build spare capacities which can be used in the case of the failure of one link or node.

Path restoration This model has also no restrictions on the routing in the normal operating state. It maintains the unaffected route flows of the normal operating state and *partially reroutes* the (end-to-end) demand affected by the failure.

Link restoration This heuristic reroutes merely the link flow of the failing edge.

Before these heuristics can be discussed with respect to their implications on minimum investment cost, we need some technical details to capture the properties of failure states. Let $\mathcal{G} = (\mathcal{V}_0, \mathcal{E})$ be an undirected graph of existing nodes $\mathcal{V}_0$ and potential edges $\mathcal{E}$. Each edge $e \in \mathcal{E}$ is represented by an existing or non-existing fiber cable having identical capacities, where parallel links are explicitly taken into account. The cost q_e of establishing the connection e is fixed and depends merely on the length of the cable. As a preliminary approach, an optimal network design asks for the minimal cost of edge construction such that the removal of a maximum number of links or nodes preserves a connected network, i.e., every OD-pair is joined by some path. As a first refinement observe that the destruction of one node implies also the removal of all edges incident to it; the converse, however, is not true. The second refinement takes link capacities into account. Suppose that an OD-pair (v_s, v_t) is connected by a route or, more precisely, chain $\rho \in \mathcal{R}(s, t)$. The maximum flow along that route is bounded by the capacities of the links involved. Thus, for sufficiently large demand $x^D_{(s,t)}$, we need more than one independent route, where the term "independent" underscores different aspects. (1) Two routes that include the same link must share its capacity. (2) A failing link suspends all route flows that use this link: we are, therefore, interested in all *edge-disjoint paths* in the set $\mathcal{R}(s, t)$ which have no edge in common. (3) The same holds true for failing nodes, except the

source v_s and the sink v_t. *Node disjoint paths* are those chains in $\mathcal{R}(s, t)$ that have no intermediate node in common.

Let $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ be a realized subgraph of $\mathcal{G}$. A *failure state* $j = 1, 2, \ldots$ can be characterized by a non-empty set of removed links and nodes resulting in the *surviving network* (or reduced graph) $\mathcal{G}_j = (\mathcal{V}_j, \mathcal{E}_j)$ with the surviving set of paths $\mathcal{R}_j(s, t)$ where $v_s, v_t \in \mathcal{V}_j$. In this sense, the demand $x^D_{(s,t)}$ "dies" if the source v_s or the sink v_t is removed from the network. The task of designing survivable networks should be clear now: construct a network at minimum cost such that any failure of a maximum number of elements results in a reduced network that must be able to accommodate the surviving demand or at least a prespecified part of it. Figure 5.14 depicts a simple example to explain failures of network components. The OD-pair is connected by three link-disjoint paths but only two node-disjoint paths. If one link fails, at least two link-disjoint paths survive. Two link failures cannot disconnect the OD-pair. In the worst case three link failures disrupt the demand $x^D_{(s,t)}$. A demolition of the intermediate node removes at once the four incident links, one path survives, however. The demand breaks down if we close v_s or v_t. If all links have the same

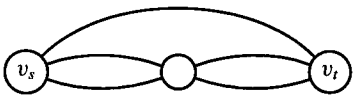


Fig. 5.14
Network survivability

capacity κ, then $x^D_{(s,t)} \leqq 3\,\kappa$. Each disrupted edge-disjoint path reduces the maximum flow by κ, while a failure of the intermediate node eliminates $2\,\kappa$ of potential flow. Stoer (1992) distinguishes therefore node and edge survivability conditions by which the network at hand must contain a minimum number of node-disjoint and edge-disjoint paths, respectively. The following statements on worst cases are trivial by intuition, but it is not easy to prove them (see Jungnickel (1994, Chap. 7)). (1) The maximum number of link-disjoint (v_s, v_t)-paths equals the minimum number of failing links that disrupt the flow between v_s and v_t. (2) Menger's theorem states that the minimum number of failing nodes that disrupt the flow between two non-adjacent[59] nodes equals the maximum number of node-disjoint (v_s, v_t)-paths. Given the above so-called connectivity requirements, one has to determine a cost minimal network design that is immune against a given number of simultaneous failures, where the probability of worst cases may be taken into account.

5.3.3 Minimum Requirements on Failure States

5.3.3.1 Diversification

Corresponding to the suggested heuristics, the minimum investment cost problem (5.17.P)

$$\min_{\boldsymbol{\delta}} \left\{ \mathbf{q}^{\mathsf{T}} \boldsymbol{\delta} \mid \boldsymbol{\Lambda}\, \mathbf{h} = \mathbf{x}^D,\ \boldsymbol{\Delta}\, \mathbf{h} \leqq \boldsymbol{\kappa} + \boldsymbol{\delta},\ \mathbf{h} \geqq \mathbf{0},\ \boldsymbol{\delta} \geqq \mathbf{0} \right\}$$

[59] Direct connections between v_s and v_t are ruled out.

is now augmented by further restrictions with respect to "plausible" failure states. Furthermore, it is useful to think of $\kappa_e + \delta_e$ as multiples of a given capacity per copper or fiber cable. This argument allows for parallel links which may fail independent of each other although a cable cut by a digger can destroy numerous bundled fiber optics at once. Nevertheless, we restrict ourselves to the case of a single failing network element for the sake of simplicity; recall that the destruction of a node implies the removal of all incident links. The diversification approach ignores rerouting and relies therefore on a network design that permits to distribute flows among alternative paths. Based on the OD/path incidence matrix $\mathbf{\Lambda}$ we can collect all route flows that constitute the demand $x^D_{(s,t)}$,

$$\mathbf{\Lambda}_\rho \mathbf{h} = \sum_{\rho \in \mathcal{R}(s,t)} h_\rho = x^D_{(s,t)}.$$

Diversification requires that all route flows sharing a certain link $e \in \mathcal{E}$ or node $v \in \mathcal{V} \setminus \{v_s, v_t\}$ (except the initial node v_s and the final node v_t) do not exceed the fractional part $\theta_{(s,t)} x^D_{(s,t)}$. Thus any failure of a network component disrupts at the most a given fractional part of each OD-specific traffic demand. Let $\mathcal{W}_\rho$ be the collection of all *edges and nodes* other than v_s and v_t constituting the path ρ. Then, diversification introduces

$$\sum_{\rho \in \mathcal{R}(s,t):\ w \in \mathcal{W}_\rho} h_\rho \leqq \theta_{(s,t)} x^D_{(s,t)}.$$

Consequently, if we neglect intermediate nodes, a route $\rho \in \mathcal{R}(s, t)$ consists only of the source v_s, the sink v_t, and a link joining them such that the route (or edge) flow constraint becomes $h_\rho \leqq \theta_{(s,t)} x^D_{(s,t)}$. Then we need at least $\lceil 1/\theta_{(s,t)} \rceil$ parallel edges so that any link failure cannot disrupt more than the maximum edge flow $\theta_{(s,t)} x^D_{(s,t)}$. The analogous reasoning implies for each OD-pair the existence of at least $\lceil 1/\theta_{(s,t)} \rceil$ routes not using the same network element. Inversely, $\theta_{(s,t)} = 50\%$ is the maximum value that can be guaranteed by the network in Figure 5.14 due to a possible failure of the intermediate node. This value can only be decreased by alternative routes not including the intermediate node. To sum up, the structure of the initial linear program remains the same, but the feasible set of cost minimization becomes tighter due to numerous additional flow constraints.

5.3.3.2 Reservation

The first step in the reservation heuristic is that the system designer specifies the fractional parts $\sigma_{(s,t),j}$ of any surviving trip table $\mathbf{x}^D$ that must be routable after the failure of *one* link or node with $j = 1, 2, \ldots$ denoting the failure state. For the sake of convenience we choose the same value for all OD-pairs and all failure states, say $\bar{\sigma}$. In order to distinguish this value from the normal operating state $j = 0$ with $\sigma_0 = 1$ we write $\bar{\sigma} = \sigma_j$, $j = 1, 2, \ldots$

The second step introduces an auxiliary variable $\mathbf{x}^S$ which can be interpreted as the supplied output of the network. The demand equation $\mathbf{\Lambda}_0 \mathbf{h}_0 = \mathbf{x}^D_0$ of the normal operating state $j = 0$ can hence be split into three conditions:

$$\begin{aligned}
&\mathbf{\Lambda}_0\mathbf{h}_0 = \mathbf{x}_0^S && \text{produced end-to-end flows,}\\
&\mathbf{x}_0^S \geqq \sigma_0\mathbf{x}_0^D && \text{we have to produce at least } \sigma_0\mathbf{x}_0^D \text{ with } \sigma_0 = 1,\\
&\mathbf{x}_0^S \leqq \mathbf{x}_0^D && \text{we cannot sell more than } \mathbf{x}_0^D.
\end{aligned}$$

The normal operating state indeed requires $\mathbf{x}_0^S = \mathbf{x}_0^D$.

The third step is to enumerate all failure states by eliminating one element at a time from the initial network. The set $\mathcal{J}$ of all network states includes particularly the normal operating state $j = 0$. Each state $j \in \mathcal{J}$ is characterized by the restrictions

$$\text{(5.54)}\quad \mathbf{\Lambda}_j\mathbf{h}_j = \mathbf{x}_j^S,\ \ \mathbf{\Delta}_j\mathbf{h}_j \leqq \boldsymbol{\kappa}_j + \boldsymbol{\delta}_j,\ \ \mathbf{x}_j^S \geqq \sigma_j\mathbf{x}_j^D,\ \ \mathbf{x}_j^S \leqq \mathbf{x}_j^D,\ \ \mathbf{h}_j \geqq \mathbf{0},\ \ \mathbf{x}_j^S \geqq \mathbf{0},$$

where a subscript $j \neq 0$ indicates that all elements with respect to the missing network element have been removed. Recall again that a node failure may be interpreted as the simultaneous breakdown of all edges incident to it. The parameter σ_j equals always $\bar{\sigma}$ except for the normal operating state with $\sigma_0 = 1$.

Regarding the cost minimal network design problem, the augmented problem with respect to a survivable network is

$$\text{(5.55)}\qquad \min\left\{\mathbf{q}^\mathsf{T}\boldsymbol{\delta}_0 \mid \text{(5.54) holds good for all } j \in \mathcal{J},\ \ \boldsymbol{\delta}_0 \geqq \mathbf{0}\right\}.$$

The variables of this problem are $\boldsymbol{\delta}_0$ and $\mathbf{h}_j$, $j \in \mathcal{J}$. Once the capacity increments $\boldsymbol{\delta}_0$ have been chosen, each system state j may have a different routing $\mathbf{h}_j$ which is not necessarily unique. The complexity of this program becomes apparent if we consider that the determination of an optimal routing for the normal operating state is difficult enough. But now we need routing patterns for every failure state, too.

Having selected an investment program $\hat{\boldsymbol{\delta}}_0$, the normal operating state shows installed capacities $\bar{\boldsymbol{\kappa}}_0 := \boldsymbol{\kappa} + \hat{\boldsymbol{\delta}}_0$. The feasibility conditions for the reservation model can be discussed in terms of the collection of constraints (5.54). Note that an optimal routing $\hat{\mathbf{h}}_j$ in any failure state $j \neq 0$ must satisfy $\mathbf{\Lambda}_j\hat{\mathbf{h}}_j \geqq \bar{\sigma}\mathbf{x}_j^D$ and $\mathbf{\Lambda}_j\hat{\mathbf{h}}_j \leqq \mathbf{x}_j^D$. As long as we look for an admissible (rather than an optimal) solution $\mathbf{h}_j$, both restrictions can be reduced to $\mathbf{\Lambda}_j\mathbf{h}_j = \bar{\sigma}\mathbf{x}_j^D (\leqq \mathbf{x}_j^D)$. *Survivability* means that the multicommodity flow problem has a feasible solution for every failure state j:

$$\text{(5.56)}\qquad \begin{aligned}
\mathbf{\Lambda}_j\mathbf{h}_j &= \bar{\sigma}\mathbf{x}_j^D &&\\
\mathbf{\Delta}_j\mathbf{h}_j + \mathbf{k}_j &= \bar{\boldsymbol{\kappa}}_j && \text{(installed new capacities)}\\
\mathbf{h}_j &\geqq \mathbf{0} &&\\
\mathbf{k}_j &\geqq \mathbf{0} && \text{(excess capacities)}
\end{aligned}$$

Each of these systems corresponds to (5.10.A) on p. 240 and the capacity vector $\bar{\boldsymbol{\kappa}}_0$ is said to be feasible if (5.56) has a feasible solution for every system state including the normal operating state. Applying Farkas' Lemma with variables $\mathbf{v}_j$ and $\mathbf{u}_j$ in accordance with (5.10.B), feasibility applies if and only if the condition (5.12) in Proposition 5.3 holds true for every $j \in \mathcal{J}$, that is

$$\text{(5.57)}\qquad \bar{\boldsymbol{\kappa}}_j^\mathsf{T}\mathbf{v}_j \geqq \sigma_j\,\mathbf{x}_j^{D\mathsf{T}}\mathbf{u}_j \qquad \forall\,\mathbf{v}_j \geqq \mathbf{0},\ \ \forall\, j \in \mathcal{J}.$$

It is important to notice that the shortest path values $\bar{\mathbf{u}}_j$ must be recalculated for every edge weight $\mathbf{v}_j$ and all system states j. That is, $\mathbf{w}_j := \boldsymbol{\Delta}^\top \mathbf{v}_j$ and

$$\bar{u}_{(s,t),j} := \min_{\rho \in \mathcal{R}_j(s,t)} w_{\rho,j}$$

It should be clear that even small networks require remarkable computational effort to solve such a problem. For example, the network in Figure 5.14 can have $3 + 5$ failure states, where 6 of them deal with a surviving demand $x^D_{(s,t)}$. Nevertheless, if we choose some $0 < \bar{\sigma} < 1$ and if the first failure state $j = 1$ denotes a break down of the transmission node, then at least the (redirected) part $\bar{\sigma}\, x^D_{(s,t),j}$ of the demand must be served by the route consisting of the direct connection $e_1 = (v_s, v_t)$, that is $\bar{\kappa}_{e_1,j} \geqq \bar{\sigma}\, x^D_{(s,t),j}$ with $j = 1$. Of course, this observation is confirmed by (5.57), because the only surviving route $\rho_1 = e_1$ has the shortest path value $\bar{u}_{(s,t),j} = w_{\rho_1,j} = v_{e_1,j}$ with $j = 1$. Since we cannot expect to obtain such trivial results for larger networks, it is useful to relax the problem particularly with respect to the ability of the network carrier to reroute the surviving demand in any failure state.

5.3.3.3 Path Restoration

The path restoration model reduces the reservation model by the way of rerouting: the route flows of the normal operating state that do not include the failing network component are maintained and those surviving route flows that are affected by the failure can be rerouted, see Wessäly (2000, p. 44 ff.). Beginning with the normal operating state $j = 0$ we know form the OD/path incidence matrix $\boldsymbol{\Lambda}_0$ all routes that connect each OD-pair $(s, t) \in \mathcal{B}_0$, namely, $\rho \in \mathcal{R}_0(s, t)$. Furthermore, the edge/path incidence matrix $\boldsymbol{\Delta}_0$ encodes all routes in the network, i.e., $\rho \in \mathcal{R}$, that use one particular link e. Consequently, we can compute all link flows φ_e, $e \in \mathcal{E}_0$, which are bounded to capacity restrictions.

$$\varphi_{e0} = \sum_{\rho \in \mathcal{R}:\, e \in \rho} h_{\rho 0} = \sum_{(s,t) \in \mathcal{B}_0} \sum_{\rho \in \mathcal{R}_0(s,t):\, e \in \rho} h_{\rho 0} \leqq \kappa_e + \delta_{e0} = \bar{\kappa}_{e0} \qquad \forall\, e \in \mathcal{E}_0$$

The rerouting makes use of the excess capacities at the normal operating state $\bar{\boldsymbol{\kappa}}_0 - \boldsymbol{\varphi}_0$ *plus* the released capacities of those routes which cannot be used any longer because of a failing network element. So rerouting includes disrupted path flows of surviving demand as well as path flows that correspond to canceled demand (the origin or destination died). Let $h_{\rho j}$ be the flow *after path restoration* on a surviving route ρ. Then, letting $\mathcal{R}^*(s, t) := \mathcal{R}_0(s, t) \setminus \mathcal{R}_j(s, t)$, the flow increments $\mathrm{d}h_{\rho j} = h_{\rho j} - h_{\rho 0} \geqq 0$ must satisfy

$$\sum_{(s,t) \in \mathcal{B}_j} \sum_{\rho \in \mathcal{R}_j(s,t):\, e \in \rho} \mathrm{d}h_{\rho j} \leqq \bar{\kappa}_{e0} - \sum_{(s,t) \in \mathcal{B}_0} \sum_{\rho \in \mathcal{R}_0(s,t):\, e \in \rho} h_{\rho 0} + \sum_{(s,t) \in \mathcal{B}_0} \sum_{\rho \in \mathcal{R}^*(s,t):\, e \in \rho} h_{\rho 0}$$

$$\forall e \in \mathcal{E}_j,\ j \neq 0$$

Rearranging the sums on the right hand side yields

$$\sum_{(s,t)\in\mathcal{B}_j}\ \sum_{\rho\in\mathcal{R}_j(s,t):\, e\in\rho} \mathrm{d}h_{\rho j} \leqq \bar{\kappa}_{e0} - \sum_{(s,t)\in\mathcal{B}_0}\ \sum_{\rho\in\mathcal{R}_j(s,t):\, e\in\rho} h_{\rho 0}.$$

Notice that there is no route connecting v_s with v_t if $(s, t) \notin \mathcal{B}_j$. Hence, the relevant part of $\mathcal{B}_0$ is $\mathcal{B}_j$. The preceding capacity restriction and the corresponding demand constraints become now

$$\sum_{(s,t)\in\mathcal{B}_j}\ \sum_{\rho\in\mathcal{R}_j(s,t):\, e\in\rho} \left[h_{\rho 0} + \mathrm{d}h_{\rho j}\right] \leqq \kappa_e + \delta_{e0} \qquad \forall e \in \mathcal{E}_j,\ j \in \mathcal{J}, \tag{5.58a}$$

$$\sum_{\rho\in\mathcal{R}_j(s,t)} \left[h_{\rho 0} + \mathrm{d}h_{\rho j}\right] \geqq \sigma_j x^D_{(s,t)} \qquad \forall\, j \in \mathcal{J}. \tag{5.58b}$$

The first condition restricts the link flows for every systems state with a surviving link e including the normal operating state $j = 0$ with $\mathrm{d}h_{\rho j} = 0$. The second condition specifies the minimum of a given demand that must be served in any case. At last the corresponding network design problem turns out to be

$$\min\left\{\mathbf{q}^{\mathsf{T}}\boldsymbol{\delta}_0 \mid (5.58) \text{ holds true},\ \boldsymbol{\delta}_0 \geqq \mathbf{0}\right\}. \tag{5.59}$$

As long as an optimal rerouting can be implemented at low cost, an optimal solution to (5.59) is suboptimal to the corresponding solution of (5.55). But with increasing networks the effort to detect an optimal rerouting and to execute it within reasonable time can be tremendous. In these cases a partial rerouting predominates a complete traffic reassignment especially when the failure is expected to be corrected after a reasonable period of time.

5.15 Summary Network design problems tend to concentrate traffic volumes on a few fully developped interconnections. This observation becomes particularly evident with the rise of modern technologies such as fiber optics. We then have to ask for the vulnerability of networks. When failing network components cannot be precluded at reasonable costs, further strategies are needed to protect survivability and reliability of networks. The approach suggested here is to provide excess capacities that can be used in failure states in order to reroute the surviving traffic. Due to the complexity of this network design problem several heuristics are provided to obtain approximately cost minimal investment programs. Besides stochastic reasons as have been discussed in queuing networks this is the second argument to build up excess capacities. While these buffers ensure regular network operations, a third cause for excess capacities may be associated with the competitive situation of a network carrier. The potential supply of network services is frequently used as a credible threat against (potential) competitors to widen the supply at decreasing prices. □

5.4 Summary

Network design starts on the basis of the results on optimal traffic assignments that have been derived in Chapter 3. While the former analysis takes the network structure

network in such a manner that it fits best to the needs of traffic demand. According to Chapter 3 we started with a technical problem that determines the maximum flow under the assumption that the prespecified link capacities can be augmented by means of a common pool of resources. This task corresponds to the determination of a maximum product combination as it is known from the theory of the firm. Consequently we analyse also the minimum cost combination which follows from the "dual" investment cost problem. The next step is to generalize the results for a single good to a multicommodity flow problem, cf. Section 5.2. In this setting the maximum flow problem becomes void. Instead we have discussed conditions of high order complexity that ensure the existence of a technically feasible flow pattern. The subsequent investment problem differs from a pure network extension problem by the fact that it minimizes the sum of transport costs and investment costs. Hence, capacity investments ensure not only a feasible flow pattern but they serve also to reduce transport costs by redirecting flows to cheaper routes. This so-called traffic diversion is closely related to traffic generation which has been subsumed under the case of a price sensitive demand for traffic.

Aside from congestion effects it should be clear that bundling of flows incorporates two opposite forces. On the one hand, cost reducing effects lead to hierarchical networks, such as hub-and-spoke networks, where most of the traffic is concentrated on a few interhub connections. On the other hand, sparse networks become more vulnerable against failures of network components. While survivability aims at a diversified network design that ensures the functionality of the network in case of failing network components, hub-and-spoke networks try to exploit link specific economies of scale. Queuing introduces another cost category besides transport costs and investment costs, namely, costs of waiting. The corresponding network design problem then pursues the objective to minimize the overal system delay. Contrasted with waiting systems are loss systems where all or parts of the rejected demand for traffic is lost for the service provider. In this case network design aims at a proper adjustment of the network such that a prespecified level of quality of service can be assured at minimum costs. In what follows Chapter 6 proceeds with an observation that is also included in queuing. The main cause for system delays is that a stochastic traffic demand accidently overwhelms network resources. One way to coordinate these processes is queuing, another option is to synchronize network activities by schedules such that network resources are utilized more uniformly.

6

Synchronization of Network Processes

6.1 Scheduling Discrete Event Dynamic Systems

6.1.1 Resource Utilization of Simultaneous Processes

Queuing and its service disciplines indicate one way to organize the utilization of scarce network resources by transport processes. The basic observation is that individual customers (or jobs) generally do not cooperate so that they arrive at service stations according to some stochastic process, which may overwhelm the capacity of the respective service unit. As a consequence customers have to wait in queues especially in periods of peak demand. This chapter analyses an attempt to *synchronize flows* by timetables in order to omit overloads of particular network elements. The aim is to determine timetables such that the whole system is in tune with its weakest parts. When all jobs are processed on a "network of perfectly coordinated assembly lines" we have to ask furthermore for cost minimal solutions as will be done in Section 6.4.3. Since the formal analysis of Section 6.2 refers to some nonstandard concepts – namely, max-plus algebra – all considerations will be accompanied by numerical examples that can be interpreted in terms of railway systems. Readers may check these results by the *MATHEMATICA*® modules presented in Section 6.3.

Many transportation phenomena can be described as so-called *discrete event dynamic systems* (DEDS). A DEDS is a dynamic asynchronous system where the state transitions are initiated by events that occur at discrete instants of time. An event corresponds to the start or the end of an activity. Examples are scheduled transportation services on networks or the timing of a system of coordinated traffic signals along an urban arterial or in an urban road network. A common property of such systems is that the start of an activity depends on the termination of several other activities. As an example, we refer to railway systems where departures of trains require the arrivals or departures of certain other trains. These restrictions ensure, on the one hand, that passengers can change over and, on the other hand, that tracks become free. The same reasoning applies to road networks, where the start of a green time phase of traffic lights at the entry of a crossing depends on platoon

arrivals as well as on the blocking of conflicting processes by corresponding red phases. Although numerous processes, journeys say, can be performed by a network at a time, single network components have limited capacity to serve more than one activity. When resource sharing is bounded above, certain activities have to wait until all conditions of their service are fulfilled. Such systems cannot conveniently be described by differential or difference equations, but they naturally exhibit a periodic behavior.

Typical questions to be answered in DEDS are given in the following list.

Precedence constraints There are precedence constraints that restrict the order of jobs to be performed. For instance, a journey along a route is defined by a necessarily successive sequence of passing roads. Such sequential jobs are contrasted with parallel tasks. On road networks with mutually independent trip choices and one vehicle per trip, the typical service discipline is FIFO. This is different in railway systems where groups of passengers share the same trains. The order in which customers enter service mutates at every interchange station, where passenger groups intermix.

Scheduling A schedule for a bundle of jobs is a specification of start times such that certain precedence constraints and other restrictions are met. The efficient use of network resources requires synchronization particularly for repeated tasks. In the end a timetable is sought that minimizes cost and/or some measure of time such as the overall project completion time (when the last job is finished), the tardy time (the amount by which the completion time exceeds a given deadline), or the overall system delay as defined on p. 282 with regard to packet switched data networks.

Bottlenecks Increasing the throughput of a network can be reached in principle by faster timetables. This process is limited by the slowest parts of the network which form the critical graph. An expansion of that "network capacity" requires to determine the corresponding arcs and nodes and the costs of disposing these bottlenecks.

Stability Similar to data networks it is strongly recommended not to run the network at its capacity limit. The reason is that the transportation system should be able to cope with a certain amount of irregularity without severe consequences. The idea is to build in buffer times at transshipment nodes so that delays phase out as time goes by. Buffer times increase the reliability of the system at the cost of reducing the throughput and additional waiting times.

Sensitivity analysis Stability analysis needs theory on delay propagation including the quantification of sensitivity to perturbations. While small delays vanish immediately by virtue of buffer times, larger delays disappear slowly and proceed from node to node so that the carrier has to decide on thresholds whether to cancel connecting trains or not. One important indicator is the maximum delay at particular nodes that do not reach certain other nodes so that the system returns to the normal operating schedule.

Costs of waiting Operational tasks are to be contrasted with costs of waiting that are mainly borne by passengers. The service provider ameliorates its level of

service by minimizing the overall waiting times of the customers taking schedule reliability[1] into account. This type of cost minimization is to be supplemented with objectives of network design. Increasing frequencies, clearing bottlenecks, and suspending precedence constraints (such as canceling connecting trains) are some instruments that aim at improving the level of service.

The above aspects suggest in accordance with Bowman, Turnquist (1981) that *frequency* and *reliability* of network services are the most important features of passenger transport networks that determine their comfort of usage. In the presence of competitive transport modes and especially of private cars almost all carriers are concerned with improving their quality of service. In doing so private as well as public passenger carriers have to relate the costs of service improvement measures to the users' willingness to pay. In a first step, this requires deep knowledge on the production technology of the network which is described here by some DEDS. The next step is to determine those measures which improve frequency and reliability. Last but not least the research aims at the provision of some type of cost-benefit analysis by which the costs of improvement measures are to be confronted with the corresponding benefits to the users.

Many frameworks exist to study DEDS. Examples are queuing theory, Petri nets, computer simulation, and max-plus algebra. The framework of max-plus algebra will be introduced in Section 6.2, where the results are explained by the synchronization of railway systems. The corresponding numerical computations refer to several modules that are provided by making use of *MATHEMATICA*®; see Section 6.3. Applying max-plus algebra essentially reveals dynamic characteristics of transportation systems such as minimum cycle times, critical cycles, and stability. The basics of the relationship between max-plus algebra and Petri nets will be discussed in Section 6.4. This discussion demonstrates the basic coherence between the number of events and their dates. Observe furthermore that queuing systems have been discussed in Section 5.2.3.3 of the preceding chapter. On the max-plus formulation of such stochastic queuing systems see Baccelli et al. (1992).

6.1.2 Real World Applications

Before introducing the concept of max-plus algebra – a problem-adequate method of calculation – let us take a look at some applications to get a better understanding of what synchronization of network processes means. One major example is presented in Braker (1993) who applies max-plus algebra to the connections and timetables of the Netherlands intercity railway system. It is shown that this large scale system is stable and the degree of stability is quantified. Moreover, critical lines are indicated explicitly which are crucial in the propagation of delays and the design of faster timetables. Further details will be reported in Example 6.19 below after the introduction of necessary analytical instruments. Another research project at the *Max-*

[1] Reliability means here the degree of schedule adherence so that passengers can trust in the announced transit services.

Planck-Institut für Dynamik komplexer technischer Systeme, Magdeburg[2] applies the max-plus framework to the suburban streetcar system of the city of Stuttgart, Germany, as represented in Figure 6.1.[3] The same method has also been employed to analyze inland navigation.

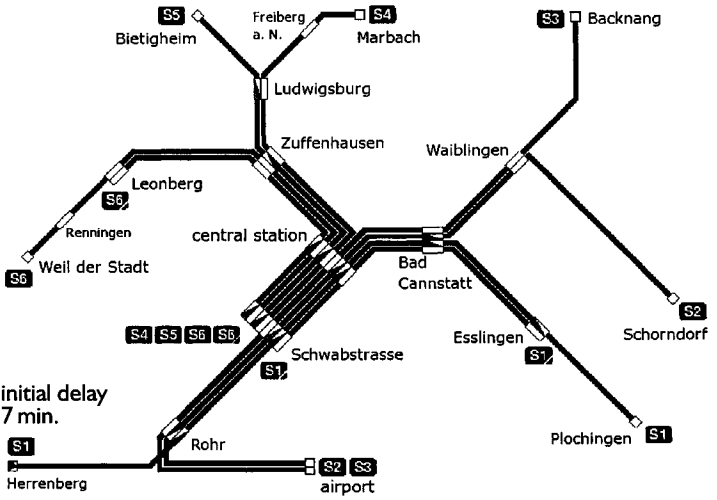


Fig. 6.1 Suburban streetcar system of the city of Stuttgart, Germany

Having detected a feasible schedule that satisfies all precedence constraints, the next step is to analyze the power of the system to cope with more or less severe irregularities. Public transport is very disturbance sensitive due to weather conditions, the numbers of passengers getting in and out, traffic jams, accidents, etc. These circumstances lead to many delays as remarked by van Egmond (1998). Such obstructions propagate through the entire network and they slow down travel speed not only for delayed trains but also for connecting trains. In the worst case travelers miss their canceled connecting trains so that they have to wait for the next departure at cold and drafty transfer points. If trains or buses wait for each other, the same problem is shifted to passengers in the waiting trains. Delayed passengers benefit from maintained connections, while waiting travelers are discriminated. As a consequence, service providers have to analyze delay propagations in consideration of appropriate buffer times. They increase reliability at the cost of longer waiting times that provide scheduled operation. When buffer times do not suffice to even out delays, carriers have to find decision rules by which they maintain or cancel connecting trains in order to minimize the overall system delay.

[2] Further details on this project can be found at `www.mpi-magdeburg.mpg.de` or `ifatwww.et.uni-magdeburg.de/systemtheorie`.

[3] For the corresponding map of the Netherlands intercity railway system see Braker (1993).

The instrument of synchronization control times – that are the times the trains wait for delayed feeder services – is discussed in Goverde (1998a, 1998b). He deals with the problem of minimizing the overall passenger waiting time, where guaranteeing scheduled connections is complicated by delays during operation. The model can be utilized to evaluate the optimal synchronization control policy with respect to given arrival delays: secure or dissolve a transfer. Irrespective of excessive delays there are various conflicting interests which are to be weighted in order to define a plausible transfer quality measure.

- Arrival and departure times affect various passenger groups and their respective waiting times in different ways. Just compare transit travelers sitting in a train and waiting for other possibly delayed transfer passengers. Bear in mind that customers may be patient and willing to wait, but fretful passengers may switch the next time to competing transport modes such as private cars.
- Minimum headways resulting from safety criteria and conflicting movements of train routes through stations restrict the arrival and departure times at and from each single station. The same argument holds true at least for neighboring stations and in the end it influences the whole network.
- As it is typical for queuing networks, the network operator has to decide on an appropriate service discipline. That may usually be FIFO, but prioritization of various train services is frequently used to improve efficiency and thus quality of service.
- Service capacity is confined by several technical aspects. Speed limits and minimum headways restrict track utilization. Stations have a bounded number of tracks that limit the count of train movements per period. The capacity of trains can be varied only within narrow limits in particular when the operator tries to avoid empty return transports. Last but not least the capacities of waiting rooms restrict the number of customers. This fact is relevant not only for narrow subway systems but also for traffic light control.
- Increasing traffic intensity can be expected to raise the number and the extent of delays. Such congestion effects impose high risk on a scheduled network operation and undermine the stability of scheduled train circulation times. In this sense delays have more severe consequences for running a network on schedule than for road traffic where individual motorists are less restrictively tied to each other.

Besides railway connections various alternative problems do exist showing the same formal properties. Baccelli et al. (1992), for example, present results on *manufacturing systems* consisting of several machines. One task is the determination of the speed with which manufacturing systems can work. The same framework applies to *parallel computation* where different processors are required to operate synchronously. Such systems can be described by so-called "heaps of pieces" as can be found in van Egmond (1999). With regard to systems of elements of material infrastructure, each "piece" is a staircase-shaped block that describes the time intervals at which a vehicle movement (or task) occupies certain resources. The corresponding "heaps" result from a graphic representation in the distance-

time diagram. Max-plus algebra is then used to solve the sequencing problem by composing the pieces in a proper way in order to find an efficient usage of network resources. Two similar examples for discrete event systems (DES) are given in Cassandras et al. (1995), namely, telephone systems and dispatching control in an elevator system. Both cases focus on the ordering of events and not on the times at which these events occur.

The aforementioned properties of railway networks suggest to apply max-plus algebra also to other *queuing systems*. One example with finite capacities is presented in Baccelli et al. (1992). In this queuing system a customer (or task) has to pass a series of servers, where each server needs a certain time to work on the respective job. Three conditions must be fulfilled before a server can start helping a customer. (1) The preceding server must have finished the corresponding job. (2) The actual server must have completed the preceding task. (3) In the case of no or full buffers, a customer can release a server only if he can enter an idle succeeding server. This principle applies also to urban road traffic as can be found in Mancinelli et al. (2001) who deal with the *traffic light control* of regular towns. The authors describe a deterministic model of traffic light synchronization based on Petri net modeling and max-plus algebra. In the next step such deterministic queuing systems are to be augmented by stochastic event patterns; a task which is left for future research.

6.1.3 Auxiliary Transportation Problems

For the sake of intuition let us continue with some technical details by making use of the above *railway examples*. All transportation processes within a discrete event dynamic system (DEDS) take place on a weighted digraph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ of order $\#\mathcal{V} = n$. Arc weights are given by an $n \times n$ matrix $\mathbf{T}$ with elements τ_{ij} denoting the time needed to pass the arc $e = (v_j, v_i)$. Observe that this movement starts at v_j and ends at v_i, while the subscripts of τ_{ij} are interchanged. This unusual notation turns out to be useful with regard to the algebra introduced below. A journey from a source v_s to a destination v_t corresponds to a motion along some path $\rho \in \mathcal{R}(s, t)$ that comprises a sequence of consecutive, forward directed arcs. Each movement requires that the preceding steps have been finished. As an example, we refer to Figure 6.2 that represents the weighted digraph of a network with two nodes.

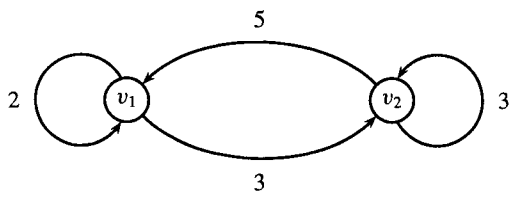


Fig. 6.2
Weighted digraph with loops
The digraph explains basic features of DEDS with arc weights

$$\mathbf{T} = \begin{pmatrix} 2 & 5 \\ 3 & 3 \end{pmatrix}.$$

A movement from v_2 to v_1 requires 5 units of time, i.e., $\tau_{12} = 5$. Similarly we have $\tau_{21} = 3$; once an activity starts in node v_1 (departure of a train) it takes 3 time units before the result reaches node v_2 (arrival of the train). Furthermore, the

digraph permits loops, where (v_1, v_1) has a time need of $\tau_{11} = 2$. Such loops are particularly useful to describe circular tours through suburbs, the details of which are of minor importance to the main schedule. Another example may be maintenance work including switches of crews that are necessary at the termini of a tour. To sum up one possible train route might start with a loop (v_1, v_1). It then passes to v_2, follows the loop (v_2, v_2) and returns to v_1.

