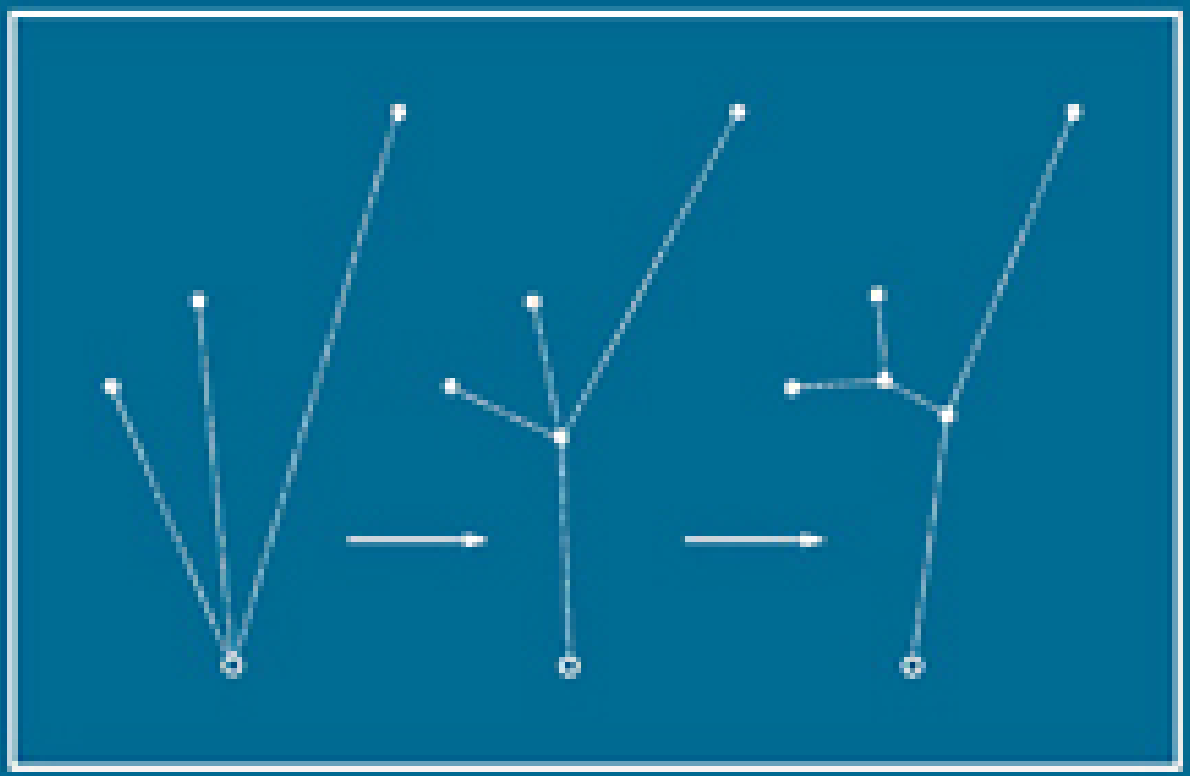
Cambridge Series in Statistical and Probabilistic Mathematics



Networks

Optimisation and Evolution

Peter Whittle

CAMBRIDGE www.cambridge.org/9780521871006

This page intentionally left blank

Networks: Optimisation and Evolution

Point-to-point vs hub-and-spoke. Questions of network design are real and involve many billions of dollars. Yet little is known about the optimisation of design – nearly all work concerns the optimisation of flow for a given design. This foundational book tackles the optimisation of network structure itself, deriving comprehensible and realistic design principles.

With fixed material cost rates, a natural class of models implies the optimality of direct source–destination connections. However, considerations of variable load and environmental intrusion then enforce trunking in the optimal design, producing an arterial or hierarchical net. Its determination requires a continuum formulation, which can however be simplified once a discrete structure begins to emerge. Connections are made with the masterly work of Bendsøe and Sigmund on optimal mechanical structures and also with neural, processing and communication networks, including those of the Internet and the Worldwide Web. Technical appendices are provided on random graphs and polymer models and on the Klimov index.

PETER WHITTLE is a Professor Emeritus at the University of Cambridge. He is a Fellow of the Royal Society and the winner of several international prizes. This is his 11th book.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

Editorial Board

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, second edition, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand, R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett

Networks: Optimisation and Evolution

Peter Whittle
Statistical Laboratory, University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521871006

First published in print format 2007

ISBN-13 978-0-511-27559-3 eBook (NetLibrary)
ISBN-10 0-511-27559-5 eBook (NetLibrary)

ISBN-13 978-0-521-87100-6 hardback
ISBN-10 0-521-87100-X hardback

To Käthe

Contents

Acknowledgements *page* viii
Conventions on notation ix

Tour d'Horizon 1

Part I: Distributional networks **7**
1 Simple flows 9
2 Continuum formulations 24
3 Multi-commodity and destination-specific flows 42
4 Variable loading 47
5 Concave costs and hierarchical structure 66
6 Road networks 85
7 Structural optimisation: Michell structures 95
8 Computational experience of evolutionary algorithms 116
9 Structure design for variable load 126

Part II: Artificial neural networks **135**
10 Models and learning 137
11 Some particular nets 146
12 Oscillatory operation 158

Part III: Processing networks **167**
13 Queueing networks 169
14 Time-sharing processor networks 179

Part IV: Communication networks **191**
15 Loss networks: optimisation and robustness 193
16 Loss networks: stochastics and self-regulation 199
17 Operation of the Internet 211
18 Evolving networks and the Worldwide Web 219

Appendix: 1 Spatial integrals for the telephone problem 227
Appendix: 2 Bandit and tax processes 234
Appendix: 3 Random graphs and polymer models 240
References 261
Index 268

Acknowledgements

I am grateful to Frank Kelly for generous orientation in some of the more recent communication literature. My references to his own work are limited and shaped by the theme of this text, a text totally different in aspiration and coverage from that which I have long encouraged him to write, and which we await.

I am also grateful to Michael Bell for piloting me through the post-Lighthill literature on traffic flow models.

Lastly, I might well not have been able to finish this work had I not, after retirement, kindly been granted continued enjoyment of the facilities and activities of the Statistical Laboratory, University of Cambridge. The advantage is all the greater, in that the Laboratory is now housed with the rest of the Faculty in the resplendent new Centre for Mathematical Sciences.

References to the literature can be regarded as a continuing stream of formal acknowledgement, as well as of association. If I make no reference on a given piece of work, then this is an indication that I regard it as either standard or new, with the greater likelihood of misapprehension in the second case.

Conventions on notation

The fact that we cover a wide range of topics, each with its own established notation, makes it difficult to hold to uniform conventions, and we do not do so entirely. We do consistently use x to denote the state variable of a system, but this will be the set of flows in Part I and the set of node occupation numbers in Part III, for example. We are forced then to use ξ to denote Cartesian co-ordinates. In general we follow the mathematical programming literature in using y to denote the variable dual to x, but in Part II bow to the conventions of control theory and use λ to denote this dual variable, releasing y to denote the observations (i.e. the information input).

The treatment is in general mathematical, although scarcely rising above the sophistication of 'mathematical methods'. The use of the theorem/proof presentation is then simply the tidiest and most explicit way of summarising current conclusions, implying neither profundity nor the pretence of it. The three appendices collect the material that is densest technically.

Equations, theorems and figures are numbered consecutively through a chapter, and also carry a chapter label. Equation (5.4) is thus the fourth equation of the fifth chapter.

Tour d'Horizon

Whither? Why?

The contents list gives a fair impression of the coverage attempted. Networks, both deterministic and stochastic, have emerged as objects of intense interest over recent decades. They occur in communication, traffic, computer, manufacturing and operational research contexts, and as models in almost any of the natural, economic and social sciences. Even engineering frame structures can be seen as networks that communicate stress from load to foundation.

We are concerned in this book with the characterisation of networks that are optimal for their purpose. This is a very natural ambition; so many structures in nature have been optimised by long adaptation, a suggestive mechanism in itself. It is an ambition with an inbuilt hurdle, however: one cannot consider optimisation of design without first considering optimisation of function, of the rules by which the network is to be operated. On the other hand, optimisation of function should find a more natural setting if it is coupled with the optimisation of design.

The mention of communication and computer networks raises examples of areas where theory barely keeps breathless pace with advancing technology. That is a degree of topicality we do not attempt, beyond setting up some basic links in the final chapters. It is well recognised that networks of commodity flow, electrical flow, traffic flow and even mechanical frame structures and biological bone structures have unifying features, and Part I is devoted to this important subclass of cases.

In Chapter 1 we consider a rather general model of commodity distribution for which flow is determined by an extremal principle. This may be a natural physical principle (e.g. the minimal dissipation principle of electrical flow) or an imposed economic principle (e.g. cost minimisation in the classic transport problem). The principle minimises a convex cost function, subject to balance constraints; classic Lagrangian methods are then applicable. These lead to the concept of a potential, and find their most pleasing development in the Michell theory of optimal structures, for which the optimal design is characterised beautifully in terms of the relevant potential field.

All this material is classic and well known, at least as far as flow determination is concerned – design optimisation may be another matter. However, we do find one class of cases for which the analysis proceeds particularly naturally. These are the models with *scale-seminvariant costs*, for which the cost of operating a link is the volume of the link (its length times some notion of a cross-sectional area) times a convex function of the flow density. The corresponding dual cost is volume times the dual function of

potential gradient. Let us refer to these simply as 'seminvariant-cost models' or, even more briefly, as 'SC models'. They constitute quite a general class of cases, which indeed includes most of the standard examples. However, they turn out to have a property that is disconcertingly dramatic.

The SC models for simple flow all reduce under a free design optimisation to the net that solves the classic transport problem: an array of direct straight-line links from sources to sinks. If the commodity is not destination-specific, then these routes will never cross in the plane, and there is a corresponding property in higher dimensions.

This is a gratifying reduction, and one that is acceptable in some cases. In Chapter 2 we solve a continuum problem by these means, and the corresponding solutions of the structural problems in Chapters 7–9 are less naive. However, in most cases this simple solution is totally unrealistic. A practical road net could never be of that pattern, and to formalise the reasons why is an instructive exercise. One would expect 'trunking': that traffic of different types should share a common route over some distance, as manifested by local traffic's feeding into major arteries and then leaving them as the destination is approached.

If the network is subject to a variable load, then this indeed has the effect of inducing some trunking in the optimal design; see Chapter 4. The reason is that there is then an incentive for links to pay their way by carrying as steady a load as possible. Links will then be installed which can carry traffic under a variety of loads, usually making route-compromises in order to do so. However, once this steadiness of flow has been achieved, there is no incentive for further trunking.

The real incentive to trunking must be that a link of whatever capacity carries a base cost. By its simple existence it has harmed some amenity, and incurred an environmental cost. More generally, one can say that environmental invasion implies that the installation cost of a link is a concave rather than a linear function of its capacity. For SC models it turns out then that the capacity-optimised sum of installation and operating costs is itself now a concave rather than a convex function of the flow rate. Otherwise expressed, the effect is that a single high-capacity link over a route section is to be preferred to a bundle of low-capacity links. The criterion then favours trunking, and indeed has a dramatic effect; see Chapter 5. The optimal network is inclined to a tree pattern, with traffic from neighbouring source nodes converging together to what is indeed a trunk; this trunk then branching successively as it approaches a cluster of sink nodes.

Trunking can be seen as inducing a hierarchical structure, in which local traffic is carried locally; traffic to distant but clustered destinations is gathered together on to a trunk route, and there can be such gatherings at several levels. Such a structure can be seen at its clearest in telephone networks, in which there can be exchanges at several levels, and a call will be taken up to just that level of exchange from which there is a route down to the receiver. This hierarchical structure, differing radically from the array of direct connections derived on a naive criterion, is the pattern for optimal nets under an environmentally conscious criterion. In Chapter 5 we consider the optimal choice of exchange levels in an isotropic environment. It turns out that ideally the trunking rate should be constant up to a certain level, although the continuous gradation of trunking implied by this assertion is not practical. The investigation throws up some questions in what is essentially geometric probability, treated in Appendix 1.

Road networks (Chapter 6) also aim to be hierarchical, but labour under the difficulty that roads actually do take a great deal more physical space than do telephone links, for example. They are particularly constrained by the fact that they are largely confined to the two-dimensional plane, from which they can depart only at great expense. They suffer also from the statistical effect of congestion, and it is a moot point whether or not this is trunk-inducing. Congestion in queueing networks is relatively well understood; congestion on the continuum of a multi-lane highway is a much subtler matter, discussed briefly in Chapter 6.

The study of engineering structures in Chapters 7–9 takes us back to the 'naive' case, when the cost of building to a certain capacity is proportional to that capacity. However, solution for the optimal structure is now considerably less naive, because of the vector nature of the potential. The theory of Michell structures is not only one of the most beautiful examples of optimisation but, published in 1904, one of the earliest. It generalises to the SC case, as we demonstrate. Evolutionary methods really do make a computational appearance here, as the only way of performing the optimisation in all but the simplest cases. We review the striking results of Bendsøe and Sigmund in this context, also of Xie and Stevens.

However, natural evolutionary optimisation (on the scale of years) is evident in a continuum version of the problem: the formation of bone. Bone is continually subject to a general wasting, but also to reinforcement in those locations and directions where there is stress. The effect is then to achieve a structure optimal for the conditions. Load-bearing elements, such as the femur (thighbone), demonstrate not only a shape but also an anisotropic structure of the form that the Michell theory would predict. Moreover, they demonstrate their response to variability of load in the replacement of interior solid bone by cancellous (spongy) bone; see Chapter 9.

Stochastic variation plays a greater role in succeeding chapters. Part II is devoted to the topic of artificial neural networks (ANNs), whose study constitutes the most concerted attempt yet to both mimic and understand the logical performance of natural neural networks, and to perhaps approach it by adaptive rules. By comparison with the distributional nets of Part I, adaptability to variable loading is of the essence; it is required of the net that it should make an appropriate response to each of a great variety of significant inputs. In contrast to the models of Part I, the nodes of the network now incorporate a nonlinear response element, a partial analogue of the animal neuron, seemingly necessary for the degree of versatility demanded. Lagrangian methods survive in the technique of 'back propagation' for the iterative improvement of input/output response.

The ANN literature is now an enormous and diffuse one, and we treat only a few specific cases that make definite structural points. In a reversal of the historical view, many standard statistical procedures can be seen as a consequence of ANN evolution. In a back-reversal, statistical analysis demonstrates that a simple neuronal assembly (a McCulloch–Pitts net) is insufficient for the functions demanded; the net has also to be able to rescale signals and to choose between competing alternatives. To achieve a dynamic version of the Hamming net, in fact. When such a net is elaborated to deal with compound signals one begins to see structures identifiable (and in no facile sense) with natural neural systems.

Actual biological systems have to operate over a very wide range of input intensities, and absolute levels of signals mean very little. Such systems achieve internal communication by oscillatory patterns, and W. Freeman has clarified the basic neuronal oscillator by which these are generated and sensed. When one couples the oscillator with the normalising mechanism for signal strength one finds an explanation for 'neuronal bursting'; the regular bursts of oscillation which are so pronounced an observational feature.

The treatment of processing networks in Part III does not in fact have a large network content, as manufacturing requirements generally specify the sequence of operations to be performed fairly closely. There can be allocation problems, however, when congestion occurs and different job streams are competing for attention. The classic Jackson network, which serves so well under appropriate conditions, does not address this point, and it is now realised that the introduction of local precedence rules into queueing networks can do more harm than good to performance. We then leap straight from the exact and amenable treatment of the Jackson network to the exact and fairly amenable treatment of the Klimov index. This certainly solves the problem in principle when processing resources can be freely redeployed, and gives indications of how to proceed when they cannot. Appendix 2 gives some background treatment of the Klimov index and related matters.

When we come to communication networks in Part IV we mostly take for granted that networks will be hierarchical, and that the aim is to extract as much performance from the net as possible under varying load, which can take the form of both short-term statistical variation and longer-term secular variation. The treatment of loss networks owes a great deal to F. P. Kelly, who passed through this topic on his long trek through evolving technology and methodology. In Chapter 15 we consider the fluid limit, for which the relationships of the various shadow prices are evident and the optimal admission/routing policy easily derivable. These features survive in a first stochasticisation of the problem. Chapter 16 considers also the second stochasticisation: that in which one uses state feedback to regulate the system. As Kelly has shown, the ideal instrument for real-time control of admissions and routing is that of trunk reservation. By this, reaction to small-scale features (the numbers of free circuits) achieves effective control of the large-scale features (the make-up of the traffic that occupies the busy circuits).

However, control rules of this type have their limits when there may be sources of congestion deep in the system that are subject neither to direct observation nor to direct control. This is the case for the Internet, which also by its nature operates in a very decentralised fashion. Some account is given in Chapter 17 of the protocols which have proved so successful in controlling this remarkable giant organism, and of the more relaxed optimality concepts which have guided thinking.

While the state of the net for either the Internet or the Worldwide Web may be inaccessible to observation at a given time, the nature of the network that develops can be discerned, and has awoken much interest. One particular feature remarked is that the distribution of node degree (the number of immediate links made with a given Web page) follows an inverse power law. This is often spoken of as the network's having a 'scale-free' character (not to be confused with the scale seminvariance of costs defined earlier). Such a law could just be explained on the classical theory of random graphs, although not particularly naturally. A more plausible mechanism has been suggested by

Barabási and Albert: that of 'preferential attachment'. However, any conjecture that the mechanism, whatever it is, might prove self-optimising is weakened by evidence that nets generated by a random mechanism show extremely poor performance by comparison with a properly designed hierarchical system. These matters are discussed in Chapter 18. The supplementary Appendix 3 describes some aspects of random graph theory of which people working in the Web context seem to be unaware.

I

Distributional networks

By ‘distributional networks’ we mean networks that carry the flow of some commodity or entity, using a routing rule that is intended to be effective and even optimal. The chapter titles give examples. Operational research abounds in such problems (see Ball *et al.*, 1995a,b and Bertsekas, 1998), but the versions we consider are both easier and harder than these. In operational research problems one is usually optimising flow or operations upon a given network, whereas we aim also to optimise the network itself. Even the Bertsekas text, although entitled *Network Optimization*, is concerned with operation rather than design. The design problem is more difficult, if for no other reason than that it cannot be settled until the operational rules are clarified. On the other hand, there may be a simplification, in that the optimal network is not arbitrary, but has special properties in its class.

The ‘flow’ may not be a material one – see the Michell structures of Chapters 7–9, for which the entity that is communicated through the network is stress. The communication networks of Part IV are also distributional networks, but ones that have their own particular dynamic and stochastic structure.

1

Simple flows

1.1 The setting

By ‘simple flow’ we mean those cases in which a single divisible commodity is to be transferred from the *source nodes* of a network to the *sink nodes*, and the routing of this transfer is determined by some extremal principle. For example, one might be sending oil from production fields to refineries in different countries, and would wish to achieve this transfer at minimal cost. (For the ideas of this chapter to be applicable one would have to assume that all oil is the same – if different grades of crude oil are to be distinguished then the more general models of Chapter 3 would be needed.) This is the classical single-commodity *transportation problem* of operations research (see e.g. Luenberger, 1989). However, we mean to take it further: to optimise the network as well as the routing. Further, in Chapters 4 and 5 we consider the radical effect when the design must be optimised to cope with several alternative loading patterns (i.e. patterns of supply and demand) or with environmental pressures.

Another example would concern the flow of electrical current through a purely resistive network. This has virtually nothing to do with the practical distribution of electrical power, which is achieved by sophisticated alternating-current networks, but the model is a very natural one, having practical implications. For given input/output specifications the flow through the network is determined physically: by balance relations and by Ohm’s law. However, Ohm’s law can be seen as a consequence of a ‘minimal dissipation’ criterion, so that one again has an extremal principle, this time a natural physical principle rather than an imposed economic one.

This example is the simplest of a whole class of models, for which the flow is characterised by an extremal principle, and which lead to the classic and fruitful concept of complementary variational principles. Within this class we shall find a realistic subclass, that of seminvariantly scaled costs, for which design optimisation turns out to be particularly simple.

Yet another example we shall come to in Chapters 7–9 is that of optimal structural frameworks, for which the entity that is transferred is *stress*. This is the complete analogue of flow; it is transferred from the points at which load is applied through the framework to the *foundation*: the backstop that accepts all load.

The simple flow model is no longer adequate for the road networks of Chapter 6. This is because stochastic effects begin to make themselves felt, and also because there are several classes of traffic – classification by destination alone is already enough to

change the situation radically. Other issues also arise as we develop the theme. If we are to optimise networks freely then we are forced to consider a much more general class of models: those allowing any pattern of flow on the continuum of physical space; see Chapter 2. We are also forced to recognise environmental constraints, which completely change the character of the optimal solution; see Chapter 5.

1.2 Flow optimisation

Denote the nodes of the network by j, taking values in the set $\{1, 2, \ldots, N\}$. Let f_j be the prescribed constant rate at which the commodity is supplied to node j from the external world, so that f_j is positive for source nodes and negative for sink nodes. Balance then requires that $\sum_j f_j = 0$, although this is a point we return to. Let x_{jk} be the rate of flow of the commodity from node j to node k. This can be nonzero only if there is a direct link between the nodes. If one thinks of the network as a graph, then the link would be termed an arc – we shall use the two terms interchangeably. In this chapter we shall assume the link to be undirected, in that flow can be in either direction, and x_{jk} can be of either sign. We must then adopt the convention that $x_{kj} = -x_{jk}$. When we come to road or communication traffic, then flows in opposite directions must be distinguished, and given separate directed links.

The flow x must obey the balance relation

$$\sum_k x_{jk} = f_j \tag{1.1}$$

at all nodes of the network. This is not in general sufficient to determine the flow, and so room is left for optimisation. Let us take as criterion that the flow is required to minimise the expression

$$C(x) = \sum_{j,k} c_{jk}(x_{jk}) \tag{1.2}$$

subject to the balance conditions (1.1). Here $c_{jk}(x_{jk})$ is to be regarded as the 'cost' of carrying flow x_{jk} along the jk link. We shall assume that direction is immaterial, so that the value of $c_{jk}(x_{jk})$ is unchanged if we reverse the direction of flow or the order of j and k. The symbol $\sum_{j,k}$ denotes a summation over the range $1 \le j < k \le N$, so that each undirected link is counted just once. If we really wished to sum over all ordered combinations of j and k we would sum with respect to the two variables separately.

The cost function $c_{jk}(x_{jk})$ can represent some economic criterion on which one bases optimisation of the flow, or it can arise as expression of a physical extremal principle; we shall see examples of both. Let us denote a specimen such cost function simply by $c(x)$. Then the basic properties we shall demand of this function are that it be convex and nondecreasing for positive x with $c(0) = 0$. The second and third assumptions are natural, but so is the first. It corresponds to the idea that the marginal cost of an increase in flow increases with flow. If the assumption fails, then one has a novel and significant physical phenomenon.

Let us denote the class of such functions by $\mathcal{C}$. There are occasions when it is useful to suppose c differentiable, with a one-to-one relationship between its argument x and its gradient $c'(x)$. Let us therefore define the class $\mathcal{C}_s$ of *strictly convex* cost functions,

a subclass of $\mathcal{C}$ for which the differential of $c(x)$ exists and is strictly increasing for positive x, with

$$\lim_{x\downarrow 0}\frac{c(x)}{x}=0, \qquad \lim_{x\uparrow+\infty}\frac{c(x)}{x}=+\infty. \tag{1.3}$$

If we consider directed links then we need consider only nonnegative x. If we consider undirected links, as we do for the moment, then we simply add the assumption of evenness: $c(x)=c(-x)$.

We have then to minimise the convex function (1.2) of flow pattern x subject to the linear constraints (1.1). This is the *primal problem*, which can be solved by Lagrangian methods in the strong form of convex programming (see e.g. Bertsekas *et al.*, 2003). Define the Lagrangian form

$$L(x,y)=\sum_{j,k}c_{jk}(x_{jk})+\sum_{j}y_j(f_j-\sum_{k}x_{jk}), \tag{1.4}$$

where y_j is the Lagrangian multiplier associated with constraint (1.1). Denote the unconstrained minimum of L with respect to x by $D(y)$. The *dual problem* is that of finding the value of y that maximises $D(y)$. Denote this maximising value by $\overline{y}$. Then one key assertion is: that the value of x minimising the Lagrangian form $L(x,\overline{y})$ solves the primal problem.

The other key conclusion is the interpretation of $\overline{y}_j$ as a marginal price: the rate of change of minimal cost incurred with change in f_j. Let $M(f)$ be the minimal value of total transport cost $C(x)$ for prescribed f. Then one can loosely state that

$$\overline{y}_j=\frac{\partial M(f)}{\partial f_j}. \tag{1.5}$$

This is an identification which must be stated more exactly if it is to be true, and there are caveats when f is subject to constraints such as the balance constraint $\sum_j f_j=0$. We cover these points in Section 1.7. In the meanwhile, it is helpful to be aware that (1.5) holds in some sense.

One advantage of the Lagrangian approach is the reduction in dimensionality. The vector y is of dimension N, the number of nodes, whereas the dimensionality of x equals the number of links, which could be as high as $N(N-1)/2$. However, the question is of course whether these calculations can be performed at all. A useful concept is that of the *Fenchel transform*

$$c^*(y)=\max_x[xy-c(x)] \tag{1.6}$$

of a cost function $c(x)$, a stronger (if more narrowly applicable) version of the classic Legendre transform. The square bracket can be regarded as the net profit one would make if one incurred a cost $c(x)$ by accepting flow x, but received a subsidy at rate y for doing so. Expression (1.6) is then the maximal profit one could make by choosing the value of x. Relation (1.6) defines a transform, converting a function c of flow to a function c^* of subsidy rate. If c is convex then so is c^*, and one has in fact that $c^{**}=c$; relation (1.6) still holds if one switches the roles of c and c^* (see e.g. Bertsekas *et al.*, 2003).

Note then the evaluation

$$D(y) = \sum_j y_j f_j - \sum_{j,k} c^*_{jk}(y_j - y_k). \tag{1.7}$$

The relevant value $\bar{y}$ of the multiplier vector y is that maximising this expression.

A case that allows very explicit treatment is that of electric current flowing through a network of resistors. For this the assumption is that

$$c_{jk}(x_{jk}) = \frac{1}{2} R_{jk} x_{jk}^2, \tag{1.8}$$

where R_{jk} is the resistance of link jk. Expression (1.8) is the rate of energy dissipation in the link, and we are appealing to the principle that the actual flow minimises dissipation. In this special case the Lagrangian form (1.4) is minimal with respect to x at

$$x_{jk} = \frac{y_j - y_k}{R_{jk}}. \tag{1.9}$$

Relation (1.9) expresses Ohm's law if y_j is interpreted as the potential (voltage) at node j. In fact, Ohm's law and the minimal dissipation principle are equivalent, in that each implies the other. One has still to determine the potentials y_j by appeal to the constraints (1.1) or by maximisation of the dual form (1.7), which now becomes

$$D(y) = \sum_j y_j f_j - \sum_{j,k} \frac{(y_j - y_k)^2}{2R_{jk}}.$$

As always, the case of quadratic costs and linear constraints holds a special place, both as being relatively amenable and as showing unexpected connections (see e.g. Doyle and Snell, 1984).

1.3 Seminvariantly scaled costs

When it comes to optimisation of the network, we would like to know how the cost $c_{jk}(x_{jk})$ of flow x_{jk} on link jk depends upon the physical size of the link. Let us suppose that the link has 'length' d_{jk} and 'cross-section' or 'rating' a_{jk}. We put these terms in quotation marks because they are not yet well-defined; interpretation is best kept elastic for the moment. Then we shall suppose that the cost function has the form

$$c_{jk}(x_{jk}) = a_{jk} d_{jk} \phi(x_{jk}/a_{jk}), \tag{1.10}$$

where ϕ is a convex function. The physical argument for this is that (dropping the jk subscript for the moment) $p = x/a$ is a flow density, so that $\phi(p)$ can be regarded as the cost per unit conductor volume of maintaining a flow density p. The factor ad is then the total volume of the link. The variable p is certainly meaningful; in the electrical context it is a a flow density, and it has equally meaningful roles in other contexts (see Section 1.6).

Given relation (1.10), which we shall write simply as $c(x) = ad\phi(x/a)$, we find that

$$c^*(y) = ad\psi(y/d), \tag{1.11}$$

where $\psi = \phi^*$. In the electrical context one would regard y as a potential difference over the link. Then y/d is a potential gradient, and, in this particular case (see the cautions of Section 1.6), $\psi(q)$ is the energy dissipation per unit volume if a potential gradient q is maintained.

The reduced class of cost functions specified by (1.10) turns out to include most cases of interest and to imply a particularly simple structure for the optimal design (at least until other constraining factors are included). We shall term this the class of *seminvariantly scaled costs*, intermittently abbreviated to SC.

We shall also in general demand that the function ϕ should belong to the class $\mathcal{C}_s$ of strictly convex functions defined above, as will then ψ. Having said that, we shall immediately consider some cases just failing this demand.

An even more special family of cost functions, which we shall find convenient, are those for which the dependence on flow density is expressed by a power law

$$\phi(p) = \frac{\kappa}{\alpha}|p|^\alpha. \tag{1.12}$$

Here κ is a coefficient, included for generality, and strict convexity requires that $\alpha > 1$. We find then that

$$\psi(q) = \phi^*(q) = \frac{\kappa^{-\beta/\alpha}}{\beta}|q|^\beta. \tag{1.13}$$

Here β is the index conjugate to α in that

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1, \tag{1.14}$$

so that $\beta > 1$ also.

Expression (1.12) defines a family of cost functions which includes a number of cases of interest. The case $\alpha = 1$ makes the cost of operating a link simply proportional to the absolute value of the flow through it, an assumption implicit in the classical transportation problem of operations research. This case lies just outside $\mathcal{C}_s$, with implications which will be evident in a moment.

In the case $\alpha = 2$, expression (1.12) would represent the rate of energy dissipation per unit volume in a material of conductance $1/\kappa$ carrying an electric current of density p. Otherwise expressed, the resistance R of formula (1.8) is proportional to d/a, which is what we would expect.

A third case of interest is the limit case $\alpha \to +\infty$. Define the function

$$H(p) = \begin{cases} 0 & (|p| \le 1), \\ +\infty & (|p| > 1). \end{cases} \tag{1.15}$$

Then the cost function (1.12) converges to $H(p)$ as $\alpha \to +\infty$. This then represents a link that will carry a flow of rate up to 1 without cost, but incurs an infinite cost the

moment this bound is exceeded. That is, it represents a link of a definite 'rating' $a = 1$, which will fail once this rating is exceeded. This is again a common situation for, for example, road or communication links, representing congestion in its crudest form. It follows from (1.13) and (1.14) that $\psi(q)$ converges to $|q|$ as α becomes infinite, exactly the Fenchel transform of $H(p)$. So, the piecewise linear function $|q|$ and the function (1.15), clearly both outside $\mathcal{C}_s$, are nevertheless mutual Fenchel transforms.

1.4 A first consideration of optimal design

Let us distinguish between the *external* nodes of the network, for which $f_j \neq 0$ and the *internal* nodes, for which $f_j = 0$, so that there is no immediate contact with the external world. This is a nomenclature taken over from neural networks, but is convenient. Then the external nodes are prescribed (in what we can think of as physical space, although we have not yet set up this identification) and the other nodes could presumably be chosen freely. One can also choose the links, in that one can choose the rating a of the link between prescribed nodes. Let us assume a total cost function

$$C(x, a) = C(x) + \gamma \sum_{j,k} a_{jk} d_{jk} = \sum_{j,k} a_{jk} d_{jk} [\phi(x_{jk}/a_{jk}) + \gamma]. \tag{1.16}$$

That is, a sum of the flow–cost function (1.2), with the additional assumption (1.10) of seminvariant scaling, and of a term proportional to the volume of material in the links, which we regard as a measure of total installation cost. This may seem like the sum of a continuing cost (that of flow) and a capital cost (that of installation), but we can make them both continuing if we regard γ as a leasing cost. For the moment we shall regard the nodes as fixed and the link-ratings a as open to choice. The cost function $C(x, a)$ is then to be minimised with respect to both x and a, and conclusions are independent of the order in which these minimisations are performed.

Lemma 1.1 *Suppose that $\phi \in \mathcal{C}_s$, and is not necessarily symmetric. Then the minimal value of $a[\phi(x/a) + \gamma]$ with respect to a is achieved at $a = p^{-1}x$ and is equal to qx, where*

$$q = \psi^{-1}(\gamma), \quad p = \psi'(q), \tag{1.17}$$

in that q is the root of $\psi(q) = \gamma$ that has the sign of x. In the case of symmetric ϕ we thus have a minimal value of $q|x|$ achieved at $a = p^{-1}|x|$, where q is the positive evaluation in (1.17).

Proof The evaluations asserted plainly hold at $x = 0$. Assume $x > 0$. If we take the transformed variable $p = x/a$ then the expression to be minimised becomes

$$\frac{\phi(p) + \gamma}{p} x.$$

The minimising value of p is evidently independent of x, and in fact satisfies

$$p\phi'(p) - \phi(p) = \gamma. \tag{1.18}$$

Let p and q be conjugate values under the transformation $\phi = \psi^*$, so that $p = \psi'(q)$ and $q = \phi'(p)$. We have then from (1.18)

$$\gamma = p\phi'(p) - \phi(p) = pq - \phi(p) = \psi(q), \tag{1.19}$$

so that both relations of (1.17) hold, with p necessarily being positive and so q being the positive root of (1.19). Relation (1.17) also implies that

$$\frac{\phi(p) + \gamma}{p} = \phi'(p) = q.$$

All the assertions of the lemma then follow for nonnegative x. The proof for negative x follows the same line, but with negative evaluations required for both p and q. $\Diamond$

The lemma leads to an immediate reduction of the optimisation problem.

Theorem 1.2 *Suppose the problem seminvariantly scaled with symmetric $\phi \in \mathcal{C}_s$. Let p and q have the positive values determined in (1.17). Then:*

(i) The flow densities $|x_{jk}|/a_{jk}$ have the constant value p in the optimal design. Equivalently,

$$a_{jk} = p^{-1}|x_{jk}|. \tag{1.20}$$

(ii) The flow optimisation is independent of ϕ once the ratings have been optimised, in that the cost expression (1.16) reduces to

$$\overline{C}(x) = \min_a C(x, a) = q \sum_{j,k} d_{jk}|x_{jk}|, \tag{1.21}$$

to be minimised with respect to x subject to the balance conditions (1.1).

This is very close to saying that the flow problem for any seminvariantly scaled flow cost reduces, when ratings are optimised, to the classical transportation problem. That is, the problem of economically transporting a commodity from source nodes to sink nodes when transportation costs are proportional to flow rates. However, the classical transportation treatment makes at least one assumption: that the only possible links are direct links from source nodes to sink nodes, so that there is no transportation within these two groups and there are no internal nodes. We shall prove that these assumptions follow as properties of the optimal design if the system is set in a uniform physical space and freely optimised. (By 'uniform' we mean that it is a Euclidean space, invariant under rigid translations and rotations.)

The essence of physical (Euclidean) space is that the shortest distance d_{jk} between two points is the straight-line distance, and that this distance measure obeys the triangular inequality

$$d_{jk} \le d_{jl} + d_{lk}$$

for any triple i, j, k of points.

The fact that we have supposed the links undirected and that we have a single unlabelled commodity forces us to make the following observation.

Lemma 1.3 *It may be assumed that all flows along a given link are in the same direction.*

Proof The conclusion follows from a substitution of flows. Suppose that we identify a flow of magnitude v_1 from source 1, which traverses link jk in that direction and a flow of magnitude v_2 from source 2, which traverses the link in the opposite direction: that is, in the direction kj. Suppose $v_1 \geq v_2$. Then change the flows through the link from sources 1 and 2 to $v_1 - v_2$ and 0 respectively. The flow of v_2 from source 1, which now has surplus is used to replace the other flows out of node j, which were previously derived from source 2, and there is a similar substitution from source 1 by source 2 at node k. The counterflow in link jk has thus been eliminated, without introducing a counterflow in any other link and without changing net flows in any link. Continued application of the procedure will thus eliminate all counterflows without affecting net flows. ◊

Theorem 1.4 *Suppose, in addition to the seminvariant assumptions of Theorem 1.2, that the d_{jk} of the problem specified in (1.16) are Euclidean distances, and that there are m source nodes and n sink nodes. Then:*

(i) The optimal network need consist of no more than $m+n-1$ straight-line links from source nodes to sink nodes.
(ii) The region that is served wholly by a given source node in the optimal design is star-shaped; likewise for the region that wholly supplies a given sink node.

Assertion (i) essentially amounts to the statement that the problem reduces to the classical transportation problem, and assertion (ii) generalises the statement that links of the optimal design will not cross in two dimensions. A star-shaped region with centre A is such that, if a point B belongs to it, then so do all points on the line segment AB.

Proof (i) Label source nodes by h and sink nodes by i. Then the commodity distribution can be seen as a flow $\overline{x}_{hi}$ from h to i by some set of routes for every pair hi, all flows being in the same direction on every link and the values $\overline{x}_{hi}$ being consistent with the specified f_j. A specified flow by some route from h to i will then clearly incur least cost if it is shifted to the straight-line route between source and destination. This implies then for the semi-optimised criterion function (1.21) that

$$\sum_{(j,k)} d_{jk}|x_{jk}| \geq \sum_h \sum_i d_{hi}\overline{x}_{hi},$$

with equality if the whole flow $\overline{x}_{hi}$ is carried on the straight-line route. We can then, without increase in cost, assume that links are directly from source to sink, and can identify $\overline{x}_{hi}$ with x_{hi}.

The problem is thus reduced to one of minimising $q\sum_h \sum_i d_{hi}x_{hi}$ subject to the conditions $x_{hi} \geq 0$ and

$$\begin{aligned} \textstyle\sum_i x_{hi} &= f_h \\ \textstyle\sum_h x_{hi} &= -f_i \end{aligned} \qquad (1.22)$$

for all source nodes h and sink nodes i. But this is the familiar linear programme of the transportation problem (see e.g. Luenberger, 1989). Having $m+n-1$ linearly independent constraints, it has a solution $\{x_{hi}\}$ with at most $m+n-1$ nonzero flows.

(ii) The assertion is best proved by considering the dual of the linear programme just derived. This would be: to choose potentials y_h and y_i to maximise $\sum_h y_h f_h + \sum_i y_i f_i$ subject to

$$y_h - y_i \leq q d_{hi} \tag{1.23}$$

for all source–destination pairs. The direct route hi will be used only if equality holds in (1.23). We bring this into correspondence with the usual transportation conventions if we define z-variables as the negative of the y-variables and g_i as the negative of f_i. The problem then becomes: choose prices z_h and z_i to maximise $\sum_i z_i g_i - \sum_h z_h f_h$ subject to

$$z_i - z_h \leq q d_{hi}, \tag{1.24}$$

route hi then being used in the optimal solution only if equality holds in (1.24). One interprets z_h as the unit buying price for the commodity at the supply point h and z_i as the unit selling price at the destination i, while g_i represents demand at i. The dual programme in this form then requires that one chooses prices to maximise the total profit on the operation, conditional on the fact that the unit profit on any route must not exceed the unit transport cost.

Let us use ξ to denote the Cartesian co-ordinate of a point in physical space, and ξ_h to denote the co-ordinate of source h, etc. Suppose that the z_h solving the linear programme above have been determined. Then the unit cost of supplying a sink node i at ξ_i from source h would be $z_h + q|\xi_i - \xi_h|$, and the minimal cost would be $z(\xi_i)$, where

$$z(\xi) = \min_h [z_h + q|\xi - \xi_h|]. \tag{1.25}$$

Denote the set of ξ for which the minimum in (1.25) is attained at h by $\mathcal{S}_h$. Then any destination that is supplied by source h in the optimal programme must lie in $\mathcal{S}_h$; destinations in the interior must be totally supplied by source h. One verifies readily that the sets $\mathcal{S}_h$ are star-shaped, whence the first part of assertion (ii) follows. The second part follows by interchange of the roles of source and sink. $\diamondsuit$

In Figure 1.1 we sketch the regions $\mathcal{S}_h$ for the case when the y_h are equal. Any consequent conjecture that the regions might be convex is revealed as false if we consider the case of two source nodes with y_1 sufficiently much greater than y_2 (i.e. source 1 can supply that much more cheaply), when the pattern is as sketched in Figure 1.2. The boundary separating the two supply sets is the locus of points P such that the distance of P from source 1 exceeds that from source 2 by a fixed amount. This is one branch of a hyperbola, with the two source points as its foci. The region supplied by source 1 is then the whole plane except for the hyperbolic 'shadow' cast by source 2. Both sets are indeed star-shaped, but $\mathcal{S}_1$ is not convex. The $\mathcal{S}_h$ will in general have piecewise hyperbolic boundaries.

Of course, the y_h are determined by the positions and resources/demands of both the source and the destination nodes. Formula (1.25) reflects an evaluation of the potential at a possible sink node of arbitrary co-ordinate ξ, and so extends the evaluation of the

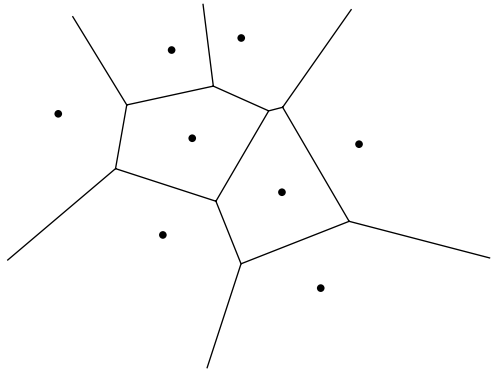


Fig. 1.1 The supply sets $\mathcal{S}_h$ in the case when all source nodes have the same potential.

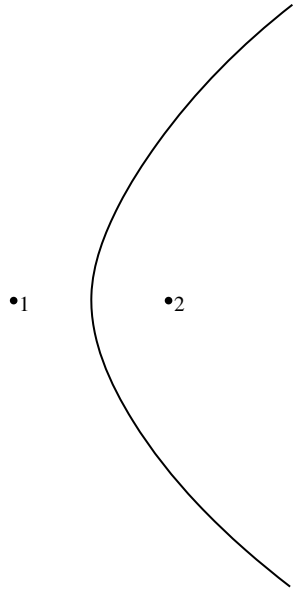


Fig. 1.2 The supply sets in the case when there are two source nodes, 1 and 2, such that $y_2 < y_1 < y_2 + qd_{12}$.

potential to all of physical space. We shall put this evaluation on a clearer footing in the next chapter.

Theorem 1.4 prompts an immediate unease. While it is interesting that the optimal design should collapse to such a simple form for such a large class of cases, one feels a justified incredulity – the solution is grossly unrealistic. If setting up a link were simply a matter of transporting the commodity by air or sea then a straight-line route might be acceptable, although there are all sorts of physical and political reasons why it might not be. However, if one is transporting over land, then the notion of an array of permanently laid, straight source-to-destination routes is unacceptable; traffic will have to be concentrated onto trunk routes. In Chapters 3–5 we study in detail various factors,

hitherto neglected, whose inclusion induces radical changes in the form of the optimal design.

1.5 The dual formulation

The belated appeal to the dual form of the linear programming problem in the proof of Theorem 1.4(ii) suggests that the whole treatment might more naturally have been carried out in the dual formulation. Indeed, if one wishes to consider all the possibilities that are latent in a free variation of the network then one is virtually forced to adopt this second approach. We were able to avoid doing so in Section 1.4 because continuum notions were effectively smuggled in through an inconspicuous back door; the acceptance of the possibility of a link between any pair of the given nodes plus appeal to the triangular inequality.

In the dual formulation the form

$$D(y, a) = \sum_j y_j f_j + \sum_{j,k} a_{jk} d_{jk} \left[\gamma - \psi \left(\frac{y_j - y_k}{d_{jk}} \right) \right] \tag{1.26}$$

is to be maximised with respect to y and then minimised with respect to the structure. That is, it is to be minimised with respect to the number and positions of the internal nodes and the ratings a of the consequent possible links, although we know that in the present case there are no internal nodes in the optimal design. One cannot in general commute these operations, but $D(y, a)$ satisfies the principal condition that would permit such commutation: it is concave in y for given a and convex (actually linear) in a for given y. Proceeding then with the a-minimisation one deduces:

Theorem 1.5 *(i) The dual form of the SC optimisation problem is: choose the potentials y_j to maximise the form*

$$D(y) = \sum_j f_j y_j \tag{1.27}$$

subject to the inequalities

$$|y_j - y_k| \le q d_{jk}, \tag{1.28}$$

for all node pairs.

(ii) A link can exist between nodes j and k in the optimal design only if equality holds in (1.28) for these values.

(iii) If the internode distances d_{jk} obey the triangular inequality then the only links that can exist in the optimal design are those carrying flow directly from source nodes h to destination nodes i. The problem then reduces further: to the maximisation with respect to y of

$$D(y) = \sum_h f_h y_h + \sum_i f_i y_i \tag{1.29}$$

subject to

$$y_h - y_i \le q d_{hi}, \tag{1.30}$$

for all source–sink pairs.

Proof A minimisation of the form (1.26) with respect to a_{jk} would yield a value of $-\infty$ unless the square bracket were nonnegative. That is, unless inequality (1.28) held, where q is the positive determination of $\psi^{-1}(\gamma)$, which we have already encountered. The maximisation with respect to y of the form (1.26) thus reduces to the maximisation of the reduced form (1.27) subject to (1.28), and a_{jk} can indeed be positive in the optimal design only if equality holds in (1.28). Assertions (i) and (ii) are thus proved.

If there is a flow by some path r from supply node h to destination node i then the equality form of (1.28) plus the fact that y is decreasing on the path imply that

$$y_h - y_i = qd(r) \geq qd_{hi}, \tag{1.31}$$

where $d(r)$ is the length of the path. But, by (1.28), strict inequality in (1.31) is forbidden. The conclusion is, then, that the flow must take place by the direct route. This implies the further reduction asserted in (iii). $\Diamond$

The reduced problem of (iii) is just the dual linear programme associated with the classical transportation problem, to which the optimisation now reduces.

To determine the optimal ratings in this approach we minimise the Lagrangian form (1.4) with respect to x_{jk}, yielding

$$\phi'(x_{jk}/a_{jk}) = \frac{y_j - y_k}{d_{jk}} = q(\text{sgn } x_{jk}). \tag{1.32}$$

Since $q = \phi'(p)$, where p and q are corresponding values, we deduce that the optimal rating of the jk link is $a_{jk} = |x_{jk}|/p$, where $p = \psi'(q)$. The ratio of flow rate to rating is thus in constant proportion, say p, and $\phi'(p) = q$. But this implies that p and q are corresponding values of the primal and dual arguments, so that $p = \psi'(q)$, as asserted in (1.17).

1.6 The primal and dual forms

The fact that the optimal-flow problem can be expressed in two forms, the primal one of minimising form (1.1) or the dual one of maximising form (1.7), permeates much of the physical, engineering and economic literature, in the form of dual extremal principles. The classic text on this subject is that by Arthurs (1970, republished 1980), but Arthurs himself admits that the work by Sewell (1987) replaces it. Sewell is especially interesting, because he considers 'cost' functions of a mixed convex/concave character, for which one must use the Legendre rather than the Fenchel transform; a structure associated with the possibility of phase transition (see Appendix 3). The additional task of optimising design is grafted on to these extremal principles, in manners to suit circumstances.

In economic contexts an optimisation of allocation subject to resource constraints always has prices as dual variables, the 'shadow prices' of the resources whose supply is critical. For this reason the concept of duality now lies deep in economic theory; the classic treatment of this topic is that by Intriligator (republished 2002). The dual variables have varying interpretations in other contexts. For the electrical resistance network they were potentials – voltages. One can then generalise the notion of a potential by definition: potentials are just the dual variables. However, there is often a clearer physical meaning.

For example, if one is considering a framework of elastic ties and struts, then the equivalent of 'flow' is the longitudinal force, tension or compression, which the member experiences. This force is related to the difference in displacement under load of the two ends of the member; and the displacement is then a vector 'potential' or dual variable. These two sets of variables become stress and strain in a continuum formulation. Classical and quantum mechanics are permeated by the concept of conjugate variables, again related to mutually complementary (i.e. dual) extremal principles.

It must not be supposed that complementary variational principles have the identical criterion function, differing only in a variable transformation. The difference in dimensionality of variables is already an indication that this is not so. It is interesting, however, that for the functions of the power-law family

$$\phi(p) = \frac{1}{\alpha}|p|^{\alpha}$$

the relation

$$\alpha\phi(p) = \beta\psi(q)$$

holds, where p and q are corresponding values. We do indeed then have $\phi(p) = \psi(q)$ in the case $\alpha = \beta = 2$, just the case of the electrical resistance net. The two complementary variational principles in this case are known as Dirichlet's and Thompson's principles. Elastic structures show the same combination of linear balance relations and quadratic 'costs', and so show the same physical identity of the optimisation criteria.

The Lagrangian approach to flow problems is plainly a venerable one. However, the author believes (until corrected) that the systematic introduction of seminvariantly scaled costs and the consequent reduction of the optimal design expressed by Theorem 1.4 are new. Bendsøe and Sigmund (2003) are however aware that optimisation of the electrical resistance net reduces to that of the transportation problem.

1.7 The multipliers as marginal prices

We can set the problem in a slightly more general version of a convex programme, one that implies seemingly more general versions, but does not engage with the complications of an infinite-dimensional case. Suppose that the problem is posed as the minimisation of a convex function $C(x)$ with respect to x, subject to the linear constraints $Ax = f$. The prescribed vector f may take values in a convex set $\mathcal{F}$. One can set up a Lagrangian form analogous to (1.4) and deduce a dual form analogous to (1.7). Let $\overline{y} = y(f)$ be the row vector of Lagrange multipliers and $M(f)$ the minimal value of C for a given value of f. Then one can assert that $M(f)$ is convex in $\mathcal{F}$ and that

$$\overline{y}(f^{\circ} - f) \leq M(f^{\circ}) - M(f) \tag{1.33}$$

for any other value f° in $\mathcal{F}$. If we choose $f^{*} = f + \epsilon u$, assuming that this belongs to $\mathcal{F}$ for all small enough scalar ϵ, and a given vector u, then (1.33) implies that

$$\overline{y}u \leq M_{u}(f) \tag{1.34}$$

where $M_u(f)$ is the derivative of $M(f)$ in direction u. This expresses the fact that $\bar{y}$ is a subgradient to M at f. If M possesses a gradient ∇M (a row vector) at f then (1.34) implies that

$$\bar{y}u \leq [\nabla M(f)]u. \tag{1.35}$$

If, further, any direction of movement from f is feasible, then (1.35) implies that

$$\bar{y} = \nabla M(f), \tag{1.36}$$

which is just assertion (1.5). However, under the constraint $\sum f_j = 0$, the potentials y_j are determined only up to an additive term independent of j, and the strongest conclusion that can be drawn from (1.35) is that

$$\bar{y}_j - \bar{y}_k = \frac{\partial M(f)}{\partial f_j} - \frac{\partial M(f)}{\partial f_k}.$$

1.8 Balance of supply and demand

One can ask what the mechanism may be that achieves the balance of supply and demand represented by the relation $\sum_j f_j = 0$. If one is dealing with an economic problem then the full-dress approach would be to introduce demand functions for suppliers at the source nodes and for consumers at the sink nodes. That is, let f_+ and f_- be the vectors of positive and negative f_j values; respectively the vector of supplies and the negative vector of consumptions. Then one would choose f and x to minimise a total cost function

$$\mathbf{C}(f, x) = C_+(f_+) + C(x) + C_-(f_-) \tag{1.37}$$

subject to (1.1). The first term in expression (1.37) is the cost to the suppliers of providing amounts f_+ at the source points, the second is just the transport cost (1.1), and the third is the cost to the consumers (actually, their negative utility) of receiving amounts $-f_-$. The balance relation then follows from (1.1). The fact that C_- is the negative of a utility function, i.e. that consumers actually want the commodity, supplies an incentive for transport to take place.

The cost components C_+ and C_- express strength of demand, and relate the transport system to the real world. However, their specification often takes one into wider issues than one wishes to consider, and a simpler approach is to remedy any mismatch in supply and demand by resort to the guillotine. So, suppose that $\sum_j f_j > 0$, so that there is excess supply. Then a crude assumption (but one that is easily refined) is to suppose that the excess can be disposed of without cost. The balance relation (1.1) then becomes

$$\sum_j x_{jk} = f_j - s_j \leq f_j, \tag{1.38}$$

where s_j is the amount of the commodity that is jettisoned at node j. The Lagrangian form (1.4) then becomes

$$L(x, y) = \sum_{j,k} c_{jk}(x_{jk}) + \sum_j y_j(f_j - s_j - \sum_k x_{jk}), \tag{1.39}$$

and the minimisation with respect to s_j then implies that $y_j \leq 0$, with equality if $s_j > 0$. Since there is a positive transport cost, one should jettison at the earliest opportunity. This will be at an entry point, so the multiplier y_j will be zero at those source nodes at which it is optimal to dispose of some of the excess. There is an analogous statement in the case of deficient supply, when we have $y_j \geq 0$, with equality at those sink nodes that it is optimal to choose to undersupply.

For the classic electrical network the physics of the situation forces balance, both at individual nodes and overall. It will often be the potentials rather than the flow rates f that are prescribed. If a capacitor is placed at a node then an imbalance of current flow there is possible, leading to a build up of charge, which cannot, however, continue indefinitely. The analogue in the case of commodity distribution would be the provision of a store of some kind at the node. The variable demand patterns considered in Chapter 4 will lead to a requirement for internal nodes in the optimal design. These constitute the natural sites for flow-smoothing stores.

2

Continuum formulations

Networks are usually seen as having a discrete set of nodes but, if one is to optimise the number and positions of such nodes, then one is virtually forced into the limit case: of envisaging flows upon a continuum. It may be that design optimisation leads one back to a discrete structure, as was the case in Theorem 1.4, but one can have freedom of manouevre only if one is willing to at least pass through a phase of continuum formulation. Moreover, many problems are by their nature set in the continuum.

We shall assume then that the potential 'nodes' are the points of a 'design space' $\mathcal{D}$, which is a bounded subset of physical space. We shall suppose this physical space to be $\mathbf{R}^\ell$, an ℓ-dimensional Euclidean space, and $\mathcal{D}$ is then the part of that space to which the structure is to be confined. Points of $\mathcal{D}$ will be denoted by ξ, the ℓ-dimensional column vector of the Cartesian co-ordinates of the point. The more usual 'x' or 'r' would be happier choices, but these symbols are pre-empted.

2.1 The primal problem

We shall consider a total cost function, which combines both flow and structure costs. In the discrete seminvariantly scaled case this would be

$$C(x, a) = \sum_{j,k} a_{jk} d_{jk} [\phi(x_{jk}/a_{jk}) + \gamma], \tag{2.1}$$

to be minimised with respect to x and the design variables, subject to the flow–balance condition (1.1). The continuum analogue of (2.1) would be

$$C(x, \rho) = w \int \int \rho(\xi, u) \{\phi[x(\xi, u)/\rho(\xi, u)] + \gamma\} \, \mathrm{d}\xi \, \mathrm{d}u. \tag{2.2}$$

Here $x(\xi, u)$ is the component of flow at ξ in direction u, and $\rho(\xi, u)$ is the density of conducting material dedicated to carrying that flow. These directed components of ρ can be physically distinguished, because there is an implicit assumption that physically separated conductors can be supplied to carry them. The point is that the double summation in (2.1) should carry over to a double integral in (2.2), representing direct connections between separated points ξ and ξ', say. However, these jump connections can be reduced to a sequence of local connections, and so to allowance of multiple directions u of flow at a given point ξ. The integral with respect to ξ is over the design space $\mathcal{D}$ and that with

respect to u is a uniform one over the spherical shell $|u| = 1$. The constant w normalises the u-distribution.

We shall see that matters simplify greatly in the optimal structure. Flow is then always carried in a single direction, except at junction points, so that in the end there is no need for physically separated (i.e. mutually insulated) conductors.

The flow $x(\xi, u)$ is thus the analogue of the link flow x_{jk}, and $\rho(\xi, u)$ is the analogue of the link rating a_{jk}. The link-length d_{jk} has now been absorbed into the differential element $\mathrm{d}\xi$. We shall assume for the moment that ϕ is even, so that the direction of flow does not affect cost. In later contexts it does: a given road link will carry traffic only in one prescribed direction, and structural materials behave differently in tension than in compression (which is the analogue of a flow or its reverse). The ratio x/ρ, which occurs as the argument of ϕ in (2.2), is what one might term the 'relative flow density': the ratio of the spatial density of flow to the spatial density of conducting material.

Expression (2.2) is to be minimised with respect to $x(\xi, u)$ and $\rho(\xi, u)$, subject to the balance condition

$$\operatorname{div} x = f \tag{2.3}$$

in $\mathcal{D}$. Here x is the total flow vector

$$x(\xi) = \int x(\xi, u) u \, \mathrm{d}u. \tag{2.4}$$

Note that the $\mathrm{d}u$ of expression (2.4) is not itself a vector; it is the area of an infinitesimal element of the spherical shell $|u| = 1$.

The term $f(\xi)$ in (2.3) represents the rate of flow into the system at ξ from the external world. This is a rate per unit time and also per unit volume, and may have either sign. The divergence term in (2.3) has the explicit form

$$\operatorname{div} x = \sum_k \frac{\partial x^k}{\partial \xi^k},$$

where the superscript k denotes the kth element of the appropriate vector. It represents the imbalance in flow x at ξ, which is to be compensated by the inflow $f(\xi)$.

The formulation expressed in (2.2) and (2.3) presumes a continuity which may well not hold. For example, if external flow is fed only via a discrete set of points ξ_j at rates f_j then we would have, formally,

$$f(\xi) = \sum_j f_j \delta(\xi - \xi_j),$$

where $\delta(\xi)$ is the Dirac δ-function: zero for nonzero ξ, but of unit integral. Further, we know from Theorem 1.4 that, if the external flows are of this form, then the optimal system reduces to a finite net. That is, both flow density x and material density ρ will be zero almost everywhere and infinite elsewhere. A rigorous treatment of these matters, while ultimately necessary, is more obscuring than enlightening; we continue in the hope

that intuition will be an adequate guide and the well-established theory of generalised functions an adequate reassurance. We can certainly replace the form $\sum_j f_j y_j$ of the discrete case by an integral $\int f(\xi)y(\xi)\mu(\mathrm{d}\xi)$ with respect to a general measure μ, although the balance equation (2.3) must then be modified to

$$\operatorname{div} x = f\frac{\mu(\mathrm{d}\xi)}{\mathrm{d}\xi}.$$

Note, in any case, an immediate deduction from (2.2): that the optimal values of x and ρ are in the simple proportional relation

$$\rho(\xi, u) = p^{-1}x(\xi, u), \tag{2.5}$$

where the scalar p has the determination (1.17). However, we need to go into the dual formulation to deduce the essential simplification: that at all points ξ, save possibly at junction points, flow is in a unique direction $u(\xi)$. That is, it forms a vector field with $x(\xi) = |x(\xi)|u(\xi)$.

2.2 The dual formulation

The continuum analogue of the dual cost function (1.26) is

$$D(y, \rho) = \int fy\mu(\mathrm{d}\xi) + w\int\int \rho[\gamma - \psi(\nabla y.u)]\,\mathrm{d}\xi\,\mathrm{d}u, \tag{2.6}$$

where the row vector ∇y is the gradient of y. Here u is a unit column vector indicating a conceivable direction of flow, f and y are both functions of ξ alone and ρ is a function of ξ and u. We include the dot in the inner product $\nabla y.u$ to make the expression unambiguous. The argument now goes very much as it did for discussion of the discrete case in Section 1.5. The y field must maximise the reduced form

$$D(y) = \int fy\mu(\mathrm{d}\xi)$$

subject to the constraint

$$|\nabla y.u| \le q \tag{2.7}$$

for all relevant ξ and u. Here q is again the positive root of $\psi(q) = \gamma$. In the optimal design, material will be laid down in direction u at position ξ only for those values of ξ and $u(\xi)$ for which equality holds in (2.7).

Inequality (2.7) and its related condition imply that material can be laid down at ξ in the optimal design only if equality holds in

$$|\nabla y| \le q, \tag{2.8}$$

and that it will then be laid down in the direction of the gradient of y. The gradient is of course nonzero when equality holds in (2.8), and its direction is determined if the gradient exists (i.e. if $y(\xi)$ is locally linear). We can thus assert:

Theorem 2.1 *At all points ξ for which the gradient ∇y exists, material can be laid down only if equality holds in (2.8), and it will then be laid down in the direction of the gradient, which will also be the direction of the flow. At such points the density ρ and the vector flow x are then functions only of ξ.*

Relation (2.5) then simplifies to

$$\rho(\xi) = p^{-1}|x(\xi)|. \tag{2.9}$$

We also obtain the reduction

$$C(x) = \min_{\rho} C(x, \rho) = q \int |x(\xi)|\, \mathrm{d}\xi, \tag{2.10}$$

analogous to (1.21), which indeed implies that the problem reduces to a version of the classical transport problem.

This reduction enables us to say something about the form of the optimal potential field $y(\xi)$.

Theorem 2.2 *Suppose that the system has a finite number of source and sink nodes at co-ordinates ξ_h and ξ_i respectively. Then the potential field for the optimal design has the form*

$$y(\xi) = \max_{h}[y_h - q|\xi - \xi_h|], \tag{2.11}$$

where $y_h = y(\xi_h)$.

Proof This is just assertion (1.25), now expressed in terms of potentials rather than prices. However, we now give the complete argument, which could scarcely be simpler, but does demand the full continuum setting.

The potential must be such that a flow from some source node h to a hypothetical node at ξ is possible. This then requires that

$$y(\xi) = y_h - q|\xi - \xi_h|, \tag{2.12}$$

by (2.8) and the accompanying equality condition for possibility of a link. Suppose that the potential is in fact determined by (2.12), and that the right-hand expression in (2.11) is maximal at $h = H$. Then

$$y(\xi) \leq y_H - q|\xi - \xi_H|. \tag{2.13}$$

But, by condition (2.8), strict inequality cannot hold in this last relation. There must then be equality; i.e. h must be such as to maximise expression (2.12). $\Diamond$

2.3 Evolutionary algorithms

We know from Theorem 1.4 that, on present assumptions, the optimal design reduces to a set of straight source-to-sink links, whose detailed form is settled by solution of a linear programme. There may seem to be little point, then, in pursuing design optimality by a direct extremisation of either of the continuum forms (2.2) or (2.6) of the primal and dual objective functions. However, matters become both more difficult and more interesting once we bring in the complications and constraints of the real world. It is then as well to introduce immediately the evolutionary approaches to optimisation that will survive these variations and will increasingly become the main recourse.

The criterion (2.7) associated with the dual approach seems to characterise the optimal design in a very explicit and appealing manner, and this characterisation will indeed prove fundamental when we come to consider the structural problems of Chapter 7. Nevertheless, the evolutionary approach is couched in terms of the primal form. Suppose we consider this form in the discrete case (2.1) and take the particular choice

$$\phi(p) = \frac{|p|^\alpha}{\alpha}. \tag{2.14}$$

We then have the criterion function

$$C(\hat{x}, a) = \sum_{j,k} a_{jk} d_{jk} [\alpha^{-1} (|\hat{x}_{jk}|/a_{jk})^\alpha + \gamma]. \tag{2.15}$$

Here $\hat{x}$ is the value of x optimising the flow; i.e. the value that minimises $C(x, a)$ subject to the balance conditions and for the given link capacities a. We assume that this has been calculated, which may seem like a breezy assumption that an essential part of the problem has been solved. However, in some cases one has a physical mechanism that does the calculation itself, in other cases one must simply resort to a direct computation (both massive and refined, and often appealing to the dual in its course; see Chapter 8).

A direct minimisation of expression (2.15) is difficult, because $\hat{x}$ is itself an implicit function of a. However, one can take a steepest-descent path

$$\dot{a}_{jk} \propto -\frac{\partial C}{\partial a_{jk}} \propto d_{jk} \left[(|\hat{x}_{jk}|/a_{jk})^\alpha - \beta\gamma \right]. \tag{2.16}$$

Here the dot on a indicates a time derivative; one regards the descent through the a-contours of C as describing a path which is followed in time. Derivatives of C with respect to $\hat{x}$ do not appear, because C is stationary with respect to this variable, which one assumes recalculated at all points on the path.