For larger networks not all arcs (v_j, v_i) can be expected to exist. This case is expressed by an abstract symbol ε or more precisely

$$\tau_{ij} = \varepsilon,$$

that will be explained soon in more detail. Now each $n \times n$ square matrix $\mathbf{T}$ corresponds to a *precedence graph*[4] $\mathbf{G(T)}$ with an existing arc (v_j, v_i) if $\tau_{ij} \geqq 0$ and no connection if $\tau_{ij} = \varepsilon$. An example with three nodes is given in Figure 6.3, where the loop (v_2, v_2) in Figure 6.2 is split into two arcs connecting a third node to the network.

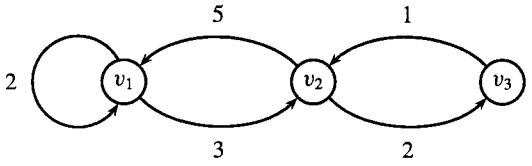


Fig. 6.3
Precedence graph of $\mathbf{T}$
The precedence graph $\mathbf{G(T)}$ with

$$\mathbf{T} = \begin{pmatrix} 2 & 5 & \varepsilon \\ 3 & \varepsilon & 1 \\ \varepsilon & 2 & \varepsilon \end{pmatrix}$$

has four possible links missing.

The example of Figure 6.3 suggests to accommodate the demand for traffic by three vehicles such as trains, buses or something else. One shuttles on the loop (v_1, v_1), one plies between v_1 and v_2, and the last plies between v_2 and v_3. A real world example with regard to the streetcar system of Stuttgart has been given in Figure 6.1.

The basic features of such a DEDS follow from a list of precedence constraints. The easiest case results from the loop (v_1, v_1). Each trip of the respective shuttle requires at least that the preceding trip has been finished. To be more concrete, the k^{th} (realized) departure time $\mathrm{t}_1(k)$ of a train at station v_1 is given by a timetable with an announced departure time $d_1(k)$, but not earlier than its arrival time $t_a(k)$ plus 2 minutes to allow passengers to leave and enter the train. The earliest date of the occurrence is then

$$\mathrm{t}_1(k) = \max\{t_a(k) + 2, d_1(k)\}.$$

Notice that the arrival time $t_a(k)$ is given by the $k - 1^{\text{st}}$ departure time at station v_1 plus the travel time τ_{11}, i.e., $t_a(k) = \mathrm{t}_1(k-1) + \tau_{11}$. (If node v_1 in the preceding

[4] The relationship between the matrix $\mathbf{T}$ and the precedence graph is closely related to the relationship between a digraph $\mathbf{G}$ and its *adjacency matrix* $\mathbf{A}(\mathcal{G})$. This $n \times n$ matrix has elements $a_{ij} = 1$ if j is a predecessor of i and $a_{ij} = 0$ otherwise. Notice again the order of i and j which is interchanged in most literature on graph theory.

examples has started an activity, the next activity cannot start within $\tau_{11} = 2$.) If the shuttle considered has to wait for the $k - 1^{st}$ train coming from station v_2, the departure time includes a further restriction yielding

$$\mathrm{t}_1(k) = \max\{\mathrm{t}_1(k-1) + \tau_{11}, \mathrm{t}_2(k-1) + \tau_{12}, d_1(k)\}. \tag{6.1}$$

More generally, the k^{th} departure time at station v_i is the scheduled departure time $d_i(k)$ unless a feeder train is delayed, in which case the departing train has to wait. For transport systems – such as the streetcar system in Figure 6.1 – with n stations there are n equations of the form[5]

$$\mathrm{t}_i(k) = \max\{\mathrm{t}_1(k-1) + \tau_{i1}, \ldots, \mathrm{t}_n(k-1) + \tau_{in}, d_i(k)\}, \quad i = 1, \ldots, n. \tag{6.2}$$

In order to analyze such confusing systems, let us substitute the maximum operator by $\oplus$ and let $\otimes$ denote addition. These steps transform the above system into a system of "linear" equations

$$\mathrm{t}_i(k) = \mathrm{t}_1(k-1) \otimes \tau_{i1} \oplus \cdots \oplus \mathrm{t}_n(k-1) \otimes \tau_{in}, d_i(k), \quad i = 1, \ldots, n,$$

which reminds us of a first order difference equation. Section 6.2 will show indeed that the new operators $(\oplus, \otimes)$ are similarly related as addition and multiplication, say $(+, \times)$, in conventional algebra. Using vector notation, the preceding system takes the form

$$\mathbf{t}(k) = \mathbf{T} \otimes \mathbf{t}(k-1) \oplus \mathbf{d}(k), \tag{6.3}$$

where we are particularly interested in the determination of *optimal schedules* $\mathbf{d}$ *that are to be reiterated after a while.*

In what follows the analysis solely deals with directed graphs $\mathbf{G}$, paths ρ and cycles ζ, rather than their undirected counterparts, namely, graphs, chains, or circuits. Hence, we can drop the vector symbol from our notation for convenience. Furthermore, it is necessary to distinguish between the length and the weight of a path as well as the distance between two points. Notice that paths do not necessarily coincide with railway lines, but they refer to journeys of travelers who can switch between lines.

Length: The length $|\rho|_\ell$ of a path ρ is the number of its arcs. Thus, the length $s_e = s_{ij}$ of an arc $e = (v_j, v_i)$ equals 1 unless otherwise specified.

Weight: The weight $|\rho|_w$ of a path ρ is the sum of the weights $w_e = w_{ij}$ of all arcs $e = (v_j, v_i)$ constituting the path.

$$|\rho|_w := \sum_{e \in \rho} w_e$$

In the case of travel times the arc weights w_{ij} are denoted by τ_{ij} and the matrix $\mathbf{W}$ of general arc weights goes over to $\mathbf{T}$.

[5] For explanatory purposes we omit here systems of higher order, where $\mathrm{t}_i(k)$ depends, for example, on $\mathrm{t}_1(k-2)$. This would be the case when there are two buses that potentially attend to the loop (v_1, v_1). A justification for ignoring this case is given in Remark 6.3 below.

Mean weight: The mean weight $\mu(\rho)$ of a path ρ is defined as the weight of the path divided by its length,

$$\mu(\rho) := \frac{|\rho|_w}{|\rho|_\ell}.$$

Distance: The distance between two nodes (v_s, v_t) is defined by the shortest interconnecting path, that is the path with the smallest weight.

$$\hat{\rho}_{(s,t)} := \min_{\rho \in \mathcal{R}(s,t)} |\rho|_w$$

This corresponds to minimum route costs, when substituting arc weights w_e by link costs c_e as has been done in Chapter 3 with respect to a user optimal trip assignment.

The same definitions apply to cycles, where a cycle ζ is a closed path which terminates at its starting node. The length $|\zeta|_\ell$, the weight $|\zeta|_w|$, and the mean weight $\mu(\zeta)$ are defined by the respective term of the path ρ that corresponds to ζ. A *loop* denotes the special case of a cycle of length one. The mean weight of a cycle is also called *cycle mean* and the *maximum cycle mean* $\mu_{\max}(\mathbf{W})$ in the precedence graph $\mathcal{G}(\mathbf{W})$ will be of particular interest. It signifies the maximum of the cycle means, where the maximum is taken over all cycles in the graph.

Most results of max-plus algebra will be explained by means of the examples given in Figure 6.2 or Figure 6.3. In the general case the $n \times x$ matrix of arc weights is denoted $\mathbf{W}$. Whenever we refer to travel times the matrix $\mathbf{T}$ is used. In simple cases the numerical results can be computed by hand. At least for large scale networks the use of computer software becomes indispensable. Section 6.3 introduces modules for *MATHEMATICA*® that can be applied to check the numerical outputs in the subsequent analysis. These tools demonstrate the power of *MATHEMATICA*® when solving general problems of symbolic algebra.

6.2 Max-Plus Algebra

6.2.1 Basic Operations

Precedence constraints as indicated by (6.1) require mainly two operations. First, the earliest possible time an event can occur follows from the preceding process that finalizes last (*parallel composition*). For two preceding activities, say a and b, this property is captured by an operator $\oplus$ defined by

$$a \oplus b := \max\{a, b\},$$

where we shall see that operations based on $\oplus$ follow similar rules as the $+$ operator in conventional notation. The second operator $\otimes$ of max-plus algebra is used to compute cumulative time needs (*series composition*),

$$a \otimes b := a + b,$$

and it corresponds to multiplication, i.e., $\cdot$ or $\times$, in conventional algebra. The symbol ε, which has been introduced to indicate a non-existing connection, is now assigned the value $-\infty$ so that ε plays the part of a *zero element* for the extended set of real numbers $\mathbb{R}_\oplus := \mathbb{R} \cup \{\varepsilon\}$,

$$\varepsilon \otimes a = a \otimes \varepsilon = \varepsilon \qquad \forall a \in \mathbb{R}_\oplus.$$

Moreover, the convention that $\varepsilon - \varepsilon := \varepsilon$ is needed, which is not defined in conventional algebra. Similarly, $\bar{e} := 0$ assures the existence of an *identity element* with $a \otimes \bar{e} = \bar{e} \otimes a = a$. Table 6.1 comprises the structure[6] of max-plus algebra and compares it to conventional results. Line four and five are referred to as *commutativity* of addition and multiplication, respectively. Line ten and eleven state *associativity* of addition and multiplication. The rule of the last line is called *distributivity* of multiplication over addition.

Table 6.1 Max-plus algebra. Comparison of operations in conventional and max-plus algebra

conventional operations	operations in max-plus algebra
$a + b$	$a \oplus b := \max\{a, b\}$
$a \cdot b$	$a \otimes b := a + b$
a^b	$a^{\otimes b} := a \cdot b$
$a + b = b + a$	$a \oplus b = b \oplus a$
$a \cdot b = b \cdot a$	$a \otimes b = b \otimes a$
0	$\varepsilon := -\infty$
1	$\bar{e} := 0$
$a + 0 = a$	$a \oplus \varepsilon = a$
$a \cdot 0 = 0$	$a \otimes \varepsilon = \varepsilon$
$a \cdot 1 = a$	$a \otimes \bar{e} = a$
$(a + b) + c = a + (b + c)$	$(a \oplus b) \oplus c = a \oplus (b \oplus c)$
$(a \cdot b) \cdot c = a \cdot (b \cdot c)$	$(a \otimes b) \otimes c = a \otimes (b \otimes c)$
$a \cdot (b + c) = a \cdot b + a \cdot c$	$a \otimes (b \oplus c) = a \otimes b \oplus a \otimes c$

In the light of (6.2) the next step is to ask for matrix operations that are capable of representing systems of equations. As indicated by (6.3) let us define the inner product $\mathbf{y} = \mathbf{A} \otimes \mathbf{x}$ with an $m \times n$ matrix $\mathbf{A}$ as short hand notation for

$$\begin{aligned} & y_i = \max\{a_{i1} + x_1, \ldots, a_{in} + x_n\} && i = 1, \ldots, m \\ :\Longleftrightarrow \quad & y_i = a_{i1} \otimes x_1 \oplus \cdots \oplus a_{in} \otimes x_n && i = 1, \ldots, m \\ :\Longleftrightarrow \quad & \mathbf{y} = \mathbf{A} \otimes \mathbf{x}. \end{aligned}$$

[6] Readers interested in further details of max-plus structure may refer to Baccelli et al. (1992) who exactly define properties of groups, subdioids, dioids (or semirings), etc.
Jungnickel (1994) discusses shortest path problems on the basis of a min-plus algebra which is also referred to as *path algebra*. In this framework $\ominus$ denotes the minimum operation and the zero element is $\ni = +\infty$.

When an $m \times n$ matrix $\mathbf{A}$ can be multiplied by an $n \times 1$ vector $\mathbf{x}$, we can also ask for $\mathbf{A} \otimes \mathbf{B}$, where $\mathbf{B}$ is an $n \times p$ matrix. Analogously to conventional algebra, the result is given by the following rule

$$(\mathbf{A} \otimes \mathbf{B})_{ij} = \bigoplus_{k=1}^{n} a_{ik} \otimes b_{kj} = \max_{k=1,\ldots,n} a_{ik} + b_{kj},$$

where $\bigoplus$ corresponds to $\sum$ in conventional notation. Observe that in general $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$. Reiterating (6.3) shows that the k^{th} event depends not only on the $k-1^{\text{st}}$ event but also on the complete history. This observation is particularly important for delay management and requires to define the k^{th} power of a square matrix $\mathbf{A}$ by

$$\mathbf{A}^{\otimes k} = \underbrace{\mathbf{A} \otimes \cdots \otimes \mathbf{A}}_{k \text{ times}},$$

where $k \in \mathbb{N}_0$. In all unambiguous cases we shall simplify notation by $\mathbf{A}^k$ rather than $\mathbf{A}^{\otimes k}$. For the special case of $k = 0$, $\mathbf{A}^0$ equals the $n \times n$ *identity matrix* $\mathbf{I}$ by definition. This square matrix has zeros on the main diagonal and ε's elsewhere, i.e., for $n = 3$

$$\mathbf{A}^0 = \mathbf{I} = \begin{pmatrix} 0 & \varepsilon & \varepsilon \\ \varepsilon & 0 & \varepsilon \\ \varepsilon & \varepsilon & 0 \end{pmatrix}.$$

It is not too hard to verify that $\mathbf{A} \otimes \mathbf{A}^0 = \mathbf{A}$. The meaning of $\mathbf{A}^k$ becomes apparent when computing $\mathbf{A}^2$ for the 2×2 matrix $\mathbf{A}$.

$$\begin{aligned} \mathbf{A}^2 = \mathbf{A} \otimes \mathbf{A} &= \begin{pmatrix} a_{11} \otimes a_{11} \oplus a_{12} \otimes a_{21} & a_{11} \otimes a_{12} \oplus a_{12} \otimes a_{22} \\ a_{21} \otimes a_{11} \oplus a_{22} \otimes a_{21} & a_{21} \otimes a_{12} \oplus a_{22} \otimes a_{22} \end{pmatrix} \\ &= \begin{pmatrix} \max\{a_{11} + a_{11}, a_{12} + a_{21}\} & \max\{a_{11} + a_{12}, a_{12} + a_{22}\} \\ \max\{a_{21} + a_{11}, a_{22} + a_{21}\} & \max\{a_{21} + a_{12}, a_{22} + a_{22}\} \end{pmatrix} \end{aligned}$$

6.1 Note (Meaning of $\mathbf{T}^k$ in the railway example) When τ_{ij} denotes the trip time from v_j to v_i, then $(\mathbf{T}^2)_{ij}$ represents the maximum trip time of all paths consisting of two arcs (or paths of length 2) that start at node v_j and end at v_i. If no such path exists, then $(\mathbf{T}^2)_{ij} = \varepsilon$. With regard to Figure 6.2 one has

$$\mathbf{T}^2 = \begin{pmatrix} \max\{2+2, 5+3\} & \max\{2+5, 5+3\} \\ \max\{3+2, 3+3\} & \max\{3+5, 3+3\} \end{pmatrix} = \begin{pmatrix} 8 & 8 \\ 6 & 8 \end{pmatrix}.$$

The same reasoning applies to $\mathbf{T}^k$ with respect to paths of length k. Even $\mathbf{T}^0$ may be interpreted in this way. The consequences for journeys on a corresponding railway network may be explained best by Figure 6.2. A trip from v_1 to v_3 and going back home includes four links with pure cumulated travel times of 11. The decisive fact is that this trip time cannot be realized due to precedence constraints. The cycle (v_1, v_2, v_1) retards the journey at hand giving $(\mathbf{T}^4)_{1,1} = 16$; see Figure 6.7. This implies a total waiting time of 5 for every passenger. □

For the border case with $n = 1$ the matrix $\mathbf{A}$ reduces to a number $a \in \mathbb{R}_{\oplus}$ and we gain

$$a^{\otimes k} \equiv a^k = a \otimes \cdots \otimes a$$

and this is ka in conventional algebra. Bear in mind that the literature on max-plus algebra frequently uses a^{-1} to realize conventional subtraction although the power is only defined for $k \in \mathbb{N}_0$. The following table collects the relevant max-plus matrix operations and their conventional counterparts, where max-plus operations result from substituting $+$ and $\times$ by $\oplus$ and $\otimes$, respectively.

Table 6.2 Max-plus matrix algebra

conventional operations	operations in max-plus algebra
$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$	$\mathbf{A}\oplus\mathbf{B}=\mathbf{B}\oplus\mathbf{A}$
$(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$	$(\mathbf{A}\oplus\mathbf{B})\oplus\mathbf{C}=\mathbf{A}\oplus(\mathbf{B}\oplus\mathbf{C})$
$(\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C}=\mathbf{A}\cdot(\mathbf{B}\cdot\mathbf{C})$	$(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$
$\mathbf{A}\cdot(\mathbf{B}+\mathbf{C})=\mathbf{A}\cdot\mathbf{B}+\mathbf{A}\cdot\mathbf{C}$	$\mathbf{A}\otimes(\mathbf{B}\oplus\mathbf{C})=\mathbf{A}\otimes\mathbf{B}\oplus\mathbf{A}\otimes\mathbf{C}$

6.2 Example Two 2×2 matrices may illustrate the rules of Table 6.2.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

Addition in conventional and max-plus algebra then yields

$$\mathbf{A}+\mathbf{B} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix},$$

$$\mathbf{A}\oplus\mathbf{B} = \begin{pmatrix} a_{11}\oplus b_{11} & a_{12}\oplus b_{12} \\ a_{21}\oplus b_{21} & a_{22}\oplus b_{22} \end{pmatrix} = \begin{pmatrix} \max\{a_{11}, b_{11}\} & \max\{a_{12}, b_{12}\} \\ \max\{a_{21}, b_{21}\} & \max\{a_{22}, b_{22}\} \end{pmatrix}.$$

The corresponding matrix multiplication turns out to be

$$\mathbf{A}\cdot\mathbf{B} = \begin{pmatrix} a_{11}\cdot b_{11}+a_{12}\cdot b_{21} & a_{11}\cdot b_{12}+a_{12}\cdot b_{22} \\ a_{21}\cdot b_{11}+a_{22}\cdot b_{21} & a_{21}\cdot b_{12}+a_{22}\cdot b_{22} \end{pmatrix},$$

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\otimes b_{11}\oplus a_{12}\otimes b_{21} & a_{11}\otimes b_{12}\oplus a_{12}\otimes b_{22} \\ a_{21}\otimes b_{11}\oplus a_{22}\otimes b_{21} & a_{21}\otimes b_{12}\oplus a_{22}\otimes b_{22} \end{pmatrix}.$$

The reader may now verify that the distributivity rule as given in the fifth line of Table 6.2 holds true. □

6.2.2 Discrete Event Dynamic Systems

6.2.2.1 Timetables without Delays

Given the rules of may-plus algebra, we can now start with the analysis of dynamic systems such as the first order difference equation (6.3). Without delays there is no

need to implement buffer times by virtue of some exogenous schedule $\mathbf{d}$. In order to derive a timetable $\mathbf{t}(k)$, $k = 1, 2, \ldots$, it is therefore useful to ignore the schedule $\mathbf{d}$ for the moment and to augment the system by the initial condition $\mathbf{t}(0) = \mathbf{t}_0$. The discrete time state-space description with general arc weights incorporated by $\mathbf{W}$ is then given by

$$\begin{aligned} \mathbf{t}(k) &= \mathbf{W} \otimes \mathbf{t}(k-1) \qquad (k \geqq 1), \\ \mathbf{t}(0) &= \mathbf{t}_0. \end{aligned} \tag{6.4}$$

This (homogeneous) DEDS corresponds to a first order linear difference equation in max-plus algebra.

6.3 Remark (Systems of higher order) The system

$$\mathbf{t}(k) = \mathbf{W}_1 \otimes \mathbf{t}(k-1) \oplus \mathbf{W}_2 \otimes \mathbf{t}(k-2)$$

is of order two. It can be rewritten as a set of first order difference equations by augmenting the state space. This is a standard trick in system theory, it sets $\mathbf{y}_1(k) = \mathbf{t}(k)$ and $\mathbf{y}_2(k) = \mathbf{t}(k-1)$ so that

$$\begin{pmatrix} \mathbf{y}_1(k) \\ \mathbf{y}_2(k) \end{pmatrix} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{I} & \mathbf{E} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{y}_1(k-1) \\ \mathbf{y}_2(k-1) \end{pmatrix},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{E}$ is an $n \times n$ matrix with all elements equal to ε. □

In order to solve (6.4), recursive substitution yields

$$\mathbf{t}(k) = \mathbf{W}^k \otimes \mathbf{t}_0 \qquad k = 1, 2, \ldots \tag{6.5}$$

Suppose $\mathbf{t}(0) = \mathbf{0}$, then in accordance with Note 6.1

$$\mathrm{t}_i(k) = \max_{j=1,\ldots,n} (\mathbf{W}^k)_{ij} \qquad i = 1, \ldots, n$$

is the maximum weight of all paths of length k, starting at some node j and ending at node i. By intuition, we are interested in those processes with the highest time need because one can expect that they determine the dynamic behavior of the whole DEDS. Dually we may say that $\mathrm{t}_i(k)$ is the earliest date for the k^{th} event at node v_i when $\mathbf{t}(0) = \mathbf{0}$, and $(\mathbf{W}^k)_{ij}$ represents the fastest[7] routes of length k from any v_j to v_i. Any feasible deadline $\mathrm{d}_i(k)$ for the k^{th} event at node v_i requires therefore $\mathrm{t}_i(k) \leqq \mathrm{d}_i(k)$.

As it is common to spectral theory, the *periodic behavior* of (6.5) will be discussed on the basis of the *eigenvalues* and *eigenvectors* of the $n \times n$ square matrix $\mathbf{W}$, where both terms are redefined with regard to max-plus algebra. Suppose in analogy to linear algebra that the *spectral problem*

$$\mathbf{W} \otimes \mathbf{z} = \lambda \otimes \mathbf{z} \tag{6.6}$$

[7] Although faster routes may exist, they are forbidden by precedence constraints.

has a solution for $\mathbf{z} \neq (\varepsilon, \ldots, \varepsilon)^{\mathrm{T}}$, then λ is an eigenvalue of $\mathbf{W}$ and $\mathbf{z}$ is an n dimensional eigenvector corresponding to λ. The *spectrum* of a matrix signifies its set of eigenvalues. Observe, if (6.6) holds true, then

$$\mathbf{W}^k \otimes \mathbf{z} = \lambda^k \otimes \mathbf{z}$$

is also satisfied by virtue of the following recursive computation.

$$\mathbf{W}^2 \otimes \mathbf{z} = \mathbf{W} \otimes \mathbf{W} \otimes \mathbf{z} = \mathbf{W} \otimes \lambda \otimes \mathbf{z} = \lambda \otimes \mathbf{W} \otimes \mathbf{z} = \lambda \otimes \lambda \otimes \mathbf{z} = \lambda^2 \otimes \mathbf{z}$$

With regard to (6.5) an eigenvector $\mathbf{z}$ corresponds to an initial state $\mathbf{t}_0$ resulting in a dynamic event pattern, where λ will be shown to be the *average interdeparture time*. Starting at the initial event times $\mathbf{t}_0 = \mathbf{z}$, all subsequent events (in the absence of delays) are scheduled modulo[8] λ, that is

$$\mathbf{t}(k) = \lambda^k \otimes \mathbf{z} \qquad k = 0, 1, 2, \ldots$$

or in conventional notation

$$\mathbf{t}(k) = k\lambda + \mathbf{t}_0 \qquad k = 0, 1, 2, \ldots$$

This expression corresponds to a timetable with the information that trains depart every λ minutes as of $\mathbf{t}_0$. For example, when railway transportation starts at 6 o'clock with $\lambda = 15$, we obtain $k = 12$ at 9 o'clock.

6.4 Example Without having shown how to compute eigenvalues and eigenvectors, the railway example represented by Figure 6.2 has the eigenvalue $\lambda = 4$ and an eigenvector $\mathbf{z} = \mathbf{t}(0) = (1, 0)^{\mathrm{T}}$ so that

$$\mathbf{T} \otimes \mathbf{z} = \begin{pmatrix} 2 & 5 \\ 3 & 3 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \end{pmatrix} = 4 \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda \otimes \mathbf{z}$$

The event pattern $\mathbf{t}(k)$, $k = 0, 1, 2, \ldots$, for both nodes turns out to be

$$\begin{array}{ll} \text{first node} & \\ \text{second node} & \end{array} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 5 \\ 4 \end{pmatrix} \rightarrow \begin{pmatrix} 9 \\ 8 \end{pmatrix} \rightarrow \cdots,$$

where two vehicles ply between v_1 and v_2. The first train is scheduled $(\mathrm{t}_1(0), \mathrm{t}_2(1), \mathrm{t}_1(3)) = (1, 4, 9)$, and $(\mathrm{t}_2(0), \mathrm{t}_1(2), \mathrm{t}_2(3)) = (0, 5, 8)$ is the timetable for the second train. □

6.2.2.2 Average Interdeparture Time

The results so far obtained suggest to interpret eigenvalues as average interdeparture times. Nevertheless, the results of Example 6.4 have been found by accident rather than logical computation. This lack of reasoning can be closed by introducing one

[8] Given the time t as of t_{j0} we say that t modulo λ denotes the remainder from dividing t by λ. Hence, the term $\lambda - (t \text{ modulo } \lambda)$ gives the time until the next event occurs.

further network property. Let $\mathbf{W}$ be a square matrix and $\mathcal{G}(\mathbf{W})$ be the corresponding precedence graph. We call $\mathbf{W}$ to be *irreducible*, if the precedence graph $\mathcal{G}(\mathbf{W})$ is strongly connected whereby each ordered pair of nodes is connected by some path. This means that every passenger can reach any node wherever he starts his journey. Without going into the technical details of spectral theory, Baccelli et al. (1992), Braker (1993), de Schutter (1996) provide among others the following results with regard to the existence, uniqueness and computation of eigenvalues.

Existence Every square matrix $\mathbf{W}$ has at least one eigenvalue. If $\mathbf{W}$ happens to have more than one eigenvalue, the largest one is called the *spectral radius* $\rho(\mathbf{W})$ of $\mathbf{W}$.

Uniqueness An irreducible square matrix $\mathbf{W}$ has a unique eigenvalue and therefore

$$\lambda = \rho(\mathbf{W}).$$

Maximum cycle mean The eigenvalue of an irreducible square matrix $\mathbf{W}$ equals the maximum cycle mean of the graph $\mathcal{G}(\mathbf{W})$, that is

$$\lambda = \mu_{\max}(\mathbf{W}). \tag{6.7}$$

There are several ways to express the eigenvalue of an irreducible matrix $\mathbf{W}$. Baccelli et al. (1992, Theorem 3.23) use the immediate definition of cycle means

$$\mu_{\max}(\mathbf{W}) := \max_{\zeta} \frac{|\zeta|_w}{|\zeta|_\ell}, \tag{6.8}$$

where ζ ranges over the set of cycles of $\mathcal{G}(\mathbf{W})$. Given a cycle with maximum cycle mean, *the eigenvalue λ denotes the average travel time per arc on that cycle.* Consequently, the reciprocal value of the maximum cycle mean is referred to as the *throughput* of the system. The throughput $1/\lambda$ gives the average number of tasks that can be processed per unit of time. Gaubert, MaxPlus (1997, Theorem 11) refer to the so-called Max-Plus Perron Frobenius Theorem by which

$$\mu_{\max}(\mathbf{W}) = \bigoplus_{k=1}^{n} \left[\bigoplus_{i=1}^{n} (\mathbf{W}^k)_{ii} \right]^{1/k}. \tag{6.9}$$

In accordance with Note 6.1 the term in brackets denotes a cycle from v_i to v_i of length k with maximum weight. This corresponds to the numerator of (6.8). The length k ranges from 1 to the number of nodes n, that is the longest conceivable cycle. The power of $1/k$ corresponds to the division by the path length k (the denominator in (6.8)) in conventional notation. Hence, $\mu_{\max}(\mathbf{W})$ denotes the maximum of the mean weights of all cycles. An even older result is given in Karp (1978). Karp's theorem states

$$\mu_{\max}(\mathbf{W}) = \max_{i=1,\ldots,n} \min_{k=0,\ldots,n-1} \frac{(\mathbf{W}^n)_{ij} - (\mathbf{W}^k)_{ij}}{n-k} \tag{6.10}$$

for an arbitrary index $j \in \{1,\ldots,n\}$. In this equation, $\mathbf{W}^n$ and $\mathbf{W}^k$ are to be evaluated in the max-plus algebra; the other operations (subtraction and division) are conventional ones. A proof can also be found in Baccelli et al. (1992, pp. 47–48).

6.5 Example The railway example given in Figure 6.3 implies

$$\mathbf{T}^0 = \begin{pmatrix} 0 & \varepsilon & \varepsilon \\ \varepsilon & 0 & \varepsilon \\ \varepsilon & \varepsilon & 0 \end{pmatrix}; \quad \mathbf{T}^1 = \begin{pmatrix} 2 & 5 & \varepsilon \\ 3 & \varepsilon & 1 \\ \varepsilon & 2 & \varepsilon \end{pmatrix}; \quad \mathbf{T}^2 = \begin{pmatrix} 8 & 7 & 6 \\ 5 & 8 & \varepsilon \\ 5 & \varepsilon & 3 \end{pmatrix}; \quad \mathbf{T}^3 = \begin{pmatrix} 10 & 13 & 8 \\ 11 & 10 & 9 \\ 7 & 10 & \varepsilon \end{pmatrix}.$$

Equation (6.9) refers to the diagonal elements of $\mathbf{T}^k$, $k = 1, 2, 3$, and divides them by the respective value of k. Hence, $\lambda = \mu_{\max}(\mathbf{T}) = 8/2 = 4$. Karp's theorem (6.10) computes firstly $\mathbf{T}^3 - \mathbf{T}^0$, $\mathbf{T}^3 - \mathbf{T}^1$, and $\mathbf{T}^3 - \mathbf{T}^2$ and results in the same eigenvalue. □

To sum up we have learned that the eigenvalue for a strongly connected precedence graph determines the minimum time need so that the corresponding timetable is feasible with respect to the slowest cycle that results from the precedence constraints. Given this maximum cycle mean, the next step is to find those cycles whose cycle mean equals $\mu_{\max}(\mathbf{W})$. Such cycles are called *critical* because they are the bottlenecks for a faster timetable. The graph consisting of all critical cycles (if there happen to be more than one) is called the *critical graph* of $\mathbf{W}$ and denoted by $\mathcal{G}^c(\mathbf{W})$. An arc (v_i, v_j) is critical if it belongs to some critical cycle. Furthermore, the nodes v_i and v_j are critical. Critical nodes and arcs form the critical graph as illustrated in Figure 6.5 below. Be aware that we cannot expect the critical graph to be strongly connected. Observe also that the computation of eigenvalues does not give the corresponding critical cycles. Their computation relies heavily upon the determination of eigenvectors.

6.2.2.3 Critical Cycles

When a faster timetable is desired, trains on critical cycles have to "speed up". This can be accomplished, for instance, by adding extra trains to a line (or lines) which forms part of this critical cycle, by infrastructural changes to shorten relevant segment trip times, or by redesigning the line structure. Solutions of this type require to determine firstly the critical cycles which are characterized by the eigenvectors of the spectral problem (6.6).

This section shows that an irreducible matrix with a unique eigenvalue has at least one eigenvector. An example confirms, however, that there may be several (non proportional) eigenvectors. Each critical node corresponds to one eigenvector. Regarding eigenvectors the main result is given in Theorem 6.9 which needs the construction of two auxiliary matrices denoted by $\mathbf{W}^\diamond$ and $\mathbf{W}^+$.

The *diamond operation* for a square matrix $\mathbf{W}$ is given by

$$\mathbf{W}^\diamond := \bigoplus_{n \in \mathbb{N}_0} \mathbf{W}^n = \mathbf{W}^0 \oplus \mathbf{W}^1 \oplus \mathbf{W}^2 \oplus \cdots .$$

In comparison to the initial DEDS (6.4), the use of the diamond operation follows from the fact that $\mathbf{t} = \mathbf{W}^\diamond \otimes \mathbf{d}$ solves the *fixed point problem*

$$\mathbf{t} = \mathbf{W} \otimes \mathbf{t} \oplus \mathbf{d}. \tag{6.11}$$

This is the static version of (6.4) and it includes now the schedule $\mathbf{d}$. To prove the fixed point just substitute $\mathbf{t}$ giving

$$\mathbf{W}^\diamond \otimes \mathbf{d} = \mathbf{W} \otimes \mathbf{W}^\diamond \otimes \mathbf{d} \oplus \mathbf{d} = \underbrace{(\mathbf{W} \otimes \mathbf{W}^\diamond \oplus \mathbf{W}^0)}_{=\mathbf{W}^\diamond} \otimes \mathbf{d}.$$

The structure of (6.11) reminds us of the difference equation (6.3). Be aware, however, that we cannot set $\mathbf{T} = \mathbf{W}$. An immediate consequence of the definition of $\mathbf{W}^\diamond$ is that when all entries of $\mathbf{W}$ are positive (and that can be expected for the travel times comprised by $\mathbf{T}$), then all elements of $\mathbf{W}^\diamond$ go to infinity and leave eventually the domain for which max-plus algebra is defined. We are therefore interested in conditions for which $\mathbf{W} \in (\mathbb{R}_\oplus)^{n\times n}$ implies $\mathbf{W}^\diamond \in (\mathbb{R}_\oplus)^{n\times n}$. The subsequent theorem is taken from Gaubert, MaxPlus (1997); it characterizes the convergence of $\mathbf{W}^\diamond$ in $(\mathbb{R}_\oplus)^{n\times n}$ and yields $\mathbf{W}^\diamond$ in a finite number of steps.

6.6 Theorem (Convergence) *Let* $\mathbf{W} \in (\mathbb{R}_\oplus)^{n\times n}$. *The entries of* $\mathbf{W}^\diamond$ *belong to* $\mathbb{R}_\oplus$ *if and only if there are no cycles with positive weight in the graph of* $\mathbf{W}$. *Then,*

$$\mathbf{W}^\diamond = \mathbf{W}^0 \oplus \mathbf{W} \oplus \cdots \oplus \mathbf{W}^{n-1}.$$

The last formula suggests to compute also

$$\mathbf{W}^+ := \mathbf{W} \otimes \mathbf{W}^\diamond = \mathbf{W} \oplus \mathbf{W}^2 \oplus \cdots \oplus \mathbf{W}^n$$

as well as

$$\mathbf{W}^\diamond = \mathbf{W}^0 \oplus \mathbf{W}^+.$$

In accordance with Note 6.1 the interpretation of $\mathbf{W}^+$ is quite simple. The elements $\mathbf{W}^+_{ij}$ represent the maximum path weight for all paths leading from v_j to v_i, where the path length ranges from 1 to n. Theorem 6.6 focuses in fact on $\mathbf{W}^+_{jj}$; it precludes cycles with positive weight that are typical for the precedence graph $\mathcal{G}(\mathbf{T})$ with $\mathbf{T}$ having only positive entries, i.e., travel times. At least for an irreducible square matrix $\mathbf{W}$, this disadvantage can be corrected by constructing a normalized matrix $\underline{\mathbf{W}}$ where we "divide" every element of $\mathbf{W}$ by the maximum cycle mean $\lambda = \mu_{\max}(\mathbf{W})$; the division operator $\oslash$ denotes conventional subtraction.

$$\underline{\mathbf{W}} := \mathbf{W} \oslash \mu_{\max}(\mathbf{W}) \tag{6.12}$$

Recall that λ is the average weight per arc on every critical cycle (with maximum cycle mean) so that the division by λ results in a matrix with non-positive cycle weights, where particularly the maximum cycle mean is zero,

$$\mu_{\max}(\underline{\mathbf{W}}) = 0.$$

The division by any $b \neq \varepsilon$ can alternatively be expressed by $\mathbf{W} \oslash b = b^{-1} \otimes \mathbf{W}$ or $\tilde{e} \oslash b \otimes \mathbf{W}$, where the last term underscores that $\oslash$ is an operator rather than the sign of b.

In the railway example the normalized matrix of trip times $\underline{\mathbf{T}}$ determines a matrix $\underline{\mathbf{T}}^+$ which will be shown to include the eigenvectors of $\mathbf{T}$. These eigenvectors in turn characterize the critical graph and indicate the bottleneck for a faster timetable. The construction of $\underline{\mathbf{T}}$ is explained in the subsequent Example 6.7, while Example 6.10 shows how to derive the critical graph from $\underline{\mathbf{T}}^+$.

6.7 Example As has already been computed, the eigenvalue of $\mathbf{T}$ as given in Figure 6.3 is $\lambda = 4$. Figure 6.4 represents the corresponding precedence graph $\mathcal{G}(\underline{\mathbf{T}})$ of the normalized matrix $\underline{\mathbf{T}}$. According to (6.12) the weight of every path of length k is reduced by $k\,\lambda$.

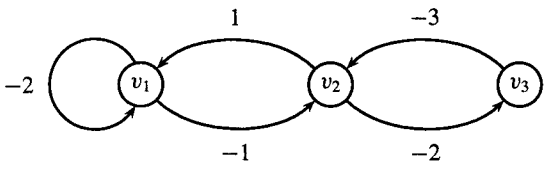


Fig. 6.4
Normalized precedence graph
The precedence graph $\mathcal{G}(\underline{\mathbf{T}})$ with

$$\mathbf{T} \oslash \lambda = \underline{\mathbf{T}} = \begin{pmatrix} -2 & 1 & \varepsilon \\ -1 & \varepsilon & -3 \\ \varepsilon & -2 & \varepsilon \end{pmatrix}$$

has again four possible links missing.

It may be helpful to verify that there are no cycles with positive weights in the graph $\mathcal{G}(\underline{\mathbf{T}})$ as needed by Theorem 6.6. □

The fixed point problem (6.11) may be seen as special case of the following theorem with $b = 0$; it is taken from Quadrat, MaxPlus (1995).

6.8 Theorem *The equation* $\beta \otimes \mathbf{t} = \mathbf{W} \otimes \mathbf{t} \oplus \mathbf{d}$ *has a least solution* $\hat{\mathbf{t}} = (\mathbf{W} \oslash \beta)^{\diamond} \otimes (\mathbf{d} \oslash \beta)$ *when the spectral radius satisfies* $\rho(\mathbf{W}) \leqq \beta$. *The solution is unique when* $\rho(\mathbf{W}) < \beta$.

The proof that $\hat{\mathbf{t}}$ solves the problem at hand follows from the same reasoning as in the sequel of (6.11). Recall that $\rho(\mathbf{W}) = \lambda$, with λ being the eigenvalue of an irreducible matrix $\mathbf{W}$. The eigenvalue of the normalized matrix $\underline{\mathbf{W}}$ is zero so that $\rho(\underline{\mathbf{W}}) = \mu_{\max}(\underline{\mathbf{W}}) = 0$. Going back to the spectral problem (6.6), i.e., $\mathbf{W} \otimes \mathbf{z} = \lambda \otimes \mathbf{z}$, we are now able to hand in the eigenvector $\mathbf{z}$ by solving the deduced fixed point problem $\underline{\mathbf{W}} \otimes \mathbf{z} = \mathbf{z}$ with the eigenvalue $\mu_{\max}(\underline{\mathbf{W}}) = 0$.

6.9 Theorem (Eigenvector)[9] *Let* $\mathbf{W}$ *be an irreducible square matrix with eigenvalue* λ. *Compute the normalized matrix* $\underline{\mathbf{W}} = \mathbf{W} \oslash \lambda$ *and then*

$$\underline{\mathbf{W}}^+ = \underline{\mathbf{W}} \oplus \underline{\mathbf{W}}^2 \oplus \cdots \oplus \underline{\mathbf{W}}^n.$$

Each column $\mathbf{z}_j$ *of* $\underline{\mathbf{W}}^+$ *such that* $\underline{\mathbf{W}}^+_{jj} = 0$ *is an eigenvector of* $\mathbf{W}$. *The eigenvector* $\mathbf{z}_j$ *is assigned to the critical node* v_j. *The matrix* $\underline{\mathbf{W}}^+$ *includes at least one such column.*

An interpretation of $\underline{\mathbf{W}}^+$ follows from the observation that $(\mathbf{W}^k)_{ij}$ denotes the maximum path weight of all paths of length k that start at node v_j and end in node v_i. The value $\underline{\mathbf{W}}^+_{ij}$ is therefore the maximum weight of all such paths but with length k

[9] The theorem is a combination of Baccelli et al. (1992, Theorem 3.23) and Braker (1993, Theorem 1.1.31).

ranging from 1 to n in the graph corresponding to $\underline{\mathbf{W}}$. An immediate consequence is that all components of $\underline{\mathbf{W}}^+$ are different from ε, because any pair (v_j, v_i) of a strongly connected graph $\mathcal{G}(\mathbf{W})$ can be connected by a path of at the most $n-1$ arcs. Observe furthermore that $\underline{\mathbf{W}}^+_{jj} \leqq 0$ due to the normalization of $\mathbf{W}$ by its maximum cycle mean λ. Hence, all nodes v_j constituting a critical cycle (with cycle mean λ) are characterized by $\underline{\mathbf{W}}^+_{jj} = 0$. In other words, if we normalize the weight of the cycle having maximum cycle mean to zero, then $\underline{\mathbf{W}}^+_{jj} = 0$ for at least one node v_j.

The idea of constructing $\underline{\mathbf{W}}^+$ follows from its definition $\underline{\mathbf{W}}^+ = \underline{\mathbf{W}} \otimes \underline{\mathbf{W}}^\diamond$ with $\rho(\underline{\mathbf{W}}) = \mu_{\max}(\underline{\mathbf{W}}) = 0$. Theorem 6.8 now states for $b = 0, \mathbf{z}_j = \mathbf{t}$ and $\mathbf{W}$ substituted by $\underline{\mathbf{W}}$ that $\mathbf{t} = \underline{\mathbf{W}}^\diamond \otimes \mathbf{d}$ solves $\mathbf{t} = \underline{\mathbf{W}} \otimes \mathbf{t} \oplus \mathbf{d}$. In the last equation the vector $\mathbf{d}$ can be omitted if its components are sufficiently small or at least $\mathbf{d} \leqq \mathbf{t}$. The $\oplus$ operation suggests indeed to consider the largest possible $\mathbf{d}$, namely, $\mathbf{d} = \mathbf{t}$, in which case we have

$$\mathbf{t} = \underline{\mathbf{W}}^\diamond \otimes \mathbf{t} \quad \text{and} \quad \underline{\mathbf{W}} \otimes \mathbf{t} \geqq \mathbf{t}.$$

Since λ is an eigenvalue of $\mathbf{W}$ by assumption (see Theorem 6.9), the last inequality must hold as an equation so that the fixed point problem becomes

$$\mathbf{t} = \underline{\mathbf{W}} \otimes \mathbf{t} = \underline{\mathbf{W}}^\diamond \otimes \mathbf{t} = \underline{\mathbf{W}}^+ \otimes \mathbf{t},$$

where the last equation results from $\mathbf{t} = \underline{\mathbf{W}} \otimes \underline{\mathbf{W}}^\diamond \otimes \mathbf{t}$. What do we know about the columns of $\underline{\mathbf{W}}^+$? From $\underline{\mathbf{W}}^\diamond = \underline{\mathbf{W}}^0 \oplus \underline{\mathbf{W}}^+$ we deduce that $\underline{\mathbf{W}}^+$ differs from $\underline{\mathbf{W}}^\diamond$ at the most by the diagonal elements $\underline{\mathbf{W}}^+_{jj}$. As $\underline{\mathbf{W}}^+_{jj} \leqq 0$, the j^{th} column of $\underline{\mathbf{W}}^+$, say $\mathbf{z}_j$, equals the j^{th} column of $\underline{\mathbf{W}}^\diamond$ if and only if $\underline{\mathbf{W}}^+_{jj} = 0$. Using $\underline{\mathbf{W}}^+ = \underline{\mathbf{W}} \otimes \underline{\mathbf{W}}^\diamond$ we obtain the statement of Theorem 6.9 from

$$\mathbf{z}_j = \underline{\mathbf{W}} \otimes \mathbf{z}_j \implies \lambda \otimes \mathbf{z}_j = \mathbf{W} \otimes \mathbf{z}_j.$$

Braker (1993, p. 16) puts it in a slightly different way. Every node v_j in a critical cycle corresponds to an eigenvector $\mathbf{z}_j$ which is the j^{th} column of $\underline{\mathbf{W}}^+$ having the element $\underline{\mathbf{W}}^+_{jj} = 0$. The maximum weight of all paths from v_j to v_j is $\mu_{\max}(\underline{\mathbf{W}}) = 0$ for any critical node v_j.

6.10 Example Computing the matrix $\underline{\mathbf{T}}^+$ for $\mathbf{T}$ as given in Figure 6.3, p. 311, yields

$$\underline{\mathbf{T}}^+ := \underline{\mathbf{T}} \oplus \underline{\mathbf{T}}^2 \oplus \underline{\mathbf{T}}^3 = \begin{pmatrix} \boxed{0} & 1 & -2 \\ -1 & \boxed{0} & -3 \\ -3 & -2 & -5 \end{pmatrix}, \qquad \begin{matrix} \uparrow & \uparrow \\ \mathbf{z}_1 & \mathbf{z}_2 \end{matrix}$$

where the eigenvectors $\mathbf{z}_1$ and $\mathbf{z}_2$ correspond to the nodes v_1 and v_2, respectively. The two eigenvectors confirm the statement of (6.6) with $\lambda = 4$ as derived in Example 6.5.

$$\mathbf{T} \otimes \mathbf{z}_1 = \lambda \otimes \mathbf{z}_1 = (4, 3, 1)^\mathsf{T}$$
$$\mathbf{T} \otimes \mathbf{z}_2 = \lambda \otimes \mathbf{z}_2 = (5, 4, 1)^\mathsf{T}$$

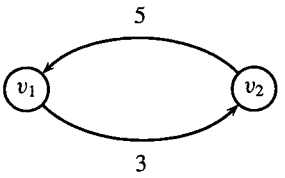


Fig. 6.5
Critical graph of $\mathbf{T}$
The critical graph $\mathcal{G}^c(\mathbf{T})$ consists of the critical nodes v_1, v_2 and a critical cycle with $\mu_{\max}(\mathbf{T}) = 8/2 = 4$.

In the example at hand the only cycle with maximum cycle mean $\mu_{\max}(\mathbf{T}) = 4$ that includes v_1 and/or v_2 is (v_1, v_2, v_1). Figure 6.5 depicts the corresponding critical graph $\mathcal{G}^c(\mathbf{T})$ although we have found the critical cycle by chance rather than consequent computation. □

The results so far obtained give timetables of the form

$$\mathbf{t}(k) = \lambda^k \otimes \mathbf{z} \qquad k = 0, 1, 2, \ldots$$

that solve the DEDS of a railway system without delays, that is

$$\begin{aligned} \mathbf{t}(k) &= \mathbf{T} \otimes \mathbf{t}(k-1) \qquad (k \geqq 1), \\ \mathbf{t}(0) &= \mathbf{t}_0. \end{aligned}$$

A corresponding eigenvalue λ indicates the average interdeparture time on critical cycles, that is the minimum time need for regularly operating timetables. The eigenvectors of the system can be used, on the one hand, to find the bottlenecks for a faster timetable. On the other hand, one eigenvector $\mathbf{z}$ is picked out to fix the initial state by $\mathbf{t}_0 = \mathbf{z}$. The next step will be to determine event series $\mathbf{t}(k)$, $k = 0, 1, 2, \ldots$, that repeat after a while in order to permit a user friendly provision of network services.

6.2.2.4 Periodically Reiterated Timetables

The eigenvalue λ and the eigenvectors determine the asymptotic behavior of $\mathbf{W}^k$ as $k \to \infty$ and this behavior passes on to $\mathbf{t}(k)$ by (6.5) with an offset of $\mathbf{t}_0$. With special regard to scheduling jobs on transport networks such as railway systems we are therefore interested in those values of $\mathbf{W}^k$ that reiterate at constant time intervals, i.e., $\mathbf{W}^k = \mathbf{W}^{k+d}$ or with an additional trend $\lambda^d \otimes \mathbf{W}^k = \mathbf{W}^{k+d}$. Such a behavior can be described as follows.

A matrix $\mathbf{W}$ is said to be *cyclic* if there are two numbers[10] k_W and d such that

$$\mathbf{W}^{k+d} = \lambda^d \otimes \mathbf{W}^k \qquad \forall k \geq k_W, \tag{6.13}$$

where the quantity λ equals the maximum cycle mean $\mu_{\max}(\mathbf{W})$. The least such d is called the *cyclicity of the matrix* $\mathbf{W}$. If a matrix $\mathbf{W}$ has cyclicity d, it is also called *order-d periodic*. Contrasted with the cyclicity of a matrix, the cyclicity of a strongly connected graph $\mathcal{G}(\mathbf{W})$ is defined by the greatest common divisor of the length of

[10] Recall that the k^{th} power of $\mathbf{W}$ has been defined only for $k \in \mathbb{N}_0$.

all its cycles. Having in mind that the critical graph needs not be strongly connected, the cyclicity of an arbitrary graph though requires a more general concept as can be found in Cassandras et al. (1995). Given such an appropriate definition, the authors note

6.11 Theorem (Cyclicity) *Any irreducible matrix is cyclic. The cyclicity of an irreducible matrix* $\mathbf{W}$ *equals the cyclicity of its critical graph* $\mathcal{G}^c(\mathbf{W})$.

As we have restricted ourselves to cases with strongly connected graphs, Theorem 3.112 in Baccelli et al. (1992) applies, by which an irreducible matrix $\mathbf{W}$ is order-d periodic if the strongly connected critical graph $\mathcal{G}^c(\mathbf{W})$ has cyclicity d. An extension of the cyclicity theorem for the reducible case is given in Gaubert, MaxPlus (1997, Theorem 17).

6.12 Example The strongly connected precedence graph $\mathcal{G}(\mathbf{T})$ in the railway system of Figure 6.3 includes a loop which is a cycle of length one. Hence, the greatest common divisor of the length of all its cycles is one. The corresponding critical graph $\mathcal{G}^c(\mathbf{T})$ has been depicted in Figure 6.5. The only (critical) cycle $\zeta = (v_1, v_2, v_1)$ is of length two, i.e., $|\zeta|_\ell = 2$. Thus the cyclicity of $\mathcal{G}^c(\mathbf{T})$ is two and $\mathbf{T}$ is order-2 periodic. Observe also $|\zeta|_w = \tau_{21} + \tau_{12} = 3 + 5$ which confirms the result $\lambda = \mu_{\max}(\mathbf{T}) = |\zeta|_w / |\zeta|_\ell = 8/2 = 4$ that has already been derived in Example 6.5.

Using the cyclicity of $\mathbf{T}$, that is $d = 2$, simple numerical computation shows for $k = 5$

$$\mathbf{T}^{5+2} = \begin{pmatrix} 24 & 23 & 22 \\ 21 & 24 & 19 \\ 21 & 20 & 19 \end{pmatrix} = \lambda^2 \otimes \mathbf{T}^5 = 8 \otimes \begin{pmatrix} 16 & 15 & 14 \\ 13 & 16 & 11 \\ 13 & 12 & 11 \end{pmatrix}.$$

Repetitive multiplication by $\mathbf{T}$ acknowledges the statement of (6.13) so that $k_W = 5$. In terms of the railway example a cyclicity d implies that – irrespective of the initial departure times $\mathbf{t}(0)$ – a periodic behavior is obtained after d periods of length λ, i.e., the average interdeparture time. □

The idea of cyclic matrices with regard to some DEDS is quite easy. Just multiply (6.13) by $\mathbf{t}(0)$, then

$$\begin{aligned} & \mathbf{W}^d \otimes \mathbf{W}^k \otimes \mathbf{t}(0) = \lambda^d \otimes \mathbf{W}^k \otimes \mathbf{t}(0) & \\ \Longleftrightarrow \quad & \mathbf{W}^d \otimes \mathbf{t}(k) = \lambda^d \otimes \mathbf{t}(k) & \text{see (6.5)} \\ \Longleftrightarrow \quad & \mathbf{t}(k+d) = d\lambda + \mathbf{t}(k) & \text{in conventional notation} \end{aligned}$$

for any $k \geqq k_W$. With special regard to the normalized matrix $\underline{\mathbf{W}}$, we know $\lambda = \mu_{\max}(\underline{\mathbf{W}}) = 0$ whereby (6.13) reduces to

$$\underline{\mathbf{W}}^{k+d} = \underline{\mathbf{W}}^k$$

for sufficiently large k. Moreover, $\lambda = 0$ results in the event pattern

$$\mathbf{t}(k+d) = \mathbf{t}(k).$$

In the railway example such behavior describes a schedule that is repeated, for instance, every hour or every day.

6.2.2.5 Synchronization

The results so far obtained omit exogenous schedules which can be used to slow down the above event pattern. Although we aim at the fastest timetable, a deceleration by buffer times may be necessary because otherwise the system would run at the risk that delays cause a breakdown of the whole schedule. In this sense a schedule $\mathbf{d}(k)$ with $k = 1, 2, \ldots$ gives permanent clock pulses to the network in order to obtain a timetable $\mathbf{t}(k)$, $k = 1, 2, \ldots$, that works on time.

The determination of an exogenous schedule $\mathbf{d}(k)$ for $k = 1, 2, \ldots$ in the DEDS

$$\begin{aligned} \mathbf{t}(k) &= \mathbf{W} \otimes \mathbf{t}(k-1) \oplus \mathbf{d}(k) \qquad (k \geqq 1), \\ \mathbf{t}(0) &= \mathbf{t}_0. \end{aligned} \tag{6.14}$$

may be seen as special case of synchronization between tasks in the following dynamic system.[11]

$$\begin{aligned} \mathbf{t}(k) &= \mathbf{W} \otimes \mathbf{t}(k-1) \oplus \mathbf{B} \otimes \mathbf{u}(k-1) \qquad (k \geqq 1) \\ \mathbf{y}(k) &= \mathbf{C} \otimes \mathbf{t}(k) \\ \mathbf{t}(0) &= \mathbf{t}_0 \end{aligned} \tag{6.15}$$

The system expands (6.14) by considering an input vector[12] $\mathbf{u} \in \mathbb{R}^m$ (multiplied by an $n \times m$ matrix $\mathbf{B}$) and an output vector $\mathbf{y} \in \mathbb{R}^p$ (multiplied by an $p \times n$ matrix $\mathbf{C}$). Given m source nodes, $\mathbf{u}(k)$ indicates the date when certain resources become available for the k^{th} time including an initial state $\mathbf{u}(0)$. Similarly, $\mathbf{y}(k)$ refers to the times when the k^{th} final products are ready. As before the numbers $\mathrm{t}_i(k)$ can be interpreted as the earliest date of the k^{th} occurrence of the event (for example the starting time of a task on a machine in manufacturing[13]). The max operator models the fact that tasks can be performed as soon as all preconditions are fulfilled.

6.13 Example It is straightforward to expand the DEDS $\mathbf{t}(k) = \mathbf{T} \otimes \mathbf{t}(k-1)$ as given in Figure 6.2 by one source u_2 and one sink y_1. The source works like a measure that gives clock pulses to the system. The sink represents outputs of node v_1 with a delay of 3 time units. The dashed circles indicate nodes which turn out to be irrelevant for the system because they are not connected, neither to v_1 nor to v_2.

This approach can also be used to synchronize processes in manufacturing systems, where several (more or less parallel) tasks are to be performed by a few resources, for instance, machines. For an appropriate formulation in max-plus algebra see van Egmond (1999). The task is the same as in project planning and control, where PERT (Program Evaluation and Review Technique) and CPM

[11] In general, $\mathbf{W}$, $\mathbf{B}$, and $\mathbf{C}$ may depend also on k or may even be stochastic, cf. Cassandras et al. (1995).

[12] The input vector $\mathbf{u}$ may be seen as exogenous clock pulses to the system, while the output vector $\mathbf{y}$ is interpreted as events emanated by the system.

[13] Precedence constraints may tolerate some reordering of jobs. Such sequencing problems have the task to find a feasible permutation of a finite collection of objects like jobs that satisfies certain conditions including precedence constraints.

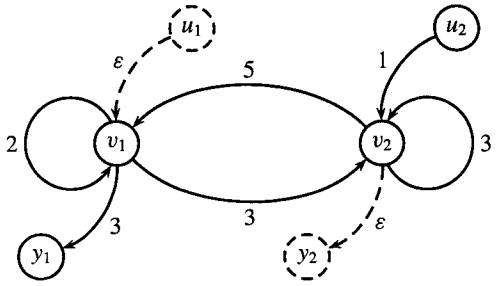


Fig. 6.6
DEDS with inputs and outputs

$$\mathbf{B} = \begin{pmatrix} \varepsilon & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 3 & \varepsilon \\ \varepsilon & \varepsilon \end{pmatrix}$$

(Critical Path Method) are frequently used procedures to determine (1) the latest time of an event to meet deadlines, (2) the earliest possible completion dates, and (3) the critical path with all "buffer times" of an activity being zero; see Hillier, Lieberman (1995) for more details. □

The system (6.15) takes into account a vector $\mathbf{u}$ that models the timing of input preconditions. There are two effects that determine the functionality of $\mathbf{u}$. When the DEDS is too slow to cope with $\mathbf{u}(k)$, it just "ignores" the input vector. This is the case where more tasks have been started than can be handled by the system. In the converse case with $\mathbf{u}(k)$ sufficiently large, the input vector works like a clock that emanates pulses to the system in order to schedule the various tasks. Going back to (6.14) with $\mathbf{d}(k) = \mathbf{B} \otimes \mathbf{u}(k-1)$, this is the reason to give $\mathbf{d}$ the interpretation of a timetable. Without delays such a schedule describes the event pattern if

$$\mathbf{t}(k) = \mathbf{d}(k) \qquad k = 0, 1, 2, \ldots$$

which requires

$$\mathbf{W} \otimes \mathbf{d}(k-1) \leqq \mathbf{d}(k) \qquad k = 1, 2, \ldots$$

An input vector $\mathbf{d}$ with these properties is referred to as *realistic timetable* and the corresponding system (6.14) is also said to be realistic.

Suppose now all events are scheduled modulo T,

$$\mathbf{d}(k) = T^k \otimes \mathbf{d}(0).$$

What is the shortest time interval for which a realistic timetable exists? Clearly, we need

$$\mathbf{W} \otimes \mathbf{t}(k-1) = \mathbf{W} \otimes \mathbf{d}(k-1) \leqq \mathbf{d}(k) = \mathbf{t}(k) \qquad k = 1, 2, \ldots$$

In the most concise case this inequality is satisfied as an equality with

$$\begin{aligned} & \mathbf{t}(k) = \mathbf{W} \otimes \mathbf{t}(k-1) \implies \mathbf{t}(k) = \mathbf{W}^k \otimes \mathbf{t}(0) \\ \text{and} \quad & \mathbf{t}(k) = \mathbf{d}(k) \implies \mathbf{W}^k \otimes \mathbf{t}(0) = T^k \otimes \mathbf{t}(0). \end{aligned}$$

The last equation corresponds to the spectral problem (6.6). Consequently, the maximum cycle mean $\mu_{\max}(\mathbf{W})$ (or the eigenvalue of $\mathbf{W}$) is the minimum time

interval T for which a realistic timetable exists. Any smaller interval forecloses $\mathbf{t}(k) = \mathbf{d}(k)$. Note also that $\mathbf{d}(0)$ must be set to some eigenvector of the system.

6.14 Example We can now close the railway example given by the DEDS (6.3). From a graph-theoretic view the eigenvalue $\lambda = \mu_{\max}(\mathbf{T})$ is the average interdeparture time of a train on the critical cycle. Viewed this way, the critical cycle is the slowest cycle in the network. The mean cycle time of this cycle determines the minimum cycle time for the entire system. Based on the numerical values as given in Figure 6.3, we have already computed $\lambda = \mu_{\max}(\mathbf{T}) = 4$, which is now used for the interarrival time T. From Example 6.10 we know two eigenvectors $\mathbf{z}_1$ and $\mathbf{z}_2$, such that there are two timetables $\mathbf{t}(k) = \mathbf{d}(k)$ starting either with $\mathbf{d}(0) = \mathbf{z}_1$ or with $\mathbf{d}(0) = \mathbf{z}_2$.

$$
\begin{array}{lccccccc}
k & 0 & & 1 & & 2 & & 3 & \cdots \\
\mathbf{d}(k) & \begin{pmatrix} 0 \\ -1 \\ -3 \end{pmatrix} & \rightarrow & \begin{pmatrix} 4 \\ 3 \\ 1 \end{pmatrix} & \rightarrow & \begin{pmatrix} 8 \\ 7 \\ 5 \end{pmatrix} & \rightarrow & \begin{pmatrix} 12 \\ 11 \\ 9 \end{pmatrix} & \rightarrow \cdots \\
\mathbf{d}(k) & \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} & \rightarrow & \begin{pmatrix} 5 \\ 4 \\ 2 \end{pmatrix} & \rightarrow & \begin{pmatrix} 9 \\ 8 \\ 6 \end{pmatrix} & \rightarrow & \begin{pmatrix} 13 \\ 12 \\ 10 \end{pmatrix} & \rightarrow \cdots
\end{array}
$$

The vector $\mathbf{d}(1) = (4, 3, 1)^{\mathsf{T}}$ indicates for the first station v_1 that the first event happens at time $t = 4$. Two trains depart, one using the loop (v_1, v_1) and one heads for v_2. It may be interesting that the latter train departs from the second station v_2 at time $t = 7$, so that two trains share the same arc (v_2, v_1) for one period of time. To see this take a look at Figure 6.7 which represents the timetable $\mathbf{t}(k)$ for $k = 1, 2, 3, 4$ with $\mathbf{t}(0) = \mathbf{z}_1 = (0, -1, -3)^{\mathsf{T}}$. In accordance with Figure 6.3 the two topmost patterns describe the two trains starting from v_1. The third and the fourth pattern refer to the trains that leave the second node v_2 and the lowest pattern indicates the departures from v_3. All trains have an interdeparture time of $T = 4$ so that the travel times on the link with link weight 5 overlap.

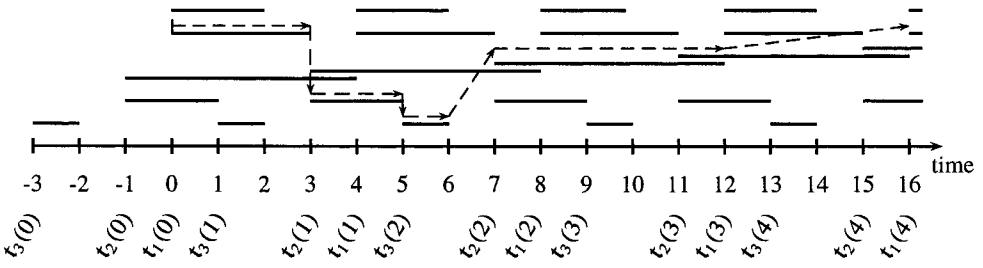


Fig. 6.7 Timetable of the precedence graph

The dashed line represents the cycle $(v_1, v_2, v_3, v_2, v_1)$ after which all trains are in the same position as before. The corresponding departure pattern is $(\mathrm{t}_1(0), \mathrm{t}_2(1), \mathrm{t}_3(2), \mathrm{t}_2(2), \mathrm{t}_1(4))$. Be aware that the last event is *not* $\mathrm{t}_1(3)$. At this point of time the cycle cannot end because the two trains commuting between v_1 and v_2 are

interchanged as compared to $t_1(0)$. Numerical computation shows indeed $(\mathbf{T}^4)_{1,1} = 16$, that is, the maximum trip time of all cycles consisting of four links that start and end at v_1 amounts to 16 time units. The last delay between $t_1(3)$ and $t_1(4)$ may be excluded from the passenger's point of view. The railway company, however, must take it into account because the same train staff cannot start the next cycle earlier. □

6.2.2.6 Stability and Sensitivity

An important issue in transportation systems is the sensitivity to delays that occur during operation. The transportation system should be able to cope with a certain amount of irregularity (delays) without severe consequences. A scheduled transport service system is stable if any initial delay disappears after a finite number of periods (subsequent departures). As an example, Figure 6.1 includes a delay of 7 minutes at Herrenberg. A useful result has been proved in Braker (1993); it refers to a DEDS of order one.

6.15 Theorem (Stability) *A system of order one as given in* (6.14) *with timetable* $\mathbf{d}(k) = T^k \otimes \mathbf{d}(0)$ *is stable if and only if*

$$\lambda < T,$$

where T is the scheduled interarrival time and the eigenvalue λ corresponds to the maximum cycle mean $\mu_{\max}(\mathbf{W})$, that is the mean interarrival time on a critical cycle.

According to $\lambda = T$, the timetable represented in Figure 6.7 has no buffer times to cope with delays so that any delay persists without chance to vanish. Of course, the larger $T - \lambda$ the faster delays disappear even though at the cost of increasing waiting times. A formal statement using $\lambda = T$ is provided by van Egmond (1998, Lemma 1):

6.16 Theorem *Let $\mathbf{W}$ be an irreducible matrix with eigenvector $\mathbf{z}$ and eigenvalue λ. Then, the maximum delay at node v_j that does not reach node v_i is $\sigma_{ij} = z_i - z_j - \underline{\mathbf{W}}^+_{ij}$.*

The idea may be explained best by starting with

$$\mathbf{t}(k) = \lambda^k \otimes \mathbf{z} = \mathbf{W}^k \otimes \mathbf{z}.$$

Division by λ^k yields $\mathbf{z} = \mathbf{W}^k \oslash \lambda^k \otimes \mathbf{z} = \underline{\mathbf{W}}^k \otimes \mathbf{z}$ and thus

$$z_i = \max_j (\underline{\mathbf{W}}^k)_{ij} + z_j.$$

For every j that does not solve the maximum problem, we know that a delay $z_i - z_j - (\underline{\mathbf{W}}^k)_{ij}$ does not reach z_j on any paths of length k. Theorem 6.16 follows from repeating the same argument for all k by which $\underline{\mathbf{W}}^k$ goes over to $\underline{\mathbf{W}}^+$. A plausibility check is to compute σ_{jj} for some critical node v_j with $\underline{\mathbf{W}}^+_{jj} = 0$; see Theorem 6.9. Hence, $\sigma_{jj} = 0$ so that any delay on a critical cycle cannot disappear.

6.17 Example It is relatively easy to check maximum delays for the precedence graph of Figure 6.3 with $\lambda = 4$. The maximum delay at v_1 and v_2 is zero as these nodes form the critical graph, see Figure 6.5. The maximum delay for v_3 should be

5 because $v_3(1) + \tau_{32} + 5 = 1 + 1 + 5 = 7 = v_2(2)$. Indeed, $\sigma_{23} = z_2 - z_3 - (\underline{\mathbf{T}}^+)_{23}$ yields the same result. □

Comprising the delays σ_{ij} by a delay matrix Σ, one can also consider buffer times $\mathbf{B}$ which improve the reliability of the system. In the easiest case $\mathbf{W}$ is substituted by $\mathbf{W} + \mathbf{B}$ where $b_{ij} = \beta$. This extends the maximum cycle mean by β, i.e., $\mu_{\max}(\mathbf{W} + \beta) = \mu_{\max}(\mathbf{W}) + \beta = \lambda + \beta$. It is not too hard to verify that the eigenvectors of $\mathbf{W} + \beta$ equal those of $\mathbf{W}$ so that the only effect is that of an extended interdeparture time. For $\beta = 0$ the average gain per pulse is $\lambda - w_{ij}$ and this is zero on the critical cycle. A positive buffer $\beta > 0$ yields $\lambda + \beta - w_{ij}$. This argument passes on to paths of length k and, consequently, the maximum delay at node v_j that does not reach node v_i becomes

$$\sigma_{ij} = z_i - z_j - (\mathbf{W} \oslash \lambda \oslash \beta)^+_{ij}.$$

Buffer times are useful, on the one hand, to obtain user friendly interdeparture time, for instance, one, two, or three departures per hour. On the other hand, the critical cycles are assigned sufficient buffer times to cope with delays on that cycle.

6.18 Example For $\beta = 0$ the maximum delays for the precedence graph of Figure 6.3 are

$$\Sigma = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 5 \end{pmatrix}$$

as indicated by Example 6.17. An additional buffer time of $\beta = 0.5$ yields

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 6 \\ 0.5 & 1 & 5.5 \\ 1 & 0.5 & 6 \end{pmatrix}.$$

□

6.19 Example A case study considering the Dutch intercity network is proposed by Braker (1993). Here we have a large scale example similar to the streetcar system of Stuttgart as given in Figure 6.1. The railway system consists of 11 lines which correspond to cycles as the trains ply between the two terminal nodes. The DEDS is of order 4 so that it must be transformed into a system of order 1 in accordance with Remark 6.3. This step results in a network with 79 nodes. The critical cycle[14] has a maximum cycle mean of $\lambda \approx 27.08$ while the trains are scheduled modulo $T = 30\,\text{min}$. Hence, the system is stable. Furthermore, Braker discusses several scenarios of delayed trains including the consequences of increased transfer times, for instance, due to maintenance work. For further discussions see also Cassandras et al. (1995) and Goverde et al. (2000). □

[14] Don't mix the critical cycle with the lines of the railway system. The critical cycle uses parts of several lines by switching trains at certain stations.

6.3 Solving the Railway Example by *MATHEMATICA*

Elementary examples of max-plus algebra are relatively easy to understand and numerical results can be computed by hand. Nevertheless, when examples grow in their extent it is useful to apply software such as *MATHEMATICA* that is capable of performing symbolic and numeric computations. Readers may use the modules below to control the numerical examples and to apply them to more complex networks.

It is advisable for the application of *MATHEMATICA* to define the zero element ε in a first step.

```
1                    ε = -Infinity;
```

Be aware not to use any operations that include $\varepsilon - \varepsilon$ until *MATHEMATICA* is told what to do with the undefined term. Otherwise the software sends back a warning message that it has found an indetermined expression.

The $\oplus$ operator is relatively easy to define, even though the module must check that the objects to be combined fit to each other. The objects `list1` and `list2` can be matrices, vectors, or scalars which is tested by the built-in functions `MatrixQ`, `VectorQ`, and `AtomQ`. Line 11 makes the operator $\oplus$ active within the *MATHEMATICA* front end.

```
1   OPlus::arg="Error: you need 2 arg's of the same type.";
2   OPlus[list1_, list2_] :=
3     If[(MatrixQ[list1] && MatrixQ[list2]),
4       MapThread[Max, {list1, list2}, 2],
5       If[(VectorQ[list1] && VectorQ[list2]),
6         MapThread[Max, {list1, list2}, 1],
7         If[(AtomQ[list1] && AtomQ[list2]),
8           Max[list1, list2],
9           Message[OPlus::arg]
10    ] ] ];
11  list1_⊕list2_ := OPlus[list1, list2];
```

MATHEMATICA is now able to compute $A \oplus B$ for properly defined lists such as a 2×2 matrix `A = {{a, b}, {c, d}}`.