Relation (2.16) is appealing in its form. The term $-\beta\gamma$ represents a wasting of the link, which is counteracted by the stimulating effect of relative flow density, as expressed by the term $(\hat{x}/a)^\alpha$. When equilibrium is reached in equation (2.16) the optimality relations (1.20) will hold. One could of course deduce a version of (2.16) for general ϕ, but when we come to consider variable loadings the choice (2.14) has special properties.

The formal continuum analogue of relation (2.16) would be

$$\dot{\rho} = \kappa[(|\hat{x}|/\rho)^\alpha - \beta\gamma], \tag{2.17}$$

where both optimised flow $\hat{x}$ and material density ρ are functions of position ξ and orientation u. Relation (2.17) may seem like simplicity itself. However, there is a great deal of structure concealed in it, through the dependence of $\hat{x}$ upon prescribed in- or outflows at external nodes. There is also the point that we know that, if the set of external nodes is discrete, then so is the optimal design. The values of both $\hat{x}$ and ρ must then ultimately be either zero or infinity. This is a prospect both simplifying and disquieting, but one that can be dealt with. Note that relation (2.17) is more powerful than (2.16), in that ultimately it determines the nature and positioning of internal nodes if these prove necessary (which they will).

At those points where the dual field $y(\xi)$ has a gradient, we know that the flow has a definite direction (if nonzero). We can then understand $\hat{x}$ and ρ in (2.17) as simply $\hat{x}(\xi)$ and $\rho(\xi)$: the vector optimal flow at ξ and the material density at ξ (which does not need to be seen as directed).

There are circumstances in which there are direct physical proxies for some of the expressions appearing. For example, suppose that one is concerned with the distribution of electrical current in a system of conductors. The 'excitation' term $(|\hat{x}_{jk}|/a_{jk})^{\alpha}$ in (2.16) might then be roughly proportional to the degree of overheating in that part of the net. This relation could then be seen simply as

$$\dot{a}_{jk} = \kappa(T_{jk} - T_w),$$

where T_{jk} is the temperature of the jk-conductor and T_w is the desired working temperature.

The simplicity of this last example reminds us of the commoner associations of the word 'evolutionary'. In the ecological context the only criterion is that of survival, and a species has to adapt in order to do so in a world initially neutral, later competitive. In the study of evolutionary automata one proposes a simple adaptation rule, and sees where it leads. In the present case, the adaptation rule is not so simple, but is derived from a postulated model and optimality criterion.

2.4 A trivial continuum example

The insight we carry over from the discrete model is that the optimal design will consist simply of straight-line links from source points to sink points. This is enough to gain a grip even on problems that are of a continuous character. On the other hand, as we shall see, there are considerations that will change the character of the optimal design completely.

The only new feature we can explore within the present framework is that in which there is a continuous spatial distribution of input to or output from the system. A plane example whose solution is immediately evident is that in which we have a ring of uniform sources and a concentric ring of uniform sinks, of radii r_1 and r_2, say. By 'uniform' we mean that supply and demand rates are constant around the rings, that they are consistent in that total supply equals total demand, and that there is only a single commodity, so that distance and overall balance are the only criteria on routing. Then in a discrete approximation it is clear that the links will be equi-spaced radii, as in Figure 2.1. The links will be

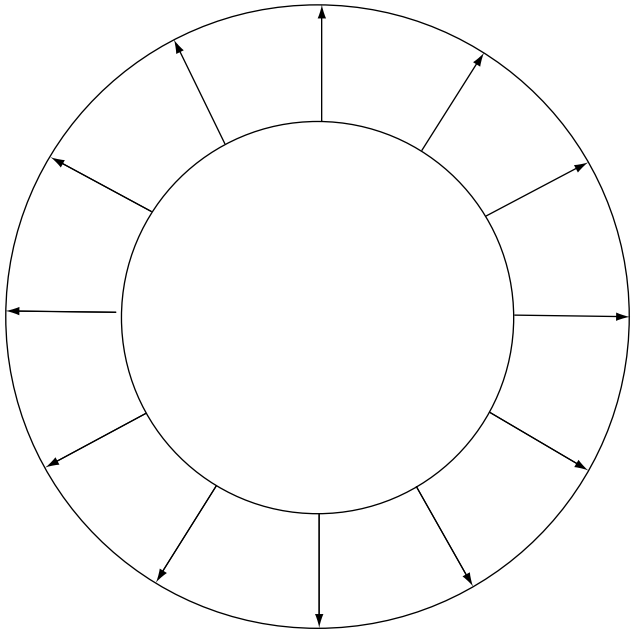


Fig. 2.1 The simple symmetric allocation of flow and material for the example of Section 2.4.

of the same capacity, so that the average density of material at an intermediate radius r will be proportional to r^{-1}. In the continuous limit we can drop the qualification 'average'.

The corresponding result in ℓ dimensions (when the rings become concentric spherical shells) would be that the density $\rho(r)$ of material at an intermediate radius of r should be proportional to $r^{1-\ell}$. This follows simply from the assertions that flow through a given intermediate shell is normal and uniform, with an integral independent of radius, and that spatial density of material is proportional to spatial density of flow.

2.5 Optimal cooling

A more substantial problem is that solved computationally on page 112 of Bendsøe and Sigmund (2003). This concerns a flat rectangular plate, which is heated uniformly over its surface; the problem is to introduce elements that will conduct the heat away to a heat sink placed on a central part of the lower edge. Heat conduction follows essentially the same rules as electricity conduction, with temperature taking the role of potential.

This continuum problem presents one interesting difference from the discrete problems we considered earlier: the conductors are 'leaky', in that they do not simply take heat from one specified location to another, but they also leak heat laterally on the way. This is because it must be supposed that the material of the plate within which the conducting elements are embedded is itself conducting, although poorly so. If this were not so, the embedded system of conductors would have to be infinitely dense if it were to cool the

continuum of the plate 'almost uniformly'. Indeed, there would not be a solution for some later versions of the problem.

The one-dimensional version of the problem is easily solved explicitly, and yields some intuition for the two-dimensional case. Suppose the plate is replaced by the interval $0 \leq \xi \leq L$, with $\xi = L$ being the point at which the heat sink is placed. If heat is supplied to the line segment at rate f per unit length then it must be extracted by the sink at rate fL. The flow at ξ will be the integral $f\xi$ of the input up to that point. If density is unconstrained then we can apply formula (2.9) to deduce that $\rho(\xi) = f\xi/p$, where p again has the determination (1.17). There is thus a simple linear growth of density to cope with the linearly increasing flow.

However, realism would demand that ρ be restricted to varying from ρ_{min}, which is the ρ-equivalent of the conductivity of the embedding material, and ρ_{max}, which is a physical upper limit on the density of the heat-distributing element. Let us for explicitness suppose that $\phi(p) = p^2/2$, which indeed represents the physics of heat conduction. The functional to be maximised with respect to temperature y and minimised with respect to density ρ is then

$$D(y, a) = \int_0^L [fy + \gamma\rho - \rho(y')^2/2]\mathrm{d}\xi - fLy(L), \tag{2.18}$$

where $y' = \mathrm{d}y/\mathrm{d}\xi$. The condition that y should maximise expression (2.18) yields the differential equation

$$\rho y'' + \rho' y' + f = 0, \tag{2.19}$$

with boundary conditions: y' equals 0 at $\xi = 0$ and $-fL/\rho$ at $\xi = L$. Under an optimal allocation of material the gradient y' will be constant for all ξ for which $\rho_{min} < \rho < \rho_{max}$. It will in fact have the negative value that makes the coefficient of ρ in expression (2.18) zero:

$$y' = -\sqrt{2\gamma}. \tag{2.20}$$

We see from (2.19) that ρ' is also constant on this set:

$$\rho' = -f/y' = f/\sqrt{2\gamma}, \tag{2.21}$$

so that ρ increases linearly with ξ at this rate.

We see from (2.19) and (2.20) that y varies linearly with ξ in the range of intermediate ρ, and quadratically in the regions of constant ρ. The values of y' at $\xi = 0$ and L are determined above, and are determined by continuity at the transition points. Collecting these conclusions, we find that the optimal configuration is as presented in Figures 2.2 and 2.3. Density follows the course $\rho(\xi) = f\xi/\sqrt{2\gamma}$ of the unrestricted case, except that it is set equal to the bounding values when it would overstep these. There is a region in which density reaches its upper limit only if $L > \rho_{max}\sqrt{2\gamma}/f$.

The geometry of the two-dimensional case is set out in Figure 2.4. One expects that this case will demonstrate the same behaviour: that ρ will increase from ρ_{min} to ρ_{max} as one moves from the bare edges of the plate to the heat sink. In fact, we can determine the optimal density distribution in implicit but simple form in the case when no bounds are imposed upon density. We know from the case of discrete external nodes that the optimal

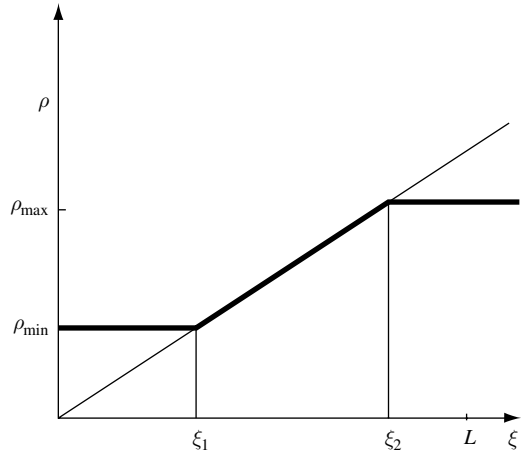


Fig. 2.2 The optimal density $\rho(\xi)$ of material for the one-dimensional cooling problem.

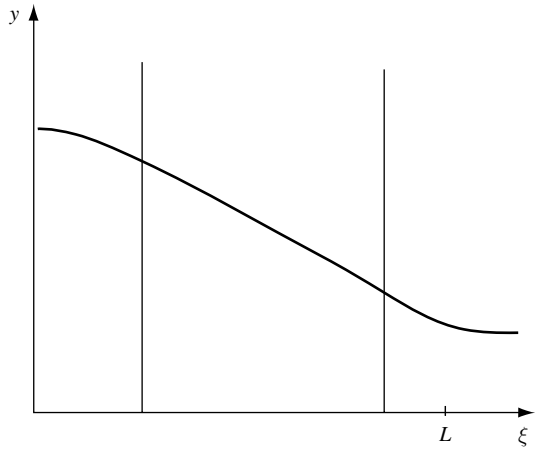


Fig. 2.3 The profile of potential (temperature) for the one-dimensional cooling problem.

configuration consists of direct linear links from source to sink nodes, without crossings. In the present case this will mean linear links from points in the plate interior to points on the interval that constitutes the heat sink, as in Figure 2.4. In the continuum limit the superposition of all these links will provide an evaluation of the optimal density.

Consider one particular such link as illustrated in Figure 2.5, extending from a given point P of the interior to a point in the sink interval distant s from its centre. We suppose that this link makes an angle θ with the normal to this interval. There is still some latitude in the design; heat is generated uniformly over the plate, but what will be the spatial pattern of its extraction over the sink? This is equivalent to the question: what should the relationship be between s and θ?

Such a relationship would be specified if the temperature (i.e. the potential) were expressed as a function $y(s)$ of s on the sink. If the point P has Cartesian co-ordinate

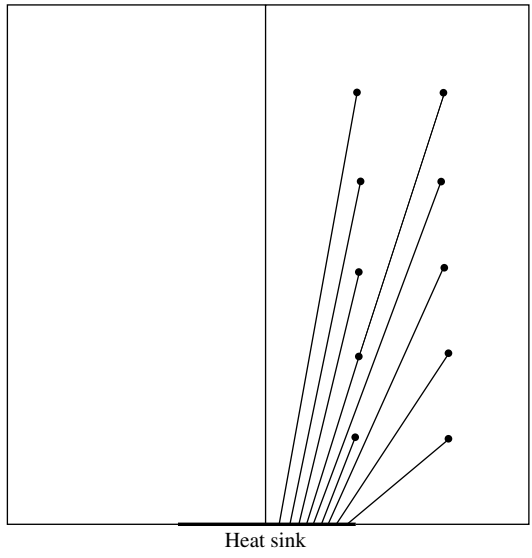


Fig. 2.4 The rectangular plate of the two-dimensional cooling problem with its boundary heat sink. The internal lines indicate the optimal path array for a discrete approximation to the problem.

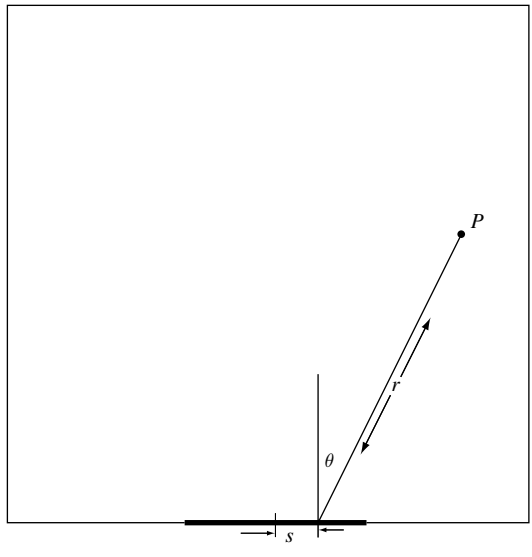


Fig. 2.5 The (s, θ, r) co-ordinates of the path from P to the heat sink.

ξ relative to an origin at the centre of the sink then we know from Theorem 2.2 that s should minimise

$$q\sqrt{(\xi_1 - s)^2 + \xi_2^2} + y(s),$$

where $q = \sqrt{2\gamma}$ in the present case. The condition that it should do so is

$$y'(s) = q \sin\theta. \tag{2.22}$$

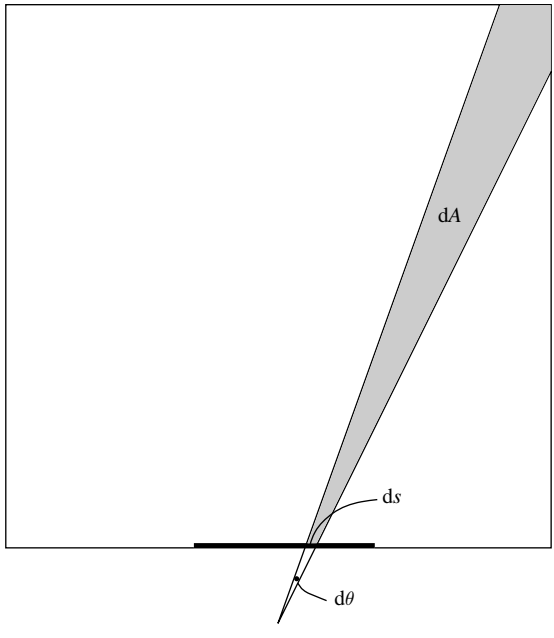


Fig. 2.6 Increments of s, θ and area A between a close pair of paths.

Relation (2.22) would determine the s/θ relationship if one indeed knew the temperature profile at the sink. Conversely, it determines the profile if the relationship is known. We shall see that the relationship can be determined, at least implicitly, if one adds some criterion on conditions at the sink.

Consider Figure 2.6, in which we have considered an increment $\mathrm{d}s$ and a corresponding increment $\mathrm{d}\theta$. Then the total flow that is to be extracted through the element $\mathrm{d}s$ is proportional to the area between the two chords of the plate that we have drawn; we can assume variables normalised so that it is equal. Let $R(s,\theta)$ be the length of the chord specified by s and θ, a known function of these variables. Then, to first order in infinitesimals, the area of the region between the two chords is

$$\mathrm{d}A = \frac{1}{2}R^2\,\mathrm{d}\theta + R\cos\theta\,\mathrm{d}s.$$

The two terms represent contributions to the area, first by a rotation of the chord by an angle $\mathrm{d}\theta$, and then by its horizontal displacement by an amount $\mathrm{d}s$. Suppose we impose the condition that heat flow should be uniform over the sink. This would then imply that $\kappa\mathrm{d}s = \mathrm{d}A$ for some constant κ, and so that

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = 2\frac{\kappa - R\cos\theta}{R^2}, \tag{2.23}$$

where the dependence of R upon s and θ is understood. Solution of this first-order differential equation would determine the s/θ relationship. The numerical integration of equation (2.23) has a very simple graphical expression. Figure 2.7 presents a plate of more general form, to emphasise that the method is not restricted to the case of a rectangular

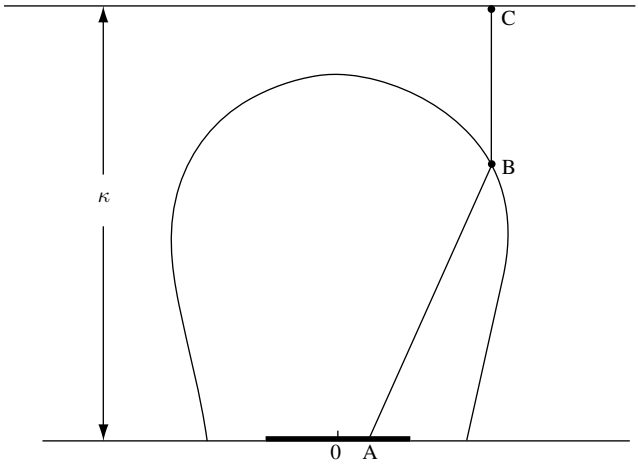


Fig. 2.7 The measurements appearing in the graphical version (2.24) of the differential equation (2.23).

plate. We have assumed, however, that the plate is symmetric about a vertical axis and that the sink remains linear. Then equation (2.23) can be given the incremental form

$$\Delta\theta = 2\frac{|BC|}{|AB|^2}\Delta s, \tag{2.24}$$

where $|BC|$ indicates the length of the line segment BC, etc. Since we have assumed symmetry, the chord at $s = 0$ will be vertical, and one can then take progressive increments of s to determine the corresponding increment in θ from (2.24).

The required value of κ can be determined by the integration of the equation $\kappa \mathrm{d}s = \mathrm{d}A$ to yield

$$\kappa = \frac{A}{S}, \tag{2.25}$$

where A is the area of the plate and S the length of the sink. For the success of the method we must have κ greater than the maximal height of the plate. This is certainly true for a rectangular plate, but not for the triangular plate of Figure 2.8. This narrows so quickly that the rate of extraction from the sink is necessarily greater at the centre. Note also, a somewhat excessive condition that the pattern of Figure 2.4 be achievable: that the straight line joining any point of the plate to any point of the sink should lie totally within the plate. This is a weaker demand than that of convexity (of the plate), but still stronger than necessary.

We are now in a position to calculate the density ρ implied by (2.23). Suppose that the position of the point P in Figure 2.5 is specified by s and r, where r is the length of the link from P to the sink point at s. Then the amount of flow at r between the two chords of Figure 2.6 is equal to the area beyond distance r. This is

$$\mathrm{d}A(r) = \frac{1}{2}R^2\,\mathrm{d}\theta + Rc\,\mathrm{d}s - \frac{1}{2}r^2\,\mathrm{d}\theta - rc\,\mathrm{d}s,$$

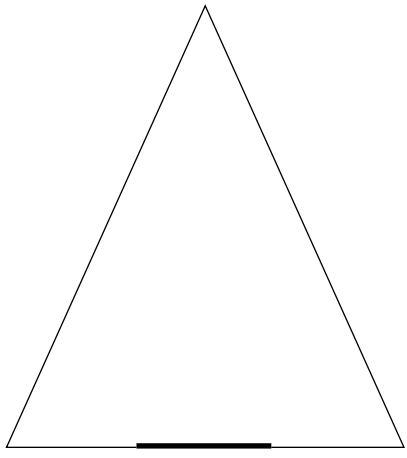


Fig. 2.8 A plate for which the analysis given is not applicable.

where we have abbreviated $\cos\theta$ to c. The width of cross-section to the flow at this point is

$$\mathrm{d}\sigma = r\,\mathrm{d}\theta + c\,\mathrm{d}s. \tag{2.26}$$

The density of material at this point is thus

$$\rho(r,s) = \frac{\mathrm{d}A(r)}{\mathrm{d}\sigma} = (R-r)\frac{(R+r)(\mathrm{d}\theta/\mathrm{d}s)/2 + c}{r(\mathrm{d}\theta/\mathrm{d}s) + c}. \tag{2.27}$$

Substituting expression (2.23) for $\mathrm{d}\theta/\mathrm{d}s$ into (2.26) we obtain a relatively explicit expression for the material density at the point (r,s) under the assumption of uniform flow at the sink:

$$\rho(r,s) = (R-r)\frac{cR^2 + (R+r)(\kappa - Rc)}{cR^2 + 2r(\kappa - Rc)}. \tag{2.28}$$

Here $R = R(s,\theta)$ and $c = \cos\theta$. Density thus equals zero for $r = R$ and $\kappa/\cos\theta(s)$ for $r = 0$. Application of relation (2.24) for an appropriate value of Δs will generate a sequence of chords, and on each chord one can use relation (2.28) to evaluate the conductor density a distance r from the sink.

We have envisaged the flow as being carried on discrete strands, these merging to a continuum as the nets of source and sink points are taken ever finer. For this to be legitimate it is necessary that there should be no incentive for 'shorting' between bare strands; i.e. neighbouring strands should be at the same temperature. This is indeed the case, because, by construction, the temperature gradient is in the direction of the strand.

An alternative criterion would be to require that the spatial density of flow should be uniform over the sink, with the consequence that the conductor density would also be so. The width of the cross-section of the element of flow towards the element $\mathrm{d}s$ is $\cos\theta\mathrm{d}s$, and so the condition now becomes $\kappa\cos\theta\mathrm{d}s = \mathrm{d}A$. Relation (2.23) is thus modified to

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = 2\frac{(\kappa - R)\cos\theta}{R^2} \tag{2.29}$$

and relation (2.28) becomes

$$\rho(r, s) = (R - r)\frac{R^2 + (R + r)(\kappa - R)}{R^2 + 2r(\kappa - R)}. \tag{2.30}$$

Expression (2.30) equals zero for $r = R$ and κ for $r = 0$, so κ can be identified with the constant value of material density at the sink. By integrating the equation $\kappa \cos\theta \mathrm{d}s = \mathrm{d}A$ we obtain the evaluation:

$$\kappa = \frac{A}{\int \cos\theta(s)\mathrm{d}s} \tag{2.31}$$

where the integral is over the line segment of the sink and $\theta(s)$ is the evaluation of θ in terms of s deduced from (2.29).

The differential equation (2.29) can also be given a graphical form; see Figure 2.9. Extend the chord AB out to length κ and then take a horizontal to cut the vertical axis at the point C. Then (2.29) can be given the incremental form

$$\Delta\theta = 2\frac{|CD|}{|AB|^2}\Delta s.$$

The Bendsøe–Sigmund computation yields, not this smoothly graded variation of density, but rather the pattern of Figure 2.10. In this the decreasing density of conducting material as one moves away from the heat sink is replaced by tapering fingers of full-density material, which themselves branch into finer and finer structures. The reason for this is a factor built into the computation. Engineers have traditionally disliked what they call 'grey' structures, meaning structures with elements of less than full density. They prefer 'black and white' structures, in which one has either full-density material or

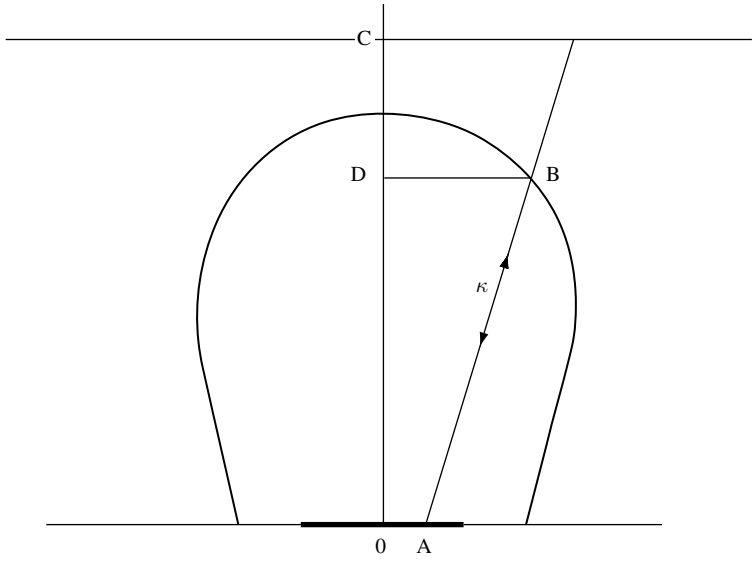


Fig. 2.9 The measurements appearing in the graphical version of the differential equation (2.29).

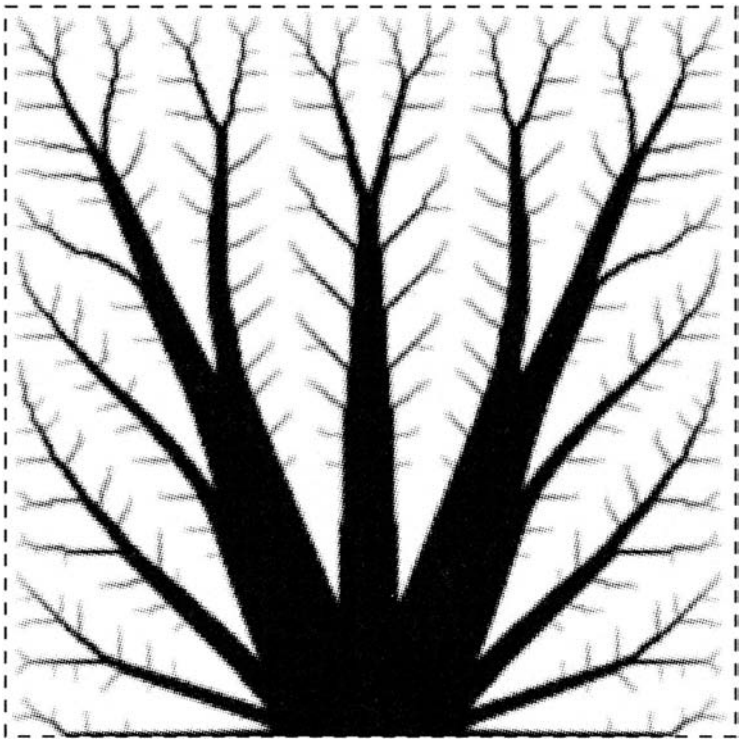


Fig. 2.10 The Bendsøe–Sigmund 'black-and-white' solution of the two-dimensional cooling problem. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

nothing. To achieve this, they modify the calculations by replacing the factor ρ where it multiplies ψ in expression (2.6) by a factor ρ^k, where k may take a value of about 2 or 3. This is a modification which is found empirically to yield a black and white design, although we shall see in Chapter 5 that it has a more fundamental basis.

Increasing sophistication of materials used leads of course to an acceptance of 'grey' designs, and indeed a growing interest. The composite and expanded materials which have been developed find increasing use. Nature reached this point aeons ago; the spongy or 'cancellous' bone that makes up much of the interior of the load-bearing elements of the animal skeleton represents just such an expanded material, with its own clear advantages; see Section 8.3.

Returning to the heat-extraction example, note the crucial role played by the presence of a background embedding material, with its small but positive conductance. As this conductance diminishes the 'black-and-white' design would have had to show an ever finer structure, becoming ever less realisable. In an evolutionary version of the calculations (see the next section), it is the feebly conducting background material which allows the signals to pass which guide growth of the conducting network and bring further growth to a halt when a sufficiently uniform heat extraction rate has been reached. One may see similar mechanisms in biological contexts; the growth of a neural network takes place

upon the scaffold of a relatively dense assembly of the more primitive glial cells, these being largely allowed to wither once the neural structure is complete.

2.6 More on evolutionary optimisation

Actual computations of the type that produced the configuration of Figure 2.10 are performed on a fine array of finite elements, approximating the continuum, and work towards the optimum by a steepest-descent method. The approach is thus an iterative one in which the design is given a small nudge in the apparent direction of optimality, the flow or potential fields are allowed to settle, and the step is repeated. It is not too much to term this approach 'evolutionary', although the guiding evolutionary pressure is supplied by a model rather than by a natural environment, and the variations of design permitted are restricted to those envisaged in the model. In Chapters 4 and 5 we shall extend the treatment of Section 2.3 to include the features of varying load and environmental considerations, and in Chapter 8 shall quote examples from some of the remarkable structural studies.

An alternative to the steepest-descent minimisation of the primal form considered in Section 2.3 is to work from the dual. The dual criterion function is

$$D(y,\rho)=\int\{fy+\rho[\gamma-\frac{1}{2}|\nabla y|^2]\}\mathrm{d}\xi, \tag{2.32}$$

if we suppose that $\phi(p)=p^2/2$, and so $\psi(q)=q^2/2$. This choice is convenient, because the condition that D be maximal with respect to y is then a linear equation in y:

$$\rho\nabla.\nabla y+\nabla\rho.\nabla y+f=0 \tag{2.33}$$

(cf. (2.19)). There will also be conditions at the boundary of $\mathcal{D}$ (of the type $\nabla y.u=f$, where u is the outward normal to the boundary).

As mentioned in the last section, Bendsøe and Sigmund modify criterion (2.32) to

$$D(y,\rho)=\int\{fy+\rho\gamma-\frac{\rho^k}{2}|\nabla y|^2]\}\mathrm{d}\xi,$$

where k is about 2 or 3. The minimising value of density would then be

$$\rho=\left(\frac{2\gamma}{k|\nabla y|^2}\right)^{1/(k-1)},$$

this in its turn affecting the stationarity condition (2.33). The modification was introduced heuristically, as being found to encourage passage to a black-and-white design; see Chapter 8. There is however a good theoretical motivation for it; see Chapter 5.

2.7 Nonuniform spaces

Suppose that one would wish the paths of the network to avoid certain areas. For example, if one is indeed transporting some material commodity through physical space then one might wish to avoid the physical obstacles of mountains, lakes and indeed towns. One

could represent this effect by making the structural cost γ a function $\gamma(\xi)$ of position. The effect of this will be that the minimal-cost paths are no longer straight, but are geodesics determined by the variable cost $\gamma(\xi)$.

This purely topographic variation is perhaps something of a side-issue at the moment, and the reader can well pass over this section without later effect. Of greater fundamental importance are those issues that may indeed be termed topological, in that they affect the optimal design by more than a spatial distortion.

Consider the cost

$$C(x, a, r) = \int_r a[\phi(x/a) + \gamma]\mathrm{d}s \tag{2.34}$$

of taking a fixed flow of x along an arbitrary path r. Here s measures the Euclidean distance along the curved path from the starting point. The cost γ and the rating a are both position-dependent, the first dependence being prescribed and the second now to be optimised. We find then, just as in Lemma 1.3, that when the cost (2.34) is minimised with respect to a it becomes

$$C(x, r) = x \int_r q\mathrm{d}s, \tag{2.35}$$

where $q(\xi)$ is the positive root of $\psi(q) = \gamma(\xi)$ and, in the integral, q is given its value on the path.

Now that the effective point cost $q(\xi)$ of transportation is determined, we can in principle determine the optimal point-to-point route. This is a routine matter, but we give the details.

Theorem 2.3 *Let $G(\xi, \eta)$ be the minimal cost of transporting unit amount of the commodity from ξ to η. Then this obeys the equation*

$$|\nabla_\xi G(\xi, \eta)| = q(\xi), \tag{2.36}$$

with terminal condition $G(\eta, \eta) = 0$. The optimal path is generated by moving from the current co-ordinate ξ in the direction of $-\nabla_\xi G(\xi, \eta)$.

Proof Here $\nabla_\xi G$ is the gradient vector of G as a function of ξ. It follows from the definition of G that

$$G(\xi, \eta) = \min_\zeta [G(\xi, \zeta) + G(\zeta, \eta)].$$

Taking $\zeta = \xi + u\delta$, where δ is a small positive scalar and u a unit vector, we find that this equation becomes, in the limit of small δ,

$$\min_u (q + u^\top \nabla G) = 0. \tag{2.37}$$

Here the active argument ξ is understood. The minimising u is in the direction of $-\nabla G$ and, of course, of unit modulus, whence both assertions of the theorem follow. $\Diamond$

The solution of the uniform case derived in Section 1.4 will now carry over to this case, with the geodesic paths just determined replacing the straight-line paths. The allocation of flow over these paths is now determined by the minimisation of $\sum_h \sum_i G(\xi_h, \xi_i) x_{hi}$

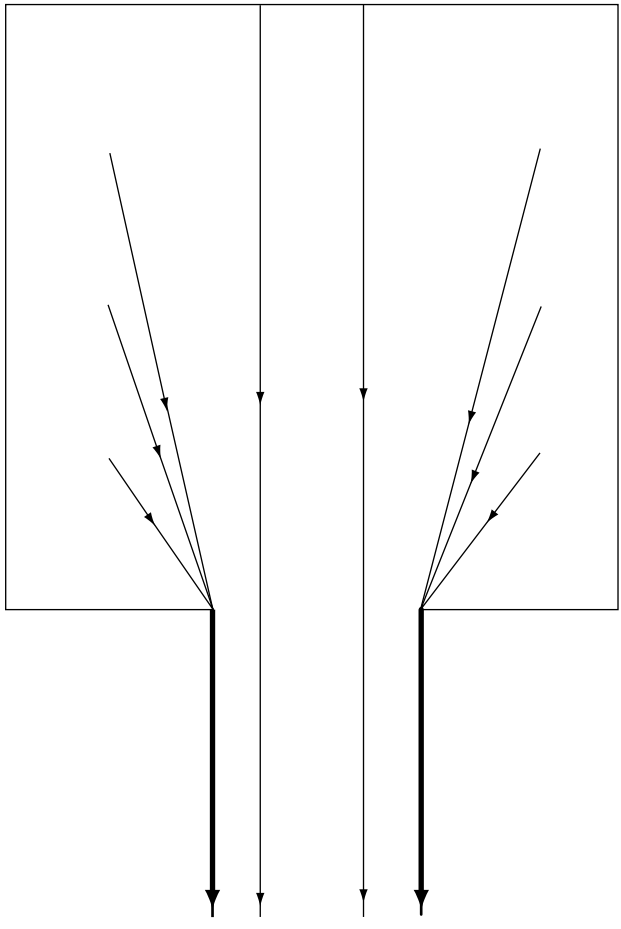


Fig. 2.11 Some of the 'optimal' paths for the two-dimensional cooling problem with a uniform exhaust-port condition at infinity.

with respect to the x_{hi}, subject to input/output constraints. Of course, this mathematical determination of the geodesic paths and their costs will seldom be practical – a more qualititative approach with a detailed map will be the usual course. However, if the area is uniform apart from some insurmountable obstacles, the tightening of a string threaded between or around the obstacles will yield the geodesic.

However, even this last case can present difficulties, as one can see by return to the cooling example. Suppose that the heat sink is prolonged into a 'heat exhaust' as in Figure 2.11, and that we phrase the problem as one of transferring the uniform inflow of heat to the plate to a uniform outflow of heat across a distant cross-section of the exhaust conductor. Then the geodesics will fall as illustrated (the continuum limit must be understood). They give rise to infinite flow densities (and so infinite conductor densities) along the edges of the exhaust conductor.

3

Multi-commodity and destination-specific flows

3.1 Reduction of the problem

In realistic cases one wishes to transport several types of commodity simultaneously over the network. A special case, which arises in both road traffic and communication contexts, is that in which the traffic is labelled by its destination, which we can take to be one of the nodes of the network. This is then again a 'multi-commodity' flow, with the simplifying feature that traffic of a given type has a unique destination.

If we assume seminvariant scaling of traffic costs, then the reduction of Section 1.4 carries over, yielding an optimal structure of the same simple form. Indeed, not only simple, but naive; the real world imposes constraints which we have yet to consider.

Let us index the type of commodity by i. This has a degree of consistency; if traffic is labelled by its destination and nothing else then the destination node i also defines its class. We shall now suppose that arcs are directed, so that the flow $j \to k$ is to be distinguished from the flow in the opposite direction, which indeed must be regarded as being carried on a separate (even if twinned) directed arc $k \to j$. The flow rate of commodity i on the jk link will be denoted x_{jki}, and the flow rate of commodity i from the external world into node j will be denoted f_{ji}. This last expression will be positive or negative according as j is a source or a sink node for commodity i. We have then the balance relation

$$f_{ji} + \sum_k (x_{kji} - x_{jki}) = 0$$

at node j.

Assume the measure of i-traffic so scaled that

$$x_{jk} = \sum_i x_{jki}$$

is a proper measure of the traffic burden on the jk link. Assume also that traffic costs are given by the seminvariantly scaled measure

$$C(x|a) = \sum_j \sum_k a_{jk} d_{jk} \phi(x_{jk}/a_{jk}), \tag{3.1}$$

and plant-leasing costs by

$$C(a) = \gamma \sum_j \sum_k a_{jk} d_{jk}.$$

Here d_{jk} is the Euclidean distance between nodes j and k and a_{jk} is, as ever, the 'cross-section' or rating of the jk link. Note that the summation $\sum_{j,k}$ has been replaced by a full double summation $\sum_j \sum_k$, since the links jk and kj are now to be distinguished.

It follows, then, as in Section 1.4, that the minimum with respect to the ratings a_{jk} of the sum $C(x, a)$ of these two expressions is

$$\overline{C}(x) = q \sum_j \sum_k d_{jk} x_{jk} \geq q \sum_i \sum_j \sum_k d_{jk} \overline{x}_{jki}, \tag{3.2}$$

the first equality being attained for $a_{jk} = x_{jk}/p$. Here p and q have the values

$$p = \psi'(q), \quad q = \psi^{-1}(\gamma)$$

determined in Section 1.4 and $\overline{x}_{ijk}$ is the amount of the inward flow f_{ji} that emerges from the net at node k. Equality will hold in the last inequality of (3.2) if such traffic takes the straight-line path.

The final stage of optimisation is then to choose the nonnegative allocations $\overline{x}_{ijk}$ so as to minimise the expressions

$$\overline{C}_i(x) = q \sum_j \sum_k d_{jk} \overline{x}_{jki}, \tag{3.3}$$

subject to the constraints

$$\left.\begin{array}{ll} \sum_k \overline{x}_{jki} = f_{ji} & (\text{if } j \text{ is a source node for class } i), \\ \sum_j \overline{x}_{jki} = -f_{ki} & (\text{if } k \text{ is a sink node for class } i). \end{array}\right\} \tag{3.4}$$

This linear programming problem is to be solved separately for each i. Summarising, we have then

Theorem 3.1 *The optimal transportation policy is solved by:*

(i) Determining the optimal allocations $\overline{x}_{jki}$ by solving the linear programming problems of minimising the form (3.3) subject to conditions (3.4), for each i.
(ii) Routing the traffic $\overline{x}_{jki}$ by the straight-line path from node j to node k.
(iii) Allowing a rating of $a_{jk} = \sum_i \overline{x}_{jki}/p$ on this path.

For the special case of destination-specific traffic (which we shall assume to be uniform in every property but destination) the result is even simpler, because we can omit stage (i) of the theorem.

Theorem 3.2 *For the case of destination-specific traffic we can identify i with the prescribed destination. The rules of Theorem 3.1 then simplify to:*

(i) If j is a source node for i-traffic (i.e. $f_{ji} > 0$) then all f_{ji} of this traffic is routed to the destination i by the straight-line path.
(ii) This path is allocated a rating $a_{ji} = f_{ji}/p$.

This theorem is an immediate corollary of Theorem 3.1, but we state it in full to demonstrate the lack of realism it reveals. One could never have a road network that gave a straight-line path between every possible origin/destination pair: we formalise the expression of this point in Chapter 5. Obviously our formulation neglects fundamental factors. One might indeed allow straight-line paths if one were considering sea- or air-routes, with no barriers in the form of land masses, navigation hazards or political considerations. However, one could not possibly have a permanent road network that cut up space as proposed. Land (and large unbroken blocks of it) is required for the hundred other purposes of urban and rural life, not to mention the civil engineering problems posed by rivers, mountains and difficult terrain generally. The design of a road network is then just one facet of a much larger planning problem, taking aesthetic as well as utilitarian and physical considerations into account.

Constraints on design and design space obtrude in almost every application, and the question is whether these can be given a manageable form (i.e. one short of considering the complete planning problem) that allows a natural resolution of the problem. One phenomenon to be expected is that of trunking: that active paths, which run sufficiently close and parallel to each other, will be merged into a single path of greater capacity, which can of course be entered and left where appropriate. This combined path constitutes what one would term a trunk route, and one would imagine that, if realistic assumptions were made, then the optimal design would incorporate such routes.

We shall see in the next chapter that trunking indeed occurs in the optimal design if the load pattern (as specified by the prescribed input/output flows f_{ji}) varies in time. This is because the flow in the trunk route can be steadier (i.e. show less proportional variation) than it would be on minor routes, and so allow better utilisation of network capacity.

In Chapter 5 we face up to the deeper consideration: that a route laid across open country in a sense blights that country.

3.2 The dual and its interpretation

In the last section we demonstrated, by an argument purely in terms of the primal formulation, that the optimal design showed the same kind of collapse as was observed for the distribution problem of Section 1.4. However, the fact that traffic is now labelled by its destination does add structure, which is revealed by the formulation of the dual problem for a general net: i.e. one that has not collapsed to the trivial structure of the naively optimised net.

Let us then adapt the Lagrangian treatment of Section 1.2 to the particular case of destination-specified traffic. We assume that the optimal flow rates for a given net can somehow be enforced. The intriguing question of enforcement is touched on in Section 6.5. If the system assumes link costs $c_{jk}(x_{jk})$ dependent only upon total link flow x_{jk} then the sum $\sum_j \sum_k c_{jk}(x_{jk})$ is to be minimised with respect to the x_{jki} subject to the constraints $x_{jki} \geq 0$ and

$$x_{jk} = \sum_i x_{jki}$$

for all relevant i, j, k, and also to the balance equation

$$f_{hi} + \sum_j (x_{jhi} - x_{hji}) = 0 \quad (h \neq i).$$

There is no need for a balance relation when $h = i$; all i-traffic entering the i-node simply departs, and none continues in the net.

We shall not suppose seminvariant scaling to begin with, but merely that the cost functions $c_{jk}(x_{jk})$ fall in the class $\mathcal{C}_s$ of convex functions defined in Section 1.2. The Lagrangian form $L(x, y, z)$ is then

$$\sum_j \sum_k [c_{jk}(x_{jk}) + z_{jk}(\sum_i x_{jki} - x_{jk})] + \sum_h \sum_i y_{hi}[f_{hi} + \sum_j (x_{jhi} - x_{hji})], \tag{3.5}$$

with the understanding that $y_{ii} = 0$ for all i. The appropriate values of the multipliers have their usual shadow-price interpretations: y_{hi} is the marginal cost of accepting more i-traffic at node h, and z_{jk} is the marginal cost of accepting more traffic on link jk. We derive the optimality conditions

$$c'_{jk}(x_{jk}) \geq z_{jk}, \tag{3.6}$$

with equality if $x_{jk} > 0$, and

$$y_{ki} - y_{ji} + z_{jk} \geq 0, \tag{3.7}$$

with equality if $x_{jki} > 0$. To see what these relations imply, define the Fenchel transform

$$c^*_{jk}(z_{jk}) = \sup_{x \geq 0} [z_{jk} x - c_{jk}(x)],$$

which can be regarded as the maximal net profit that the system could make on the jk link if traffic on that link were compelled to pay a toll of z_{jk}. If we then regard the z_{jk} as tolls, to be determined, we can summarise conclusions as follows.

Theorem 3.3 *(i) The multipliers y_{ji} are determined in terms of the tolls z_{jk} by the recursion*

$$y_{ji} = \min_k (z_{jk} + y_{ki}) \tag{3.8}$$

with the terminal condition y_{ii}=0.

(ii) Relation (3.8) establishes y_{ji} as the minimal total toll payment that will cover passage (of a unit flow) from node j to node i. Any optimal direct move of i-traffic from node j can only be to a node k that achieves the minimum in (3.8).

(iii) The tolls z_{jk} are determined as the values maximising the expression

$$\sum_h \sum_i f_{hi} y_{hi} - \sum_j \sum_k c^*_{jk}(z_{jk}) \tag{3.9}$$

with y determined in terms of z as in (i).

Proof Assertions (i) and (ii) follow from inequalities (3.6) and (3.7), with the cases of equality noted. Relation (3.8) is the dynamic programming equation that would be obeyed if one were looking for the route from node h to node i that minimised the total toll over the route. Assertion (iii) is then just a statement of the dual problem. ◊

For given toll values, the first term in expression (3.9) is the minimum (over routings) of total toll payments, and the term then subtracted is the maximum (over flows) of the net profit from flow in the system. Note that both double summations in (3.9) run over all node pairs of the net, for we have assumed that any node of the net may be a destination node for traffic starting from any other node.

The appearance of the dynamic programming equation (3.8) is interesting, with its picture of a unit of traffic (a motorist, say) who seeks a path through the net to his destination that minimises his total toll bill.

Expression (3.9) is just the dual objective function. If we make the seminvariant-scaling assumption $c(x) = ad\phi(x/a)$ then this form becomes

$$\sum_h \sum_i f_{hi} y_{hi} - \sum_j \sum_k a_{jk} d_{jk} \psi(z_{jk}/d_{jk}). \tag{3.10}$$

The expression

$$\frac{z_{jk}}{d_{jk}} = \frac{y_{ji} - y_{ki}}{d_{jk}}$$

is just the rate of toll per unit distance on the jk link, and the second expression again identifies it as the gradient of a potential. However, note from (3.8) that this second expression can be asserted only if the jk link is part of an optimal route for i-traffic.

3.3 An alternative formulation

The labelling of traffic by destination leads to another description of the state of the net, which has its advantages. Suppose for concreteness that the flow consists of vehicles on a road net. One may postulate that a vehicle is seeking a route r through the network, one compatible with its entrance node h and its exit node i. Then specification of the flow rates x_r along the possible routes ($r \in \mathcal{R}$) of the net gives a full description of the state of the network (in the deterministic equilibrium sense); a description which automatically satisfies the balance equations at those nodes that have no external input.

If we take the formulation in terms of route flows x_r then the Lagrangian form (3.5) becomes

$$L(x, y) = \sum_j \sum_k c_{jk} \Big(\sum_r a_{jkr} x_r \Big) + \sum_h \sum_i y_{hi} \Big(f_{hi} - \sum_{r \in \mathcal{R}_{hi}} x_r \Big).$$

Here a_{jkr} is 1 or 0 according as route r traverses segment jk or not, and $\mathcal{R}_{hi}$ is the set of routes that begin at node h and end at node i. Minimising with respect to x_r, we obtain the condition

$$\sum_j \sum_k a_{jkr} z_{jk} - y_{hi} \geq 0 \quad (r \in \mathcal{R}_{hi}), \tag{3.11}$$

with equality if $x_r > 0$. Here $z_{jk} = c'_{jk}(x_{jk})$, as before. Condition (3.11) and its equality case again imply that, for all optimal routes in $\mathcal{R}_{hi}$, the sum of z-values over the route $h \to i$ is minimal.

4

Variable loading

4.1 Variable loading and the primal problem

The common situation is that loading is variable, manifested in variation of the loading vector f, possibly systematic or possibly random. In generalising to this case, we shall for simplicity consider just the case of simple flows. The analogous analysis of the multi-commodity case is then fairly evident, and we shall in fact take destination-specific traffic as the standard case in the next chapter.

Assume then, as in Chapter 1, a discrete network with seminvariantly scaled flow costs and undirected arcs. If structural costs are added in, then we have a total cost function

$$C(x, a) = \sum_{j,k} d_{jk} a_{jk} [\phi(x_{jk}/a_{jk}) + \gamma]. \tag{4.1}$$

The d_{jk} are assumed to be Euclidean distances in physical space. Associated with the cost function (4.1) is the Lagrangian form

$$L(x, y, a) = C(x, a) + \sum_j y_j (f_j - \sum_k x_{jk}), \tag{4.2}$$

which takes account of the balance constraints at the nodes.

It is the exogenous flow rates f that express the loading. In most practical situations the system will have to operate under a variety of loading patterns, and so under a variety of f-values. This can have a profound effect on the form of the optimal structure. We shall not consider the case for which f is changing on a short time scale, which would involve us in a full-dress control problem. Rather, we shall assume that the f_j adopt constant values $f_j(\omega)$ for a considerable time, a time that is long relative to the time needed for the network flow x to settle to its cost-minimising value $\hat{x} = x(\omega)$. We can then speak of the system as being in 'regime ω', and can average costs over regimes. That is, if regime ω has probability $\pi(\omega)$ and z is some numerical-valued quantity that is a function $z(f)$ of f, then the expectation of z (also its average value over time) will be

$$\mathrm{E}(z) = \sum_\omega \pi(\omega) z[f(\omega)].$$

Note that the roman E is used to denote the expectation operator. We have assumed that f takes a discrete set of values. The formal extension to more general cases is obvious,

and we make no apologies for a probabilistically naive presentation when pathology-free versions of the model already encapsulate the real problems. The ω notation is one traditionally used for the 'elementary outcomes' of a probabilistic model and, in our notationally challenged state, we seize upon it gratefully.

The design, as expressed by a, will not depend upon f, for we suppose it fixed in time. It will, however, depend upon the distribution of f. We shall assume that the design is to be chosen to minimise the expected cost, although we shall see that there are cases for which this criterion needs discussion. The cost-minimising flow $\hat{x}$ is a function of both a and regime ω. It is the second dependence which makes it a random variable. The problem is then to choose a to minimise $\mathrm{E}[C(\hat{x}, a)]$. If we take the dual formulation then a should be chosen to minimise $\mathrm{E}[\max_y \min_x L(x, y, a)]$, where separate x and y extremals are taken for every value of ω.

For explicitness of treatment, it is now convenient to specialise to the case, itself a substantial and typical one, when ϕ is a power function:

$$\phi(p) = \frac{1}{\alpha} p^{\alpha}. \tag{4.3}$$

The fact that $a\phi(x/a)$ then factorises into a function of x and a function of a makes the optimisation of a under random variation of x much simpler. We require that $\alpha > 1$, for strict convexity. Any numerical multiple of expression (4.3) would do as well, but this can be standardised to unity by a rescaling of γ and of total cost.

If we begin with the assumption that $\hat{x} = x(\omega)$ has been determined for every ω (possibly by the physical processes of the system itself) then we can quickly obtain a partial analogue of Theorem 1.2.

Theorem 4.1 *The assumptions (4.1) and (4.3) and the criterion of minimal expected cost imply the following assertions:*

(i) The optimal value of a_{jk} must satisfy

$$a_{jk}^{\alpha} = \frac{1}{\beta\gamma} \mathrm{E}(|\hat{x}_{jk}|^{\alpha}). \tag{4.4}$$

Its substitution leads to the reduction

$$\min_a \mathrm{E}[C(\hat{x}, a)] = \beta\gamma \sum_{j,k} a_{jk} d_{jk} = \beta\gamma S, \tag{4.5}$$

say. Here β is the value conjugate to α; see (1.14).

(ii) If node j is an internal node of the system whose position in physical space may be optimised, and if the ratings a_{jk} have been determined by (4.4), then a necessary condition for position-optimality is

$$\sum_k a_{jk} u_{jk} = 0, \tag{4.6}$$

where u_{jk} is the unit vector in the direction of $\xi_k - \xi_j$.

Proof Minimisation of $\mathrm{E}[C(\hat{x}, a)]$ with respect to a_{jk} yields (4.4) immediately, and the reduction (4.5) is easily verified. Of course, $\hat{x} = x(\omega)$ is also a function of a, but the fact that we have evaluated expression (4.1) at its minimum with respect to permissible variations in x for each regime means that this dependence will not affect the stationarity condition with respect to a. Note the implication of (4.5): that, if the ratings a have been optimised (although not necessarily anything else), then the minimised cost is proportional to the consequent material cost.

Now, if ξ_j is perturbed by a small vector amount δ, then d_{jk} is perturbed by an amount

$$|\xi_j + \delta - \xi_k| - |\xi_j - \xi_k| = \frac{\delta^\top(\xi_j - \xi_k)}{|\xi_j - \xi_k|} = \delta^\top u_{jk}$$

to within first order terms in δ. Expression (4.5) is already stationary with respect to variations in a, and stationarity with respect to variations in ξ_j then implies condition (4.6). ◊

Direct determination of the optimum is still distant, since the $\hat{x}$ in (4.4) is a function of the existing a. However, the theorem has an immediate and significant implication: that variation in load will in general create the need for internal nodes in the optimal design. We discuss this matter in the next section, but note now one other immediate conclusion.

Corollary 4.2 *Any internal node of the optimal design lies in the convex hull of the external nodes: those nodes j for which $f_j(\omega)$ is nonzero for some ω.*

Proof We can rewrite relation (4.6) as

$$\sum_k w_{jk}(\xi_k - \xi_j) = 0,$$

where $w_{jk} = a_{jk}/|\xi_k - \xi_j|$ is positive. We then have

$$\xi_j = \frac{\sum_k w_{jk}\xi_k}{\sum_k w_{jk}},$$

which implies that ξ_j lies in the convex hull of the ξ_k for nodes k to which node j is directly linked. Iteration of this conclusion then exhibits node j as lying in the convex hull of the external nodes, prescribed by the statement of the problem. ◊

4.2 Internal nodes and trunking

The simplest of examples demonstrates the need for internal nodes in a freely optimised design (i.e. one allowing variation in the number and positions of nodes as well as in ratings). Consider the plane graph set out in Figure 4.1, with nodes 1, 2 and 0 at co-ordinates (0,1), (0, −1) and $(d_0, 0)$ respectively. We suppose that there are two regimes, $\omega = 1$ or 2, each of probability 1/2. Only one source node is active at a time: under regime ω node ω sends unit flow to node 0. If there were only a single regime, in which nodes 1 and 2 were simultaneous sources, then the optimal network would consist of direct 10 and

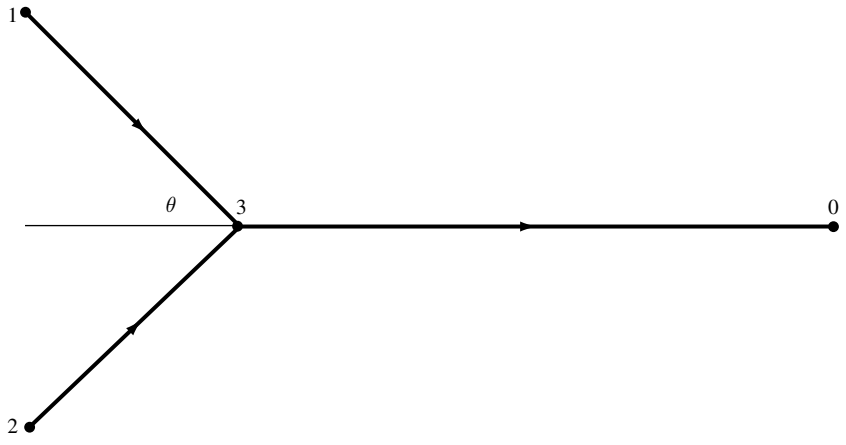


Fig. 4.1 The simplest example of trunking.

20 links. Consider the network suggested in the figure, with an internal node 3 at $(d_3, 0)$, say. We have then

$$\mathrm{E}(x_{30}^{\alpha}) = 1, \quad \mathrm{E}(x_{13}^{\alpha}) = \mathrm{E}(x_{23}^{\alpha}) = 1/2,$$

so that, by (4.4), a_{30}, a_{13} and a_{23} are proportional to 1, $2^{-1/\alpha}$ and $2^{-1/\alpha}$ respectively. Relation (4.6) will then imply that $2.2^{-1/\alpha}\cos\theta = 1$, where θ is the angle indicated in the diagram. We thus deduce that

$$\cos\theta = 2^{-1/\beta}. \tag{4.7}$$

If $d_0 > \cot\theta$ then the network is indeed improved by the insertion of the internal node 3; in the other case one will stay with direct 10 and 20 links. We shall demonstrate in the next section that these configurations are indeed optimal. The phenomenon is one of 'trunking': that is, if the sink node 0 is sufficiently distant then one gains a material economy by pooling the flows of the two source nodes into a compromise route, the trunk route 30.

In the cases of increasing convexity of the cost function, α = 1, 2 and $+\infty$, we find that θ = 0, 45° and 60° respectively, so that trunking takes place ever earlier on the route. In the case of linear costs, $\alpha = 1$, there is indeed no trunking – some strict convexity of the cost function is required if trunking is to be advantageous.

We have considered this symmetrical two-source, one-sink model for simplicity, but the asymmetric case of Figure 4.2 gives little more trouble. Suppose that the unit flow comes from source i with probability p_i $(i = 1, 2)$. We have then the optimality balance relations

$$\tau_1 \sin\theta_1 = \tau_2 \sin\theta_2, \quad \tau_1 \cos\theta_1 + \tau_2 \cos\theta_2 = 1,$$

where $\tau_i = p_i^{1/\alpha}$. From these relations it follows that the angles θ_i are determined by $\sin\theta_i = \tau/\tau_i$, where τ is determined by

$$\sqrt{\tau_1^2 - \tau^2} + \sqrt{\tau_2^2 - \tau^2} = 1.$$

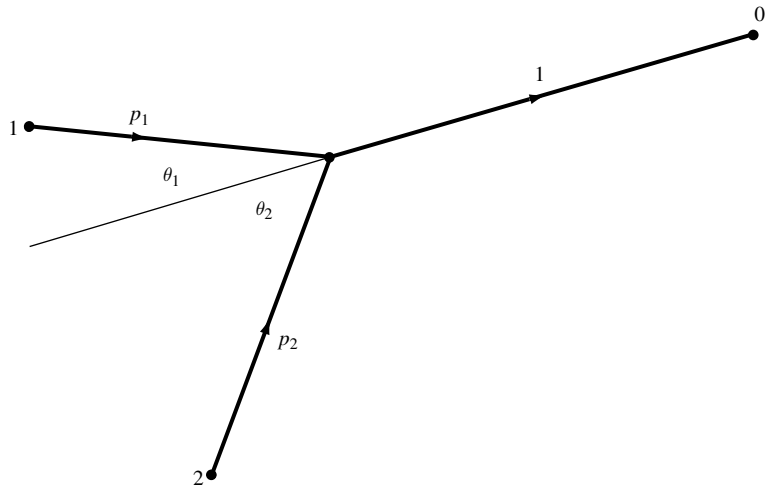


Fig. 4.2 The asymmetric version of the simple trunking problem.

This determination of course holds only if the geometry of the configuration allows it, in that the sink node is far enough from the other two that the configuration of Figure 4.2 with the recommended angles can be fitted in. In other cases there will be no trunking.

The reason why variable input can make trunking advantageous is that already stated in Section 3.1. A steady flow utilises the capacity of the links along which it travels better than would a variable flow. If one can then pool variable flows in such a way as to achieve a proportionately steadier total flow, one has made an economy which might justify the expense of a longer path. In the simple example above, the 10 and 20 flows were exactly complementary, in that their sum had the constant value of unity. We can express the point mathematically by considering a slightly more general example.

Suppose that there are random flows x_j to be conveyed from nodes j to a distant node 0, as in Figure 4.3. There will be an incentive to trunk these flows if the sum of costs per unit distance of individual flows is greater than the cost per unit distance of the sum of the flows. We speak of 'per unit distance' because the assumption that node 0 is distant from the group of j-nodes implies that the proportional variation in the distances d_{j0} is small. The criterion for trunking is then that the margin of inequality in the relation

$$\sum_j \{\mathrm{E}(|\hat{x}_j|^\alpha)\}^{1/\alpha} \geq \{\mathrm{E}|\sum_j \hat{x}_j|^\alpha\}^{1/\alpha} \tag{4.8}$$

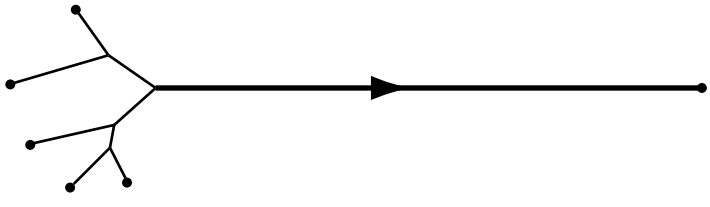


Fig. 4.3 Trunking to a distant destination.

should be sufficiently large. More exactly: that the ratio of the right-hand member of (4.8) to the left-hand member should be sufficiently small. Inequality (4.8) is certainly valid; it is Minkowski's inequality. Equality holds in (4.8) if and only if the x_j are all proportional to a single scalar random variable. The implication is worth stating as a theorem.

Theorem 4.3 *Assume the properties (4.1) and (4.3) of seminvariantly scaled and power-law costs. Then, if the variable loading reduces to a variable scaling of a fixed loading pattern, there is no incentive to trunk. The optimal net in this case thus consists of direct source-to-sink links with dimensions determined by (4.4).*

The case when trunking is of the greatest benefit is that for which the $\hat{x}_j$ are random, but their sum is not.

The trunking induced by a variable load then has a rather limited character; it stabilises when flows have been aggregated to the degree that further aggregation brings no worthwhile smoothing of flow. The simple example of Figure 4.4 makes the point. Suppose that there are two regimes of probability 1/2; in regime 1 nodes 1 and 3 wish to send unit flows to node 0, and in regime 2 nodes 2 and 4 wish to send such flows. With the geometry indicated there is then an incentive to aggregate the 10 and 20 flows, and likewise the 30 and 40 flows, as in the figure. However, there is no incentive to trunk further, as the two aggregated flows each have the constant value of unity.

It seems likely that the internal nodes located in this fashion would also provide the natural locations for storage depots, used to smooth out statistical fluctuations in supply or demand. The case for trunking is strengthened in the next chapter, when we consider environmental factors. The internal nodes thus determined would again serve as natural depots, but now justified by a need to switch between heavy and light modes of transport as well as to cushion against fluctuations.

There is one other point that should be made. The moment (4.4) may be infinite for α larger than some critical value. The clearest case of this is the case $\alpha \to +\infty$, when $(x_{jk}/a_{jk})^{\alpha}$ will become infinite if flow x_{jk} exceeds rating a_{jk}. In this case relation (4.4) becomes

$$a_{jk} = \max(|\hat{x}_{jk}|),$$

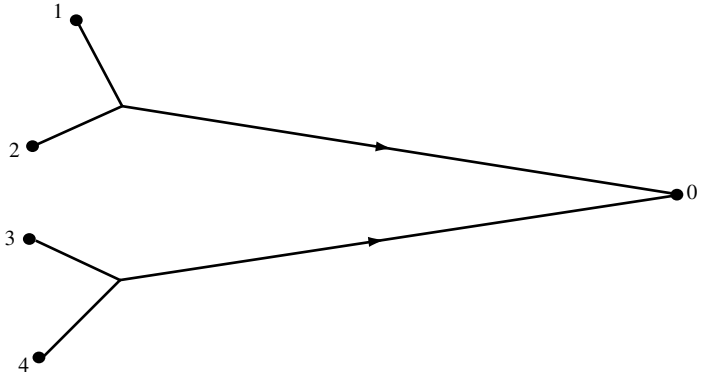


Fig. 4.4 The cessation of trunking once steady flows are attained.

where the maximum is over all values of the random variable $x_{jk}(\omega)$ that have positive probability. In other words, the ratings of the network links must be adequate for all possible eventualities. This may well be an impractical requirement. An engineer, for example, may exclude in his design those eventualities that are expected to occur less than once in 50 or 100 years. That is, his criterion is to bring the probability of failure over a prescribed period down to a tolerable value.

For values of α for which the moments (4.4) exist it is a question of overstrain rather than of failure, and in minimising expected costs one is making a rational assessment of this strain.

4.3 Optimality and load splitting

We shall see in the next section that the dual formulation provides, in principle, a test of the optimality of a proposed structure. However, one can often reason more directly. Consider the situation in which the commodity handled by any of a number of randomly supplied source nodes is to be conveyed to a single specified destination node. The aim is then to make source-to-sink routes as direct as possible, while taking advantage of the benefit of combining flows. This benefit is expressed by inequality (4.8), which holds whatever the statistics of the flows. All arcs of the net must then be straight, and all nodes must be trunking nodes. The optimal net is thus a tree with straight arcs, rooted at the destination node and with its tips at the source nodes.

The optimised four-node nets of the last section are thus indeed optimal. For more general examples one must apply condition (4.6) at the nodes, or apply a dynamic version of this condition to seek the optimal configuration; see Section 4.8.