The $\otimes$ operator needs similar checks of what types of objects can be combined (for the sake of brevity we omit here to define corresponding warning messages as in line 1 above).

```
1 OTimes[list1_, list2_] :=
2   If[(AtomQ[list1]||AtomQ[list2]),
3     Plus[list1, list2],
4     If[(MatrixQ[list1] && (MatrixQ[list2]||VectorQ[list2])),
5       Inner[Plus, list1, list2, Max],
6       If[(VectorQ[list1] && MatrixQ[list2]),
```

```
7           Inner[Plus, list1, Transpose[list2], Max],
8           If[VectorQ[list1],
9             Inner[Plus, list1, list2, Max],
10            Message[OTimes::arg1]
11        ] ],
12        Message[OTimes::arg2]
13    ] ];
14  list1_ ⊗ list2_ := OTimes[list1, list2];
```

The decisive lines are 5, 7, and 9, where `Inner` is a built-in function by which the two lists are added before the `Max` function is applied. Line 14 enable the $\otimes$ operator for the front end.

The subsequent power module merely simplifies notation so that we can use `OPower[A, 3]` rather than $\mathbf{A} \otimes \mathbf{A} \otimes \mathbf{A}$. The power operation is only defined for square matrices, see line 5. Furthermore, the module provides the identity matrix $\mathbf{I} = \mathbf{A}^0$ with corresponding dimensions, cf. line 9 and 10. (The definition of warning messages is again omitted.) This module uses several local variables, namely, *m*, *tmpList*, *i*, and *j*.

```
1   OPower[list_, n_] := Module[{m, tmpList, i, j},
2     If[(NumberQ[list] && NumberQ[n]),
3         n list,
4         If[(MatrixQ[list] && IntegerQ[n] && n ≥ 0),
5           If[Dimensions[list][[1]] ≠ Dimensions[list][[2]],
6             Message[OPower::matrix],
7             m = Dimensions[list][[1]];
8             If[n == 0,
9               tmpList = Table[ε, {i, m}, {j, m}];
10              Do[tmpList[[i, i]] = 0, {i, m}];
11              Return[tmpList],
12              tmpList = list;
13              Do[tmpList = OTimes[tmpList, list], {i, n - 1}];
14              Return[tmpList]
15            ] ],
16          Message[OPower::arg]
17    ] ] ];
```

The eigenvalue is computed by virtue of the Max-Plus Perron Frobenius Theorem as given in (6.9); it requires particularly a square matrix. It is not too hard to define an alternative module based on Karp's theorem; cf. (6.10). But this is only useful when *MATHEMATICA*® knows how to handle the expression $\varepsilon - \varepsilon$.

```
1   OEigenvalue[list_] := Module[{m, tmpList = list, EV, i, j},
2     m = Dimensions[list][[1]];
```

```
3     If[(MatrixQ[list] && m == Dimensions[list][[2]]),
4       EV = ε;
5       For[i = 1, i ≤ m, i++,
6         For[j = 1, j ≤ m, j++, EV = Max[EV, tmpList[[j, j]]/i]]
7         tmpList = OTimes[tmpList, list];
8       ];
9       Return[EV],
10      Message[OEigenvalue::arg]
11    ]];
```

This module returns either an eigenvalue EV of the square matrix `list` or a corresponding error message.

The computation of eigenvectors refers to the auxiliary matrix $\underline{\mathbf{W}}^+$ of $\mathbf{W}$. The next module requires an eigenvalue being computed in line 2 which must be an explicit number, see line 3. We provide a separate module for the determination of $\underline{\mathbf{W}}^+$ that can also be used to compute maximum delays.

```
1   OMatrixPlus[list_] := Module[{tmpList1, tmpList2, m, n, i},
2     n = OEigenvalue[list];
3     If[NumberQ[n],
4       tmpList1 = list - n;
5       tmpList2 = tmpList1;
6       m = Dimensions[list][[1]];
7       Do[tmpList2 = tmpList2 ⊗ (OPower[tmpList1, i]), {i, 2, m}];
8       Return[tmpList2],
9       Message[OEigenvalue::msg]
10  ]];
```

The last module `OEigenvectors` needs the eigenvalue of the square matrix $\mathbf{W} = \mathit{list}$, cf. line 3. It then computes $\underline{\mathbf{W}}^+ = \mathit{OMatrixPlus}[\mathit{list}]$ and checks for its diagonal elements, see line 8. All columns with $\underline{\mathbf{W}}^+_{jj} \neq 0$ will be removed by virtue of the `Drop` function and *MATHEMATICA* gives back a list of eigenvectors that have been checked to satisfy $\mathbf{W} \otimes \mathbf{z} = \lambda \otimes \mathbf{z}$; see line 13 with a corresponding error message invoked by line 16. The transposition is done so that $z = \mathit{OEigenvectors}[\mathit{list}][[j]]$ denotes the j^{th} eigenvector. Such an object is treated in *MATHEMATICA* "by means of plausibility", i.e., z is effectively treated as column vector in `list.z` but as row vector in `z.list`.

```
1   OEigenvectors::msg1="Error in eigenvectors: W⊗λ ≠ λ⊗z";
2   OEigenvectors[list_] := Module[{n, m, i, tmpList, Flag = 0},
3     n = OEigenvalue[list];
4     If[NumberQ[n],
5       m = Dimensions[list][[1]];
```

```
6         tmpList = Transpose[OMatrixPlus[list]];
7         For[i = m, i ≥ 1, i--,
8           If[tmpList[[i, i]] ≠ 0,
9             tmpList = Drop[tmpList, {i, i}];
10        ]];
11        m = Dimensions[list][[1]];
12        For[i = 1, i ≤ m, i++,
13          If[list ⊗ tmpList[[i]] ≠ n ⊗ tmpList[[i]],
14            Flag = i;
15        ]];
16        If[Flag ≠ 0, Message[OEigenvectors::msg1,Flag]];
17        Return[tmpList]
18    ]];
```

Of course, one can think of many other desirable tools, which are, for instance, able to compute explicitly the critical graph. Nevertheless, the modules suffice to check for our numerical results that are based on Figure 6.2 and Figure 6.3. Further tasks are left to the software users.

6.4 Economic Theory of Synchronization

6.4.1 Production Technology

6.4.1.1 Output Measurement

The results so far obtained describe the functionality of transport networks by means of DEDSs that are restricted by sets of precedence constraints. The analysis has revealed timetables that schedule transport processes particularly in consideration of delays. Furthermore, we have noted several properties of critical graphs which are the bottlenecks for a faster timetable. Nevertheless, it is hard to describe exactly what outputs are offered by the network at hand. On the demand side, passengers understand network services as journeys from an origin v_s to a destination v_t that are finalized at date x_t. The quality of service increases, the shorter the trip times are. The operator represents the supply side and measures the output of the network by the number of departures or arrivals that are scattered through space and time. Notice that this point of view lays emphasis on the number of events instead of their dates. In doing so, a trip along a cycle ζ of length $|\zeta|_\ell$ may be seen as $|\zeta|_\ell$ outputs, where the arc weights imply that different connections involve arc specific efforts. Furthermore, the carrier must take into account that each train movement is usually endowed with a given number of seats. Although capacity constraints for the trains have been left out, empty seats indicate an excess supply of transport services. In order to determine the costs per (supplied or sold) network service, it is necessary to know the underlying technology. In the present context this technology describes network outputs – that are node and time specific arrivals – by means of the only input, namely, time. The

arc specific travel times may then be interpreted as a substitute for a bundle of inputs, e.g., train staff, energy, etc., which are consumed at a constant rate.

The distinction between the number of events and their dates becomes especially apparent when transforming the previous digraphs into Petri nets. Petri nets are directed bipartite graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the set of nodes $\mathcal{V}$ is partitioned into two disjoint sets $\mathcal{V}_1$ and $\mathcal{V}_2$ such that all arcs in $\mathcal{E}$ have one node in $\mathcal{V}_1$ and the other in $\mathcal{V}_2$. The nodes of the first set are called *transitions*, while all other nodes are referred to as *places*. They are marked by boxes and circles, respectively. As an example, we refer to the digraph represented in Figure 6.2, p. 310. The two nodes (or railway stations) v_1 and v_2 become transitions and each arc (or track) is represented now by a corresponding place. The result is given by Figure 6.8. The indicated travel times τ_{ij} may be seen as (minimum) *holding times* of a place a train must spend before contributing to the next event. In principle, holding times can differ from so-called *firing times* which represent the period between the start (departure) and the completion (arrival) of an event. Similar to queuing, one reason that an event cannot be completed immediately after the holding time may be that a train cannot enter the station until a track becomes available.[15]

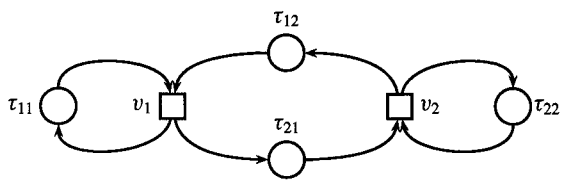


Fig. 6.8
Petri net
The precedence graph of Figure 6.2 can be transformed into a Petri net with corresponding transitions and places.

The event pattern of this Petri net has been described in (6.2), p. 312, by two evolution equations, namely,

$$\begin{aligned} \mathrm{t}_1(k) &= \max\{\tau_{11} + \mathrm{t}_1(k - \kappa_{11}), \tau_{12} + \mathrm{t}_2(k-1)\}, \\ \mathrm{t}_2(k) &= \max\{\tau_{21} + \mathrm{t}_1(k-1), \tau_{22} + \mathrm{t}_2(k-1)\}, \end{aligned} \tag{6.16}$$

where the integer κ_{11} is not necessarily one. As an example one can think of two trains that commute on the loop (v_1, v_1) so that $\mathrm{t}_1(k)$ depends on the $k - 2^{\text{nd}}$ event. System (6.16) lays emphasis on the dates when the k^{th} event occurs at one of the two nodes. Inversely, one can ask for the number of events $\mathbb{k}(t) = (\mathbb{k}_1(t), \mathbb{k}_2(t))^{\mathsf{T}}$ that take place up until time t. With regard to the loop (v_2, v_2), the number $\mathbb{k}_2(t)$ requires that the $k - 1^{\text{st}}$ departure has happened not later than τ_{22} time units ago. Similarly, $\mathbb{k}_2(t) = k$ requires $\mathbb{k}_1(t - \tau_{21}) = k - 1$ arrivals from v_1, where the last departure from v_1 takes τ_{21} time units to arrive at v_2. Summing up, we obtain a new system

$$\begin{aligned} \mathbb{k}_1(t) &= \min\{\kappa_{11} + \mathbb{k}_1(t - \tau_{11}), 1 + \mathbb{k}_2(t - \tau_{12})\}, \\ \mathbb{k}_2(t) &= \min\{1 + \mathbb{k}_1(t - \tau_{21}), 1 + \mathbb{k}_2(t - \tau_{22})\}, \end{aligned} \tag{6.17}$$

[15] For further details see Baccelli et al. (1992, Chap. 2).

where $\mathfrak{h}_j(t)$ indicates the number of departures at node v_j, $i = 1, 2$, which have taken place up to and including time t. While (6.16) has been analyzed by means of max-plus algebra, the same can be done for (6.17) by defining the corresponding min-plus algebra with $\ominus$ denoting the minimum operation, $a \ominus b := \min\{a, b\}$, (and $\ni = +\infty$ as zero element.) The relationship between (6.16) and (6.17) becomes apparent if we assume that two trains can operate on the loop (v_1, v_1), both with the same travel time τ_{11}. In this case $\kappa_{11} = 2$ indicates two potential events for the period τ_{11}. This observation is represented in (6.16) by $\tau_{11} + \mathfrak{t}_1(k - \kappa_{11})$ and in (6.17) by $\kappa_{11} + \mathfrak{h}_1(t - \tau_{11})$.

The same concept applies indeed to more general networks with κ_{ij} indicating that transition v_j is capable of dispatching κ_{ij} trains per period τ_{ij} and send these vehicles to transition v_i. The decisive word is "capable". The transition has the power to emit a maximum number of events, but it can do so only if the rest of the network is able to process these jobs. As before, let $\mathfrak{t}_i(k)$ denote the earliest time instant at which transition v_i has "fired" k times, then (6.16) takes the more general form

$$\mathfrak{t}_i(k) = \max_{j \in \mathcal{V}^-(i)} \tau_{ij} + \mathfrak{t}_j(k - \kappa_{ij}), \quad i = 1, \ldots, n, \tag{6.18}$$

where the set $\mathcal{V}^-(i)$ includes all immediate predecessor nodes of v_i. This set can be omitted when $\tau_{ij} = \varepsilon$ for all non-existing arcs. The system (6.18) is sometimes referred to as the *dater equations* (with daters $\mathfrak{t}$) in order to distinguish it from the so-called *counter equations* (with counters $\mathfrak{h}$) defined as[16]

$$\mathfrak{h}_i(t) = \min_{j \in \mathcal{V}^-(i)} \kappa_{ij} + \mathfrak{h}_j(t - \tau_{ij}), \quad i = 1, \ldots, n, \tag{6.19}$$

where $\mathfrak{h}_i(t)$ denotes the number of firings of transition v_i, $i = 1, \ldots, n$, which have taken place up to and including time t. Again the set $\mathcal{V}^-(i)$ could be omitted by using $\kappa_{ij} = \ni$ for all non-existing connections. Equations (6.18) and (6.19) describe the same underlying system and one equation is called the dual of the other. Note that k and $\mathfrak{h}_i$ are integer-valued. The functions $\mathfrak{h}_i(t)$ and $\mathfrak{t}_i(k)$ are each others inverse in a way.

One reason why we have focused on (6.18), rather than (6.19), is that the previous analysis refers to systems of order one, i.e., all κ_{ij} are equal to one. In accordance with Remark 6.3, systems of higher order with integers κ_{ij} can be reduced to systems of order one by augmenting the state space. This trick is harder to apply to (6.19) with continuous time. Nevertheless, having solved

$$\mathfrak{t}(k) = \mathbf{T} \otimes \mathfrak{t}(k-1) = \mathbf{T}^k \otimes \mathfrak{t}_0 \tag{6.20}$$

reveals also $k = \mathfrak{h}_i(t)$ with $t = \mathfrak{t}_i(k)$ for every node $i = 1, \ldots, n$. With that the network carrier knows the maximum number of "pulse beats" the network is able

[16] Mancinelli et al. (2001) use counter equations in the modeling of traffic light synchronization, where the length of green phases must be assigned such that distinct inflows can be served. Of course, junctions have bounded capacities limiting the amount of cars that can cross.

to perform in the period $[0, t]$. Observe that all nodes must work in step with the entire network. An ideal case has been described in Example 6.14 with regard to a DEDS of order one. Here every station shows the same event pattern as all other transitions with the only difference of a constant offset. Notice furthermore that each pulse takes a period λ determined by the maximum cycle mean. The inverse $1/\lambda$ gives the *throughput* representing the maximum number of pulses per unit of time.

6.20 Remark The DEDS (6.20) may also be interpreted in a reverse sense. The previous meaning has been that fast transport processes have to wait for their slower counterparts. This induces waiting times besides pure travel times. If we let link weights w_{ij} denote the accretion assigned to a movement from v_i to v_j, then $\mathbf{t}(k) = \mathbf{W} \otimes \mathbf{t}(k+1)$ selects those transitions with the maximum accretion. Gaubert, MaxPlus (1997) shows that this relationship (here with hidden control variables) corresponds to the backward dynamic programming equation in a corresponding maximum problem of optimal control. Recursive substitution yields now $\mathbf{t}_0 = \mathbf{W}^{\bar{k}} \otimes \mathbf{t}(\bar{k})$ for some prespecified event level $\bar{k}$, where $(\mathbf{W}^{\bar{k}})_{ij}$ determines the maximum reward for all paths of length $\bar{k}$. □

6.4.1.2 Time Limits

The evolution equation (6.20) exposes a further aspect underlying the present transportation problem: the $k - 1^{\text{st}}$ output determines the input of the subsequent event with daters $\mathbf{t}(k)$. The equation says that there is a mapping $\mathbf{t}(k-1) \mapsto \mathbf{t}(k)$. This suggests to ask for the inverse relationship. Is it always possible to find an evolution that concludes with a predefined final state, say $\mathbf{t}(k) = \mathbf{d}$. The answer is clearly no, but we can determine that evolution which ends up with a state $\mathbf{t}(k)$ "closest" to $\mathbf{d}$. From an economic point of view it is immediate to look for the greatest "subsolution" $\mathbf{t}(k) \leqq \mathbf{d}$ where $\mathbf{d}$ has the meaning of deadlines. Similar technical results can be obtained regarding the smallest "supersolution" satisfying $\mathbf{t}(k) \geqq \mathbf{d}$. For further details on this so-called residuation theory see Baccelli et al. (1992). We focus here on the carrier's problem to start processes that meet certain deadlines or, alternatively, on the maximum number of events that can be produced until date $\mathbf{d}$. Nevertheless, the greatest subsolution and the smallest supersolution affect also the travelers' schedule delay costs. We shall come back to this point in the subsequent Section 6.4.2.

The equation $\mathbf{W} \otimes \mathbf{t} = \mathbf{d}$ has not always a solution regarding $\mathbf{t}$, but its maximal subsolution – the so-called residual $\hat{\mathbf{t}}$ – can explicitly be computed by the *inverse problem*,[17]

$$\hat{\mathbf{t}} := \max \{\mathbf{t} | \, \mathbf{W} \otimes \mathbf{t} \leqq \mathbf{d}\}$$
$$\Longrightarrow \hat{\mathbf{t}}_j = \min_i \mathbf{d}_i - w_{ij} \qquad j = 1, \ldots, n\,.$$

The implication follows from transferring the restrictions into conventional algebra

[17] The relaxation of the equality can be formalized in terms of *residuation*, a term that is borrowed from ordered set theory.

$$\left.\begin{array}{l}\max\{w_{11}+\mathrm{t}_1,\ldots\}\leqq \mathrm{d}_1\\ \quad\vdots\\ \max\{w_{m1}+\mathrm{t}_1,\ldots\}\leqq \mathrm{d}_m\end{array}\right\}\Longrightarrow \hat{\mathrm{t}}_1=\min\{\mathrm{d}_1-w_{11},\ldots,\mathrm{d}_m-w_{m1}\}. \tag{6.21}$$

This computation defines a new[18] operator $\oslash$ which can be seen as the dual operator of $\otimes$.

$$\hat{\mathbf{t}}=\mathbf{W}\oslash\mathbf{d}=\max\left\{\mathbf{t}|\ \mathbf{W}\otimes\mathbf{t}\leqq\mathbf{d}\right\}$$

In the former dynamic system $\mathbf{t}(k)=\mathbf{W}^k\otimes\mathbf{t}_0$, the number $\mathrm{t}_j(k)$ represents the *earliest date of the k^{th} occurrence at node v_j*. Conversely, the problem $\mathbf{t}(k)\leqq\mathbf{d}$ with k held constant corresponds to computing $\hat{\mathbf{t}}_0=\mathbf{W}^k\oslash\mathbf{d}$, where $\hat{\mathrm{t}}_j(0)$ determines the *latest date to initiate the first job at node v_j*. Alternatively, one can fix $\mathbf{t}_0$ and solve for k, that is the *maximum number of events that can occur before the deadline* $\mathbf{d}$ *is reached.*

6.21 Remark (Production technology) The formal structure of (6.21) reminds us of a Leontief production technology. In this sense (6.21) might be interpreted as the underlying technology of a discrete event system. Given the "inputs" $\mathbf{d}$, the system on the left hand side suggests output efficiency, where $\hat{\mathbf{t}}$ is the maximum "output" that can be produced within the time limits $\mathbf{d}$. Similarly, the right hand side could be interpreted in the sense of input efficiency. The minimum problem seeks for the shortest periods $\mathbf{d}$ that suffice to produce $\hat{\mathbf{t}}$. In this sense the arc specific weights w_{ij} have the meaning of unavoidable disposals. In accordance with (6.21) increasing disposals require either to extend time limits $\mathbf{d}$ or to reduce $\hat{\mathbf{t}}$.

A more adequate interpretation follows from the railway example with $\mathbf{T}^k\otimes\mathbf{t}_0=\mathbf{t}(k)\leqq\mathbf{d}$. The solution $\hat{\mathbf{t}}_0$ represents the latest departure times so that all passengers who want to make trips of length k arrive in due time. In other words, the network produces trips of a specified length that arrive not later than certain specified dates. These outputs require departure times $\hat{\mathbf{t}}_0$ plus travel times (including waiting times) as inputs. The smaller the journey time $\mathrm{d}_j-\hat{\mathrm{t}}_j(0)\geqq 0$ the higher is the level of service delivered to the customers. Observe that $\mathrm{d}_j-\mathrm{t}_j(k)>0$ corresponds to the final waiting time just before d_j. Such waiting times may be valued by the passengers in the same way as pure travel times when sitting in a comfortable train or even worse due to uncomfortable waiting rooms such as cold and drafty railway stations.

The crucial point is here that the travel times $\tau_{1j},\ldots,\tau_{mj}$ are not only additively combined, but they run also parallel to each other. Series composition is essentially expressed by the $\otimes$ operation. It computes journey times by adding the corresponding sequence of arc specific travel times. In contrast, $\oplus$ takes parallel composition into account. This means all tasks have to wait for the process that finalizes last. The corresponding waiting times could only be omitted in the very limited case where all parallel jobs fit exactly to each other. In reality this perfect matching is usually suspended by precedence constraints. □

The implications of solving $\hat{\mathbf{t}}=\mathbf{W}\oslash\mathbf{d}$ for the DEDS (6.15), p. 326, can be explained best by setting $\mathbf{C}=\mathbf{I}$ (the identity matrix). Fixing some final event $k=\bar{k}$

[18] Do not mix the former $\oslash$ with $\oslash$. However, when $\mathbf{A}$ is reduced to a scalar a, then $a\oslash\mathbf{t}=\mathbf{t}\oslash a=\mathbf{t}-a$.

yields

$$\begin{aligned}\mathbf{t}(k) &= \mathbf{W} \otimes \mathbf{t}(k-1) \oplus \mathbf{d}(k) \qquad k = 1, 2, \ldots \\ \mathbf{y}(k) &= \mathbf{I} \otimes \mathbf{t}(k) \\ \mathbf{t}(0) &= \mathbf{t}_0 \\ \mathbf{y}(\bar{k}) &\leqq \mathbf{d}\end{aligned}$$

Let $\mathbf{W}$ be irreducible with eigenvector λ and set $T = \lambda$ so that $\mathbf{d}(k) = T^k \otimes \mathbf{d}(0)$, $k = 0, 1, 2 \ldots$, is a realistic timetable with $\mathbf{t}(k) = \mathbf{d}(k)$ though not stable. Then, for the predefined $\bar{k}$,

$$\mathbf{W}^{\bar{k}} \otimes \mathbf{d}(0) \leqq \mathbf{d}$$

has the following greatest subsolution (in the absence of delays)

$$\hat{\mathbf{d}}(0) = \mathbf{W}^{\bar{k}} \oslash \mathbf{d}$$

or, equivalently,

$$\hat{d}_j(0) = \min_{i=1,\ldots,n} \mathrm{d}_i - (\mathbf{W}^{\bar{k}})_{ij} \qquad j = 1, \ldots, n.$$

In effect we have to shift the event pattern as indicated by Example 6.14 to the left until a feasible solution $\mathbf{t}(\bar{k}) = \mathbf{W}^{\bar{k}} \otimes \mathbf{d}(0) \leqq \mathbf{d}$ has been found. The result gives the *minimum time needed to accomplish the deadlines* $\mathbf{d}$ *or the latest dates at which certain events must occur to finish tasks in due time.*

6.22 Remark (Systems of linear equations in max-plus algebra) To sum up, the results obtained up until now may be seen as special cases of a general system of linear equations in the sense of max-plus algebra

$$\mathbf{A} \otimes \mathbf{x} \oplus \mathbf{a} = \mathbf{B} \otimes \mathbf{x} \oplus \mathbf{b}.$$

In accordance with Gaubert, MaxPlus (1997), we have seen that such systems with unknowns $\mathbf{x}$ are frequently not solvable. Nevertheless, there is satisfactory theory to deal with the following list of subclasses of equations:

1. The inverse problem attends to $\mathbf{A} \otimes \mathbf{x} = \mathbf{b}$ by computing second best solutions.
2. The spectral problem $\mathbf{A} \otimes \mathbf{z} = \lambda \otimes \mathbf{z}$ asks for an eigenvector $\mathbf{z}$ and the eigenvalue λ of $\mathbf{A}$ which can be used to describe the evolution of a corresponding DEDS.
3. The fixed point problem $\mathbf{x} = \mathbf{A} \otimes \mathbf{x} \oplus \mathbf{a}$ has been analyzed with special regard to the determination of the critical graph. This is the bottleneck for a faster timetable.

If the operators $\ominus$ and $\oslash$ are endowed with similar properties as $\oplus$ and $\otimes$, one can derive corresponding "dual" results[19] as in the previous sections considering $\mathbf{A} \oslash \mathbf{x} \ominus \mathbf{a} = \mathbf{B} \oslash \mathbf{x} \ominus \mathbf{b}$. This is an issue that goes beyond the scope of this book. Interested readers are referred to Baccelli et al. (1992) and Quadrat, MaxPlus (1995). □

[19] As an example, Quadrat, MaxPlus (1995) compares $\beta \otimes \mathbf{x} = (\mathbf{A} \otimes \mathbf{x}) \oplus \mathbf{a}$ with $\alpha \oslash \mathbf{x} = (\mathbf{A} \oslash \mathbf{x}) \ominus \mathbf{a}$.

6.4.2 Assessment of Synchronized Timetables

6.4.2.1 Passenger Waiting Times

The assessment of synchronized timetables together with their inherent event patterns is basically a question for their goodness of fit with respect to the needs of the customers. Travelers essentially request journeys from some origin to some destination involving two aims. They want to arrive at a preferred date and the individual travel times should be at a minimum. The latter criterion has been employed in Chapter 3 with regard to road traffic to determine user optimal path flows, where route choice is left to the motorists; cf. Wardrop's first principle. In contrast to road networks, railway systems require a coordinated behavior of all movements. The individual route choice is substituted by a prespecified schedule that ties the customers to the corresponding event pattern. As an additional aspect the carrier must take into account that passengers hate to change trains including the risk to miss a connecting train or to wait on less comfortable platforms. All these details determine the quality of service customers pay for.

Given fixed arc travel times, a necessary condition for the overall value of a timetable to attain a maximum is that the total waiting times are minimized. In this sense precedence constraints balance stopping times of trains and their passengers with the disadvantage that otherwise travelers in a feeder train would have to wait for the next departure. For a detailed list of passenger waiting times see Goverde (1998b). In the present context it is enough to consider the timetable in Figure 6.7 which refers to the precedence graph $\mathcal{G}(\mathbf{T})$ as depicted in Figure 6.3. Node v_2 is particularly useful to introduce several passenger classes and their respective waiting times. Figure 6.9 supports these explanations by representing the node v_2 together with arriving and departing connections. The first train comes in via arc 1 and outgoes via arc 2, while the second train comes via arc 3 and leaves via arc 4. Without delays the regular timetable says that the first train arrives and departs, for instance, at $t = 3$. By contrast, the second train has a stop, it arrives at $t = 2$ and departs at $t = 3$.

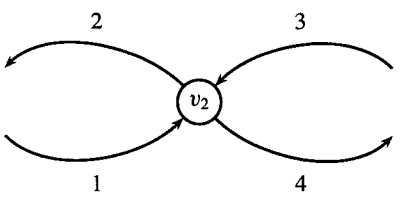


Fig. 6.9
Waiting times
The node v_2 of the precedence graph $\mathcal{G}(\mathbf{T})$ represented in Figure 6.3 suits to demonstrate four passenger classes.

In consideration of Figure 6.9 four passenger classes are to be distinguished.

(a) Passengers who start their journey at station v_2: these travelers may differ by those who know the scheduled departure times and those who do not. The first case is realistic for timetables with relatively long interdeparture times so that we can assume that these passengers arrive in due time. They have to wait only for delayed trains. The second case is plausible for timetables with

higher frequencies such as those of urban subway systems or road traffic with light control. Here the travelers arrive according to a continuous random flow. Bowman, Turnquist (1981) ascertain, however, that passenger waiting time is much more sensitive to schedule reliability and much less sensitive to service frequency than previously believed. Empirical evidence suggests that passengers coordinate their arrival times with the schedule. Any accidental behavior indeed turns out to be suboptimal as compared to passengers who are aware of the service schedule and utilize this information in an attempt to reduce their waiting time.

(b) Passengers who terminate their journey at station v_2: we usually assume that all travelers have an individually preferred arrival time which is generally incompatible with scheduled arrivals. Network users then have to assess the individual disadvantage of arriving early or late.

(c) Transit passengers must wait until all precedence constraints for a departure are satisfied. These travelers have though the amenity that they can wait within their train. Considering Figure 6.3 two cases may occur in the absence of delays.
   - Passengers in the first train have to wait for the minimum stopping time necessary for other people to enter and leave the train.
   - Persons in the second train have to wait for the first train plus a minimum time needed for changing from the first to the second train.

(d) Transferring passengers have to do the most unpleasant job.
   - Changing over from the first to the second train incurs no immediate waiting time, but switching between platforms within scarce transfer times requires additional effort.
   - Customers in the second train have no need to hasten, but less comfortably they must spend their breaks outside a train.

Observe that the actual case is based on a precedence constraint by which the second train has to wait on schedule one period for the first train. Such a constraint prevents the operator from the case where the first train arrives just after the departure of the second. As a consequence some passengers would have to wait for its next departure. Nevertheless, it is recommended to suspend the precedence constraint if hundreds of travelers in the second train are waiting for only few transferring persons.[20] Be aware that giving up the precedence constraint by splitting v_2 into two separate nodes results in a reducible precedence graph which has been ruled out in the previous analysis of regular transportation problems.

6.4.2.2 Instruments of the Carrier

Irrespective of the use of faster trains, the network provider can take advantage of several instruments in order to improve its quality of service by modifying the timetable.

[20] Schöbel (2001) aims at the minimization of the overall delay of all passengers in the transportation system. In the corresponding mixed-integer approach all link flows are weighted by the respective number of customers.

Buffer times In order to cope with ubiquitous delays the carrier has to detect proper buffer times so that the network can be expected to run on schedule. In the easiest case arc specific travel times include buffer times to consider irregularities that occur on the way. Furthermore, buffer times are necessary for travelers to alight and board the train as well as for changing trains. These processes take place within the nodes of the network. In this sense buffer times reduce the risk to catch connecting trains (for small delays), but they increase waiting times for all other passengers.

As long as delays do not exceed regular buffer times, the network is running on schedule. This case is to be contrasted with increasing delays so that the operator has to decide whether to hold or to cancel a connecting train. The corresponding synchronization control time denotes the time a transportation service waits for a delayed feeder train. Again two problems are at hand. The advantage of transferring passengers must be compared with the disadvantage of waiting passenger groups. Moreover, the operator runs the risk of ever increasing delays at subsequent stations. Thus delay propagation particularly limits the applicability of synchronization control times.

Precedence constraints The set of feasible timetables is determined by precedence constraints that can be relaxed or tightened by the carrier. A relaxation is needed, for example, when maintenance work restricts the operability of some connection. Tighter restrictions become necessary, for instance, when interconnecting two distinct networks.

Temporal variation in demand Busy hours especially demand for a higher frequency of services. In consideration of minimum headways, an increasing demand can be accommodated by repeating the timetable at a given time shift.

Network design The previous analysis has shown that the critical graph is the bottleneck for a faster timetable. Having detected the critical elements of the network indicates measures to improve the timetable. That may be additional trains on the critical cycles, infrastructure investments improving functionality, relaxation of precedence constraints, and so on.

Suppose for the moment that the network carrier has chosen some "efficient" schedule endowed with appropriate buffer times that suffice to ensure network operation according to schedule. The term "efficient" means that the overall passenger waiting time is at a minimum which requires minimal buffer times. Without excessive delays such a schedule indicates all trip times in advance; see, e.g., the timetable given in Figure 6.7, p. 328. Here every journey $(v_1, v_2, v_3, v_2, v_1)$ takes a period of 12 time units plus a final waiting time of 4 time units (this is necessary for the alternative trip starting at v_2 rather than v_1). Hence, the only variable costs are the travelers' schedule delay costs which denote the difference between preferred arrival times and on schedule arrivals. (In terms of max-plus algebra, one might think of the greatest subsolution $\mathbf{t}(k) \leqq \mathbf{d}$ and the smallest supersolution $\mathbf{t}(k) \geqq \mathbf{d}$.) Travelers with individual, most preferred arrival times are interested in those services that end up closest to their needs. As an example pick out one passenger with preferred arrival time $\bar{t}$ at station v_j. In general, we cannot expect that $\bar{t} = \mathrm{t}_j(k)$ for any k. Hence,

the customer has to choose between arriving early or late. This decision can be represented by a schedule delay cost function C^d, which is assumed here to have a piecewise linear form.

$$C^d(t|\bar{t}) = \begin{cases} \gamma_1 (\bar{t} - t) & \text{if } t \leqq \bar{t} \\ \gamma_2 (t - \bar{t}) & \text{if } t > \bar{t} \end{cases}$$

The parameters γ_1 and γ_2 represent the schedule delay cost per minute of arriving early ($t < \bar{t}$) or late ($t > \bar{t}$); see also Vickrey (1969) who applies the same concept to queuing delays at a bottleneck in road traffic. Given $\bar{t}$, there must be some interval of the form $\mathrm{t}_j(k) \leqq \bar{t} \leqq \mathrm{t}_j(k+1)$ so that the traveler chooses the $k + 1^{\text{st}}$ arrival if $\gamma_1 \left(\bar{t} - \mathrm{t}_j(k)\right) > \gamma_2 \left(\mathrm{t}_j(k+1) - \bar{t}\right)$ and the k^{th} arrival otherwise. Figure 6.10 represents the schedule delay cost function C^d for one passenger with preferred arrival time $\bar{t} = t'$ by the dashed curve. Although this trip-maker is indifferent between $\mathrm{t}_j(k)$ and $\mathrm{t}_j(k+1)$, he is assigned to the earlier event by convention. The same reasoning shows that all passengers with a preferred arrival time in the interval $]\,t'', t'\,]$ choose the arrival time $\mathrm{t}_j(k)$. Having detected the arrival time, the corresponding departure just follows from backward computation.

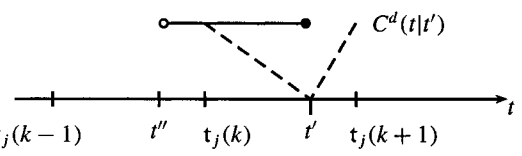


Fig. 6.10
Schedule delay costs
Any traveler with a preferred arrival time $\bar{t}$ in the marked interval $]\,t'', t'\,]$ chooses the k^{th} arrival.

It is now natural to assume identical schedule delay cost functions for all travelers except that they differ in their preferred arrival times. Moreover, trip-makers differ by their respective OD-pair. This is important if we want to model the aggregate behavior of network users. Computing delay costs requires to determine passenger groups that arrive (at constant interarrival times) at the same node via the same line. If we assume for these groups that their preferred arrival times are uniformly distributed on a given time interval, then the operator observes platoons of the same size exiting every incoming train. It should be clear that such a general approach to compute the overall schedule delay costs is at the most practicable for very simple schedules on small networks. A sufficiently simple model can be found in de Palma, Lindsey (2001) who analyze the inverse task of the carrier: given some distribution of desired arrival times, determine the optimal timetables $\mathrm{t}_j(k)$, $k = 1, 2, \ldots, K$, that minimize the overall schedule delay costs of the travelers.[21] Observe that the authors refer to a *single transit line* by which essential aspects of timed networks – that are particularly the precedence constraints – are ruled out. Nevertheless, the basic idea remains valid.

[21] The problem is formulated as an K-median problem by which K arrival dates are to be optimally distributed over a period of given length. Technically, this location problem with respect to time is the same as the p-hub median problem with respect to space. The latter has been introduced in the context of hub-and-spoke networks.

When the density of desired arrival times increases, for instance, during the busy hours, then – irrespective of capacitated vehicles – optimal timetables should provide higher throughput by reducing interarrival times. This problem is hard to solve for timed networks as compared to road traffic with untied individual motorists. The unanswered question is then what system is more robust in dealing with congestion.

Regarding the rigid schedule of Figure 6.7, the task of increasing throughput could be achieved for instant by doubling the number of trains with the same precedence constraints as before. Provided the new schedule meets minimum time headways between vehicles, the former system of order one $\mathbf{t}(k) = \mathbf{T} \otimes \mathbf{t}(k-1)$ then becomes a system of order two $\mathbf{t}(k) = \mathbf{T} \otimes \mathbf{t}(k-2)$. In accordance with Remark 6.3, p. 317, this system can be transformed into a system of order one by letting $\mathbf{y}_1(k) = \mathbf{t}(k)$ and $\mathbf{y}_2(k) = \mathbf{t}(k-1)$. Then,

$$\begin{pmatrix} \mathbf{y}_1(k) \\ \mathbf{y}_2(k) \end{pmatrix} = \begin{pmatrix} \mathbf{E} & \mathbf{T} \\ \mathbf{I} & \mathbf{E} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{y}_1(k-1) \\ \mathbf{y}_2(k-1) \end{pmatrix}.$$

If λ is an eigenvalue of $\mathbf{T}$ with an eigenvector $\mathbf{z}$, then it is natural to test $\tilde{\lambda} = \lambda/2$ and $\tilde{\mathbf{z}} = \begin{pmatrix} \mathbf{z} \\ \mathbf{z}-\lambda/2 \end{pmatrix}$ for the new systems. The reader may verify that $\tilde{\lambda}$ and $\tilde{\mathbf{z}}$ indicate indeed an eigenvalue and an eigenvector of the new system with the consequence of two timetables as represented in Figure 6.7, one of which having the offset $\lambda/2$. In other words, all we have done is to double the frequency of the systems.

The Example 6.14, p. 328, has shown that the interarrival time can be set to $\lambda = T = 4$. Hence, for a minimum headway of one period it is plausible to apply the same reasoning to a system of order 4, i.e., $\mathbf{t}(k) = \mathbf{T} \otimes \mathbf{t}(k-4)$. Using $\mathbf{y}_1(k) = \mathbf{t}(k)$, $\mathbf{y}_2(k) = \mathbf{t}(k-1)$, $\mathbf{y}_3(k) = \mathbf{t}(k-2)$, and $\mathbf{y}_4(k) = \mathbf{t}(k-3)$, yields the subsequent DEDS with 12 nodes.

$$\begin{pmatrix} \mathbf{y}_1(k) \\ \mathbf{y}_2(k) \\ \mathbf{y}_3(k) \\ \mathbf{y}_4(k) \end{pmatrix} = \begin{pmatrix} \mathbf{E} & \mathbf{E} & \mathbf{E} & \mathbf{T} \\ \mathbf{I} & \mathbf{E} & \mathbf{E} & \mathbf{E} \\ \mathbf{E} & \mathbf{I} & \mathbf{E} & \mathbf{E} \\ \mathbf{E} & \mathbf{E} & \mathbf{I} & \mathbf{E} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{y}_1(k-1) \\ \mathbf{y}_2(k-1) \\ \mathbf{y}_3(k-1) \\ \mathbf{y}_4(k-1) \end{pmatrix}$$

The solution can be computed by using the *MATHEMATICA* modules in Section 6.3 and the result shows what one might have expected. The new interarrival time follows from the eigenvalue $\tilde{\lambda} = \tilde{T} = \lambda/4 = 1$. One eigenvector turns out to be

$$\tilde{\mathbf{z}} = \begin{pmatrix} \mathbf{z} \\ \mathbf{z}-\lambda/4 \\ \mathbf{z}-\lambda/2 \\ \mathbf{z}-3\lambda/4 \end{pmatrix} = (0, -1, -3, -1, -2, -4, -2, -3, -5, -3, -4, -6)^{\mathsf{T}}.$$

This eigenvector can be used to initialize the timetable by

$$\begin{pmatrix} \mathbf{y}_1(0) \\ \mathbf{y}_2(0) \\ \mathbf{y}_3(0) \\ \mathbf{y}_4(0) \end{pmatrix} = \begin{pmatrix} \mathbf{z} \\ \mathbf{z}-\lambda/4 \\ \mathbf{z}-\lambda/2 \\ \mathbf{z}-3\lambda/4 \end{pmatrix}$$

so that the original timetable is repeated four times all with an offset of one period. That is, the frequency has been quadruplicated. At the first node, for instance, the original event $z_1 = 0$ is now substituted by $\tilde{z}_1 = \tilde{z}_4 + 1 = \tilde{z}_7 + 2 = \tilde{z}_{10} + 3 = 0$. Although these outcomes might have been found by simple reflections, it demonstrates the capability of max-plus algebra with respect to more complex problems.

To sum up, let us assume that the operator provides an efficient schedule that can be repeated within limits following from the underlying infrastructure. Irrespective of any fixed network costs and of any congestion effects, doubling the frequency clearly doubles also transport costs. This is advantageous when the corresponding savings in schedule delay costs outweigh additional transport costs. Once again it must be emphasized that this argument does not refer to any vehicle capacities. All passengers can be served at low and high frequency. Nevertheless, increasing the transport capacity may be the most important reason to raise frequency during the busy hour.

6.4.3 Cost Measurement

6.4.3.1 A Retrospect of Network Cost Evaluation

This last section works out essential properties of the cost structure of transport networks, where the reflections actually serve as an outlook for future research. In a first instance, the analysis goes back to the production technology underlying road networks as has been discussed in Chapter 3. In this setting network services focus on spatial relationships where origin-destination flows are collected by *trip tables*. Consequently, we start with basic features of cost functions that result from producing trip tables at minimum costs. Afterwards the analysis is refined with respect to incremental costs that arise from additional flows. Conditional incremental costs emphasize that the effects depend in general on the prevailing traffic conditions. Finally, the derived concepts will be augmented by aspects of time. Traffic patterns possess not only a spatial dimension but are also characterized by precedence constraints. Given some realistic *timetable*, we are then interested in the cost structure that follows from prespecified deadlines for the network activities.

In any case, cost minimal network operations require technically efficient activities that do not waste scarce resources. The first step must therefore be to describe feasible network activities that determine the transformation of network resources (inputs) into network services (outputs). In order to describe the production technology of a transport network we have started in Chapter 3 with a set of nodes $\mathcal{V}$ and a set of links $\mathcal{E}$ determining the basic ingredients of a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ possibly with all links $e \in \mathcal{E}$ directed. Aside from nodes and arcs, several classes of capacities with respect to links, nodes, vehicles, etc. restrict the production of network services. With special regard to link capacities κ let $\Phi = \mathbb{R}_+^m$ denote the *input space* of a network with m arcs. This means that when $\kappa \in \Phi$, then all link flows $\varphi \in \Phi$ are feasible if $\varphi \leqq \kappa$. Moreover, a non-existing connection e might be represented by $\kappa_e = 0$. The input space Φ is now to be confronted with the

output space $\mathcal{X}$. It includes trip tables $\mathbf{x}$ the components $x_{(s,t)} \geqq 0$ of which denote the flow between a connected OD-pair $(v_s, v_t) \in \mathcal{V} \times \mathcal{V}$. It is common practice to assume a strongly connected network such that every OD-pair is represented by one component of $\mathbf{x}$. These elements suffice to define an *output correspondence* $P\colon \Phi \to \mathfrak{P}(\mathcal{X})$ (with $\mathfrak{P}$ denoting the power set) that describes what trip tables are technically feasible given a capacitated network $(\mathcal{G}, \kappa)$. The level set

$$P(\kappa) = \left\{ \mathbf{x} \mid \mathbf{x} \text{ is a feasible trip table on the capacitated network } (\mathcal{G}, \kappa) \right\}$$

is referred to as *production possibility set*. Of course, this is a crude description of a network technology since it neglects several important aspects. For example, it says nothing about the level of service which decreases in the presence of congestion. Moreover, there is no explicit dimension of time which is necessary to characterize the interdependence of serial and parallel network processes. Nevertheless, such an output correspondence suffices to explain basic properties of transport networks.

The root of the analysis is to identify each flow unit by an individual motorist or the corresponding vehicle. Then three classes of flow variables are to be distinguished: trip tables $\mathbf{x}$, path flows $\mathbf{h}$, and link flows $\boldsymbol{\varphi}$, where the latter may be interpreted as intermediate inputs of different network processes. The relationship between these flow variables follows from decomposing OD-flows $x_{(s,t)}$ into path flows h_ρ which in turn can be composed of link flows φ_e. If we know (1) the set $\mathcal{R}(s,t)$ of all paths ρ connecting the OD-pair (v_s, v_t) and (2) the links e constituting each path ρ, then

$$x_{(s,t)} = \sum_{\rho \in \mathcal{R}(s,t)} h_\rho \quad \text{and} \quad \varphi_e = \sum_{\rho\colon e \in \rho} h_\rho$$

These formulas are essential, particularly for the distinction between consumers and providers of network services. Travelers accept only complete trips from an origin to a destination (i.e., path flows). Any failing component (i.e., a link) breaks all path flows using that link. Network users are to be confronted with network operators who are mainly interested in the operability of links and the corresponding provision of link capacities. This difference becomes more obvious below when we distinguish transport costs from costs of network provision.

The relationships between flow variables are determined by the topology of the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which can be encoded by the arc/path incidence matrix $\boldsymbol{\Delta}$ and the origin destination/path incidence matrix $\boldsymbol{\Delta}$. The preceding production possibility set then takes the form

$$P(\kappa) = \left\{ \mathbf{x} \mid \mathbf{x} = \boldsymbol{\Lambda}\, \mathbf{h},\ \boldsymbol{\Delta}\, \mathbf{h} = \boldsymbol{\varphi},\ \kappa \geqq \boldsymbol{\varphi} \geqq \mathbf{0},\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{x} \geqq \mathbf{0} \right\}.$$

Notice that $\mathbf{x} \in P(\kappa)$ requires as an intermediate step that there is a path flow $\mathbf{h}$ and a link flow $\boldsymbol{\varphi}$ being compatible with the trip table $\mathbf{x}$ and the capacities κ. The redefined feasible set reveals also one further shortcoming of such a technology. It seems as if every flow unit uses all path related links at the same time during the complete period of observation. This property is suitable for circuit switched telephone networks but

hard to justify for road networks. An auxiliary interpretation may be that each flow unit is immediately substituted by its successor.

Having determined an appropriate network technology represented by the output correspondence P, the next task is to find the minimum costs of producing a trip table $\mathbf{x}$. This objective includes transport costs $\mathbb{C}_1$ as well as costs of network provision $\mathbb{C}_2$.

$$C(\mathbf{x}) = \inf_{\varphi,\kappa} \{\mathbb{C}_1(\varphi, \kappa) + \mathbb{C}_2(\varphi, \kappa) \mid \mathbf{x} \in P(\kappa)\}$$

The preceding notation is somewhat inconvenient because we search implicitly for a feasible flow $\varphi \leqq \kappa$ that permits the provision of the trip table $\mathbf{x} \in P(\kappa)$ given the topology of the network. This problem results from the fact that link flows play the role of intermediate goods needed to transform link capacities into route flows. The two parts of the objective function depend on link flows as well as link capacities in order to incorporate further technological aspects of service provision such as congestion or maintenance costs. Assuming sufficiently large link capacities, we may ignore congestion effects, in which case it is natural to assume constant link costs $\mathbf{c}$ per flow unit. Transport costs then turn out to be $\mathbb{C}_1(\varphi) = \mathbf{c}^\top \varphi$. The analysis has shown that this approach is insufficient in many respects and especially when modeling congestion effects represented, for instance, by some speed-flow-density relationship. Congestion essentially means that the level of service – say speed – goes down with an increasing utilization of network resources κ and this corresponds to a progressive increase in link specific transport costs $\mathbf{c}(\varphi, \kappa)$ and thus in $\mathbb{C}_1(\varphi, \kappa)$. Recall that the inverse case has been justified for hub-and-spoke networks with economies of scale on the interhub connections. Overall network costs are eventually completed by the costs of network provision $\mathbb{C}_2(\varphi, \kappa)$. They include not only the annual costs of capital formation[22] due to network design (i.e., capacity expansions) but also maintenance costs that can be expected to increase with network utilization.

The cost function C features several attributes that are determined by the properties of the input correspondence P. Such regularity properties have already been introduced, see [P1]–[P8] on p. 95 ff., and they can be reused in the present context. In doing so let us suppose sufficiently large capacities such that $\varphi < \kappa$. This assumption has the advantage that the intermediate goods $\varphi \in \Phi$ become the only inputs. Moreover, it is natural to rule out congestion effects for this setting and to assume constant or even zero costs of network provision, $\mathbb{C}_2 = 0$. We thus concentrate on transport costs and define a *generalized output correspondence* $\mathbb{P}\colon \Phi \to \mathfrak{P}(\mathcal{X})$ by

$$\mathbb{P}(\varphi) = \left\{\mathbf{x} \mid \mathbf{x} = \mathbf{\Lambda}\,\mathbf{h},\ \mathbf{\Delta}\,\mathbf{h} \leqq \varphi,\ \mathbf{h} \geqq \mathbf{0},\ \mathbf{x} \geqq \mathbf{0}\right\}.$$

This production possibility set is independent of link capacities and we relax the former constraints by using $\mathbf{\Delta}\,\mathbf{h} \leqq \varphi$ rather than the strict equality. In a sense, the inequality unties generated link flows from consumed route flows and may be

[22] The static calculation at hand is difficult to apply to investment in durable physical capital. We thus refer to the concept of rental costs of infrastructure capital.

interpreted as taxi drivers searching for customers or drivers having lost their way. The intention is not to give up flow conservation, but to expose technically inefficient activities with $\mathbf{\Delta h} \leq \boldsymbol{\varphi}$ which will never be cost minimal.

In order to derive cost functions, economists are more familiar with production functions or their set valued equivalents – namely, *input correspondences*. The inverse input correspondence $\mathbb{L}\colon \mathcal{X} \to \mathfrak{P}(\Phi)$ is required to satisfy

$$\mathbf{x} \in \mathbb{P}(\boldsymbol{\varphi}) \iff \boldsymbol{\varphi} \in \mathbb{L}(\mathbf{x})$$

and the *input requirement set* $\mathbb{L}(\mathbf{x}) \subseteq \Phi$ is given by

$$\mathbb{L}(\mathbf{x}) = \left\{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \geqq \mathbf{\Delta h},\ \mathbf{\Lambda h} = \mathbf{x},\ \mathbf{h} \geqq \mathbf{0},\ \boldsymbol{\varphi} \geqq \mathbf{0} \right\}.$$

The input correspondence $\mathbb{L}$ should possess regularity attributes that are equivalent to the properties [P1]–[P8] of $\mathbb{P}$; see p. 95 ff. As can be found in Bobzin (1998), the most important properties are

Possibility of Inaction Any non-negative link flow suffices to produce an empty trip matrix, $\mathbb{L}(\mathbf{0}) = \Phi$.

No Land of Cockaigne Every non-zero trip matrix $\mathbf{x}$ requires some positive link flows, $\mathbf{0} \notin \mathbb{L}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}$.

Attainability of Each Production Every trip matrix $\mathbf{x} \in \mathcal{X}$ is supported by some feasible link flow[23] so that $\mathbb{L}(\mathbf{x}) \neq \emptyset$.

Disposability of Inputs When a non-negative link flow $\boldsymbol{\varphi}_1$ suffices to produce the trip matrix $\mathbf{x}$, i.e., $\boldsymbol{\varphi}_1 \in \mathbb{L}(\mathbf{x})$, then every flow $\boldsymbol{\varphi}_2 \in \Phi$ with $\boldsymbol{\varphi}_2 \geqq \boldsymbol{\varphi}_1$ yields also $\boldsymbol{\varphi}_2 \in \mathbb{L}(\mathbf{x})$.

Disposability of Outputs When $\boldsymbol{\varphi} \in \mathbb{L}(\mathbf{x}_1)$ then every trip matrix $\mathbf{x}_2 \in \mathcal{X}$ with $\mathbf{x}_1 \geqq \mathbf{x}_2 \geqq \mathbf{0}$ can also be produced so that $\boldsymbol{\varphi} \in \mathbb{L}(\mathbf{x}_2)$ or, equivalently, $\mathbb{L}(\mathbf{x}_1) \subseteq \mathbb{L}(\mathbf{x}_2)$.

Boundedness There is no link flow $\boldsymbol{\varphi}$ that suffices to provide an ever increasing sequence of trip matrices $\{\mathbf{x}^\nu\}$, that is more precisely, $\|\mathbf{x}^\nu\| \to \infty$ implies $\bigcap_{\nu=1}^{\infty} \mathbb{L}(\mathbf{x}^\nu) = \emptyset$.

Closedness The input correspondence $\mathbb{L}$ is a closed mapping (with closed level sets).

For the special case with constant link costs per flow unit $\mathbf{c} > \mathbf{0}$, that is, $\mathbb{C}_l(\boldsymbol{\varphi}) = \mathbf{c}^\mathsf{T}\boldsymbol{\varphi}$, the cost function defined by

$$C(\mathbf{x}) = \inf_{\boldsymbol{\varphi}} \{\mathbb{C}_l(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathbb{L}(\mathbf{x})\} \tag{6.22}$$

is equivalent the support function of $\mathbb{L}(\mathbf{x})$. Both functions show, therefore, the same properties with respect to variations of $\mathbf{c}$; see Bobzin (1998, Proposition III.7) for further details. As constant link costs are too restrictive for transport networks, let us consider more general functional forms of $\mathbb{C}_l$. We then have to ask for the minimum requirements on $\mathbb{C}_l$ and thus on C with respect to a variation of trip tables $\mathbf{x}$.

[23] This is only useful when the components of the trip matrix refer to connected OD-pairs. If no such connection exists, then $x_{(s,t)} > 0$ would imply $\mathbb{L}(\mathbf{x}) = \emptyset$. We avoid this formal problem by assuming a strongly connected network.

1. The possibility of inaction $\mathbf{x} = \mathbf{0}$ is supported by zero flows $\boldsymbol{\varphi} = \mathbf{0}$ and implies, therefore, zero transport costs, $C(\mathbf{0}) = \mathbb{C}_l(\mathbf{0}) = 0$.
2. Any trip table $\mathbf{x} \notin \mathcal{X}$ is technically infeasible, $\mathbb{L}(\mathbf{x}) = \emptyset$, and forbidden network activities are represented by infinite costs, $C(\mathbf{x}) = +\infty$. The same rule holds true for $\mathbb{C}_l$ when trying to realize an infeasible link flow $\boldsymbol{\varphi}$, for example, on a non-existing connection.
3. Additional services cannot reduce overall network costs. More precisely, any two trip tables $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ with $\mathbf{x}_1 \geqq \mathbf{x}_2$ imply $C(\mathbf{x}_1) \geqq C(\mathbf{x}_2)$. This is an immediate consequence of disposability of outputs with $\mathbb{L}(\mathbf{x}_1) \subseteq \mathbb{L}(\mathbf{x}_2)$. Similarly, transport costs $\mathbb{C}_l$ should satisfy $\mathbb{C}_l(\boldsymbol{\varphi}_1) \geqq \mathbb{C}_l(\boldsymbol{\varphi}_2)$ whenever $\boldsymbol{\varphi}_1 \geqq \boldsymbol{\varphi}_2$ and $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \Phi$.
4. The boundedness of outputs implies for every increasing sequence of trip tables $\{\mathbf{x}^\nu\} \subset \mathcal{X}$ with $\|\mathbf{x}^\nu\| \to +\infty$ that network costs approach infinity, $C(\mathbf{x}^\nu) \to +\infty$. The same property must apply to $\mathbb{C}_l$ going to infinity whenever a sequence $\{\boldsymbol{\varphi}^\nu\} \subset \Phi$ outruns any limit, $\|\boldsymbol{\varphi}^\nu\| \to +\infty$.

Further properties of the cost function C are less obvious and need detailed analysis to work out typical characteristics of networks. As an example, we refer to the additivity of the input correspondence $\mathbb{L}$. This means

$$\mathbb{L}(\mathbf{x}_1) + \mathbb{L}(\mathbf{x}_2) = \mathbb{L}(\mathbf{x}_1 + \mathbf{x}_2) \qquad \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X},$$

where the addition of two input requirement sets is defined by

$$\mathbb{L}(\mathbf{x}_1) + \mathbb{L}(\mathbf{x}_2) = \left\{ \boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2 \mid \boldsymbol{\varphi}_1 \in \mathbb{L}(\mathbf{x}_1), \boldsymbol{\varphi}_2 \in \mathbb{L}(\mathbf{x}_2) \right\}.$$

In this case one has an additive cost function satisfying

$$C(\mathbf{x}_1) + C(\mathbf{x}_2) = C(\mathbf{x}_1 + \mathbf{x}_2).$$

In the presence of congestion with corresponding capacity restrictions, additivity is not plausible because we observe then $C(\mathbf{x}_1) + C(\mathbf{x}_2) \leqq C(\mathbf{x}_1 + \mathbf{x}_2)$ (at least locally), which is equivalent to having a (locally) subadditive input correspondence; cf., e.g., Bobzin (1998). A proper redefinition of $\mathbb{L}$ capturing congestion effects should therefore satisfy $\sum_\nu \mathbb{L}(\mathbf{x}^\nu) \supseteq \mathbb{L}(\sum_\nu \mathbf{x}^\nu)$. The opposite case may be referred to as synergy or just network effect, which is claimed particularly for hub-and-spoke networks. A subadditive cost function with $C(\mathbf{x}_1) + C(\mathbf{x}_2) \geqq C(\mathbf{x}_1 + \mathbf{x}_2)$ requires here a superadditive input correspondence holding $\mathbb{L}(\mathbf{x}_1) + \mathbb{L}(\mathbf{x}_2) \subseteq \mathbb{L}(\mathbf{x}_1 + \mathbf{x}_2)$.

6.4.3.2 Prospective Approaches of Measuring Transport Costs

In fact these forms of additivity reflect how network processes interact, where we have to distinguish at least between parallel, sequential, or just independent activities. This issue may be explained best by decomposing the input requirement set $\mathbb{L}(\mathbf{x})$ into a set of feasible *path flows*

$$\mathcal{H}(\mathbf{x}) = \left\{ \mathbf{h} \mid \Lambda\, \mathbf{h} = \mathbf{x},\ \mathbf{h} \geqq \mathbf{0} \right\}$$

and the underlying sets of *link flows*

$$\mathbb{L}^{\updownarrow}(\mathbf{h}) = \left\{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi} \geqq \Delta \mathbf{h},\ \boldsymbol{\varphi} \geqq \mathbf{0} \right\}.$$

Compared with $\mathbb{L}$, we refer to $\mathbb{L}^{\updownarrow}$ as *generalized input correspondence*. The original set can then be regained via

$$\mathbb{L}(\mathbf{x}) = \bigcup_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} \mathbb{L}^{\updownarrow}(\mathbf{h}).$$

It might be useful to interpret $\boldsymbol{\varphi}$ as "potential" link flows which are bounded by link capacities $\boldsymbol{\kappa}$ such that $\boldsymbol{\varphi} \leqq \boldsymbol{\kappa}$. Without congestion effects and perfectly divisible capacity units, the minimum capacities satisfying $\boldsymbol{\kappa} = \boldsymbol{\varphi} = \Delta \mathbf{h}$ then determine transport costs $\mathbb{C}_{\mathrm{I}}(\boldsymbol{\varphi})$. Observe, however, that capacities signify in general fixed inputs so that $\mathbb{C}_{\mathrm{I}}(\boldsymbol{\kappa})$ with excess capacities $\boldsymbol{\kappa} > \boldsymbol{\varphi}$ would not denote realized transport costs but the costs of potential transportation.

We now turn our attention to the generalized input correspondence $\mathbb{L}^{\updownarrow}$ which transforms link flows $\boldsymbol{\varphi}$ into path flows $\mathbf{h}$ representing the network services. As long as we stick to the above definition of generalized input requirement sets, additivity of $\mathbb{L}^{\updownarrow}$ is ensured. This is not necessarily true for more general input correspondences where link flows $\boldsymbol{\varphi}$ (e.g., number of trains) and path flows $\mathbf{h}$ (e.g., number of passengers) are correlated by a more complex interrelationship. Given an input correspondence $\mathbb{L}^{\updownarrow}$ it is natural to define a *generalized cost function* $\mathbb{K}$ by letting

$$\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h})\right) := \inf_{\boldsymbol{\varphi}} \left\{ \mathbb{C}_{\mathrm{I}}(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathbb{L}^{\updownarrow}(\mathbf{h}) \right\}.$$

6.23 Notational Remark The notation $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h})\right)$ stresses that the functional properties of the cost function $\mathbb{K}$ depend on the input correspondence $\mathbb{L}^{\updownarrow}$. This relationship draws attention to input requirement sets and allows to consider terms like $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cup \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$ or $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$. Alternatively, one might use $\mathbb{K}(\mathbf{h})$ rather than $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h})\right)$ as it is common practice in economics; see, for instance, (6.22). The new notation emphasizes outputs rather than input requirement sets. In this case $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cup \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$ could be noted as

$$\mathbb{K}(\mathbf{h}_a \vee \mathbf{h}_b) = \inf_{\boldsymbol{\varphi}} \left\{ \mathbb{C}_{\mathrm{I}}(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cup \mathbb{L}^{\updownarrow}(\mathbf{h}_b) \right\},$$

which is quite intuitive. But substituting $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$ by the notation

$$\mathbb{K}(\mathbf{h}_a \wedge \mathbf{h}_b) = \inf_{\boldsymbol{\varphi}} \left\{ \mathbb{C}_{\mathrm{I}}(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b) \right\},$$

is misleading as we look for some link flow $\boldsymbol{\varphi}$ that leaves the choice open between $\mathbf{h}_a$ and $\mathbf{h}_b$. Such a choice does not mean though that both path flows can be provided at the same time as it is suggested by $\mathbf{h}_a \wedge \mathbf{h}_b$. □

The choice between two *alternative*[24] path flows, say either $\mathbf{h}_a$ *or* $\mathbf{h}_b$, can now be represented by

[24] In general the two path flows exclude each other when focusing on link flows per period. If we refer to potential link flows, say capacities $\boldsymbol{\kappa}$, then both path flows might be realized one after the other by repeating $\boldsymbol{\varphi} \in \mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)$.

$$\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cup \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) = \inf_{\mathbf{h}\in\{\mathbf{h}_a,\mathbf{h}_b\}} \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h})\right). \tag{6.23}$$

Clearly, this decision is only sensible if the two flows $\mathbf{h}_a$ and $\mathbf{h}_b$ are comparable, for otherwise $\mathbf{h} = \mathbf{0}$ would dominate all other path flows. One important case where the operator is indifferent between $\mathbf{h}_a$ and $\mathbf{h}_b$ is $\mathbf{h}_a, \mathbf{h}_b \in \mathcal{H}(\mathbf{x})$; both flow patterns support the same trip table $\mathbf{x}$. The preceding cost minimization problem then represents the optimal choice between all link and path flows yielding a given trip table $\mathbf{x}$. Notice also that an observed link flow $\boldsymbol{\varphi}$ does not permit the conclusion of a unique path flow $\mathbf{h}$. A graphical representation of the above OR-statement is given in Figure 6.11. The alternative path flows h_{1a} and h_{2b} represent the two cases $\mathbf{h}_a = (h_{1a}, 0)$ and $\mathbf{h}_b = (0, h_{2b})$ such that the carrier is indifferent between these alternatives if $h_{1a} = h_{2b}$.

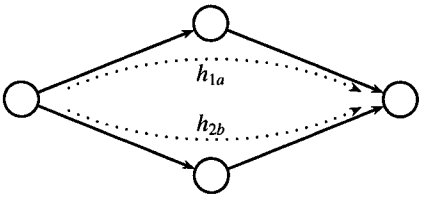


Fig. 6.11
Alternative flows
Realize h_{1a} *or* h_{2b}, provided that $h_{1a} = h_{2b}$.

The actual cost function C can now be derived from $\mathbb{K}$ by a final step, namely,

$$\begin{aligned} C(\mathbf{x}) &= \inf_{\boldsymbol{\varphi}} \{\, \mathbb{C}_{\mathrm{l}}(\boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathbb{L}(\mathbf{x}) \} \\ &= \mathbb{K}\left(\cup_{\mathbf{h}\in\mathcal{H}(\mathbf{x})} \mathbb{L}^{\updownarrow}(\mathbf{h})\right) \\ &= \inf_{\mathbf{h}} \left\{ \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h})\right) \mid \mathbf{h} \in \mathcal{H}(\mathbf{x}) \right\}. \end{aligned} \tag{6.24}$$

Aside from the OR-statement, the more interesting case is to determine a corresponding AND-statement by which we search for some cost minimal link flow $\boldsymbol{\varphi}$ that permits the provision of $\mathbf{h}_a$ *and* $\mathbf{h}_b$. It is reasonable to declare two path flows to be *independent* whenever the production of $\mathbf{h}_a$ *and* $\mathbf{h}_b$ at the same time yields additivity of their costs:

$$\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)\right) + \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) \tag{6.25}$$

Bear in mind that a realized link flow $\boldsymbol{\varphi} \in \mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)$ does in general not mean that $\mathbf{h}_a$ and $\mathbf{h}_b$ can be realized at the same time, but that the choice between $\mathbf{h}_a$ and $\mathbf{h}_b$ is kept open. Independency dispenses with this distinction which is particularly useful for an additive input correspondence satisfying $\mathbb{L}^{\updownarrow}(\mathbf{h}_a) + \mathbb{L}^{\updownarrow}(\mathbf{h}_b) = \mathbb{L}^{\updownarrow}(\mathbf{h}_a + \mathbf{h}_b)$. In this case $\boldsymbol{\varphi} = \boldsymbol{\Delta}(\mathbf{h}_a + \mathbf{h}_b)$ is also an element of the intersection $\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)$ due to the disposability of outputs given on p. 348. A more appealing interpretation would be to focus on link capacities $\boldsymbol{\kappa}$ rather than link flows $\boldsymbol{\varphi}$ so that $\mathbf{h}_a$ and $\mathbf{h}_b$ can be realized not necessarily at the same time but one after the other. The path flows are then independent over time.

Independency states here that "alternatives" do not exclude each other but can be realized at the same time with additive costs. Figure 6.12 represents this case by two parts of a common network that are not interconnected and thus independent of each other. On road networks without congestion effects, this AND-statement would

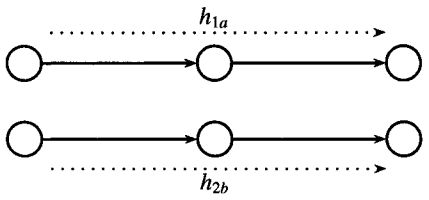


Fig. 6.12
Independent flows
Realize two independent flows h_{1a} *and* h_{2b}.

also hold true for overlapping path flows as given in Figure 6.13. The flow φ_3 on the last link is then given by $\varphi_3 = h_{1a} + h_{2b}$ and the corresponding transport costs are $c_3\varphi_3 = c_3h_{1a} + c_3h_{2b}$.

The easy case of independency is now to be continued with the crucial point of interacting flows that suspend additivity of network costs. One example follows from congestion effects. Another example is indicated by railway networks where passenger groups are bundled by trains so that transport costs depend on the number of trains and only mediately on the number of passengers. Synchronized timetables are just another example for network processes that interact also over time. If we want to provide the choice between two flows $\mathbf{h}_a$ *and* $\mathbf{h}_b$, *given* $\mathbf{h}_b$, then

$$\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) + \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)|\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right), \tag{6.26}$$

where $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)|\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$ is referred to as *conditional incremental cost.* This can definitely be expected to determine the core element of pricing network services because incremental (marginal) flows can be expressed this way. Consistency with (6.25) requires for independent flows that

$$\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)|\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)\right).$$

If an additional flow $\Delta\mathbf{h} = \mathbf{h}_a$ is independent of the prevailing flow $\mathbf{h}_b$, then incremental costs are $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\Delta\mathbf{h})\right)$. This is the easiest case with regard to pricing network services.

The use of conditional incremental costs may be explained best by the example given in Figure 6.13. The two path flows $\mathbf{h}_a = (h_{1a}, 0)$ and $\mathbf{h}_b = (0, h_{2b})$ overlap on the last connection. The additivity result for independent flows has been explained above. By contrast, congestion is reflected by $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)|\mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right) > \mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a)\right)$. The opposite inequality is reasonable, for instance, for two passenger groups represented by h_{1a} and h_{2b} that use a common train on the last link. Such bundling of traffic is omnipresent in many transport networks and particularly airline passenger carriers take advantage of this instrument.

Some plausibility checks for the generalized cost function in conjunction with its conditional incremental costs sustain the power of that concept. A flow $\mathbf{h}_a$ that excludes the provision of $\mathbf{h}_b$ implies $\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b) = \emptyset$ and thus $\mathbb{K}\left(\mathbb{L}^{\updownarrow}(\mathbf{h}_a) \cap \mathbb{L}^{\updownarrow}(\mathbf{h}_b)\right)$

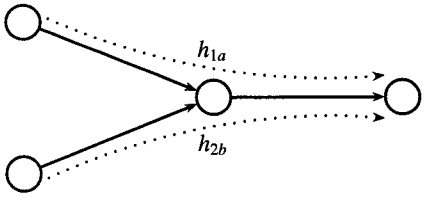


Fig. 6.13
Interdependent flows
Realize two interdependent flows h_{1a} *and* h_{2b}.

$= +\infty$. This is the case of Figure 6.13 if it is assumed that either h_{1a} or h_{2b} can be realized but not both.

The two equations (6.25) and (6.26) are statements on incremental costs if the path flows $\mathbf{h}_a$ is augmented by $\mathbf{h}_b$. One special case arises from the reverse view where some path flow $\mathbf{h}_a$ is decomposed in a smaller flow $\mathbf{h}_b$ and a residuum $\Delta\mathbf{h}$. Disposability of outputs then yields a non-decreasing cost function satisfying

$$\mathbf{h}_a \geqq \mathbf{h}_b \implies \mathbb{L}^{\dagger}(\mathbf{h}_a) \subseteq \mathbb{L}^{\dagger}(\mathbf{h}_b) \implies \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a)\right) \geqq \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_b)\right).$$

At the same time one concludes that[25]

$$\mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a) \cap \mathbb{L}^{\dagger}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a)\right).$$

This is compatible with (6.25) only if $\mathbf{h}_b = \mathbf{0}$ and $\mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{0})\right) = 0$. In all other cases the two path flows cannot be independent such that (6.26) gives the conditional incremental costs

$$\mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a)|\mathbb{L}^{\dagger}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a)\right) - \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_b)\right) \geqq 0$$

or

$$\mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_b)|\mathbb{L}^{\dagger}(\mathbf{h}_a)\right) = 0.$$

These statements must be interpreted with caution as we do not talk about $\mathbf{h}_a + \mathbf{h}_b$. Any realized link flow $\boldsymbol{\varphi}$ that supports a prevailing path flow $\mathbf{h}_a$ suffices also to provide the smaller path flow $\mathbf{h}_b$ (but not necessarily $\mathbf{h}_a + \mathbf{h}_b$) at zero conditional incremental costs. This result may be explained best if we substitute the link flow vector $\boldsymbol{\varphi}$ by link capacities $\boldsymbol{\kappa}$. On the assumption that the carrier supplies $\boldsymbol{\kappa}$ in the sense of potential flows at given transport cost, any flow pattern within these limits can be performed at zero conditional incremental costs. As an example, one might think of train capacities $\boldsymbol{\kappa}$; as long as there are empty seats any additional passenger can be carried at zero incremental transport cost. The two path flows can now be added having in mind that they must have different date stamps. These observations conform with the results of *peak-load pricing* where different prices are to be determined for peak and off-peak periods. In general, peak-load pricing separates capacity and operating costs and allocates the marginal opportunity cost of capacity to peak users while off-peak users pay merely marginal operating costs; cf., e.g., Wiseman (1987).