For an example of a somewhat different character, consider the net of Figure 4.5. There are four nodes, placed at the corners of a square of side $\sqrt{2}$, so that the diagonals are of length 2. In regime 1 there is a unit flow from node 1 to node 3; in regime 2 there is a unit flow from node 2 to node 4; the two regimes are equally probable. If one takes the links along the diagonals, as in Figure 4.5(i), they carry a unit flow half the time, and so must have cross-section proportional to $2^{-1/\alpha}$. If one takes them along the sides, as in Figure 4.5(ii), then they carry flow 1/2 all the time, and so must have cross-section proportional to 1/2. The total costs in the two cases are proportional to the material costs, by relation (4.5), and these are readily verified to be proportional to $2^{2-1/\alpha}$ and $2^{3/2}$, respectively. Hence the first design is preferable for $\alpha \leq 2$, the second for $\alpha \geq 2$. There

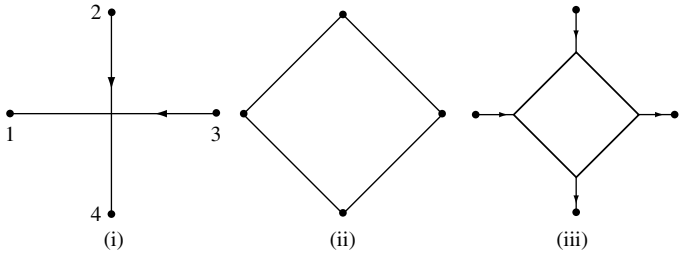


Fig. 4.5 Optimal configurations for the four-node example of the text in the cases: (i) $\alpha \leq 2$, (ii) $\alpha \geq 2$ and (iii) $\alpha = 2$.

is a trade-off between length of path and variability of flow, and increasing convexity of the flow–cost function implies increasing averseness to variability. For the transition case $\alpha = 2$ all the intermediate structures of Fig. 4.5(iii) incur the same cost.

These suggested configurations are in fact optimal, as can be seen by appeal to the dual formulation. To run ahead of ourselves a little, in this formulation one must calculate a dual field $y(\xi, \omega)$ for every regime ω, these being subject to the condition

$$\mathrm{E}[|\nabla y.u|^{\beta}] \leq \beta\gamma \tag{4.9}$$

for all ξ in the design space and all unit vectors u. If one can find such fields for which line-elements of the proposed net lie only at (ξ, u) such that equality holds in (4.9), then one can derive from these a lower bound for minimal cost, and so test the proposed net for optimality.

In the present case, consider a field y wholly in the direction of the intended transportation in the two regimes, and so in the direction of the ωth co-ordinate axis in regime ω. If we suppose u inclined at an angle θ to one of the co-ordinate axes then we have

$$\mathrm{E}[|\nabla y.u|^{\beta}] \propto |\cos\theta|^{\beta} + |\sin\theta|^{\beta}.$$

For $\alpha < 2$ (and so $\beta > 2$) this expression is maximal at θ equal to 0 or $\pi/2$. For $\alpha > 2$ (and so $\beta < 2$) it is maximal at θ equal to $\pm\pi/4$. These directions correspond to the nets proposed in the two cases, and we can complete the argument in the next section to demonstrate optimality of the suggested nets. In the transitional case $\alpha = \beta = 2$ the expression above is constant in θ, consistent with the high degree of indeterminacy of the optimal net in this case.

The second case of this example is interesting, in that the incoming load is split between two routes in the optimal design – the reverse of trunking. Part of the aim is again to achieve steadiness of flow rates, and in the case $\alpha > 2$ this consideration outweighs the disadvantage of longer routes.

4.4 Variable loading in the dual formulation

In the dual formulation the ratings a are to be chosen to minimise the expression $\mathrm{E}[\max_y \min_x L(x, y, a)]$. By the argument of Section 1.5 it is legitimate to commute the operations of y-maximisation and a-minimisation, and the conclusion of Section 1.5 has its analogue.

Theorem 4.4 *The dual problem for the optimised network is: choose y as a function $y(\xi, \omega)$ to maximise $\mathrm{E}[\sum_j y_j f_j]$ subject to the constraints*

$$d_{jk}^{-\beta}\mathrm{E}[|y_j - y_k|^{\beta}] \leq \beta\gamma. \tag{4.10}$$

A jk-link can exist in the optimal design only if equality holds in (4.10) for those nodes.

Note that it follows from the relation

$$\left(\frac{x_{jk}}{a_{jk}}\right)^{\alpha-1} = \frac{y_j - y_k}{d_{jk}}$$

that

$$\mathrm{E}\left(\frac{x_{jk}}{a_{jk}}\right)^{\alpha} = \mathrm{E}\left(\frac{y_j - y_k}{d_{jk}}\right)^{\beta}.$$

Thus, if equality holds in (4.4) it will also do so in (4.10), and vice versa. That is, if link ratings have been optimised for a given design then the bound (4.10) will be exactly attained on those links, and vice versa. However, what remains open in the first case is that one could with advantage introduce or vary internal nodes, and, in the second, that the inequality (4.10) could be violated by some *jk* pair.

The continuum form of the dual follows as in Section 2.2.

Theorem 4.5 *(i) The continuum form of the dual problem for the optimised network is: choose the function $y(\xi, \omega)$ to maximise*

$$D(y) = \mathrm{E}\int yf\mu(\mathrm{d}\xi) \tag{4.11}$$

subject to the constraints

$$\mathrm{E}[|\nabla y.u|^{\beta}] \le \beta\gamma, \tag{4.12}$$

for all ξ in $\mathcal{D}$ and all unit vectors u. It is understood in the integral of (4.11) that y and f are corresponding random functions $y(\xi, \omega)$ and $f(\xi, \omega)$ of ξ.

(ii) A line element of material can exist at (ξ, u) in the optimal design only if equality holds for these values in (4.12).

The necessity of preserving the directional argument u is seen from relation (4.12). The averaging over regimes ω is carried out for a given value of u. It will, in most cases, be true, however, that the left-hand expression in (4.12) will reach its extreme only for a single value of u (or its negative). Consider, for example, the power-law case (4.3) with $\alpha = 2$. Then $\mathrm{E}\psi(\nabla y.u) \propto u^{\top}Q(\xi)u$, where Q is what will later be seen as the 'principal strain matrix' $\mathrm{E}[(\nabla y)^{\top}(\nabla y)]$, evaluated at ξ. Then the extreme is reached in (4.12) only when u is proportional to an eigenvector of Q corresponding to a maximal eigenvalue (all eigenvalues are nonnegative). If this eigenvalue is simple then the flow is simple, in that it constitutes a vector field at ξ in the direction of the corresponding eigenvector of Q. See the discussion at the end of Section 2.1.

The introduction of the dual variables creates the possibility of testing a proposed design for optimality, a technique well established in the case of a fixed load. In this case we have the sequence of relations

$$C(\hat{x}, a) = L(\hat{x}, y, a) \ge \min_{x} L(x, y, a) = D(y, a),$$

where the minimisation in the third expression is unconstrained. The value of the dual form $D(y, a)$ for any y is thus a lower bound to the total cost when flow is optimised (and in fact one for which equality holds for y under regularity conditions). This continues to hold when design is optimised,

$$\min_{a} C(\hat{x}, a) \ge \min_{a} D(y, a). \tag{4.13}$$

However, if the minimal cost in the first expression is bounded below, then the bound (4.13) is useful only for y such that the right-hand side is also bounded below. This means, in the continuum treatment of Section 2.2, that the fields $y(\xi)$ obey the condition $|\nabla y.u| \leq q$. The lower bound of (4.13) becomes simply the criterion function $\sum_j y_j f_j$, or its integral equivalent in more general cases. The restriction on the choice of y decreases the maximal value of the bound, reflecting exactly the effect of optimisation.

This argument has a complete analogue in the variable load case, and we need merely state the conclusion.

Theorem 4.6 *Consider a given design, whose cost must then be an upper bound to the cost of the optimal design. Then expression (4.11) provides a lower bound to this minimal cost, which will be bounded below if the random field $y(\xi, \omega)$ is chosen to satisfy condition (4.12).*

If one conjectures that the proposed design is optimal then, in looking for a trial field $y(\xi, \omega)$, one should add the condition that equality should hold in (4.12) for ξ at which the design has material.

One verifies easily in this way that the solutions suggested for the final example of Section 4.3 are optimal, by appealing to the y-fields suggested, with an appropriate choice of the multiplicative constant.

4.5 Degree of trunking

In this section and the next we consider partial optimisations of the multiple-source, single-sink model, designed to illuminate a couple of points. In Section 4.7 we consider the full-dress evolutionary approach to determination of the optimal structure in the general case, and in Section 4.8 the reduction of this when it is established that the optimal structure is discrete.

Suppose we have a cluster of source nodes with supply rates f_j and co-ordinates ξ_j, all supplying a relatively distant sink node at co-ordinate ξ_0, as in Figure 4.3. We shall suppose that the f_j are independently and identically distributed. The question is then: if one chooses to trunk n of the source nodes together, what would be the optimal value of n? The point of considering the case of independent sources is that trunking will then decrease the relative variability of the flow, but not so powerfully as in cases for which an increase in the flow of one source is invariably balanced by a decrease in another.

Results come easiest if we take the case $\alpha = 2$. The rating of a link fed from n sources will then be

$$a_n \propto [\mathrm{E}(\sum_{j=1}^{n} f_j)^2]^{1/2} = [(n\mu)^2 + n\sigma^2]^{1/2},$$

where μ and σ are the mean and standard deviation of a single source rate.

Suppose that we now amalgamate n sources at an internal node placed at $\bar{\xi}$. The total length of links within the cluster of n and between this cluster and the sink are then

$$D_n = \sum_{j=1}^{n} |\xi_j - \bar{\xi}|, \qquad d = |\bar{\xi} - \xi_0|.$$

One imagines that $\overline{\xi}$ will be chosen to minimise D_n, at least approximately, and that the value of d will be relatively insensitive to this choice.

By (4.6) the cost of this part of the network will be proportional to

$$C_n = a_1 D_n + a_n d.$$

If N supply nodes are to be connected to the sink by N/n such trunkings then the total cost is NC_n/n and we should choose n to minimise C_n/n. (We have neglected the fact that one choice of a domain of n nodes will restrict the choices for the rest; such restrictions will affect the shape of the domains of n and so the form of D_n.)

If one is in a space of ℓ dimensions, if sources have a relatively constant density within the cluster, and if domains are approximately of the same shape, then a domain of n sources will have a scale of $n^{1/\ell}$ and D_n will be of the order of $n^{(\ell+1)/\ell}$. With this approximation we shall have

$$C_n/n \propto [\kappa n^{1/\ell}(1+v)^{1/2} + d(1+v/n)^{1/2}]. \tag{4.14}$$

Here the constant κ reflects the degree of spread in the source cluster, and $v = \sigma/\mu$ is the coefficient of variation of a single source flow. Minimising expression (4.14) with respect to n we obtain

$$n^{(\ell+1)/\ell} \approx \left(\frac{dv}{\kappa}\right) \frac{\ell}{2\sqrt{(1+v)(1+v/n)}}. \tag{4.15}$$

We see from (4.15) that the optimal value of n tends to increase with both increasing d/κ, the ratio of sink distance to source spread and v, the coefficient of source variation. For fixed v and increasing d/κ the optimal n is of the order of $(d/\kappa)^{\ell/(\ell+1)}$.

4.6 Trunking from a continuous source line

Conider the situation of an equispaced array of N sources along a straight line, feeding the sink point 0 of Figure 4.6. We shall suppose that only one of the sources is active at a given time, although all have the same probability $1/N$ of being so. One will then have a tree of links somewhat as sketched in the figure, and one asks for the optimal configuration. Analytical evaluation of the variable geometry required is forbidding, so we shall consider something of the type of Figure 4.7, with a fixed geometry throughout any given layer. We shall then not have the 'waisting' depicted in Figure 4.6. At each node of Figure 4.6 one would have an optimality condition of type (4.6), and it is the fact that flow leaves the node obliquely which pulls the net into its waisted shape. We shall apply condition (4.6) as though the flow left the node vertically, which is not so, but the condition nevertheless gives a guide to the desired ratio of spacing between layers to spacing within layers. We shall assume that the ultimate sink node 0 lies far enough from the source line that condition (4.6) can be applied at all stages, so that trunking is complete before node 0 is reached, as in Figure 4.7.

Actually, there are constraints on the value of N if a regular geometry is to be possible. For this reason, it is better to characterise the flow from the bottom up, and to see what the pattern of the higher reaches of the network should be if it is to fit in with the already

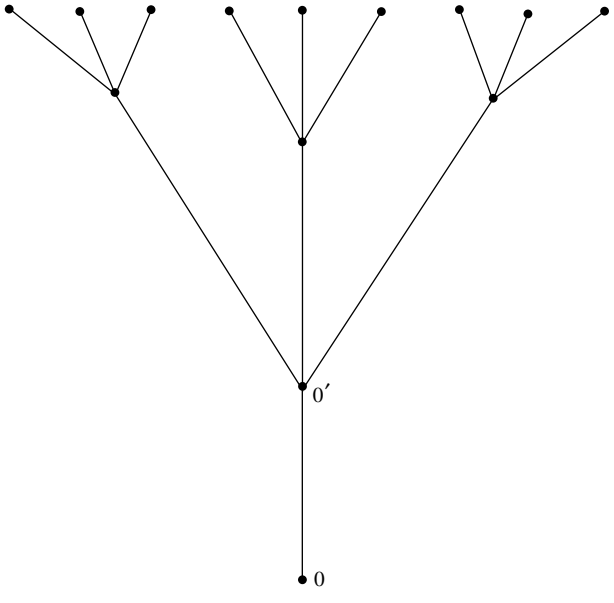


Fig. 4.6 The configuration for a many-to-one flow, exhibiting ‘waisting’.

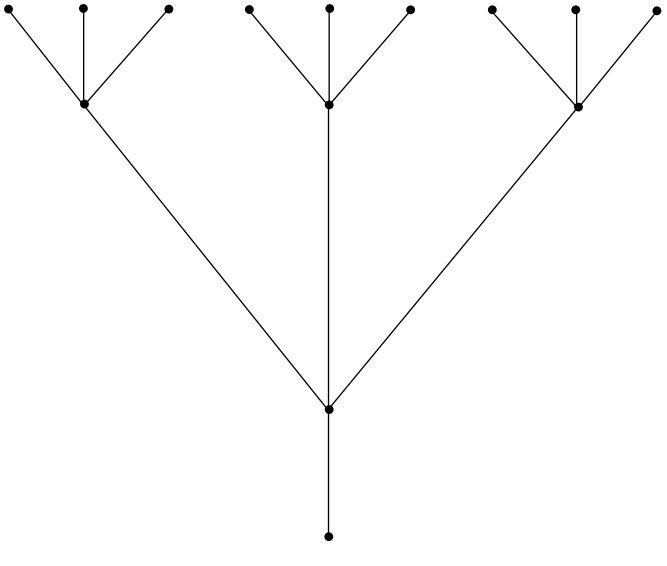


Fig. 4.7 The sub-optimal but tractable reconfiguration of Fig. 4.6, with regular geometry.

established pattern of the lower reaches. In this way we shall have successive layers of ever-denser evenly spaced nodes as we rise in the network, until the ‘continuous’ limit of zero spacing is reached at a finite height.

We thus visualise a sequence of layers of intermediate nodes, rising from the sink to the ultimate layer of the source line. The local scale is specified by the spacing s between nodes in the layer. If each node on a given layer takes the entire feed from n nodes on the layer above we shall speak of a ‘reduction factor’ of $1/n$.

Lemma 4.7 *The cost of transfer to a single node of the next lower layer has the form $ag(n)s$, where s is the spacing in that layer, $1/n$ the reduction factor and $a = [\mathrm{E}(\hat{x}^\alpha)]^{1/\alpha}$ is the rating required for links carrying flows $\hat{x}$ from the upper layer.*

Proof The ratio of effective flow rate from the lower layer to that from the upper layer is $n^{1/\alpha}$. For given n the optimality condition (4.6) at the node thus involves only the orientations of the links, and so is independent of scale. The optimal spacing between the layers will then be proportional to s, as will the total length of the n-to-one links. ◊

The function $g(n)$ is just the length of all the n-to-one links required if $s = 1$ and if the condition (4.6) is applied. It is thus in principle calculable; see the end of the section.

Theorem 4.8 *Consider a node in layer r with spacing s, which has probability p of receiving unit flow for further transmission. Let $F_r(p, s)$ be the minimal cost of that upper part of the network that feeds this node. Then:*

(i) The value function $F_r(p, s)$ obeys the dynamic programming equation

$$F_r(p, s) = \min_n [g(n)(p/n)^{1/\alpha} s + nF_{r+1}(p/n, s/n)]. \tag{4.16}$$

(ii) Consider the limit of indefinite progression to the case $s = 0$. Then this limit layer lies a finite distance above the sink node; the optimal value of n is fixed and equals the integer $\overline{n}$ minimising the expression

$$h(n) = \frac{g(n)}{n^{1/\alpha} - 1}.$$

The geometry of the net thus varies only in scale as one progresses through the layers.

(iii) The value function in this limit case has the form

$$F(p, s) = \overline{h} p^{1/\alpha} s, \tag{4.17}$$

where $\overline{h} = h(\overline{n})$.

Proof The form (4.16) of the dynamic programming equation is evident. If one assumes the value of n fixed then indefinite application of the recursion (4.16) in the case of r-independent F yields the evaluation

$$F(p, s) = h(n) p^{1/\alpha} s.$$

This would then imply the evaluation (4.17) if the optimal value of n were indeed fixed. To test the point, substitute solution (4.17) into the square-bracketed expression in the right-hand member of (4.16). If the integer value of n minimising this expression is still $\overline{n}$ then one has verified optimality of the fixed rule $n = \overline{n}$. But this expression is $p^{1/\alpha} s$ times

$$[g(n) + \overline{h}] n^{-1/\alpha} = \overline{h} + (1 - n^{-1/\alpha})[h(n) - \overline{h}],$$

which indeed attains its minimum for integral n at $\overline{n}$.

The assertion that the limit layer is reached in a finite distance follows immediately from the exponential diminution in scale as one progresses up the layers. ◊

The discrete net thus meets the continuous source line by a succession of layers of identical discrete geometry but exponentially decreasing scale. One can consider a finite-horizon case, which terminates at layer v, so that $F_v = 0$. Discussion of this case can supply any rigour felt to be lacking in the treatment of the case $v = \infty$. Suppose that the spacing at the level of the root $0'$ of the tree is s, meaning that we are considering an array of systems of the type of Fig. 4.7, side by side, with sink nodes spaced at distance s. Suppose we also assume a fixed value of n for the moment. Then the material cost of transmission to each root is

$$F_0(1, s) = \frac{g(n)(1 - n^{-v/\alpha})}{n^{1/\alpha} - 1} s,$$

as may be determined by direct calculation. The minimising value of n may well then be less than $\overline{n}$, giving an indication that the optimal value of n may then also decrease as the termination layer v is approached.

For explicitness we can determine the form of $g(n)$, at least approximately. Suppose that the layer with spacing s is separated from that above it by an amount t. Then the expression for the cost incurred in flow between these two layers is (to within a proportionality constant that we can suppose normalised to unity)

$$C = \sum_{j=-m}^{m} \sqrt{t^2 + (js/n)^2}. \tag{4.18}$$

Here we have supposed that n is odd and equal to $2m + 1$. The corresponding optimality condition (4.6) is (if we suppose the outflow from the lower layer to be vertical)

$$n^{1/\alpha} = \sum_{j=-m}^{m} \frac{t}{\sqrt{t^2 + (js/n)^2}}. \tag{4.19}$$

If we suppose the relevant value of n so large that we can approximate these sums by integrals, then evaluation (4.18) becomes

$$C \approx (2nt^2/s) \int_0^z \sqrt{1 + u^2}) \mathrm{d}u = (nt^2/s)[z\sqrt{1 + z^2} + \ln(z + \sqrt{1 + z^2})], \tag{4.20}$$

where $z = 2t/s$. The condition (4.19) correspondingly reduces to

$$n^{1/\beta} \approx \frac{z}{\ln(z + \sqrt{1 + z^2})}. \tag{4.21}$$

Relation (4.21) determines the ratio $z = 2t/s$ purely in terms of n, so that inter-layer spacing follows the scale of intra-layer spacing. We then see from (4.20) that cost C follows the scaling for given n asserted in Lemma 4.5, and relations (4.20) and (4.21) between them implicitly determine the function $g(n)$.

4.7 Evolutionary algorithms

The quantity being minimised is now the expectation of those considered in Section 2.5:

$$\mathrm{E}[C(\hat{x}, a)] = \mathrm{E}\sum_{j,k} a_{jk} d_{jk} [\phi(\hat{x}_{jk}/a_{jk}) + \gamma]$$

in the discrete case, and

$$\mathrm{E}[C(\hat{x}, \rho)] = \mathrm{E}\int\int \rho[\phi(\hat{x}/\rho) + \gamma]\mathrm{d}\xi\mathrm{d}u$$

in the continuum version. Here $\hat{x}$ is again the optimised flow for the current design and regime. Its dependence upon regime makes it a random variable.

The evolutionary equation (2.17) then becomes

$$\dot{\rho} = \kappa[\mathrm{E}(|\hat{x}|/\rho)^{\alpha} - \beta\gamma], \tag{4.22}$$

if we assume again the power-law form (4.3) for ϕ. The presence of the expectation in (4.22) may now induce the development of internal nodes in the optimal design. This expectation must be computed, requiring then the computation of $\hat{x}$ for the current value of ρ for all regimes ω. An alternative would be to delete the expectation sign in (4.22) and to change the regime randomly (according to the known statistics) as the calculation progresses, taking κ so small that there is an effective averaging over regimes during the computation. Such a course would be acceptable if one were simulating an evolution in a random environment of unknown statistical characteristics, but almost certainly one saves computation in the present case by simply computing the expectation from the known statistics.

Again, there is a great deal that is latent and unevident in relation (4.22). It of course implies the analogue $\mathrm{E}(|\hat{x}|/\rho)^{\alpha} = \beta\gamma$ of the optimality condition (4.4) once equilibrium is reached, but it must imply so much more. It should yield the details of trunking and the placement of internal nodes, and it should imply the analogue

$$\int \rho u \mathrm{d}u = 0$$

of the optimality condition (4.6).

There are also many points to be observed if one wishes to make the computational solution of (4.22) a practical proposition – matters we return to in Chapter 8 when we cover some of the techniques developed by Bendsøe and Sigmund (2003) in their remarkable structural optimisations. To make the computation practical, the continuum is approximated by an array of small cells: the 'finite element' method. The practicalities of actual construction materials set an upper limit on the density; external links must be compatible with these. For mathematical reasons one sets a small but positive lower bound on density. This makes it possible to evaluate the potential y at all points of the design space, and ensures that structures not initially envisaged can nevertheless come into contention. This minimal density is of course interpreted as zero in the actual final structure. In fact, values of density intermediate between the minimal and maximal extremes are often excluded, for reasons and by means which we shall consider in Chapters 5 and 8.

There are evolutionary methods based directly upon use of the dual approach. One can compute the potential fields $y(\xi, \omega)$ maximising the dual criterion for a given design and regime, and then add or subtract material at ξ according as $\gamma - \max_u \mathrm{E}\psi(\nabla y.u)$ is negative or positive at this value. Xie and Steven (1997) find this technique effective, as do other authors, but Bendsøe and Sigmund (2003) criticise it for not exploiting valuable gradient information; see Chapter 8.

Nevertheless, the calculation by either of these approaches remains one of very high dimension. The best hope of an economy seems to be that, if the emerging design is plainly discrete, then one should be able to move to a discrete but flexible description, whose evolutionary relations may be less elegant, but greatly reduced. We sketch such a reformulation in the next section.

4.8 Node migration and node splitting

The evolutionary policies described in the last section are blind and slow, while making computational demands of astronomical proportions. It is both their strength and their weakness that they forswear insight in favour of a prolonged grinding encounter with reality. An approach that is vastly more economical and insightful is to appeal to a dynamic version of Theorem 4.1. Briefly, if the number of external nodes is finite, then sooner or later the evolving network will reduce to one with a finite number of nodes linked by straight arcs. We can assume that link ratings adjust to variable flows at such a rate that relations (4.4) and (4.5) hold almost all the time. The number and positions of the internal nodes are then coaxed towards optimality by a dynamic version of condition (4.6).

We are thus envisaging in fact at least four different time scales. The fastest is that on which the cost-minimising flow $\hat{x} = x(\rho, \omega)$ is determined, for a given design (as specified by ρ) and regime ω, either by computation or by physical relaxation of the actual system. The next is that on which regimes change, slowly enough that $x = \hat{x}$ almost all the time. After a time that is doubtless considerable a discrete structure will emerge, if the external links are discrete. At this stage the structure ρ will reduce to specification of the positions ξ_j of nodes and of the ratings a_{jk} of links between them. The next time scale is that on which ratings a adapt to flows $\hat{x}$, in a regime-averaged sense. The slowest, and that with which we are now concerned, is that on which the node numbers and positions develop. It may be that these last three stages can be telescoped into one, as they are in the evolutionary algorithm of the last section, but the successive reductions become clearer if we distinguish them.

Consider the simple example of Figure 4.8, in which three source nodes 1, 2 and 3 feed a single sink node 0. Assume that only one source node is active at a time, so that trunking is encouraged. The naive solution is that drawn in (i), of direct connections. We suppose capacities of links adapted to flows by the rule (4.4) of the primal problem. One could now apply test (4.6). A first application at node 0 would show the benefit of a move to configuration (ii), with creation of an internal node 4. A second application at 4 would show the benefit of a move to configuration (iii), with creation of a second internal node 5. One will then need a final adjustment of the positions of nodes 4 and 5 if condition (4.6) is to be satisfied exactly at these nodes, and the net to be fully optimised.

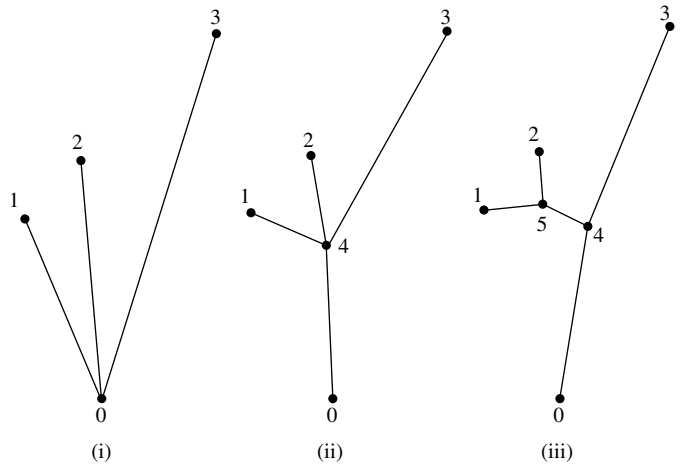


Fig. 4.8 Stages in the evolution of a trunked flow.

The optimal configuration was evident in this case; we need an objective and quantitative guide for cases when it is not. For repositioning of a given internal node j the obvious dynamic version of (4.6) is

$$\dot{\xi}_j = -\kappa \sum_k a_{jk} u_{jk}. \tag{4.23}$$

The expression $S = \sum_{j,k} a_{jk} d_{jk}$, now the effective criterion function, will then decrease at rate

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\kappa |\sum_k a_{jk} u_{jk}|^2.$$

Consider now the splitting of node j into a number of nodes that we shall label by i, and let J_i be the set of nodes previously attached to j that are now to be attached to i. We can then define

$$F_{ij} = \sum_{k \in J_i} a_{jk} u_{jk} \tag{4.24}$$

as the 'force' that the nodes of J_i exert on node j. We shall suppose that node j is initially fixed, in that it is either an external node or an internal node that has come to rest under the forces expressed by (4.23), so that $\sum_i F_{ij} = 0$. If the force F_{ij} is now to be exerted on the potentially detaching node i then we shall initially have an equation $\dot{\xi}_i = \kappa F_{ij}$. However, as soon as there is any detachment this becomes

$$\dot{\xi}_i = \kappa (F_{ij} - a_{ij} u_{ij}), \tag{4.25}$$

where u_{ij} is a unit vector in the direction of the initial displacement, and so in the direction of F_{ij}. The additional term in (4.25) represents the restoring force to the initial position ξ_j caused by the fact that the total flow from J_i must still be transferred back to the initial joint node at ξ_j. The magnitude of the force dragging node i from its initial position ξ_j is thus $|F_{ij}| - a_{ij}$, and, unless this is positive, node i will simply not detach itself from

this initial position. We see from (4.24) and this last observation that the splitting will reduce the criterion at rate

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\kappa \sum_i (|F_{ij}| - a_{ij})_+^2, \tag{4.26}$$

where $x_+ = \max(x, 0)$. The decomposition $\{J_i\}$ of the nodes initially attached to node j should be chosen to maximise the sum in (4.26).

Once the nodes i that can separate from the parent node j have done so, they will seek to optimise their positions by following the rule (4.23). This will continue until their positions equilibrate, when the splitting option can be tested again, and so on until no further trunking is indicated. The reverse of node splitting is of course node merging. If this is to occur during evolutionary optimisation we would expect it to do so as a natural consequence of the node migration equation (4.23).

We saw from Section 4.3 that the other phenomenon we should expect is that of load splitting. This would correspond to the creation of new arcs. The need for the creation of new arcs would be indicated by an excessive value of $\mathrm{E}\{\psi[(y_j - y_k)/d_{jk}]\}$ between nodes j and k of the current design. Alternatively, if one begins with a design with direct links between all external nodes, then one would expect to shed links rather than gain them.

Our use of the term 'forces' between nodes, prompted naturally by the dynamics of node movement under minimisation of the criterion S, might lead to confusion with the so-called 'gravity' models (see Rougham *et al.*, 2003 and Li *et al.*, 2004). There is in fact no relation. The gravity approach is a heuristic one, in which one assumes that the total traffic X_{jk} between source/destination pairs j and k is proportional to the product of demands at these two nodes, and then optimises routing by maximising a linear form in the X_{jk} subject to linear constraints in these variables.

4.9 The case of general convex ϕ

If we had retained a general convex ϕ then the cost component for the jk link would have been

$$\Phi(\hat{x}_{jk}, a_{jk}) = a_{jk}[\phi(\hat{x}_{jk}/a_{jk}) + \gamma].$$

For a manageable analysis one would like the expectation of this function to be a function of a_{jk} and of the expectation of some function of $\hat{x}_{jk}$. This will be the case if the function ϕ is homogeneous of some degree; we specialised to the particular case of a power law in this chapter. If, however, we persist with the general case, we do have the mild satisfaction of deriving a generalised Minkowski inequality very quickly.

Optimality demands that a_{jk} should have the value minimising the expression $\mathrm{E}[\Phi(\hat{x}_{jk}, a_{jk})]$. If this minimal value is denoted M_{jk} then an internal node j must satisfy the condition

$$\sum_k M_{jk} u_{jk} = 0$$

if its position is to be optimal.

Now consider again the example of Section 4.2, in which a number of random flows x_j were directed towards a distant node 0. The condition for these flows to be trunked in the optimal solution is that the relation

$$\sum_j \min_{a_j} \mathrm{E}[\Phi(\hat{x}_j, a_j)] \geq \min_a \mathrm{E}[\Phi(\sum_j \hat{x}_j, a)] \tag{4.27}$$

should hold with a sufficiently large margin of inequality. Inequality (4.27), if valid, would be a generalised version of Minkowski's inequality (4.8). It is indeed valid, because convexity of ϕ implies that

$$\sum_j a_j \phi(\hat{x}_j / a_j) \geq (\sum_j a_j) \phi[(\sum_j \hat{x}_j)/(\sum_j a_j)]$$

and hence that

$$\sum_j \mathrm{E}[\Phi(\hat{x}_j, a_j)] \geq \mathrm{E}[\Phi(\sum_j \hat{x}_j, \sum_j a_j)] \geq \min_a \mathrm{E}[\Phi(\sum_j \hat{x}_j, a)],$$

which implies (4.27). Equality again holds in (4.27) if and only if all the flows $\hat{x}_j$ are multiples of a common scalar random variable.

5

Concave costs and hierarchical structure

5.1 Incentives to trunking

The reason why trunking can be desirable in the variable-load case is because effective flows are in a sense sub-additive. To take the example of Figure 4.1, the effective flow rates out of nodes 1 and 2 are each $2^{-1/\alpha}$, but the effective rate of unity out of the amalgamating node 3 is less than the sum of these. The same effect can be produced by a number of causes. Suppose that the flow cost per unit length of a link of rating a carrying a fixed flow x is $c(x, a)$ and that the installation cost (expressed as a lease rate) per unit length is $\gamma(a)$. Then the minimal total cost per unit length of carrying the flow is

$$\chi(x) = \min_a [c(x, a) + \gamma(a)]. \tag{5.1}$$

Then, even in the case of a single load pattern, trunking can be advantageous if this function is sub-additive, in that

$$\chi(x_1 + x_2) \leq \chi(x_1) + \chi(x_2), \tag{5.2}$$

for nonnegative x_1 and x_2, with strict inequality somewhere. One might expect this. If χ is also continuous then standard arguments demonstrate that $\chi(x)/x$ is nonincreasing in x. (Note that the function $\chi(x)/x$ has a meaning; in the road traffic context it is the cost of flow plus installation per unit time and also per motorist.) This implies that, if one replaces a flow of x by s parallel flows of rate x/s, then one incurs a cost of $s\chi(x/s)$. As a function of s this is nondecreasing; if it is strictly increasing then division of the link into parallel fibres is actually costly. That is, there is an incentive to realise the link by a single conductor, which can go so far as to encourage trunking.

To see that it can do so, consider an internal node j. If this is to be optimally placed then the argument of Theorem 4.1 now implies the analogue of condition (4.6)

$$\sum_k \chi(x_{jk}) u_{jk} = 0. \tag{5.3}$$

Here u_{jk} is again the unit vector in the direction of the directed arc jk. Consider again the three-node example of Figure 4.1, with actual flows 1, 1 and 2 on links 10, 20 and 30. Then condition (5.3) requires that the angle θ of the figure should satisfy

$$2\chi(1)\cos\theta = \chi(2).$$

This has a solution for θ, by (5.2), and the solution will be positive if the strict inequality $\chi(2) < 2\chi(1)$ holds. There will then be trunking if the geometry of the fixed nodes allows this angle to be realised.

Concavity of χ implies sub-additivity if also $\chi(0) = 0$. To see this, note that if $\chi(x)$ is concave its derivative satisfies $\chi'(x_1 + u) \leq \chi'(u)$ for $u \geq 0$. Integration of this inequality with respect to u from 0 to x_2 yields

$$\chi(x_1 + x_2) + \chi(0) \leq \chi(x_1) + \chi(x_2),$$

which implies sub-additivity if indeed $\chi(0) = 0$.

We shall indeed have cause to consider the class $\mathcal{Q}$ of functions $\chi(x)$ defined on the nonnegative half-line $x \geq 0$ that are nondecreasing, concave and for which $\chi(0) = 0$. Standard arguments lead to the following assertions.

Theorem 5.1 *(i) If $\chi \in \mathcal{Q}$ then $\chi(x)/x$ is nonincreasing in x. (ii) If χ_1 and χ_2 belong to $\mathcal{Q}$ then so does the compounded function $\chi_2\chi_1$; i.e. the function $\chi_2(\chi_1(x))$.*

A strict version of (5.2) implies that we have sub-additivity for actual flows rather than for 'effective' flows, so that the trunking option remains attractive whatever the statistical character of the flows. That is, trunking will not cease at a given stage, as was the case for the variable flows of Section 4.2, but will continue as long as is geometrically possible.

The traffic models of Chapter 6 demonstrate the possibility of congestion, which can induce concavity of χ. However, a more serious factor, and an almost universal one, is that raised in Chapter 1 and emphasised in Section 3.1. This is, that the naive optimal design deduced in Chapter 1 does not allow for the blighting effects of routes on their immediate environment, a blight which must be checked by an explicit penalisation of environmental intrusion. One way to achieve this would be to choose $\gamma(a)$ as a sum of a base cost and a proportional cost:

$$\gamma(a) = c_0\delta(a) + \gamma a. \tag{5.4}$$

Here $\delta(a)$ is zero or unity according as a is zero or positive, and the coefficient γ in the right-hand member is constant. More generally, we can consider the case $\gamma(a) \in \mathcal{Q}$.

Theorem 5.2 *Suppose seminvariance of flow costs, in that $c(x, a)$ has the form $a\phi(x/a)$ where ϕ belongs to the class $\mathcal{C}_s$ of strictly convex functions defined in Section 1.2. Then, if the function $\gamma(a)$ lies in $\mathcal{Q}$, so also does the function $\chi(x)$ defined in (5.1).*

Proof The assumptions on $\gamma(a)$ imply that it has a representation

$$\gamma(a) = \min_j(\beta_j + \gamma_j a), \tag{5.5}$$

where the γ_j are positive and $\min_j \beta_j = 0$. Inverting the order of the minimisations in (5.1) we have then

$$\chi(x) = \min_j \min_a [a\phi(x/a) + \beta_j + \gamma_j a] = \min_j [\beta_j + q_j x], \tag{5.6}$$

where q_j is the positive root of $\psi(q) = \gamma_j$ (see Lemma 1.1). We see from (5.6) that χ indeed has the properties asserted. $\diamondsuit$

So, the effect of a concave costing of route rating is to convert a seminvariant convex flow cost $c(x)$ to a concave one, $\chi(x)$, once rating has been optimised for a given flow. Note that the special case (5.4) can be seen as a limiting version of (5.5):

$$\gamma(a) = \lim_{\theta\uparrow\infty} \min[\theta a, c_o + \gamma a]. \tag{5.7}$$

Corollary 5.3 *Relation (5.4) has the consequence*

$$\chi(x) = c_0\delta(x) + qx = c_0\delta(x) + \psi^{-1}(\gamma)x. \tag{5.8}$$

Proof Relation (5.8) follows formally from (5.6) and (5.7). However, a direct proof also follows easily, with appeal to the fact that the expression $a\phi(x/a)$ is either zero or tends to $+\infty$ as a tends to zero, according as x is zero or not. $\Diamond$

Another example is provided if we take $\gamma(a)$ of the power-law form: a^μ with $0 < \mu < 1$. For definiteness, let us take $\phi(x)$ also of the power-law form $|x|^\alpha$. One finds then that

$$\chi(x) \propto |x|^\nu,$$

where $\nu = \alpha(1-\mu)/(\alpha-\mu)$, which also lies between 0 and 1. This clearly has the properties asserted in Theorem 5.1.

5.2 Other consequences of cost concavity

The fact that the optimisation of ratings can convert a convex flow cost $c_{jk}(x)$ into a concave version $d_{jk}\chi(x)$ has profound implications. At least if one wants to perform the optimisation in a single pass, Lagrangian methods are now no longer applicable, and the notions of a potential and its associated dual extremal problem, which seemed so powerful, are now quite lost. The principle that must then replace the Lagrangian approach is: if one is minimising a concave function over a convex domain then this minimum will be achieved at one of the extreme points of the domain.

As far as flow cost minimisation is concerned, it is a classic result that, if one is considering the scheduling of a single commodity flow on a given network with concave arc-flow costs, then there is always an optimal flow that is confined to a spanning tree of the network (see Ahuja *et al.*, 1993). The spanning tree assignments are indeed the extreme points of the set of possible flows. When we add also the minimisation of design costs, then the network itself will become a tree, or collection of trees, which is just another way of characterising trunking.

The spanning-tree assertion for the optimal flow pattern follows by a standard argument which, however, needs elaboration when there are several types of traffic. That is, when there are several commodities to be distributed, or when traffic is labelled by its desired destination. The most realistic course in this more general case is to assume that arcs are directed, so that all flows on a given arc are in the same direction, and that the flow cost is a function $\chi_{jk}(x_{jk})$ only of *total* flow x_{jk} along the arc. Normally one will

have $\chi_{jk}(x) = d_{jk}\chi(x)$ for a universal function χ, but we can equally well consider the more general case.

Theorem 5.4 *The optimal net can contain no cycle for which some type of traffic has nonzero flow on all arcs.*

Proof Let us for simplicity label the arcs of the cycle by j. The cost of the flow on the links of the cycle is then $\sum_j \chi_j(x_j)$, where x_j is the total flow along arc j. By hypothesis there is some traffic component that contributes to flow for every j. We can perturb this by modifying the flow of this component in a clockwise direction (say) around the cycle by ϵ for every j without affecting flows on arcs outside the cycle. The cost will then be revised to $\sum_j \chi_j(x_j + \sigma_j\epsilon)$, provided ϵ is not large enough that the flow of the component is reversed on some arc. Here σ_j is +1 or −1 according as arc j is directed clockwise or anticlockwise. The perturbation in cost for small ϵ is then

$$\delta C = \sum_j \chi_j'(x_j)\sigma_j\epsilon. \tag{5.9}$$

If all σ_j are positive (negative) then costs will decrease if we choose ϵ negative (positive). One can then continue with this decrease of flow along the arcs until the component has zero flow on some arc.

In the case when the σ_j vary in sign (i.e. the arcs are not all directed the same way around the cycle) let us assume for concreteness that the coefficient of ϵ in expression (5.9) is nonnegative. We should then choose the perturbation ϵ negative if costs are not to increase. But the coefficient of the next perturbation will remain nonnegative if we thus perturb x, because concavity will imply that $\chi_j'(x_j + \sigma_j\epsilon)$ is greater or less than $\chi_j'(x_j)$ according as σ_j is positive or negative. One can continue in this way to reduce flow in a clockwise direction without increasing costs until, again, the component has zero flow on some arc. ◊

In the case of a single commodity the conclusion is that cycles need not (and usually cannot) exist in the optimal design, so that the graph of actual arcs in use is a collection of trees. This will become a single spanning tree under nondegeneracy assumptions. For the case of several types of traffic there are also immediate consequences.

Corollary 5.5 *For destination-specific traffic there need be no more than one route between source and destination in the optimal design.*

Proof Use of two routes would produce a cycle carrying this type of traffic on both routes. ◊

Corollary 5.6 *If two types of traffic share part of a route then this shared part can be unbroken in the optimal design.*

Proof That is, in the optimal design one can exclude the case in which two types of traffic share the route for a while, then part, and then meet again. For, if it were advantageous for one type to vary the route, then it would be advantageous for the second type to make the same variation, by the argument of Theorem 5.4. ◊

There can of course be cycles, in the sense that multiple routes can exist between two nodes: e.g. routes carrying different commodities in one direction or the other. The point is, that one cannot have a given component of traffic being carried on all arcs of a cycle.

We now have a picture of the optimal network, in the case of destination-specific traffic, in which a given traffic component follows a single route from source to destination. It may share part of this route with other traffic components, but the section that it shares with any other given component is unbroken. To this extent the destination-specific case demonstrates the tree properties of the single-commodity case. We would now wish to have more of a feeling for the detailed form of this route sharing.

The view we shall develop is that concave costs essentially imply a hierarchical structure. However, before that, there is another aspect to be kept in mind.

5.3 The combination of environmental penalty and variable load

We have seen that environmental considerations and variable load both induce trunking in the optimal design, the first in a more radical fashion. Both effects are practically important, however, so their combination has to be considered. Perception of pattern in this more complicated case is hard to come by, but we shall at least see how the two effects can be jointly incorporated in evolutionary calculations.

For definiteness we shall keep the telephone application in mind, which means that we are thinking of destination-specific traffic, perhaps decomposable into several 'commercial classes': call categories of differing priority or profitability. The model is then a development of that of Chapter 3, and will be taken a little further in Chapter 15.

We start with a discrete net. The destination node will again be labelled by i, and the commercial class by v, so that we can speak of iv-traffic. Let f_{hiv} be the input rate of this traffic at node h, a random variable if load is variable. Let x_{jkiv} be the flow rate of iv-traffic on the jk-link. The total traffic on this link is then

$$x_{jk} = \sum_i \sum_v x_{jkiv} \tag{5.10}$$

if we assume traffic measured in such units that the flow rates can simply be added. We then have the balance relation

$$f_{hiv} + \sum_j (x_{jhiv} - x_{hjiv}) = 0 \qquad (h \neq i) \tag{5.11}$$

at node h. The cost of flow and materials is

$$C(x, a) = \sum_j \sum_k d_{jk}[a_{jk}\phi(x_{jk}/a_{jk}) + \gamma(a_{jk})], \tag{5.12}$$

where $\gamma(a)$ is a concave function measuring the leasing and environmental cost per unit length of a link of capacity a. The extended Lagrangian in a given regime (i.e. for a given value of f) will then be

$$L(x, y, z, a) = C(x, a) + \sum_j \sum_k z_{jk}(x_{jk} - \sum_i \sum_v x_{jkiv}) + \sum_h \sum_i \sum_v y_{hiv}[f_{hiv} + \sum_j (x_{jhiv} - x_{hjiv})], \tag{5.13}$$

where z_{jk} and y_{hiv} are the multipliers for constraints (5.10) and (5.11) respectively. Lagrangian methods are applicable for fixed a, since the problem is then one of convex costs. For the moment we shall in fact bypass these methods, but the values of the multipliers will become relevant in Chapter 15.

Let us now assume, as in Chapter 4, that the optimal flow $\hat{x}$ has been determined for a given design and regime, the randomness of regime making $\hat{x}$ a random variable. The expression to be minimised with respect to the design variables is then

$$\mathrm{E}[C(\hat{x}, a)] = \sum_j \sum_k d_{jk}[a_{jk}\mathrm{E}\phi(x_{jk}/a_{jk}) + \gamma(a_{jk})]. \tag{5.14}$$

For explicitness, let us choose the familiar cost functions

$$\phi(p) = p^\alpha/\alpha, \qquad \gamma(a) = \gamma a^\mu, \tag{5.15}$$

where $\alpha > 1, 0 < \mu < 1$ and the final γ is a constant coefficient. The proof of Theorem 4.1 can then be adapted to yield

Theorem 5.7 *Under the assumptions above, the criterion of minimal expected cost implies the following assertions:*

(i) The optimal value of a_{jk} must satisfy

$$a_{jk}^{\alpha+\mu-1} = \frac{1}{\beta\gamma\mu}\mathrm{E}(\hat{x}_{jk}^\alpha). \tag{5.16}$$

(ii) Its substitution leads to the reduction

$$\min_a \mathrm{E}[C(\hat{x}, a)] = \theta \sum_j \sum_k a_{jk}^\mu d_{jk} = \theta S, \tag{5.17}$$

say. Here $\theta = [(\beta - 1)\mu + 1]\gamma$ and β is the value conjugate to α.

(iii) If node j is an internal node of the system whose position in physical space may be optimised, and if the ratings a_{jk} have been determined by (5.16), then a necessary condition for position optimality is

$$\sum_k [a_{jk}^\mu - a_{kj}^\mu]u_{jk} = 0, \tag{5.18}$$

where u_{jk} is the unit vector in the direction of $\xi_k - \xi_j$.

The whole pattern of trunking should be latent in this last assertion. However, although conditions (5.16) and (5.18) are interesting, they are not really practical, because the expectation $\mathrm{E}[\hat{x}_{jk}^{\alpha}]$ is itself a highly implicit function of the capacities a. It is through evolutionary recursions that these dependences can play themselves out. Under the assumptions above the continuous analogue of $C(\hat{x}, a)$ is

$$C(\hat{x}, \rho) = \int\int [\rho^{1-\alpha}\hat{x}^{\alpha}/\alpha + \gamma\rho^{\nu}]\mathrm{d}\xi\mathrm{d}u. \tag{5.19}$$

Here both $\hat{x}$ and ρ are functions of position ξ and orientation u. Note that $\hat{x}$ is now nonnegative, because we are assuming directed links. It is the oriented spatial density ρ that specifies the design. The evolutionary equation (2.17), modified to (4.22) when loading became variable, now becomes

$$\dot{\rho} = \kappa[\mathrm{E}(\hat{x}/\rho)^{\alpha} - \beta\gamma\mu\rho^{\mu-1}]. \tag{5.20}$$

Equation (5.20) contains two features that induce trunking. Firstly, the expectation term will encourage any pooling of flows that decreases variability. Secondly, the fact that ρ occurs to a negative power in the last term of the bracket introduces what is perhaps the subtler and more radical effect: that of environmental penalty. Its effect will be that small values of ρ are rapidly reduced to zero, in the absence of countervailing effects. This encourages trunking in discrete designs and a black-and-white character in continuous ones, and explains why the Bendsøe–Sigmund stratagem to achieve the latter effect works.

Just as in Chapter 4, if the set of external nodes is discrete, then one would expect the optimal design to be so also. The evolutionary formalism for a discrete design developed in Section 4.8 will carry over to this more general case, with the reduced criterion function S now determined from (5.17).

5.4 Hierarchical structure

The compound tree structure that environmental considerations induce in the optimal design gives one something of a purchase on the problem, in that it corresponds to a hierarchical structure. Telephone networks give the obvious example. Calls going in somewhat the same direction will, under the influence of concave costs, be gradually merged together to a common trunk. That is, they will be taken from the tips of a tree to its root. At the other end they will traverse a tree in the opposite direction, being distributed to its tips by a successive branching. The internal nodes induced by an algorithm such as (5.20) will just be the exchanges of the net, determined in size and site by the underlying optimality criterion, even if this will be complicated by a dozen other factors in any practical case.

The successive nodes at the departure end can be seen as ascending stages in a hierarchy; stages which would in practice be fairly clearly quantised. So, local calls are gathered together to a local exchange, which one might term a level-one exchange. These will in turn feed level-two exchanges, and so on. A call from a given source will then progress up the hierarchy until it reaches a level at which the exchange has downward communication to the required destination. The call will then follow this downward route until it completes connection.

If the network is serving a finite area then there will be a natural upper limit to the hierarchy, with perhaps a single exchange at this highest level. Equally plausibly, the costs at this level may be such that it is more natural to crown the system by a cabinet than by a king. That is, to have several exchanges at the highest level, linked so as to constitute a connected set. When the current United Kingdom telephone network was planned, it was envisaged that there would be three levels of exchanges. These would consist of some 31,000 *line concentrators* at level one, some 750 *digital local exchanges* at level two and some 70 *digital main switching units* at level three, with complete interconnection at this top level. In fact, spatial heterogeneity and provision of special services have required some elaboration of this basic scheme, so that there are now six levels with some departure from strict hierarchy at the top. We return to the question of how the hierarchy should be terminated in the absence of such complications in Sections 5.6 and 5.7.

We are asserting that a system that trunks in reaction to concave costs becomes essentially hierarchical. One gropes for a more exact statement and proof of this assertion. Certainly, if the source and sink sets of the system are well nucleated, then the optimal system connecting them will condense down to a clear set of trunk routes, at various levels, without cycles, and this seems to amount to a loose hierarchy. The best course is perhaps to consider systems in which this nucleation is *not* so clear. Consider, for example, the situation of Figure 5.1, in which the source/sink nodes cluster closely and somewhat randomly around a linear locus, in a uniform fashion along its length. Then a trunk route will certainly develop along this locus, but the feeder system connecting the nodes to the trunk will tend to be fairly direct, and show little amalgamation. The system is indeed hierarchical, but in a rather dispersed sense: the trunk route defines the system, but as a chain of small exchanges rather than as one master exchange. In the case of a road network, at least, one would encourage amalgamation of the feeders by restricting the number of access points to the trunk; i.e. by setting a cost on access nodes.

Let us nevertheless develop the idea of a hierarchical network; the telephone example is a natural one to take for definiteness. We shall assume that each link of the network consists of two directed links of the same rating. One of these serves ‘up’ traffic, the other ‘down’ traffic, and we shall assume symmetry, in that these two traffic streams have the same rate. Furthermore, we are trying to perceive structure, and so assume for simplicity that all traffic is carried by land-lines, thus abjuring the technological advantages of, for example, radio or satellite transmission. Traffic will be labelled by its destination electronically, and so can be carried on a common conductor and appropriately routed at the exchanges.

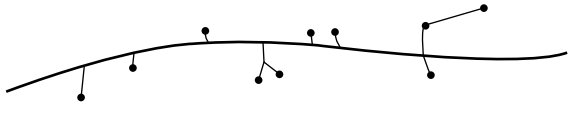


Fig. 5.1 A case in which the hierarchical structure is weak: that of nodes distributed somewhat randomly and uniformly in the neighbourhood of a linear locus.

One can scarcely obtain analytic results for an arbitrary geographical distribution of callers. We shall then go to the other extreme, and assume an isotropic system. That is, one in which the statistical characterisation of the system is invariant under translations or rotations of physical space, this being the space in which the callers are distributed. Such assumptions are most unrealistic, but offer the best chance of some insight. In view of the observations above, however, it is interesting to note that the optimal hierarchy turns out to be a continuous one, in that there is a continuous gradation rather than a quantisation of levels.

One might make a passing observation on the general case. Human settlements have developed historically, by growth around given nuclei of population (the risk-averse option) and by the quest for favourable sites for new nuclei (the risk-seeking option). At every stage, service points that play the role of exchanges in communication and commerce have come into existence in a semi-heuristic, semi-planned fashion. How close is the consequent pattern to optimal, in that it is optimally adapted to the final population distribution? Otherwise expressed, do the adaptive algorithms considered in Chapters 2 and 8 reflect the adaptive mechanisms of real life?

5.5 The outflow function

We shall continue with the telephone example, for definiteness. Callers are supposed uniformly distributed in a subset $\mathcal{P}$ of physical space. We earlier used the symbols p and q as specimen arguments of the functions ϕ and ψ. We shall in fact not need to do so in the future, so it is very convenient notationally if we can henceforth use p to denote the dimension of the Euclidean space $\mathbf{R}^p$ within which $\mathcal{P}$ lies, and q for a related purpose. We use ξ and η to denote the vector co-ordinates of two specimen points within this space.

Calls from source ξ for destination η are assumed to occur at rate $\rho(|\xi-\eta|)$, in that $\rho(|\xi-\eta|)\mathrm{d}\xi\mathrm{d}\eta$ is the expected number of calls per unit time from an infinitesimal volume element $\mathrm{d}\xi$ centred on ξ to a similar element $\mathrm{d}\eta$. These assumptions imply isotropy, in that they imply invariance of the model under rigid translations or rotations. Such assumptions are of course unrealistic, but the model nevertheless enables one to identify features which may persist in less idealised circumstances.

We shall make the assumption that the integral

$$\int \rho(|\xi|)\mathrm{d}\xi \propto \int_0^{+\infty} \rho(r) r^{p-1}\mathrm{d}r$$

is finite, reflecting the finite density of generated traffic. In fact, we are compelled to the stronger assumption: that

$$\int \rho(|\xi|)|\xi|\mathrm{d}\xi \propto \int_0^{+\infty} \rho(r) r^{p}\mathrm{d}r < \infty. \tag{5.21}$$

This reflects the assumption that density of cost incurred is also finite, based on a naive direct point-to-point routing. Condition (5.21) sets a lower bound on the decay rate of the function $\rho(r)$ for large r. If it decays as a power law, r^{-q}, then necessarily $q > p+1$.

The first stage of aggregation, in which subscribers are connected to a local exchange (the level-one exchange, or line concentrator), is exceptional, in that the subscribers' lines are active only very intermittently. This connection then collects a considerable number of weak and statistically variable flows to produce a steady aggregate flow. Flows from this point can be regarded as steady and deterministic (until, of course, the reverse stage of connection to the destination).

We shall then take the array of level-one exchanges as the starting point, and shall assume that they constitute a regular lattice in $\mathbf{R}^p$. The simplest course is to assume that they constitute a cubic lattice so that, in the usual case of two dimensions, each exchange is at the centre of its square catchment area, its 'cell'. Hexagonal cells might seem preferable, in that the mean distance from caller to exchange is then less for a cell of given area. However, when we come to aggregate cells to form a level-two cell, then the cubic lattice has advantages. Briefly, if we aggregate several level-one cells to form a square level-two cell, as in Figure 5.2, we can do so without leaving any of the level-one exchanges on the boundary of the new cell. If we try to work with hexagonal cells, then this is not possible.

We suppose then that the cells are the cubes (actually hypercubes, but we shall use the familiar term) of a cubic lattice, with exchanges placed at their centres. We take N and n as being scale factors, so that (again using the three-dimensional terms in a p-dimensional setting) a cube of scale N has volume N^p, contains N^p level-one exchanges, and has a surface area of $2pN^{p-1}$.

Suppose that we take the edge of the level-one cell as the unit of distance, and form the cubic level-two cells from the cube of side n_1, containing n_1^p level-one cells. Thus n_1 measures the 'scale' at level two. The scale of level-t cells will then be $N_t = n_1 n_2 ... n_{t-1}$. The design problem is then to choose the successive scaling factors n_t optimally, and also to terminate the hierarchy optimally. At least, optimally for the regular geometry we are imposing; the truly optimal structure will not be so regular. (If a level-two exchange is placed at the centre of a cell of scale n, as in Figure 5.2, then true optimality would demand that all the n^p level-one exchanges should be displaced towards this centre. It would then also demand that the shapes of the level-one cells should be modified, so destroying the regular geometry.)

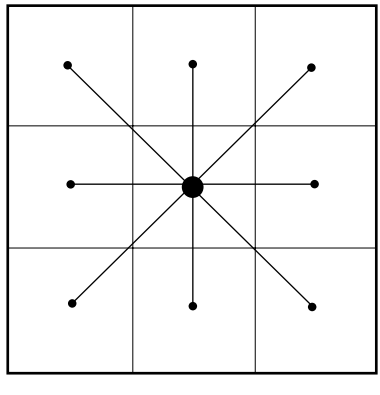


Fig. 5.2 The aggregation of nine level-one cells to form a 3×3 level-two cell.

Let us denote a cell, of unspecified scale, by Γ. Then a significant quantity is the *outflow* from Γ:

$$x = \int_{\xi\in\Gamma} \mathrm{d}\xi \int_{\eta\notin\Gamma} \mathrm{d}\eta\rho(|\xi-\eta|). \tag{5.22}$$

This is just the rate of calls from subscribers in Γ to destinations outside Γ. Let us denote a cell of scale N by Γ_N, and the corresponding outflow rate by $x(N)$. Then $x(N)$ is an important quantity, which we shall term the *outflow function*. It is the flow that must be carried from an exchange associated with a cell of scale N to the next exchange in the hierarchy. The costs incurred at this stage will be a function just of $x(N)$ and the total length of the links making the connection to the next level. However, at the final stage we do have to consider how the outflow is distributed, so it is useful to define $x_j(N)$ as the flow from one N-cell to another displaced from it by a vector amount labelled by j.

Despite the fact that the succession of levels that we have considered will always produce a cell whose scale N is integer-valued, we can very well define a cell of any scale, and so consider $x(N)$ as a function of a continuous variable. We can develop some feeling for the behaviour of $x(N)$ by beginning with the one-dimensional case in this and the next section. The one-dimensional case is degenerate, because all the inter-level links lie along the same line, and could be realised by a single conductor carrying a flow of $\sum_j |j|x_j(N)$, where j runs through the signed integers. However, let us assume flows carried in dedicated conductors, so that the picture is consistent with that of higher-dimensional cases. The function $\rho(r)$ must be even (reflection being the nearest thing to a rotation in one dimension). Define the function

$$M(r) = 2\int_r^{+\infty} \rho(|\eta|)\mathrm{d}\eta, \tag{5.23}$$

the rate of flow from a source point to destinations distant r or more. Then

$$x(N) = 2\int_0^N \mathrm{d}\xi \int_N^{+\infty} \mathrm{d}\eta\rho(|\eta-\xi|),$$

leading to the alternative expressions

$$x(N) = \int_0^N M(r)\mathrm{d}r = 2\int_0^{+\infty} \min(N, r)\rho(r)\mathrm{d}r. \tag{5.24}$$

Since $M(r)$ is nonincreasing (5.24) implies:

Lemma 5.8 *In the one-dimensional case $p = 1$ the function $x(N)$ belongs to $\mathcal{Q}$ and has the limit value*

$$x(\infty) = 2\int_0^{+\infty} r\rho(r)\mathrm{d}r. \tag{5.25}$$

The outflow must necessarily increase with N, because callers are being added, and even the numbers within a given distance of the cell boundary are increasing. The limit (5.25) is finite because the only subscribers making a substantial contribution to outflow when N is large are those relatively close to the boundary. The limit thus reflects the

number of such subscribers. It is a symptom of what seems to be an unavoidable anomaly: that however large one makes the cells (i.e. however high one goes in the hierarchy), there will be source–destination pairs that are close, but can communicate only via an exchange at this level. More generally, one will have an asymptotic term proportional to the surface area of Γ_N, and so to N^{p-1}.

Consider, for example, the exponential case, $\rho(r) \propto \exp(-\alpha r)$. Then $M(r) \propto \exp(-\alpha r)$, and

$$x(N) \propto (1 - \mathrm{e}^{-\alpha N}).$$

Contrast this with the case of an inverse power dependence: $\rho(r) \propto |r|^{-q}$ for $|r| \geq h$, and zero otherwise. Condition (5.21) implies that $q > 2$, and one finds indeed that

$$x(N) \propto \frac{h^{2-q}}{q-2} - \frac{N^{2-q}}{(q-1)(q-2)}.$$

5.6 Optimisation of the trunking rate

Full determination of the structure requires determination of the aggregation factors n_t at each stage t and determination of when and how the system should be capped. That is, at what stage should aggregation be halted and what system of final interconnections should be installed at that stage. It is convenient to term $u_t = \log n_t$ the *trunking rate* at stage t.

We shall again begin with consideration of the one-dimensional case. Suppose that the system has reached a scale of N, with a corresponding outflow rate per N-cell of $x(N)$. Define $c(N) = \chi(x(N))$; this is the cost per unit length of line of carrying the outflow, rating a having been optimised. Suppose we trunk by a factor of $n = 2m+1$ at this stage; it is convenient to assume n odd. Then the total cost per unit length of achieving this trunking is

$$(Nn)^{-1} c(N) \sum_{j=-m}^{m} N|j| = \frac{m(m+1)}{2m+1} c(N) \approx (n-1)c(N)/4.$$

We divide through by Nn because each exchange at the next level draws from Nn level-one exchanges, and the factor N occurs in the first sum because this is the distance between adjacent exchanges at the level that is now being trunked. We have assumed n large enough that the final approximation is tolerable, but have in any case written $n-1$ rather than n, since this connection cost does reduce to zero if there is no aggregation, i.e. if $n = 1$. The total cost of trunking from level one up to a termination at level h is then approximately (and to within some numerical factor)

$$C = \sum_{t=1}^{h-1} (n_t - 1) c(N_t) + K(N_h), \tag{5.26}$$

where $N_0 = 1$ and $N_t = n_1 n_2 \ldots n_{t-1}$. The final term $K(N_h)$ denotes the termination cost per level-one exchange, which has the general form

$$K(N) = \sum_j |j| \chi(x_j(N)). \tag{5.27}$$

It is permissible to write this as a function purely of N, because the individual flows $x_j(N)$ are determined by N. The distance between N-cells separated by an amount j is $N|j|$, but we also divide by N to reduce the cost to a rate per unit volume.

To form an idea of the form of $K(N)$, consider again the case $\rho(r) \propto \exp(-\alpha r)$. We find that

$$x_j(N) \propto \left[\frac{1-\theta^2}{\alpha}\right]^2 \theta^{2(j-1)},$$

where $\theta = \exp(-n\alpha/2)$. Suppose now that $\chi(x) = \sqrt{x}$. Then expression (5.27) becomes

$$K(N) \propto \frac{1-\theta^2}{\alpha\theta}\frac{\theta}{(1-\theta)^2} = \frac{1+\theta}{\alpha(1-\theta)} = \frac{1+\mathrm{e}^{-N\alpha/2}}{\alpha(1-\mathrm{e}^{-N\alpha/2})}. \tag{5.28}$$

Expression (5.28) decreases convexly with increasing N, and then approaches a positive limit, as illustrated in Figure 5.3(ii). The decrease is because increased trunking brings increased economy. The convergence to a positive limit is because of the convergence of outflow to a finite limit with increasing N; the limit $K(\infty)$ then represents the connection cost for sender/receiver pairs who are physically close but fall into different catchment areas. The arguments of Appendix 1 indicate that, under natural hypotheses, $K(N)$ will have this convexly decreasing form, even in the multi-dimensional case.

We are now interested in minimising expression (5.26) with respect to the scaling factors n_t and the number of stages h. The problem is one immediately expressible as a dynamic programme. Let $G(N)$ denote the minimal operating cost for a system whose first level of exchanges serves N-cells. This is what we would term the *value function*; it obeys the dynamic programming equation

$$G(N) = \min\{\min_n[(n-1)c(N) + G(nN)], K(N)\}, \tag{5.29}$$

where the minimisation operations determine the optimal action. The first minimum is over the choice of continuation or termination: i.e. of adding at least one more layer of exchanges or of setting up a terminal system of links between the existing exchanges. The second minimisation is over the scale factor n to be adopted at the next stage if one

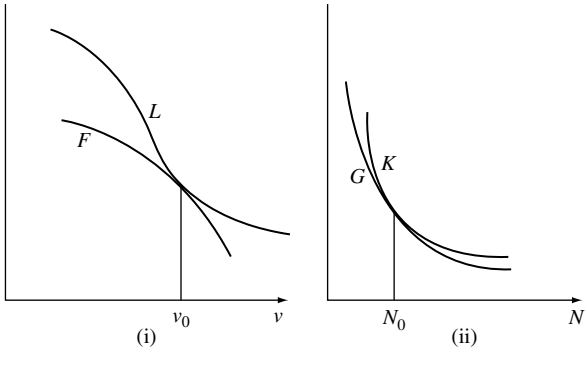


Fig. 5.3 (i) The determination of the optimal termination level $v_0 = \ln N_0$. (ii) The equivalent diagram in terms of the scale variable N.

does indeed continue. This formulation then incorporates an optimisation over both the number of exchange levels and the scalings between them.

The quantities n and N are both positive integers. It is mathematically advantageous to change to new variables $u = \ln n$ and $v = \ln N$, and to assume that these can take any real value. In so doing we make a continuous approximation to the discrete problem. If we write $c(N)$, $K(N)$ and $G(N)$ as $g(v)$, $L(v)$ and $F(v)$ in terms of the new variables then the dynamic programming equation (5.18) takes the form

$$F(v) = \min\{\min_{u \geq 0}[(e^u - 1)g(v) + F(v+u)], L(v)\}. \tag{5.30}$$

We shall in fact find solution of this equation and of the problem gratifyingly simple, thanks to the following observation.

Lemma 5.9 *The function $c(N)$ of the continuous variable N belongs to $\mathcal{Q}$, with the consequence that*

$$c(nN) \leq nc(N). \tag{5.31}$$

Proof The function $\chi(x)$ belongs to $\mathcal{Q}$, by hypothesis, and so does the function $x(N)$, by Lemma 5.8. By Theorem 5.1, the function $c(N) = \chi[x(N)]$ then enjoys both the properties asserted above. $\Diamond$

Theorem 5.10 *Suppose the function $L(v)$ monotone nonincreasing. Then:*

(i) Termination will occur as soon as $v \geq v_0$, where v_0 is the solution of the equation

$$g(v) + L'(v) = 0. \tag{5.32}$$

(ii) The solution of the dynamic programming equation (5.30) is

$$F(v) = \int_v^{v_0} g(z)\mathrm{d}z + L(v_0). \tag{5.33}$$

This corresponds to the optimal policy of a continuous trunking at a constant scaling rate until the termination level v_0 is reached.

Proof (i) This is a case in which the optimal stopping value of v is determined by the 'one-step look-ahead' or OSLA rule; as the value for which the cost $L(v)$ of stopping immediately just equals the cost $L(v + \delta u) + g(v)\delta u$ of taking a small further scaling increment δu and then stopping. (See e.g. Whittle, 1982 or Bertsekas, 1987.) Equation of these two expressions yields the equation (5.32), determining the optimal stopping point. (ii) In the continuation region $v < v_0$ equation (5.30) reduces to

$$F(v) = \min_{u \geq 0}[(e^u - 1)g(v) + F(v+u)]. \tag{5.34}$$

Consider the cost of moving from the value v to the terminal value v_0 in a single step u of size $2\Delta = v_0 - v$ with that of taking two steps each of size $u = \Delta$. The second course will be less costly if

$$(e^{2\Delta} - 1)g(v) > (e^{\Delta} - 1)[g(v) + g(v+\Delta)].$$

This reduces to the inequality

$$g(v+\Delta) \leq e^{\Delta} g(v),$$

which is just inequality (5.31), with $n = \exp\Delta$. Since the particular v interval chosen was immaterial, the conclusion is that it will always be advantageous to break an aggregation step into two smaller steps; i.e. into two stages with an equivalent effect. This has the implication that the aggregation steps must be infinitely dense on the y-axis, and trunking must in fact be continuous. If the optimal increment u in (5.34) is infinitesimal but positive, then the equation reduces to

$$g(v) + F'(v) = 0.$$

This equation with the boundary condition (5.32) has the solution (5.33), which is exactly the value of the total cost if trunking takes place at a constant continuous scaling rate. ◊

The solution is almost disconcerting in its simplicity. Trunking from a dense mass of source nodes will in fact be almost continuous if the rules of Theorem 5.10 are followed. For telephone exchanges and the like, this is not practical, and exchange levels will be quantised and, indeed, few in number. However, the recommendation does seem to be that the aggregation factor should be relatively constant over stages. There are, of course, good reasons for exceptions. The first stage is exceptional in that it aggregates many variable and largely inactive sources to arrive at a statistically stable flow, and the last stage is one of direct bridging rather than aggregation. If different technologies are used at the different levels then that will of course also make a difference.

The OSLA terminal condition (5.32) implies a tangential coincidence $F' = L'$ of value function F and terminal cost L at the optimal termination value v_0. This is illustrated in Figure 5.3(i): $F(v)$ is concave decreasing (because of (5.33) and the fact that $g(v)$ is increasing) and $L(v)$ is decreasing and becomes convex as L approaches its limit value with increasing v. The corresponding relation for $G(N)$ and $K(N)$ is illustrated in Figure 5.3(ii). Both functions are now convex decreasing, as we shall see in a moment, but the tangent point of Figure 5.3(i) survives.

In terms of N (treated as a continuous variable) relations (5.32) and (5.33) become respectively

$$N^{-1}c(N) + K'(N) = 0, \qquad G(N) = \int_N^{N_0} w^{-1}c(w)\mathrm{d}w + K(N_0), \tag{5.35}$$

where $N_0 = \exp(v_0)$. The factor w^{-1} in the integral for G comes about because of the distinction between scaling factor and scaling rate. Suppose that we reverted to the discrete

case of adopting h steps of equal scaling factor. Expression (5.33) for the cost incurred under this policy, starting from $v = 0$, would then become

$$F(0) = \sum_{t=0}^{h-1} (\mathrm{e}^{\lambda} - 1) g(\lambda t) + L(v_0),$$

where $\lambda = v_0/h$. The corresponding equation for G would be

$$G(1) = \sum_{t=0}^{h-1} (\mathrm{e}^{\lambda} - 1) c(\mathrm{e}^{\lambda t}) + K(N_0), \tag{5.36}$$

which makes clear the constant scaling at each transition. The w^{-1} in the integral of (5.35) corresponds to the increasing spacing of the argument $\mathrm{e}^{\lambda t}$ in (5.36) as t increases. Expression (5.35) for $G(N)$ is convex decreasing, because $c(N)/N$ is decreasing.

5.7 The multi-dimensional case

There are substantial extra complications in evaluating the outflow in more than one dimension. However, the final results generalise those of the one-dimensional case in a rather direct fashion. The first task is to determine the dependence of the various cost terms upon N. More explicitly, to determine the exact order of dependence. There will be a number of proportionality factors which we shall assume effectively normalised to unity.

Consider the progression from cells of scale N to those of scale Nn. The n^p lines from the exchanges of the N-cell to the next-level exchange of the Nn-cell will have scale N and so a total length of order Nn^{p+1}. They will thus incur cost of order $Nn^{p+1}c(N)$, where $c(N)$ is again the cost per unit distance of carrying the outflow $x(N)$. However, these costs pertain to a cube of scale Nn, and so of volume $(Nn)^p$. The cost per unit volume of this aggregation step is thus of order $nN^{1-p}c(N)$ and formula (5.26) then becomes modified to

$$C = \sum_{t=1}^{h-1} (n_t - 1) N_t^{1-p} c(N_t) + K(N_h). \tag{5.37}$$

Here $K(N)$ is again the cost per unit volume of termination at scale N. One will expect that it has the form

$$K(N) = N^{-p} \sum_j N d_j \chi(x_j(N)) = N^{1-p} \sum_j d_j \chi(x_j(N)), \tag{5.38}$$

where j labels the final-level exchanges and Nd_j is the distance of exchange j from some fixed exchange, say 0.

It is the behaviour of the outflow $x(N)$ which needs special analysis, set out in Appendix 1. We retain the assumption that $\rho(r)$ decays at least as fast as r^{-q} for large r, where $q > p + 1$. The conclusion is then that, for cubic cells of the type we have considered, the outflow has the form

$$x(N) = 2pN^{p-1} z(N). \tag{5.39}$$

Here $z(N)$ is the outflow density (i.e. outflow per unit surface area of the cell) averaged over the surface of the cell. As a function of N it belongs to the class $\mathcal{Q}$; i.e. it increases from zero monotonically and concavely as N increases from zero. Furthermore, it increases to the finite limit

$$z(\infty) = (\omega_p/2)\int_0^\infty r^p \rho(r)\mathrm{d}r, \tag{5.40}$$

where ω_p is the volume of the unit ball in p dimensions. Relations (5.39) and (5.40) hold in fact for any convex shape of cell, if the factor $2p$ in (5.39) is replaced by the surface area of the cell at unit scale. The reason for this behaviour is, of course, that under our assumption on the decay rate of $\rho(r)$, the main contribution to outflow for large N comes from callers just inside the N-cell, and so is proportional to the surface area of the cell.

We can now be more specific about the character of $K(N)$. The following assertions are proved in Appendix 1.

Theorem 5.11 *Suppose that $\chi(x)$ has the form x^ν, and that $\rho(r)$ decays at least as fast as r^{-q} for large r, where*

$$q > \frac{p+1}{\nu}. \tag{5.41}$$

Then: (i) If $\nu = 1$ the terminal cost function $K(N)$ increases from $2z$ to $2pz$ as N increases from zero to ∞, where $z = z(\infty)$ is the limiting outflow density evaluated in (5.40).

(ii) If $0 < \nu < 1$ and $p > 1$ then $K(N)$ decreases from ∞ to 0 over the corresponding range.

It follows from (i) that $K(N)$ increases with N if there is no environmental cost, with the consequence that the optimal choice is to terminate at as small a value of N as possible. This is, to return to the direct source–sink connections required by the naive optimisation of Chapter 1. However, even then, the effect of intermittent demand at the individual level would be to require a degree of initial trunking sufficient to give a smooth aggregated flow. That is, to establish 'line concentrators', which would then form direct mutual connections.

A cost function

$$\chi(x) = c_\nu x^\nu + c_1 x \tag{5.42}$$

is realistic in that it contains both concave and proportional components if $0 < \nu < 1$. The theorem implies that the corresponding terminal cost $K(N)$ decreases from ∞ to a multiple of $z(\infty)$ as N increases from 0 to ∞. One would expect this decline to be monotonic and convex, but see from assertion (i) of the theorem that this can be so only if environmental penalties bite sufficiently.

The cost expression (5.37) now takes the form

$$C = \sum_{t=1}^{h-1}(n_t - 1)N_t^{1-p}\chi[N_t^{p-1}z(N_t)] + K(N_h). \tag{5.43}$$

If the environmental cost function $\chi(x)$ becomes linear for large x, as in the case (5.42), then the terms in N^{p-1} will cancel in expression (5.43), and one is essentially back in the one-dimensional case.

Suppose, however, that the proportional term in (5.42) is missing. Then, if the N_t values are large enough that $z(N_t)$ is close to its limit value z the sum (5.43) essentially becomes

$$C = \sum_{t=1}^{h-1} (n_t - 1)(2pz)^{\nu} N_t^{-\kappa} + K(N_h), \tag{5.44}$$

where $\kappa = (p-1)(1-\nu)$. Because the summands are now decreasing so powerfully, it is possible to have a scheme in which one continues to trunk indefinitely while only incurring finite cost. Specifically, if q is large enough, then to the same degree of approximation we have

$$K(N) \approx 2pz^{\nu} N^{-\kappa},$$

this being the contribution from cells neighbouring the reference cell and sharing a face of dimension $p-1$. The terms neglected are those relating to neighbours of lower-order contact and nonneighbours. The OSLA condition determining the stopping value of N is

$$NK'(N) + c(N) = 0.$$

In the present case we have

$$NK'(N) + c(N) \approx z^{\nu} N^{-\kappa} [(2p)^{\nu} - 2p\kappa],$$

so immediate stopping or continuation is advised according as the expression

$$(p-1)(1-\nu) - (2p)^{\nu-1} \tag{5.45}$$

is negative or positive. In the case $p = 1$ it is negative, indicating that stopping is long overdue if N is so large that $z(N)$ has approached its limit value. For larger p immediate stopping is indicated if $1 \geq \nu > \nu_p$, where ν_p approaches 1 from below as p increases. It does seem, then, that aggregation at successive exchange levels will continue indefinitely if $\nu < \nu_p$, although one can be sure of this only after having excluded the possibility of an optimal stop at moderate value of N.

Of course, in the real world one would hit a constraint of some kind as scale were increased. If the physical domain $\mathcal{P}$ were the surface of a sphere rather than an infinite plane then one could still have isotropy, but would meet a limit in aggregation at some level, possibly when the highest-level exchanges were placed at the vertices of a simplex inscribed to the sphere. If $\mathcal{P}$ were a bounded land-mass then one would also run out of room in which to expand, but would also lose isotropy. In this case, termination would certainly provide connection, and perhaps complete connection, between all highest-level exchanges.

As particular cases of outflow evaluation we can give the expressions for cubic cells in dimensions 1 and 2. For $p = 1$ we have the evaluation

$$x(N) = 2z(N) = 2\int_0^{+\infty} \min(N, r)\rho(r)\mathrm{d}r$$

of (5.24), with the 2 being seen as σ_1. For $p = 2$ we find in Appendix 1 that

$$z(N) = \int_0^\infty f(N, r)\rho(r)\mathrm{d}r$$

where $f(N, r)$ has the evaluation

$$\left.\begin{array}{ll} (r^2/2)(4 - r/N), & (0 \le r \le N) \\ 2Nr\theta + (r/2N)[r^2 + 2N^2 - 4N\sqrt{r^2 - N^2}], & (N \le r \le N\sqrt{2}] \\ 0, & (r \ge N\sqrt{2}) \end{array}\right\}$$

and $\theta = \cos^{-1}(N/r)$.

6

Road networks

6.1 The setting

We are daily and directly convinced of the importance of traffic networks, and more particularly of road networks, and of the desirability that the factors that ensure their smooth running should be well understood. There is a vast literature on the subject, and a substantial body of theory, which must indeed take account of new factors.

The road network is a graph, whose streets constitute the arcs and whose junctions the nodes. However, traffic does not consist of just a single 'commodity' to be shunted around the network, but breaks up into classes labelled by their destination, and possibly by other characters as well. For simplicity, we shall suppose that destinations coincide with exit nodes.

This seems then just a particular case of the destination-specific distribution net of Chapter 2. However, we saw there that a naive optimisation led to a totally unrealistic solution, in which every origin–destination pair for which there was demand was connected by a direct route, of rating matched to this demand. Such a solution ignores a multitude of issues, arguable only in the order in which one should list them. From the aesthetic or environmental point of view, one could never allow the countryside to suffer this death of a thousand cuts. From the utilitarian point of view, land, and substantial blocks of it, is needed for other purposes (e.g. homes, offices, shopping, industry, agriculture, leisure activities, plus access to some sites and total protection of others), and its usage must be planned. From the point of view of practicality in its narrowest terms, there will be physical obstacles to direct linkage, and millions of direct links (most of them lightly loaded) would produce milliards of crossings – of which more anon.

We can loosely class all such issues as 'environmental' in one sense or another, and argued in Chapter 5 that such considerations implied concave costs, if roads were given capacities adapted to the traffic they were required to bear. These in turn implied a network with a high degree of trunking and, indeed, a hierarchical system. This is just what we see in existing road networks, of course, with city roads being classified as main roads, distribution roads or access roads, and with traffic between cities being carried on a relatively sparse network of highways.

One might then see the hierarchical road structure as analogous to the hierarchical telephone network considered in Chapter 5 (before we specialised to the isotropic case). There are differences, however. Roads take up much more space than telephone lines, and are largely confined to the surface of the Earth (idealised as a plane). We shall expand on these points in the next section.

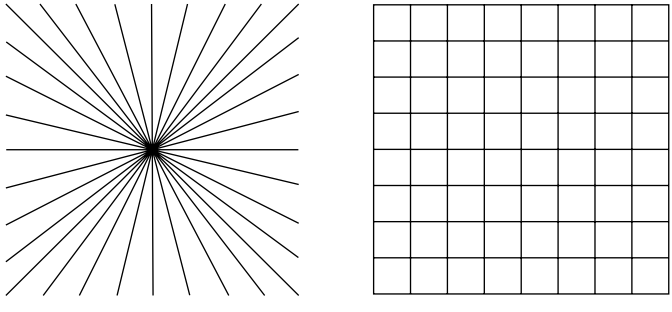
Moreover, road networks present other special features. The reduction in Chapter 3 of the optimal network to a set of direct origin–destination connections occurred because we assumed that flow costs were seminvariantly scaled. At its simplest, we might have assumed a cost $c(x) = adH(x/a)$ for the passage of flow x along a road segment of length d and capacity (rating, cross-section) a. Here H is the step function defined in equation (1.15), so the implication is that any flow up to a can be accepted without immediate cost, but that a flow greater than a cannot be accepted. However, the phenomenon of *congestion*, familiar to every motorist, will induce an earlier growth of $c(x)$ with x. In brief, congestion is a statistical phenomenon, which leads to the crowding and slowing of traffic, even for flow rates within the theoretical capacity of the link. We follow up this point and its implications in Sections 6.3 and 6.4.

Another new feature is the fact that motorists are independent agents, who choose their own routes and timing, although of course subject to the pressures or restrictions that the system may impose. The entities making up the flow are thus active decision-makers, whose personal criteria for action may not be consistent with the system criterion. We consider this point in Section 6.5.

6.2 Structural considerations

Optimality may indeed indicate that road networks should, like telephone networks, have a hierarchical structure, but there are differences between the two cases. We sketch two levels of the tree structure of the telephone network in Figure 6.1(i). Roads could never be organised wholly on this pattern, because roads take up space, whereas cables, relatively speaking, do not. Moreover, the economics of cables are such that it is rational for someone to call his neighbour by a route that travels some miles to the local exchange and back again, whereas he would never make the equivalent drive if he wished to pay a visit. There are local housing developments that have something of a tree character (with closes, etc.), but considerations of both internal travel times and utilisation of ground area limit the amount of branching that is practical.

The familiar grid pattern, Figure 6.1(ii), offers complete area utilisation and also fairly direct point-to-point internal access, and could be regarded as the best traffic equivalent of the zero-level fan of connections in Figure 6.1(i). Like the fan, it is only efficient



(i) (ii)

Fig. 6.1 (i) The fan of level-zero connections for the telephone network of Chapter 5.
(ii) A recognised urban packing: the grid.

up to a certain size, however. Once that size is exceeded, one would wish to make connections at a higher level between units of this type, grid patches. The difficulty with a grid is, of course, the great number of intersections. Elevated (or underground) roadways and railways represent attempts to avoid the conflicts of an intersection and to achieve a physical layering of trunk routes by a modest escape from the constraint of two dimensions. However, cost puts a fairly stern limit on such endeavours, and throughways, ringways and the like represent attempts to realise these higher levels of trunk connection, while staying in two dimensions. The problem of efficient transport is then in considerable degree one of realising an ideal three-dimensional structure on a two-dimensional living space.

The constraints on flight are of course much weaker, and it is interesting that the most recent designs for long-haul passenger aircraft under construction represent the two conflicting philosophies. The Boeing 777-200LR carries only a relatively modest 300 passengers, but is advertised as being able to fly nonstop between virtually any two cities in the world. By contrast, the Airbus A380 can take a passenger load of up to 1000, but is restricted to flying between 'hubs', airports large enough to accept such a behemoth. So, for air transport, the relative advantages of single-stage or multi-stage passage between two given points are still being debated.

6.3 Congestion: the queueing analogue

Once one has a road that is smooth and obstacle-free, the main impediment to free passage is other vehicles. Vehicles travel at different speeds, indeed at variable speeds, and passing may or may not be possible. One may then find oneself in a cohort of vehicles, most of which are travelling at a speed less than they would wish. This is the phenomenon of congestion. Sometimes there is an evident reason for congestion, such as a bottleneck ahead. But sometimes there is seemingly none; such an accumulation can take on a life of its own, and persist for considerable periods for no apparent reason.

A model of the situation would require one to consider both dynamic and stochastic effects. Moreover, the 'state variable' of the model would have to be the state of the whole long file of traffic rather than of just a few adjacent vehicles. There have been many attempts to formulate and analyse such models, which we shall consider briefly in the next section. However, it is illuminating to consider a simple queueing version that, while certainly not an adequate model, provides a partial analogue.

Let us regard a unit length of a single-lane road, a 'cell', as a barrier that vehicles, arriving at rate x, must pass. Suppose that this has capacity a. By speaking of 'capacity' we are implying that we are in the case in which traffic passes freely if $x < a$ but suffers an indefinite build-up if $x > a$. Since a vehicle will suffer either zero or infinite delay in the two cases, the cost is just the function $H(x/a)$ defined in equation (1.15). This should be multiplied by ad to be consistent with our previous conventions, but is in fact invariant under such a multiplication.

However, let us take a stochastic version of the model, something we shall formulate more generally in Chapter 13. Suppose that the constant arrival rate of x is replaced by a Poisson stream of individual vehicles of rate x. Suppose also that the lane segment can deal with a vehicles per unit time, in the sense that the times τ taken to 'process'

vehicles arriving at the barrier are independent random variables with expectation $1/a$. This processing time refers to delay rather than to the time taken for uninterrupted passage. Then a queue of vehicles will form at the barrier, and the probability that there are n vehicles in the queue is given by the geometric distribution

$$P(n) = (1-R)R^n \qquad (n = 0, 1, 2, \dots), \tag{6.1}$$

where $R = x/a$ is the *traffic intensity* (see Section 13.1). Relation (6.1) holds only if $x < a$; if $x \geq a$ then queue size n is infinite with probability one. The expected queue size is

$$\mathrm{E}(n) = \frac{R}{1-R} = \frac{x}{a-x} \tag{6.2}$$

and the expected delay experienced by a vehicle at the barrier is

$$\mathrm{E}(n)\mathrm{E}(\tau) = \frac{x}{a(a-x)}.$$

If we regard this delay as the cost for the motorist, then the cost for the whole stream is x times the expected delay:

$$\frac{x^2}{a(a-x)} = \frac{R^2}{(1-R)} = \phi(R), \tag{6.3}$$

say. This is then a function purely of the traffic intensity $R = x/a$.

However, suppose that the handling rate of a is achieved by using a parallel and independent lanes, each of handling rate unity and receiving a fraction x/a of the input. (In that this is the rate of the Poisson stream into the lane; we shall see in Chapter 13 that the traffic stream from one cell to the next is Poisson. We assume a integer-valued for convenience, but the idea carries through more generally.) Then the queues in the cell represented by the a lanes are independent with traffic intensities x/a, and so the total cost incurred in the cell per unit time is

$$c(x, a) = a\phi(x/a) = \frac{x^2}{(a-x)}. \tag{6.4}$$

The point of this calculation is to show that achievement of capacity by adding lanes rather than by increasing the handling rate within a lane yields the seminvariant cost function (6.4), with the implication that, in this case at least, congestion does not induce trunking. The cost function does represent the effect of congestion, however, in that it increases continuously from zero to an infinite value at $x = a$, rather than suddenly increasing from zero to infinity at that value. In other words, the limitations on capacity begin to be sensed as flow increases towards the design limit. These were matters early taken into account by Kleinrock (1976) in this context.

One may say that the introduction of separate lanes allows passing, at least between lanes. We have not allowed lane changing; if we did so then one would virtually be back in the case of a single lane of handling rate a.

The model gives a correct description of congestion while this remains local, but the local character of the model means that it cannot be adequate. When a queue becomes so large that it backs into the preceding cell, then a whole different class of phenomena arises. We explore these briefly in the next section.

6.4 Congestion: fluid and discrete models

We require a more adequate model. A seminal paper in this direction which remains influential is that by Lighthill and Whitham (1955). This regards the flow of vehicles as analogous to the flow of a compressible fluid in one dimension: along the axis constituted by the road. One can then set up a continuity equation (relating the temporal gradient of density at a point in the fluid to the spatial gradient of flow rate). However, this has to be supplemented by something more specific to the problem; how do drivers respond to the distance and speed of the car in front?

One envisages a long stretch of road with statistically uniform traffic conditions along it. Let v be the speed of cars on the road, in a locally averaged sense, and let c be the concentration of vehicles: the number of vehicles in a unit length of road if one could take a snapshot. The notation c is in fact an appropriate one, since it turns out that the flow cost incurred is just the concentration. The flow rate is determined as $x = cv$.

Lighthill and Whitham then appealed to an empirical fact: that velocity v seems to be virtually determined by the concentration c, and that the dependence is well approximated by the piecewise-linear relation

$$v = v(c) = \begin{cases} v_0, & (0 \le c \le ka), \\ v_0 - a^{-1}(c - ka), & (c \ge ka). \end{cases} \tag{6.5}$$

Here a is a measure of the capacity of the road and k is a constant. So, as c increases, v is initially equal to a prescribed speed limit v_0 for the road. However, when concentration reaches a critical value ka, speed declines linearly, until it reaches zero when $c = a(v_0 + k)$. This behaviour is supported by observation; something like it is also predicted by the models discussed later in the section.

In Figure 6.2 we sketch the function (6.5) and in Figure 6.3 the function

$$x(c) = cv(c). \tag{6.6}$$

This graph of flow against concentration is known as the 'fundamental diagram'. According to it, flow increases as concentration increases, first linearly and then parabolically. The parabola rises to a maximum, x_{max}, say, and then declines to zero. In the case (6.5) we have

$$x_{max} = a(v_0 + k)^2/4,$$

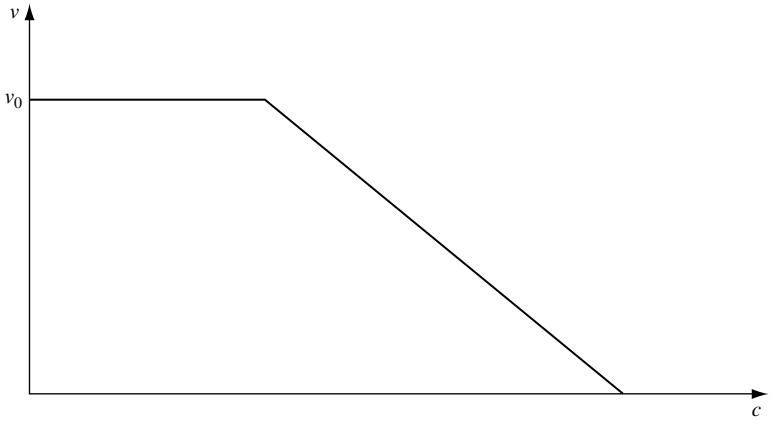


Fig. 6.2 The relation between velocity v and concentration c postulated in relation (6.5).

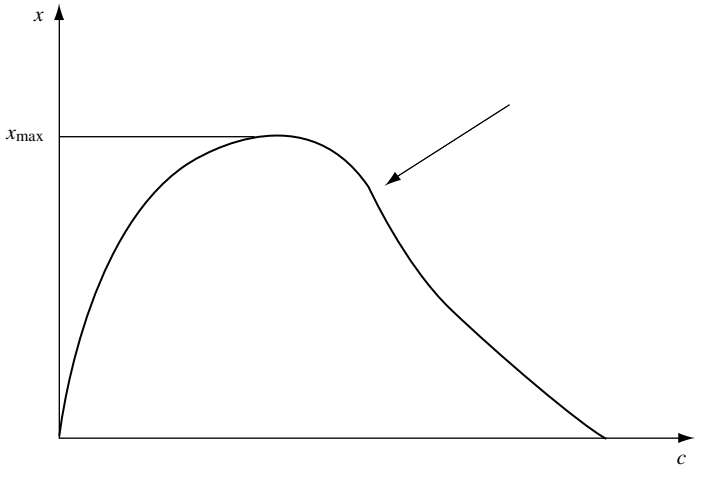


Fig. 6.3 The 'fundamental diagram', representing the relation between flow x and concentration c implied by relation (6.5).

and this is the effective capacity of the road. On the plots of observed flow/concentration pairs this maximal value is reached at a concentration of about 40 vehicles per kilometre. The plot at lower concentrations follows a rising curve quite closely. However, the plot at greater concentrations shows a very considerable scatter about a falling curve. The interpretation is that in the first case the flow is free, orderly and predictable, but in the second is congested and much less predictable. However, traffic in the congested case does show a regularity in that it is 'synchronised'; congestion forces all lanes to proceed at about the same speed.

One should not see the function $x(c)$ as expressing a causal relationship. It is rather its inverse $c(x)$ which does so. But this inverse function is two-valued. If one attempts to send a flow x along the road then this is bound to fail if $x > x_{max}$. If $x < x_{max}$ then two possible values of c are possible. At the lower value flow is free and homogeneous, corresponding to laminar flow of a fluid. At the upper value the system has become trapped in a regime

in which jams have developed. Needless to say, one would wish to operate only on the ascending branch; whether one can do so is dictated by the rate at which traffic is admitted to the system and by subtleties of stability which we shall discuss shortly.

An immediate point of interest is that, if the cost of passage is taken as the total vehicle time incurred in passage through unit distance, then this is $x/v = c$, and so identical with concentration, c. It is determined as a function $c(x)$ of x by simply inverting relation (6.6). The two roots yielded,

$$c(x) = (a/2)[v_0 + k \mp \sqrt{(v_0 + k)^2 - 4(x/a)}],$$

are the cost rates for a free-flow regime and a congested regime, respectively. The first is the relevant one; it is seminvarient in form, and so yields a concave flow-plus-plant cost $\chi(x)$ if plant costs are concave in a.

Lighthill and Whitham (1955) develop the argument further, to establish a number of interesting conclusions. One of these is that small deviations from uniformity are propagated along the road with velocity $\mathrm{d}v/\mathrm{d}c$. This will be less than v if we are considering the laminar regime, indicating that such disturbances propagate back along the traffic stream. If there is a point in the traffic stream at which the concentration ahead is greater than that behind, then a 'shock wave' of suddenly increased concentration can be propagated back along the traffic stream. Development of this argument shows how a passing constriction in the road can produce those stretches of congestion from which one can suddenly emerge without ever seeing its cause.

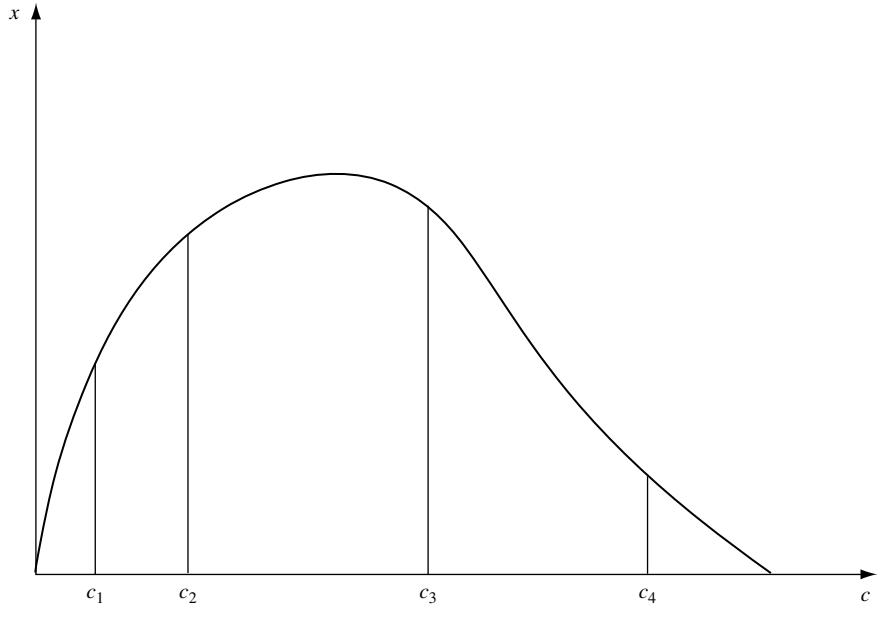


Fig. 6.4 The concentration/flow regimes that were recognised by Kerner and Könhauser (1994) and Bando *et al.* (1995). For $c < c_1$ the free-flow regime is stable. For $c > c_4$ the completely jammed regime is stable. For $c_2 < c < c_3$ there is no stable regime. The free-flow and completely jammed regimes are metastable in the intervals (c_1, c_2) and (c_3, c_4) respectively; see the text.

There is a large literature examining models that might give a theoretical basis for the observed fundamental diagram $x(c)$, excellently and conveniently surveyed by Nagel *et al.* (2003). Many of these models are discrete, in that individual vehicles are identified in the flow. The string of such vehicles can be regarded as a cellular automaton, and so the term 'cellular' is often used. This term rather confuses the moving cells of the traffic stream with the fixed cells of the road we employed in the last section, and so we shall keep to the description 'discrete'. In any case, the rules of the automaton are set up as the reaction rules of a driver to the speed and distance of the vehicle in front of him. In some cases an element of randomness is built into these rules.

Models based on a realistic choice of these reaction rules have not yet yielded markedly to mathematical analysis. However, computation confirms a relationship $x(c)$ of the form of the fundamental diagram, but with a finer regime structure, indicated in Figure 6.4. Kerner and Könhauser (1994) and Bando *et al.* (1995) deduce the existence of four transition values c_1 to c_4. For $c < c_1$ the free-flow regime is stable. For $c > c_4$ the completely jammed regime is stable. There is no stable regime in the interval (c_2, c_3). The free-flow regime is metastable in the interval (c_1, c_2) in that it is stable against small perturbations, but a sufficiently large perturbation can take it into an inhomogeneous regime, in which there are intervals of free flow (of concentration c_1) and intervals of jammed traffic (of concentration c_4). The completely jammed regime is likewise metastable in the interval (c_3, c_4).

6.5 When motorists choose

We come now to the interesting question: what happens when motorists are themselves able to choose what they believe to be their most favourable route, rather than accepting rules laid down by the system? This is, of course, the case in practice, although the system may by restrictions and tolls be able to limit or bias the motorist's choice. We shall assume for simplicity that all motorists are statistically identical, and so shall not make distinctions such as that between heavy and light vehicles.

If the system accepts the motorist's valuation of the cost $\tilde{c}_{jk}(x_{jk})$ of traversing segment jk then the relation

$$c_{jk}(x_{jk}) = x_{jk}\tilde{c}_{jk}(x_{jk})$$

must hold. We are speaking of equilibrium conditions, so that a motorist is well aware of conditions on all sectors, which he probably would not be in a dynamic context. If we make our standard assumption of seminvariance

$$c_{jk}(x_{jk}) = a_{jk}d_{jk}\phi(x_{jk}/a_{jk}) \tag{6.7}$$

then we have

$$\tilde{c}_{jk}(x_{jk}) = d_{jk}\tilde{\phi}(x_{jk}/a_{jk}),$$

where $\tilde{\phi}(R) = R^{-1}\phi(R)$. The conditions expressed in relations (1.3) ensure that $\tilde{\phi}(R)$ will have a limiting value of zero as R approaches zero, as one would expect.

We assume then that a motorist will choose a route that minimises the sum of costs $\tilde{c}_{jk}$ between his current position h and his destination i. His decision will then depend

upon current traffic x, and so upon the decisions of other road users. This is then a noncooperative game (with an infinite number of players, since we have supposed the x_r to be continuous variables). The concept of an 'optimum' is much more qualified in such cases. A 'strategy' or 'policy' for all the players is specified by the set of flows $x = \{x_r\}$; a single player (motorist) can make his small contribution to one of these. A policy x is 'Pareto-optimal', or 'on the Pareto boundary', if there is no alternative policy that leaves nobody worse off and makes somebody better off. This can be a large set, and one needs other criteria to determine a preference within this set. More to the point, it may be impossible, without co-operation between players, to reach a Pareto-optimal solution. A weaker concept is that of a 'Nash equilibrium', which is an equilibrium in the sense of some kind of balance between competing players rather than the temporal equilibrium of a dynamic system. A policy (in this case, a flow specification) constitutes a Nash equilibrium if there is no incentive for any individual player, working on his own, to vary the decision specified for him by the policy.

Even a Nash equilibrium may not exist; the 'prisoner's dilemma' provides the well-known and archetypal example. The classic example in the transportation context is that provided by Braess (1968). It is classic to the point that it is the example always given, and we shall not deviate. It concerns a net of four nodes, with directed arcs between them as indicated in Figure 6.5. There are six players (i.e. motorists), so flows can take integer values from 0 to 6. All of them wish to get from node 1 to node 3. The user cost functions for the different links are specified as

$$\begin{aligned}
\tilde{c}_{12} &= 10x_{12} \\
\tilde{c}_{43} &= 10x_{43} \\
\tilde{c}_{23} &= 50 + x_{23} \\
\tilde{c}_{14} &= 50 + x_{14} \\
\tilde{c}_{24} &= 10 + x_{24}.
\end{aligned}$$

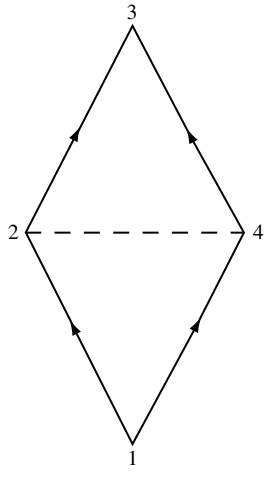


Fig. 6.5 The road pattern illustrating Braess' paradox.

One can regard the constant terms as the cost of fixed passage times, the terms in x_{jk} as representing the cost of congestion.

Suppose, to begin with, that the cross-link 24 is not available. Then just the two routes are available, 123 and 143, and one confirms easily that the symmetric solution, with flows of 3 on both routes, gives both the system optimum and the only Nash equilibrium. This solution gives a cost per motorist of 83 on either route. The system cost is thus 498.

Suppose now that the cross-link 24 is opened. This certainly can only improve the system optimum, and one might think that it could only improve the costs for individual motorists. Braess' paradox is that it in fact makes matters worse for all motorists. A driver on route 123 can now take advantage of the cross-link and switch his route to 1243, lowering his route cost from 83 to 71. However, this increases the congestion on the 43 sector, with the consequence that it becomes advantageous for a driver on the 143 route to switch to the 1423 route. This leaves two drivers on each of the two original routes, and one driver on each of the cost-equivalent routes 1243 and 1423, all drivers suffering an individual cost of 92. One verifies that this configuration is the only Nash equilibrium, and so that is where matters will settle unless motorists agree to co-operate.

The effect is, then, that the opening of a new opportunity has left everybody worse off. The net is so balanced that the opening of the new link gives an opportunity which, if taken by one motorist, improves matters for him at the expense of a deterioration for others. When these others try to recover their position then everybody ends up worse off, exactly as for the prisoner's dilemma. The pointlessness of introducing the cross-link 24 could have been predicted from study of the system optimum. The marginal cost of shifting traffic from the symmetric configuration by a shift from route 123 to 1243 is $c'_{24}(0) + c'_{43}(3) - c'_{23}(3) = 14 > 0$, ditto for a shift from route 143. The cross-link is thus never used in the system-optimised configuration.

A certain amount of work has been done on the form of road networks that are guaranteed not to incorporate this perverse feature. That is, if one assumes that drivers will take an apparent advantage, then there is no link that could be removed without immediate cost to the individual. One can formulate sufficient conditions for this property, but these cover only a few situations.

However, there is one way of inducing motorists to behave so that the system optimum is realised, and that is to levy an explicit link toll. This must be one that brings the actual cost of passage up from the motorist's face valuation $\tilde{c}_{jk}$ to the 'effective toll' c'_{jk} of the system optimisation (see Section 3.2). We thus have

Theorem 6.1 *Assume that motorists choose the routes that minimise their sum of link costs $\tilde{c}_{jk}(x_{jk})$. Then the system optimum will be realised if the link cost is supplemented by a link toll of $c'_{jk}(x_{jk}) - \tilde{c}_{jk}(x_{jk})$. If the seminvariance property (6.7) holds then this becomes $x_{jk}\tilde{c}'_{jk}(x_{jk})$, plainly nonnegative.*

Much of the literature is concerned with explicit pricing, and so takes a much more overtly economic approach than we have adopted here. In particular, rather then prescribed traffic demands f_{hi} being assumed, demand functions are postulated, relating demand to prices levied (see Section 1.8). As examples we note the classic text by Beckmann *et al.* (1956) and the much more recent text by Nagurney (1993).

7

Structural optimisation: Michell structures

We come now to a most remarkable piece of work. This concerns the optimisation of load-bearing structures in engineering, such as frameworks consisting of freely jointed struts and ties – we give examples. The purpose of such structures is to communicate *stress*, a vector quantity, in an economic fashion from the points where external load is applied to the points where it can be off-loaded: the load-accepting *foundation*.

The problem of optimising such structures was considered in a paper, remarkable for its brevity and penetration as much as for its prescience, published as early as 1904 by the Australian engineer A. G. M. Michell. He derived the dual form of the problem and exhibited its essential role in determining the optimal design. This was at a time when there was no general understanding of the role or interpretation of the dual variable – Michell uses the concept of a 'virtual displacement', a term that we shall see as justified. He then went on to derive the continuous form of the dual, corresponding to the unrestricted optimisation of structures on the continuum. This opened the way to the study of structures made of material with directional properties (e.g. the use of resin-bonded fibre-glass matting in yacht hulls, the laying down of bone in such a way as to meet the particular pressures and tensions to which it is subjected). It also turned out to offer an understanding of materials that behave plastically rather than elastically under load – i.e. that yield catastrophically when load reaches a critical value.

Michell's theory was largely seen by engineers as an interesting curiosity of limited practical application. However, with much more powerful computing facilities available these days direct optimisation of continuum structures has become possible (see, for example, Bendsøe and Sigmund, 2003; Xie and Steven, 1997), and Michell's early insights suddenly seem both modern and applicable.

7.1 Force and deformation; stress and strain

The analysis fits exactly into the framework we have set up for commodity flows in Chapters 1 and 2. One must just become accustomed to the idea that a *flow* through a transport link is now to be interpreted as a *force* (tension or compression) suffered along its length by a mechanical link, and the *potential* which drove the flow is now to be interpreted as the *displacement* of a node under the influence of the applied forces, with its consequent deformation of the structure. The terms 'stress' and 'strain' are largely equivalent to what we have termed 'force' and 'deformation' but require a more explicit definition, especially in continuum contexts.

However, the nodes of the network (the joints of the framework) are sited in physical space, and force and displacement both have a direction in that space. We must then see them both as vectors. There is also the difference that the material of a link may behave differently under tension or compression. For example, concrete is well known to be strong under compression but weak under tension. This is in distinction to the symmetry we have supposed up to now: that the cost of a flow through a link was independent of its direction. We must then amend the analysis of Chapter 1 to take account of these generalisations, and reinterpret it for this new context.

Let us consider a one-dimensional case first, in that q is regarded as the deformation of a link along its length (i.e. a stretching or shortening) induced when the link is clamped at one end and subjected to a force of p at the other, in the direction of the link from the clamped end. If we introduce functions Φ and Ψ analogous to the ϕ and ψ of Chapter 1 then we have

$$\Phi(p) = \max_q [pq - \Psi(q)]. \tag{7.1}$$

Here $\Psi(q)$ is the *strain potential*, interpretable as the energy stored in the link by its having suffered the deformation q. The relation between p and q expresses a generalised Hooke's law, the relation between stress and strain. This relation is determined by the maximising condition in (7.1):

$$p = \Psi'(q), \tag{7.2}$$

where the prime indicates a differentiation. The law as classically announced by Hooke (*Ut tensio, sic vis*) for elastic bodies prescribes a linear relation. It follows from (7.2) that this would imply a quadratic form for the strain energy $\Psi(q)$. We shall consider more general functions. We must also allow asymmetric functions if the link behaves differently in tension and compression: p and q must then be regarded as signed quantities.

The function $\Phi(p)$ is known as the *complementary potential*; this is the analogue of the flow–cost function of previous chapters. Before we consider design optimisation we have to determine the equilibrium state of the loaded structure, just as we had to determine the equilibrium flow pattern for the distribution networks of Chapter 1. This state can be expressed either in terms of the displacements of the nodes (joints) or of the forces within the members. The most immediately physical approach is to find the displacements that minimise the stored potential energy. This is the analogue of the dual formulation of Chapter 1, and appeals to the energy function $\Psi(q)$. The analogue of the primal formulation is to determine the forces that minimise the complementary potential subject to a balance condition at the nodes. Much as in previous chapters, we shall switch between the two approaches, each of which has some conceptual or algorithmic advantage.

We assume that the function $\Psi(q)$ of the signed scalar q belongs to the class $\mathcal{C}_s$ defined in Section 1.2, but make no assumption of symmetry. Then $\Phi(p)$ also belongs to $\mathcal{C}_s$, and the defining relation (7.1) has the inversion

$$\Psi(q) = \max_p [pq - \Phi(p)]. \tag{7.3}$$

Relation (7.3) implies the alternative expression of Hooke's law

$$q = \Phi'(p), \tag{7.4}$$

which must of course be equivalent to (7.2).

Consider now the vector case, where links, forces and displacements may have all possible orientations. Relations (7.1) and (7.3) still hold if we take q as a column vector and p as a row vector, so that pq is then the inner product of the two vectors. Relations (7.2) and (7.4) will also hold if we take the convention that the vector of differentials of a scalar with respect to a column (row) vector is a row (column) vector. However, if we again consider one end of the link pinned (so that this end of the link is fixed, but the other is free to take any position) then the only part of a displacement of the other end that contributes to the strain energy is that which changes the length of the link. In other words, we can write

$$\Psi(q) = \psi(u^\top q) \tag{7.5}$$

where u is a unit vector in the direction of the link (taken from the pinned end). The inner product $u^\top q$ is then the component of displacement q along the length of the link – the actual deformation in length. The function ψ is the complete analogue of the ψ of Chapter 1. That is, it is a scalar function of a scalar belonging to the class $\mathcal{C}_s$, but no longer presumed symmetric.

It follows then from the vector version of (7.1) that

$$\Phi(p) = \begin{cases} \phi(pu) & \text{if } p \parallel u, \\ +\infty & \text{otherwise,} \end{cases} \tag{7.6}$$

where ϕ is again the Fenchel transform of ψ. Here by $p \parallel u$ we mean that p is a scalar multiple of $u^\top$, so that pu will be just that multiple. This is the quantity which we shall denote by x, consistently with the flow notation of previous chapters. The function $\Phi(p)$ then essentially reduces to the function $c(x) = \phi(x)$, since relation (7.6) expresses the fact that one can only envisage a force in a link that is parallel to that link.

Finally, we shall build the physical dimensions of a link into the cost specifications by making the same seminvariant assumptions that we made in Chapter 1: that

$$\Phi(p) \;\rightarrow\; C(x) = ad\phi(x/a), \tag{7.7}$$

$$\Psi(q) \;\rightarrow\; D(y) = ad\psi(y/d), \tag{7.8}$$

where a and d are respectively the rating and the length of the link, x is the force exerted along its length and y is the deformation suffered along its length (i.e. the total extension, positive or negative). The arguments of ϕ and ψ are then the force per unit cross-section and the extension per unit length respectively.

The power function again supplies an important special case for these functions. For ψ this would correspond to the assumption

$$\psi(y) = \frac{\kappa}{\beta}|y|^\beta \tag{7.9}$$

where the constant σ and the power β may depend upon the sign of y. The Fenchel transform of expression (7.9) is

$$\phi(x) = \frac{\kappa}{\alpha}(|x|/\kappa)^{\alpha} \tag{7.10}$$

where α and β are again related by the conjugacy condition $\alpha^{-1} + \beta^{-1} = 1$. Since corresponding x and y are of the same sign, the dependences of α, β and κ upon the sign of x are the same as they were for y.

The force–extension relation between corresponding values

$$y \propto x^{\alpha-1} \tag{7.11}$$

(for positive x, say) is linear for the case $\alpha = \beta = 2$. This the case of perfectly elastic materials. It corresponds to the case of a simple electrical resistance in Chapter 1, the linearity of Hooke's law corresponding to the linearity of Ohm's law.

The parameter β reflects the rate of increase of stiffness of the link with increasing stress. The point is illustrated by the relation graph of (7.11) in Figure 7.1. For $\beta = 2$ also $\alpha = 2$, the stiffness is constant in that the relation is linear, at least for x of a given sign. For $\beta > 2$ ($\alpha < 2$) the rate of increase of deformation decreases with increasing x, i.e. the link shows increasing stiffness. For $\beta < 2$ ($\alpha > 2$) the reverse is the case. Indeed, as β approaches unity (α approaches infinity) there is catastrophic failure of the link as the force x exceeds unity in magnitude. (Or, if we substitute the ϕ of (7.10) back into (7.7), as x exceeds the critical value of κa.) This is the equivalent of flow's exceeding a hard limit of capacity which we have already seen in Chapter 1. By allowing the family (7.9) we are thus able to discuss a

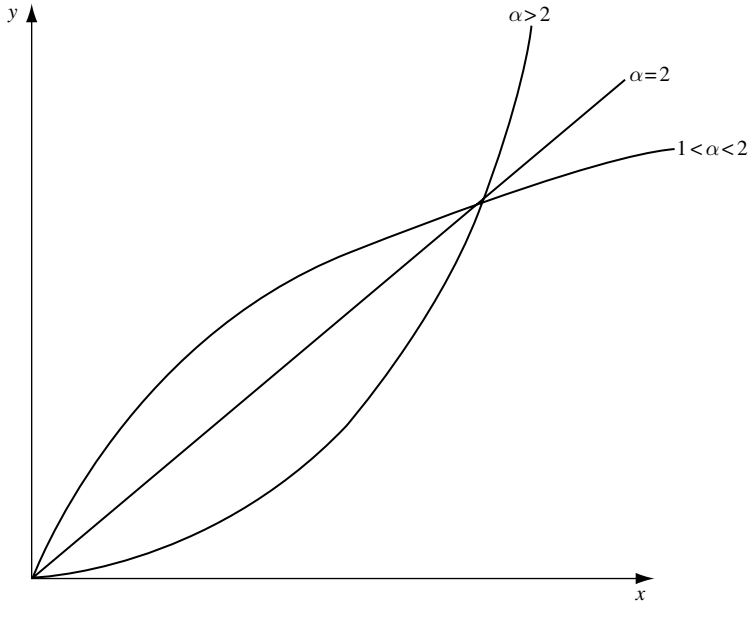


Fig. 7.1 Graphs of the generalised Hooke's law (7.11) relating force x and extension y. This is associated with the power law (7.9) for potential energy stored under extension y.

wide range of behaviours, ranging from plastic ($\beta = 1$) through elastic ($\beta = 2$) to completely unyielding ($\beta = +\infty$).

The elastic case reflects the behaviour one expects of structural metals and of rubber, for example, at least within a certain range of deformation. The case $\alpha \to +\infty$ is the *plastic* case, referring not to the great variety of synthetic materials now termed 'plastics', but to materials such as lead and putty that yield at moderate deformations. Note another point, that we have measured deformations y on the same scale as length d. For a medium such as rubber this would indeed be the case, but for building materials such as steel any deformation would, under normal circumstances, be not more than 1% of the linear dimensions of members. The deformation denoted by y is then a considerable multiple of the actual physical deformation; the physical deformation is such as to store a substantial amount of strain energy without appreciably changing the geometry of the structure.

7.2 Braced frameworks: formulation

Suppose that there are n points in physical space at which one can place the hinged joints of a framework; we shall denote their vector co-ordinates by ξ_j $(j = 1, 2, \ldots, n)$. That is, these points, or *nodes* can be connected by members whose ends are freely hinged, and so are in pure tension (ties) or compression (struts or braces).

There will be other nodes that are given and fixed in space, in that they will not move under any load likely to be placed on them. These are the *foundation points*. The potential foundation points often form a surface that will be referred to as the *foundation arc*. A node that is not a foundation point will be referred to simply as a *joint*.

An external prescribed load is applied to some or all of the nodes; the function of the framework is to communicate this to the foundation. The problem is to find the most economical framework within a prescribed region of space (the 'design space') that will do this, without failure or excessive deformation.

As examples, a bridge is a framework that communicates a load distributed along its deck to a foundation consisting of the available part of the river banks and bed; see Figures 7.2 and 7.3. A coat-hook could be conceived as a framework carrying a vertical point load, see Figure 7.4, which it communicates to the foundation consisting of the wall. The optimal design in fact turns out to be quite compact in form, as one would hope for aesthetic, practical and constructional reasons.

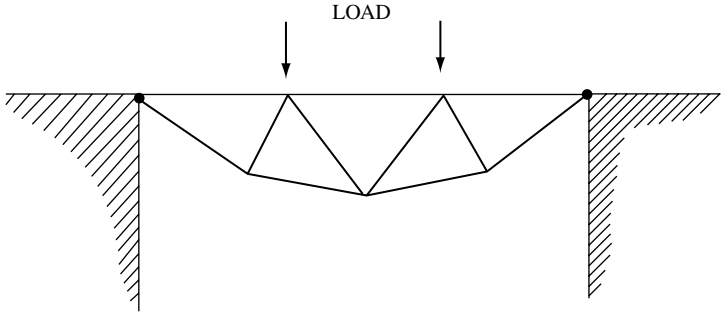


Fig. 7.2 A truss bridge, viewed as a framework communicating the load applied to its deck to foundation points on the river banks.

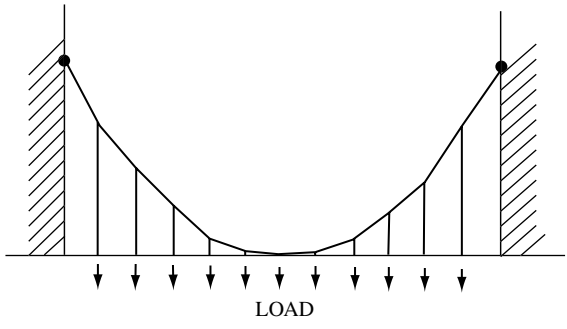


Fig. 7.3 A suspension bridge, viewed as a framework communicating the load applied to its deck to foundation points on the walls of a gorge.

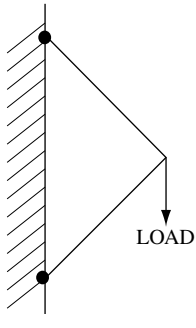


Fig. 7.4 A schematic coat-hook, viewed as framework communicating a vertical point-load to the foundation arc constituted by a vertical wall.

The measures open to us in our optimisation are: (a) to vary the cross-sections of the members connecting the nodes; (b) to vary the positions and number of the internal nodes, and (c) to vary the materials of the framework. When we come to a continuum treatment all these variations will be combined in a single material allocation over space. For the moment we consider a discrete framework, however, and shall consider at most two materials: one for tension members and one for compression members.

Let f_j be the force of the external load applied at node j, a vector quantity. Let x_{jk} be the signed scalar force in member jk: positive for a tension and negative for a compression. Then the equilibrium equation at node j is

$$\sum_k x_{jk} u_{jk} = f_j, \tag{7.12}$$

where u_{jk} is the unit vector in the direction of $\xi_j - \xi_k$. Relation (7.12) holds only at joints, and not at foundation points, where there will be an equilibrating reaction from the foundation.

The assumed cost of stress x_{jk} in the jk member is, by relation (7.7),

$$c_{jk}(x_{jk}) = [ad\phi(x/a)]_{jk}.$$

Here we have taken what will be a very helpful notational convention: that if the subscript jk is appended to a bracketed expression then it will apply to all terms in that expression. Thus, x is to be interpreted as x_{jk} and quantities such as κ and α, if they occur, are to be given the values appropriate to the sign of x_{jk}. We then have a total cost function

$$C(x, a) = \sum_{j,k} [c(x) + \gamma a d]_{jk}, \tag{7.13}$$

where the unit material cost γ may also vary with the sign of x_{jk}, since different material may be used in the two cases.

We have then formulated the optimisation problem as one of minimising expression (7.13) with respect to x and a subject to conditions (7.12). In arriving at this point we have made several assumptions which might not always be acceptable. The self-weight of the structure has been neglected, and also the costs of joints and of fabrication generally. To include these factors would not be impossible, but they can reasonably be neglected in a first treatment. There is another point, which could be serious. To assume that it is the magnitude of forces per unit cross-section which dictates failure might be reasonable for tension members, but not for members in compression. A strut can fail by buckling and, the longer the strut, the smaller the compressive force at which it will do so. This is a feature that should be represented in the cost function. It is true that the solution to the problem as posed tends to avoid long members, but we return to the point in Chapter 8.

7.3 Reduction of the primal problem

The first point of interest is the analogue of Theorem 1.2: the reduction of the problem if we optimise the ratings a in expression (7.13) for given x. Define γ_+ and γ_- as the values of material cost rate γ for members in tension and in compression respectively, and γ_{jk} as the appropriate value for the jk member. Define q_+ as the positive root of $\psi(q) = \gamma_+$ and q_- as the negative root of $\psi(q) = \gamma_-$. Finally, define $p_+ = \psi'(q_+)$ and $p_- = \psi'(q_-)$. The value of q appropriate for the jk member may be denoted q_{jk}, etc.

Theorem 7.1 *(i) The value of rating a_{jk} minimising expression (7.13) is*

$$a_{jk} = x_{jk}/p_{jk}, \tag{7.14}$$

where p_{jk} takes the value p_+ or p_- according as the jk member is in tension or compression.

(ii) The value of the minimised criterion (7.13) is

$$C(x) = \min_a C(x, a) = \sum_{j,k} (qdx)_{jk}. \tag{7.15}$$

The proof follows exactly as for Theorem 1.2, but must now take account of the possible asymmetry of ψ allowed for in Lemma 1.1.

Recall that q has the same sign as x. We could write expression (7.15) as

$$C(x) = q_+ \sum_{+} (dx)_{jk} + q_- \sum_{-} (dx)_{jk}, \tag{7.16}$$

where the summations $\sum_+$ and $\sum_-$ cover members in tension and in compression respectively. The problem has thus been reduced to minimisation of expression (7.16) with respect to x subject to the balance conditions (7.12) and with respect to design (i.e. node number and placement). The influence of ψ thus lingers only in the value of the ratio $|q_+/q_-|$.

We see that the problem reduces tremendously, just as the flow problems of Chapter 1 reduced to a simple transportation problem. However, the fact that the balance conditions (7.12) have a vector nature and involve the link orientations means that the reduced problem is now one of greatly increased sophistication. We do, however, have the analogue of Theorem 4.1, which would be empty for a fixed-loading distribution network, but is not so now.

Theorem 7.2 *Suppose that x has already been optimised for a given design, and a then optimised by (7.14). Then:*

(i) The minimised primal cost function has the form

$$\min_a \min_x C(x, a) = \sum_{j,k} (pqad)_{jk}. \tag{7.17}$$

(ii) A necessary condition that an internal node j be optimally placed is

$$\sum_k (pqau)_{jk} = 0. \tag{7.18}$$

Relation (7.18) can also be written

$$\sum_k (qxu)_{jk} = 0,$$

whereas the balance equation (7.12) can be written

$$\sum_k (xu)_{jk} = 0$$

at an internal node. These two relations then imply that, at an optimally placed internal node, tension forces are statically balanced on their own; likewise for compression forces.

7.4 The dual form of the problem

The Lagrangian form for the minimisation of the convex function (7.13) subject to the linear conditions (7.12) is

$$L(x, y, a) = C(x, a) + \sum_j y_j [f_j - \sum_k x_{jk} u_{jk}], \tag{7.19}$$

where the Lagrangian multipliers y_j are now row vectors, of the dimensionality of physical space. Minimising with respect to x_{jk} we find that

$$[d\phi'(x/a)]_{jk} = [c'(x)]_{jk} = (y_k - y_j)u_{jk}. \tag{7.20}$$

Comparing this with the force–extension relation for the jk member, we see from it that $y_j^\top$ can indeed be identified, at least to within a factor, with the displacement of node j. Michell's characterisation of the dual variable as a virtual displacement is thus completely justified. The value of y_j is zero at a foundation point. This is so formally because the balance relation (7.12) does not hold at a foundation point. The foundation supplies whatever forces are needed to meet the load placed upon it, and the balance condition is self-fulfilling. It is also consistent with the displacement interpretation: that the foundation can by definition accept any load, and does not itself deform.

Minimising the extended Lagrangian form (7.20) with respect to x we find the expression

$$D(y,a) = \min_x L(x,y,a) = \sum_j y_j f_j + \sum_{j,k} \{ad[\gamma - \psi(\Delta u/d)]\}_{jk} \tag{7.21}$$

for the dual cost function. Here

$$\Delta_{jk} = y_j - y_k.$$

This differs from relation (1.26) only in that there are some vector arguments and ψ may be asymmetric. However, these differences are enough to take one into new country.

Expression (7.21) is then to be maximised with respect to y and minimised with respect to all design variables. Minimising with respect to a we deduce

Theorem 7.3 *The dual problem for the a-optimised framework is: choose y to maximise* $\sum_j y_j f_j$ *subject to the constraints* $y = 0$ *on the foundation and*

$$q_- \le \frac{(y_j - y_k)u_{jk}}{d_{jk}} \le q_+ \tag{7.22}$$

for all node pairs. A jk link in tension (compression) can exist in the optimal design only if equality holds in the right-hand (left-hand) inequality of (7.22).

The proof goes exactly as for that in Section 1.5, with the a-minimisation taken before the y-maximisation. The only additional point to watch concerns the possible asymmetry of ψ.

As in Chapter 1, the dual formulation lends itself immediately to passage to the continuum case, formally at least. The continuum version permits a totally free optimisation; i.e. a free disposition of structural material (or materials) in the design space. There is even the prospect of studying a lattice or cellular structure of spatially variable properties and indefinitely fine scale. We do however have to find the continuum analogue of

$$\frac{(y_j - y_k)u_{jk}}{d_{jk}} = \frac{(y_j - y_k)(\xi_j - \xi_k)}{|\xi_j - \xi_k|^2} = \frac{\Delta y \Delta \xi}{|\Delta \xi|^2},$$

where $\Delta\xi$ is a vector increment in position and Δy is the corresponding increment in the row vector y. We formally define then the limiting ratio

$$\frac{Dy}{D\xi} = \lim_{|\Delta\xi| \to 0} \frac{\Delta y \Delta \xi}{|\Delta \xi|^2},$$

the rate of change in a certain ξ direction (the direction of $\Delta\xi$) of the component of y in that direction. Theorem 7.3 then has an obvious limit form:

Theorem 7.4 *The problem dual to the continuous design problem is: choose the vector field $y(\xi)$ to maximise the espression*

$$\int_{\mathcal{D}} yf\mu(\mathrm{d}\xi) \tag{7.23}$$

subject to $y(\xi) = 0$ on the foundation and

$$q_- \leq \frac{Dy}{D\xi} \leq q_+ \tag{7.24}$$

for all relevant ξ and all directions of $\Delta\xi$. Tension (compression) members may exist only on line elements for which the right-hand (left-hand) bound is attained in (7.24).

In expression (7.23) we have allowed a general distribution of load. As previously, in the continuum treatment we consider links of infinitesimal length only – links between clearly separated nodes will be built up from these if the optimisation calls for it.

The theorem is very close to that obtained by Michell in his 1904 paper. Michell necessarily used rather different methods, but then carried them through to detailed design conclusions in a masterly fashion. The paper would have been impressive had it appeared 50 years later, and must rank as one of the foremost and most original contributions to the optimisation literature. Theorem 7.4 has some immediate implications, also deduced by Michell.

Theorem 7.5 *Suppose that the vectors ξ and y have Cartesian components ξ^h and y^h. Then*

(i) The differential coefficient $Dy/D\xi$ has the evaluation

$$\frac{Dy}{D\xi} = \frac{u^\top E(y)u}{u^\top u} = u^\top E(y)u, \tag{7.25}$$

where u is a unit vector in the direction of $\Delta\xi$ and $E(y)$ is the symmetric matrix with jkth element

$$e_{jk} = \frac{1}{2}\left(\frac{\partial y^j}{\partial \xi^k} + \frac{\partial y^k}{\partial \xi^j}\right).$$

(ii) Expression (7.25) achieves its distinct extrema in directions u that are mutually orthogonal. Hence, if tension members meet compression members they must do so at right angles.