In summary form we have derived several results on the effects of additional flows. A sum of flows, say $\mathbf{h}_a + \mathbf{h}_b$, may conversely be seen as an increment $\mathrm{d}\mathbf{h} =$

[25] It is worthwhile to compare this equation with (6.23) by which $\mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a) \cup \mathbb{L}^{\dagger}(\mathbf{h}_b)\right) = \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_b)\right) \leqq \mathbb{K}\left(\mathbb{L}^{\dagger}(\mathbf{h}_a)\right)$.

$\mathbf{h}_b \geqq \mathbf{0}$ of the prevailing flow $\mathbf{h} = \mathbf{h}_a$. The *costs for the incremented flow* always turn out to be

$$\mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h}) \cap \mathbb{L}^{\downarrow}(\mathrm{d}\mathbf{h})\right) = \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})\right).$$

Incremental costs, however, depend on the fact whether flows interact or not. The latter case leads to additive cost shares

$$\mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})\right) = \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h})\right) + \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathrm{d}\mathbf{h})\right)$$

so that (conditional) incremental costs can be computed as

$$\mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})|\mathbb{L}^{\downarrow}(\mathbf{h})\right) = \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})\right) - \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h})\right) = \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathrm{d}\mathbf{h})\right) \geqq 0.$$

In comparison the last equation cannot be expected to hold true for interdependent flows, where conditional incremental costs turn out to be

$$\begin{aligned}\mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})|\mathbb{L}^{\downarrow}(\mathbf{h})\right) &= \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h}) \cap \mathbb{L}^{\downarrow}(\mathrm{d}\mathbf{h})\right) - \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h})\right) \\ &= \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h}+\mathrm{d}\mathbf{h})\right) - \mathbb{K}\left(\mathbb{L}^{\downarrow}(\mathbf{h})\right).\end{aligned}$$

This last relationship is not new but just a restatement of (6.26).

6.24 Remark Quadrat, MaxPlus (1995) and Akian (1995) suggest similar cost measures which follow a formalism analogous to probability calculus adapted to optimization. This approach needs a decision space Φ with subsets $A_1, A_2, \ldots$ that correspond to the above input requirement sets. The cost function $\mathbb{K}$ can now be defined in the same way as before by using a "cost density" $\mathbb{C}$. In accordance with max-plus algebra, let $\ominus$ denote the minimum operation and let $\otimes$ signify addition in conventional algebra. The above reasoning then suggest the following list of basic rules.[26]

OR-statement (6.23): $\mathbb{K}(A_1 \cup A_2) = \mathbb{K}(A_1) \ominus \mathbb{K}(A_2)$
AND-statement (6.26): $\mathbb{K}(A_1 \cap A_2) = \mathbb{K}(A_1|A_2) \otimes \mathbb{K}(A_2)$
Independency (6.25): A_1 and A_2 are said to be independent if $\mathbb{K}(A_1 \cap A_2) = \mathbb{K}(A_1) \otimes \mathbb{K}(A_2)$.

These axioms determine not only properties of the generalized cost function $\mathbb{K}$ but also of the actual cost function C by virtue of (6.24). Furthermore, network analysts are interested in the cost minimal link and path flows that are feasible with regard to a given trip table. □

6.4.3.3 The Railway Example Revisited

One may now ask how to apply the above cost measurement to synchronized timetables with fixed service or travel times on all arcs which are to be completed by

[26] Probability theory refers to similar rules, where $P(A)$ denotes the probability of an event A. The addition rule is $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. The conditional probability of A_1 given A_2 is $P(A_1|A_2) = P(A_1 \cap A_2)/P(A_2)$. Two events, A_1 and A_2, are said to be independent if and only if $P(A_1 \cap A_2) = P(A_1)\,P(A_2)$.

Every time lag represents, on the one hand, a waste of time for waiting processes. On the other hand, they are needed to finalize other processes in order to continue on schedule. Let us assume that time is the only input needed to produce network services; this can be justified when all other inputs are consumed at a constant rate of time. We then have to ask how to compute the overall time (or the corresponding cost) that is needed to provide numerous independent as well as interacting network processes.

According to Figure 6.13 above, the task can be explained best by Figure 6.14 which represents a distinct part of a network with four nodes $\mathcal{V} = \{v_1, v_2, v_3, v_4\}$ and three links $\mathcal{E} = \{e_1, e_2, e_3\}$. The link travel times τ_e, $e = 1, 2, 3$, are constants and they may be interpreted as distances between nodes provided all trains drive at a constant speed. The travel times are to be completed by waiting times σ_e, $e = 1, 2, 3$, which are determined by the schedule that synchronizes services. In order to keep

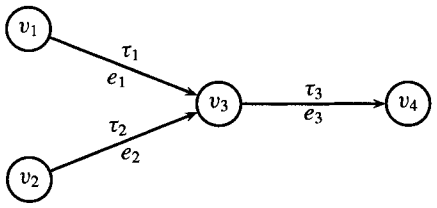


Fig. 6.14
Flows on a railway network
Railway networks are characterized by interacting processes that are to be performed in series or at the same time.

explanations as simple as possible we abstain here from synchronization control times that are necessary when a feeder train is delayed and all connecting trains have to wait up to a certain prespecified time limit. Taking such delays into account requires deep insight into delay propagation as discussed in Goverde (1998b). Furthermore, let us abstain from transit trips that enter the subnetwork at one node and leave it at some other node. Now the attention can be restricted to five classes of routes with path specific services h_ρ, $\rho = 1, \ldots, 5$, and corresponding sojourn times $\tilde{\tau}_\rho$ as given in the following table.

Table 6.3 Service sojourn times

service	route	nodes	links	sojourn time
h_1	ρ_1	$= (v_1, v_3)$	$= e_1$	$\tilde{\tau}_1 = \tau_1$
h_2	ρ_2	$= (v_2, v_3)$	$= e_2$	$\tilde{\tau}_2 = \tau_2$
h_3	ρ_3	$= (v_3, v_4)$	$= e_3$	$\tilde{\tau}_3 = \tau_3$
h_4	ρ_4	$= (v_1, v_3, v_4)$	$= (e_1, e_3)$	$\tilde{\tau}_4 = \tau_1 + \sigma_1 + \tau_3$
h_5	ρ_5	$= (v_2, v_3, v_4)$	$= (e_2, e_3)$	$\tilde{\tau}_5 = \tau_2 + \sigma_2 + \tau_3$

Observe that a train on the third link is capable of supporting three services – namely, h_3, h_4, and h_5 – at the same time. This sharing of network resources by different services is one of the most important reasons for synchronization.

The next step is to define an appropriate technology that describes feasible network activities. Such a pattern transforms travel times τ and waiting times σ into service bundles $\mathbf{h}$ by virtue of a realistic timetable $\mathbf{d}(k)$, $k = 1, 2, \ldots$ The idea has been explained in Section 6.2.2 with regard to a system of order one and the precedence matrix $\mathbf{T}$ comprising the link specific travel times. A realistic timetable then generates an event pattern $\mathbf{t}$ of the form

$$\mathbf{T} \otimes \mathbf{t}(k-1) \leqq \mathbf{d}(k) = \mathbf{t}(k) \qquad k = 1, 2, \ldots$$

As $\mathbf{t}(k)$ has been interpreted as departure times of the k^{th} event, the service h_1 is finished k times at $\mathbb{t}_1(k) = \mathrm{t}_1(k) + \tau_1$. The same reasoning holds good for h_2 and h_3. Similarly, the service h_4 (respectively h_5) is finalized for the k^{th} time at $\mathbb{t}_4(k) = \mathrm{t}_3(k+1) + \tau_3$, where the waiting time σ_1 (and correspondingly σ_2) per service unit follows from $\mathrm{t}_3(k+1) - \mathrm{t}_1(k) - \tau_1 = \sigma_1$. Notice that there are more services than nodes so that service bundles $\mathbf{h}$ cannot be interpreted as counters defined by (6.19). The same argument applies to node specific start times $\mathbf{t}$ which are to be contrasted with completion times $\mathbb{t}$ of a service bundle $\mathbf{h}$. Different timetables indicate that the same service bundle can be provided at different completion times including different cumulated waiting times. These facts must be taken into account when defining level sets $\mathbb{L}(\mathbf{h}(\mathbb{t}))$ for every service bundle $\mathbf{h}$ with end times $\mathbb{t}$. Such level sets determine the time need with regard to all schedules $\mathbf{d}(k)$, $k = 1, 2, \ldots$, that suffice to realize $\mathbf{h}(\mathbb{t})$. As usual every infeasible output $\mathbf{h}(\mathbb{t})$ is represented by $\mathbb{L}(\mathbf{h}(\mathbb{t})) = \emptyset$. At least for fixed completion times any correspondence $\mathbb{L}$ should be checked against the regularity conditions on p. 348. For example, every period suffices to produce a zero output $\mathbf{h}(\mathbb{t}) = \mathbf{0}$ independent of the deadlines $\mathbb{t}$. Moreover, there is no service bundle $\mathbf{h}(\mathbb{t}) \neq \mathbf{0}$ that can be provided without time consumption.

Although the railway example aims at a timetable that can be reiterated after a while, let us assume that all services are provided a finite number of times. In a very simple case the carrier may accept any activity pattern that generates at least one unit of every service, $\bar{\mathbf{h}} = (1, 1, 1, 1, 1)^{\mathsf{T}}$, within an available period $[0, \mathbf{d}]$ but irrespective of the exact starting or ending times. Figure 6.15 depicts three exemplary timetables all of which provide $\bar{\mathbf{h}}$ or, more precisely, $\mathbf{h}(\mathbb{t}_a)$, $\mathbf{h}(\mathbb{t}_b)$, and $\mathbf{h}(\mathbb{t}_c)$ not exceeding the time slot $[0, \mathbf{d}]$. The first two cases differ merely by shifting τ_3 backwards. The third case avoids waiting times by employing an additional train on the third link.

Table 6.3 in conjunction with Figure 6.15 suggest to compute the cost of providing $\mathbf{h}(\mathbb{t}_a)$ respectively $\mathbf{h}(\mathbb{t}_b)$ and $\mathbf{h}(\mathbb{t}_c)$ by

$$\begin{aligned}
\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_a))) &= \tau_1 + \tau_2 + \sigma_2 + \tau_3, \\
\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_b))) &= \tau_1 + \sigma_1 + \tau_2 + \tau_3, \\
\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_c))) &= \tau_1 + \tau_2 + 2\,\tau_3,
\end{aligned}$$

where the price per unit of time has been normalized to one. This result includes, for example, that the collected time need $u = \{\tau_1, \tau_2, \tau_3, \sigma_2\}$ is feasible with respect to $\mathbf{h}(\mathbb{t}_a)$, i.e., $u \in \mathbb{L}(\mathbf{h}(\mathbb{t}_a))$. In a more general setting it may be necessary to define an objective function $\mathbb{C}$ that maps u to costs possibly by weighing the components

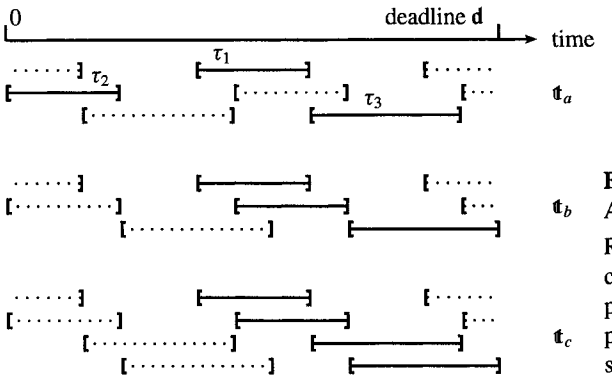


Figure 6.15
Alternative timetables
Railway networks are characterized by interacting processes that are to be performed in series or at the same time.

of u with the number of passengers involved. Although it seems as if there is no optimization in the current setting, bear in mind that the examples in Figure 6.15 represent merely small parts of some large scale timetable. Within such a time table every passenger seeks the shortest connection fitting to his needs. The more general definition of a cost function

$$\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}))) = \inf\{\mathbb{C}(u) \mid u \in \mathbb{L}(\mathbf{h}(\mathbb{t}))\}$$

thus takes into account that every service (i.e., every OD-connection) is performed in accordance with a shortest path which can change with alternative timetables.

Having found cost minimal solutions $\hat{u}$ for every service bundle $\mathbf{h}(\mathbb{t})$ with respect to all feasible timetables, the next step is to choose one timetable that minimizes the overall service costs. Similar to the reasoning of (6.24), this procedure requires to define classes of comparable service bundles that restrict the set of feasible timetables. One such class has been indicated in the above example, where the carrier accepts any service pattern $\mathbf{h}(\mathbb{t})$ satisfying certain minimum standards of a reference vector $\bar{\mathbf{h}}$. In this sense *an optimal timetable provides a service bundle with predefined minimum standards at minimum costs.* Although the computation of optimal timetables goes beyond the scope of this book, let us hand in some basic concepts on the composition and decomposition of service bundles that have been introduce in the preceding Section 6.4.3.2.

Following Figure 6.15 it turns out to be useful to consider five special cases where $\mathbf{h}_\rho$, $\rho = 1,\ldots,5$, signifies the service bundle with $h_\rho = 1$ and all other services being zero; that is, e.g., $\mathbf{h}_1 = (1,0,0,0,0)^\top$. In accordance with Table 6.3 the costs of services follow from $\mathbb{K}\big(\mathbb{L}(\mathbf{h}_\rho)\big) = \tilde{\tau}_\rho$, $\rho = 1,\ldots,5$; the end times $\mathbb{t}_\rho(1)$ are ignored here because they have no impact on the following reasoning. The same argument applies to multiples of $\mathbf{h}_\rho$, where the first service unit is reiterated after a while or the carrier makes use of additional trains. We are thus able to explain simple relationships between network services.

(I) Independent schedules. The basic services h_1, h_2, and h_3 are independent in the sense of additive costs; for instance,

$$\mathbb{K}(\mathbb{L}(\mathbf{h}_1) \cap \mathbb{L}(\mathbf{h}_2)) = \mathbb{K}(\mathbb{L}(\mathbf{h}_1)) + \mathbb{K}(\mathbb{L}(\mathbf{h}_2))$$

gives the costs of providing h_1 and h_2. Similarly, h_4 is independent of h_2, and h_5 is independent of h_1. Observe, however, that an independent offer of h_1 and h_3 differs from h_4 as the latter service includes the waiting time σ_1.

(2) Conditional schedules. Independency is to be contrasted with interacting flows. Offering h_4 given the service h_1 yields conditional incremental costs

$$\mathbb{K}(\mathbb{L}(\mathbf{h}_4)|\mathbb{L}(\mathbf{h}_1)) = \sigma_1 + \tau_3.$$

This is unequal to $\mathbb{K}(\mathbb{L}(\mathbf{h}_4)) = \tau_1 + \sigma_1 + \tau_3$ indicating that the services interact with each other. Similarly the provision of h_1 given h_4 implies

$$\mathbb{K}(\mathbb{L}(\mathbf{h}_1)|\mathbb{L}(\mathbf{h}_4)) = 0 \neq \tau_1.$$

The conditional incremental costs are zero reflecting that additional passengers on the first link do not increase transport costs as long as the corresponding train shows excess capacities. The same reasoning implies $\mathbb{K}(\mathbb{L}(\mathbf{h}_5)|\mathbb{L}(\mathbf{h}_4)) = \tau_2 + \sigma_2$ and $\mathbb{K}(\mathbb{L}(\mathbf{h}_4)|\mathbb{L}(\mathbf{h}_3)) = \tau_1 + \sigma_1$.

(3) Alternative schedules. An alternative offer of basic services $h_1, \ldots, h_5$ is meaningless insofar as they connect different OD-pairs and hence cannot be compared. Nevertheless, we can compare bundles $\mathbf{h}(\mathbb{t}_i)$, $i = a, b, c$, that include essentially the same services though based on different timetables; see Figure 6.15. Each timetable corresponds to different waiting times which in turn determine the goodness of the timetable. In the first case denoted by $\mathbb{t}_a$, the departure of h_3 suits to the arrival of h_1, i.e., $\sigma_1 = 0$. As a consequence, service h_5 includes a long waiting time σ_2, which is substituted in the second scenario $\mathbb{t}_b$ by a shorter waiting time σ_1 with regard to the service h_4. All we have done is to shift the second service to the right so that $\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_a))) > \mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_b)))$ or

$$\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_a)) \cap \mathbb{L}(\mathbf{h}(\mathbb{t}_b))) = \min\{\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_a))), \mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_b)))\} = \mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_b))).$$

Hence, it is advantageous to substitute $\mathbb{t}_a$ by $\mathbb{t}_b$ at least when the number of passengers that consume h_4 and h_5 coincide.[27] The third case $\mathbf{h}(\mathbb{t}_c)$ may be interpreted as the cost of independency. The services h_4 and h_5 are untied by adding one further train such that $\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_c))) = \tau_1 + \tau_2 + 2\tau_3$. This result is to be contrasted with the costs $\mathbb{K}(\mathbb{L}(\mathbf{h}(\mathbb{t}_b))) = \tau_1 + \sigma_1 + \tau_2 + \tau_3$ for the synchronized case.

6.5 Summary

The core problem of the Chapters 3–5 has been traffic assignment, which basically deals with origin-destination flows and their components. In doing so, it concentrates on spatial aspects of traffic flows and their sharing of network resources. One approach that takes time explicitly into account is queuing theory. If the number of jobs arriving at some service station exceeds the capacity, they are queued and

[27] The result may be reversed when many people have to wait for a small group.

have to wait for their turn. Chapter 6 now has shown how to reduce such waiting times by synchronizing network processes in advance. The task is to determine timetables that transform precedence constraints into coordinated traffic patterns which reveal a periodic behavior in the ideal case. The analysis in Section 6.2 has determined technically efficient solutions in the sense of "fastest timetables". They are characterized by a constant average interdeparture time and critical cycles which form the bottlenecks for faster timetables. Furthermore, buffer times have been implemented to improve the stability of the timetable. Slowing down a timetable makes the transportation system able to cope with a certain amount of delays without breaking down. The opportunity costs of deceleration are increasing waiting times of the passengers.

Section 6.4 starts with basic considerations on how to measure outputs (i.e., services) that feature two properties: all events happen at distinct nodes at distinct points of time. Consequently output measurement has to distinguish two views. On the one hand, we can ask for node specific points of time at which the k^{th} event such as the departure of a train happens. The dual view, on the other hand, refers to the number of events that occur up until a given point of time. Given any event pattern the next step is to assess the overall quality of service in terms of passengers' travel times including less comfortable waiting times. This evaluation must take into consideration that passengers coordinate their individual behavior with the schedule in order to realize the most preferred arrival times. The chapter ends with basic expositions on the cost structure of service provision. With special regard to pricing network services, the analysis concentrates on (conditional) incremental costs that follow from additional flows. In a first instance, the concepts are derived and explained on the basis of road networks which omit here aspects of time. Afterwards, the results are transferred to synchronized timetables.

7

Critical Review

7.1 Synopsis of Network Economics

Economic principles of transport networks, which constitute the core of network economics, essentially deal with methods of attaining bundles of network services with least possible expenditure of means. Network services, however, are generally not desired on their own, but they designate a derived demand resulting from the desire to realize spatially separated activities. In this sense, transportation indicates a wide variety of problem classes that involves among others the transportation of goods, persons, vehicles, electricity, voice, and data. The microeconomic analysis of certain transport phenomena over physical networks gives rise to network models that represent the physical infrastructure and aim to compute the flows of one or more commodities with respect to the capacitated elements (i.e., nodes and links) of the network. More specifically, the economic problem is a question of how to allocate scarce network resources to rival flows without a waste. It takes no wonder that the answers differ due to the technical peculiarities of the network concerned. Road traffic is substantially characterized by the decisions of individual motorists, railway networks are guided by timetables, regular voice telephone networks use circuit-switched as well as packet-switched traffic, and packet delivery systems exploit advantages of bundling flows.

(a) Network Infrastructure

In the light of this variety of aspects, we have started with basic properties that are common to all transport systems; see Chapter 2. *Networks consist of nodes and links forming paths* which are used to send one or more goods or messages from their points of origin to other places of destination. The corresponding flows are restricted to balance constraints to ensure the *conservation of flow* (Kirchhoff's Law) by which no flow unit is lost accidentally. The distinction between origin-destination flows, path flows, and link flows is necessary to record serial and parallel network processes, where the flows will move several usually capacitated links. The sharing of network resources binds flows together, where we have to differentiate strictly between two

classes of network services. While the demand refers to origin-destination flows (e.g., trips), the supply side is mainly interested in flows on capacitated links. Given a capacitated flow network in the sense of a microeconomic production technology, the next step is to ask for technically efficient flow patterns that do not dissipate network resources (i.e., link capacities); see Chapter 3. The corresponding *maximum flow* problem has been discussed at first on the basis of a single commodity, and we have learned that the *bottlenecks of a network* can be described in terms of minimum cuts. Later, the same problem has been discussed with respect to multicommodity flow problems which arise when several commodities share the same network. Here the commodities are differentiated by their origin-destination pairs, which may be refined by further physical characteristics – see, e.g., multiclass-user networks.

(b) Traffic Assignment

Technical feasibility is to be contrasted with *cost efficiency*, where analysts must take account of several cost categories: transport costs, costs of waiting, investment costs, etc. In accordance with microeconomic theory of the firm Chapter 3 starts with the problem of minimizing transport costs in its simplest form for a single flow on a capacitated network. This problem assumes constant average link costs and neglects therefore any congestion phenomena although the least cost routes (or shortest paths) cannot carry the whole traffic. Correspondingly, the analysis yields link specific shadow prices that indicate the relative scarcity of the respective links. In the sense of element based charging, shadow prices increase link costs per flow unit in such a way that the carrier becomes indifferent between alternatively used paths. This outcome is twofold; on the one hand, indifference reflects the Law of One Price while, on the other hand, the carrier pays in fact only pure transport costs. The difference indicates the producer's surplus because all flow units are valued at marginal costs. Analogous results apply to multicommodity flows with constant average link costs and a fixed demand for trips. Again the *cost minimal traffic assignment* is intertwined with a system of shadow prices that makes all consumers of network services indifferent between those paths that are in use for the respective service (or end-to-end connection). Be aware that the Law of One Price for every service is a construct which assigns property rights to a benevolent network operator; he seeks to minimize overall flow costs resulting in a system optimal traffic assignment. If the derived price system matches the perceived travel costs of price taking network users, then – in accordance with *Wardrop's first principle* – no traveler can take advantage from altering his route choice (user optimum). Hence, we have found in principle a price system that makes non-cooperative self-interested network users behave in a system optimal manner. A similar result may be derived when property rights of road usage are claimed by motorists. In this setting pure travel costs are to be augmented by costs of waiting whenever demand exceeds capacity. Again we can expect to obtain a user optimal travel pattern with the same costs on all used routes.

(c) Overloaded Networks

The situation of user optima changes completely when users' perception of travel costs differs from marginal social costs. This is the case of *congestion* where

additional drivers force others to slow down their speed contingently to a standstill. The fundamental diagram of traffic represents this correlation between the level of service (i.e., speed) and the traffic density for a single road. Again we have shown in principle that there is a price system which guides travelers to behave in a system optimal way. In the sense of Pigou, external congestion costs require a toll/subsidy schedule that implements marginal cost pricing – possibly besides a toll revenue. Be aware that marginal cost pricing is in general an extremely complex problem due to the problems of measuring traffic flows respectively congestion. For example, fixing one point in time and taking a photo of the network, we are able to measure all relevant variables but not their evolution. By contrast, observing one point in space over a certain period of time, we can measure evolution of trafic processes, but the rest of the network remains hidden. Although systematic traffic counts may solve this problem of time and space related road pricing[1], marginal cost pricing is of limited capability to solve problems of the following type. Suppose a perturbation occurs at a certain time at a certain place. Depending on traffic conditions, the corresponding shock wave can just dissolve or cause a traffic break down. But more interestingly, such waves may trigger further waves, and they can move through space and time. Aside from capacity expansions, all that can be done to avoid these problems is to limit traffic volumes which ensure a certain level of service including a reduced risk of perturbations. In other words, it is strongly recommended not to operate the network at the limit. At present, the effectiveness of a congestion charge can be observed for the City of London.

(d) Toward More Realistic Consumer Behavior

Traffic assignment as has been discussed in Chapter 3 features various simplifications that are to be relaxed in the subsequent analysis. Deterministic assignment models are restricted to having full information about the technical details of the network and about traffic conditions. For cases without perfect knowledge Chapter 4 suggests to define *stochastic user equilibria* where the travelers' behavior is described through probabilistic modeling. Network economists are then interested in the determination of the most likely traffic patterns which combine user optimality with aspects of entropy maximization. This approach is closely related to *gravity models* which try to explain – among other things – the attractiveness of a target relative to other regions and the relative importance of a source region for a target area.

The next generalization refers to pricing network services with special regard to a *price sensitive demand* for traffic – a result which can be derived within microeconomic theory of the household. In the standard assignment model with a fixed demand for traffic, element based charging aims at uniform network utilization. This effect emphasizes the role of prices in guiding traffic flows – i.e., route choice – with respect to resource allocation. Nevertheless, congestion often befalls the entire network so that *traffic diversion* is no longer a practicable solution. Instead, congestion charges as well as peak-load pricing simply targets *diminution of traffic*

[1] A generalized fine-tuning pricing scheme is unlikely to be ever introduced, be it for efficiency criteria or for lack of political acceptability.

volumes. This effect of trip choice is realistic and requires to account for a price sensitive demand for traffic. Of course, we implement at the same time traffic creating effects by any measure that improves traffic conditions. As before, it is shown in principle that a link specific toll system does exist which transforms a user optimal traffic assignment into a system optimum. Moreover, it is shown what prices will prevail in user optima that have certain regularity conditions missing.

(e) Adjusting Network to Consumer Needs

Having dropped the assumption of constant average costs per link and the assumption of a fixed trip matrix, it is consequent in the next step to abandon the fiction of a given physical infrastructure. This relaxation corresponds to the distinction between the short run and the long run in microeconomic theory of the firm. While in the first case some of the factors of production must be used in fixed amounts, a firm is allowed to adjust all of its factors in the long run. In this sense *network design* pursues the target to develop a production, inventory, and transportation plan to meet anticipated demands over a specified time horizon. Although investment problems have an inherently dynamic character, Chapter 5 starts with a sequence of static problems. They deal with the efficient usage of resources that are dedicated to network improvements such as capacity expansions or insertions of new links and nodes. An example is the projected Gotthard railway tunnel which is claimed to increase the capacity of the existing connection, to reduce the travel time between Zurich and Mailand, to improve the route, and to increase quality of life in the surrounding area. At least in part the analysis has shown that investment programs can be "self-financing" in the sense that savings in transport costs offset investment costs. Regarding road networks, for instance, we have to admit though that those participants who profit from network improvements do not necessarily coincide with the parties who have to pay. By contrast, this question is of minor importance to telecommunications networks where transport costs can be ignored. *Optimal investment programs* focus here, for example, on the minimum cost of capacity installation that suffice to accommodate a fixed demand. Of course, rerouting given traffic volumes is in general superimposed by taking into account a price sensitive demand. Network improvements then affect traffic assignment in opposite direction as compared to congestion charging. Aside from routing problems, network design serves further tasks such as reliability and survivability of networks. Both aspects supplement network structures that result from minimum investment costs by issues of their vulnerability. The corresponding risk of canceled orders determines one part of the level of service that is delivered to the customers. The construction of redundant connections helps therefore to reduce the consequences of failing network elements that destroy certain network services. This deterministic case of network failures is referred to as *survivability*. By contrast, network *reliability* addresses issues of random events (e.g., weather) which determine the probability of the availability of a service.

(f) Analytical Procedures

At the latest when accounting for a price sensitive demand in the context of network design, traffic assignment problems arrive at a degree of complexity that demands for model simplification. One strategy to tackle this problem is to separate the problem at hand into several interdependent but smaller problems. The procedure of travel forecasting in Chapter 2 refers to a *sequential decision process* the network users pass through. In the ideal case such a sequence consists of a list of modules all of which explain a distinct behavior under a ceteris paribus condition. These results can then be used as inputs for the next module, where feedback control is needed to ensure consistency. With special regard to the travelers decision procedure, traffic assignment turns out to be the core problem of the travel forecasting procedure which starts with locational choice. A somewhat more pretentious approach is needed when a second group of decision-makers is taken into consideration. This is the case of a network operator who has the task of designing a network that fits to the needs of traffic demand. The difficulty in determining a desirable market solution is as obvious as complex: any change in the network infrastructure modifies traffic conditions and leads thus to a changing behavior of network users. Any investment program of the network designer that is optimal for a given travel behavior may be proved to be suboptimal when accounting for reactions of the opposite group of travelers. *Bilevel programming* is one approach to simplify a complex problem by separating it into two interdependent but smaller problems with a hierarchic structure. The upper level problem refers to the network designer as a von Stackelberg leader who makes the first move in the corresponding game. By assumption he can observe the response of the dependent group of travelers who act in the sense of a von Stackelberg follower (lower level problem). The idea is to solve repeatedly both types of problems given the strategy of the other group. Provided that this sequence converges, we have found a *von Stackelberg solution* where the strategy of the leader determines the most preferred answer in the set of best responses of the followers.[2] In the short term such a solution may be used to find an optimal system design with respect to route and trip choice; see Chapter 5. In the long term, however, the procedure must also account for locational choices such as those locating facilities on networks.

(g) Correlation of Space and Time

Aspects of time pertain by far not only to investment problems. Every movement of goods or messages between two places has always a *temporal dimension*, too. The time need between the departure time and the arrival time depends not only on the distance between the origin and the destination but also on the prevailing traffic conditions. Congestion increases travel times because it forces to reduce speed, and synchronization of flows induces further waiting times. The concept of a space-time flow diagram with queuing has been introduced at the end of Chapter 4. It

[2] It seems to be worth noting that the von Stackelberg model refers to sequential games where the von Stackelberg leader may take advantage of making the first move. If we want to emphasize that both groups give mutually best responses to the chosen strategy of the other group at the same time, it would be more adequate to speak of Cournot-Nash equilibria.

suggests to improve our analysis by *dynamic traffic assignment* including a time-of-the-day choice that helps to avoid peak travel with high congestion. As this issue goes beyond the scope of this book, it has been widely ignored. Instead, Chapter 5 introduces *queuing*[3] which occurs when the capacities of network elements are overloaded by a (stochastic) demand for services. One task of network design is to minimize the overall system delay – which usually refers to a busy hour – by an appropriate capacity assignment. Another strategy could aim at the design of service disciplines through pricing or flow controls. These instruments would lead to a more efficient usage of network resources by dispersing the demand including its peaks more uniformly over time. Contrasted with continuous flows on road networks or in telecommunications systems, Chapter 6 lays emphasis on discrete events that are restricted by precedence constraints. Railway networks, to give an example, bundle their customers by trains and each transport process serves a distinct "batch of passengers". Such service networks improve their level of service by *synchronization through timetables* in order to minimize overall waiting times. In doing so, the service provider has to design a package of scheduled transportation services which must then be communicated to the market (and modified dynamically in response to the market's feedback).

7.2 Closing Comments

From an economic point of view there can be no doubt that transport systems must satisfy three requirements. First, all network services are to be produced *without waste of resources*. Second, *consumer sovereignty* requires that the network performance including the quantity and the quality of services suits to the needs of its customers. And third, the *network users have to pay the full cost of consumed services*. Regarding packet delivery systems or telecommunications systems – where customers can be excluded from consuming services – all these aims seem, at least in principle, to be within reach. Competition between rival network operators (including potential rivals) forces them to produce any service range in a non-wasteful manner with the most suitable methods of production.[4] As the success of an individual service range depends essentially on its goodness of fit in relation to consumers' needs, only the most competitive suppliers will survive provided they can realize a nonnegative profit. By contrast, consumers strive for scarce network services that all cannot obtain. Hence, only those consumers win out over their rivals whose willingness to pay outranges less important needs. Whenever the consumers are not willing or not able to bear the full cost of service provision, supply will be reduced possibly until it disappears.

The role of pricing network services results basically from its effects on selection and allocation. When one firm is able to earn a nonnegative profit at given prices, then

[3] Queuing requires that jobs are able and willing to wait. Regarding loss systems, network design targets to minimize lost traffic with given expenditure of means.

[4] From Schumpeter's point of view one may add the introduction of new methods of production and new services by redesigning transport networks.

all competitors have to leave the market who cannot cover their costs at the same prices. Similarly all potential consumers are sorted out whose willingness to pay for a service is insufficient to bear the respective cost of production. These arguments are to be refined when talking about multicommodity networks with differentiated services in the sense that each transport service with respect to an origin-destination pair denotes a distinct good. Moreover, we have learned that each of these services represents a composite good as it uses a sequence of certain network elements. Here, element based charging is one way to find a fair allocation of costs among customers of communications, transportation and other networks. Although we have to realize that there is still a significant mismatch between prices paid by individual transport users and the costs they cause, adequate solutions seem to be feasible even for urban road systems. This is particularly true in the light of computerized pricing systems. However, one conclusion to be drawn is clear: policies to curb congestion must be differentiated in time and space. The corresponding price system then serves also as an information system which supports the search for traffic equilibria even when starting at a disequilibrium.

Unfortunately, there is a long list of caveats against a satisfiable market solution that must be handled with care. For example, it is frequently argued that railroads or telecommunications systems possess to some degree the characteristics of a natural monopoly so that competition becomes deactivated. This argument seems to be disproved at least for telecommunications networks, and railways are likewise not necessarily absolute monopolists at least if we account for highway competition or alternative transport modes such as water transport. In any case we should bear in mind that monopolies – be they natural or not – sustain due to barriers that deter potential rivals from market entry. The causes for these barriers can be manifold, but in the end we frequently observe incumbent network monopolies which ignore consumer needs and are incapable of introducing modern technologies. Instead, they just repair an obsolete system including a bad design that has been grown over the years. This observation seems to be particularly true for publicly owned networks such as road infrastructure. Here, externalities internal and external to the network (i.e., congestion, environmental consequences, pollution, etc.) seem to be unsolvable problems mainly because the public opinion insists that roads are to be publicly provided[5] and that they are non-rivalrous and non-excludable goods. Nevertheless, we cannot too strongly emphasize that *roads are no public goods*. Furthermore, feasible instruments of road pricing do exist that internalize the external costs of road usage according to the polluter-pays-principle; examples are road damage pricing, congestion pricing (including peak-load pricing), fuel excise duties, and differentiated vehicle taxes in accordance with the vehicle's environmental and noise characteristics, possibly linked with electronic kilometre charges. Nevertheless, costs are seldom imputed at the point of use and the prices paid for a journey rarely reflect the true costs of that journey. In the end, the inefficiencies of state-owned and regulated systems remain the major problem.

[5] If public roads are financed by means of tax revenues, a further problem arises: those who pay the tax do not necessarily benefit in an adequate way from road services.