Proof If we write the components of $\Delta\xi$ as Δ^h then we have

$$\frac{\Delta y \Delta\xi}{|\Delta\xi|^2} = \frac{\sum_h \Delta^h \sum_j \left(\frac{\partial y^h}{\partial \xi^j}\Delta^j + o(\Delta)\right)}{\sum_h (\Delta^h)^2},$$

whence (7.25) follows. The extreme values of the ratio of quadratic forms (7.25) are just the eigenvalues of E, and the values of u at which these are achieved are just the corrsponding eigenvectors. Eigenvectors corresponding to distinct eigenvalues are certainly mutually orthogonal, whence the assertions of the theorem follow. $\Diamond$

The elements of the matrix E form the 'pure strain tensor'. They represent distinct components of genuine deformation, incorporating no element of rigid translation or rotation.

The theorem implies that it is impossible that, for example, three members of an optimal structure should form a triangle, unless they were all in tension or all in compression. For such structures there is a very simple result, due to Clerk Maxwell.

7.5 The dual field: Hencky–Prandtl nets

Theorem 7.4 has further implications, which will be investigated only for the case of two-dimensional structures.

Theorem 7.6 *The following are the possibilities for a point in an optimal two-dimensional framework.*

(i) The eigenvalues of E at a given point ξ lie strictly between the bounds q_- and q_+. Then no member passes through the point.

(ii) One eigenvalue only attains one of the bounds, say q_+ (q_-). Then any member through the point is in the direction of the corresponding eigenvector and is under tension (compression). It thus either terminates at the point or passes through it in a straight line.

(iii) The eigenvalues both equal the same extreme value, say q_+ (q_-). Then members may extend from the point in any direction, but are all in tension (compression).

(iv) The two eigenvalues equal q_+ and q_-. Then tension and compression members may extend from the point in the directions of the corresponding eigenvectors, and are consequently orthogonal. If this condition holds over a connected spatial set $\mathcal{R}$ then tension and compression members form mutually orthogonal families of curves in $\mathcal{R}$. These two families have the additional property: that any tension (compression) curve crossing two given compression (tension) curves turns through a constant angle in passing from one to the other (see Fig. 7.5).

All statements except the last are immediate consequences of previous theorems. The last assertion sets a significant further limitation on the form of the optimal structure.

Note that 'members' are now infinitesimal line elements, but that a chain of such elements forms a curve, these being the tension and compression curves referred to in the theorem. It is these two families of curves ('slip-lines') which constitute the mutually orthogonal families in $\mathcal{R}$.

Nets consisting of two mutually orthogonal families of curves with the constant turning property expressed in Theorem 7.6 (iv) are termed *Hencky–Prandtl nets*. They seem to have been recognised first by Michell, in this context, in his pioneering paper of 1904. However, the shear-lines in perfectly plastic solids are also described by such nets (for

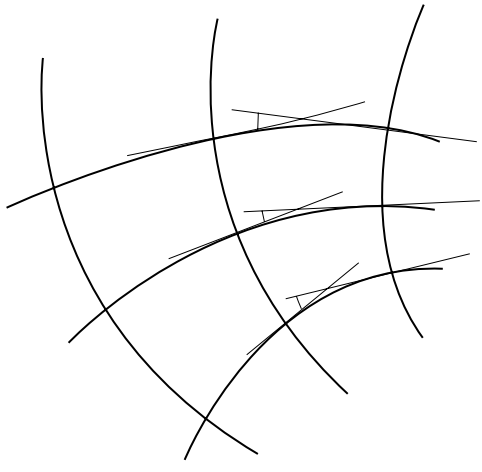


Fig. 7.5 Curves of a Hencky–Prandtl net, illustrating the property that all curves of one family turn through a constant angle along the arc lying between two given curves of the other family.

very good reasons), and it was in this context that Hencky and Prandtl rediscovered them in 1923. The proof of the turning property requires further consideration of the properties of the vector field $y(\xi)$. We have now seen that Michell's interpretation of $y^\top$ as a 'virtual displacement' has a physical basis, and it is helpful to carry this interpretation in mind.

Consider the matrix

$$\Lambda = (\lambda_{jk}) = \left(\frac{\partial y^j}{\partial \xi^k}\right)$$

in the two-dimensional case. If the ξ-space is rotated by the operation $\xi \to H\xi$ (so that H is orthogonal) then Λ suffers the transformation $H\Lambda H^\top$. The invariants of such a transformation – which must be the physically significant local aspects of the v-field – are just

$$w = \frac{1}{2}(\lambda_{12} - \lambda_{21})$$

and the eigenvalues of $E = \frac{1}{2}(\Lambda + \Lambda^\top)$.

The quantity w measures the *rotation* suffered by a line element in ξ-space when the y-deformation is applied. The eigenvalues of E measure the *dilation* (i.e. contraction or extension) that is suffered, in the canonical directions in which this dilation is extremal. These are just the directions in which elements of the optimal structure may be placed.

Suppose that these canonical directions (necessarily orthogonal) make angles υ and $\upsilon + \frac{1}{2}\pi$ with the ξ^1 axis, and correspond to dilations q_+ and q_-. If we make the convenient abbreviations

$$c = \cos\upsilon, \qquad s = \sin\upsilon,$$

then E must have the spectral representation

$$E = q_+ \begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} c \\ s \end{pmatrix}^\top + q_- \begin{pmatrix} -s \\ c \end{pmatrix} \begin{pmatrix} -s \\ c \end{pmatrix}^\top,$$

so that

$$\begin{aligned} e_{11} &= q_+\cos^2 \upsilon + q_-\sin^2 \upsilon, \\ e_{22} &= q_+\sin^2 \upsilon + q_-\cos^2 \upsilon, \\ e_{12} &= e_{21} = (q_+ - q_-)\cos \upsilon \sin \upsilon. \end{aligned}$$

Consider now how υ varies as we move along a tension or compression member (i.e. move in a direction of maximal stretching or shortening under the deformation). In doing so we shall be moving along a curve of one of the two families of mutually orthogonal curves. It is convenient to transform to curvilinear co-ordinates $\chi(\xi)$ and $\eta(\xi)$, so that these two families of curves are specified by constant values of χ and η.

Suppose we are at a point at which the χ and η directions coincide locally with the ξ^1 and ξ^2 directions respectively, so that $\upsilon = 0$; this can always be achieved by an initial rotation of the co-ordinate system. Then E takes the form

$$E = \begin{pmatrix} q_+ & 0 \\ 0 & q_- \end{pmatrix}$$

at the point. Suppose now that we move an infinitesimal distance $\delta\xi^1$ along the ξ^1 axis (i.e. in the direction of the tension member) and that E then suffers the increment

$$\delta E = \delta\xi^1 \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix} + o(\delta\xi^1).$$

Then $\tau_{jk} = \partial e_{jk}/\partial\xi^1$, so that

$$\begin{aligned} \tau_{11} &= \lambda_{111}, \\ \tau_{22} &= \lambda_{221}, \\ \tau_{12} &= \tau_{21} = \frac{1}{2}(\lambda_{121} + \lambda_{211}), \end{aligned}$$

where

$$\lambda_{ijk} = \frac{\partial^2 y^i}{\partial\xi^j \partial\xi^k}.$$

Now, by the properties of the υ-field, $E + \delta E$ must have the same eigenvalues as E: namely q_+ and q_-. So, in particular, the two matrices must have the same trace and determinant. These conditions imply that $\tau_{11} = \tau_{22} = 0$ or

$$\lambda_{11i} = \lambda_{22i} = 0$$

for $i = 1$ and, by the same argument in the ξ^2 direction, also for $i = 2$.

The expression for e_{12} in terms of ϕ above implies the identification

$$\tau_{12} = \frac{\partial e_{12}}{\partial\xi^1} = (q_+ - q_-)\frac{\partial\phi}{\partial\xi^1}.$$

Collecting these statements, we conclude that

$$(q_+ - q_-)\frac{\partial \phi}{\partial \xi^1} = \frac{1}{2}(\lambda_{121} + \lambda_{211}) = \frac{1}{2}(\lambda_{211} - \lambda_{121}) = -\frac{\partial w}{\partial \xi^1}.$$

Now, the only significance of the ξ^1 axis is that it coincides locally with the χ axis; the universal expression of the last relation would then be

$$\frac{\partial}{\partial \chi}(w + \sigma v) = 0,$$

where

$$\sigma = q_+ - q_-.$$

By a similar argument along the ξ^2 axis it is seen that

$$\frac{\partial}{\partial \eta}(w - \sigma v) = 0.$$

That is, $w + \sigma v$ is a function of η alone, say $B(\eta)$, and $w - \sigma v$ is a function of χ alone, say $A(\chi)$. The angle v then, as a function of χ and η, has the representation

$$v(\chi, \eta) = \frac{1}{2\sigma}[B(\eta) - A(\chi)]. \tag{7.26}$$

Relation (7.26) implies the constant-turning property asserted in Theorem 7.6 (iv).

7.6 Some examples of Michell structures

An analytic solution for the y-field characterised in Theorem 7.4, and hence for the structure, can be obtained in only a few cases. However, the more explicit properties of the field on the continuum determined in Theorems 7.5 and 7.6 give an intuitive lead in the building up of almost or wholly optimal solutions.

Consider a connected part of a plane structure that is not directly (i.e. externally) loaded, and contains both tension and compression members. By Theorems 7.5 and 7.6 these two sets of members then constitute a Hencky–Prandtl net. Examples of such nets are the rectangular and circular fan-shaped nets of Figure 7.6 and the families of conjugate spirals of Figure 7.7. In this last case the spirals have equations

$$r = r_1 e^{\kappa\theta}, \qquad r = r_2 e^{-\theta/\kappa}$$

in polar co-ordinates (r, θ) centred on their common origin. The constant κ is common and the two families of spirals are parameterised by r_1 and r_2. The optimal structure can sometimes be built up by piecing together sections of such nets. For example, the net around a load point may well be locally of the circular fan form, with origin at the load point.

As a specific example, consider the 'coat-hook' problem: that of communicating a downward vertical point-load of magnitude f (the coat) to a vertical foundation (the wall). If the need to cope with nonvertical loads is ignored, the problem is a two-dimensional one.

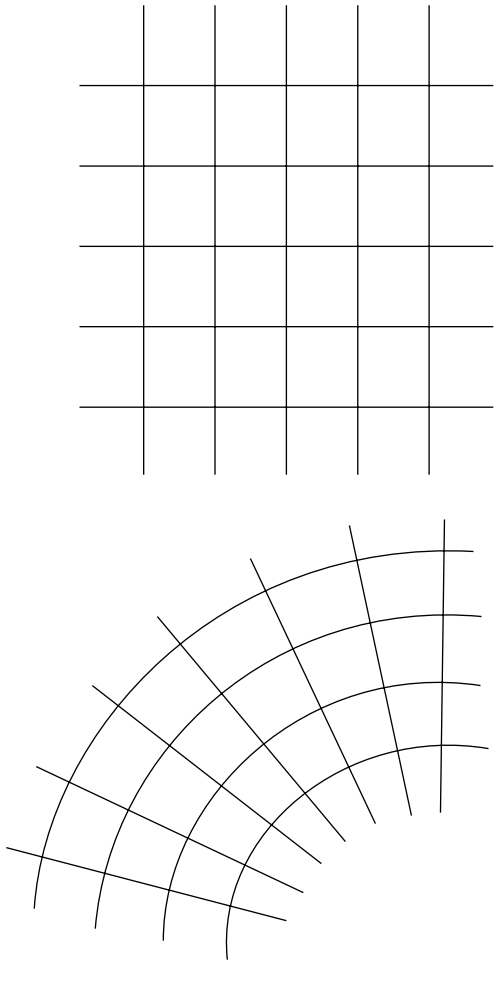


Fig. 7.6 Examples of Hencky–Prandtl nets: rectangular nets and circular fans. See also Fig. 7.7.

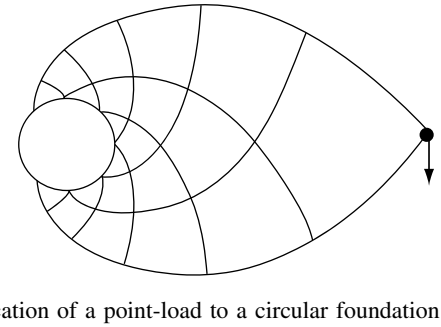


Fig. 7.7 The communication of a point-load to a circular foundation arc: the design of a crank. The Hencky–Prandtl net consists of two conjugate families of equi-angular spirals.

By Theorem 7.4, y has value zero on the wall, and its downward component at the load point should be as large as possible, consistent with (7.24). The obvious invariance and scale features of the problem make it clear that y has a constant direction, with magnitude depending only upon distance from the wall, and increasing linearly with this distance. That is, y is of the form

$$y(\xi) = v\xi^1 \tag{7.27}$$

for some fixed vector $v = (v^1, v^2)$. It follows then that

$$E = \begin{pmatrix} v^1 & \frac{1}{2}v^2 \\ \frac{1}{2}v^2 & 0 \end{pmatrix}.$$

Since the structure will certainly contain both tension and compression members, E must have eigenvalues q_+ and q_-; this condition determines v^1 and v^2. For simplicity, take the case $q_+ = -q_- = q$. One finds then that $v^1 = 0$ and $v^2 = \pm 2q$. Since y is to be downward directed we must take the negative option. This yields $2qfd$ for the maximal value of $D(y)$, where d is the distance of the load point from the wall.

Consider now to what structure this corresponds. The extreme values of $Dy/D\xi$ for the field (7.27) with $v^1 = 0$ are attained for directions at the constant angles $\pm 45°$ to the co-ordinate axes. The structure must thus take the form of a rectangular net with members in these directions; see Figure 7.8. One surmises that the portion of the net extending outside the two heavily drawn members will not be used; static considerations imply then that the members inside these two will carry no load, and so can also be deleted.

The presumption is, then, that the optimal framework will simplify to the two-member structure of Figure 7.8. This is indeed optimal. Resolution of forces at the load point shows that the members will be carrying longitudinal forces of magnitude $f/\sqrt{2}$ (the upper in tension, the lower in compression) and so will cost $qf/\sqrt{2}$ per unit length. But the total length of member is $(2\sqrt{2})d$, so the total cost is $2qfd$. This lower bound for the minimal cost coincides with the upper bound derived earlier, so the proposed structure is indeed optimal.

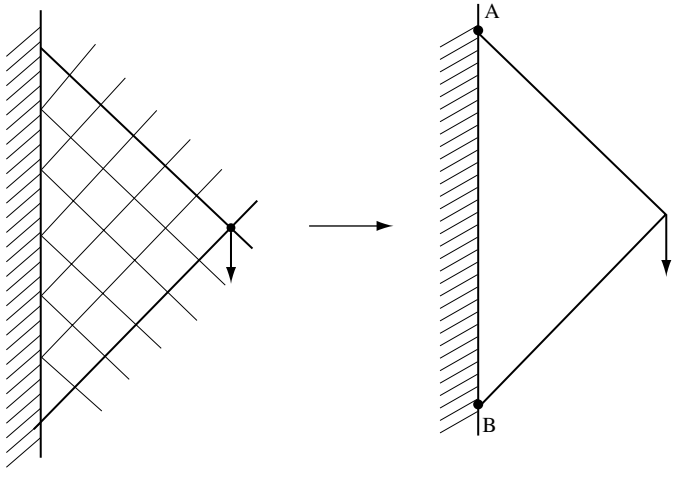


Fig. 7.8 The Hencky–Prandtl net and its simplification for the coat-hook problem.

If the foundation is of circular rather than linear cross-section, then the optimal net is no longer rectangular, but is made up of equi-angular spirals. This solves the problem of transferring a point load to the circumference of a circle, i.e. of designing a crank.

The net extends to infinity, but in fact the only part needed for the optimal design is the part bounded by the two spirals passing through the load point, heavily drawn in Figure 7.7. Since the boundaries of the net are now curved (a consequence of the curvature of the foundation arc) we cannot dispense with the inner members; these will be required to equilibrate stresses normal to the boundary. However, they will carry relatively light loads, except at points where the curvature of the boundary is relatively large (and the curves of the net are packed more densely). The outer members, which carry the principal load, would be represented by ribs, and the inner members would be represented by a web whose thickness increases towards the axle-end of the crank. The curvature and packing of the lines of the net will be greatest near the axle, and so the structure will be heavier and denser there. Indeed, if the radius of the axle is small enough relative to the length of the crank-arm the bounding ribs will meet there and wrap around the axle to form a boss to the crank.

Suppose one modifies the original coat-hook problem of Figure 7.4 by restricting the upper limit of the foundation arc to a point A' that is below the point A of Figure 7.8, used as the upper anchor-point of the unrestricted optimal design. Then the point A' is in a sense singular, because upon it must be concentrated all the load-carrying capacity that could otherwise have been distributed over the foundation arc above A'. One guesses that it might be the origin of a circular-fan component of the optimal net, and that the optimal net might have the composite rectangular/fan character of Figure 7.9(i). Deleting all members outside those passing through the load point, and deleting all internal members that meet a straight boundary (and so carry no load), one derives the simplified framework of Figure 7.9(ii). This consist of three straight members, a circular-arc boundary member CD and a continuum of radial members in tension between $A'C$ and $A'D$. This framework is statically determined and optimal.

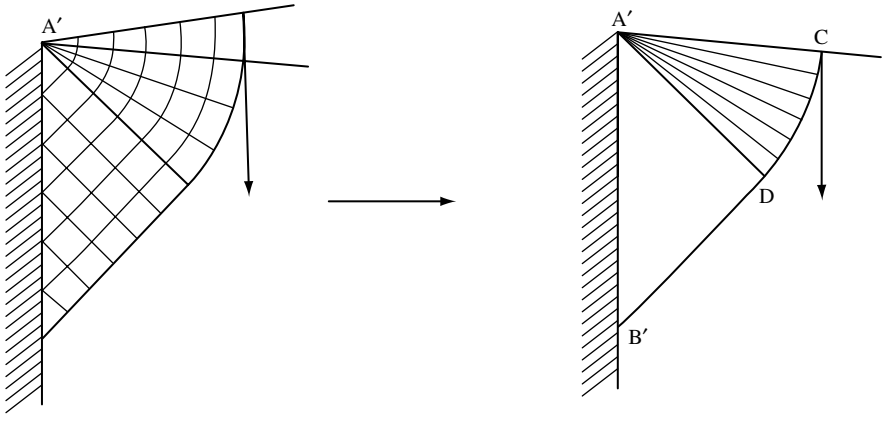


Fig. 7.9 The Hencky–Prandtl net and its simplification for the coat-hook problem in the case when the foundation arc does not reach as high as the point A of Fig. 7.8.

A further restriction in the same direction is to restrict the foundation to just a pair of arbitrarily chosen points, as in Figure 7.10. This yields the design of a cantilever framework. A graphical method that gives an approximately Michell structure is illustrated in the figure, and simplified to the actual structure in Figure 7.11. Assume that the framework will be locally fan-shaped in the neighbourhood of the two foundation points. Draw then upon each, as centre, a circular arc of such radius that the two arcs meet at right angles. The net is then continued out from this compound arc by choosing a number of equally spaced points along each circular arc (the significant point is not the equality in spacing, but equality in increment of the angle ϕ) and from each continuing the radius vector. Where these vectors of the two systems meet, the net is continued by observing the right-angle requirements marked by a dot in the figure.

The net thus generated is a discrete approximation (with straight members) to a Hencky–Prandtl net. Complete orthogonality at all nodes is, of course, not possible with straight members and a nonrectangular net. However, incorporation of the right-angle requirement supplies an approximate mutual orthogonality, and the fact, easily proved, that every quadrilateral mesh of the net has the same geometry (i.e. the same set of corner angles) supplies an approximate equivalent of the constant ϕ-increment property.

The net is then completed, and adapted to the case of a given load point, by drawing in the radial members from the two foundation points, approximating the elements of arc by

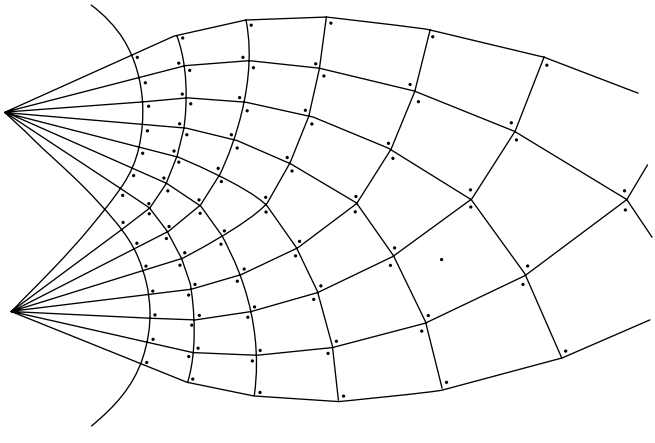


Fig. 7.10 The coat-hook problem with two assigned foundation points, leading to the graphical determination of an optimal cantilever structure. Right angles are marked with a dot.

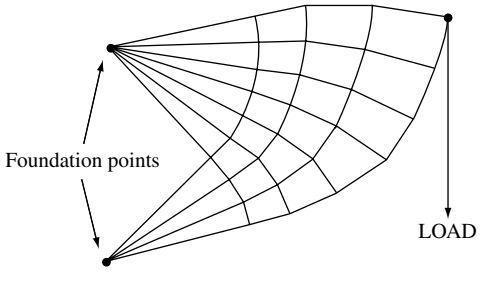


Fig. 7.11 The simplification of Fig. 7.10 to the actual cantilever structure.

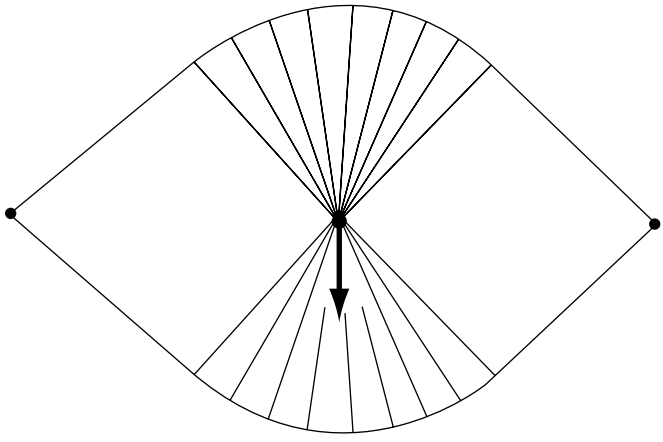


Fig. 7.12 The Michell structure for acceptance of a vertical load midway between two foundation points.

chords, and deleting all members outside the generators passing through the load point. The result is a statically determined structure, in which the loads (and so cross-sections) can be determined recursively from the load point. The constant geometry at each node eases this calculation greatly.

A very special case of such a structure is the Michell solution for the support of a vertical load midway between two level foundation points; see Figure 7.12. This appeals to a combination of fan and rectangular dual fields. It is optimal, but is unstable to lateral perturbations of the load; see Chapter 9.

7.7 The shaping of bone structure

Nowhere does the Michell analysis find a more graphic confirmation and illustration than in the observed internal structure of bones. We shall take the human example, although these observations doubtless hold for most animals with an internal skeleton. Bones are regarded as the most durable and fixed part of the anatomy, and yet in fact they exist in a dynamical state of permanent adaptation, with constant bone loss being compensated by bone gain at stress points. The phenomenon of bone loss is rapidly evident in astronauts experiencing zero gravity, or indeed in bedridden patients. It is a benevolent mechanism, however, which permits a continuing adaptation to the varying physical demands of life. In the following we shall refer repeatedly to the text by Carter and Beaupré (2001). Anatomical entities introduced in inverted commas can be regarded as defined by the functions ascribed to them.

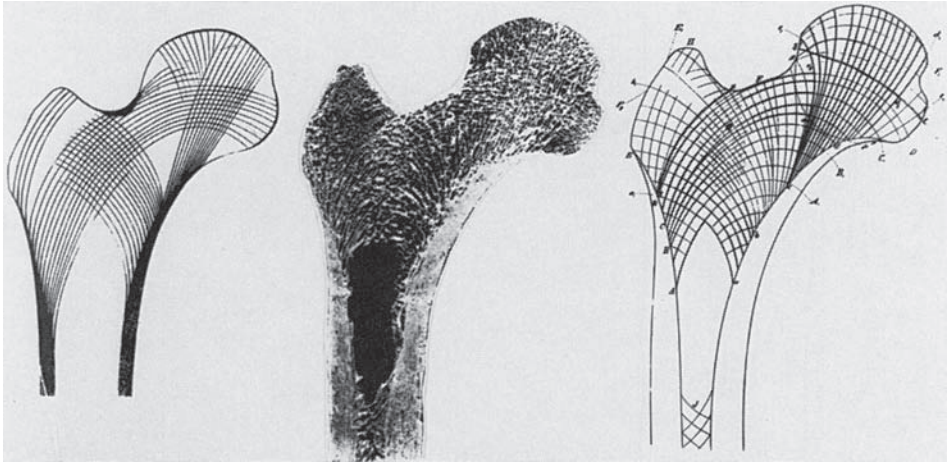
In humans, bone begins as cartilage, which is then calcified, but is then subject to resorption by the ‘chondroclasts’. In the small erosion bays thus created the ‘osteoblasts’ immediately form small rod-like structures of genuine bone, termed trabiculae. These are the line elements of an emerging structure; they align themselves in such a way as to meet whatever stresses are experienced locally. One then sees the importance of such stresses in directing the pattern of renewal. The structure thus generated is most evident

in the cancellous (spongy, open-textured) bone that forms the interior of a number of skeletal features; the cortical (surface) bone is dense and less revealing.

The effect of this organising principle is graphically demonstrated in the cross-sections of the upper human femur (thighbone) shown in Figure 7.13. The third sketch gives in clearer form the structure observable in the preceding photograph. The existence of the mutually orthogonal tension and compression lines formed by the trabiculae is evident. It is even evident in the sketch made a century earlier, by von Meyer in 1867, long before Michell had propounded his ideas. The structure is exactly as the Michell analysis might have predicted. The publication from which these figures are taken (with the kind permission of the copyright-holders, Springer-Verlag) is the text *The Law of Bone Remodelling* by Wolff (1986). The German original was published as early as 1892, and was the pioneer publication in the area. It speaks for itself that the text was republished after a lapse of 94 years; both Wolff and Michell appear as men of remarkable prescience and ability.

The distinguishing feature of the upper femur is the head, to which is attached the ball of the weight-bearing ball-and-socket hip joint. The ball must accept this weight, but is offset laterally from the shaft of the femur. The tension and compression lines which are apparent in Figure 7.13 very much resemble those of the cantilever structure of Figure 7.11. They differ in that the upper foundation point (which anchors the tension members) is now replaced by a foundation arc along the cortical surface opposed to the ball, and the lower foundation point (which anchors the compression members) is now replaced by a foundation arc along the cortical surface below the ball.

Carter and Beaupré express this more anatomically by seeing the compression lines as aligned ‘from the superior contact surface to the calcar region of the medial cortex’ and the tension lines as ‘a secondary arcuate system of trabeculae arching from the



(i) (ii) (iii)

Fig. 7.13 (i) von Meyer’s 1867 drawing illustrating the bone architecture at the head of the femur (thighbone). (ii) Wolff’s photograph of a vertical cross-section. (iii) Wolff’s schematic representation of the bone architecture of the section shown. Reproduced from Wolff (1986) with the kind permission of Springer Science and Business Media.

infero-medial joint surface through the superior neck and into the lateral metaphyseal region'.

There is a substantial literature on bone remodelling, addressing both anatomical observation and computer simulation. The text by Carter and Beaupré contains an extensive bibiography, to which one might add Bagge (2000) and Huiskes (2000).

A point often made is that the observed structure is not exactly optimal for any given static load. One contributing reason for this might be that bones serve other biological purposes and are subject to other biological constraints. For example, their interiors must be permeable to the passage of fluids. However, what is certainly a much weightier consideration is that they must cope with a variety of mechanical loads, and indeed with dynamic loads. We know from Chapter 4 what a difference this can make to the optimal structure. The point is taken up in Chapter 9.

8

Computational experience of evolutionary algorithms

The subject of structural design is of course a long-established and sophisticated one, with an enormous literature. Explicit optimisation was however long regarded as too ambitious a target, whose attempt would entrain risky over-simplifications and the neglect of the wealth of intuition gained by practical experience. However, the advent of powerful computers has made the ambition a realisable one, and revived interest in the class of ideas revealed by Michell.

These developments have now generated a substantial literature in their turn; see Section 8.4 for references. We shall draw particularly, however, on the recent text by Bendsøe and Sigmund (2003). These authors have cast the optimality equations into a form suitable for efficient computation, and have acquired a vast computational experience on a range of practical problems, many of them highly sophisticated, to the point that their programmes virtually deliver production-ready working drawings. In many respects their work may be said to lead theoretical insight rather than follow it.

We refer to the Bendsøe–Sigmund text often enough that it is convenient to use the abbreviation BS. The text by Xie and Stevens (1997) uses somewhat different techniques, and so supplies interesting comparisons; we shall refer to it as XS.

8.1 Solid materials

BS consider largely the case of a linear, elastic medium (so that $\alpha = \beta = 2$ and the medium follows the same linear Hooke's law in tension and in compression). Terms like 'work' and 'strain energy' then have their conventional mechanical interpretations. They also consider initially the case of a medium that is isotropic, except insofar as its density may vary in space. The cost function we set up in the dual formulation of Section 7.4 could then be written

$$D(y, \mathbf{D}) = D(y) - C_1(y, \rho) + C_2(\rho), \tag{8.1}$$

where $\mathbf{D}$ indicates design and

$$D(y) = \int y f \mu(\mathrm{d}\xi)$$

is twice the work done by the external force f as it deforms the structure,

$$C_1(y, \rho) = \frac{\kappa}{2} \int \int \rho(\xi) |u^\top E(y) u|^2 \mathrm{d}\xi \mathrm{d}u \tag{8.2}$$

is the strain energy stored in the structure, and

$$C_2(\rho) = \gamma \int \rho(\xi) \mathrm{d}\xi \tag{8.3}$$

is the material cost of the structure. Integrals with respect to ξ are over the design space $\mathcal{D}$, and those with respect to u in (8.2) constitute a uniform integration over the spherical shell $u^\top u = 1$. $E(y)$ is the matrix of linearised strains

$$e_{ij}(y) = \frac{1}{2}\left(\frac{\partial y_i}{\partial \xi_j} + \frac{\partial y_j}{\partial \xi_i}\right)$$

already introduced in Section 7.4.

Expression (8.1) is to be maximised with respect to y and minimised with respect to ρ. The maximum with respect to y of $D(y) - C_1(y, \rho)$ would be termed the 'compliance' in this context, so the design aim is to minimise the sum of compliance and material costs.

Note that expressions (8.2) and (8.3) are in terms of an unoriented density $\rho(\xi)$ rather than the oriented density $\rho(\xi, u)$ that appeared in Chapters 2, 4 and 5 as the natural analogue of the explicit paths of the discrete case. Later, when one allows the option of the anisotropic materials of modern materials science, matter will be both oriented and with a microstructure which gives it less than full density.

BS characterise the extremal problem very nearly in the terms above. The expression (8.2) that we have written for $C_1(y, \rho)$ they write as the quadratic form $(1/2)Q(y, y, \rho)$ where the corresponding bilinear form is

$$Q(y, z, \rho) = \int \rho(\xi)^v F_0^{ijkl} e_{ij}(y) e_{kl}(z) \mathrm{d}\xi. \tag{8.4}$$

Here F is the stiffness tensor for unit density of material. (Recall the tensor convention, that if an index such as i appears in both superscript and subscript positions then a summation over that variable is understood.) Our expression (8.2) then implies an evaluation of F. There is one clear difference, however, in that the density $\rho(\xi)$, appearing under the integral to the first power in (8.2), appears to the vth power in (8.4). The reason for this is a heuristic but significant one, which we shall explain in a moment. BS require that the field y should satisfy the equation system

$$Q(y, z, \rho) = D(z)$$

for all z. This is just the stationarity condition expressing the demand that y should maximise the quadratic form (8.1) in y. The minimised value of the quadratic form is just $(1/2)D(y)$, with the maximising value of y substituted, and BS demand that $\rho(\xi)$ should minimise this expression, conditional on a given mass of material and fixed upper and lower bounds on density. They thus limit material by quantity rather than by price γ, but there is of course a Lagrangian equivalence. The construction materials one is using, such as steel, set an upper limit on density, but could occur at a lower density if they were employed in some expanded form. However, an almost infinitesimal lower bound is set on density to preclude the trouble foreshadowed in Chapter 4. This is, that one might

never otherwise discover that addition of material could be advantageous at a certain point if it has never been tested there.

BS speak of structures for which the material is present at maximal or minimal density as 'black-and-white' structures. Structures at which the material appears at some intermediate density are termed 'grey'. BS wish to discourage 'greyness' (which reflects some expanded version of the material, possibly difficult to fabricate), and find that they can do so by giving v a value of 3 or so. A change of the variable ρ shows that this is equivalent to replacing ρ by $\rho^{1/v}$ in the material cost expression (8.3). We have already seen in Chapter 5 that this represents a distaste for an all-pervading grid of links, and encourages trunking, which would amount in the present case to a consolidation of material. That is, there is a concentration of flow (and so of stress, in the present case) on to a few large members rather than many small ones. On the other hand, if the load is finely distributed (or variable), then there must be something in the nature of a graduated fine structure that communicates the load to these heavier members.

Actually, intermediate values of ρ can occur very easily if, for example, one is making the approximation of giving a two-dimensional treatment of a 'flat' three-dimensional object. In this case varying ρ can represent varying thickness, and one is making the approximation of neglecting stresses and strains in the direction of thickness. We have already seen this in the two-dimensional treatment of a crank, Fig. 7.6. One would regard the spacing of the slip-lines as inversely proportional to density, and so we see the tight spacing around the axle as indicating a boss and the increasing spacing with increasing radius as indicating a web of diminishing thickness. (Although the fact that the bounding curves in the figure must replace an extended web implies the presence of rim there.) We shall see in the next chapter that 'greyness' can also come about if the structure must cope with variable loads, when to replace several starkly alternative structures by one incorporating a region of genuinely low-density material may be the most economic way of accepting all stresses.

If one modifies the compliance measure by adopting the BS form (8.4) then the optimality condition for the distribution of material would be that $\rho(\xi)$ should take the prescribed minimum or maximum value according as the value of strain $F_0^{ijkl}e_{ij}(y)e_{kl}(y)$ at ξ falls above or below some constant. This constant would be proportional to the unit cost γ of material in our treatment, but would be a Lagrange multiplier for the material constraint in the BS treatment. BS adopt a finite element analysis, in that they divide the design space into small cells or 'finite elements' within each of which ρ is presumed to be constant. They then adopt a version of the familiar evolutionary algorithm, in which material is shifted as suggested by infraction of the optimality conditions.

BS refer to their method as SIMP (solid isotropic material with penalisation), but the complete package for application has a considerable degree of refinement and precaution built into it. Calculations can be greatly accelerated if one adopts some mathematical programming techniques locally. Anomalous 'chess-board' patterns can appear in the solution, because the finite-element approximation attributes a greater strength to such patterns than they in fact have. Special precautions must be taken against this. Also, solution can fail because the finite-element approximation is struggling to represent a structure with infinitely dense detail. This manifests itself in a solution which does not converge as the mesh (the cell representation of the design space) becomes finer.

The phenomenon need not be a phantom one, however: it can correspond to the need to accept a continuously distributed load.

The approach taken to optimal design by Xie and Steven (1997) is deflatingly simple. They evaluate a scalar expression σ_i for stress (the *von Neumann stress*) for every element i of a finite-element approximation, and then wholly delete those elements for which the ratio σ_i/σ_{max} falls below a prescribed threshold value R. Here $\sigma_{max} = \max_i \sigma_i$. The process continues, with redetermined values of the σ_i, until it stabilises. The value of R is initially taken quite low, and then is progressively raised until it is such that the total amount of material in the structure has fallen to the prescribed limit. The authors refer to their approach as ESO (evolutionary structural optimisation).

The approach indeed has a theoretical basis. We saw in Theorem 7.1 that every link of the optimal network should have a rating proportional to the stress it carried, and XS are invoking the continuum analogue of this condition. However, the complete continuum analogue would require that the designer should be free to vary both the density and the directional properties of the material of the structure, and it is only with this freedom that one can achieve complete uniformity of a continuum measure of absolute stress. XS assume, at least initially, that one is limited to an isotropic material of given density, and so can hope only to minimise stress variation in the structure, rather than eliminate it. BS quote Zhou and Rozvany (2001), cautioning against the neglect of gradient information.

8.2 Examples

Both BS and XS confirm that their calculations converge on to the Michell structure in cases where this is known (except, of course, that bars of infinitesimal width and infinite density are replaced by bars of positive width and finite density). So, XS confirm the two-bar 'coat-hook' structure of Figure 7.8 and its constrained version Figure 7.9, and both sets of authors confirm the Michell structure of Figure 7.12 for the support of a vertical point load midway between two foundation points. BS confirm the Michell construction of an optimal crank generated from equi-angular spiral slip-lines (Fig. 7.7), except that their 'black-and-white' condition results in a lattice structure rather than a continuous one. The corresponding XS structure is generally consistent, although it does differ in some detail from the Michell structure.

BS consider the design of a long cantilever beam, required to support a vertical load at its tip, and deduce the structure of Figure 8.1, for varying costs of material. The design is restricted to a long, narrow volume, which makes the Michell character of the solution less obvious. The Michell analysis predicts cycloidal slip-lines, with some effect from end-conditions (see Hemp, 1973, pp. 95–7), and the trusses produced by BS do approximate these very fairly. Tension and compression members that are tied to the same side of the structure do indeed meet at right angles (except at the tip, which is a load point). XS produce a solid solution for the same problem; the evolution of shape and stress contours during the calculation, presented in Figure 8.2, is interesting. (Note that only the upper transverse half of the structure is drawn.)

Querin (1997, and reported in XS) uses ESO to determine the optimal shape for an object hanging from a stalk by its own weight, and deduces a form which Newton would have recognised: see Figure 8.3.

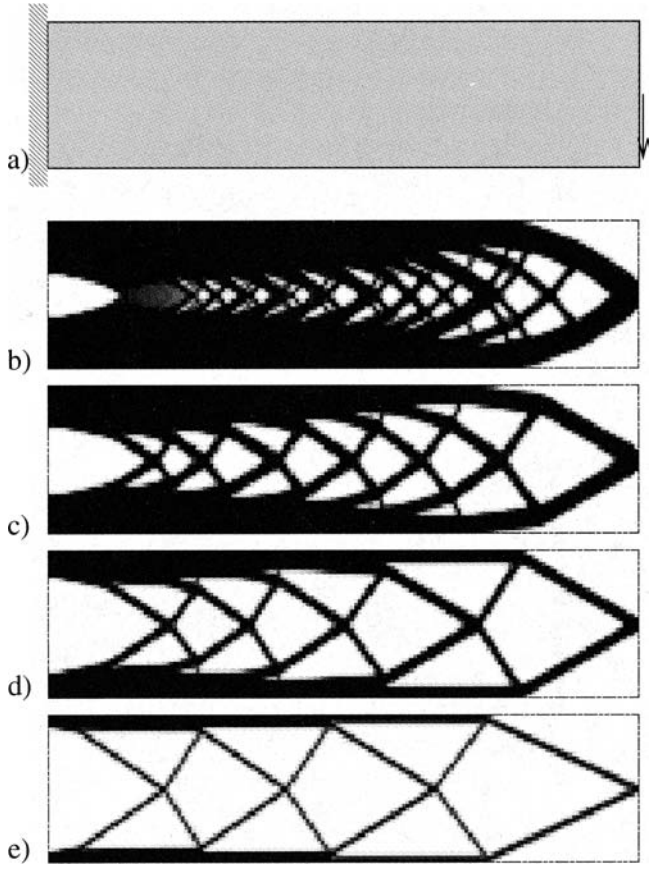


Fig. 8.1 The optimal form of a long cantilever beam (loaded at the tip) for increasing cost of material. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

A classic problem has been the optimisation of the MBB beam, a beam that carries the fuselage floor of an Airbus passenger plane. It is supported along its length and carries a central load. This is a simple problem, but it is one for which the aircraft application demands a sophisticated optimisation. BS deduce the structure of Figure 8.4, which indeed demonstrates the mutual orthogonality of tension and compression members. An interesting point is the degree to which the design actually changes as the mesh size decreases, not simply by the addition of refinements, but by the distribution of load over ever slighter and more numerous members as the opportunity arises to do so. XS deduce a broadly similar structure, but one restricted in its detail by a relative coarseness of its mesh.

BS deduce optimal designs for a number of genuinely three-dimensional and rather special problems. Both sets of authors also give a detailed consideration of other design criteria (e.g. optimisation of frequency response, wave propagation or stability against buckling). However, the two general considerations which loom ahead in our context are those of anisotropic materials and variable load.

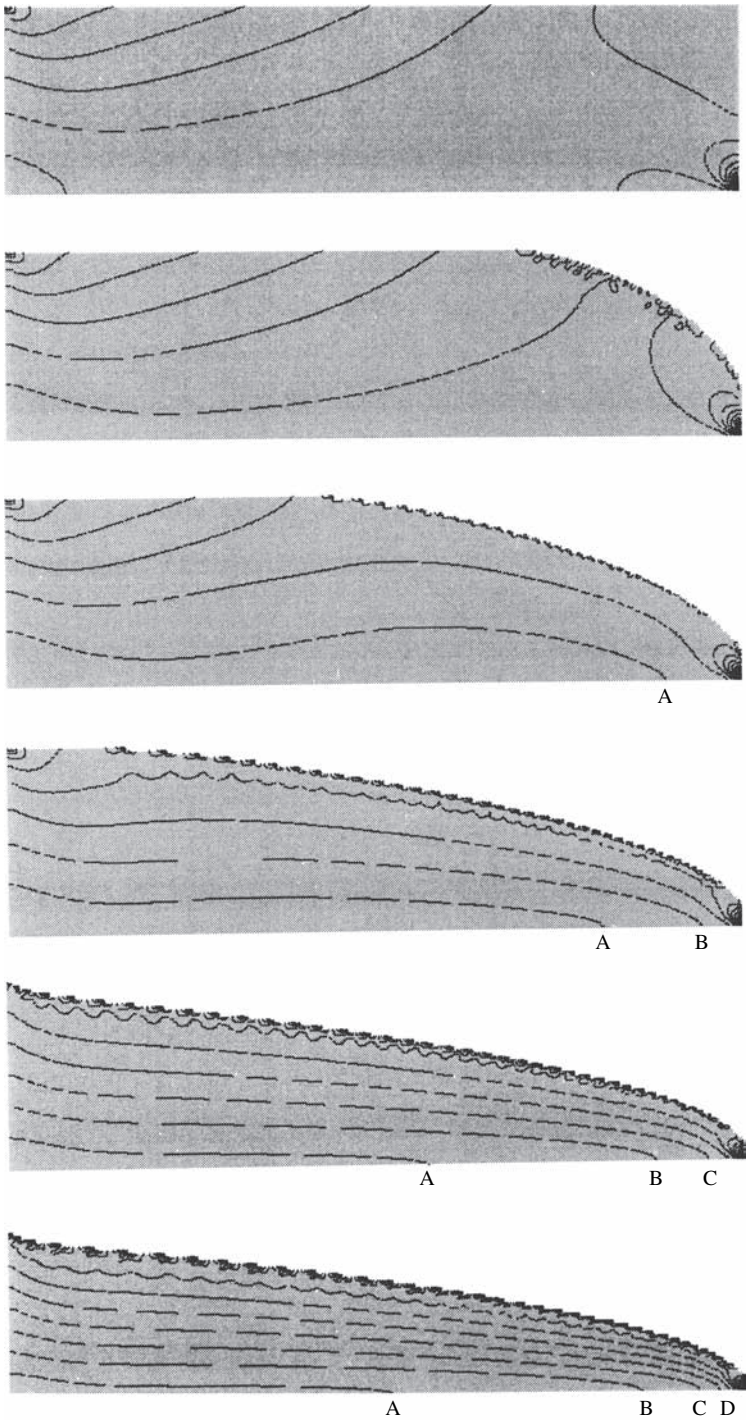


Fig. 8.2 Evolution towards optimality of a solid version of the cantilever problem, with the stress contours dotted. Only the upper half of the beam is illustrated. Reproduced from Xie and Steven (1997) with the kind permission of Springer Science and Business Media.

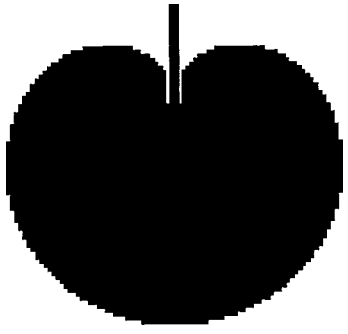


Fig. 8.3 Querin's solution for the optimal shape of a solid object hanging from a stalk under its own weight. Reproduced from Xie and Steven (1997) with the kind permission of Springer Science and Business Media.

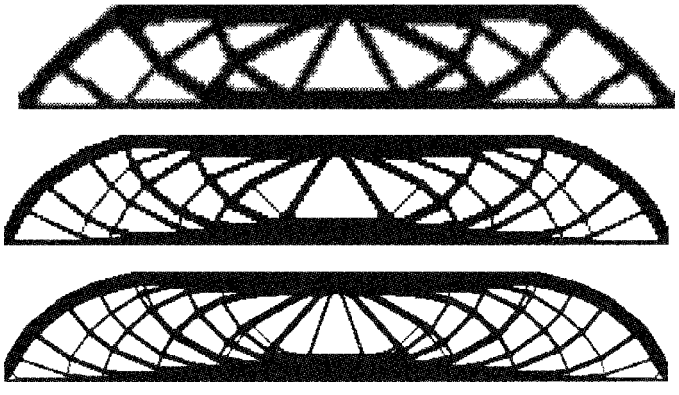


Fig. 8.4 The optimal form of the MBB beam for increasing refinement of the discrete approximation. Note that new members appear as the computation is refined. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

8.3 Expanded materials

We can allow anisotropy in the structure immediately by the modification of expressions (8.2) and (8.3) for strain energy and material cost to

$$C_1(y,\rho) = \int\int (\kappa\rho/\beta)|u^\top E(y)u|^\beta \mathrm{d}\xi \mathrm{d}u \tag{8.5}$$

and

$$C_2(\rho) = \int\int \gamma(\xi,u)\rho(\xi,u)\mathrm{d}\xi\mathrm{d}u. \tag{8.6}$$

Here the quantities ρ, κ, β and γ (representing respectively material density, material strength, rate of stiffness variation with load and unit cost) will depend upon the material chosen at position ξ and orientation u, and so will in turn depend upon these variables. However, if we assume the linear-elastic case $\beta = 2$ and normalise κ to unity by defining

new variables $\rho' = \kappa\rho$ and $\gamma' = \gamma/\kappa$ (and immediately dropping the primes) then (8.5) becomes

$$C_1(y,\rho) = \frac{1}{2}\int\int \rho(\xi,u)|u^\top E(y)u|^2 \mathrm{d}\xi \mathrm{d}u \tag{8.7}$$

and (8.6) remains as it is.

One could allow for an optimisation over materials, but under present assumptions one would use only a single material for members in tension (that whose price/effectiveness ratio was least) and ditto for members in compression. If for simplicity we assume the same linear rule in compression and tension then one will use only a single material. However, expression (8.7) implies that we can now orient the material. It also follows, as in Section 7.4, that for optimality it will be oriented along the axes of strain: those directions u maximising the virtual strain $e_{vir} = |u^\top E(y)u|$. The strain field $y(\xi)$ will then again be a Hencky–Prandtl field, as in Chapter 7.

The assumption that one can orient the material implies some degree of microstructure. We must now also face the possibility (and the necessity, if a solution is to be found in the class of structures envisaged) that there may be a call for material at less than full density. This also implies some degree of microstructure, if the material is intrinsically solid. Special techniques are employed in the literature to represent such forms of material analytically. One of these is the technique of 'homogenisation'. This is expounded at length in chapter 3 of Bendsøe and Sigmund (2003). To sketch the idea, assume a discrete structure whose strain energy under a vector of strains y can be written as a quadratic form $y^\top Ky$, for some positive definite matrix K. Suppose now that the structure consists of a periodic array of small identical cells in physical space, these cells being separated by the embedding material. Then the energy functional can be written

$$y^\top Ky = y_0^\top K_{00} y_0 + 2\sum_{i>0} y_0^\top K_{0i} y_i + \sum_{i>0} y_i^\top L y_i,$$

where y_i is the subvector of strains in cell i for $i > 0$ and of strains in the embedding material for $i = 0$. The cell strain matrix, which we might have written as K_{ii}, we have written as L. This is to emphasise that all the cells have the same internal structure, which we assume given. There are no cross-terms, because the cells are separated. If we now minimise out the cell strains then we are left with the strains in the embedding material, of energy

$$y_0^\top [K_{00} - \sum_{i>0} K_{0i} L^{-1} K_{i0}] y_0,$$

where $K_{i0} = K_{0i}^\top$. In principle, one can now take this as the energy functional for a voidless material.

There is a very technical and still open-ended literature on these topics; even more so if one begins to consider composite materials. However, we can note one question and conclusion of great interest. What is the cell structure (as specified by the matrix L or its continuum equivalent) that yields the expanded macro-material of greatest strength (as measured by its bulk modulus) for a given density? There are known bounds on this quantity, the Hashin–Shtrikman bounds, and there are several cell configurations that yield a bulk modulus within a few per cent of this bound.

Bendsøe and Sigmund give four known such configurations for the two-dimensional case under the conditions that the macro-structure is required to be isotropic. Three of these are strictly periodic in character. In Figure 8.5 we illustrate the fourth, which appeals as a 'biological' solution, and obviously transfers to higher dimensions. In three dimensions it is an assembly of 'coated spheres' – empty adhering spheres of randomly varying size which can be packed to leave very little interstitial void. It is strongly reminiscent of the cancellous bone, mentioned at the end of Chapter 7, which fills some of the internal voids of major bones near a joint. For that purpose it would have to be modified to acquire a degree of porosity, however. Fluid circulation must be allowed if bone maintenance and the more general functions of bone marrow are to be possible.

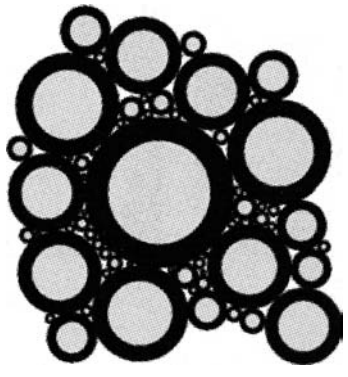


Fig. 8.5 The coated sphere assemblage. One of a number of known microstructures that very nearly maximise the bulk modulus of the material for a prescribed density, and the only one that is nonperiodic. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

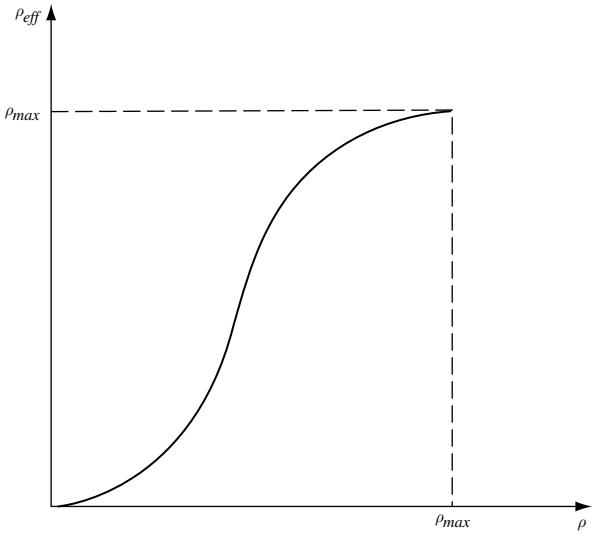


Fig. 8.6 A conjectured relation between actual density ρ and effective density ρ_{eff}.

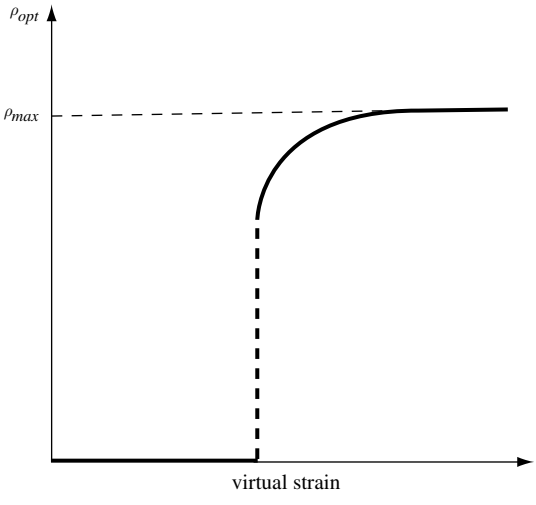


Fig. 8.7 The form that the relation of Fig. 8.6 implies for the dependence of optimal density ρ_{opt} upon virtual strain.

The fact that the density ρ appears linearly in integral (8.7) implies that the effective strength (however measured) of the expanded structure is proportional to its density. This is a consequence of the fact that we reached the continuum as a limit of discrete frameworks, which showed a simple scaling. It seems likely, however, that the 'effective density' ρ_{eff} is a sigmoid function $R(\rho)$ of actual density, as illustrated in Figure 8.6. The optimal value ρ_{opt} of ρ is then that maximising $R(\rho)(\text{virtual strain})^2 - \gamma\rho$, subject to $\rho \leq \rho_{max}$, and so shows the dependence upon virtual strain $|u^\top E(y)u|$ indicated in Figure 8.7.

8.4 The literature on structural optimisation

There has of course been an empirical interest in structural optimisation throughout history, and attempts at mathematical analysis whenever this seemed feasible. However, only in more recent years has the ambition of optimising over all feasible distributions of matter over the continuum design space seemed realistic. The first, sweeping advance in this direction was that by Michell (1904), and the highest current reach of the powerful wave that has developed since is to be seen in the text by Bendsøe and Sigmund (2003).

The latter two authors see this wave as having been initiated by workers such as Wasiutynski (1960), Prager and Taylor (1968), Taylor (1969) and Masur (1970), continued with the development of algorithms by, for example, Taylor and Rossow (1977), Cheng and Olhoff (1982), Bendsøe (1986), Bendsøe and Kikuchi (1988), Rozvany and Zhou (1991) and Rozvany *et al.* (1994), followed by mathematical treatments such as those by Levy (1991), Svanberg (1994), Cheng and Pedersen (1997) and Toader (1997).

9

Structure design for variable load

In the structural engineering literature the term 'multiple load' is often used in preference to 'varying load'. We shall continue to use the second term, however, since one is supposing that different loading patterns are used at different times rather than that these patterns should be superimposed. Note the distinction between the use of roman E for the expectation operator and italic E for the strain matrix.

The formal generalisation to the varying load case is immediate if we take the problem in the dual continuum form of Chapter 4. We can also formally include the option of various materials at the same time. The criterion function (8.1) now becomes

$$D(y, \mathbf{D}) = \mathrm{E}[D(y) - C_1(y, \rho) + C_2(\rho)], \tag{9.1}$$

where the expectation is over regimes ω on which f (and consequently the strain y) is dependent. The components of this expression are

$$D(y) = \int fy\mu(\mathrm{d}\xi)$$

(twice the work done by external forces), the strain potential

$$C_1(y, \rho) = \int \int (\kappa\rho/\beta)|u^\top E(y)u|^\beta \mathrm{d}\xi \mathrm{d}u$$

and the cost of materials

$$C_2(\rho) = \int \int \gamma\rho \mathrm{d}\xi \mathrm{d}u.$$

It is understood that the ξ-integration is over the design space $\mathcal{D}$ in all cases, and that ρ is a function of ξ and u. It is also supposed that the parameters κ, β and γ (representing respectively material strength, rate of stiffness variation with load and unit cost) may be so too, in that these will depend upon the particular material chosen. This choice is part of the design, and will depend upon conditions prevailing at a given co-ordinate and orientation.

Expression (9.1) must first be maximised with respect to y as a function of ω and then minimised with respect to the design as specified by the allocation of density and type of materials. As in Chapter 2, these operations may be commuted. However, the ω-dependent displacement field y is now a random field over which averages must be taken, and the Michell theory of Chapter 7 no longer applies. For example, it is no longer true in general

that tension and compression members must meet orthogonally in the optimal design. The different structures that would be individually optimal for the individual alternative load patterns must effectively be condensed to a single structure, and it is of considerable interest that this condensation often results in nonsolid designs.

The variable-load formulation is salutory in that it can expose weaknesses in designs optimised for a single load pattern. We have already noted that the Michell structure of Fig. 7.12 will fail if the load, supposed vertical, in fact has a lateral component; see Thompson and Hunt (1973, p. 292).

The passage from the scalar potentials of Chapter 1 to the vector potentials of the structure problem has a profound effect, in that the simplification induced by the combination of seminvariant scaling and optimisation of ratings no longer implies the disconcerting collapse described in Theorem 1.4. The parallel assertion would be Theorem 7.1, but this is different enough that one is left with a sophisticated and highly structured problem to solve. There are however parts of the variable-load treatment of Chapter 4 which transfer en bloc. For example, if one knows that the optimal structure is discrete, then Theorem 4.1, with its assertions concerning optimality of ratings and the positioning of internal nodes, has an immediate analogue. So too has its evolutionary version, set out in Section 4.7.

The dual assertions of Theorem 4.4 now take the form: that the ω-dependent dual field $y(\xi, \omega)$ is to be chosen to minimise

$$\mathrm{E}[D(y)] = \mathrm{E} \int y f \mu(\mathrm{d}\xi) \tag{9.2}$$

subject to

$$\mathrm{E}|u^{\top} E(y) u|^{\beta} \leq \beta \gamma \tag{9.3}$$

for all relevant co-ordinates ξ and directions (unit vectors) u. Furthermore, that line elements of the optimal structure can be laid down only where equality holds in (9.3). In the evolutionary/conjectured passage to an optimal structure one will add material where $(\beta\gamma)^{-1}\mathrm{E}|u^{\top}E(y)u|^{\beta}$ is greatest, paying attention to which material is to be used for given circumstances.

9.1 Return to the coat-hook

It is hard to state general conclusions for an arbitrary set of load regimes. However, as in Chapter 4, the effective aim of an optimisation is to decrease the relative variability of stress in the structure's main members. In some cases this will lead to the sharing of load between members; in some to the concentration of alternative loads on to larger compromise members. In the next section we shall collect the results of some of Bendsøe and Sigmund's numerical optimisations. In this we shall see what can be said of the simplest nontrivial problem: the coat-hook problem of Section 7.6 in the case when the point load applied varies about the vertical.

Suppose in fact that the load can be applied in directions varying by $\pm\theta$ from the vertical, as illustrated in Figure 9.1. The two alternatives are supposed equally probable. For simplicity we suppose the materials used of identical properties and cost in tension or compression. The dual fields in the individual single-load cases would simply tip the

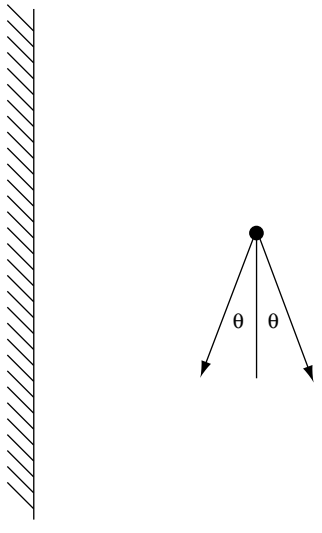


Fig. 9.1 The coat-hook problem with two load patterns.

two-bar solution of Figure 7.8 by the amount $\pm\theta$ (relative to the fixed vertical foundation), the rectangular dual fields for the solution being rotated similarly. Let us take these two rotated fields as test fields for the two random alternatives, as we did for the example of Figure 4.5.

If we take axes parallel to the slip-lines of the rectangular field that is optimal in the case of a vertical load, then the strain matrix E will take the simple form

$$E = E_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

to within a proportionality constant that we have supposed normalised to unity. If the field is rotated by an amount θ then we shall have

$$E = H^\top E_0 H,$$

where the orthogonal matrix H achieves this two-dimensional rotation.

Suppose that u is a unit vector in the direction χ, in that it has components $\cos\chi$ and $\sin\chi$. Then Hu is a unit vector in the direction $\chi+\theta$, so that

$$u^\top E u = \cos^2(\chi+\theta) - \sin^2(\chi+\theta) = \cos(2\chi+2\theta).$$

The measure of expected strain in direction χ is thus

$$\mathrm{E}|u^\top E u|^\beta = (1/2)[|\cos(2\chi+2\theta)|^\beta + |\cos(2\chi-2\theta)|^\beta]. \tag{9.4}$$

We look for directions χ in which expression (9.4) is maximal; one would expect these to be the directions in which a link should be laid. For β small (i.e. for materials whose stiffness decreases with increasing strain) expression (9.4) will be maximal for χ equal

to zero or $\pi/2$, when the two expressions in the square brackets are equal. (At least this is true for θ less than $\pi/8 = 22.5°$, when both the cosine functions in (9.4) are positive.) That is, the members are aligned with the slip-lines of the unrotated field, and we are back to the two-bar solution of Figure 7.8. In other words, it is better to have a single substantial member (one each in tension and in compression) that takes the whole strain, and takes a compromise position between the two directions of strain. For β large (i.e. for materials whose stiffness increases with strain) the reverse is the case. For some value of β the maximum of expression (9.4) at the central value $\chi = 0$ will split into two symmetrically placed maxima, which will move to $\chi = \pm\theta$ as β becomes indefinitely large. There will also be maximising values displaced from these by $\pi/2$; one pair of solutions is for members in tension, the other for members in compression. The solution suggested, then, is that of Figure 9.2, in which the members divide, the better to meet the stresses of individual regimes. The larger β, the closer the structure is to two separate structures, one optimal for one load pattern and the other for the other. The supposed good behaviour of the material under stress allows one to make this duplication with rotated (and somewhat lightened) versions of the single-load structure.

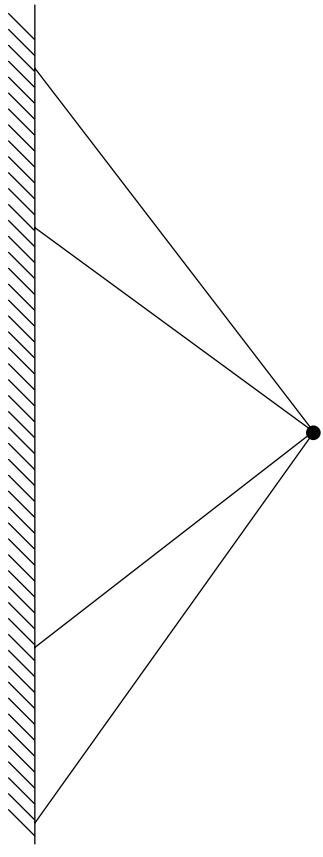


Fig. 9.2 The deviation of the optimal solution from the simple two-bar structure that is suggested if β or θ are large enough (i.e. if the material is sufficiently stiff or if the two alternative loadings are sufficiently different).

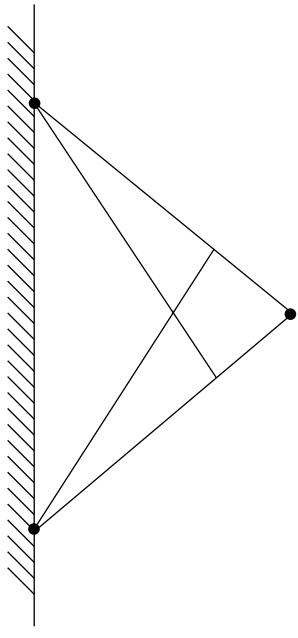


Fig. 9.3 The modification of the solution of Fig. 9.2 if the foundation is restricted.

If θ were large enough to make the rotations of the basic solution unacceptably large, one would restrict the foundation and take a solution of the form of Figure 9.3, verging on to a cantilever structure. Indeed, the problems of accommodating a variable load are not dissimilar to those of meeting foundation restrictions.

A case of interest is of course that of constant stiffness: $\beta = 2$. In this case expression (9.4) is a linear increasing function of $\cos(4\chi)\cos(4\theta)$, as long as $\theta < \pi/8 = 22.5°$, so that $\cos(4\theta)$ is positive. It is thus maximal with respect to χ at $\chi = 0$, indicating that this is a case in which one should stay with the simple two-bar solution of Figure 7.8.

9.2 Numerical experience

Bendsøe and Sigmund (2003, p. 54) carry out the computations for a solid, one-material, 'elastic' ($\beta = 2$) example; the optimisation of a truss that must accept load at two or three points. It is supported by a fixed bearing at one end and a rolling bearing at the other, so that the foundation bears only vertical loads. In Figure 9.4 we reproduce their pictures of the optimal structures for the cases when all loads are applied simultaneously and when they are applied as equally probable alternatives. The single-load version results in an unstable structure based on square frames, whereas the variable-load case results in a stable structure based on triangular frames.

On their page 234 they also consider worst-case optimisation, in which the expectation with respect to ω in expression (9.1) is replaced by a maximisation. They consider optimisation of a truss with two fixed foundation points and three possible load points. Figure 9.5 reproduces their solutions for the case when all three loads are applied simultaneously, and a worst-case design for when the loads are alternative. The structure has much

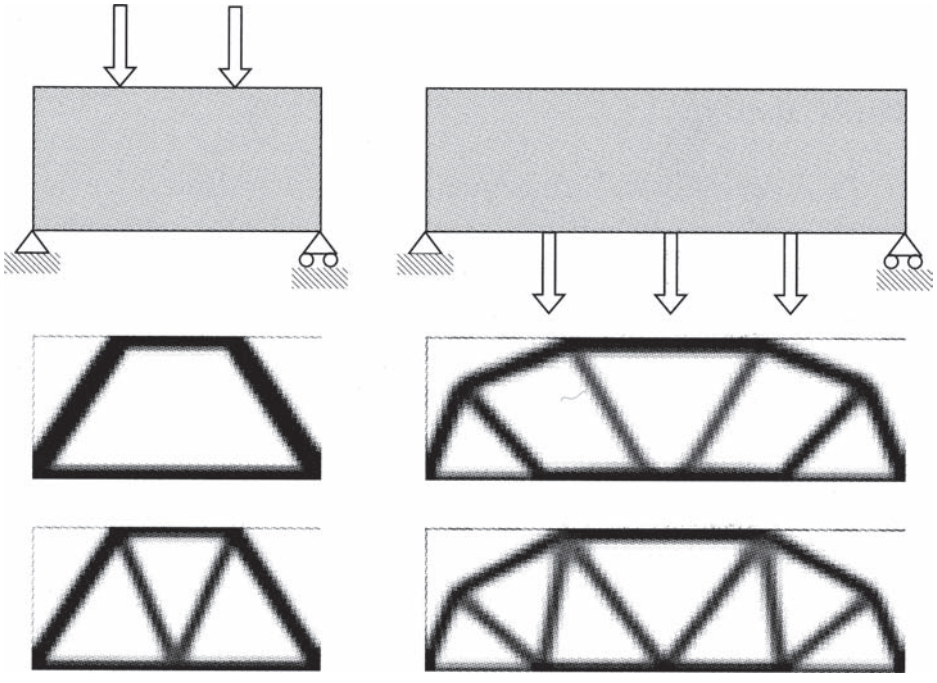


Fig. 9.4 Structures that are optimal when the separate loads indicated are applied simultaneously (*upper*) or as alternatives (*lower*). Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

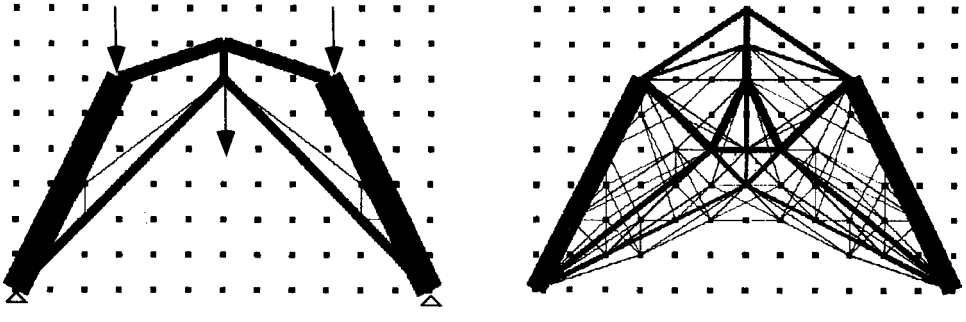


Fig. 9.5 (a) The optimal structure for three loads applied simultaneously. (b) The optimal worst-case structure if the three loads are alternatives. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

more infill in the second case, and there seems to be a hierarchy of structure: main load-bearing members on the outside, a substantial internal skeleton coupling these, and then a fine structure of much smaller members providing local coupling. On their diagram on page 238, they contrast two truss designs, one optimal on an averaged criterion and the other on a worst-case criterion, reproduced in Figure 9.6. The designs are not so very different – both have a considerable infill of secondary members – but the inner

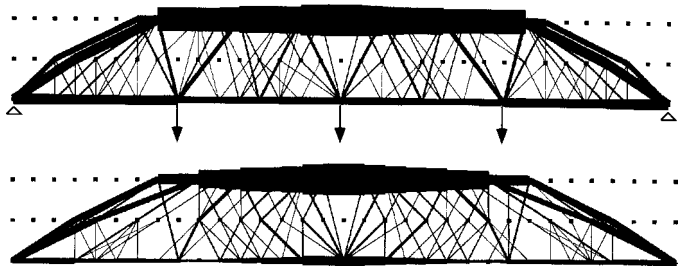


Fig. 9.6 Structures that are optimal under three alternative loads if (a) an average criterion is taken or (b) a worst-case criterion is taken. Reproduced from Bendsøe and Sigmund (2003) with the kind permission of Springer Science and Business Media.

members of the average-optimal design tend to be straighter and the outer members more curved.

Xie and Stevens take what is in effect a worst-case approach in that, when considering the removal of material from the design under the σ/σ_{max} criterion, they remove material only if this is indicated under all load patterns. They evolve a design for a bridge subject to nine alternative point loads in this way. The result is a very solid multi-arched structure with some indication of 'greyness' in the apex of the outer arches.

9.3 The shaping of bone structure

We return to the topic of Section 7.7: the observed pattern of bone deposition induced by a general wasting plus a reinforcement at stress points. The pattern of the trabiculae (the small directed elements of bone) in the upper femur strikingly confirmed the pattern deduced by Michell on an optimality criterion: of a system of mutually orthogonal tension and compression members. However, while there may be a dominant load pattern, there is also evidence that variation of load pattern has played some part in forming the structure. In Figure 9.7 we reproduce figures adapted from Carter *et al.* (1989) obtained by a finite-element simulation of the reinforcement/wasting mechanism. These are to be compared with the observed structure of Figure 7.13.

The first three diagrams illustrate the bone densities produced under loads (a), (b) and (c). The whiter the area, the denser the bone. Case (a) represents the dominant load pattern, corresponding to the midstance phase of gait. This evidently accepts the compressive forces of the body weight, in that it develops a strong dense bridge of bone from the head to the calcar (the solid shell of the femur). However, in the actual femur one observes also a strong development along the oblique upper neck (the arcuate system) with a rather empty area below it (Ward's triangle); neither of these are observed in the simulation. Under loads (b) and (c) such developments do appear, however, in tension and compression respectively. When the calculation is run under an alternation of these loadings then the pattern of the fourth diagram emerges; very faithful to the observed pattern of Figure 7.13.

What is also striking is the extensive area of cancellous bone, of density about half that of the distal bone (the solid shell of the mid-femur).

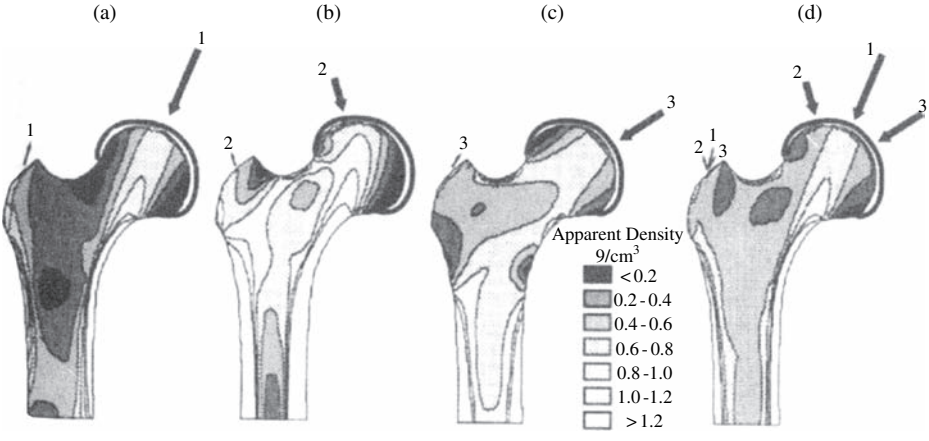


Fig. 9.7 The computed optimal structures for the femur under (a, b, c) each of three possible load patterns, and (d) when all three patterns must be met as alternatives. Reproduced from Carter *et al.* (1989) with the kind permission of Elsevier.

9.4 Buckling

The question of buckling is one that should be addressed at some point, and the fact that buckling can take place in several modes might be said to relate it to the question of variable loads. We have assumed that a strut taking a compressive load along its length could deform only by, in fact, becoming compressed. However, it could also deform by flexing, as shown in Figure 9.8. The unflexed (and so purely compressed) state is in fact an equilibrium one, but the question is whether or not this equilibrium is stable. If it is not stable, then a flexing will grow, and lead to buckling.

An even simpler case would be that of a 'knee', a strut that can flex at only one point. Suppose that the 'leg' constituting the strut must support a mass of M and is of length L, with the knee at the mid-point; see Figure 9.9. If the leg flexes by an angle of θ then it falls at the hip by an amount $L\cos\theta$, losing then a potential energy of $LMg\cos\theta$, where g is the acceleration due to gravity. On the other hand the system gains a potential energy of $a\theta^2/2$ for some constant a, the energy stored by the stretching of the thigh muscle (the quadriceps muscle) which stabilises the knee. (We imagine the muscle behaving as a simple spring, and so neglect any conscious reaction from the owner of the leg.) For small θ then the

$$\text{gain of potential energy} \approx (a - LMg)\theta^2/2.$$

The knee is or is not stable to buckling according as this expression is positive or negative; i.e. according as a is greater or less than LMg.

Fig. 9.8 Flexure of a strut under compression.

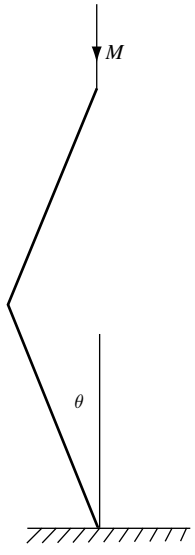


Fig. 9.9 The simplest example of buckling: a knee.

A consideration of buckling then introduces a secondary model, and this induces the associated concept of a ‘geometric stiffness matrix’. These are matters rather too specialised for our purposes. Both Bendsøe and Sigmund (2003) and Xie and Stevens (1997) give an account of the topic and of structural optimisation against buckling.

II

Artificial neural networks

Artificial neural networks (with the recognised abbreviation of ANN) embody a fascinating concept: a stripped-down idealisation of the biological neural network, with the promise of genuine application. The state of the net is described by the activities at its nodes, rather than, as hitherto, by the flows on its arcs. There are such flows, however: the output of a node is a particular nonlinear function (the *activation function*) of a linear combination of its inputs from other nodes. It is the coefficients of these linear functions (the *weights*) that parameterise the net. The specimen aim to which we shall largely restrict ourselves is: to choose the weights so that the system output of the network shall approximate as well as possible a prescribed function of system input. Here by 'system output' we mean the output from a recognised subset of 'output nodes', and by 'system input' the environmental input to a recognised subset of 'input nodes'. However, this design problem is solved, not by a deliberate optimisation algorithm, but by an adaptive procedure that modifies the weights iteratively in such a way as to bring net input and net output into the desired correspondence.

To clarify the biological parallel: the activation function provides a crude version of the operation of a biological neuron; the weightings represent the strengths of interneuronal connections, the axons, and so define the net. The animal neural system is a remarkable processor, converting sensory inputs into what one might term inferences on the environment and then into stimuli for appropriate action. The input/output relation of the ANN is supposed to represent this processing operation, although the sophistication and complexity of the natural system can scarcely be matched in any foreseeable future. However, one is encouraged by the fact that this impressive system is actually the product of evolution: an evolution responding to environmental challenges and opportunities. The adaptive algorithm of the ANN is intended to mimic this response: to represent both the shaping effect of evolutionary selection over thousands of generations and the finishing touch given by 'learning' during a single lifetime.

Of course, evolution works on the whole organism, and the neural net constitutes only the logical component of such an organism. Nevertheless, the logical component is the essential one; it is realised by the aggregation of relatively standard subcomponents (the neurons), and can well be partially decoupled from the 'rude mechanics' and studied on its own. Indeed, the evolving ANN holds out the promise of a simple example of morphogenesis – of the emergence of structure under simple shaping forces.

The appeal of the approach at the superficial level is that it seems both concept-free and painless. There is apparently no need for mathematical analysis or deliberate optimisation; one simply applies an improvement algorithm and lets it run, possibly on a digital or an analogue computer. In fact, the approach is concept-*generating*, in that one is forced to seek insights which prove elusive in high degree. How should one choose an adaptive rule? Where will a given rule lead? How can one interpret the structure that has evolved in cases where the rule does indeed prove successful? Only very partial answers have been found to these questions. The study of ANNs is attractive and self-evidently fundamental. It also provides a context to which people from a range of disciplines can bring ideas and suggestions, enjoying a freedom to play a game whose rules, while unannounced and unevident, are subtle and unforgiving.

ANN models lie at a particular extreme in comparison with the models of Part I. Their very essence is that they must be able to cope with a great variety of inputs – what we would before have termed load patterns. The extreme versatility thus demanded also demands nonlinear operation at the nodes, and invites a statistical formulation. Indeed, the whole aspiration is higher: one of achieving logical operation, in some statistical sense, rather than just simple distribution.

In this brief compass we can give only a quick exposition of some standard material, followed by a foray into some less familiar areas where specific points can be made, relevant to our theme.

The finiteness of the alphabet and the differing conventions of different disciplines also cause a few notational changes. The variable x continues to represent the state of the network, but this is now state at the nodes rather than flow on the links. We follow the convention of control models, by using y to represent observations on the real world, which constitute input data for the net. The variables conjugate to x are then denoted by λ rather than, as previously, by y. The symbol f is now used to represent the activation function, with which we shall make an immediate acquaintance.

10

Models and learning

10.1 The McCulloch–Pitts net

The McCulloch–Pitts net is the basic ANN model, although subject to endless variation and elaboration. We assume a discrete net with nodes labelled $j = 1, 2, \ldots, n$. As noted above, the state of the net $x = \{x_j\}$ is described by measures of activity x_j at the nodes, rather than by measures of flow on the arcs (which would now be termed *axons* if one is to maintain the neural analogy). More specifically, we shall take x_j as the scalar output from node j; in the neural analogy it would be the rate at which neuron j discharges pulses along the axon into which it feeds. These variables are assumed linked by the equations

$$x_j = f\left(\sum_k w_{jk} x_k + y_j\right) \tag{10.1}$$

where f is the so-called *activation function*. The argument of the function can be regarded as the total input to the neuron: a sum of inputs from other neurons plus a term y_j, the input from the environment. Relation (10.1) represents the physics of the situation in that, in a dynamic version of this simple model, it would express x_j in terms of a slightly time-lagged version of the right-hand expression.

Equation (10.1) looks familiar, but with differences. For one thing, it is not, like (1.1), a conservation relation, but is rather a propagation relation. The sum $\sum_k w_{jk} x_k$ is indeed a combination of outputs from other nodes, but these outputs are modulated by the 'synaptic weights' w_{jk}. The 'pulses' of the flow are thus not conserved.

The other new feature is the presence of the activation function f, which converts the excitation provided by the total input rate into a neuronal output rate. The function is usually assumed to have the sigmoidal form of Figure 10.1. The original motivation for this choice is that such a function gives a crude quasi-static approximation to the behaviour of a biological neuron. It turns out however that the family of networks (10.1) provides a versatile computational tool, in that by variation of the weights w one can obtain a great variety of functional relationships between net inputs y and a given subset of node outputs x.

A feature of the sigmoidal form is that, if one feeds the output of a single neuron back as input, then this self-exciting neuron can have two stable equilibrium states, one of low excitation and one of high excitation. The equation $y = \kappa f(y)$ can have three roots, as one sees graphically, and the two outer of these correspond to stable equilibria of the temporal recursion $y_t = \kappa f(y_{t-1})$.

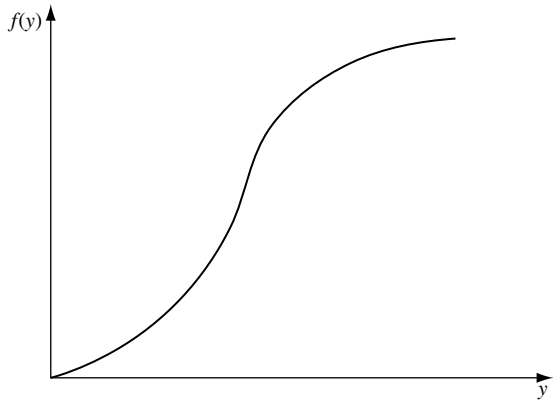


Fig. 10.1 The activation function, usually supposed of the classic sigmoid form.

Finally, one can view equation (10.1) as describing a discrete reaction–diffusion system and, as Turing (1952) first demonstrated, such models have potential morphogenetic properties.

It is convenient to define the 'synaptic matrix', the matrix $W = (w_{jk})$ of weights, which specifies the network structure. In the ANN context, one wishes to choose W in such a way as to achieve a specified functional relationship $x_O = g(y)$ between the input vector y and the output vector x_O of the network, which we regard as the values of x_j for j in a subset O of *output nodes*. However, for this problem the notion that the network should be able to cope with a multiplicity of 'loads' y is of the essence. It is then natural to regard y as a random variable, whose distribution may or may not affect the adaptive algorithm, but will certainly affect its course. This suggests the further relaxation of the problem: to replace the requirement that the output vector should approximate a given function of y by the requirement that it should approximate a random variable z, where the joint distribution of y and z is given. We suppose that there is a measure $C(z, \hat{z})$ of the discrepancy between z and the estimate $\hat{z}$ of it yielded by the output nodes of the network. Let us write this estimate more explicitly as $\hat{z}(y, W)$, to emphasise its dependence upon network input and network structure. Then one would wish to choose W, subject to physical constraints, to minimise the average cost

$$\overline{C}(W) = \mathrm{E}\{C[(z, \hat{z}(y, W)]\}. \tag{10.2}$$

The input variable y will be zero on all but a given subset of nodes, the *input nodes*, just as $C(z, \hat{z})$ depends on x_j only for j in the set O of output nodes. It is convenient, however, to write relations (10.1) and (10.2) in their more general forms.

One generally sees the net as operating by a progression (e.g. a directed pulse flow) leading by some route from input nodes to output nodes. If this progression is strict, in that all paths start from the input set and end in the output set without returning on themselves (and so contain no cycles), then the net is termed a *feedforward* net. Feedforward nets are sufficient until one brings other mechanisms into play. The notation w_{jk} is the reverse of what we used in earlier chapters, in that it refers to a pulse flow in the direction kj. It is convenient in the present context, however. Nodes lying neither in the input set nor

the output set are again termed *internal* nodes, and the question of how many such are needed for a given level of performance is again central.

'Neural space', in which the nodes are embedded, is only rather elastically related to physical space, so the ideas of distance or of axon length are not as prominent as they were for the nets of Part I. The weights w play something of the role of the cross-sections a considered there. However, the upper limits on these are set more by effectiveness than by material cost. Because the activation function f saturates when total input to the neuron increases beyond a certain point, effectiveness is lost rather than gained by excessive increases in the weights.

The need for formulation of concepts becomes acute as soon as one begins to consider ANN performance. For example, the form of the input–output relation that is to be realised affects operation considerably. At one extreme, if the function is generated by some simple model then there is hope that the evolving ANN might come to mimic that model. At the other, if the input vectors are generated by a simple randomising mechanism then asymptotic evaluations are possible.

10.2 Back-propagation and Hebb's rule

We wish to minimise expression (10.2) with respect to W. For the moment we assume the number of nodes to be fixed, although this should also be optimised. The minimisation of the expected discrepancy (10.2) is to be achieved by an analogue of the evolutionary algorithms suggested in Section 4.6, in which one takes into account the effect of a new observed input value y and then perturbs weights in such a direction as to decrease the discrepancy. In the present case one is assuming 'supervised learning', and so is told the value of the target output variable z as well as of y. One then evaluates the discrepancy measure $C[z, \hat{z}(y, W)]$ for the current value of W and revises w_{jk} by the amount

$$\delta w_{jk} = -\kappa \frac{\partial}{\partial w_{jk}} C[z, \hat{z}(y, W)], \tag{10.3}$$

unless this should make w_{jk} negative. One could regard this as a gradient 'learning' algorithm, averaging over varying (y, z). ANN theory considers a variety of learning rules, of which this would be seen as a rather special model-based example.

As in Part I, the form of the revision (10.3) becomes more transparent if we adopt Lagrangian methods to account for the propagation relation (10.1). Let us write this relation in the vector form

$$x = f(Wx - y), \tag{10.4}$$

where it is to be understood that f as an operation is applied to a vector by applying it to every element of that vector. Define the Lagrangian form

$$L(x, y, z, W, \lambda) = C(z, x_O) + \lambda[x - f(Wx - y)] \tag{10.5}$$

where λ is a row vector of Lagrange multipliers λ_j. The interposition of the sigmoidal operation f in the constraint (10.1) now means that Lagrangian extremal assertions hold only locally rather than globally. With this understanding, we can write

$$C[z, \hat{z}(y, W)] = \max_{\lambda} \min_{x} L(x, y, z, W, \lambda). \tag{10.6}$$

The values of x and λ have to be recalculated for every new (y, z). The extremal condition with respect to λ in (10.6) gives the x-determining condition (10.4). That with respect to x then gives the λ-determining conditions

$$\frac{\partial}{\partial x_j} C(z, x_O) + \lambda_j - \sum_k w_{kj} \gamma_k \lambda_k = 0, \tag{10.7}$$

where γ_k is the differential of $f(\xi_k)$ with respect to its argument $\xi_k = \sum_i w_{ki} x_i - y_k$.

Relation (10.4) is to be regarded a a forward recursion, in that it determines the x-values from the known (zero) values in the input set through to the values in the output set. Relation (10.7) is a backward recursion, determining first the λ-values in the output set and then, by reverse progression, back to the values in the input set. Relation (10.3) then becomes

$$\delta w_{jk} = \kappa \gamma_j \lambda_j x_k. \tag{10.8}$$

Rule (10.8) is remarkably simple and intriguing, in that the recommended increment in w_{jk} is a product of a term for node j and a term for node k. In this it recalls the celebrated Hebb's rule, derived from physiological observation, but much invoked in the ANN context. Hebb (1949) observed that a particular synaptic connection tended to be reinforced if the two neurons it connected were simultaneously excited. The product in (10.8) represents the notion of simultaneous excitation. (One would also expect a decline in the absence of such events; this would have to come about by an explicit wasting mechanism, or by a shift of operating point if it turns out that some neurons are operating at a low-gradient point of their activation function.)

However, while the factor x_k can indeed be regarded as measuring degree of excitation at node k, the factor $\gamma_j \lambda_j$ has a different interpretation. If relation (10.4) were modified to

$$x = f(Wx - y) + u, \tag{10.9}$$

where u is a vector of perturbations, then λ_j would have the interpretation

$$\lambda_j = \frac{\partial}{\partial u_j} C[z, \hat{z}(y, W)]. \tag{10.10}$$

It is thus an indication of the direction and degree in which x_j should ideally be changed if one wishes to decrease the degree of mismatch between desired and actual outputs: z and $\hat{z}$. We might then regard it as measure of local mismatch. The factor $\lambda_j x_k$ is then the product of activity at node k and mismatch at node j. The factor γ_j reflects the degree of saturation at node j: the degree to which x_j will actually respond to a change in node input.

The 'back-propagation algorithm', often denoted simply as BP, is just the use of the forward recursion (10.4) to determine x, of the backward recursion (10.7) to determine λ and then of the 'Hebbian' formula (10.8) to revise the weights. It was proposed by Rumelhart *et al.* (1986), the essential novelty (which gave it its name) being the introduction of the dual variables λ and their determination by the backward recursion (10.7). The technique has proved remarkably successful in some cases. One of the classic examples is that in which the y-values correspond to points in a Euclidean space and are to be assigned to sets labelled by z. If the sets are such that they can be separated by a number of hyperplanes then the back-propagation algorithm will solve the problem relatively quickly; this is the reason why the linear inference rules of Section 10.3 are generated so easily by the technique. However, it has also been successful in cases for which there is no such linear separation (Hassoun, 1995 gives examples) and indeed for cases in which the sets to be distinguished show a remarkable degree of mixing.

Even for problems that have been successfully solved, the solution itself can defy comprehension. That is, one has a network that works, but whose functioning or structure cannot be interpreted. On the other hand, there have been cases in which the investigators have claimed to see analogues of biological features in the final structure. Whether there is any biological evidence of the back-propagation mechanism itself is a matter for discussion. It seems that there is no obvious such structure, but this could be because the biological pathways that achieve the back-propagation are more diffuse and primitive than the axonal pathways that carry the forward propagation. Mechanisms have been suggested; see Frégnac (1999). It could also be that biological systems use a different learning rule, although the simplest alternative, that of success reinforcement, shows performance greatly inferior to that of the BP rule's mismatch-correcting action.

ANN theory has suggested a great variety of 'learning' rules, but one would expect that the rather natural derivation of the BP rule would imply a favoured status. However, uncritical acceptance of the particular form (10.3) of the steepest descent rule does lead to troubles that, it turns out, can be avoided by a refinement of the approach.

There is, inevitably, a capacity limit; the BP algorithm will fail to converge if limitations on the network prevent it from realising a transfer function of the required complexity. However, the nonlinear character of the dynamic equation (10.4) (and so of (10.7)) has its own consequences. One is that the methods of convex programming are inapplicable, and so we cannot follow the route of Part I and derive a dual problem (in the λ-variables alone) that might partially characterise the optimal structure. The combination of the presence of nonlinearities and the absence of a length-cost for axons manifests itself in another way. The BP algorithm is slowed down greatly by the occurrence of 'plateaux': regions in which the recursion (10.3) sticks for a long time before finding a direction that yields a definite improvement. Rattray and Saad (1999) explained this by the fact that the network is invariant under a relabelling of the internal nodes (just because labelling supplies no positional information of consequence). The difficulty was resolved by Amari (1998), appealing to concepts he had developed in statistical inference (see Amari and Nagaoka, 2000), but also familiar in numerical optimisation generally.

Suppose we write the time-averaged version of the steepest descent relation (10.3) as

$$\dot{w} = -\kappa \nabla C(w). \tag{10.11}$$

Here w is now the vector of the stacked elements of the synaptic matrix W, and $C(w)$ is the expected cost incurred as a function of w. Then a consequence of the nonlinear character of the net dynamics plus the multiplicity of equivalent solutions is that the function $C(w)$ has many saddlepoints as well as equivalent minima, and it is just at these saddlepoints that the rule (10.11) becomes directionless and stagnates – the plateau phenomenon. On the basis of statistical performance (performance under a variety of inputs) one can define a 'distance' between nets with parameter values w and w', say. If the two values of w differ by an infinitesimal amount δw then this distance squared is of the form $(\delta w)^\top G(w)\delta w$, where $G(w)$ is just the Fisher information matrix. This matrix function $G(w)$ thus defines a metric on the 'neuromanifold', the space of nets. Amari then showed that considerations very similar to those on which Newton derived his method for locating the minimum of a function suggest that relation (10.11) should be replaced by

$$\dot{w} = -\kappa G^{-1} \nabla C(w). \tag{10.12}$$

He termed this modified version the 'natural gradient' method. It supplies the correction that gives gradient descent true minimum-seeking character, and speeds up convergence by a factor of a hundred. Roughly, if a perturbation of w in a given direction u does not affect performance then the perturbation has zero distance, and $Gu = 0$. The matrix $G(w)$ is thus singular at this point, and the dynamics represented by (10.12) will steer w away from such values. One does have the additional task of calculating $G(w)^{-1}$, but, just as in the Fletcher–Powell theory of numerical optimisation (see Fletcher, 1987), there is an efficient updating algorithm for this expression as well.

Thoughts of a biological analogue must prompt thoughts of a biological realisation. In Part I we imagined starting optimisation on a physical continuum that was a blank slate, but was at least a slate, in that a flow-bearing or stress-bearing medium was everywhere present. The biological–neural equivalent of this exists. One can scarcely have a continuum of neurons but, before neural structures begin to develop in the embryo, there is a scattering through the latent region of glial cells – precursor versions of neurons. Axons are observed to grow through the particle-skeleton thus provided, obviously guided by electrical or chemical gradients of some kind that one might see as a version of the λ-field.

Pending further insights, one can make advances in two directions. One is to consider the linear cases exemplified in the next section. These may appear trivial, and indeed they fall far short of the aspiration level of this section, but they do offer both interest and guidance. The second approach is exemplified in the next chapter. This is to consider the solution-structure of specific cases that can be optimised, and to consider whether sequential adaptation of a McCulloch–Pitts net would or could have led to such a solution.

10.3 Linear least-squares approximation

Adaptive neural networks were initially regarded as providing an alternative to the methods derived from classical statistical inference. However, one of the few explicit demonstrations that adaptive optimum-seeking algorithms work is that they lead to just these classical methods, under suitable hypotheses. To take the simplest example, suppose

that one seeks the linear predictor $\hat{z} = Wy$ of an unobserved random vector z in terms of an observed random vector y, optimal in that it minimises $D = \mathrm{E}|z - Wy|^2$. The solution is classical and immediate: the optimal value of the 'matrix regression coefficient' W is $V_{zy}V_{yy}^{-1}$, where $V_{zy} = \mathrm{E}(zy^\top)$, etc.

However, let us mechanically follow the formalism of the last section. We regard y as the n-vector of excitations on the input nodes, $\hat{z}$ as the corresponding m-vector on the output nodes, and W as the set of weights in the linear input–output relation $\hat{z} = Wy$. There are no internal nodes, and the activation function is linear. The Lagrangian function (10.5) reduces to

$$L = \frac{1}{2}|z - \hat{z}|^2 + \lambda(\hat{z} - Wy) \tag{10.13}$$

and the updating relation (10.8) to

$$\delta W \propto (z - Wy)y^\top. \tag{10.14}$$

In its time-averaged form this becomes

$$\dot{W} = \kappa(V_{zy} - WV_{yy}) \tag{10.15}$$

for some constant κ. The passage from (10.14) to (10.15) expresses what amounts to an estimation of the matrices V_{zy} and V_{yy}. Relation (10.15) is, of course, exactly what we would have obtained by following the steepest-descent path of $\mathrm{E}|z - Wy|^2$ as a function of W. One verifies readily that, if the nonnegative definite matrix V_{yy} is nonsingular (equivalent to the assumption that no linear relation holds between the elements of the random vector y) then the solution $W(t)$ of (10.15) converges exponentially fast to the optimal value. The adaptive algorithm is thus successful.

Consider now the more interesting case, in which W is constrained to be of the form $W = BA^\top$, where A and B are $n \times r$ and $m \times r$ matrices respectively, and $r < \min(m, n)$. The assumption is, effectively, that the information flow must pass through a set of r internal nodes, which restrict this flow. The question is, then: which components of y are to be selected to provide the final basis for inference? In statistics this is known as the problem of restricted-rank regression. The solution which emerges involves something like the statistical concept of a canonical correlation, although it is in fact rather a canonical predictor – the fact that we are trying to predict z from y means that the roles of the two vectors are not symmetric.

We shall state conclusions formally.

Theorem 10.1 *Let the scalars θ_j and column r-vectors α_j $(j = 1, 2, \ldots, n)$ be the solutions of the eigenvalue problem*

$$(V_{yz}V_{zy} - \theta V_{yy})\alpha = 0, \tag{10.16}$$

ordered so that θ_j is nonincreasing in j. Then any A whose columns are linearly independent linear combinations of $\alpha_1, \alpha_2, \ldots, \alpha_r$ is optimal; the corresponding B gives the unconstrained optimal predictor $BA^\top y$ of z in terms of excitations $A^\top y$ at the r internal nodes.

The solution of the adaptive equations converges to such an optimal value from almost all initial values.

Proof Let us, for notational simplicity, denote V_{yy} and V_{zy} by V and S respectively. The time-averaged adaptive equations are just the steepest-descent equations

$$\dot{B} = \kappa(S - BA^{\top}V)A, \tag{10.17}$$

$$\dot{A} = \kappa(S^{\top} - VAB^{\top})B. \tag{10.18}$$

Define the two $r \times r$ matrices $L = A^{\top}VA$ and $M = A^{\top}S^{\top}SA$. Then from the equilibrium version of (10.17) we deduce that

$$B = SAL^{-1}, \tag{10.19}$$

which is just the characterisation asserted for the optimal B. Substituting this expression for B in the equilibrium version of (10.16) we find that A obeys the equation

$$S^{\top}SA - V\,AL^{-1}M = 0. \tag{10.20}$$

Now, the product $BA^{\top}$ is unchanged under the transformations $A \to A^{*} = AH^{\top}$ and $B \to B^{*} = BH^{-1}$, where H is an arbitrary nonsingular $r \times r$ matrix. But we can choose H so that the consequent matrices L^{*} and M^{*} are both diagonal. Let us assume this normalisation already performed, so that L and M are both diagonal. Equation (10.18) then becomes

$$S^{\top}SA - V\,AR = 0, \tag{10.21}$$

where R is the diagonal matrix LM^{-1}. Let a_j and ρ_j be the jth column and the jth diagonal element of A and R respectively. Then (10.21) implies that

$$S^{\top}Sa_j - \rho_j V a_j = 0. \tag{10.22}$$

Comparing (10.16) and (10.22), we see that we can identify ρ_j and a_j with one of the eigenvalues θ of the equation system (10.16) and its corresponding eigenvector α. (At least, to within a scaling of α and a more general linear combination in the case of repeated eigenvalues.) This evaluation is consistent with the diagonal nature of L and M. From the values of A and B thus deduced we also find easily that

$$\mathrm{E}|z|^2 - \mathrm{E}|z - BA^{\top}y|^2 = \sum_{j=1}^{r} \rho_j, \tag{10.23}$$

so that the ρ_j must be chosen as large as possible for optimality. Since the a_j must be linearly independent, we see that we must choose the r largest eigenvalues θ_j, an s-fold eigenvalue being counted as s distinct eigenvalues. The value of A (and so of B) determined in this way can be regarded as a canonical form of the solution, the general form of the solution being obtained by the modification of $BA^{\top}$ to $(BH^{-1})(HA^{\top})$. The first part of the theorem is thus proved.

The set of equilibrium solutions of equations (10.17) and (10.18) obtained in this way form a manifold $\mathcal{M}$ in (A, B) space. There are as many such manifolds as there are choices of r eigenvalues from the set of n, but a simple stability analysis shows that $\mathcal{M}$ is the only one that is stable (i.e. is attracting, in that all paths entering a neighbourhood of $\mathcal{M}$ terminate in $\mathcal{M}$). Since $\mathrm{E}|z - BA^\top y|^2$ decreases under the steepest-descent rules (10.15) and (10.16) (A, B) must ultimately enter one of these manifolds, but the stability characterisation implies that all paths but a vanishing proportion end in. $\Diamond$

One concludes that the adaptive algorithm will indeed ultimately converge to an optimal solution, although closer analysis would be needed to determine the time needed to both smooth out statistical variation and allow deterministic dynamics to converge. However, the limit structure is far from determinate. The procedure might terminate at any point in $\mathcal{M}$, and the solution may well then not be particularly revealing. There are means of ensuring that the procedure converges to a solution which is canonical in one sense or another, if there are good reasons to impose direction upon an algorithm whose naive striving was supposed to be its virtue.

A special case is that in which $z = y$, so that one is attempting to reconstruct y from the best choice of r scalar linear functions of the vector. We see then from (10.16) that the θ_j and α_j are just the eigenvalues and corresponding eigenvectors of V_{yy}. The forms $\alpha_j^\top y$, with α_j normalised to unit modulus, are the 'principal components' of y. The canonical form of the optimal reduction Ay, in which this consists of just the first r principal components, is that which would be obtained under the normalisation $B = A$.

It was Oja (1982) who first pointed out that the principal components could be determined by evolutionary methods. An extensive literature has since developed, generalising the problem and seeking automatic methods of normalisation. However, it is the nonlinear version of such ideas which presents the real challenge: the questions of which dimension-reducing nonlinear transformations retain the most information and, a subtler issue, of how 'significant' features of the input are to be recognised.

11

Some particular nets

We review some cases of genuine interest in which analysis predicts a particular form of net, with a view to seeing whether the net thus derived belongs to the McCulloch–Pitts family, and whether it could have emerged by an evolutionary process.

11.1 Recognition, feedback and memory

Suppose that the data vector y that an ANN receives reflects, at least partially, the state of the environment within which the ANN finds itself, and that it is the task of the ANN to infer that state from y. More specifically, suppose that the environment can be in just one of m possible states, which we shall label by $j = 1, 2, \ldots, m$. Suppose also that the net realises an input/output relationship

$$x_O = h(y),$$

which should identify the state as well as possible. For example, the output vector x_O might itself be an m-vector, which takes the value e_j if it is intended to indicate an inference that the environment is in state j; here e_j is a vector with a unit in the jth place and zeros elsewhere. From the statistical point of view, the problem is one of choosing between m exhaustive and mutually exclusive hypotheses on the basis of data y; the net then embodies a particular inference rule. This inference is what would in other circumstances be termed 'recognition'; e.g. the recognition of an individual from a glimpse, or of an alphabetic character in unfamiliar handwriting.

The data vector y will usually give only a fuzzy indication of state, in that there will be observational noise and distortion, but suppose that for each state there is a value $y = a_j$ that should be unequivocally recognisable. The a_j are termed 'traces', and the minimal demand one would make of the net is that

$$h(a_j) = e_j$$

for all j. In fact, however, one would also wish for faithful identification even with some degree of noise and distortion, so that the net should satisfy the stronger demand

$$h(y) = e_j \quad (y \in \mathcal{A}_j), \tag{11.1}$$

for all j. Here $\mathcal{A}_j$ is a set of y-values, necessarily including a_j, which should be recognised as reasonable variants of a_j. One can be more definite only after having made statistical assumptions about the relation between y and environmental state.

One might demand that the net should give a_j rather then e_j as an output when it decides for hypothesis j. That is, that it would convert the observed pattern to one of m standard patterns – one might say that a_j is a 'memory' evoked by the observation y that approximates it. This modification could be achieved by giving the net the response function

$$g(y) = Ah(y)$$

where A is the matrix whose columns are the traces. However, we shall simply assume that $g(y)$ obeys the appropriate analogue of (11.1). We shall also find it preferable to use an operator rather than a functional notation, so that we shall write gy rather than $g(y)$. The analogue of (11.1) is then the requirement that

$$gy = a_j \quad (y \in \mathcal{A}_j), \tag{11.2}$$

for all j.

A state identifier of this type can be converted into a memory device by feeding its output back as input. By a 'memory device' we mean a system that will hold one of a number of given states when it is placed in that state. Since a_j lies in $\mathcal{A}_j$, it follows from (11.2) that the equation

$$gy = y \tag{11.3}$$

has all the traces a_j as solutions. The device can thus 'hold' a memory a_j once it is evoked. However, one would demand a certain robustness: that the device should home on to the memory trace a_j if started from a value of y close in some appropriate sense to a_j. That is, if we consider the dynamic form of (11.3)

$$y_t = gy_{t-1} \tag{11.4}$$

then one would demand that the sequence $y_t = g^t y_0$ should converge to a_j if the initial value y_0 lies in a set such as $\mathcal{A}_j$, which would then lie in the 'basin of attraction' of a_j. Property (11.2) would imply such a convergence in a single step. However, the weaker demand, of ultimate convergence, would be sufficient for the memory property.

If one is designing a memory device then one would wish the basins of attraction to be as large and as numerous as possible, these of course being conflicting demands if material resources are limited. In practice one must allow for the possibility that the dynamic equation (11.4) of the device is subject to random perturbations, in the form of injected noise. One must then also require that the basins should be 'deep', in that a path beginning in a given basin should have a very long expected escape time.

We shall refer to g as the 'autoassociative operator'. The description 'associative' refers to the fact that the trace is chosen that best matches the data vector in some sense; the prefix 'auto' to the fact that the inference is presented in the form of the trace itself.

11.2 The Hamming net

The simplest statistical assumption one could make is that the input vector has the form

$$y = a_j + \epsilon \tag{11.5}$$

if hypothesis j holds. Here ϵ is a column n-vector of observational noise variables. For definiteness we shall suppose ϵ normally distributed with zero mean and covariance matrix vI. That is, the elements ϵ_j of the vector are distributed normally and independently with zero mean and variance v. One could just as well have supposed a general covariance matrix $V_{\epsilon\epsilon}$, but we specialise for simplicity.

One must also make assumptions about the hypotheses (environmental states) themselves. We shall assume that there are m hypotheses, lablled by $j = 1, 2, \ldots, m$, and that hypothesis j has prior probability π_j. The rule that minimises the probability of an error in inference is to choose a value of j maximising $\pi_j P_j(y)$, where $P_j(y)$ is the probability (or probability density relative to a j-independent measure) of y conditional upon the truth of hypothesis j. Under assumption (11.5)

$$P_j(y) \propto \exp[-\frac{1}{2v}|y - a_j|^2],$$

so that one should choose a value of j maximising

$$a_j^\top y - \frac{1}{2}|a_j|^2 + v\log\pi_j = a_j^\top y - b_j, \tag{11.6}$$

say.

This corresponds to a net whose autoassociative operator has the action

$$gy = AM(A^\top y - b), \tag{11.7}$$

where M operates on a vector by substituting a unit for the greatest element (or one of them) and substituting zeros elsewhere. The net is termed the *Hamming net*: it is based on decoding techniques suggested by Hamming, but seems to have been recognised as defining a type of neural net first by Lippmann (1987). It is founded upon statistical principles and is optimal under the stated assumptions. It has long been familiar in communication contexts, with the linear form $a_j^\top y$ being regarded as the output of a 'matched filter'.

The action of g is not quite that of a McCulloch–Pitts net, but would be so if we replaced the maximisation operator M by the operator equivalent of an activation function:

$$gy = Af(A^\top y - b). \tag{11.8}$$

Such a modification would be legitimate if $f(x)$ increased from 0 to 1 and had its threshold so placed on the x-axis that $f(x)$ took a value close to unity for the largest argument and close to zero for the rest. This can be possible only if the greatest value is sufficiently distinct from the rest, but it also demands a scaling of the input signal to f (by a shifting of the f-threshold) that achieves this separation. We shall later see this additive scaling

by b implicit in (11.6) and (11.7) can imply a multiplicative scaling of f, something that is quite natural in a biological context. A biological organism must cope with a vast range of intensities in its sensory input, and must consequently be able to scale the variables so as to gain sensitivity, at one end of the scale, or avoid saturation, at the other.

We thus see that there are at least two features that must be available to the net. One is that of feedback, which is required if a system is to act as a memory device. Feedback is of course perfectly achievable in a McCulloch–Pitts net; the moral is simply that one cannot restrict attention to feedforward nets. The other feature is that of a data-dependent scaling, which is not achievable in a McCulloch–Pitts net, and which we now see to be essential for operation. In fact, the need for some such mechanism was perceived early by von der Malsburg (1973), and re-emphasised by Grossberg (see Grossberg, 1988 and references). It was apparent to von der Malsburg that Hebbian reinforcement would lead to an all-round strengthening of synapses and loss of discrimination unless there was a countervailing effect. To produce this he suggested a renormalisation of signals, and realised that this produced a sharpening effect. In fact, he went so far as to enhance this by introducing lateral inhibition, so that if a neuron (or neural mass) was doing a certain job then its neighbours should be discouraged from trespassing on that role. However, the Hamming net achieves what is necessary in this direction automatically.

If we neglect the shift b in (11.8) (which we may do if the shift is the same for all elements of the vector) then the autoassociative operator for the modified Hamming net is $AfA^{\top}$. The corresponding operator for the well-known Hopfield net is $fAA^{\top}$, and there is no transformation of the problem that exhibits these two operators as equivalent.

The Hopfield net had long acclaim as a 'physical system with emergent collective computational abilities', to quote the title of the paper in which it was proposed (Hopfield, 1982). It is surprising that the stronger claims of the Hamming net were not realised at the time – its concepts had been so well absorbed by the engineers that it was not even seen as a neural net. The author has elsewhere (Whittle, 1998) stated his belief that the Hopfield net has enjoyed an undeserved vogue. The following list of considerations summarises points both for and against this view.

(1) The Hopfield net seems to lack an obvious derivation that would establish its rationale. Lippman (1987) makes the point that it must necessarily suffer by comparison with the Hamming net. He further asserts that this difference in performance is borne out in tests on character recognition, pattern recognition and bibliographic retrieval.
(2) The Hamming net achieves its final inference in a single pass – the Hopfield does so only after several iterations, if at all. During these passes, the data with which it was initially furnished is degrading.
(3) The Hopfield net will operate as a successful associative memory only if the traces are almost orthogonal. If one sets it up with nonorthogonal traces then it can fail completely. For example, Kumar *et al.* (1996) report cases in which the net annihilates the memory of all encoded vectors. That is, even if y_0 is set equal to one of these traces y_t will diverge from it with increasing t.
(4) The Hopfield net does obey a restricted extremal principle (that if f is the signum function then the quadratic form $y^{\top}Wy$ is nondecreasing under action of the Hopfield

operator for vectors y with ± 1 elements). However, this seems too special and fortuitous a property to be regarded as a motivating extremal principle.

(5) The Hopfield net shows what are regarded as 'spurious' equilibria: that equation (11.3) may have not only the traces as solutions, but also linear combinations of these traces. If these are indeed regarded as spurious then m can grow only linearly fast with n for performance to be reliable. In fact, these 'spurious' solutions are the strangled manifestation of what for generalisations of the Hamming net can be seen as proper and meaningful entities, compound traces, whose acceptance permits exponential growth of m with n.

(6) If one is not interested in compound traces then the Hopfield net seems to have an unecessarily high dimension: n rather than the m of the Hamming net.

However, we shall keep options open by using the inclusive term 'H-net' to designate either a Hamming or a Hopfield net.

11.3 The probability-maximising algorithm

The operators h and g were introduced in the last section under the assumption that recognition was achieved in a single step. In fact, as became increasingly evident as discussion continued, the more natural view is that the recognition device has its own internal dynamics. It is fed with a data vector y and then, by the iteration of a fixed operation, modifies this so that it approaches one of the traces a_j, hopefully the correct one in most cases. This view and the analysis which follow were presented in Whittle (1998). We shall assume operation in continuous time.

Denote the value of this modified vector at time t, a transitional form between the raw input vector y and the final trace inference, by $\eta(t)$. One can view the dynamics as proceeding by a steepest descent from the initial value $\eta(0) = y$ into the nearest potential well, centred on one of the traces. The potential well also defines the basin of attraction of this trace, and one wishes to choose the potential so that the inference has minimal probability of error.

Under the assumptions of the last section the probability density of y unconditioned by knowledge of j is

$$P(y) = \sum_j \pi_j P_j(y) \propto \sum_j \pi_j \exp[-\frac{1}{2v}|y - a_j|^2]. \tag{11.9}$$

This is a function with a mode near each of the traces if the noise variance v is small enough. (Actually, the critical quantities are the noise/signal power ratios $v/|a_j|^2$, which we imagine to be of order v/n.) As the noise/signal ratio increases some of these nodes will coalesce, reflecting the unavoidable fact that discrimination between hypotheses becomes more difficult. A natural choice would then be to generate the developing estimate $\eta(t)$ by following the steepest ascent path of the function $P(\eta)$ from the initial value y. This path will end at a mode and effectively identify a trace. One would imagine that this procedure is asymptotically optimal in the appropriate sense, and the point is proved in Whittle (1998).

Equivalently, one could follow the steepest descent path of the potential function $U(\eta) = -\log P(\eta)$, whose wells correspond to the modes of $P(\eta)$. The evolution equation is then

$$\dot{\eta} = -\kappa v \frac{\partial U(\eta)}{\partial \eta} \tag{11.10}$$

where the derivative on the right-hand side is interpreted as the column vector of derivatives with respect to the elements of η. No particular condition is placed upon the coefficient κ at the moment; the factor v is included for convenience.

Theorem 11.1 *Under the assumption (11.9) the steepest descent equation (11.10) becomes*

$$\dot{\eta} = \kappa(\sigma A D f A^{\top} \eta - \eta). \tag{11.11}$$

Here A is again the matrix whose columns are the traces a_j. The activation operator f induces the exponential transform

$$fx = \mathrm{e}^{x/v} \tag{11.12}$$

on a scalar x. The operator D is simply a diagonal matrix with jth element $D_j = \pi_j \exp[-|a_j|^2/(2v)]$ and σ is the scaling factor

$$\sigma = \left(\sum_k D_k \mathrm{e}^{x_k/v} \right)^{-1},$$

where x_k is the kth element of $A^{\top}\eta$.

Proof is a matter of direct verification. The interesting and pleasantly surprising point is that the steepest ascent of $\log P$ has led us to relation (11.11): the continuous-time evolutionary equation for a network. The net would belong to the McCulloch–Pitts family, but for the presence of the rescaling factor σ.

We shall term this algorithm the 'probability-maximising algorithm', a term abbreviated to PMA. The factors A, $A^{\top}$ and f in the operator of (11.11) give the algorithm very much the effect of a McCulloch–Pitts net. One might be perturbed by the fact that the activation function f behaves as an exponential rather than as the classical saturating sigmoid. However, saturation is supplied by the presence of the scaling factor σ, which nullifies the exponential growth for large inputs. This data-dependent scaling emerges as a necessary feature of the inference process. It is one that cannot be realised in a pure McCulloch–Pitts net, but is certainly both evident and essential in biological contexts.

If v/n is small then the vector $\sigma D f A^{\top} y$ is virtually $M(A^{\top} y - b)$; the vector e_j corresponding to the best-matching trace. The PMA net thus behaves as a dynamic and autoassociative version of the Hamming net, in that it swings η from its initial value y to a value near the best-matching trace. The selection effect is fuzzy if v/n is not small, but we recognisably have the simple Hamming net.

One gains insight if the network implied by equation (11.11) is seen in block form. Consider first the equilibrium version

$$A \sigma D f A^{\top} \eta = \eta \tag{11.13}$$

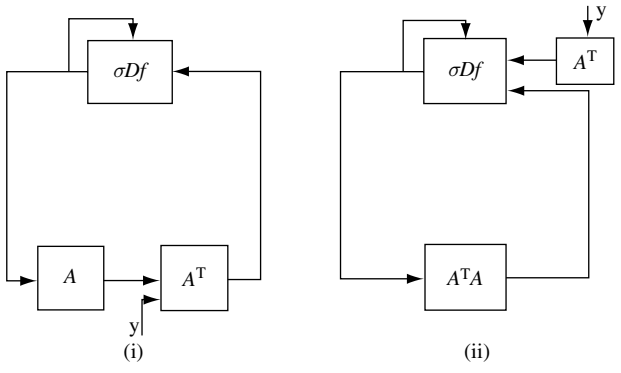


Fig. 11.1 Alternative block versions of the network represented by equation (11.13).

of the equation. This is equivalent to the block diagram (i) of Figure 11.1, the blocks representing operations and the signals travelling between them being vector-valued. So, the block labelled A represents simply a net that achieves the linear operation of multiplication by the matrix A. The block labelled σDf achieves the nonlinear operation indicated, but we have included a subsidiary feedback loop to reflect the fact that the normalising scalar σ is to be calculated from the output of the Df part of the operation. The structure as a whole constitutes a feedback loop, but we have indicated where y, the initial value of η, is to be injected.

Premultiplication of relation (11.13) by $A^\top$ transforms it to the relation

$$(A^\top A)\sigma Df\zeta = \zeta, \tag{11.14}$$

in terms of the variable $\zeta = A^\top\eta$. This is represented in block form by (ii) of Figure 11.1. The effect is to achieve a cyclic permutation of the factors of the operator $A\sigma DfA^\top$, because one is now expressing the feedback structure from a different point in the loop. Because of this the initial data are now injected in a different form, that of $A^\top y$, and at a different point in the loop.

The point of these observations will be seen in the next section, when we generalise the model slightly, and find that the block structure agrees in detail with what is believed to be an anatomical analogue.

11.4 The PMA with compound traces

Suppose that the traces a can be combined additively, so that one can observe compound traces of the form $\sum_j r_j a_j$. Here the r_j are nonnegative scalar mixing coefficients. It might be argued that such additive combination is against biological reality. One might have a trace for 'cat' and a trace for 'mat' but 'cat on mat' is not the same as 'cat plus mat'. This opens the whole question of what we mean by a 'trace'.

Let us note first that there are cases where there is no such problem. Consider the olfactory system, developed for the perception of smells. One odour can very well be simply added to another, in any proportion. Whether the compound odour is perceived as

the sum of its parts is quite another matter – the character of perfumes and foods often depends upon the fact that they may not be.

There are more general arguments for the view that compounding can take the loose form of aggregation in certain circumstances. As we have noted, optical images form compounds in a nonadditive manner, which is one reason why their processing presents such problems. However, one observes 'features', and the total impression formed is often more a list of the features observed than of a total spatial image. The eye cannot take in a whole image (and nor can the brain perhaps recall one, pictorially) – it scans the image in a free, sequential fashion, looking for such features. The notion of a 'feature' is a tricky one – a local structure that is recognised and seen as significant despite variations in size, orientation or a degree of distortion or obscuration. A high proportion of the ANN literature is concerned with the fundamental problems of characterising and recognising features. However, if one accepts that features can be characterised and recognised, then there is a case for arguing that a list of features observed is often a sufficient summary of the total observation, and can be learned as a sufficient characteriser of the complex entity.

The model (11.5) will now take the more general form

$$y = Ar + \epsilon, \tag{11.15}$$

where r is the column m-vector of weightings r_j. Suppose that attention can be confined to q linearly independent such compoundings; let R be the $m \times q$ matrix whose columns are the corresponding r-values. Then the matrix whose columns are the q possible traces is AR. So, formally, we just generalise to the compound case by replacing A by AR.

The action of the feedback loop in the static (equilibrium) case is then represented by the relation

$$AR\sigma DfR^{\top}A^{\top}\eta = \eta, \tag{11.16}$$

where f has again the definition (11.12) and the factors D and σ are modified to take account of the fact that we now have q rather than m possible traces. So, D is a $q \times q$ diagonal matrix with rth diagonal element $D_r = \pi_r \exp[-|Ar|^2/(2v)]$, where π_r is the prior probability that the m-vector of mixing coefficients takes the value r. The scaling factor σ is again the reciprocal of the sum of outputs from the f operation, i.e.

$$\sigma = \left(\sum_r D_r \mathrm{e}^{x_r/v} \right)^{-1},$$

where x_r is the rth element of $R^{\top}A^{\top}\eta$.

Let us again change the injection point and rewrite relation (11.16) as

$$(A^{\top}A)\sigma RDfR^{\top}\zeta = \zeta, \tag{11.17}$$

in terms of the variable $\zeta = A\eta$. This is in fact the form in which we want it; we have set out the block diagram representing relation (11.17) in Figure 11.2. A new input y is injected as $A^{\top}y$ just before the block of simple connections representing $R^{\top}$. We have taken the output from the R-block as also an output from the system, because this is just $\hat{r}$, the estimate of the mixing vector r yielded by the net. It thus represents the inference

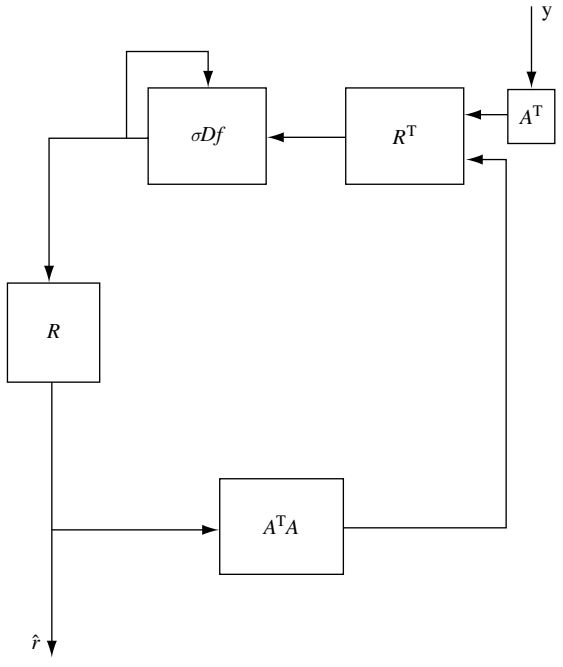


Fig. 11.2 A block version of the network represented by equation (11.16).

yielded by the net in its essential form. The return arrow on the σDf block represents the rescaling operation.

11.5 Comparisons with the olfactory system

In the author's view, at least, the structure derived above from the PMA principle supplies a surprisingly immediate and detailed match to the initial stages of processing in the animal olfactory system; those constituted by the *olfactory bulb* (OB) and the *anterior olfactory nucleus* (AON). The match between the block diagram of Figure 11.2 and the schematic block diagram of the actual anatomy in Figure 11.3 is only part of the argument. A more detailed discussion of the anatomy shows that there is also agreement in form and function at block level.

We should note, first, that two types of neurons occur in this part of the neural system: *mitral* cells and *granule* cells, whose outputs are respectively excitatory and inhibitory. The signals from the nasal receptors first enter the *glomuleri*, a set of densely interconnected mitral cells, where they undergo an initial reduction. We regard this as analogous to the operation of $A^\top$ on y, although the positive internal feedback induced by the interconnections also has the effect of prolonging the memory of a brief sensory exposure. The OB consists of a bank of neurons so large that a considerable 'divergence' (i.e. branching of axons) is required as the OB is entered; this we would regard as equivalent to the operation of the $q \times m$ matrix $R^\top$ in the PMA net. Interconnections between the neurons of the OB consist largely of a pooling of the inhibitory outputs, now interpretable as part of the mechanism to achieve (by internal negative feedback) the

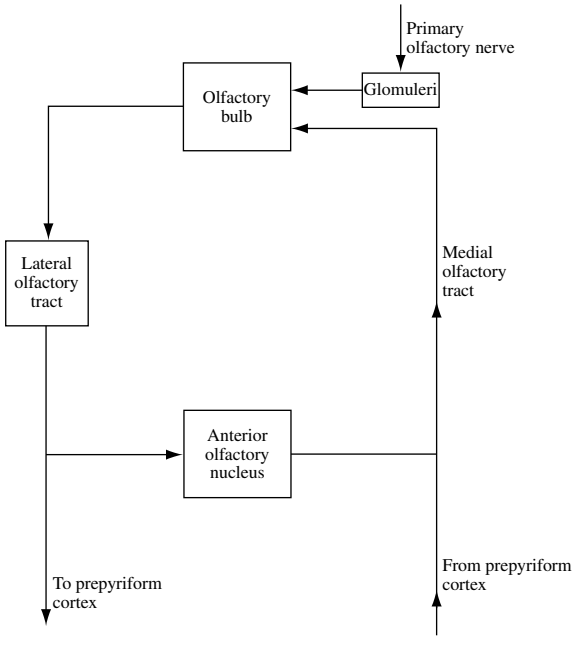


Fig. 11.3 A block diagram of the animal olfactory system.

standardising rescaling operation represented by σ in the analogue. The *lateral olfactory tract* (LOT), which takes the output of the OB, achieves a powerful 'convergence', i.e. a mapping on to considerably fewer axons, which we see as analogous to the operation R of the PMA net, reducing q lines to m. The OB also manifests a symmetry of connections, which we see as matched by the mutual symmetry of the $R^\top$ and R operations of the PMA net.

The LOT passes this reduced information on to the AON, but also to the *prepyriform cortex* (PC), which integrates the olfactory information with other brain functions. In this it would seem to convey the equivalent of the $\hat{r}$ output delivered by the net of Figure 11.2, which represented the essential condensation of the sensory input. The AON consists of a set of neurons with internally symmetric connections, a symmetry mirrored in that of the $A^\top A$ block of the PMA net, which we regard as the analogue of the AON. A further bundle of axons, the *medial olfactory tract* (MOT) completes the feedback loop.

If this analogy is correct then the AON can be seen roughly as determining which elementary traces a are to be detected in y, and the OB as determining the weights r to be assigned to these in the compound trace.

Let us now make a more deliberate comparison, listing the points of agreement between the observed physiology and the PMA mechanism, and hoping that their unforced nature will be incidentally apparent. We shall refer to the $\sigma DfR^\top$ block and the $A^\top A$ block as 'OB' and 'AON' respectively, for convenience of notation and to indicate the analogy we hope to demonstrate.

(1) The two subsystems 'OB' and 'AON' form a natural loop, just as we have observed the OB and AON to do; this presumably achieves the aim that fixation of a data vector evolves to fixation on an environmental state, an identification then passed on to higher-order processors.
(2) The occurrence of the product matrix AR in the PMA automatically leads to two recognisable stages in this feedback loop. One is associated with the factor A, listing the basic traces; the other with the factor R, listing the weights defining recognised compound traces. These stages then respectively hold the basic trace information and form the inference on the weight vector r. One is then led to identify them with the AON and OB respectively, the precise identification being determined by the placing of input and output lines in the PMA.
(3) Each of the two subsystems is formed from a symmetric H-net, consistently with physiological observation. In the case of the 'AON' there is symmetry in the conventional sense, in that a feed from excitor j to excitor k is balanced by an equal feed in the reverse direction, as observed in the AON. The structure of the 'OB' agrees with that of the OB in that there is divergence of axonal connections on entry and a convergence on exit, also in that there is only a moderate degree of internal connection. Such connections as there are seem to correspond to a pooling of the inhibitor outputs, which is just what would be needed if an internal negative feedback is to realise the essential standardising rescaling. If we are to believe in the identification of 'OB' and OB, then symmetry must hold in OB in the sense that an input feed from line j to excitor k is balanced by an equal feed from excitor k to output line j.
(4) Freeman (1992) emphasised the strongly reducing character of the LOT – an observation repeated in the literature, although at least one set of authors find evidence of some structure (Scott *et al.*, 1980). The 'LOT' of the analogue performs the operation R, a reduction from a high dimension q (the total number of possible traces) to the lowest meaningful dimension m (the number of basic traces).
(5) The continuation of 'LOT' as an output would be consistent with the continuation of the LOT to the PC. The information conveyed is in its most condensed and transparent form: an estimate $\hat{r}$ of the weightings of the basic traces.
(6) Inhibitory effects (which we have mentioned only cursorily, and whose essential role we shall discuss in Chapter 12) enter the analogue roughly as they are observed physiologically: in the mitral/granule units which make up the basic oscillators (see Chapter 12), in the global inhibition required for standardisation of the output from the OB and as an aid to phase-matching in the return loop from the AON to the OB.
(7) The preprocessing achieved by the glomuleri can be seen as a fair analogue of the transformation of the input vector y by the operation $A^{\top}$. The positive internal feedback of the glomuleri can also be seen as a way of increasing contrast and of retaining a fleeting sensory experience.

The comparison at this level makes the essential points, and we shall not depart too much from the themes of the text by going into issues which are special for this particular application. Nevertheless, it must be emphasised that the picture given by Figure 11.2 is skeletal in the extreme, and does not constitute a complete mathematical model as it stands, let alone an adequate analogue of biological reality. To begin with, one should

develop a dynamic version of the model, implying also a more adequate model of the neuron. Related to this is the recognition of the fact that the absolute level of a signal can mean very little when, typically, the strength of the input signal can vary by several orders of magnitude. Information must then be conveyed by temporal patterns in the signal, notably by varieties of oscillation, rather than by simple indicators of current signal strength. We turn to these points in Chapter 12.