All these arguments are recognized by governments. At least we can find in the "Green Paper on Fair and Efficient Pricing" of the European Commission the follwing comment: *The crucial importance of transport externalities arises from the fact that, in a market economy, (economic) decisions are heavily dependent on market prices. However, when market prices fail to reflect existing scarcities (clean air, absorptive capacity of the environment, infrastructure etc.), the individual decisions of consumers and producers no longer add up to an outcome that provides maximum benefits to society as a whole. Thus, pricing on the basis of full social costs is a key element of an efficient and sustainable transport system.* Nevertheless, we do not agree with the conclusion of a "market failure". Of course, an individual infrastructure user does not compare the private benefits of his decision with the total costs this decision imposes on society as a whole (the so-called marginal social costs). But before accusing the market of any failure we should be aware that the society has agreed not to implement basic rules – such as the polluter-pays-principle – that are necessary for a proper working of the market. As long as we do not abstain from foolish positions such as "free drive for free riders", the tragedy of the commons including the free-rider problem continues to be unsolved. In this sense we should trust much more in Adam Smith's idea that competition can harmonize the pursuit of self-interested individuals with economic efficiency.

References

Aashtiani, H. Z., *The Multi-Modal Traffic Assignment Problem,* PhD thesis, Alfred P. Sloan School of Management. Cambridge, Mass.: MIT Press, 1979.

Aashtiani, H. Z. and Magnanti, T. L., Equilibria on a Congested Transportation Network. *SIAM Journal on Algebraic and Discrete Methods*, 2 (1981), pp. 213–226.

Abdulaal, M. and LeBlanc, L. J., Continuous Equilibrium Network Design Models. *Transportation Research*, 13B (1979a), pp. 19–32.

Abdulaal, M. and LeBlanc, L. J., Methods for Combining Modal Split and Equilibrium Assignment Models. *Transportation Science*, 13 (1979b), pp. 292–314.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., *Network Flows: Theory, Algorithms, and Applications*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

Akcelik, R., A Graphical Explanation of the Two Principles and Two Techniques of Traffic Assignment. *Transportation Research*, 13A (1979), pp. 179–184.

Akian, M., Theory of Cost Measures: Convergence of Decision Variables. Rapport de recherche N 2611, Institut National de Recherche en Informatique en Automatique, INRIA, 1995.

Alcaly, R. E. and Klevorick, A. K., A Note on the Dual Prices of Integer Programs. *Econometrica*, 34 (1966), pp. 206–214.

Alevras, D., Grötschel, M., and Wessäly, R., Cost-Efficient Network Synthesis from Leased Lines. Technical report, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1997.

Alonso, W., Gravity Models. In EATWELL ET AL. (1987), pp. 561–562.

Anderson, D., Models for Determining Least-Cost Investment in Electricity Supply. *Bell Journal of Economics and Management Science*, 3 (1972), pp. 267–299.

Anderson, R. C. and Claus, A., Cost Allocation in Transportation Systems. *Southern Economic Journal*, 43 (1976), pp. 793–803.

Añez, J., de la Barra, T., and Pérez, B., Dual Graph Representation of Transport Networks. *Transportation Research*, 30B (1996), pp. 209–216.

Arnott, R., de Palma, A., and Lindsey, R., Economics of a Bottleneck. *Journal of Urban Economics*, 27 (1990), pp. 111–130.

Arnott, R. and Kraus, M., Financing Capacity in the Bottleneck Model. *Journal of Urban Economics*, 38 (1995), pp. 272–290.

Ashton, W. D., *The Theory of Road Traffic Flow*. London: Methuen, 1966.

Awduche, D. O., Agogbua, J., and McManus, J., An Approach to Optimal Peering Between Autonomous Systems in the Internet. In *Proceedings of the 7th International Conference on Computer Communication and Networks*, pp. 346–351, 1998.

Baccelli, F. L., Cohen, G., Olsder, G. J., and Quadrat, J.-P., *Syncronization and Linearity: An Algebra for Discrete Event Dystems*. Chichester: Wiley & Sons, 1992.

Bailey, E. E., Graham, D. R., and Kaplan, D. P., *Deregulating the Airlines*. 2 ed. Cambridge, Mass.: MIT Press, 1986.

Ball, M. O., Colbourn, C. J., and Provan, J. S., Network Reliability. In BALL ET AL. (1995B), pp. 673–783.

Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., eds., *Network Models: Handbooks in Operations Research and Management Science, Vol. 7*. Amsterdam: North-Holland, 1995.

Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., eds., *Network Routing: Handbooks in Operations Research and Management Science, Vol. 8*. Amsterdam: North-Holland, 1995.

Bäßler, R. and Deutsch, A., *Nachrichtennetze: Handbuch Nachrichtenelektronik*. 2 ed. Berlin: Verlag Technik, 1989.

Batten, D. F. and Roy, J. R., Entropy and Economic Modelling. *Environment and Planning*, 14A (1982), pp. 1047–1061.

Baumol, W. J., Indivisibilities, Vol. 2. In EATWELL ET AL. (1987), pp. 793–795.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M., *Nonlinear Programming: Theory and Algorithms*. 2 ed. New York: Wiley, 1993.

Beckmann, M. J., *Lectures on Location Theory*. Berlin: Springer, 1999.

Beckmann, M. J., McGuire, C. B., and Winsten, C. B., *Studies in the Economics of Transportation*. New Haven: Yale University, 1956.

Beckmann, M. J. and Golob, T. F., Traveler Decisions and Traffic Flows: A Behavioral Theory of Network Equilibrium. In BUCKLEY (1974), pp. 453–482.

Beckmann, M. J. and Kapur, K. C., Conjugate Duality: Some Applications to Economic Theory. *Journal of Economic Theory*, 6 (1972), pp. 292–302.

Beckmann, M. J. and McGuire, C. B., *The Determination of Traffic in a Road Network – An Economic Approach*. Santa Monica: The RAND Corporation, 1953.

Bell, M. G. H., Stochastic User Equilibrium Assignment in Networks with Queues. *Transportation Research*, 29B (1995), pp. 125–137.

Bell, M. G. H. and Iida, Y., *Transportation Network Analysis*. Chichester: Wiley, 1997.

Ben-Akiva, M., Dynamic Network Equilibrium Research. *Transportation Research*, 19A (1985), pp. 429–431.

Ben-Akiva, M. and Bowman, J. L., Activity Based Travel Demand Model Systems. In MARCOTTE, NGUYEN (1998), pp. 27–46.

Ben-Akiva, M., Hirsh, M., and Prashker, J., Probabilistic and Economic Factors in Highway Geometric Design. *Transportation Science*, 19 (1985), pp. 38–57.

Ben-Akiva, M. and Lerman, S. R., *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambdridge, Mass.: MIT Press, 1985b.

Bergendorff, P., Hearn, D. W., and Ramana, M. V., Congestion Toll Pricing of Traffic Networks. In Pardalos, P. M., Hearn, D. W., and Hager, W. W., eds.: *Network Optimization,* Proceedings of the Network Optimization Conference, Florida, 1996, pp. 51–71. Springer: Berlin, 1997.

Bern, M. W. and Graham, R. L., Das Problem des kürzesten Netzwerks. In Faltings, G., ed.: *Moderne Mathematik*, pp. 52–58. Berlin: Spektrum der Wissenschaft, 1996.

Bertsekas, D. P., *Linear Network Optimization: Algorithms and Codes*. Cambridge, Mass.: MIT Press, 1991.

Bertsekas, D. P., *Network Optimization: Continuous and Discrete Models*. Belmont, Mass.: Athena Scientific, 1998.

Bertsekas, D. P. and Gallager, R., *Data Networks*. 2 ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1992.

Bienstock, D., Chopra, S., Günlük, O., and Tsai, C.-Y., Minimum Cost Capacity Installation for Multicommodity Network Flows. *Mathematical Programming*, 81 (1998), pp. 177–199.

Blackorby, C. and Donaldson, D., Market Demand Curves and Dupuit Marshall Consumers' Surpluses: A General Equilibrium Analysis. *Mathematical and Social Sciences*, 37 (1999), pp. 139–163.

Blackorby, C., Primont, D. A., and Russel, R. R., *Duality, Separability, and Functional Structure:Theory and Economic Applications*. Amsterdam: North-Holland, 1978.

Blessing, J., Efficient Network Design Using Heuristic and Genetic Algorithms. In CORNE ET AL. (2000), pp. 35–55.

Bobzin, H., *Indivisibilities: Microeconomic Theory with Respect to Indivisible Goods and Factors*. Heidelberg: Physica, 1998.

Bowman, L. A. and Turnquist, M. A., Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops. *Transportation Research*, 15A (1981), pp. 465–471.

Boyce, D. E., A Framework for Constructing Network Equilibrium Models of Urban Location. *Transportation Science*, 14 (1980), pp. 77–96.

Boyce, D. E., Is the Sequential Travel Forecasting Paradigm Counterproductive? *Journal of Urban Planning and Development*, 128 (2002), pp. 169–183.

Boyce, D. E. and Daskin, M. S., Urban Transportation. In ReVelle, C. and McGarity, A. E., eds.: *Design and Operation of Civil and Environmental Engineering Systems*, pp. 277–341. New York: Wiley, 1997.

Boyce, D. E., Janson, B. N., and Eash, R. W., The Effect on Equilibrium Trip Assignment of Different Link Congestion Functions. *Transportation Research*, 15A (1981), pp. 223–232.

Braess, D., Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, 12 (1968), pp. 258–268.

Braess, D. and Koch, G., On the Existence of Equilibria in Asymmetrical Multiclass-User Transportation Networks. *Transportation Science*, 13 (1979), pp. 56–63.

Braker, J. G., *Algorithms and Applications in Timed Discrete Event Systems, Dissertation.* Doctoral thesis, Faculty of Technical Mathematics and Informations, TU Delft, Delft, 1993.

Branston, D., Link Capacity Functions: A Review. *Transportation Research*, 10 (1976), pp. 223–236.

Breyer, S. and MacAvoy, P. W., Regulation and Deregulation, Vol. 4. In EATWELL ET AL. (1987), pp. 128–134.

Bryan, D. L. and O'Kelly, M. E., Hub-and-Spoke Networks in Air Transportation: An Analytical Review. *Journal of Regional Science*, 39 (1999), pp. 275–295.

Buckley, D. J., ed., *Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney, 1974.* New York: Elsevier, 1974.

Buhr, W. and Pauck, R., *Stadtentwicklungsmodelle: analytische Instrumente empirisch orientierter Simulationsansätze zur Lösung von Projektions- und Planungsproblemen der Städte.* Baden-Baden: Nomos, 1981.

Buhr, W. and Wagner, H. J., Line Integrals in Applied Welfare Economics: A Summary of Basic Theorems. Discussion Paper No. 54–95, University of Siegen, Siegen, Germany, 1995.

Campbell, J. F., Integer Programming Formulations of Discrete Hub Location Problems. *European Journal of Operational Research*, 72 (1994), pp. 387–405.

Cantarella, G. E., A Genereal Fixed-Point Approach to Multimode Multi-User Equilibrium Assignment with Elastic Demand. *Transportation Science*, 31 (1997), pp. 107–128.

Caputo, M. R., Lagrangian Transposition Identities and Reciprocal Pairs of Constrained Optimization Problems. *Economics Letters*, 66 (2000), pp. 265–273.

Carey, M., Integrability and Mathematical Programming Models: A Survey and a Parametric Approach. *Econometrica*, 45 (1977), pp. 1957–1976.

Carey, M., The Dual of the Traffic Assignment Problem with Elastic Demands. *Transportation Research*, 19B (1985), pp. 227–237.

Carey, M., Network Equilibrium: Optimization Formulations with Both Quantities and Prices as Variables. *Transportation Research*, 21B (1987), pp. 69–77.

Carey, M., Nonconvexity of the Dynamic Traffic Assignment Problem. *Transportation Research*, 26B (1992), pp. 127–133.

Cassandras, C. G., Lafortune, S., and Olsder, G. J., Introduction to the Modelling, Control and Optimization of Discrete Event Systems. In Isidori, A., ed.: *Trends in Control: a European Perspective*, pp. 217–291. Berlin: Springer, 1995.

Chakravarty, S., Cost-benefit Analysis, Vol. 1. In EATWELL ET AL. (1987), pp. 687–690.

Chao, G. S. and Friesz, T. L., Spatial Price Equilibrium Sensitivity Analysis. *Transportation Research*, 18B (1984), pp. 423–440.

Chen, H.-K., *Dynamic Travel Choice Models: A Variational Inequality Approach.* Berlin: Springer, 1999.

Chen, M. and Alfa, A. S., Algorithms for Solving Fisk's Stochastic Traffic Assignment Model. *Transportation Research*, 25B (1991), pp. 405–412.

Chipman, J. S. and Moore, J. C., The Scope of Consumer's Surplus Arguments. In Tang, M., Westfield, F., and Worley, J., eds.: *Welfare and Time in Economics: Essays in Honor of Nicholas Georgescu-Roegen*, pp. 69–123. Lexington: Heath, 1976.

Chipman, J. S. and Moore, J. C., Compensating Variation, Consumer's Surplus, and Welfare. *The American Economic Review*, 70 (1980), pp. 933–949.

Clarke, F. H., Generalized Gradients and Applications. *Transactions of the American Mathematical Society*, 205 (1975), pp. 246–262.

Clegg, J., Smith, M. J., Xiang, Y., and Yarrow, R., Bilevel Programming Applied to Optimising Urban Transportation. *Transportation Research*, 35B (2001), pp. 41–70.

Collatz, L. and Wetterling, W., *Optimierungsaufgaben.* 2 ed. Berlin: Springer, 1971.

Conrad, K. and Heng, S., Financing Road Infrastructure by Savings in Congestion Costs: A CGE Analysis. *The Annals of Regional Science*, 36 (2002), pp. 107–122.

Cooper, R. B., *Introduction to Queueing Theory.* New York: Macmiallan, 1972.

Corne, D. W., Oates, M. J., and Smith, G. D., eds., *Telecommunications Optimization: Heuristic and Adaptive Techniques.* New York: Wiley, 2000.

Dafermos, S. C., *Traffic Assignment and Resource Allocation in Transportation Networks,* PhD thesis. Baltimore, MD: John Hopkins University, 1968.

Dafermos, S. C., An Extended Traffic Assignment Problem with Applications to Two-Way Traffic. *Transportation Research*, 5 (1971), pp. 366–389.

Dafermos, S. C., The Traffic Assignment Problem for Multiclass-User Transportation Networks. *Transportation Science*, 6 (1972), pp. 73–87.

Dafermos, S. C., Traffic Equilibrium and Variational Inequalities. *Transportation Science*, 14 (1980), pp. 42–54.

Dafermos, S. C., The General Multimodal Network Equilibrium Problem with Elastic Demand. *Networks*, 12 (1982), pp. 57–72.

Dafermos, S. C. and Nagurney, A., Stability and Sensitivity Analysis for the General Network Equilibrium-Travel Choice Problem. In VOLMULLER, HAMERSLAG (1984), pp. 217–231.

Dafermos, S. C. and Nagurney, A., Supply and Demand Equilibrium Algorithms for a Class of Market Equilibrium Problems. *Transportation Science*, 32 (1989), pp. 118–124.

Dafermos, S. C. and Sparrow, F. T., The Traffic Assignment Problem for a General Network. *Journal of Research of the National Bureau of Standards*, 73B (1969), pp. 91–118.

Dafermos, S. C. and Sparrow, F. T., Optimal Resource Allocation and Toll Patterns in User-Optimised Transport Networks. *Journal of Transportation Economy and Policy*, 5 (1971), pp. 184–200.

Daganzo, C. F., *Multinomial Probit: The Theory and Its Application to Demand Forecasting*. New York: Academic, 1979.

Daganzo, C. F., Unconstrained Extremal Formulation of Some Transportation Equilibrium Models. *Transportation Science*, 16 (1982), pp. 332–360.

Daganzo, C. F., *Fundamentals of Transportation and Traffic Operations*. Oxford: Pergamon-Elsevier, 1997.

Daganzo, C. F. and Sheffi, Y., On Stochastic Models of Traffic Assignment. *Transportation Science*, 11 (1977), pp. 253–274.

Dantzig, G. B., *Linear Programming and Extensions*. Princeton, New Jersey: Princeton University, 1963. (11th printing 1998).

Dantzig, G. B., Harvey, R. P., Landsdowne, Z. F., Robinson, D. W., and Maier, S. F., Formulating and Solving the Network Design Problem by Decomposition. *Transportation Research*, 13B (1979), pp. 5–17.

Dantzig, G. B. and Thapa, M. N., *Linear Programming 1: Intoduction*. New York: Springer, 1997.

Daskin, M. S., A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. *Transportation Science*, 17 (1983), pp. 48–70.

Deneubourg, J. L., de Palma, A., and Kahn, D., Dynamic Models of Competition Between Transportation Modes. *Environment and Planning*, 11A (1979), pp. 655–673.

Diewert, W. E., *The Measurement of the Benefits of Infrastructure Services*. Berlin: Springer, 1986.

Dijkstra, E. W., A Note on two Problems in Connexion with Graphs. *Numerische Mathematik*, 1 (1959), pp. 269–271.

Domencich, T. A. and McFadden, D., *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North Holland, 1975.

Drezner, T. and Drezner, Z., A Note on Applying the Gravity Rule to the Airline Hub Problem. *Journal of Regional Science*, 41 (2001), pp. 67–73.

Drissi-Kaitouni, O., A Variational Inequality Formulation of the Dynamic Traffic Assignment Problem. *European Journal of Operational Research*, 71 (1993), pp. 188–204.

Dupuit, P. and Nagurney, A., Dynamical Systems and Variational Inequalities. *Annals of Operations Research*, 44 (1993), pp. 9–42.

Eatwell, J., Milgate, M., and Newman, P., eds., *The New Palgrave, A Dictionary of Economics*. London: Macmillan, 1987.

Economides, N., The Economics of Networks. *International Journal of Industrial Organization*, 14 (1996), pp. 673–700.

van Egmond, R.-J., Propagation of Delays in Public Transport. *Delft University*, (1998).

van Egmond, R.-J., Capacity Assessment of Transport Networks, an Algebraic Approach. *Proceedings 5th NECTAR Conference, Delft*, (1999), pp. 14 ff.

Eisenkopf, A., Zur Rolle von sektorspezifischen Regulierungsinstitutionen in Netzindustrien. *Perspektiven der Wirtschaftspolitik*, 4 (2003), pp. 449–465.

Ekeland, I. and Témam, R., *Convex Analysis and Variational Problems*. Philadelphia: SIAM, 1999.

Elster, K.-H., Reinhardt, R., Schäuble, M., and Donath, G., *Einführung in die nichtlineare Optimierung*. Leipzig: Teubner, 1977.

Elzinga, J., Hearn, D., and Randolph, W. D., Minimax Multifacility Location with Euclidean Distance. *Transportation Science*, 10 (1976), pp. 321–336.

Erber, G. and Hagemann, H., Economics of Networks. In Zimmermann, K., ed.: *Frontiers in Economics*, pp. 235–268. Berlin: Springer, 2002.

Erlander, S., *Optimal Spatial Interaction and the Gravity Model*. Berlin: Springer, 1980.

European Commission, *Green Paper, Towards Fair and Efficient Pricing in Transport*, 1995.

European Commission, *White Paper, European Transport Policy for 2010: Time to Decide*, 1998.

Evans, S. P., *Some Applications of Optimisation Theory in Transport Planning*. PhD thesis, Research Group in Traffic Studies, London, 1973.

Evans, S. P., Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research*, 10 (1976), pp. 37–58.

Evans, A. W., Road Congestion Pricing: When is it a Good Policy? *Journal of Transport Economics and Policy*, 26 (1992), pp. 213–243.

Faragó, A., Blaabjerg, S., Ast, L., Gordos, G., and Henk, T., A New Degree of Freedom in ATM Network Dimensioning: Optimizing the Logical Configuration. *IEEE Journal on Selected Areas in Communications*, 13 (1995), pp. 1199–1206.

Färe, R., Grosskopf, S., and Yoon, B. J., A Theoretical and Empirical Analysis of the Highway Speed-Volume Relationship. *Journal of Urban Economics*, 12 (1982), pp. 115–121.

Färe, R. and Grosskopf, S., Measuring Congestion in Production. *Journal of Economics*, 43 (1983), pp. 257–271.

Federal Highway Administration (FHWA), Washington, D.C., *Highway Statistics*, 2000. `www.fhwa.dot.gov/ohim/hs00/`.

Fernandez, J. E. and Friesz, T. L., Equilibrium Predictions in Transportation Markets: the State of the Art. *Transportation Research*, 17B (1983), pp. 155–172.

Ferrari, P., The Reliability of the Motorway Transportation System. *Transportation Research*, 22B (1988), pp. 291–310.

Ferrari, P., The Control of Motorway Reliability. *Transportation Research*, 25A (1991), pp. 419–427.

Ferrari, P., Road Pricing and Network Equilibrium. *Transportation Research*, 29B (1995), pp. 357–372.

Ferrari, P., Capacity Constraints in Urban Transport Networks. *Transportation Research*, 31B (1997), pp. 291–301.

Fiacco, A. V. and McCormick, G. P., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: Whiley, 1968.

Fisk, C. S., Some Developments in Equilibrium Traffic Assignment. *Transportation Research*, 14B (1980), pp. 243–255.

Fisk, C. S., A Conceptual Framework for Optimal Transportation Systems Planning with Integrated Supply and Demand Models. *Transportation Science*, 20 (1986), pp. 37–47.

Fisk, C. S., Spatial Price Equilibrium on Congested Networks. *Transportation Research*, 21B (1987), pp. 175–182.

Fisk, C. S., On Combining Maximum Entropy Trip Matrix Estimation with User Optimal Assignment. *Transportation Research*, 22B (1988), pp. 69–79.

Fisk, C. S., Trip Matrix Estimation from Link Traffic Counts: The Congested Network Case. *Transportation Research*, 23B (1989), pp. 331–336.

Fisk, C. S., Traffic Assignment and the Lane Choice Problem. *Transportation Research*, 24B (1990), pp. 375–389.

Fisk, C. S. and Boyce, D. E., Alternative Variational Inequality Formulations of the Network Equilibrium-Travel Choice Problem. *Transportation Science*, 17 (1983a), pp. 454–463.

Fisk, C. S. and Boyce, D. E., A Note on Trip Matrix Estimation from Link Traffic Count Data. *Transportation Research*, 17B (1983b), pp. 245–250.

Fisk, C. S. and Nguyen, S., Solution Algorithm for Network Equilibrium Models with Asymmetric User Costs. *Transportation Science*, 16 (1982), pp. 316–381.

Florian, M., A Traffic Equilibrium Model of Travel by Car and Public Transit Modes. *Transportation Science*, 11 (1977), pp. 166–179.

Florian, M., Nguyen, S., and Ferland, J., On the Combined Distribution-Assignment of Traffic. *Transportation Science*, 9 (1975), pp. 43–53.

Florian, M., Wu, J. H., and He, S., A Multi-Class Multi-Mode Variable Demand Network Equilibrium Model with Hierarchical Logit Structures. *INRO Solutions, Cer for Research on Transportation, Universite de Montreal*, (1999), pp. 1–12.

Florian, M. and Hearn, D., Network Equlibrium Models and Algorithms. In BALL ET AL. (1995C), pp. 485–550.

Florian, M. and Nguyen, S., A Method for Computing Network Equilibrium with Elastic Demand. *Transportation Science*, 8 (1974), pp. 321–332.

Florian, M. and Nguyen, S., A Combined Trip Distribution Modal Split and Trip Assignment Model. *Transportation Research*, 12 (1978), pp. 241–246.

Frank, A., Shepherd, B., Tandon, V., and Végh, Z., Node-Capacitated Ring Routing. Technical report, Egrerváry Research Group, `www.cs.elte.hu/egres`, 2001.

Friedman, M. and Boorstin, D. J., Epilogue: How to Plan and Pay for the Safe and Adequate Highways We Need. In GABRIEL ROTH: Roads in a Market Economy, pp. 223–245. Brookfield: Ashgate, 1996.

Friesz, T. L., Transportation Network Equilibrium, Design and Aggregation: Key Developments and Research Opportunities. *Transportation Research*, 19A (1985), pp. 413–427.

Friesz, T. L., Harker, P. T., and Tobin, R. L., Alternative Algorithms for the General Network Spatial Price Equilibrium Problem. *Journal of Regional Science*, 24 (1984), pp. 475–507.

Friesz, T. L., Gottfried, J. A., and Morlok, E. K., A Sequential Shipper-Carrier Network Model for Predicting Freight Flows. *Transportation Science*, 20 (1986), pp. 80–427.

Friesz, T. L., Luque, J., Tobin, R. L., and Wie, B.-W., Dynamic Network Traffic Assignment Considered as a Continuous Time Optimal Control Problem. *Operations Research*, 37 (1989), pp. 893–901.

Friesz, T. L., Bernstein, D., Smith, T. E., Tobin, R. L., and Wie, B.-W., A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Operations Research*, 41 (1993), pp. 179–191.

Friesz, T. L., Bernstein, D., Mehta, N. J., Tobin, R. L., and Ganjalizadeh, S., Day-to-Day Dynamic Network Disequilibria and Idealized Traveler Information Systems. *Operations Research*, 42 (1994), pp. 1120–1136.

Friesz, T. L., Shah, S., and Bernstein, D., Disequilibrium Network Design: A New Paradigm for Transportation Planning and Control. In Lundqvist, L., Mattsson, L.-G., and Kim, T. J., eds.: *Network Infrastructure and the Urban Environment*, pp. 99–111. New York: Springer, 1998.

Fukushima, M., On the Dual Approach to the Traffic Assignment Problem. *Transportation Research*, 18B (1984), pp. 235–245.

Gabszewicz, J. J. and Thisse, J.-F., Spatial Competition and the Location of Firms. In Arnott, R., ed.: *Location Theory*. Chur: Harwood Academic, 1986.

Gale, D., A Theorem on Flows in Networks. *Pacific Journal of Mathematics*, 7 (1957), pp. 1073–1082.

Gallego, G. and van Ryzin, G. J., A Multiproduct Dynamic Pricing Problem and its Applications to Network Yield Management. *Operations Research*, 45 (1997), pp. 24–41.

Gartner, N. H., Optimal Traffic Assignment with Elastic Demands: A Review, Part I. Analysis Framework and Part II. Algorithmic Approaches. *Transportation Science*, 14 (1980), pp. 174–208.

Gaubert, S. and MaxPlus, Methods and Applications of (max,+) Linear Algebra. *14th Symposium on Theoretical Aspects of Computer Science, Hansestadt Lübeck, Germany, Feb. 27 March 1, 1997 (to appear in Springer Lecture Notes in Computer Science)*, (1997).

Georgescu-Roegen, N., Entropy, Vol. 2. In EATWELL ET AL. (1987), pp. 153–156.

Gibson, C. and Pavlak, T. J., Policy Options for Financing Georgia's Highway Infrastructure. Technical report, Carl Vonson Institute of Government, University of Georgia, 2002.

Gomory, R. E. and Baumol, W. J., Integer Programming and Pricing. *Econometrica*, 28 (1960), pp. 521–550.

Götz, G., Der deutsche Telekommunikationsmarkt zwei Jahre nach der vollständigen Marktöffnung: Eine Bestandsaufnahme aus Sicht der Verbraucher. *Perspektiven der Wirtschaftspolitik*, 2 (2001), pp. 167–183.

Goverde, R. M. P., Optimal Scheduling of Connections in Railway Systems. *Paper prepared for the 8th WCTR, Antwerp, July, 1998, revised version, Delft University of Technology*, (1998).

Goverde, R. M. P., *Synchronization Control of Scheduled Train Services to Minimize Passenger Waiting Times*. Delft: TRAIL Research School, 1998.

Goverde, R. M. P., Bovy, P. H. L., and Olsder, G. J., The Max-Plus Algebra Approach to Transportation Problems. In Meersman, H., van de Voorde, E., and Winkelmans, W., eds.: *World Transport Research Vol.3: Transport Modelling/Assessment*, pp. 377–391. New York: Elsevier, 2000.

Greenberg, H. J., *Mathematical Programming Glossary*. `www.cudenver.edu/~hgreenbe/glossary/`: World Wide Web, 1996–2003.

Grötschel, M., Monma, C. L., and Stoer, M., Design of Survivable Networks. In BALL ET AL. (1995B), pp. 617–672.

Hakimi, S. L., Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems. *Operations Research*, 13 (1965), pp. 462–475.

Hall, M. A., Properties of the Equilibrium State in Transportation Networks. *Transportation Science*, 12 (1978), pp. 208–216.

Hall, R. W., Graphical Interpretation of the Transportation Problem. *Transportation Science*, 23 (1989), pp. 37–45.

Hanke, S. H., Privatization, Vol. 3. In EATWELL ET AL. (1987), pp. 976–977.

Hardy, W. C., *QoS: Measurement and Evaluation of Telecommunications Quality of Service*. New York: John Wiley, 2001.

Harker, P. T., Multiple Equilibrium Behaviors on Networks. *Transportation Science*, 22 (1988), pp. 39–46.

Hearn, D. W., Lawphongpanich, S., and Nguyen, S., Convex Programming Formulations of the Asymmetric Traffic Assignment Problem. *Transportation Research*, 18B (1984), pp. 357–365.

Hearn, D. W. and Ramana, M. V., Solving Congestion Toll Pricing Models. In MARCOTTE, NGUYEN (1998), pp. 109–124.

Helgason, R. V. and Kennington, J. L., Primal-Simplex Algorithms for Minimum Cost Network Flows. In BALL ET AL. (1995B), pp. 85–133.

Heuser, H., *Lehrbuch der Analysis*, Bd. 2. 7 ed. Stuttgart: Teubner, 1992.

Heydecker, B. G., On the Definition of Traffic Equilibrium. *Transportation Research*, 20B (1986), pp. 435–440.

Hillier, F. S. and Lieberman, G. J., *Introduction to Operations Research*. 6 ed. New York: McGraw-Hill, 1995.

Hochman, O. and Pines, D., Cost of Adjustment and Demolition Cost in Residential Constructuion, and their Effects on Urban Growth. *Journal of Urban Economics*, 7 (1980), pp. 2–19.

Horner, M. W. and O'Kelly, M. E., Embedding Economies of Scale Concepts for Hub Network Design. *Journal of Transport Geography*, 9 (2001), pp. 255–265.

Hotelling, H., Stability in Competition. *The Economic Journal*, 39 (1929), pp. 41–57.

Huff, D. L., A Programmed Solution for Approximating an Optimum Retail Location. *Land Economics*, 42 (1966), pp. 293–303.

Inman, R. P., A Generalized Congestion Function for Highway Travel. *Journal of Urban Economics*, 5 (1978), pp. 21–34.

Iri, M., On an Extension of the Maximum-Flow Minimum-Cut Theorem to Multicommodity Flows. *Journal of the Operations Research Society of Japan*, 13 (1971), pp. 129–135.

Janson, B. N. and Zozaya-Gorostiza, C., The Problem of Cyclic Flows in Traffic Assignment. *Transportation Research*, 21B (1987), pp. 299–310.

Jarvis, J. J. and Martinez, O. M., A Sensitivity Analysis of Multicommodity Network Flows. *Transportation Science*, 11 (1977), pp. 299–306.

Jones, P. C., Zydiak, J. L., and Hopp, W. J., Stationary Dual Prices and Depreciation. *Mathematical Programming*, 41 (1988), pp. 357–366.

Jones, R. H. and Steele, N. C., *Mathematics in Communication Theory*. Chichester: Ellis Horwood, 1989.

Jones, P. C. and Theise, E. S., On the Equivalence of Competitive Transportation Markets and Congestion in Spatial Price Equilibrium Models. *Transportation Science*, 23 (1989), pp. 112–117.

Jung, V. and Warnecke, H.-J., eds., *Handbuch für die Telekommunikation*. Berlin: Springer, 1998.

Jungnickel, D., *Graphen, Netzwerke und Algorithmen*. 3 ed. Mannheim: Wissenschaftsverlag, 1994.

Kakusho, O. and Onaga, K., On Feasibility Conditions of Multicommodity Flows in Networks. *Transactions on Circuit Theory*, 18 (1971), pp. 425–429.

Karp, R. M., A Characterization of the Minimum Cycle Mean in a Digraph. *Discrete Mathematics*, 23 (1978), pp. 309–311.

Katz, M. L. and Shapiro, C., Network Externalities, Competition, and Compatibility. *The American Economic Review*, 75 (1985), pp. 424–440.

Keeler, T. E. and Small, K. A., Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways. *Journal of Political Economy*, 85 (1977), pp. 1–25.

Kinderlehrer, D. and Stampacchia, G., *An Introduction to Variational Inequalities and their Applications*. New York: Academic, 1980.

Kleinrock, L., *Queuing Systems, Volume I: Theory*. New York: Whiley, 1975.

Kleinrock, L., Creating a Mathematical Theory of Computer Networks. *Operations Research*, 50 (2002), pp. 125–131.

Knight, F., Some Fallacies in the Interpretation of Social Cost. *Quarterly Journal of Economics*, 38 (1924), pp. 582–606.

Koopmans, T. C. and Beckmann, M., Assignment Problems and the Location of Economic Activities. *Econometrica*, 25 (1957), pp. 53–76.

Korbut, A. A. and Finkelstein, J. J., *Diskrete Optimierung*. Berlin: Akademie Verlag, 1971.

Kraft, E. R., Srikar, B. N., and Phillips, R. L., Revenue Management in Railroad Applications. *Journal of the Transportation Research Forum*, 39 (2000), pp. 157–176.

Lam, W. H. K. and Huang, H.-J., A Combined Trip Distribution and Assignment Model for Multiple User Classes. *Transportation Research*, 26B (1992), pp. 275–287.

Lam, W. H. K. and Huang, H.-J., A Combined Activity/Travel Choice Model for Congested Road Networks with Queues. *Transportation*, 29 (2002), pp. 5–29.

Larsson, T. and Patriksson, M., An Augmented Lagrangean Dual Algorithm for Link Capacitated Side Constrained Traffic Assignment Problems. *Transportation Research*, 29B (1995), pp. 433–455.

Larsson, T. and Patriksson, M., Side Constrained Traffic Equilibrium Models — Traffic Management Through Link Tolls. In MARCOTTE, NGUYEN (1998), pp. 125–151.

Larsson, T. and Patriksson, M., Side Constrained Traffic Equilibrium Models — Analysis, Computation and Applications. *Transportation Research*, 33B (1999), pp. 233–264.

LeBlanc, L. J. and Abdulaal, M., Combined Mode Split-Assignment and Distribution-Model Split-Assignment Models with Multiple Groups of Travelers. *Transportation Science*, 16 (1982), pp. 430–442.

LeBlanc, L. J. and Abdulaal, M., A Comparison of User-Optimum versus System-Optimum Traffic Assignment in Transportation Network Design. *Transportation Research*, 18B (1984), pp. 115–121.

Lee, Y., Lim, B. H., and Park, J. S., A Hub Location Problem in Designing Digital Data Service Networks: Lagrangian Relaxation Problem. *Location Science*, 4 (1996), pp. 185–194.

Link, H., Dodgson, J. S., Maibach, M., and Herry, M., *The Cost of Road Infrastructure and Congestion in Europe*. Heidelberg: Physica, 1999.

Lo, H. K. and Chen, A., Traffic Equilibrium Problem with Route-Specific Costs: Formulation and Algorithms. *Transportation Research*, 34B (2000), pp. 493–513.

Lo, H. K. and Szeto, W. Y., A Cell-Based Variational Inequality Formulation of the Dynamic User Optimal Assignment Problem. *Transportation Research*, 36B (2002), pp. 421–443.

MacKie-Mason, J. K. and Varian, H. R., Pricing Congestible Network Resources. *IEEE Journal on Selected Areas in Communications*, 13 (1995), pp. 1141–1149.

Magnanti, T. L., Mirchandani, P. B., and Vachani, R., Modeling and Solving the Two-Facility Capacitated Network Loading Problem. *Operations Research*, 43 (1995), pp. 142–157.

Magnanti, T. L. and Wong, R. T., Network Design and Transportation Planning: Models and Algorithm. *Transportation Science*, 18 (1984), pp. 1–55.