12

Oscillatory operation

In Sections 11.1 to 11.3 we developed the idea of an associative memory as a device that is intrinsically dynamic, although this aspect was not emphasised in the later sections. If one is to keep the biological archetype in mind then one should also formulate a more realistic dynamic model of the neuron. In a sequence of publications W. Freeman (see the reference list) has developed a model which, while simple, captures the biological essentials with remarkable fidelity. It represents the pulse input to the neuron as driving a linear second-order differential equation, whose output constitutes the 'membrane potential'. The law by which this potential discharges and stimulates a pulse output implies, ultimately, a nonlinear activation function of the sigmoid form.

This dynamic view is related to the second point made at the close of Chapter 11: that it is unrealistic to assume that the absolute level of activation or pulse rate can convey any real information in a biological context. These variables are too fuzzy, even when aggregated, and a variable scaling is all the time being applied to keep activity at an optimal level. Neural signals are typically oscillatory, in an irregular kind of way, and it is the pattern of this oscillation which conveys information – by binding the relevant neural units into a joint dynamic state. This is a line of thought which Freeman has developed particularly, and that is combined with the ideas of Chapter 11 in Whittle (1998).

These are important network aspects which appear nowhere else in the text, and whose consequences we shall review in this chapter.

12.1 The Freeman model of the neuron

The appropriate sections of ANN texts are always prefaced by an illustrated account of the biological neuron. Being at one remove from that position, we give just the briefest of verbal sketches.

The input to the neuron consists of electrochemical pulses conducted along the incoming lines (the *axons*) to meet the cell body of the neuron at the *synapses*; there may be as many as 100,000 such inputs to a single neuron. In a first approximation these inputs are assumed to add; we shall denote the pooled pulse input by u, a function $u(t)$ of time t.

The cell body has its own internal dynamics; the chemical pulse input produces an electric cell current x. In the Freeman model x is the solution of a linear differential equation driven by u:

$$\mathcal{L}x = u. \tag{12.1}$$

$\mathcal{L}$ is then the corresponding linear differential operator. Observation led Freeman to evaluate it as the second-order operator

$$\mathcal{L} = \mathcal{D}^2 + 0.94\mathcal{D} + 0.16, \tag{12.2}$$

if the unit of time is taken as one millisecond. Here $\mathcal{D}$ is the differential operator $\mathrm{d}/\mathrm{d}t$. Both the order of $\mathcal{L}$ and its coefficients turn out to be significant.

The cell current x can virtually be identified with the potential of a membrane which is being charged up by the input. This identification is to be made, not in a literal sense for the individual neuron, but in a statistical sense for an assembly of co-operating neurons, as we shall soon clarify. When this potential reaches a critical value the membrane discharges, sending an electrochemical pulse along the single outgoing axon, which may then branch greatly to feed a pulse input into many other neurons. Freeman postulates that this output y is related to x by a static nonlinear relation

$$y = f(x). \tag{12.3}$$

One may wonder how relation (12.3) is to be interpreted, if y has a pulse-like character and x has not, and there is indeed a critical point at issue. In reality one is dealing with a cluster of neurons all doing much the same job. Wilson and Cowan (1972) refer to such an assembly as a 'localised population'; Freeman (1975, 1987) uses the term 'neural mass'. Freeman's equations apply to the mass rather than the individual neurons. The variables u and x should then be regarded as pooled values of input pulse rates and cell currents respectively, and so as continuous variables. The output pulse rate y is then likewise a continuous variable, related to x by the relation (12.3).

The function f of (12.3) is just the activation function of the more static McCulloch–Pitts formulation, presumed to have the same sigmoid character. This sigmoid form now has a physical explanation. If relation (12.3) were presumed to hold for a single neuron, then f would have to reflect the triggering or nontriggering of a pulse according as x does or does not exceed a critical value. However, there will be a statistical spread of values in the neural mass, and an aggregation of the response over the mass will blur this step response to a smooth sigmoid.

Suppose we consider an assembly of such neural masses, all obeying the dynamics expressed by (12.1), (12.3) and interconnected by a synaptic matrix W. Denote the average cell current of the jth mass by x_j, the column vector of the x_j by x and the vector with elements $f(x_j)$ by $f(x)$. Then the network thus formed will obey the equation

$$\mathcal{L}x = Wf(x) + u, \tag{12.4}$$

if we suppose for generality that the system is subject to an exogenous pulse input u.

One very special case of interest would be that in which the system reduces back to a single mass. If the mixing within this mass is homogeneous enough, then equation (12.4) reduces effectively to a scalar form. It represents a single neural mass whose pulse input consists of a feedback term $Wf(x)$ as well as an external feed u. Suppose that u is

constant in time and that we look for solutions of (12.4) with constant x. Then we are looking for solutions of the equation

$$f(x) = ax + b, \tag{12.5}$$

where a and b are constants determined by W and u. By the usual arguments for such equations, there may be as many as three solutions of (12.5) (see Figure 10.1), the smallest and greatest of these corresponding to stable equilibria of the dynamic system. The single neural mass with feedback may then have two possible equilibrium states, characterised as states of 'low' and 'high' excitation.

12.2 The Freeman oscillator

Biological neural nets are in a constant state of oscillation, of a nature, complexity and strength dependent on circumstances, and doubtless serving the purpose of conveying information or commands in a situation where slow variation will not do so. Wilson and Cowan suggested in their classic paper (1972) that a basic periodic oscillation might be produced by mutual interaction between an excitatory and an inhibitory unit: a biological and nonlinear version of the simple harmonic oscillator. Neurons with these differing characteristics are indeed observed, the mitral and the granule cells, and are observed in close association. A mitral cell will stimulate a granule cell, the effect in the reverse direction is inhibitory.

We may now often use the term 'neuron' where 'neural mass' is intended. Freeman (1975, 1987) developed the Wilson–Cowan model, notably by assuming his own model for the neuron. Denote the cell currents of the excitor and inhibitor by x and z respectively. Then the general equation (12.4) becomes, for the excitor/inhibitor pair,

$$\mathcal{L}x = -f_2(z) + u, \quad \mathcal{L}z = f_1(x), \tag{12.6}$$

if we assume that it is only the excitor that may receive external stimulation. Here f_1 and f_2 are the activation functions of excitor and inhibitor, not necessarily identical. Seeing that the behaviour of the system (12.6) depends very much on the specific character of the Freeman neural model, and seeing that Freeman took the system as the basic building block for his much more general models, we shall term this system the 'Freeman oscillator'. The model is examined in detail in the papers quoted and on pp. 130–31 of Whittle (1998); we summarise conclusions with a brief indication of the reasoning.

The frequency response $L(s)$ of the operator $\mathcal{L}$ is determined by $L(s) = \mathrm{e}^{-st}\mathcal{L}\mathrm{e}^{st}$. Assume that $\mathcal{L}$ has a stable inverse, in that the zeros of $L(s)$ lie strictly in the left half of the complex plane. If u is constant then the system (12.6) has a unique constant equilibrium, and the linearised version of (12.6) at this equilibrium value has a solution proportional to e^{st} if

$$L(s) = \pm i\lambda. \tag{12.7}$$

Here $\lambda = \sqrt{\gamma_1\gamma_2}$, where γ_j is the derivative of f_j at the equilibrium argument.

We can take λ as a coupling parameter. For λ zero the s-roots of (12.7) lie in the left half of the complex plane, by hypothesis, and the equilibrium point is stable. As λ increases stability weakens, and if ever a root crosses the imaginary axis then the equilibrium point becomes unstable. If it crosses at a nonzero value $i\omega$ then such a transition can mark a Hopf bifurcation, for which the appearance of an oscillatory instability in the linearised model marks the emergence of a limit cycle around the equilibrium point for the actual nonlinear model. Conditions that this should be so are satisfied in this case. We shall refer to the value of ω at the transition as the *threshold frequency*.

In the oscillation that develops at criticality the excitor leads the inhibitor by a quarter-cycle, just as velocity leads displacement for the simple harmonic oscillator. This predicted phase lead is confirmed by physiological observation.

The nature of the oscillation shows an interesting dependence on the order of $\mathcal{L}$. In the first-order case $L(s) = s + L_0$ (for which, by the stability assumption, $L_0 > 0$) the roots of (12.7) are $s = -L_0 \pm i\lambda$. These lie strictly in the left half-plane, whatever λ, and the equilibrium point is always stable.

Consider the second-order case, $L(s) = s^2 + L_1 s + L_0$, for which the assumption of stability would imply that L_0 and L_1 were both positive, and the additional assumption of real zeros (nonoscillatory response) would imply that $L_1 > 2\sqrt{L_0}$. We find then from (12.7) that $\omega = \sqrt{L_0}$ and that the critical value of λ is $L_1\sqrt{L_0}$. That is, if the equation $\mathcal{L}x = 0$ is regarded as the equation of a lone damped oscillator, then the excitor/inhibitor system begins to oscillate when the coupling exactly equals the damping, and at a frequency that would be that of the lone oscillator if undamped.

All these observations are due to Freeman. Freeman *et al.* (1988) derived the determination (12.2) of $\mathcal{L}$ from the temporal response of the rabbit neuron to a single pulse input. The predicted threshold frequency is thus $\omega = \sqrt{0.16} = 0.4$ radians per millisecond. In herz (cycles per second) this would be

$$\frac{1000}{2\pi}0.4 \approx 63\,\text{Hz}.$$

This lies quite centrally in what is known as the gamma band, a range of frequencies prominent in electroencephalograph (EEG) records. Activity in this band is strongest round about 40 Hz, but it turns out that the threshold frequency drops when oscillators are linked into a system.

One can vary the point of operation (the 'equilibrium' point) by varying the biasing input u (supposed constant in time). If u is either too high or too low then the operating point lies at a value at which one or other of the activation functions saturate, so that λ is small and there is no oscillation. In the intermediate range the amplitude of oscillation rises from zero at the ends of the range to a maximum in the middle. However, the frequency stays within a few per cent of the 63 Hz figure over the whole of the range; see Figure 12.1. The actual waveform is quite close to sinusoidal; see Figure 12.2.

It is this behaviour that reveals the potential usefulness of the Freeman oscillator. It converts a biasing input u to an oscillatory response, in that if u lies within a certain band of values there will be such a response, of a fairly definite frequency.

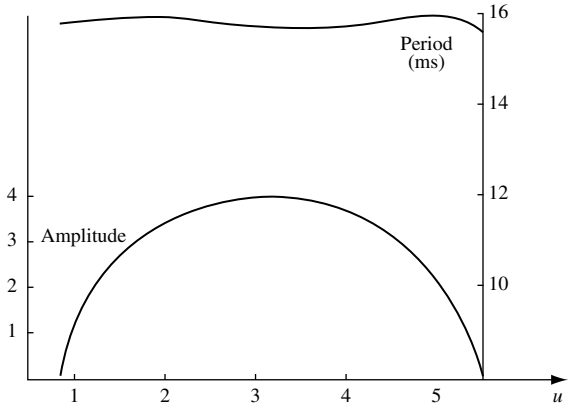


Fig. 12.1 A plot of the amplitude and period of the limit cycle of the Freeman oscillator, as functions of the biasing input u.

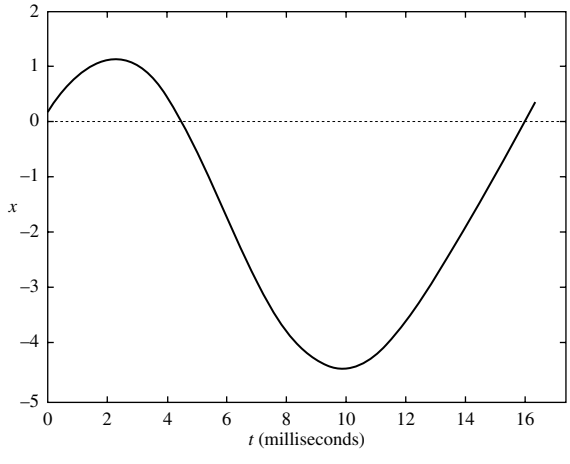


Fig. 12.2 The near-sinusoidal form of the wave generated by an isolated Freeman oscillator. This can become greatly distorted if feedback or standardising operations are introduced.

12.3 Oscillation in memory arrays

One can transfer the linearised analysis of the last section to the case of a full network, as specified by equation (12.4). One finds then that relation (12.7), testing for solutions of the linearised equations proportional to e^{st}, becomes

$$L(s) = \lambda_j. \tag{12.8}$$

Here the λ_j are the eigenvalues, in general complex-valued, of the matrix WG, where G is the diagonal matrix of the gradients of the activation functions $f_j(x_j)$ at the 'equilibrium' operating point. If one now returns to the associative memory nets of Sections 11.4 and 11.5, suggested by the PMA, and replaces the 'neurons' of the structure by Freeman oscillators, then one obtains an oscillatory net whose behaviour stands in fairly direct analogue to that of the nonoscillatory version. That is, the presence of a particular trace

in the sensory input leads to the excitation of corresponding receptors in the H-nets, this excitation now manifesting itself as an oscillation. The analysis is given in chapters 12–14 of Whittle (1998).

However, there is one new and intriguing feature, which comes about because of the rescaling operation which is intrinsic to the PMA (and to biological reality). We can exhibit the effect by considering a single Freeman oscillator which incorporates both feedback and rescaling. Consider the system

$$\begin{aligned} \mathcal{L}x &= wf(x-q)-z, \\ \mathcal{L}z &= f(x-q), \\ \mathcal{D}q &= \kappa[f(x-q)-\delta], \end{aligned} \tag{12.9}$$

where all the variables are scalar. Here δ is the value of f at the point where its gradient is maximal.

The first two relations describe a Freeman oscillator, modified in that an x-feedback has been introduced into the first relation, the effect of the inhibitor is taken as linear, and a variable q has been introduced, which biases x. The excitor feedback is something that occurs when one incorporates oscillation into a H-net, and the inhibiting effect is made linear because the granule cells often exert their effect directly on cell current rather than via the synapses. The bias term q is the real novelty: it represents the rescaling effect. The third equation of (12.9) represents q as attempting to shift the operating point of x to the value at which the activation function is most responsive; that at which the gradient of f is maximal.

The combined effect of feedback and rescaling is radical. Whether it could have been foreseen depends, as they say, upon the foresight of the seer. For values of the correction constant κ greater than a critical value there is no oscillation. As κ decreases through this value the 63 Hz oscillation breaks out. As κ decreases further the period of the oscillation increases and the waveform soon becomes almost perfectly square; see Figure 12.3. This is because the correction q overshoots into regions where $f(x-q)$ essentially saturates, and x then swings between two quasi-fixed points, the extreme solutions of an equation of type (12.5). For plausibly chosen parameter values this square wave has a frequency of about 5 Hz.

One can imagine that such an oscillation might have clear synchronising properties, and we shall indeed term it ‘the escapement oscillation’. Insofar as one can localise effects, this oscillation is generated by the rescaling box in Figure 11.1. If we couple such a box to the H-net $AA^{\top}$ which follows (to be identified with the AON) then we obtain the pattern illustrated in Figure 12.4: a 63 Hz oscillation modulated by the escapement oscillation. This is indeed an effect that is observed in actual EEGs in a diffuse kind of way, but which also appears in exactly this strict form when measurements are taken more locally. It is then known as ‘neuronal bursting’; Wang and Rinzel (2003) review a range of striking examples. The explanations given for the phenomenon are usually electrochemical in character; that there are sodium, potassium and calcium currents, changing at different rates, and oscillations of one of these may be radically affected by values of another. However, this is to explain the mechanism rather than the function of the phenomenon. The argument briefly summarised in this section demonstrates neuronal bursting as a

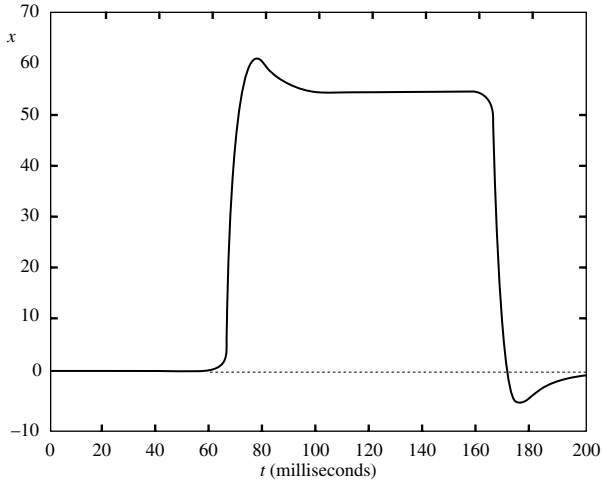


Fig. 12.3 The ‘escapement oscillation’; the almost-square slow wave generated by a Freeman oscillator with feedback and self-standardisation of the operating point.

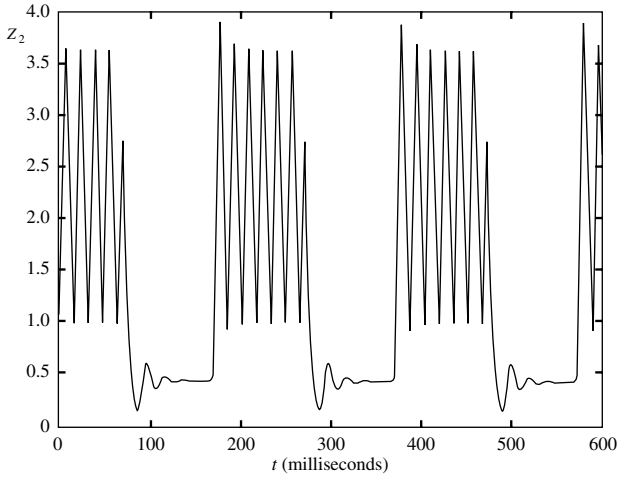


Fig. 12.4 Neuronal bursting, produced by the system of Fig. 11.2 when simple neurons in the two main blocks (the ‘OB’ and the ‘AON’) are replaced by Freeman oscillators.

consequence of oscillatory operation plus the incorporation of a rescaling unit. One can see the pattern as having a synchronising effect.

The oscillations generated by these simple models are indeed strictly periodic. However, it is a feature, both of actual EEG traces and of models linking several component structures, that the oscillations are certainly not periodic, although oscillation of a compound and statistical character is all-pervading. Freeman (1987, 1992) observes that any of the three anatomical subsystems of the olfactory bulb, the anterior olfactory nucleus and the prepyriform cortex shows a static equilibrium when unstimulated and a periodic oscillation (limit cycle) when stimulated. However, when coupled, they showed the compound, irregular oscillation so familiar from EEG traces and the like. The same behaviour was

observed for the electronic analogue that Freeman built, incorporating his own dynamic model for the neural masses (see Yao and Freeman, 1990; Kay *et al.*, 1995).

The author (1998) referred to this compound oscillation as 'chaotic', which may well be correct even in the present-day technical sense of the term, although the point was not pursued technically. It was the ubiquity of this type of signal that led Freeman to his conclusion that the transfer or processing of information in the biological system is not mediated by, for example, the mechanism of modulation and ultimate demodulation of a carrier wave of fixed frequency that is staple in radio-frequency devices. Rather, the system constitutes a dynamic whole, which can exist in many states of excitation, and communication is achieved by the locking of the component units of the system into the current dynamic state – which they might be said to 'recognise'.

III

Processing networks

By ‘processing’ we mean ‘industrial processing’, often referred to more restrictively as job-shop scheduling. Although this may be computerised, we shall not stray into the profoundly specialised and developed fields of computer processing or computer networks.

We are then concerned with the situation in which items under manufacture undergo a sequence of processing operations before they are completed. The aim is then to so schedule these that congestion at any part of the processing line is avoided, by the optimal allocation of fixed or mobile capacity.

This application does not constitute such a strong example of network optimisation, because the fact that the sequence of operations is prescribed means that the form of the network is both simple and largely preordained. Nevertheless, the application is one that in fact generated the concept of a queueing network, and raises the question of the prioritisation of competing flows.

Even here we need something of a disclaimer. The subject of scheduling is a strongly developed one in its own right, founded on both empirical experience and theoretical investigation, and to step in claiming insight would be the highest presumption. Nevertheless, the subject has itself generated at least two discernable bodies of fundamental theory. One is that which views manufacture as a queueing network, to be optimised in operation by appropriate state-dependent rules, with design optimisation then following. There are difficulties, not least those of local versus central implementation, and the possibly dire effects of misconceived priority rules.

A second development centres around the rules developed by Klimov for the optimisation of time sharing, quite closely related to the index rules deduced by Gittins in his solution of the famed multi-armed bandit problem. Recent work makes the determination of these rules relatively straightforward. These techniques supply an optimising solution in the case of extreme centralisation; extreme in that not only are all decisions made centrally, but also all processing capacity is supposed immediately reassignable to any task. This is an unreal situation, but its solution does give a clean indication of the course to be followed in the ideal case when such complete information, control and flexibility is available, and there is hope that it will supply some guide to more realistic cases.

13

Queueing networks

In this chapter we shall give an introductory coverage of those aspects of queueing models that will prove useful in the sequel, starting with the not-to-be-despised fluid model. In Chapter 14 we develop the Klimov rules for central control.

13.1 The simple queue

Let us initially regard 'work' (in the sense of work waiting to be done) as a continuous variable. Suppose that work arrives at a work-station at rate λ, and can be despatched at rate μ. Then obviously we must have $\lambda \leq \mu$ if a backlog is not to build up, and $\lambda < \mu$ if any existing backlog is to be cleared. More specifically, if we regard the backlog as a continuous variable x, then this will obey the equation

$$\dot{x} = \lambda - \mu \quad (x > 0). \tag{13.1}$$

Relation (13.1) will hold also for $x = 0$ if $\lambda \geq \mu$, but otherwise $\dot{x}$ will be zero there, corresponding to an effective work rate of λ.

This is essentially a model of a deterministic queue, of size x. It is often termed a 'fluid' model, in that work is regarded as a continuous entity, a fluid, which arrives at rate λ and is emptied through a pipe of capacity μ. This may seem rather a trivial object, but its stochastic analogue is both more realistic and more interesting. Suppose that the work arrives in quanta, i.e. as single items that demand a unit amount of work. We can regard the backlog as a queue of n items waiting for processing (or customers waiting for service), and shall replace x by the nonnegative integer n. Suppose that the random variable n follows a continuous-time Markov process in which the transitions $n \to n+1$ (arrival of a new item) and $n \to n-1$ (completion of processing of the item at the head of the queue) have probability intensities λ_n and μ_n respectively. Then μ_0 must be zero, because an empty queue cannot decrease further, but, for simplicity, we shall suppose these rates otherwise positive. Let π_n be the probability that the queue size is n when statistical equilibrium has been reached. Then the requirement of balance of the probability flux between states n and $n+1$ leads to the detailed balance relation

$$\pi_n \lambda_n = \pi_{n+1} \mu_{n+1}, \tag{13.2}$$

whence we deduce that

$$\pi_n = \pi_0 \prod_{i=1}^{n} (\lambda_{i-1}/\mu_i). \tag{13.3}$$

This will be a proper distribution (i.e. n will be finite with probability 1) if and only if expression (13.3) has finite sum. The normalising factor π_0 is then determined by equating this sum to unity.

We have actually formulated a rather more general model, a birth-and-death process. We specialise to the analogue of the queueing model by assuming constant arrival and service rates λ and μ. The distribution (13.3) then becomes, when normalised

$$\pi_n = (1-\rho)\rho^n \qquad (n = 0, 1, 2, \dots), \tag{13.4}$$

where $\rho = \lambda/\mu$ is the *traffic intensity*. Summability of expression (13.4) thus requires that $\rho < 1$. If $\rho \geq 1$ then n becomes infinite with probability one and the queue is irredeemably congested. One finds easily from (13.4) that

$$\mathrm{E}(n) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda},$$

a measure of congestion in equilibrium which we have already used in Chapter 6.

This is the simplest of all queueing models; the simple transitions allowed correspond to the assumptions that inter-arrival times are distributed independently and exponentially with expectation $1/\lambda$; correspondingly for service times, with expectation $T = 1/\mu$. There is an enormous literature, generalising the simple stochastic model in several directions. The most natural path to generalisation is to allow that service (for example) may take place in several stages; this mechanism or variants of it yields a service time whose distribution has a rational probability generating function.

In formula (13.4) we recognise the geometric distribution, with probability generating function

$$\Pi(z) = \frac{1-\rho}{1-\rho z}. \tag{13.5}$$

One can obtain other distributions by variation of the assumptions. Suppose, for example, that $\mu_n = \mu n$. This corresponds to the situation in which items, instead of waiting their turn to be processed, in fact depart independently, each with probability intensity μ. This might, for example, represent the behaviour of noninteracting gas molecules entering and leaving a chamber. In this case the distribution (13.3) becomes, when normalised, the Poisson distribution

$$\pi_n = \mathrm{e}^{-\lambda/\mu} \frac{(\lambda/\mu)^n}{n!}$$

with probability generating function

$$\Pi(z) = \mathrm{e}^{(\lambda/\mu)(z-1)}. \tag{13.6}$$

This model never saturates, because there is no interaction between individuals, and hence no congestion. In this it is most unqueuelike, its easy nature being reflected in the finiteness of the mean value λ/μ of n. This nature is reflected more fundamentally in the fact that the probability generating function (13.6) has infinite radius of convergence, in contrast to (13.5), which has radius of convergence $1/\rho$.

A point worth noting is that we have not given the items individuality, and so have not considered matters such as the waiting time (for processing) of a given item. This would be a natural consideration if the items were living, or perishable. It can only be addressed if one specifies a *queue discipline*; a rule that identifies individual items and states which item in the queue should next be processed. Classical queueing theory addresses such matters; we hope not to need to.

13.2 The multi-station case; deterministic treatment

Most installations will have several work-stations, carrying out different processes. Calculations are very simple if we again take the quasi-static deterministic formulation with which we began the last section.

Just for this section, then, we regard all variables (such as numbers of items) as continuous variables and all operations as deterministic. Suppose that several types of item are being manufactured, the types being labelled by $v = 1, 2, \ldots$ The resources devoted to manufacture consist of both work-force and plant. However, seeing the system is to be a fixed one, we can reduce these to the single resource of 'effort' (per unit time). Suppose that there are several manufacturing operations, which we can identify with work-stations indexed by $j = 1, 2, \ldots$ It is possible that items must make repeat visits to a given work-station. Let us suppose that the sth visit of an item of type v to work-station j will take a time T_{vjs}/a_{vjs} if effort is devoted at rate a_{vjs} to this combination. Seeing that items of type v are entering this phase of processing at rate λ_v we must have

$$\lambda_v T_{vjs}/a_{vjs} \leq 1$$

if progress in the phase is to keep pace. The total effort rate A available must thus satisfy the sharp bound

$$A \geq \sum_v \sum_j \sum_s \lambda_v T_{vjs}. \tag{13.7}$$

If strict inequality holds in (13.7) then there is spare capacity, and one has the latitude to begin to prioritise; to favour some class of items that one would wish to process more quickly. However, such latitude is necessary for quite other reasons. The inevitable statistical variability will cause a degree of congestion that will be unacceptable unless there is adequate capacity in hand. We seek then a multi-station version of the stochastic model of Section 13.1.

Latent in this very collapsed version of the model is a genuine dynamic structure. The backlogs of different types and stages of items at each of the work-stations constitute a

system of queues, whose sizes will obey a system of equations of type (13.1). However, it is easiest to go directly to the stochastic model, which gives a direct and general appreciation of the issues.

13.3 Jackson networks

The manufacturing example specifies a network whose nodes can be identified as work-stations, each with its attendant queue of jobs in a stochastic model. A classic model of a queuing network developed for just this application is the *Jackson network*, a model at once amenable, elegant and useful. Particular cases of it were noted by R. R. P. Jackson (1954) and a more general structure then perceived by J. R. Jackson (1963). To rediscover the Jackson distribution became almost a rite of passage at one time, realised of course only after the event. The author experienced this rite, although also observing the partial balance of the model and generalising to its open version. Advances of real novelty were made by Kelly, notably the achievement of routing by use of a changing type-label noted below, and development of the concept of quasi-reversibility; see Kelly (1979) and references.

Consider a network of m nodes whose state is described by the vector $n = \{n_1, n_2, \ldots, n_m\}$, where n_j is the number of items at node j. For the moment we assume these items statistically identical; the notion of classes is easily incorporated once the analysis of the basic model is clear. The nodes can be regarded as representing 'queues' insofar as the rates of transition between nodes depend upon n, the state of the net. Let $\lambda(n, n')$ denote the probability intensity of the transition $n \to n'$ and e_j denote the m-vector with a unit in the jth place and zeros elsewhere. Then a Jackson net is specified by the assumptions

$$\lambda(n, n+e_j) = \lambda_j, \qquad \lambda(n, n-e_j+e_k) = \frac{\Phi(n-e_j)}{\Phi(n)}\lambda_{jk}, \tag{13.8}$$

for $j = 1, 2, \ldots, m$ and $k = 0, 1, 2, \ldots, m$. The extra node value $k = 0$ corresponds to passage out of the system, and we can then consistently set $e_0 = 0$. The function $\Phi(n)$ is specified; it must be regarded as zero if any component of n is negative (so that there can be no departures from an empty queue), and it must be such as to yield a proper equilibrium distribution for n. Thus the first expression gives the arrival rate of items to node j and the second gives the transition rate of items from node j to node k, this being taken as departure from the network in the case $k = 0$. Departure is permanent in that $\lambda_{0j} = 0$. The special feature of the network is that the transition rates depend upon n through the single function $\Phi(n)$, with the consequence that this rate is affected by conditions at the donor node j but not at the receptor node k. This special dependence of rates upon system state gives the model a degree of time-reversibility, which in turn explains its amenability. The n-independent factor λ_{jk} determines the routing between nodes, and constitutes the jkth element of the *routing matrix*.

Let $w = \{w_1, w_2, \ldots, w_m\}$ be the solution of the set of linear equations

$$\lambda_j + \sum_{k=0}^{m}(w_k\lambda_{kj} - w_j\lambda_{jk}) = 0 \quad (j = 1, 2, \ldots, m), \tag{13.9}$$

where w_0 in fact makes no appearance. These are known as the *traffic equations*. They are the equilibrium equations of a deterministic and 'atomised' flow, in which work arrives at node j at rate λ_j and departs thence for node k at rate $w_j\lambda_{jk}$, where w_j is the amount of work outstanding at node j. This is not to be confused with the 'fluid limit'; a deterministic model which follows the queueing rules, as in (13.1).

It is a direct matter to confirm from the Kolmogorov forward equation for the net that the occupation vector n has the equilibrium distribution

$$\pi(n) \propto \Phi(n) \prod_j w_j^{n_j}. \tag{13.10}$$

This is the unique solution if there is no subset of nodes that is absorbing, in that items that once enter it can never leave. The remarkable feature of the Jackson network is that its equilibrium distribution is so simple. In the particular case when the function Φ has the factorisation

$$\Phi(n) = \prod_j \Phi_j(n_j)$$

relation (13.10) will in fact imply that the occupation numbers n_j are distributed independently in equilibrium. Given that the routing matrix (λ_{jk}) can take any value consistent with escape from the net's being always possible (and so certain), this conclusion is remarkable.

An assumption that particles move independently through the net would imply a transition rate $\lambda(n, n - e_j + e_k) = \lambda_{jk} n_j$, so that (13.10) holds with $\Phi(n) \propto \prod_j (n_j!)^{-1}$. That is, the n_j are independent Poisson variables, corresponding to Boltzmann statistics. In the case when $\Phi(n)$ equals unity (say) for all elements of n nonnegative, and is zero otherwise, then we see from (13.10) that the n_j follow independent geometric distributions, corresponding to Bose–Einstein statistics. This is just the case when all the nodes behave as simple queues (as understood in Section 13.1).

Finiteness of n, and so the impossibility of irrevocable congestion, will require that the distribution (13.10) be summable. We can put this another way, by defining the generating function

$$\tilde{\Phi}(z) = \sum_n \Phi(n) \prod_j z_j^{n_j},$$

where z is an m-vector of complex variables z_j. Let $\mathcal{C}$ denote the set of z for which $\tilde{\Phi}(z)$ is absolutely convergent. Then we could express the requirement that n be finite by the condition that w should lie in $\mathcal{C}$; this imposes a condition on the routing matrix for given Φ. One could say that congestion becomes ever more serious as w approaches the boundary of the closure of $\mathcal{C}$. Just where it approaches the boundary will also locate the source of congestion in some degree.

A generalisation that is as powerful as it is simple is to suppose that the items moving around the network can be of a changing type as well as changing work-station. Let $n_{\alpha j}$ be the number of items of type α at work-station j, and n be the vector of these. Then if we replace 'item state' j by the compound state αj then everything works out exactly as above. Explicitly, the assumptions

$$\lambda(n, n + e_{\alpha j}) = \lambda_{\alpha j}, \qquad \lambda(n, n - e_{\alpha j} + e_{\beta k}) = \frac{\Phi(n - e_{\alpha j})}{\Phi(n)} \lambda_{\alpha j, \beta k}$$

imply, under mild regularity conditions, the equilibrium distribution

$$\pi(n) \propto \Phi(n) \prod_{\alpha} \prod_{j} w_{\alpha j}^{n_{\alpha j}},$$

where the $w_{\alpha j}$ solve the traffic equations

$$\lambda_{\alpha j} + \sum_{\beta} \sum_{k=0}^{m} (w_{\beta k} \lambda_{\beta k, \alpha j} - w_{\alpha j} \lambda_{\alpha j, \beta k}) = 0$$

for $j = 1, 2, \ldots, m$ and all α.

For the manufacturing example this variable label α is to be identified with the item type v plus a statement of what stage the item has reached in its processing. Kelly (1979) was the first to use such a label to guide an item through a prescribed sequence of operations.

13.4 Optimisation of effort distribution

Let us return to the manufacturing example of Section 13.2. If the sequence of operations is exactly prescribed for all operations, and if any given operation is concentrated in a single work-station, then the only scope one has for optimisation of the network is the allocation of spare capacity. Under these restrictive assumptions, the rate at which items of type v enter a given phase js (the sth visit to station j) is either zero or λ_v. The rate at which they leave is $a_{vjs}\mu_{vjs}$, where μ_{vjs} can be identified with the T_{vjs}^{-1} of the deterministic treatment, and can be taken as infinite if items of type v never enter phase js. The occupation numbers n_{vjs} are thus independently and geometrically distributed with respective traffic rates $\lambda_v / a_{vjs}\mu_{vjs}$. If items of type v are given a weighting c_v, then we choose the effort variables a to minimise the criterion function

$$\mathrm{E}\left(\sum c_v n_{vjs}\right) = \sum \frac{c_v \lambda_v}{a_{vjs}\mu_{vjs} - \lambda_v}, \tag{13.11}$$

where the summation is over all relevant v, j and s. The minimisation is subject to the resource constraint

$$\sum a_{vjs} \leq A. \tag{13.12}$$

The optimising allocation of the cost resource is then

$$a_{vjs} = \frac{\lambda_v}{\mu_{vjs}} + \theta \sqrt{\frac{c_v \lambda_v}{\mu_{vjs}}}, \tag{13.13}$$

where θ is a parameter to be adjusted to yield equality in the constraint (13.12). This rule is consistent with inequality (13.7) (if we identify μ with T^{-1}), and shows very clearly how the spare capacity has been allocated in response to the minimal congestion criterion. With θ and the effort allocation a determined from (13.12) and (13.13) we find that the minimal value of the performance criterion (13.11) is

$$\frac{\left[\sum \sqrt{c_v \lambda_v / \mu_{vjs}}\right]^2}{A - \sum \lambda_v / \mu_{vjs}}. \tag{13.14}$$

The Jackson model has its limitations, however. It cannot represent assembly or disassembly of items. Neither is it 'smart', in that it might respond to current network state n by a variation of routing or a prioritisation of jobs designed to control the value of n in real time. One can favour one category of item over another by variation of the routing rates $\lambda_{\alpha j,\beta k}$, but this gives no response to the actual numbers present. The n-dependence of the rates specified in (13.8) does not amount to a genuine state-feedback, but rather to an expression of the physical laws of the queues. In particular, there is no specification of Φ that achieves a prioritisation; see the note below. Nor is there any sensitivity to numbers in the target node. In the next chapter we develop rules that do respond optimally to the types of work that are actually present in the system. These rest however on a rather unrealistic assumption: that it is possible to instantaneously focus all resources on the highest-priority job present.

Note Suppose that one has a queue of two types of item, of numbers n_1 and n_2, and that type 1 is served at a rate proportional to $\Phi(n_1-1,n_2)/\Phi(n_1,n_2)$. This ratio should then be zero for $n_2>0$, if one wishes type 2 to have priority. There is no proper Φ that achieves this.

13.5 Queueing networks more generally

The once waning topic of queueing has taken on a new lease of life since applications in communications, computation and scheduling widened the interest to queueing networks. There is now an enormous literature, with journals devoted entirely to the subject.

Queueing models are intrinsically stochastic, although they may have a natural 'fluid limit' that behaves deterministically. Any notion of acceptable performance requires convergence with time to a unique equilibrium regime, in which the queues are finite for the stochastic case, and usually zero for the deterministic case. This is then a concept of stability. Quite soon one strengthens this to a concept of optimality, in which the operational rules of the system (e.g. the queue disciplines) are chosen to optimise some criterion. The aspect over which one often has choice is the allocation of server effort over different classes of customers, implying the formulation of precedence rules. Questions of design can of course be considered only once those of operation have been clarified.

However, the matter is subtler than one might think. Seminal has been the work of Rybko and Stolyar (1992) and Bramson (1994a,b), demonstrating that the introduction of precedence rules or even of quite plausible queue disciplines can result in deadlock and an irretrievably congested network. This can be so even in the fluid limit. The Braess paradox has been observed also for queueing networks. On the positive side, Bramson (1996) develops conditions that assure convergence to an equilibrium. Dai (1995) shows that under certain conditions convergence of the fluid model to an equilibrium will imply the same for an appropriate stochastic version of the model, and there is a strong Russian literature on the subject, headed by Rybko.

However, there are examples where optimisation (of routing and of design) is more straightforward. The simplest case is that of a telephone 'loss network', for which an incoming call is accepted only if a complete path through the network to its destination can be found for it immediately. Stability is thus achieved by fiat, and there are no queues at all (if no further account is taken of rejected calls). This is a 'circuit-switched' network,

in that a complete switching path is prepared for the call.An obvious fluid limit version of the model reduces optimisation of admission, routing and design to a simple linear programme; see Chapter 15.

More generally, one can say that the material of Part I provides a basis for the optimisation of quite a general class of fluid limit models, the 'routing rules' being determined by the minimising principles for flow cost adopted there and the design rules by similar principles for the consequent material cost. The treatment of Part I also takes account of the longer-term stochastics of regime variation and of the costs of environmental intrusion.

A different type of stochastic effect is that of short-term variation in input and in the state of the network (i.e. those links or servers that are engaged). Kelly (1988a,b) has shown how routing can be guided by stochastic versions of the shadow prices generated in the linear programming formulation. Admission rules can be made optimally responsive to system state by the simple technique of 'trunk reservation' plus possibly a quick indicative observation on system state; see Chapter 16 and references there.

A 'packet-switched' network is one in which a packet (customer, item or portion of a file) can be accepted with no guarantee of route or route completion; the actual routing often being revised in the light of changing system state as the packet progresses through the system. Queues can indeed then form at the nodes of the system (the 'routers') and one has a genuine queueing network. Much effort has been put into the determination of decision rules at these nodes that will optimise some criterion of performance. Not only does one require optimality of decisions as functions of system state; one requires practicality in that both observation and control should be decentralised. That is, decisions at a node should depend as far as possible on what is easily observable at that node (so that relevant global aspects of state must be made easily observable). There have been a number of ad hoc attacks on the problem over the years; an early convincing proposal being that by Gallager (1977). Routing problems in communication networks have their own special character; see Steenstrup (1995). In the processing context, routing (i.e. the sequencing of operations) is prescribed to a high degree; latitude lies rather in the prioritisation of jobs and, to some extent, in the distribution of processing effort over the work-stations.

A good deal of effort has gone into the seeking of insight by the study of quite small and special systems. Harrison and coauthors (Harrison, 1985; Harrison and Wein, 1990; Harrison and Nguyen, 1993) have devoted particular study to systems operating in heavy traffic, when one can reasonably approximate discrete Markov models by Brownian (diffusion) models. Kelly and Laws (1993) likewise explored this approach, and observed the effective pooling of efforts by servers. This line of work has developed greatly in recent years, the applicability of Brownian methods in heavy traffic being seen as related to the phenomena of state space collapse and resource sharing (see e.g. Bramson, 1998; Bramson and Dai, 2001; Harrison, 2000, 2003; Williams, 1998; Kelly and Williams, 2004).

Resource sharing of course occurs in its purest form when there is complete flexibility in processing, in that processing effort can be switched without constraint or delay to any work-station. It is just in this case that the ideal optimum is attainable and by relatively simple rules. Klimov (1974, 1978) showed that an index could be calculated that would give a complete prioritisation list for tasks in the system; optimal performance was

guaranteed if all effort was concentrated on the job highest on the list. Whittle (2005) showed that methods previously used to solve the multi-armed bandit problem (Gittins, 1979, 1989; Whittle, 1980, 1981a) gave an efficient determination of the index. This approach is described in Chapter 16.

The catch is, of course, that processing effort can by no means be switched so readily; every work-station usually has its own fixed installation. One rather natural suggestion is made in Section 16.5 for adaptation of the priority rules to cover this point, but the problem stands. A recent paper employing programming methods is that by Stolyar (2005).

Early studies of design in the communication context are found in Kleinrock (1964) and Gerla and Kleinrock (1977).

The huge amorphous queueing networks constituted by the Internet and the Worldwide Web present special problems, because so many users are seeking to gain access, and because observation of network state, by anybody at all, is in the highest degree sketchy. One redeeming feature is that the routers that constitute the 'nodes' of present-day networks are so fast that the question of queue discipline is almost irrelevant. Moreover, observations on packet experience during passage yields enough diffuse information on the state of congestion in the system that this, together with special 'fairness' criteria written into the operational protocols, achieves remarkably smooth running. We give some detail and references in Chapters 17 and 18.

13.6 A small coda

The word 'coda' comes from the Latin 'cauda', for a tail, and so is used to denote a tail-piece in music or ballet. It survives in Italian also as the word for 'queue', and the two words are in fact cognate.

There is one approach to queueing systems which leads to striking simplifications when it is applicable, and it is applicable not infrequently. A problem that arose early but resisted analysis is that of determining the distribution of queue lengths in a system in which statistically identical arrivals can choose any one of m statistically identical queues, and so of course choose the shortest. A direct investigation of the performance of this rule nevertheless proves very difficult. However, Vvedenskaya *et al.* (1996) considered a variation of the problem that was easily treated, demonstrated the powerful improvement in performance that such choice implies, and found new application for a generally useful technique. Suppose the problem scaled up, in that there are N queues, each completing service at rate μ, and customers arrive at rate $N\lambda$. We assume a Markov model in that these rates are all transition intensities. It is supposed that a customer, on arrival, chooses a set of m queues at random and then joins the shortest of these. What then is the joint distribution of queue sizes?

It is plausible that, if N is very large and m is fixed, then in equilibrium the queue sizes in the sample of m will be independently and identically distributed. This is proved in the article quoted, and the technique, essentially a mean field technique, carries over to similar cases. Assume then that queue size n has equilibrium distribution π_n and define

$$u_n = \sum_{k \geq n} \pi_k,$$

the probability that the queue size is at least n. Then, under the proposed policy

$$\dot{u}_n = \lambda(u_{n-1}^m - u_n^m) - \mu\pi_n. \tag{13.15}$$

This is because the coefficient of λ is the probability that all queues in the sample have length not less than $n-1$ and at least one has length exactly $n-1$. This will then be the rate per arrival at which the proportion of queues not less than n in length can increase. Adding the equilibrium version of (13.15) over $n > i$ we thus deduce that

$$u_{i+1} = \rho u_i^m \qquad (i \geq 0).$$

From this it follows that

$$u_n = \rho^{(m^n-1)/(m-1)}.$$

This is of course consistent with the conventional result $u_n = \rho^n$ for the case $m = 1$ of the simple queue, but demonstrates how rapidly queue size decreases with increasing m.

14

Time-sharing processor networks

The Jackson model of Chapter 13 was a natural one, whose behaviour was easily analysed. However, it allowed no state feedback, and so could not be induced to respond to current network state. Various attempts have been made to modify the model to achieve such feedback (see Section 13.5), but these have not led to what one might regard as the natural successor in the next generation of models.

However, one can indeed find a model with a full theory if one goes to the other extreme: of a model that is completely centralised, in that the optimiser can deploy all processing resources freely and instantaneously to any part of the network. There is then a full and exact theory for the optimisation of this deployment, in the light of current network state. Of course, such assumptions are unrealistically extreme; processing units can be deployed only at the work-station to which they are attached. Moreover, one would wish for a control rule of a decentralised nature, in that operators make local decisions largely on the basis of local information.

Nevertheless, this model (based on the multi-armed bandit) does mark a genuine advance, and does suggest policies for the case when resource deployment is free only within a work-station (see Section 14.5).

The model is similar to that of Chapter 13, in that there is a queue at every node (or a multi-class queue at every work-station, in the more explicit version). However, we shall speak of it as a time-sharing network rather than a queueing network, to emphasise that processing effort is allocated centrally rather than locally.

14.1 The fluid and Markov models

We consider the fluid limit first; the treatment of the stochastic case then follows in complete analogue. We suppose initially that 'work' is flowing between the nodes ($j = 1, 2, \ldots, m$) of a network; this formulation is general enough that we can later represent the different types and stages of the items being processed. We suppose that, if effort is exerted at rate a_j at node j, then work that is present will be despatched and sent to node k at rate $a_j\mu_{jk}$. The backlogs (or queues) x_j of work at the nodes j will then obey the equations

$$\dot{x}_j = \lambda_j + \sum_k (a_k\mu_{kj} - a_j\mu_{jk}) \quad (j = 1, 2, \ldots, m), \tag{14.1}$$

where λ_j is the rate at which work arrives at node j. The sum covers also the state $k = 0$, of completion and discharge from the system. We assume that items once completed are

never returned, so that μ_{0k} is zero for all k. The rates of effort a_j constitute the control variables, and their choice as functions of network state constitute the control policy. They must in any case be chosen so that $\dot{x}_j$ is never negative if x_j is zero; queues can never run negative.

The term $a_j\mu_{jk}$ replaces the term $w_j\lambda_{jk}$ that occurred in the traffic equations of Section 13.3. The intrinsic rate of working at node j is now proportional, not to the amount of work w_j standing at the node, but to an effort rate a_j that is to be specified in terms of the state of the net by the operating policy. We have written the routing rates as μ_{jk} rather than λ_{jk}, largely to emphasise that the queues at the nodes are now all simple queues.

This is then a specification of the model in the fluid limit. A plausible formulation of the optimisation problem, starting from time $t = 0$, might then be to choose the nonnegative effort rates a_j so as to minimise the weighted integral of work in the system

$$C = \int_0^\infty \sum_j c_j x_j \mathrm{d}t \tag{14.2}$$

subject to an overall resource constraint

$$\sum_j a_j \leq A \tag{14.3}$$

where the time-dependence of x and a is understood. The coefficients c_j measure the 'cost' of backlog at the nodes j, and are constant in time.

One will not be able to achieve a finite limit on expression (14.2) unless resources are adequate, i.e. unless A is sufficiently large. One can determine how large it must be by choosing appropriate *constant* values for the rates a_j. Suppose we write equation (14.1) in vector form as

$$\dot{x} = \lambda - Ma.$$

The inflow λ would then be exactly balanced if we chose $a = M^{-1}\lambda$, so incurring a total effort rate of $\mathbf{1}M^{-1}\lambda$, where $\mathbf{1}$ is a row m-vector of units. The inequality

$$\mathbf{1}M^{-1}\lambda < A \tag{14.4}$$

is then the classic necessary and sufficient condition for stabilisability of the system. That is, that A be such that x can ultimately be reduced to zero ('the system be emptied') from any starting value.

The fixed allocation of effort $a = M^{-1}\lambda + \delta$ for some positive δ (adapted to keep x nonnegative) will work in the fluid limit, but we seek a rule that is optimal. The stochastic case will also demand that this rule be expressed in 'closed-loop' form, i.e that a should be determined as a function of current observables.

The inclusion of the factor a_j allows us to vary the effort associated with a particular node. However, one could similarly fix the effort to be devoted to a much more specific activity, e.g. the processing of an item of a given type at a given station for passage to a given destination.

In the analogous Markov model we replace x_j by n_j, the actual number of items waiting for processing at node j. The rate λ_j is now the probability intensity of the transition $n \to n + e_j$, the arrival of a new item at node j. The rate $a_j \mu_{jk}$ is the probability intensity of the transition $n \to n - e_j + e_k$ if $n_j > 0$; the transition is otherwise impossible. Criterion (14.2) is now replaced by

$$\gamma = \mathrm{E}(\sum_j c_j n_j) = \mathrm{E}(c^\top n), \tag{14.5}$$

where the expectation is that holding under equilibrium conditions. There are degrees of optimality; one may hope that not merely is the average cost γ minimised, but also the transient cost incurred in reaching equilibrium.

The resource condition (14.3) will still hold, and inequality (14.4) in fact remains the necessary and sufficient condition for the system to be stabilisable for a given value of A. Again, the model is general enough that, with slight redefinition, items are effectively labelled by their type and current stage of processing. The optimal effort–allocation rule turns out to be exactly the same for the fluid and stochastic versions of the problem.

14.2 The Gittins–Klimov index

The considerations of the last section may seem quite simple, but the optimisation problem posed is celebrated for its difficulty, and long resisted solution. We shall phrase the discussion for the Markov model, which is the more general (in that the fluid model is a limit version of it), which demonstrates more clearly the need for state feedback. It also relates more immediately to the standard treatments of the multi-armed bandit problem, now to be invoked.

If the current network state n can be observed, then the optimal equilibrium rule for the distribution of effort a will be in terms of it. The allocation a is of course subject to the constraints $a \geq 0$ for all t. If the only other constraint is the overall capacity constraint (14.3), then we have in fact a version of the so-called multi-armed bandit problem, which is now well understood. As emphasised above, effort can be reassigned neither arbitrarily nor immediately in a practical situation. Nevertheless, study of this extreme case certainly has value, and we shall consider relaxation of conditions in Section 14.5.

The classic multi-armed bandit (MAB) problem concerned the allocation of effort over a set of given ‘projects’, each of which followed Markov transitions through its various states if effort was allocated it, but otherwise remained in its current state. Projects in operation yielded a state-dependent reward, and one sought an allocation of effort that maximised the expected discounted total reward over time. A ‘project’ corresponds to an individual item in our case, and we see immediately that our work-flow case differs from the original MAB formulation in two respects. Firstly, the process is ‘open’, in that new items continue to arrive for processing and completed items are discharged. Secondly, an item going through the system does not so much yield a reward every time it receives some attention, as incur cost during the whole period it spends in the system, from entry to discharge. Problems of this nature are termed ‘tax’ problems; they are related to the MAB, but show important differences.

Both of these aspects can be dealt with. The theory is substantial but relevant; we give an outline of it in Appendix 2, together with references to the literature. The essential conclusion is that items in a given state j (i.e. at node j) can be assigned an index ν_j, the *Gittins index*, and that the optimal policy is to concentrate effort entirely on those items currently in the system that have the greatest index.

Gittins' approach to the multi-armed bandit problem, successful also for many of its variants, involved the idea that, as well as the options of selecting one from a number of projects for operation, one had the option of giving up the whole venture and retiring on an 'income' of ν. Gittins' advance was to see that the problem could be solved by considering the simple alternatives of operating a single given project or of accepting the retirement option. The value of ν at which the optimal decision changed was critical, and would depend upon the state of the project. Denote this critical value by ν_j, the *index* for this particular project when it is in state j. It then turns out that for the original problem the *Gittins index policy* is optimal: that which concentrates effort on those projects that are currently of greatest index.

This policy, appropriately modified, remains optimal under the generalisations to an open process and to a tax problem. However, the tax problem had been tackled independently by Klimov (1974, 1978), who demonstrated optimality of an index policy, although with an evaluation of the index taht was recursive rather than via the enlightening optimal stopping property perceived by Gittins. Honour should then be shared between both authors, but in the tax context it is perhaps fair to speak of the 'Klimov index'.

The treatment of these matters in Section 14.1 leads us to the point of immediate relevance. The equations determining the index in our formulation for the case $A = 1$ are

$$\max[\sum_k \mu_{jk}(\Delta_k - \Delta_j) - \nu + \chi, -\psi_j] = 0 \qquad (j = 1, 2, \ldots, m), \tag{14.6}$$

where

$$\Delta_j = \psi_j - c_j$$

and

$$\chi = \sum_k \lambda_k \Delta_k.$$

This characterises an optimal stopping problem, in which ψ_j is a measure of reward incurred before stopping and ν is a fixed reward rate which continues after stopping. The index ν_j is the value of ν that makes both terms in the square bracket of (14.6) zero, i.e. makes the choices of continuation or of stopping equally attractive. If the relation of the equation system (14.6) to the original problem seems obscure, this is very much because of the special features of a tax problem, explained in Appendix 2. For the case of general A one simply multiplies all the rates μ_{jk} by A. When it comes to determination of the indices, calculations are very much simplified if we work with the modified version

$$\max[\sum_k \mu_{jk}(\Delta_k - \Delta_j) - \alpha, -\psi_j] = 0 \qquad (j = 1, 2, \ldots, m), \tag{14.7}$$

of (14.6), where then

$$\alpha = \nu - \chi = \nu - \sum_k \lambda_k \Delta_k. \tag{14.8}$$

This then gets rid of the immigration effects, and leads to a much easier determination of the critical values of the parameter α; these are what would be the critical index values if there were no immigration. However, ultimately we must revert to the more general case, and so must determine the relation between the critical values of the two parameters, α and ν. A few preliminaries are needed.

If α is large enough then the second option will hold in (14.7) for all j, and we shall have $\psi_j = 0$, or $\Delta_j = -c_j$, for all j. As we reduce α progressively then the option will every now and again change in one of the relations (14.7). Let us denote the values of α at which this happens by $\alpha_1, \alpha_2, \ldots$: a nonincreasing sequence. We can retrospectively reorder the states so that it is ψ_i which changes option when α goes through the critical value α_i, which we interpret as what the index value for state i would be if there were no input of fresh items.

Suppose that α takes a value between α_i and α_{i+1}. Then the relations

$$\sum_k \mu_{jk}(\Delta_k - \Delta_j) = \alpha \qquad (j \le i), \tag{14.9}$$

$$\Delta_j = -c_j \qquad (j > i) \tag{14.10}$$

hold. These have solution

$$\Delta_j = -\alpha\tau_{ij} - \sum_{k>i} P_{ijk}c_k \qquad (j \le i), \tag{14.11}$$

where τ_{ij} is the expected time an item starting in state j takes to escape from the set of states $\{1, 2, \ldots, i\}$ under the transition rules μ_{jk}, and P_{ijk} is the probability that the item ends in state k when it has done so. Inserting solution (14.11) into (14.8) we deduce the relationship

$$\nu = \alpha - \sum_{j\le i} \lambda_j(\alpha\tau_{ij} + \sum_{k>i} P_{ijk}c_k) - \sum_{j>i} \lambda_j c_j \qquad (\alpha_i \ge \alpha \ge \alpha_{i+1}). \tag{14.12}$$

This equation also relates the indices α_j and ν_j for j equal to i or $i+1$. Note that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\alpha} = 1 - \sum_{j\le i} \lambda_j \tau_{ij} \qquad (\alpha_i \ge \alpha \ge \alpha_{i+1}). \tag{14.13}$$

The right-hand side is positive for all i, by the assumption that the system has the capacity to process the input. We thus see that ν increases monotonically with α.

14.3 Examples

The simplest case would be that of a single stage of processing with p types of item. This is just the case of a multi-class queue with a single server, and it has long been known that one should first process an item present for which $\mu_j c_j$ is maximal. Here c_j is the cost that an item of type j incurs per unit time while it is in the system, and μ_j is the rate at which such an item is processed, using unit effort. Let us nevertheless confirm that the formal index approach yields this result.

We take an item in state j as being an item of type j that has not yet been processed. Relation (14.7) then becomes

$$\max[\mu_j(c_j - \psi_j) - \alpha, -\psi_j] = 0, \tag{14.14}$$

since the only possible transition is for the item to leave the system after processing and incur no further cost. We see from (14.14) that the index values must be $\alpha_j = \mu_j c_j$, consistently with the known optimal policy. Let us again assume the states j ordered so that α_j is nonincreasing. Then the relation (14.12) between ν and α now becomes

$$\nu = (1 - \sum_{j \le i} \rho_j)\alpha - \sum_{j>i} \lambda_j c_j \qquad (\alpha_i \ge \alpha \ge \alpha_{i+1}),$$

where $\rho_j = \lambda_j/\mu_j$. This then leads to the index evaluation

$$\nu_i = (1 - \sum_{j \le i} \rho_j)\mu_i c_i - \sum_{j>i} \lambda_j c_j = (1 - \sum_{j<i} \rho_j)\mu_i c_i - \sum_{j \ge i} \lambda_j c_j.$$

A second case that is certainly realistic and basic in this context is that of a single product. Suppose that it requires m stages of processing, and let these be numbered in order of number of stages to go, so that items enter the system in state m (at rate λ, say) and then progress through states $m-1, m-2, \ldots$ Ultimately they reach state 0, which can be regarded as completion and discharge from the system. Suppose that items incur cost at rate c at all stages, and leave state j for state $j-1$ at rate μ_j if unit effort is exerted. Relation (14.7) then becomes

$$\max[\mu_j(\Delta_{j-1} - \Delta_j) - \alpha, -\psi_j] = 0, \tag{14.15}$$

with $\Delta_0 = 0$, effectively.

One finds readily from (14.15) that, as α is decreased from a sufficiently large value, states 1, 2, 3,... successively become active (i.e. adopt the first option in (14.14)). The indices α_j are thus decreasing in j, and the recommendation is that all effort should be concentrated on an item present that is nearest completion. We find the evaluation

$$\alpha_j = c/T_j, \tag{14.16}$$

where

$$T_j = \sum_{k=1}^{j} \mu_k^{-1} \tag{14.17}$$

is the expected time needed to complete the last j stages of processing, on unit effort.

Relation (14.12) becomes

$$\nu = \begin{cases} \alpha - \lambda c, & \text{if } \alpha < \alpha_m, \\ \alpha(1 - \lambda T_m), & \text{if } \alpha \ge \alpha_m. \end{cases} \tag{14.18}$$

From this and (14.16) we thus derive the expression

$$\nu_j = \frac{c}{T_j}(1 - \lambda T_j) \tag{14.19}$$

for the actual index, valid for all relevant j. In the case when effort is available at rate A we effectively multiply the μ_j by A, so that (14.16) becomes

$$\nu_j = \frac{c}{T_j}(A - \lambda T_j). \tag{14.20}$$

The condition $A - \lambda T_m > 0$ is just the condition that the system should be able to cope with the input.

If there are several such lines for different products then the indices for each line can be determined independently by formula (14.20). These separate index lists can then be interleaved to determine the complete order of priority for the different product/stage combinations. Of course, matters change when reallocation of effort is possible only within a work-station rather than system-wide, a point we consider in Section 14.5.

14.4 Performance of the index policy

The index policy is not unworkable; evaluation of the indices from the equation systems (14.6) or (14.7) is laborious, but is a once-and-for-all job which provides a preference listing of the options. However, for all that the policy is known to be optimal, one should have an evaluation of its performance. This is partly to see how performance depends upon parameters such as A, and partly to provide a benchmark for other, suboptimal, policies. Furthermore, it is via the performance measure that one adapts the policy to other circumstances.

There are at least two levels of aspiration to optimality. One is simply to minimise the average cost γ, defined in (14.5). A stronger aspiration is to minimise also the *transient cost* $f(n)$: the transient incurred in passage from an initial state n to the equilibrium regime. As it turns out, the index policy is optimal at both levels.

For the fluid model of Section 14.1 the average cost is zero if the system is stabilisable, because then the control will empty it. This will be the case for any stabilising policy, so one looks at the second criterion, that of minimal transient cost. Under the index policy, known to be optimal, one will at any stage apply fixed controls designed to bring the positive queue of maximal index down to zero, while holding queues of higher index at zero. Queues will then vary linearly in time during this phase, and the integrated cost incurred over the phase will be quadratic. The total transient cost will then necessarily have the quadratic form

$$f(x) = \frac{1}{2}x^\top Q x. \tag{14.21}$$

Suppose the nodes $j = 1, 2, \ldots, m$ are so ordered that ν_j is nonincreasing in j. Then it is shown in Whittle (2005) that

$$Q = \int_0^\infty \frac{\partial \psi(\nu)}{\partial \nu}\frac{\partial \psi(\nu)}{\partial \nu}^\top \mathrm{d}\nu = \left[\int_0^\infty \frac{\partial \psi_j(\nu)}{\partial \nu}\frac{\partial \psi_k(\nu)}{\partial \nu}^\top \mathrm{d}\nu\right], \tag{14.22}$$

where $\psi(\nu)$ is the solution to system (14.6) for given ν. Alternatively, the elements Q_{jk} of Q can be determined from the equations

$$c_i + \sum \mu_{jk}(Q_{ki} - Q_{ji}) + \sum_k \lambda_k Q_{ki} = 0 \qquad (1 \geq i \geq j \geq m). \tag{14.23}$$

For the Markov case, we have the evaluations

$$f(n) = R^\top n + \frac{1}{2} n^\top Q n, \qquad \gamma = \sum_k \lambda_k f(e_k) = \sum_k \lambda_k (Q_{kk}/2 + R_k), \tag{14.24}$$

where Q is the matrix determined above, and the coefficients R_j are determined by the equation system

$$\sum_k \mu_{jk}(R_k - R_j) + \frac{1}{2}\sum_k \mu_{jk}(Q_{jj} - 2Q_{jk} + Q_{kk}) = 0 \qquad (1 \le j \le m). \tag{14.25}$$

Recall that the effect of allowing for a general A is to multiply all service rates μ_{jk} by A.

Let us for example consider the case of a single product line. We find from equation (14.23) that the elements of the symmetric matrix Q are

$$Q_{jk} = \frac{cT_j}{1 - \lambda T_m}[1 - \lambda(T_m - T_k)] \qquad (j \le k), \tag{14.26}$$

and from (14.25) that

$$R_j = \frac{c}{2}\left[T_j + \frac{\lambda}{1 - \lambda T_m}\sum_{k \le j} \mu_k^{-2}\right]. \tag{14.27}$$

Recall that to recover the case of general A we multiply the rates μ_j by A, and consequently divide expression (14.17) for the expected passage times T_j by A. The formula for the minimal expected cost γ in (14.24) then yields

$$\gamma = \frac{\lambda c}{2(1 - \lambda T_m)}\left[T_m(2 - \lambda T_m) + \lambda \sum_{k=1}^{m} \mu_k^{-2}\right]. \tag{14.28}$$

Expression (14.28) may not seem particularly transparent, but it is certainly smaller than the cost

$$\gamma_J = \frac{\lambda c}{1 - \lambda T_m}\left(\sum_k \mu_k^{-1/2}\right)^2, \tag{14.29}$$

which would have held if the line had simply been treated as a sequence of balanced queues, and so a Jackson network (see formula (13.14)). We find that

$$\gamma_J - \gamma \propto \left(\sum_k \mu_k^{-1/2}\right)^2 - \left(\sum_k \mu_k^{-1}\right) + \frac{\lambda}{2}\left[\left(\sum_k \mu_k\right)^2 - \left(\sum_k \mu_k^2\right)\right], \tag{14.30}$$

the constant of proportionality being $\lambda c/(1 - \lambda T_m)$. Expression (14.30) is plainly positive, and can in fact be zero only if $m = 1$, when the line consists of a single stage.

The case of a single multi-class stage, considered in the last section, in fact works out much more laboriously. This is because the number of cases (i.e. sequences of favoured classes) that can arise is so large. We can again take 'state j' as being an item of class j that has still to be processed. Equations (14.23) for Q yield the relations

$$Q_{ji} = \mu_j^{-1}(c_i + \sum_k \lambda_k Q_{ki}) = S_i/\mu_j \qquad (i \ge j),$$

say. Inserting this evaluation into the definition of S_i, we obtain the recursion

$$S_i = G_i^{-1}(c_i + \mu_i^{-1} \sum_{j>i} \lambda_j S_j), \tag{14.31}$$

where

$$G_i = 1 - \sum_{j \le i} \rho_j.$$

We find from (14.25) that $R_j = Q_{jj}/2$, so that the formula for the minimal average cost is

$$\gamma = \sum_j \rho_j S_j.$$

Substituting the solution S of (14.31) into this last expression we find at length that

$$\gamma = \sum \frac{\rho_{j_1} \lambda_{j_2} \ldots \lambda_{j_s} c_s}{(\mu_{j_1} \mu_{j_2} \ldots \mu_{j_{s-1}})(G_{j_1} G_{j_2} \ldots G_{j_s})}, \tag{14.32}$$

where the sum is over all strictly increasing sequences $\{j_1, j_2, \ldots, j_s\}$ that can be drawn from the set $\{1, 2, \ldots, m\}$.