Mahmassani, H. S. and Chang, G.-L., On Boundedly Rational User Equilibrium in Transportation Systems. *Transportation Science*, 21 (1987), pp. 89–99.

Mahmassani, H. S. and Herman, R., Dynamic User Equilibrium Departure Time and Route Choice on Idealized Traffic Arterials. *Transportation Science*, 18 (1984), pp. 362–384.

Mancinelli, E., Cohen, G., Gaubert, S., Quadrat, J.-P., and Rofman, E., On Traffic Light Control of Regular Towns. Rapport de recherche RR-4276, Institut National de Recherche en Informatique en Automatique, INRIA, 2001. PDF.

Marcotte, P. and Nguyen, S., eds., *Equilibrium and Advanced Transportation Modelling*. Boston: Kluwer, 1998.

Mas-Colell, A., Whinston, M. D., and Green, J. R., *Microeconomic Theory*. New York: Oxford University, 1995.

Mason, L. G., Equilibrium Flows, Routing Patterns and Algorithm for Store-and-Forward Networks. *Large Scale Systems*, 8 (1985), pp. 187–209.

McBride, R. D. and Mamer, J. W., Solving the Undirected Multicommodity Flow Problem Using a Shortest Path-Based Pricing Algorithm. *Networks*, 38 (2001), pp. 181–188.

Merchant, D. K. and Nemhauser, G. L., A Model and an Algorithm for the Dynamic Traffic Assignment Problems. *Transportation Science*, 12 (1978), pp. 183–207.

Miller, T. C., Friesz, T. L., and Tobin, R. L., *Equilibrium Facility Loaction on Networks*. Berlin: Springer, 1996.

Mohring, H., ed., *The Economics of Transport, Vol. I.* Aldershot: Elgar, 1994.

Nagurney, A., *Network Economics: A Variational Inequality Approach*. 2 ed. Boston: Kluwer, 1999.

Nagurney, A., Congested Urban Transportation Networks and Emission Paradoxes. *Transportation Research*, 5D (2000), pp. 145–151.

Nagurney, A., A Multiclass, Multicriteria Traffic Network Equilibrium Model. *Mathematical and Computer Modelling*, 32 (2000), pp. 393–411.

Nagurney, A. and Dong, J., A Multiclass, Multicriteria Traffic Network Equilibrium Model with Elastic Demand. *Transportation Research*, 36B (2002a), pp. 445–469.

Nagurney, A. and Dong, J., *Supernetwork: Decision-Making for the Information Age*. Northampton, Mass.: Elgar, 2002c.

Nagurney, A., Dong, J., and Mokhtarian, P. L., Multicriteria Network Equilibrium Modeling with Variable Weights for Decision-Making in the Information Age with Applications to Telecommunication and Teleshopping. *Journal of Economic Dynamics and Control*, 26 (2002b), pp. 1629–1650.

Neumann, K. and Morlock, M., *Operations Research*. München: Carl Hanser, 1993.

Ng, P. H. and Rardin, R. L., Commodity Family Extended Formulations of Uncapacitated Fixed Charge Network Flow Problems. *Networks*, 30 (1997), pp. 57–71.

Nguyen, S., Estimating an OD Matrix from Network Data: A Network Equilibrium Approach. Technical report, Université de Montréal, Centre de recherche sur les transports, Départment d'informatique, Publication # 87, 1977.

O'Kelly, M. E., The Location of Interacting Hub Facilities. *Transportation Science*, 20 (1986), pp. 92–106.

O'Kelly, M. E., A Geographer's Analysis of Hub-and-Spoke Networks. *Journal of Transport Geography*, 6 (1998), pp. 171–186.

O'Kelly, M. E. and Bryan, D. L., Hub Location with Flow Economies of Scale. *Transportation Research*, 32B (1998), pp. 605–616.

Okutani, I., Equilibrium Flows in a Network with Congested Links. In VOLMULLER, HAMERSLAG (1984), pp. 253–271.

O'Mahony, M., Geraghty, D., and Humphreys, I., Distance and Time Based Road Pricing Trial in Dublin. *Transportation*, 27 (2000), pp. 269–283.

de Palma, A., Ben-Akiva, M., Lefévre, C., and Litinas, N., Stochastic Equilibrium Model of Peak Period Traffic Congestion. *Transportation Science*, 4 (1983), pp. 430–453.

de Palma, A. and Lindsey, R., Optimal Timetables for Public Transportation. *Transportation Research*, 35B (2001), pp. 789–813.

de Palma, A. and Nesterov, Y., Optimization Formulations and Static Equilibrium in Congested Transportion Networks. Working Paper 9717, THEMA, Université de Cergy-Pontoise, 1998.

Pas, E. I. and Principio, S. L., Braess' Paradox: Some New Insights. *Transportation Research*, 31B (1997), pp. 265–276.

Patriksson, M., *The Traffic Assignment Problem. Models and Methods*. Utrecht: VSP, 1994.

Pigou, A. C., *The Economics of Welfare*. London: Macmillan, 1920. (reprint 1978, AMS: New York).

Pöppe, C., Paradoxes Verhalten. *Spektrum Wissenschaft, Dossier: Verkehr und Auto*, (1995), pp. 40–41.

Powell, W. B., Jaillet, P., and Odoni, A., Stochastic and Dynamic Networks and Routing. In BALL ET AL. (1995C), pp. 141–295.

Powell, W. B. and Sheffi, Y., The Convergence of Equilibrium Algorithms with Predetermined Step Sizes. *Transportation Science*, 16 (1982), pp. 45–55.

Prashker, J. N. and Bekhor, S., Some Observations on Stochastic User Equilibrium and System Optimum of Traffic Assignment. *Transportation Research*, 34B (2000), pp. 277–291.

Quadrat, J.-P. and MaxPlus, Max-Plus Algebra and Applications to System Theory and Optimal Control. In *Proceedings of the International Congress of Mathematicians, Zurich, Switzerland, August 1994*, pp. 1502–1511, Boston: Birkhäuser, 1995.

Ran, B. and Boyce, D. E., *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*. 2 ed. Berlin: Springer, 1996.

Reith, S. and Vollmer, H., Wer wird Millionär. *c't*, Heft 7 (2001), pp. 240–251.

Rockafellar, R. T., *Convex Analysis*. Princeton, New Jersey: Princeton University, 1972.

Rohatgi, V. K., *An Introduction to Probability Theory and Mathematical Statistics*. New York: Wiley, 1976.

Roth, G. J., *Roads in a Market Economy*. Brookfield: Ashgate, 1996.

Safwat, K., Ali, N., and Magnanti, T. L., A Combined Trip Generation, Trip Distribution, Modal Split, and Trip Assignment Model. *Transportation Science*, 18 (1988), pp. 14–30.

Samuelson, P. A., Spatial Price Equilibrium and Linear Programming. *The American Economic Review*, 33 (1952), pp. 283–303.

Samuelson, P. A., The Transfer Problem and Transport Costs II: Analysis of Effects of Trade Impediments. *The Economic Journal*, 64 (1954), pp. 264–289.

Sasaki, T. and Inouye, H., Traffic Assignment by Analogy to Electric Circuits. In BUCKLEY (1974), pp. 495–518.

Schöbel, A., A Model for the Delay Management Problem Based on Mixed-Integer-Programming. *Electronic Notes in Theoretical Computer Science*, 20 (2001), pp. 1–10.
www1.elsevier.com/gej-ng/31/29/23/85/27/show/toc.htt.

de Schutter, B., *Max-Algebraic System Theory for Discrete Event Systems*. Doctoral thesis, Katholieke Universiteit Leuven, Leuven, 1996.

Scitovsky, T., Two Concepts of External Economies. *Journal of Political Economy*, 62 (1954), Nr. 2, pp. 143–151.

Sen, A. and Smith, T. E., *Gravity Models of Spatial Interaction Behavior*. Berlin: Springer, 1995.

Sharkey, W. W., Natural Monopoly, Vol. 3. In EATWELL ET AL. (1987), pp. 603–605.

Sharkey, W. W., Network Models in Economics. In BALL ET AL. (1995C), pp. 713–765.

Sheffi, Y., *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Englewood Cliffs, New Jersey: Prentice-Hall, 1985.

Shephard, R. W., *Theory of Cost and Production Functions*. Princeton: Princeton University, 1953.

Sherali, H. D., Sivanandan, R., and Hobeika, A. G., A Linear Programming Approach for Synthesizing Origin-Destination Trip Tables from Link Traffic Volumes. *Transportation Research*, 28B (1994), pp. 213–233.

Shy, O., *The Economics of Network Industries*. Cambridge: Cambridge University, 2001.

Simon, H., A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69 (1955), pp. 99–118.

Skorin-Karpov, D., Skorin-Karpov, J., and O'Kelly, M., Tight Linear Programming Relaxations of Uncapacitated p-Hub Median Problems. *European Journal of Operational Research*, 94 (1996), pp. 582–593.

Small, K. A., Using the Revenues from Congestion Pricing. *Transportation*, 19 (1992), pp. 359–381.

Small, K. A., Winston, C., and Evans, C. A., *Road Work: A New Highway Pricing and Investment Policy*. Washington, D.C.: Brookings, 1992.

Smith, M. J., The Existence, Uniqueness and Stability of Traffic Equilibria. *Transportation Research*, 13B (1979a), pp. 295–304.

Smith, M. J., The Marginal Cost Taxation of a Transportation Network. *Transportation Research*, 13B (1979b), pp. 237–242.

Smith, M. J., An Algorithm for Solving Asymmetric Equilibrium Problems with a Continuous Cost-Flow Function. *Transportation Research*, 17B (1983), pp. 365–371.

Smith, M. J., A Descent Algorithm for Solving a Variety of Monotone Equilibrium Problems. In VOLMULLER, HAMERSLAG (1984), pp. 273–297. Cited as Smith (1984a).

Smith, M. J., The Stability of a Dynamic Model of Traffic Assignment — an Application of a Method of Lyapunov. *Transportation Science*, 18 (1984b), pp. 245–252.

Smith, M. J., Two Alternative Definitions of Traffic Equilibria. *Transportation Research*, 18B (1984c), pp. 63–65.

Steenbrink, P. A., *Optimization of Transport Networks*. New York: Wiley, 1974.

Stoer, M., *Design of Survivable Networks*. Berlin: Springer, 1992.

Stopher, P. R. and Meyburg, A. H., *Urban Transportation Modeling and Planning*. Lexington, Mass.: Lexington Books, 1975.

Takayama, T. and Judge, G. G., *Spatial and Temporal Price and Allocation Models*. Amsterdam: North-Holland, 1971.

Talluri, K. T. and van Ryzin, G. J., An Analysis of Bid-Price Controls for Network Revenue Management. *Management Science*, 44 (1998), pp. 1577–1593.

Talluri, K. T. and van Ryzin, G. J., A Randomized Linear Programming Method for Computing Network Bid Prices. *Transportation Science*, 33 (1999), pp. 207–216.

Theil, H., *Economics and Information*. Amsterdam: North-Holland, 1967.

Tobin, R. L., Sensitivity Analysis for Variational Inequalities. *Journal of Optimization Theory and Applications*, 48 (1986), pp. 191–204.

Tobin, R. L., Sensitivity Analysis for General Spatial Price Equilibria. *Journal of Regional Science*, 27 (1987), pp. 77–102.

Tobin, R. L. and Friesz, T. L., Sensitivity Analysis for Equilibrium Network Flow. *Transportation Science*, 22 (1988), pp. 242–250.

Transportation Research Board (TRB), Washington, D.C., *Highway Capacity Manual. Special Report 209*, 3 ed., 1994.

U. S. Bureau of Public Roads (BPR), Urban Planning Division, Washington, D.C., *Traffic Assignment Manual*, 1964.

Vanderbei, R. J., *Linear Programming: Foundations and Extensions*. Boston: Kluwer, 1998.

Varian, H. R., Market Structure in the Network Age. In *Understanding the Digital Economy*, Department of Commerce, Washington, D.C., 1999.

Verhoef, E. T., Nijkamp, P., and Rietveld, P., Second-Best Congestion Pricing: The Case of an Untolled Alternative. *Journal of Urban Economics*, 40 (1996), pp. 279–302.

Vicente, L. N. and Calamai, P. H., Bilevel and Multilevel Programming: A Bibliography Review. *Journal of Global Optimization*, 6 (1994), pp. 1–16. Current version available on `solstice.uwaterloo.ca/~phcalama /papers/bilevel-review.ps`.

Vickrey, W. S., Congestion Charges and Welfare: Some Answers to Sharp's Doubts. *Journal of Transport Economics and Policy*, 2 (1968), pp. 107–118.

Vickrey, W. S., Congestion Theory and Transport Investment. *The American Economic Review, Papers and Proceedings*, 59 (1969), pp. 251–260. Reprint in Mohring (1994).

Viner, J., Cost Curves and Supply Curves. *Zeitschrift für Nationalökonomie*, 111 (1931), pp. 23–46.

Viton, P. A., Private Roads. *Journal of Urban Economics*, 37 (1995), pp. 260–289.

Volmuller, J. and Hamerslag, R., eds., *Proceeding on Transportation and Traffic Theory, Delft, 1984*. VNU Science: Utrecht, The Netherlands, 1984.

Walk, M., *Theory of Duality in Mathematical Programming*. Wien: Springer, 1989.

Walters, A. A., The Theory and Measurement of Private and Social Cost of Highway Congestion. *Econometrica*, 29 (1961), pp. 676–699. Reprint in Mohring (1994).

Walters, A. A., *The Economics of Road User Charges*. Baltimore: Hopkins, 1968.

Walters, A. A., Congestion, Vol. 1. In EATWELL ET AL. (1987), pp. 570–573.

Wardrop, J. G., Some Theoretical Aspects of Road Traffic Research. *Proceeding of the Institution of Civil Engineers, Part II*, 1 (1952), pp. 325–378. Reprint in Mohring (1994).

van Wee, B., Hagoort, M., and Annema, J. A., Accessibility Measures with Competition. *Journal of Transport Geography*, 9 (2001), pp. 199–208.

Wessäly, R., *DImensioning Survivable Capacitated NETworks (DISCNET)*, PhD thesis TU Berlin. Göttingen: Cuvillier, 2000.

Williams, H. C. W. L., On the Formulation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment and Planning*, 9A (1977), pp. 285–344.

Willig, R. D., Contestable Markets, Vol. 1. In EATWELL ET AL. (1987), pp. 618–622.

Wilson, A. G., A Statistical Theory of Spatial Distribution Models. *Transportation Research*, 1 (1967), pp. 253–269.

Wilson, A. G., *Entropy in Urban and Regional Modelling*. New York: Pion, 1970.

Wilson, A. G., Catastrophe Theory and Urban Modelling: An Application to Modal Choice. *Environment and Planning*, 8A (1976), pp. 351–356.

Wilson, J. D., Optimal Road Capacity in the Presence of Unpriced Congestion. *Journal of Urban Economics*, 13 (1983), pp. 337–357.

Wilson, A. G. and Senior, M. L., Some Relationships between Entropy Maximizing Models, Mathematical Programming, and their Duals. *Journal of Regional Science*, 14 (1974), pp. 207–215.

Wiseman, J., Peak-Load Pricing, Vol. 3. In EATWELL ET AL. (1987), pp. 822–824.

Yang, H., Evaluating the Benefits of a Combined Route Guidance and Road Pricing System in a Traffic Network with Recurrent Congestion. *Transportation*, 26 (1999), pp. 299–322.

Yang, H. and Bell, M. G. H., Traffic Restraint, Road Pricing and Network Equilibrium. *Transportation Research*, 31B (1997), pp. 303–314.

Yang, H., Bell, M., and Meng, Q., Modeling the Capacity and Level of Service of Urban Transportation Networks. *Transportation Research*, 34B (2000), pp. 255–275.

Yang, H. and Huang, H.-J., Principle of Marginal-Cost Pricing: How Does It Work in a General Road Network? *Transportation Research*, 32A (1998), pp. 45–54.

Yang, H. and Lam, W. H. K., Optimal Road Tolls under Conditions of Queueing and Congestion. *Transportation Research*, 30A (1996), pp. 319–332.

Zhang, H. M., A Note on Highway Capacity. *Transportation Research*, 35B (2001), pp. 929–937.

Zhang, D. and Nagurney, A., On the Local and Global Stability of a Travel Route Choice Adjustment Process. *Transportation Research*, 30B (1996), pp. 245–262.

Index

A

AASHO 251
accessibility 14, 18
adjacency matrix 311
arc 55
arrival rate 278
ascent
  direction 133
  steepest 133
assignment problem
  capacity 284
  trip 73
availability 18, 241

B

bandwidth 18, 110
Bellman's equations 107
bilevel programming 119, 216
birth-death model 280
blocking
  call 291
  time 291
boundedness of production 96, 348
BPR .... *acronym* U.S. Bureau of Public Roads
BPR function 113
Braess' paradox 163, 232
branch 55
budget design problem 250
buffer times 306, 329

C

capacity
  assignment problem 284
  cut 62
  link 56
  network 97
  node 122
  trip 121
cdf ..... *acronym* cumulative distribution function
centroids 55
chain 58, 312
charging, element based 115
circuit 312
circuit-switched traffic 102
circulation 63
compensating variation 195
complementary slackness 40
conditional incremental cost 352
congestion 96, 142
connectivity 295
consumer's surplus 195
correspondence
  input 348
  output 95, 346
cost function 85, 94, 105, 141
  separable 141
cost-benefit analysis 227
counter equations 336
coverage 229, 270
cut 62
  capacity 62
  flow value 62
cycle 68, 312

critical 320
mean 313
cyclicity 324

D

dater equations 336
DEDS .. *acronym* discrete event dynamic system
delay 18
overall system 283
departure rate 279
DES ..... *acronym* discrete event system
descent
direction 133
steepest 133
design problem
budget 250
fixed charge 247
network improvement 247
system optimum 249
user optimum 250
deterrence function 184
digraph 55
weighted 62
directed graph 55
direction 132
of steepest ascent 133
of steepest descent 133
dispersion factor 174
disposability
of inputs 96, 348
of outputs 96, 348
distance 106, 313
dual points, pair 207
duality
Lagrangean 44
linear programming 39
theorem, strong 40, 44
theorem, weak 39, 44

E

edge 55
eigenvalue 317
eigenvector 317
Engset
loss formula 291
rejection formula 291
entropy 172
entropy maximization 71
envelope theorem 49
equilibrium
existence 161
Nash 120
spatial price 92, 222
stability 161
stochastic user 169, 175
uniqueness 146
user 108, 190
von Stackelberg 120
Wardrop 108, 191
equivalent variation 195
Erlang 287
delay formula 292
distribution 293
loss formula 290
Euclidean norm 133
externality
pecuniary 116, 143
technologial 143

F

facility location problem 274
failure state 297
Farkas' Theorem 41
feasible set 42
Fenchel's Duality Theorem 212
Fenchel's inequality 202
fixed point problem 135, 320
flow
conservation 56, 63
equilibrated 159
network 62
user optimized 158
Frank-Wolfe algorithm 134
frequency 307, 344
function
(anti)monotone 129
closed 202
conjugate concave 202
conjugate convex 202
proper 202

G

Gordan's Theorem 41
gradient 45
graph
(strongly) connected 58
bipartite 64

critical 320
precedence 311
simple 55
gravity model 71, 184

H

Hessean matrix 46
hub-and-spoke network 269

I

identity matrix 315
incidence matrix
arc/path 60
destination/path 62
node/arc 59
OD/path 60
origin/path 62
inequality
Fenchel's 202
Mahler's 189
inflow 56
input correspondences 348
input requirement set 100, 348
input space 345
interarrival time 318

J

Jacobian matrix 48, 54
jitter 18

K

Karp's theorem 319
Karush-Kuhn-Tucker
conditions 48
vector 43
Kendall's notation 277
Kirchhoff's law
first (current) 63
second (voltage) 87
KKT ..... *see* Karush-Kuhn-Tucker
Knapsack problem 124
Kruskal's algorithm 229

L

Legendre transform 202
line integral 131
path independent 131
link 55
link performance function 112
Little's Law 279
loading 103
logit model 174
loop 313
loss formula
Erlang 290

M

marginal cost pricing 266
Markov process 279
matrix
adjacency 311
cyclicity 324
Hessean 46
incidence 59
irreducible 319
Jacobi 48
order-*d* periodic 324
positive semidefinite 46
max-plus algebra 313
maximum cycle mean 313
maximum flow problem 78
MCCI ..... *see* Minimum Capacity Cost Installation
Metcalfe's Law 24
min-hop routing 107
minimum cost capacity installation ..245
minimum cut problem 78
modal split 72
monotonicity 129
MST .. *acronym* minimum spanning tree
multiclass traffic 72
multiclass-user network 155
multicommodity flow problem 101
multicriteria traffic demand 73
multigraph 55

N

Nash equilibrium 110, 120
natural monopoly 254
net benefit of traffic 200
network 55
(strongly) connected 58
capacitated 56
hub-and-spoke 269
multiclass-user 155
neural 65
order 55
size 55

throughput 59, 282
topology 230
neural network 65
node 55
adjacent 65
sink 56
source 56
transshipment 56
normal operating state 296

O

occupation rate 279
OD ..... *acronym* origin destination
optimum
system, traffic assignment 103
user, price sensitive traffic demand 190
user, traffic assignment 103
order of a network 55
origin destination pair 58
orthogonal projection 135
outflow 56
output correspondence 346
output space 346

P

packet-switched traffic 102
pair of dual points 207
partition 62
path 57, 312
algebra 314
closed 57
elementary 57
length 312
simple 57
weight 312
pdf *acronym* probability density function
peak-load pricing 353
perturbation function 43
Petri net 335
Pigou tax 150
pmf ..*acronym* probability mass function
possibility of inaction 95, 348
potential function 132
precedence
constraint 306
graph 311
Prim's algorithm 229
probit model 174
problem
fixed point 135
nonlinear complementary 135
variational inequality 127
production possibility set 95
program
linear 39
ordinary convex 42
primal and dual 38
programming
bilevel 119, 216
projection, orthogonal 135
pseudoconvexity 47

Q

quality of service 18
queue
service discipline 276
queuing 219
delay 277

R

random utility 71
rejection formula
Engset 291
reliability 18, 111, 228, 241
revenue maximization 124
road pricing 150
route 57
closed 57
elementary 57
simple 57
route choice
stochastic 169
Wardrop (deterministic) 108
routing 103
pattern 164
Roy's theorem 188

S

saddle-point 48
saturation, of a link 115
schedule 306
service rate 278
Shannon's theorem 114
Shephard's theorem 189
shortest path problem 107
single transit 10
size of a network 55

spatial price equilibrium 92, 222
spectral problem 317
spectral radius 319
spectrum 318
stability 306
Stackelberg equilibrium, von 120
Steiner tree 229
Stirling's approximation 170
store-and-forward switching 22
strongly connected network 58
subdifferential 45
subgradient 45
superdifferential 45
supergradient 45
survivability 295
synchronization 326

T

theorem
  complementary slackness 40
  Farkas 41
  Fenchel's duality 212
  Gordan 41
  Karp 319
  Karush-Kuhn-Tucker 49
throughput 59, 282, 319
timetable 306
  realistic 327
topology, network 230
traffic
  assignment problem 105, 147
  circuit-switched 13, 22
  creation 185
  diminution 185
  diversion 246
  packet-switched 13, 22
  stochastic assignment 174
transportation problem 73
transshipment node 56
transshipment problem 73
tree 229
  minimum spanning 228
  spanning 229
trip
  assignment 73
  distribution 71
  generation 70
  table 58

U

user optimum
  fixed traffic demand 103
  price sensitive traffic demand .. 190

V

variational inequality problem 127
vertex 55

W

Wardrop equilibrium 108, 191
  stochastic 169
Wardrop's principles 108, 169
weight
  edge 62, 312
  path 312
willingness to pay 175, 194
  marginal 187
Wold's theorem 188

Lecture Notes in Economics and Mathematical Systems

For information about Vols. 1–470
please contact your bookseller or Springer-Verlag

Vol. 471: N. H. M. Wilson (Ed.), Computer-Aided Transit Scheduling. XI, 444 pages. 1999.

Vol. 472: J.-R. Tyran, Money Illusion and Strategic Complementarity as Causes of Monetary Non-Neutrality. X, 228 pages. 1999.

Vol. 473: S. Helber, Performance Analysis of Flow Lines with Non-Linear Flow of Material. IX, 280 pages. 1999.

Vol. 474: U. Schwalbe, The Core of Economies with Asymmetric Information. IX, 141 pages. 1999.

Vol. 475: L. Kaas, Dynamic Macroeconomics with Imperfect Competition. XI, 155 pages. 1999.

Vol. 476: R. Demel, Fiscal Policy, Public Debt and the Term Structure of Interest Rates. X, 279 pages. 1999.

Vol. 477: M. Théra, R. Tichatschke (Eds.), Ill-posed Variational Problems and Regularization Techniques. VIII, 274 pages. 1999.

Vol. 478: S. Hartmann, Project Scheduling under Limited Resources. XII, 221 pages. 1999.

Vol. 479: L. v. Thadden, Money, Inflation, and Capital Formation. IX, 192 pages. 1999.

Vol. 480: M. Grazia Speranza, P. Stähly (Eds.), New Trends in Distribution Logistics. X, 336 pages. 1999.

Vol. 481: V. H. Nguyen, J. J. Strodiot, P. Tossings (Eds.). Optimation. IX, 498 pages. 2000.

Vol. 482: W. B. Zhang, A Theory of International Trade. XI, 192 pages. 2000.

Vol. 483: M. Königstein, Equity, Efficiency and Evolutionary Stability in Bargaining Games with Joint Production. XII, 197 pages. 2000.

Vol. 484: D. D. Gatti, M. Gallegati, A. Kirman, Interaction and Market Structure. VI, 298 pages. 2000.

Vol. 485: A. Garnaev, Search Games and Other Applications of Game Theory. VIII, 145 pages. 2000.

Vol. 486: M. Neugart, Nonlinear Labor Market Dynamics. X, 175 pages. 2000.

Vol. 487: Y. Y. Haimes, R. E. Steuer (Eds.), Research and Practice in Multiple Criteria Decision Making. XVII, 553 pages. 2000.

Vol. 488: B. Schmolck, Ommitted Variable Tests and Dynamic Specification. X, 144 pages. 2000.

Vol. 489: T. Steger, Transitional Dynamics and Economic Growth in Developing Countries. VIII, 151 pages. 2000.

Vol. 490: S. Minner, Strategic Safety Stocks in Supply Chains. XI, 214 pages. 2000.

Vol. 491: M. Ehrgott, Multicriteria Optimization. VIII, 242 pages. 2000.

Vol. 492: T. Phan Huy, Constraint Propagation in Flexible Manufacturing. IX, 258 pages. 2000.

Vol. 493: J. Zhu, Modular Pricing of Options. X, 170 pages. 2000.

Vol. 494: D. Franzen, Design of Master Agreements for OTC Derivatives. VIII, 175 pages. 2001.

Vol. 495: I. Konnov, Combined Relaxation Methods for Variational Inequalities. XI, 181 pages. 2001.

Vol. 496: P. Weiß, Unemployment in Open Economies. XII, 226 pages. 2001.

Vol. 497: J. Inkmann, Conditional Moment Estimation of Nonlinear Equation Systems. VIII, 214 pages. 2001.

Vol. 498: M. Reutter, A Macroeconomic Model of West German Unemployment. X, 125 pages. 2001.

Vol. 499: A. Casajus, Focal Points in Framed Games. XI, 131 pages. 2001.

Vol. 500: F. Nardini, Technical Progress and Economic Growth. XVII, 191 pages. 2001.

Vol. 501: M. Fleischmann, Quantitative Models for Reverse Logistics. XI, 181 pages. 2001.

Vol. 502: N. Hadjisavvas, J. E. Martínez-Legaz, J.-P. Penot (Eds.), Generalized Convexity and Generalized Monotonicity. IX, 410 pages. 2001.

Vol. 503: A. Kirman, J.-B. Zimmermann (Eds.), Economics with Heterogenous Interacting Agents. VII, 343 pages. 2001.

Vol. 504: P.-Y. Moix (Ed.),The Measurement of Market Risk. XI, 272 pages. 2001.

Vol. 505: S. Voß, J. R. Daduna (Eds.), Computer-Aided Scheduling of Public Transport. XI, 466 pages. 2001.

Vol. 506: B. P. Kellerhals, Financial Pricing Models in Con-tinuous Time and Kalman Filtering. XIV, 247 pages. 2001.

Vol. 507: M. Koksalan, S. Zionts, Multiple Criteria Decision Making in the New Millenium. XII, 481 pages. 2001.

Vol. 508: K. Neumann, C. Schwindt, J. Zimmermann, Project Scheduling with Time Windows and Scarce Resources. XI, 335 pages. 2002.

Vol. 509: D. Hornung, Investment, R&D, and Long-Run Growth. XVI, 194 pages. 2002.

Vol. 510: A. S. Tangian, Constructing and Applying Objective Functions. XII, 582 pages. 2002.

Vol. 511: M. Külpmann, Stock Market Overreaction and Fundamental Valuation. IX, 198 pages. 2002.

Vol. 512: W.-B. Zhang, An Economic Theory of Cities.XI, 220 pages. 2002.

Vol. 513: K. Marti, Stochastic Optimization Techniques. VIII, 364 pages. 2002.

Vol. 514: S. Wang, Y. Xia, Portfolio and Asset Pricing. XII, 200 pages. 2002.

Vol. 515: G. Heisig, Planning Stability in Material Requirements Planning System. XII, 264 pages. 2002.

Vol. 516: B. Schmid, Pricing Credit Linked Financial Instruments. X, 246 pages. 2002.

Vol. 517: H. I. Meinhardt, Cooperative Decision Making in Common Pool Situations. VIII, 205 pages. 2002.

Vol. 518: S. Napel, Bilateral Bargaining. VIII, 188 pages. 2002.

Vol. 519: A. Klose, G. Speranza, L. N. Van Wassenhove (Eds.), Quantitative Approaches to Distribution Logistics and Supply Chain Management. XIII, 421 pages. 2002.

Vol. 520: B. Glaser, Efficiency versus Sustainability in Dynamic Decision Making. IX, 252 pages. 2002.

Vol. 521: R. Cowan, N. Jonard (Eds.), Heterogenous Agents, Interactions and Economic Performance. XIV, 339 pages. 2003.

Vol. 522: C. Neff, Corporate Finance, Innovation, and Strategic Competition. IX, 218 pages. 2003.

Vol. 523: W.-B. Zhang, A Theory of Interregional Dynamics. XI, 231 pages. 2003.

Vol. 524: M. Frölich, Programme Evaluation and Treatment Choise. VIII, 191 pages. 2003.

Vol. 525: S. Spinler, Capacity Reservation for Capital-Intensive Technologies. XVI, 139 pages. 2003.

Vol. 526: C. F. Daganzo, A Theory of Supply Chains. VIII, 123 pages. 2003.

Vol. 527: C. E. Metz, Information Dissemination in Currency Crises. XI, 231 pages. 2003.

Vol. 528: R. Stolletz, Performance Analysis and Optimization of Inbound Call Centers. X, 219 pages. 2003.

Vol. 529: W. Krabs, S. W. Pickl, Analysis, Controllability and Optimization of Time-Discrete Systems and Dynamical Games. XII, 187 pages. 2003.

Vol. 530: R. Wapler, Unemployment, Market Structure and Growth. XXVII, 207 pages. 2003.

Vol. 531: M. Gallegati, A. Kirman, M. Marsili (Eds.), The Complex Dynamics of Economic Interaction. XV, 402 pages, 2004.

Vol. 532: K. Marti, Y. Ermoliev, G. Pflug (Eds.), Dynamic Stochastic Optimization. VIII, 336 pages. 2004.

Vol. 533: G. Dudek, Collaborative Planning in Supply Chains. X, 234 pages. 2004.

Vol. 534: M. Runkel, Environmental and Resource Policy for Consumer Durables. X, 197 pages. 2004.

Vol. 535: X. Gandibleux, M. Sevaux, K. Sörensen, V. T'kindt (Eds.), Metaheuristics for Multiobjective Optimisation. IX, 249 pages. 2004.

Vol. 536: R. Brüggemann, Model Reduction Methods for Vector Autoregressive Processes. X, 218 pages. 2004.

Vol. 537: A. Esser, Pricing in (In)Complete Markets. XI, 122 pages, 2004.

Vol. 538: S. Kokot, The Econometrics of Sequential Trade Models. XI, 193 pages. 2004.

Vol. 539: N. Hautsch, Modelling Irregularly Spaced Financial Data. XII, 291 pages. 2004.

Vol. 540: H. Kraft, Optimal Portfolios with Stochastic Interest Rates and Defaultable Assets. X, 173 pages. 2004.

Vol. 541: G.-y. Chen, X. Huang, X. Yang, Vector Optimization. X, 306 pages. 2005.

Vol. 542: J. Lingens, Union Wage Bargaining and Economic Growth. XIII, 199 pages. 2004.

Vol. 543: C. Benkert, Default Risk in Bond and Credit Derivatives Markets. IX, 135 pages. 2004.

Vol. 544: B. Fleischmann, A. Klose, Distribution Logistics. X, 284 pages. 2004.

Vol. 545: R. Hafner, Stochastic Implied Volatility. XI, 229 pages. 2004.

Vol. 546: D. Quadt, Lot-Sizing and Scheduling for Flexible Flow Lines. XVIII, 227 pages. 2004.

Vol. 547: M. Wildi, Signal Extraction. XI, 279 pages. 2005.

Vol. 548: D. Kuhn, Generalized Bounds for Convex Multistage Stochastic Programs. XI, 190 pages. 2005.

Vol. 549: G. N. Krieg, Kanban-Controlled Manufacturing Systems. IX, 236 pages. 2005.

Vol. 550: T. Lux, S. Reitz, E. Samanidou, Nonlinear Dynamics and Heterogeneous Interacting Agents. XIII, 327 pages. 2005.

Vol. 551: J. Leskow, M. Puchet Anyul, L. F. Punzo, New Tools of Economic Dynamics. XIX, 392 pages. 2005.

Vol. 552: C. Suerie, Time Continuity in Discrete Time Models. XVIII, 229 pages. 2005.

Vol. 553: B. Mönch, Strategic Trading in Illiquid Markets. XIII, 116 pages. 2005.

Vol. 554: R. Foellmi, Consumption Structure and Macroeconomics. IX, 152 pages. 2005.

Vol. 555: J. Wenzelburger, Learning in Economic Systems with Expectations Feedback (planned) 2005.

Vol. 556: R. Branzei, D. Dimitrov, S. Tijs, Models in Cooperative Game Theory. VIII, 135 pages. 2005.

Vol. 557: S. Barbaro, Equity and Efficiency Considerations of Public Higer Education. XII, 128 pages. 2005.

Vol. 558: M. Faliva, M. G. Zoia, Topics in Dynamic Model Analysis. X, 144 pages. 2005.

Vol. 559: M. Schulmerich, Real Options Valuation. XVI, 357 pages. 2005.

Vol. 560: A. von Schemde, Index and Stability in Bimatrix Games. X, 151 pages. 2005.

Vol. 561: H. Bobzin, Principles of Network Economics. XX, 390 pages. 2006.

Vol. 562: T. Langenberg, Standardization and Expectations. IX, 122 pages. 2006. (planned)

Vol. 563: A. Seeger, Recent Advances in Optimization. XV, 455 pages. 2006. (planned)

Vol. 564: P. Mathieu, B. Beaufils, O. Brandouy, (Eds.) Artificial Economics. XIII, 237 pages. 2005.

Vol. 565: W. Lemke, Term Structure Modeling and Estimation in a State Space Framework. IX, 224 pages. 2006.

Vol. 566: M. Genser, A Structural Framework for the Pricing of Corporate Securities. XIX, 176 pages. 2006.