Expression (14.32) is quite pleasing, but evidently the combinatorial possibilities are building up. They would do so even more if one combined the two cases we have examined, by considering several parallel processing lines.

14.5 Control with fixed work-station resources

A striking feature of the optimal policy, assuming freely deployable effort, is its seemingly slight (although crucial) dependence upon network state. Item states were arranged in a fixed order of priority, and the only way in which network state affected policy was that all processing effort should be concentrated on one of those items in the system whose index currently happens to be highest.

However, the situation assumed there is rarely a realistic one; resources are seldom freely deployable in real time. Each work-station will usually have its own fixed plant and dedicated staff, and can redeploy its effort only within limits. Moreover, the model we have taken implies a complete centralisation of both information and decision which would work only in an ideal world, and is exposed by experience as a recipe for unresponsive and inefficient management.

Add to this that, although the rules derived were undoubtedly optimal under the assumptions made, Section 14.4 demonstrated the beginnings of a combinatorial explosion. That is, that the number of contingencies envisaged grows so rapidly with continuing augmentation of the model that the calculations required become impractical. One needs a view of the model that enables each agent (i.e. decision-taker) to identify the significant factors, local or global, and to act upon these. There is a large literature which attempts just this: the summary in Section 13.5 lists just some of the more relevant items.

We are in danger of being diverted from our theme (the optimisation of network design) by the need to consider the preliminary problem: the achievement of efficient

network operation. 'Operation' in this case means the formulation of sequencing rules (which job to tackle next?) and routing rules (to whom shall it be assigned?). We shall content ourselves with discussion of what must be the simplest nontrivial case, comparing conclusions derived from the literature with those suggested by a plausible adaptation of the index rules.

Of the complex of possible factors, some will emerge as significant only under extreme conditions. The realistic version of an extreme condition is that of heavy traffic, when there is only a slight working margin of capacity. More specifically, both demand and capacity are large, but capacity exceeds demand by only a small relative amount. This case is often also that in which it is legitimate to replace discrete-state Markov dynamics by Brownian dynamics (i.e. deterministic differential equations driven by white noise). Under these extreme conditions several authors have noted the phenomenon of *state space collapse*: that the equilibrium distribution of an n-dimensional state variable is confined to a manifold of dimension lower than n. (See e.g. Kelly and Williams, 2004.) This seems to be associated with an effective pooling of service resources.

The example we shall discuss has been analysed by Harrison and Wein (1990) and Kelly and Laws (1993). It illustrates the elementary level at which difficulties are encountered, although one must also admit that both sets of authors draw general conclusions from it. The model is illustrated in Fig. 14.1. Items of types A and B enter station 1, which can process either type. Items of type A then move on to a further stage of processing at station 2, whereas for items of type B processing is complete when they leave station 1. The question is then: which type of item should one choose to process at station 1, for a given state of the system?

We have the three item states: item of type A waiting at station 1, item of type A waiting at station 2, item of type B waiting at station 1. Let us label these as states 1, 2 and 3. Then equations (14.7) become

$$\begin{aligned}
&\max\,[\mu_1(\Delta_2-\Delta_1)-\alpha,-\psi_1]=0,\\
&\max\,[\mu_2(-\Delta_2)-\alpha,-\psi_2]=0,\\
&\max\,[\mu_3(-\Delta_3)-\alpha,-\psi_3]=0,
\end{aligned}$$

where the μ_j are the appropriate transition rates under unit effort. Harrison and Kelly both assume that items of all states are weighted equally, so we can set $c_j=1$. The indices

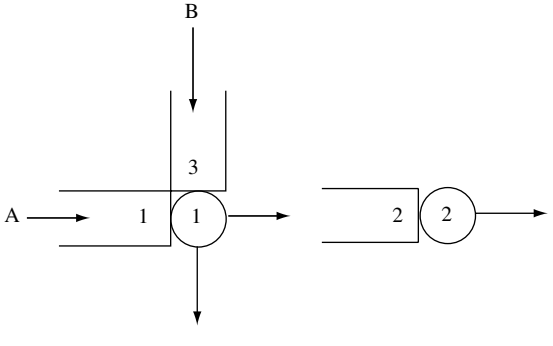


Fig. 14.1 The queueing example discussed in the text.

α_j deduced from these relations would then be $\alpha_1 = \mu_1\mu_2/(\mu_1+\mu_2)$, $\alpha_2 = \mu_2$, $\alpha_3 = \mu_3$. Harrison and Kelly assume that μ_1 and μ_3 are both twice μ_2, to allow for the fact that station 1 carries twice the traffic that station 2 does. The three index values are then respectively proportional to 2/3, 1 and 2, so that the states in descending order of priority are 3, 2 and 1. That is, one should first seek to clear items of type B from station 1, then items of type A from station 2, and only then items of type A from station 1.

However, it is now assumed that there can be no shifting of resource between stations, so that station 2 continues to process items at rate μ_2 as long as they are there, and the only latitude one has is the choice of item type to process at station 1. Harrison and Wein showed, in a heavy-traffic analysis, that the optimal policy was always to give priority to items of type B at station 1, except when the number of items at station 2 fell below a threshold, when items of type A should gain priority. In this way one cleared items as quickly as possible from the system, consistently with keeping station 2 busy.

It is possible to adapt the index approach to this case. Suppose that station h can process items whose state j falls into a set $\mathcal{P}_h$, where j now indicates both the type and the stage of processing of an item. Suppose also that any given j can lie only in one such class. That is, that at a given stage of processing there is only one choice of station for the next stage. Then a calculation given in Whittle (2005) and reproduced in Appendix 2 suggests that a good policy would be for station h to next process that item j that is present ($n_j > 0$), lies in $\mathcal{P}_h$ and minimises the index

$$I_j = \sum_{i<j} \omega_{ij} n_i. \tag{14.33}$$

Here ω_{ij} is the expression on the left in equation (14.23). We know this to be zero for $i \geq j$. For $i < j$ it is positive and can be seen as an interference measure: a measure of the degree to which the choice of j in $\mathcal{P}_h$ affects the progress of those items whose state i gives them higher priority. Criterion (14.33) does reflect congestion, in that it is concerned with actual numbers rather than just the presence or absence of an item of designated index. Its linear form makes implementation relatively easy. In equations (A2.19) and (A2.20) of Appendix 2 we give integral expressions for the coefficients ω_{ij} which offer the quickest evaluation.

Applying this to the two-station case, we find that we should choose a type A or a type B item for processing at station 1 according as $\omega_{12}n_2 + \omega_{13}n_3$ or zero is smaller and, of course, an item of the relevant type is available. The recommendation is then that one always gives a type B item priority over a type A item at station 1. This is a cruder form of the Harrison–Wein policy, in which the threshold is set equal to zero, so that no particular effort is made to keep station 2 busy.

IV

Communication networks

The topic of communication networks is now one of immense range and sophistication, concerned with the management of many competing forms of information traffic between many kinds of agents and devices. In this part we consider only some of the simplest representative cases that have implications for the design question. As ever, the problem is that design cannot be tackled before efficient operating policies have been agreed, and this first stage poses questions enough.

Loss networks are characterised by the fact that a call is accepted only if a clear path through the net is immediately available for it. As observed in Section 13.5, this makes them special in the class of queueing networks, in that they are by fiat both stable and queueless. Their operating rules then amount merely to the formulation of policies for call admission and routing. Under rather stark assumptions this reduces to solution of a simple linear programme. The analysis of Section 5.3 is invoked to show that the features of seminvariant costs, variable load and environmental penalty can all be incorporated. The hierarchical arrangement of exchanges suggested in Chapter 5 certainly gives the easiest analysis of reliability, and perhaps the best protection.

The stochastic upgrading in Chapter 16 of the basic version owes much to F. P. Kelly, although this is work he has now long left behind him. Kelly perceived the stochastic analogues of the shadow prices generated by the linear programme analysis, and their implications for routing and for updating of a network. A deeper contribution was his achievement of efficient self-regulation in the form of minimally interventionist and minimally centralised trunk reservation rules.

The phenomena of the Internet and the Worldwide Web, briefly treated in Chapters 17 and 18, have been remarkable for the revolution in thinking their development has demanded. The notion of a physical system to be optimised is now replaced by that of a game with a very large number of poorly informed players; a game whose operation is nevertheless required to be smooth and fair, in some sense. This much more diffuse system is now intrinsically evolutionary; it settles its own design in that its structure develops with demand. The statistical character of the structure thus generated is of extreme interest.

15

Loss networks: optimisation and robustness

A telephone network consists of a network of exchanges (or routers, in more modern formulations). Many of these are themselves the centre of a local star network, in that they have direct connections to individual subscribers in the region. However, it is the exchanges that we shall regard as constituting the nodes of a network. The defining feature of a loss network is that a call is accepted only if a clear route to its destination is available, otherwise the call is lost.

Random variation can enter the system in various ways. Even if all the parameters of operation and loading are constant and equilibrium has been reached, there will be statistical variation of the numbers and types of calls in progress. This is the source that receives most attention in at least that part of the literature favoured by mathematicians. One might consider it secondary in comparison with the more radical type of uncertainty which faces the system when there are massive variations in load or major internal failures. It is these that determine system structure, by setting a premium on versatility and robustness.

Nevertheless, there will always be times when a system is working close to capacity, and real-time decisions on acceptance and routing must be made to accommodate normal statistical variation. Such circumstances are associated with the decentralised realisation of policies, and will be considered in the next chapter.

We shall then begin by considering a system operating deterministically with a fixed demand pattern. The variables (e.g. numbers of calls in progress, capacity assigned to a given link) will be regarded as continuous in this treatment. We follow convention in following the simplest case: that in which flow costs are zero as long as one operates within capacity and infinite otherwise. This leads to a linear programming formulation. However, as indicated in (1.15), this zero/infinity cost function is a limiting version of a seminvariant prescription, and the whole analysis goes through for that more general case. Indeed, as demonstrated in Chapter 1, it is just for seminvariant flow costs that the naive design problem collapses to a linear programme. In Section 15.3 we indicate how the treatment of Chapter 5 can in principle be transferred to cope with the effects of variable demand and environmental penalty.

15.1 The linear programming formulation

Consider operation in a fixed regime. Suppose that calls are divided into classes, labelled by i. Calls in a given class have the same statistical parameters, the same degree of

priority and demand the same connection (i.e. between the same two given exchanges). We shall suppose that calls in class i arrive at rate ν_i and, if accepted, are completed at rate μ_i. If they were all accepted then the number of such calls in progress would be $\rho_i = \nu_i/\mu_i$. One can regard ρ_i as the rate at which work of class i arrives.

Calls can be routed in various ways through the system. A route r is a sequence of arcs (i.e. of links between exchanges) which then forms a path through the system. Denote the set of routes that realise the connection demanded by class i by $\mathcal{R}_i$. Denote the number of calls in progress that follow route r by x_r. Then necessarily

$$\sum_r R_{ir} x_r \leq \rho_i, \tag{15.1}$$

where R_{ir} equals 1 or 0 according as r does or does not belong to $\mathcal{R}_i$. We shall regard specification of the x_r as prescribing the policy, i.e. the rule by which the system operates. This specification addresses two issues: admission and routing. If strict inequality holds in (15.1) then this means that not all calls of class i are admitted – capacity limitations and priority considerations may require this.

This formulation of a network is simpler by a good margin than that with which we began in Chapter 1, and there are no surprises. The optimal operating policy is to route calls on the cheapest paths available and to accept only those calls that exceed a certain level of profit. The optimal design policy is to provide full capacity on the cheapest routes for those calls that are to be accepted, which are again those that exceed a certain level of profit. We shall nevertheless follow through the linear programming formulation, which generates the useful concept of shadow prices, which in turn persists into a fully stochastic model.

The capacity constraints fall on the links (arcs) in the conventional formulation. A given link j may contain b_j *circuits*, and a call on route r may demand a_{jr} of these circuits. (Usually a_{jr} will be either 0 or 1, but it is notationally simpler to formulate the more general case.) We then have the capacity constraints

$$\sum_r a_{jr} x_r \leq b_j. \tag{15.2}$$

The optimality criterion that is often taken is of the type: choose $x = \{x_r\}$ subject to $x \geq 0$ and conditions (15.1) and (15.2) to maximise the linear form

$$W = \sum_i \sum_r w_i R_{ir} x_r. \tag{15.3}$$

Here the weighting coefficient w_i represents the income per unit time derived from an i-call in progress. It is sometimes characterised as the reward that is given when an i-call is completed (or, equivalently, the loss incurred when such a call is refused). If this were the case then it should be replaced by $w_i \mu_i$ in (15.3).

The problem of optimising both admissions and routing has thus been reduced to a linear programme, which we can express: choose $x \geq 0$ to maximise

$$W = wRx \tag{15.4}$$

subject to the constraints

$$Rx \leq \rho, \tag{15.5}$$

$$Ax \leq b. \tag{15.6}$$

Here (15.5) and (15.6) are the vector versions of (15.1) and (15.2). Inequality (15.5) is an acceptance constraint, saying that the input rates set upper bounds on flow. Inequality (15.6) is a capacity constraint. Expression (15.4) is the matrix version of (15.3), with w the row vector of coefficients w_i.

If we maximise the Lagrangian form

$$L(x, \alpha, \beta) = wRx + \alpha(\rho - Rx) + \beta(b - Ax), \tag{15.7}$$

where α and β are the row vectors of Lagrangian multipliers associated with the vector constraints (15.5) and (15.6), then we find that

$$\alpha \geq 0, \qquad \beta \geq 0, \qquad wR - \alpha R - \beta A \leq 0, \tag{15.8}$$

together with the complementary slackness conditions

$$\alpha(\rho - Rx) = 0, \qquad \beta(b - Ax) = 0, \qquad (wR - \alpha R - \beta A)x = 0. \tag{15.9}$$

The dual problem is: choose α and β subject to (15.8) to minimise $L^* = \alpha\rho + \beta b$.

The Lagrange multipliers have their usual shadow price interpretation; α_i is the marginal increase in rate of return from an increase in the work arrival rate ρ_i, and β_j correspondingly for the link capacity b_j. Note an implication of the final conditions of (15.8) and (15.9): that

$$\alpha_i \geq w_i - \beta a_r \tag{15.10}$$

for all r of $\mathcal{R}_i$, with equality if x_r is nonzero. Here a_r is the rth column of A: that with jth element a_{jr}. Relation (15.10) and the fact that $\alpha_i \geq 0$, with $(Rx - \rho)_i = 0$ in the case of strict inequality, then imply

Theorem 15.1 *(i) The only routes r used for i-calls in the optimal solution are those that minimise the shadow cost βa_r of the circuits required by a call on this route, and α_i then equals w_i minus this minimal shadow cost. (ii) All i-calls or no i-calls are accepted according as w_i is strictly greater or strictly less than this minimal shadow cost.*

15.2 Design optimisation for a given configuration

If x specifies the admission/routing policy then the assignment of circuits b specifies the design. At least, this is the case if the node pattern, i.e. the existing set of exchanges, is regarded as fixed. We consider this more superficial optimisation in the present section, and a root-and-branch version in the next.

Suppose that additional circuits on link j can be bought at a unit cost of c_j, so that the leasing cost of design b is cb, where c is the row vector of prices c_j. Then the modified criterion

$$W = wRx - cb$$

is to be maximised with respect to both x and b. The maximisation with respect to b in the modified version of the Lagrangian form (15.7) yields $\beta_j \leq c_j$ with equality if link j is actually used in the optimal policy. Hence β is essentially identified with c, and Theorem 15.1 can be strengthened to:

Theorem 15.2 *Choose the routes r that minimise ca_r for r in $\mathcal{R}_i$, and denote this minimal value of ca_r by σ_i. Then these are the only routes that are used for i-calls in the optimal design, and input from class i is wholly accepted or wholly barred according as $w_i - \sigma_i$ is strictly positive or strictly negative. Decision in the transitional case is indifferent. That plant b is acquired that is needed to implement the policy thus determined.*

When we speak of cost-minimising routes we regard the choice as limited to those routes available with the prescribed nodes and links. If this prescription can be varied then, as in Chapter 1, the optimal net will reduce to something like straight-line source/destination links. Call variability will however force the adoption of an initial trunking (i.e. line concentrators), as in Chapter 4, and environmental considerations will force the adoption of higher levels of trunking, as in Chapter 5.

A slight variant of some interest is that in which the reward vector w and the cost vector c are not measured in the same currency. Suppose, for example, that an expenditure of S per unit time is ring-fenced for plant-leasing costs; this may be used in no other way, and no other source of finance is available. Then the capacity constraint (15.6) is replaced by a budget constraint

$$cAx \leq S. \tag{15.11}$$

The Lagrangian form (15.7) is then modified to

$$L(x, \alpha, \theta) = wRx + \alpha(\rho - Rx) + \theta(S - cAx),$$

where θ is the scalar multiplier associated with constraint (15.11).

Theorem 15.3 *Define σ_i as the minimal value of ca_r for $r \in \mathcal{R}_i$. List the categories i in order of nonincreasing w_i/σ_i. Then the optimal design/admission/routing rule for the fixed-budget version of the problem is as follows. Send whatever i-traffic is accepted on one of the cost-minimising routes just determined. Assign whatever circuit capacity is needed to meet the full demand of the early categories on the list until one reaches the point at which the cost constraint (15.11) becomes active. The last category admitted will then in general be only partially admitted.*

Proof By the same argument as for Theorem 15.1 any i-traffic accepted is sent on one of the cost-minimising routes asserted, and such traffic is definitely accepted or definitely rejected according as $w_i - \theta\sigma_i$ is strictly positive or strictly negative. But θ is

given a value such that the whole plant budget is expended, implying the admission rule asserted. ◊

The effect of fixing the budget rather than allowing open financing is then to replace the admission criterion $w_i - \sigma_i$ by w_i/σ_i.

15.3 A free optimisation

What one might term a free optimisation would be constrained only by restriction to a given physical domain, explicit costings of material, construction, operation and environmental spoliation and recognition of the need to cope with a variable load. The model of a telephone network formulated in Section 5.3 includes all these features, plus that of a seminvariant specification of flow costs. However, a loss network does bring one new feature: that there is a choice of which offered traffic to accept and so (at a cost) of which to refuse. This implies the explicit valuation of differing classes of traffic. One might add, however, that the example of a loss network does raise in acuter form a general problem we have scarcely mentioned: the implications for operation and design of the possibility of failure in some part of the system.

If we except these two features, then Section 5.3 sets out the formal general analysis, leading to the evolutionary equation (5.20) for the determination, in principle, of the optimal design. This incorporates the two features, variable load and environmental costing, which we know induce the trunking which will certainly be a feature of the network. Although equation (5.20) makes no presumptions, in that it initially allows a continuous distribution of material over the physical domain, a discrete structure will certainly emerge, and Section 4.8 gives an indication of how the evolutionary calculations could then become both much lighter and more immediately illuminating. The internal nodes that emerge will be identifiable as the exchanges of the optimal net (insofar as the model assumed is realistic). More than this, the effect of environmental costs is to produce something like a hierarchical system, consistently with both intuition and practice. The analysis of Sections 5.4–5.7 pursues this line explicitly, although admittedly for a wildly unrealistic case: that of uniform conditions over the whole of physical space.

Note the regrettable double use of the symbol ρ, as representing the work input rates $\rho = \nu/\mu$ in previous sections of this chapter and as representing the spatial density of material in references to previous chapters. Both usages are hallowed by tradition. However, we assume the first in this chapter; the second only in back-references to equations such as (5.20).

We harmonise with previous chapters if we replace the class label i by the triple label hiv, where h and i label the nodes of caller and called respectively and v labels the commercial class of the call. The coefficient w_i is then replaced by w_{hiv}, the income per unit time derived from an hiv-call in progress.

The fact that acceptance of traffic is now a matter of choice will be reflected in the fact that the effective input rate f_{hiv} of iv-traffic at node h can be chosen at any value between zero and ρ_{hiv}, the actual rate of demand. The optimal admission rule will be the analogue of that determined in previous sections: that one will accept all offered iv-traffic at node h or none according as w_{hiv} is greater or less than $y_{hiv}(\omega)$, with possible partial admission in the case of equality. Here $y_{hiv}(\omega)$ is the value determined for the multiplier y_{hiv} of the

Lagrangian form (5.13) when the system is in load regime ω. The admission rule will of course depend upon the regime, whose nature we assume either predictable or quickly observable.

Of course, the fact that an admission rule is in operation will affect the working of the evolutionary rule (5.20). To allow for this one must introduce an auxiliary evolutionary rule

$$\dot{f}_{hiv} = \kappa' \mathrm{sgn}(w_{hiv} - y_{hiv}), \tag{15.12}$$

where equation (15.12) is subject to impermeable boundary conditions when f_{hiv} reaches the extreme values zero or ρ_{hiv}. That is, f_{hiv} is not permitted to fall below the first or rise above the second. The value y_{hiv} of the multiplier is that linked with the determination of current optimal flow pattern $\hat{x}$. Relation (15.12) is essentially a learning rule for the concurrent determination of the values $y_{hiv}(\omega)$.

The analysis of the knock-on effects of failure in a general network, and of the best way to mitigate these, is an extraordinarily difficult exercise in statistical mechanics. However, for a hierarchical network the calculation is immediate. Suppose that a link at level t (i.e. one connecting an exchange at level t to one at level $t+1$) can fail with probability $1-P_t$. The link from subscriber to local exchange is counted as level zero. Then, if failures are assumed independent, the probability of an intact path between subscribers whose connection is completed at exchange level t is $(P_0 P_1 \ldots P_{t-1})^2$.

16

Loss networks: stochastics and self-regulation

The treatment of the last chapter ignored the statistical variation in input under a fixed demand pattern. However, one should check that this does not lead one to miss points of real significance, affecting both performance and operation. As far as operation goes, we simply assumed in Chapter 15 that the optimal quotas determined for admission and routing could be achieved somehow. 'How?' can only be settled by consideration of a model that is both dynamic and stochastic. A stochastic formulation forces one to develop control rules in terms of current state, which alone can extract best advantage from fortuitous variation while observing constraints.

The adaptability to stochastic state thus achieved can also provide a substantial degree of adaptability to changes in demand. However, as far as the optimisation of design is concerned, the effect of stochastic variation of state for fixed demand will be secondary in comparison with that of the major variations in demand considered in Section 15.3.

16.1 A single exchange; Erlang's formula

Before considering the stochastic version of the full net, it is useful to consider the case of a single exchange. Suppose that an exchange offers a single link to some other destination, consisting of G parallel circuits, any one of which can carry a call. Calls arrive in a Poisson stream of rate λ, and will be accepted if there is a circuit free. Once connected, the call will terminate with a probability intensity μ, independently of the state of the exchange. This state is then described by n, the number of busy circuits. This is indeed the state variable of a Markov process, in fact a birth-and-death process. Its equilibrium distribution $p(n)$ obeys the detailed balance relation

$$\lambda p(n-1) = \mu n p(n) \quad (0 < n \leq G),$$

with solution

$$p(n) \propto \frac{\rho^n}{n!} \quad (0 \leq n \leq G),$$

where $\rho = \lambda/\mu$ can be regarded as the rate at which work (i.e. time needed for processing) arrives at the exchange. If $n > G$ then $p(n)$ is zero, so n has a truncated Poisson

distribution. In particular, the probability that an incoming call will be lost is given by Erlang's formula

$$p(G) = \frac{\rho^G}{G!}\left[\sum_{n=0}^{G}\frac{\rho^n}{n!}\right]^{-1} = E(\rho, G). \tag{16.1}$$

We have taken this as defining the function $E(\rho, G)$, associated with the name of Erlang and appealed to repeatedly in this context. For given ρ the function decreases monotonically from unity to zero as G increases from zero to infinity.

More complex cases are often treated by appeal to this simple case, by arguments that are approximate but insightful; see Section 16.4. Expression (16.1) for the *blocking probability* has in fact a wider validity: it holds whatever the distribution of call lengths, provided these are independent with expectation $1/\mu$. This 'insensitivity' result is a consequence of the fact that $p(n)$ obeys a detailed balance condition. A general exposition of such matters is given in Whittle (1986).

We can make the familiar scaling assumptions: that λ and G are both proportional to a scale parameter V. The ratio $G\mu/\lambda$ of capacity to input has a scale-independent value θ. Then, if V is large and the exchange is saturated (so that $\theta < 1$) the statement that the number of busy circuits has a truncated Poisson distribution can be inverted to the statement that the number s of free circuits has the geometric distribution

$$\pi(s) = (1-\theta)\theta^s.$$

This is a view which will recur, with variations.

A first variant of the simple exchange that is worth considering is that in which there are several classes of calls, arriving in independent Poisson streams, and upon which differing priorities are set, as in Section 15.1. The optimal admission policy deduced there would accept all calls of priority above a certain level, refuse calls below that level, and partly admit calls in the transitional case. We have to consider how this policy might be modified, and certainly how it would be realised, when the stream of calls to be handled is random.

Suppose that calls of class i arrive at rate λ_i, demand a_i circuits and, if accepted, are completed at rate μ_i. The process is then the loss version of a queue with several classes of customers, except that there is no queueing. Admissions have to be decided as calls arrive, in the knowledge of what this might imply for the blocking of higher priority calls which may follow. Let us denote the vector with elements n_i by n and the vector with a unit in the ith place and zeros elsewhere by e_i. Then the equation determining the optimal admission rule is

$$\gamma = \sum_i \{\lambda_i \max[(w_i/\mu_i) + f(n+e_i) - f(n), 0] + \mu_i n_i [f(n-e_i) - f(n)]\}. \tag{16.2}$$

Here w_i/μ_i is the reward for processing an i-call, equivalent to giving reward at rate w_i during processing, γ is the maximal rate of reward that the exchange can earn, and $f(n)$ is the transient reward, the transitory reward attached to starting from a state value n. We assume that $f(n) = -\infty$ if the value n is forbidden, i.e. does not satisfy $\sum_i a_i n_i \le G$. The 'max' operator in (16.2) reflects the optimisation of the admission decision. We see

from (16.2) that the decision of whether or not to accept a given i-call in general depends upon the whole state n of the exchange. There is a simplification in one case.

Theorem 16.1 *Suppose that $a_i = 1$ and $\mu_i = \mu$ for all i. Then the optimal admission rule depends only upon s, the number of free circuits, and equation (16.2) reduces to*

$$\gamma = \sum_i \lambda_i \max[(w_i/\mu) + f(s-1) - f(s), 0] + \mu(G-s)[f(s+1) - f(s)] \qquad (16.3)$$

where $f(s) = -\infty$ for $s < 0$. The optimal policy is of the form: accept an incoming i-call if the number of free circuits s exceeds a critical value s_i, where s_i increases with decreasing w_i.

Proof Note that we have used the same notation f for a transient reward in equations (16.2) and (16.3), but that the functions are quite different in the two cases. All the points can be made verbally. The rate at which circuits are released is a function only of $\sum_i n_i = G - s$, and so optimal decisions will be based solely upon s. If calls of type i are accepted at a given value s' of s, then, *a fortiori*, calls of higher or equal priority will be accepted at values $s \geq s'$. Consequently, the admission rules have the form indicated. $\Diamond$

We thus see that, for the restricted case of Theorem 16.1 the optimal admission rule is a *trunk reservation rule*; calls of type i are accepted only if more than s_i circuits are free. One can see intuitively that the same will be true for the more general case of (16.2) in the case of large-scale full-capacity operation, because under any given nonidling policy n will settle to a relatively steady value whose variations will have only a secondary effect upon the future release of circuits.

This then gives basis to an idea developed by Kelly and his colleagues (see Gibbens, 1988; Gibbens *et al.*, 1988; Gibbens and Kelly, 1990; Kelly, 1990, 1991a, 1995; Kelly *et al.*, 1995): that a trunk reservation scheme provides the effective form of state-based control of a loss network. The remarkable feature of the trunk reservation rule is that, by reacting to the value s of the number of free circuits (a scalar variable normally O(1) in V), one is controlling the composition of the O(V) vector n.

16.2 Admission control for a single exchange

We know from Section 15.1 that the optimal admission rules are very simple in the fluid limit, even for the more general case of (16.2). In this limit i-calls are assigned a priority index w_i/a_i, and there is a sharp cut-off on the basis of this index. That is, if calls are labelled in order of decreasing index then there is a k such that all calls are accepted for $i < k$, all calls are refused for $i > k$, and k and the proportion of k-calls accepted is determined by the capacity limit G. The trunk reservation rule provides the state-based policy that realises this quota rule in the case of a random input, but in a nonrigid form, in that it is sensitive to the actual state of the exchange, and even adaptable to variation in plant and input parameters.

Suppose that the classes are ordered in decreasing priority, so that w_i/a_i does not increase with increasing i. Suppose one adopts the rule that an incoming i-call is accepted only if the current number s of free circuits exceeds a threshold s_i, the *trunk reservation parameter* for such calls. These thresholds are also supposed increasing in i, to reflect

priorities, and of course $s_i \geq a_i$. The choice of thresholds is otherwise not critical, but a reasonable choice would be $s_i \approx \kappa a_i/w_i$ for some κ, of which more anon. This is certainly simple, in that it depends upon system state only through s, but is also sensitive, in that it responds to either excessively large or excessively small s. The remarkable feature that emerges is that s has a distribution essentially independent of V, and the thresholds s_i can then be chosen similarly independently – one is controlling a large beast by twisting its very small tail.

For simplicity, consider first the case $a_i = 1$ for all i. The more general case then follows. The state of the system is described by the vector n, but we shall extend this to (n, s). This variable undergoes the transition to $(n - e_i, s + 1)$ at rate $\mu_i n_i$ and to $(n + e_i, s - 1)$ at rate $\lambda_i \mathrm{H}(s - s_i)$, where $\mathrm{H}(u)$ is the Heaviside function, taking the value zero for $u \leq 0$ and unity for $u > 0$. (We use a roman 'H' to distinguish the Heaviside function from the infinite-step function $H(u)$ defined in (1.15).)

We have of course the identity $\sum_i n_i + s = G$, but we assume full-capacity working (saturation) in that s is O(1) and so n is O(V). If we denote E(n) under a given policy by $\overline{n}$ then the assumption of saturation implies the relation

$$\sum_i \overline{n}_i \approx G.$$

The essential point is now that the rate of release of busy circuits will be determined largely by $\overline{n}$; the variation of n about this value will have only a secondary effect. If we average n-dynamics over stochastic variation of s then the transition $n_i \to n_i + 1$ has rate $\lambda_i P(s > s_i) = \lambda_i F_i$, say. Balance then requires that $\mu_i \overline{n}_i = \lambda_i F_i$ or

$$\overline{n}_i = \frac{\lambda_i F_i}{\mu_i}.$$

The last two equations then imply that

$$\sum_i \frac{\lambda_i F_i}{\mu_i} \approx G. \tag{16.4}$$

This is not a constraint, but rather a statement of the assumption that the exchange is saturated, and so s is O(1). The transition rate $s \to s - 1$ is

$$\lambda(s) = \sum_i \lambda_i \mathrm{H}(s - s_i),$$

while, under the smoothed dynamics, the transition rate $s \to s + 1$ is

$$\sum_i \mu_i \overline{n}_i = \sum_i \lambda_i F_i = F,$$

say. The consequent equilibrium distribution is thus

$$\pi(s) \propto \theta_1^{s_2 - s_1} \theta_2^{s_3 - s_2} \ldots \theta_{i-1}^{s_i - s_{i-1}} \theta_i^{s - s_i}, \quad \text{if} \quad s_i \leq s \leq s_{i+1}, \tag{16.5}$$

where $\theta_i = F/(\lambda_1 + \lambda_2 + \cdots + \lambda_i) = F/\sigma_i$, say. Thus $\pi(s)$ is increasing or decreasing according as the current θ is greater or less than unity. But as i increases then θ_i

decreases from a value greater than 1 to a value less than 1. The distribution $\pi(s)$ is thus unimodal, and will become ever more strongly so as the inter-threshold intervals $s_{i+1} - s_i$ are increased, in the sense that the distribution of the variable s/κ will become ever more concentrated about its mode as κ increases. The factor κ does not need to be large, however, before the distribution provides an effective cut-off, in that for some value k the tail probability F_i is virtually unity for $i < k$ and virtually zero for $i > k$. This critical value is not determined by the relation $\theta_k \approx 1$ or $\sum_i \lambda_i F_i \approx \sigma_k$, which we now see as an identity. It is determined by relation (16.4), which states, as usual, that the cut-off point is determined by requiring that all high-priority traffic be accepted up to the point at which all resources are committed.

Note that the trunk reservation system will automatically adapt to changing load λ and capacity G, in that the cut-off point k will adapt to that value which just slices off the amount of higher-priority traffic that can be accommodated. This is a robustness feature of the policy that makes optimisation of the reservation values s_i a secondary matter. One could achieve such an optimisation by giving a dynamic programming analogue of the argument we have just given. However, we already know the character of the priority index, and an optimisation for given input rates λ_i scarcely has much point when a feature of the policy is that it successfully adapts to varying values of these rates.

Consider now the case of i-dependent circuit demands a_i. We have then the transitions and effective transition rates: $s \to s - a_i$ at rate $\lambda_i \mathrm{H}(s - s_i)$, and $s \to s + a_i$ at rate $\mu_i \overline{n}_i = \lambda_i F_i$. Note that the saturation condition (16.4) now becomes

$$\sum_i \frac{a_i \lambda_i F_i}{\mu_i} \approx G.$$

The expected rate of change of s conditional on a current value s is now

$$R(s) = \sum_i a_i \lambda_i [F_i - \mathrm{H}(s - s_i)].$$

This is monotonic in s, decreasing from the positive value $\sum_i a_i \lambda_i F_i$ to the negative value $\sum_i a_i \lambda_i (F_i - 1)$ as s increases from zero, which is enough to demonstrate that $\pi(s)$ is unimodal. The point where the mode occurs is that where R changes sign, and constitutes the effective cut-off point at which admissions are guillotined.

16.3 Equilibrium and asymptotics for the network

We consider now a general network of exchanges, and begin with a very classic piece of work. This is the determination of the most probable state of the exchange under prescribed state-independent routing rules. The calculation yields interesting insights, but does not of itself suggest value-based admission or routing rules. Effective such rules, again developed by Gibbens and Kelly, are derived in Section 16.5.

We shall again use i to denote distinguishable input classes, but shall suppose that a route has been prescribed for each, so we know that an i-call will demand a_{ji} of the G_j circuits available on link j. A call is accepted if and only if this route is clear for it, otherwise the call is lost. It is supposed that i-calls arrive in a Poisson stream of rate λ_i and, if accepted, are completed at rate (probability intensity) μ_i. Let n_i be the number

of i-calls in progress at a given time, so that $n = \{n_i\}$ is the state of the net. Then n has remarkably simple equilibrium distribution (see Brockmeyer *et al.*, 1948; Girard and Ouimet, 1983; Burman *et al.*, 1984). Let $\mathcal{N}$ be the set of n-values that are physically possible: i.e. that are compatible with the capacity constraints upon the net. Then

$$\pi(n) \propto \prod_i \frac{\rho_i^{n_i}}{n_i!} \quad (n \in \mathcal{N}), \tag{16.6}$$

and is zero outside $\mathcal{N}$. Here the proportionality factor normalises the distribution and $\rho_i = \lambda_i/\mu_i$. Relation (16.6) follows from the fact that there is detailed balance between the only possible state transitions: $n \rightleftharpoons n + e_i$ for varying i.

The set $\mathcal{N}$ of feasible n-values is determined by the analogue of the capacity constraint (15.2),

$$An \leq G, \tag{16.7}$$

plus of course that the occupation numbers n_i are nonnegative integers.

The occupation numbers are then independent Poisson variables constrained by (16.7). One could determine the effect of the constraint by considering its effect on either the most probable value of the random variable n or its expectation. The calculation is almost identical with the calculation of the Gibbs distribution in elementary statistical mechanics: the determination of the abundances $\overline{n_i}$ of molecules of types i for prescribed atomic abundances G_j of the elements j from which the molecules can be constituted. Just as for that problem, it is helpful to consider the case of a large-scale system. Suppose that V is a large number representing the sizes of the communities that the exchanges serve. We would then expect the variables λ, G and n all to be order V, and the normal course would be to work in terms of standardised variables λ/V, G/V and $x = n/V$. Indeed, we would expect the random variable n/V to converge to a deterministic limit in almost any stochastic sense as $V \to \infty$, and shall in fact use x to denote this stochastic limit. However, there is merit in working in terms of the original variables, while keeping in mind that they are all of order V.

We shall first follow the classical direct route of approximating the factorials in (16.6) by Stirling's formula, neglecting the discrete nature of n, and then determining the most probable value $\overline{n}$ of n as the value maximising the consequent expression subject only to constraints (16.7). In doing so we slur over a number of delicate issues, but the more refined treatment of Section 16.4 takes some account of these. Note that characterisation in terms of a most probable value provides an example of how a stochastic problem can generate its own extremal principle.

Theorem 16.2 *(i) The most probable value of n in the limit of large V differs by an O(1) term from the value $\overline{n}$ determined by*

$$\overline{n}_i = \rho_i \mathrm{e}^{-ya_i} = \rho_i \prod_j p_j^{a_{ji}}. \tag{16.8}$$

Here a_i is the ith column of the matrix A and $p_j = exp(-y_j)$.

(ii) Here the row vector y can be characterised as the solution of the dual problem: as the nonnegative vector minimising the expression

$$L(y) = \sum_i \rho_i e^{-ya_i} + yG. \tag{16.9}$$

(iii) The solution $\overline{n}$ satisfies the constraints

$$\sum_i a_{ji}\overline{n}_i \le G_j, \tag{16.10}$$

with y_j zero in the case of strict inequality, i.e. when the circuits of link j are not fully utilised.

Proof We are maximising the concave function $\ln \pi(n)$ subject to the linear inequalities (16.7), and so the methods of convex programming are applicable. The value of n maximising the Lagrangian form

$$L(n, y) = \sum_i n_i[\ln(\rho_i/n_i) + 1] + y(G - An)$$

is just that given by (16.8). The Lagrange multipliers y_j will be nonnegative (because of the inequality nature of the constraints) and will solve the dual problem: i.e. will minimise the n-maximised value of the Lagrangian form, evaluated as expression (16.9). Remaining assertions follow from the complementary slackness condition: that y_j will be zero if the corresponding constraint is not active. ◊

The final expression in (16.8) follows from the one before it with p_j identified as $\exp(-y_j)$. Thus $0 \le p_j \le 1$, and $p_j = 1$ if constraint (16.10) is nonactive at j, i.e. if link j has more circuits than it would need. Expression (16.8) then seems to imply a remarkable interpretation: that p_j is the probability that a circuit is available on link j, and that these events are independent for different links, or even for repeated traverses of the same link.

These assertions appeared in Kelly (1986) and Whittle (1986). The latter reference also develops the probabilistic implications of the dual problem, which we shall examine in the next section.

The quantity

$$U_j = \sum_i a_{ji}\overline{n}_i = \sum_i a_{ji}\rho_i \prod_j p_j^{a_{ji}}$$

is the *utilisation* of link j: the number of circuits on this link actually in use in the most probable configuration. We thus have $p_j = 1$ on the unsaturated links, while the other p-values are determined by the equations

$$U_j = G_j \tag{16.11}$$

on the saturated links.

The dual problem could be stated as the unconstrained minimisation (with respect to y) of the form

$$Q(y) = \sum_i \rho_i e^{-ya_i} + \sum_j C(G_j, y_j), \tag{16.12}$$

where

$$C(G_j, y_j) = G_j(y_j)_+ = G_j \max(y_j, 0). \tag{16.13}$$

Alternatively yet again, one is choosing p to minimise

$$\phi(p) - \sum_j G_j \ln p_j$$

subject to $p_j \leq 1$ (all j), where

$$\phi(p) = \sum_i \rho_i \prod_j p_j^{a_{ji}}. \tag{16.14}$$

16.4 Refinements of the asymptotics

Consider now the alternative calculation, the evaluation of $\mathrm{E}(n|\mathcal{C})$, where the expectation is understood to be over distribution (16.6), and $\mathcal{C}$ represents the event that constraints (16.7) are satisfied. As shown in Whittle (1988), this leads to a refined evaluation of blocking.

Under the unconstrained version of distribution (16.6) the probability generating function (p.g.f.) of the n_i would be

$$\mathrm{E}\left[\prod_i z_i^{n_i}\right] \propto \exp \sum_i \rho_i z_i.$$

If with each call of type i are associated a_{ji} j-links then the joint p.g.f. of the n_v and the m_j, the numbers of j-links in operation, would be

$$\mathrm{E}\left[\prod_i \prod_j z_i^{n_i} p_j^{m_j}\right] \propto \exp\left[\sum_i \rho_i z_i \prod_j p_j^{a_{ji}}\right].$$

It follows then that

$$\mathrm{E}[n_i|m = G] \propto \rho_i \int \left[e^{\phi(p)} \prod_j p_j^{a_{ji}}\right] \left[\prod_j p_j^{-G_j - 1} \mathrm{d}p_j\right].$$

Here $\phi(p)$ is the expression defined in (16.14), and the integral evaluates the coefficient of $\prod_j p_j^{G_j}$ in the expansion of the first square bracket under the integral in nonnegative powers of the p_j. It does this by following the integration paths $|p_j| = \delta$ in the complex p_j-plane for all j and for small positive δ.

However, the conditioning event is not $m = G$, but $m \leq G$. By the same argument, we obtain the evaluation

$$\mathrm{E}[n_i|\mathcal{C}] = \frac{\rho_i I\left[\prod_j p_j^{a_{ji}}\right]}{I[1]}, \tag{16.15}$$

where the operator I has the effect

$$I[\chi(p)] = \int \chi(p)\Psi(p)\left(\prod_j \frac{\mathrm{d}p_j}{p_j}\right)$$

and

$$\Psi(p) = [\exp\phi(p)]\left[\prod_j \sum_{m=0}^{G_j} p_j^{-m}\right] = [\exp\phi(p)]\left[\prod_j \frac{1-p_j^{-G_j-1}}{1-p_j^{-1}}\right].$$

Expression (16.15) is exact. Its evaluation may seem formidable, but is in fact made for saddle-point evaluation, since $\Psi(p)$ is the sum of terms raised to the power of V. For large V expression (16.15) will yield the evaluation (16.8) for $\mathrm{E}[n_v|\mathcal{C}]$, with p determined as the saddle-point value, the value minimising $\Psi(p)$. (Recall that at this saddle point a minimum for real variation of p will be a maximum for variations along the integration path.) This is equivalent to choosing the value of y in (16.8) to be the value minimising the form (16.12), but with the evaluation

$$C(G_j, y_j) = \log\left(\sum_{m=0}^{G_j} \mathrm{e}^{my_j}\right). \tag{16.16}$$

For large V expression (16.16) does indeed reduce to expression (16.13), plus terms of lower order than V. There is no need to make this reduction, however, we can regard the function (16.12) with the definition (16.16) of C as a criterion function that makes greater concession to finiteness of scale than does (16.13). We shall term the problems of choosing y (or p) to minimise criterion (16.12) under the alternative specifications (16.13) and (16.16) of C as the 'hard' and 'soft' dual problems, respectively. The soft version has the hard version as limit, but one may expect it to be superior for moderate V.

Minimising $\Psi(p)$ with respect to p we obtain, for those links j that are saturated in that $p_j < 1$,

$$U_j = \frac{\partial C(G_j, y_j)}{\partial y_j} = G_j - \frac{\sum_{s=0}^{G_j} s p_j^s}{\sum_{s=0}^{G_j} p_j^s}. \tag{16.17}$$

These are the determining equations for the p_j on such links, the analogue of equation (16.11). If $p_j < 1$ then (16.17) reduces to

$$U_j = G_j - \frac{p_j}{1-p_j} \tag{16.18}$$

to within terms of order $p_j^{G_j}$. Relation (16.17) has an attractive statistical interpretation: that determination of p from the soft dual implies the conclusion that the number of free circuits on link j follows a truncated geometric distribution on $[0, G_j]$ with parameter p_j, if this parameter is less than unity.

There is an alternative method of determining the passing probabilities p_j that is well established in the telephony literature and has been particularly developed by Kelly. This is to use the relations

$$1 - p_j = E(U_j/p_j, G_j) \tag{16.19}$$

on the saturated links, where E is the Erlang function defined in (16.1). The argument is advanced for the case when a_j is either 0 or 1, so that U_j can be regarded as the number of busy j-circuits. The expression on the right would indeed be the blocking probability for link j if the link were subject to a Poisson input of rate U_j/p_j, each input call demanding a single link. The assumption that the input can be thus regarded is the *reduced load approximation*: if U_j is the number of busy circuits, when a request for a circuit is satisfied with probability p_j, then U_j/p_j must be the rate at which circuits are being demanded.

Kelly (1991b) regards the determination of the passing probabilities by relations (16.11), (16.18) and (16.19) as providing successively increasing degrees of refinement. It is certainly true that relation (16.19) implies (16.18) to within a term o(1) in V, so that the Erlang determination is at least as accurate; also that the Erlang approximation is felt to express some essential feature of telephone traffic. Kelly has demonstrated that relation (16.19) can also be deduced by minimisation of a form of type (16.12), although with a choice of the function $C(G_j, y_j)$ that is not readily interpretable.

16.5 Self-regulation for the network

Sections 16.3 and 16.4 determined the most probable state of the net if a route were prescribed for every class of call. However, the real aim is to develop n-dependent admission and routing rules that will maximise a value criterion $\mathrm{E}\sum_i w_i n_i$; a totally different extremal problem. For a fixed demand pattern the analysis of Chapter 15 recommended fixed admission quotas and fixed routing. In a statistical formulation there will be room for flexibility in both, a flexibility that should be exploited.

The aim is to find admission and routing rules that are sensitive to network state, but at the same time simple and practicable. These will not lay down quotas – for a self-regulatory system quotas appear as a consequence rather than either a means or an end. Rules should also not be over-centralised; decisions should ideally be made locally, using only information that is locally available. As indicated in Section 13.5, many such schemes have been proposed, but the trunk reservation rules developed by Gibbens and Kelly do seem to have succeeded best in combining these happy qualities for the case of a loss network.

These authors consider that in practice alternative routing should not go beyond an attempt at two-link connections in the case when the direct single-link connection is blocked. The successful Gibbens–Kelly scheme ('dynamic alternative routing') attempts first to make a direct connection; if this fails then a single attempt is made at a two-link connection, with random choice of the intermediate node. If this fails then the call is jettisoned. A trunk reservation is placed upon the two-link connection, so that the frequency of these is reduced to the point that there is not excessive interference with direct calls.

We shall analyse this procedure for the case of several call classes, assuming for simplicity that the network consists of a complete graph, symmetric in that all links have the same capacity G, costs are the same on all links and call rates the same for all connections. The classes i then just differ only in the values w_i attached to their connection, the rates λ_i at which they seek passage through a given link, and the rate μ_i

at which they are completed. The argument is an extension of that of Section 16.2, whose notation we shall carry over. We shall now use n_i and s to denote the number of i-calls in progress and the number of free circuits at a given link. The (n, s) statistics will of course be invariant under a permutation of nodes. Furthermore, if the network has many nodes, then, by appeal to the mean field arguments implicitly invoked in Section 13.6, the values of s on a few randomly selected links are effectively independent. We shall suppose that i-calls are subject to a trunk reservation s_i on single-link connections and u_i on each of the links of a two-link connection. If one adopted the general rule $s_i = \kappa a_i/w_i$ on a direct link, with circuit demands and costs as indicated, then u_i should be at least twice this, because each link is earning only half the reward w_i.

We postulate only the trunk reservation policy; all assertions now to be made follow from it. For simplicity we suppose all the a_{ij} either 0 or 1. The rate of the transition $(n, s) \to (n - e_i, s + 1)$ is $\mu_i n_i$, and rate of the transition $(n, s) \to (n + e_i, s - 1)$ is

$$\lambda_i(s) = \lambda_i\{\mathrm{H}(s - s_i) + 2[1 - \mathrm{H}(s' - s_i)]\mathrm{H}(s - u_i)\mathrm{H}(s'' - u_i)\}.$$

If links 1, 2 and 3 are the links with s, s' and s'' free circuits, then the first term in the curly brackets corresponds to a demand for direct passage through link 1, and the second to a demand for alternative routing through 1 and 3, after demand for direct passage through 2 has failed. This second term is multiplied by 2 because link 1 may be demanded as either the first or the second step in an alternative routing. Under smoothed dynamics the transition $n_i \to n_i + 1$ would then have intensity

$$\mathrm{E}[\lambda_i(s)] = \lambda_i[F_i + 2(1 - F_i)K_i^2] = \lambda_i J_i,$$

say. Here the expectation is over the s variables, assuming independence, so that $F_i = P(s > s_i)$ and $K_i = P(s > u_i)$. Balance implies that $\mu_i n_i = \lambda_i J_i$, so that the capacity condition (automatically fulfilled as long as the constraint is active) is

$$\sum_i \frac{\lambda_i[F_i + 2(1 - F_i)K_i^2]}{\mu_i} \approx G, \tag{16.20}$$

and the rate of the transition $s \to s + 1$ is $\sum_i \mu_i n_i = \sum_i \lambda_i J_i = J$, say. Under smoothed dynamics the rate of the transition $s \to s - 1$ is

$$\lambda(s) = \sum_i \lambda_i\{\mathrm{H}(s - s_i) + 2K_i(1 - F_i)\mathrm{H}(s - u_i)]\},$$

and s has the equilibrium distribution

$$\pi(s) \propto \prod_{q \le s} \frac{J}{\lambda(q)}. \tag{16.21}$$

The argument now goes as in Section 16.2. The rate $\lambda(s)$ is increasing in s, with the implication that the s-distribution (16.21) is unimodal. If the thresholds are scaled by a factor κ then the distribution of s/κ becomes ever more concentrated around this mode as κ increases. This scale factor does not need to be large, however, before an effective

cut-off point emerges, realising the quota policy of the last chapter. At this there is a class boundary k, fixed automatically by condition (16.20), which is an effective cut-off point, in that F_i is near unity for $i < k$ and near zero for $i > k$. All calls in the first group are accepted for direct transmission, no calls in the second group are, and k-calls are accepted with a probability determined by the random variation of s in the neighbourhood of s_k. Calls of class i are of course permitted to attempt indirect transmission if $s > u_i$, and this will be the case with near certainty if $u_i < s_k$.

However, it should again be emphasised that all these acceptance decisions are automatic, following solely from the threshold rules for s. The system not merely realises the optimal quota rule; it takes advantage of statistical variation and also adapts itself to sustained changes in plant parameters G and input parameters λ.

Gibbens (1988) found a case for alternative routing even in the fluid limit, if demand was variable. He carried out a linear-programming optimisation of flow at progressively increased levels of demand. It turned out that at low demand all routing was direct. There was then an interval in which some alternative routing was used, followed by a resumption of direct routing. This is reasonable. At low demand there is no need for alternative routing. Then there is an intermediate band where some rerouting is allowed, the increased link load being acceptable, to become unacceptable as input is increased further. It is in this intermediate band that there is a role for alternative routing, and this is the opportunity exploited by the trunk reservation scheme.

The models chosen both in Section 16.2 and in this section were special in that congestion became manifest at the very exchange to which the call first sought admission. If congestion manifests itself later on the route then one has a statistically more complicated situation; trunk reservation is being applied prospectively rather than at the point of entry. One might say that a desirable design property would be that absence of congestion at the entry port would be a guarantee of trouble-free passage all the way. This is an unrealistic expectation, however, and the Internet, in particular, has been designed with no such illusion. We give an elementary treatment of this development in Chapter 17.

17

Operation of the Internet

These last two chapters are intended as something of a corrective: to give significant examples of newer types, concepts and models of communication networks. This chapter continues the preoccupation of Chapter 16 with operation rather than design: to find rules that make effective use of poor and local information to achieve smooth operation of a rather diffuse system. The design aspect makes a natural return in Chapter 18; both the Internet and the Worldwide Web achieve their own design, partially at least, by a kind of directed evolution.

For loss networks a call was not accepted unless an open path from source to destination could be guaranteed. Once the connection was established, delivery of the message could begin. The Internet operates by packet-switching rather than circuit-switching. The 'message' is seen as a prepared file rather than a real-time conversation. This file is broken into 'packets' which are launched successively into the system, with no guarantee of an open path along the required route. The role of an 'exchange' is replaced by that of a 'router': a device that sorts packets as they arrive on incoming links and directs them on to the outgoing link appropriate for their destination. Each of these outgoing links is prefaced by a 'buffer', a queue that will store packets until there is free capacity on the link. The buffer is finite, however; once its capacity has been exceeded the overflow of packets is lost.

Final loss is not permitted, however, because the packets must be assembled at the destination to reconstitute the original file. If any packet is missing then transmission of the file is unacceptably impaired. A check is achieved by 'acknowledgement': the receipt of a packet at the destination is reported back to the sender. The sender thus learns which of his packets have been received, and so which have gone missing and must be retransmitted. This is all that the sender is presumed to know of network state, and all that he has to guide him in deciding the rate at which to attempt continued transmission. He knows only whether or not his own desired route is congested at some point; in particular, he may not know where such congestion lies. Remarkably, a simple and effective transmission policy can be achieved on the basis of this very partial information.

Note that even a lost packet represents a load on the system, because it has utilised network resources up to the point of loss. One sees then the dilemma of the system. A sender wishes to transmit his data in any case, but now has the added incentive that it is only by a trial transmission that he can learn something of changing network state. On the other hand, transmission into a congested net can only make matters worse for everybody.

'Everybody' refers to the intrinsic complication of a multitude of senders, whose individual transmission policies have to be determined in this knowledge. One effectively has a multi-person game with unshared information, and the only way to avoid unproductive anarchy is to bring a co-operative element into a situation that is otherwise competitive. This implies the imposition from above of a transmission policy or 'protocol'. The operating rules implied by these protocols are in fact realised by computers at the sending and receiving stations. For general expositions that go into some detail, see Kelly (2000, 2001).

17.1 Some background

The necessity for careful admission rules with an acknowledgement-based system became apparent for a precursor of the Internet, the Ethernet. The procedure then followed was that packets were transmitted in discrete time slots. If more than one packet was placed in a given slot then all the packets involved in the resultant 'collision' were lost. These losses were reported back to the relevant senders. Two influential papers (Kelly and MacPhee, 1987; Aldous, 1987) demonstrated the necessity for at least an 'exponential back-off' after an adverse report, if irreversible congestion was to be avoided. This found expression in the 'transmission control protocol' (TCP) since adopted, and the algorithm suggested by Jacobson (1988) that implements it. This is expressed in terms of the 'windows' of the senders. The window is the set of packets that have been transmitted but whose fate is yet unknown. If W is the window size then the TCP algorithm prescribes roughly that the window should be increased by a term of order $1/W$ after report of a successful transmission and halved after an unsuccessful transmission.

This simple rule has been outstandingly successful. However, technology has advanced continuously. Transmission is no longer synchronised to time slots. As indicated in the introduction above, packet traffic is now directed through routers, with buffers placed between router output ports and the corresponding link input. Congestion then manifests itself, not as a collision, but as an overspill of one of the buffer stores, the overspill being effectively returned to the relevant senders for retransmission. More recent technology in fact goes even further: when a packet arrives at its destination it returns an indication to the sender of the degree of congestion it experienced *en route*. Incorporation of this early warning in the admission rules provides a substantial degree of protection against serious congestion.

Theory and practice have endeavoured to play catch-up with each other during this period in a fashion more vigorous than ordered.

We should now become more specific. Possible source/destination connections are indexed by r; we shall assume for simplicity that this also specifies the route through the system. A 'sender' is associated with each value of r. We shall denote the rate at which sender r transmits work by u_r, and the size of his backlog of untransmitted work by x_r. The size of the buffer queue at the input to link j will be denoted by z_j. The variables u_r are then the control variables, to be chosen as functions of corresponding r-observables (i.e. the information currently available to sender r) in such a way as to achieve efficient running of the system. The control policy is thus expressed in terms of the transmission rates u_r rather than in terms of windows, although the two views can be related.

At this point one might contrast the aspirations of an optimal policy and of a satisficing naive policy. A full-dress optimisation demands specification of a full dynamic–stochastic model, of the information available to the various agents and of a performance criterion. After that one has the problem of 'solution': the deduction of optimal action rules. For a system as complex and diffuse as the Internet this problem is unlikely to be soluble, and a mathematical solution may well not be practical.

For a naive policy, on the other hand, one has little in the way of model specification, and then seeks a policy that is robust, and so effectively self-adapting to a variety of models. It must also be simple in that it is at least readily implementable. This may imply, as in the Internet case, a degree of decentralisation, in that local decisions are made on local information, even if the strategy is determined centrally. Models and criteria still have their place, however. One may perform a quasi-optimisation in order to suggest plausible rules, and one may test performance of a proposed rule on a plausible model.

The Internet presents fearsome difficulties for an exact analysis. The dynamics of a network of queues are in any case complicated, and much more so if the queues contain a mix of traffic, when there may be questions of priority. As we have already noted in Section 13.5, there are cases for which the local application of priority rules can lead to checkmate – a frozen system. However, there are aspects that can imply a radical simplification. Firstly, in modern systems the buffer queues are cleared so rapidly that queue discipline is almost irrelevant, and the admission rules settle priorities in a way that we shall soon describe. Secondly, there are many aspects of the problem that can be settled by considering a deterministic version of the model, the so-called fluid limit. Notably, Dai (1995) proves that convergence with time of the variables of the fluid model to unique fixed equilibrium values implies convergence of the stochastic model to a unique stochastic equilibrium (under subsidiary conditions, to be checked). Thirdly, if one is interested in variations of traffic on a slower time-scale than that of individual packet flow, then the detailed dynamics of the system can be ignored; one need only observe capacity constraints (see e.g. Lin *et al.*, 2006).

In trying to give an account of essentials, it is difficult to do justice to a large and complex literature. We follow, in simplified and somewhat modified form, the line of analysis developed by F. P. Kelly and his coworkers (see e.g. Gibbens and Kelly, 1999; Kelly, 2000, 2003; Kelly *et al.*, 1998; Kelly and Williams, 2004; Key *et al.*, 2004), which has turned out to be particularly fruitful. This starts from a quasi-optimisation of the problem, intended to express an enlightened naive policy. Suppose that the vector of input rates u is chosen to maximise a utility function

$$V(u) = \sum_r w_r \frac{u_r^{1-\alpha}}{1-\alpha} \tag{17.1}$$

subject to the feasibility conditions

$$Au \le b, \quad u \ge 0, \tag{17.2}$$

where the parameter α lies in the interval $0 \le \alpha < 1$. The concavity of the utility function (17.1) is enough to ensure that all elements of the optimising u are strictly positive; the particular power form (suggested in this context by Mo and Walrand, 2000) is one that

has proved realistic and tractable in many optimisation contexts. The criterion is chosen to express the attribute of 'fairness', in that the weighting by the coefficients w_r is the only expression of a relative valuation of the different types of traffic. There is no enforcement of priority in any other way. The perceptive reader may observe that account should also be taken of sender states, i.e. of the current backlogs x. We take up this point in the next section.

Condition (17.2) is designed to keep the policy feasible; i.e. to make sure that no link is overloaded. However, the naive controller is not aware of the conditions for noncongestion; he knows congestion only when he sees it. Suppose then that a congestion cost rate $C(u)$ is incurred when the system reaches equilibrium under a fixed input vector u and, to anticipate the argument, that observed congestion will reveal something of the character of this function. Plausibly, $C(u)$ will have the additive form

$$C(u) = \sum_j C_j(\sum_r a_{rj} u_j) = \sum_j C_j(v_j), \tag{17.3}$$

where $C_j(v_j)$ is the cost incurred at link j if this is carrying work at rate v_j.

The suggestion is now that u should be chosen to maximise the modified utility function

$$\phi(u) = V(u) - C(u)$$

freely in the positive orthant: $u \geq 0$. In fact, explicit maximisation is not a practicable operation to perform on-line in continuous time. A more realistic algorithm is to follow a continuous steepest-ascent rule of the type

$$\dot{u}_r = \kappa \frac{\partial \phi(u)}{\partial u_r}. \tag{17.4}$$

Under assumptions (17.1) and (17.3) relation (17.4) becomes

$$\dot{u}_r = \kappa[w_r u_r^{-\alpha} - \sum_j a_{rj} p_j(u)], \tag{17.5}$$

where

$$p_j(u) = \frac{\partial C_j(v_j)}{\partial v_j}.$$

The term $p_j(u)$ represents the marginal cost of additional traffic on link j, and the evolution rule (17.5) for the quasi-optimal input rates then has a very natural form. The sum over j is the sum of marginal costs incurred at links on the route r under the current input rates u. If this can be identified with the congestion warning signal that is being returned to sender r from the system, then rule (17.5) is of the TCP type.

The beauty of this formulation is that stability is guaranteed, since $\phi(u)$ is by definition a Liapunov function for these dynamics, increasing monotonically on the orbit. There are two points to be noted, however. Firstly, the rule takes no account of the work rates that are *demanded*, in that it shows no dependence on x. We deal with this point in the next section. Secondly, the approach works only because the degree of congestion in the system under input rates u is assumed to be a function of u. One would think that

it is more properly a function of z, of the actual current queue sizes in the system, and z is a complicated function of past u. One may say that this makes z a function of the current value u if u is effectively constant over the relevant time interval, but then the z-values would be simply zero if the constraint (17.2) were satisfied. The assumption of a nontrivial dependence of z upon u is a concealed appeal to a statistical version of the model: that under a fixed value of u satisfying (17.2) positive queues will form, and that these queues will increase (statistically) as u approaches the boundary of the region $\mathcal{U}$ specified by (17.2).

17.2 A stable regulation rule

We now make the model sensitive to existing backlogs x by a line of argument developed successively by Massoulié and Roberts (2000), de Veciana *et al.* (2001), Bonald and Massoulié (2001) and Key *et al.* (2004). Suppose the definition (17.1) modified to

$$V(u|x) = \sum_r w_r x_r \frac{b_r^{1-\alpha}}{1-\alpha} = \sum_r w_r x_r^{\alpha} \frac{u_r^{1-\alpha}}{1-\alpha} \tag{17.6}$$

where $b_r = u_r/x_r$ is the *bandwidth* alloted to r-traffic. The x-dependence of the coefficients increases the priority of those routes for which there is a heavy backlog of demand. The backlog vector x obeys the equation

$$\dot{x} = \rho - u \tag{17.7}$$

if no overspill is explicitly returned, and the suggestion is now that one chooses the value of u by maximising the modified criterion function

$$\phi(x, u) = V(u|x) - C(u).$$

The time-dependence of the quantities x and u is understood, and the maximisation of ϕ is carried out afresh for every value of t. The agent performing the maximisation is not supposed to be all-seeing; he knows that $V(u|x)$ has the form (17.6), but is not aware of relation (17.7) and has only observational information on $C(u)$ – a point we shall deal with later by replacing the maximisation by a steepest ascent with respect to u, as in the last section.

In this section we are concerned with performance of the rate control rule deduced in this way, under plausible conditions. In trying to prove stability of the rule we can no longer assert that ϕ is increasing along the path, because ϕ is now dependent also upon x, whose dynamics are prescribed by (17.7) rather than by steepest ascent of ϕ.

We shall assume simply that the function $C(u)$ is convex, increasing in all its individual arguments u_r and increasing to $+\infty$ as u approaches the open boundaries of the permissible region $\mathcal{U}$ specified by (17.2), itself then an open set. By *stability* of the procedure (i.e. of the choice of $u(t)$ as the value maximising $\phi[x(t), u(t)]$) we mean that under it $x(t)$ and $u(t)$ tend to finite limits, independent of initial conditions, as $t \to +\infty$. Under conditions specified by Dai (1995) this implies a similar convergence to a unique limit distribution of $x(t)$ and $u(t)$.

Theorem 17.1 *Under the assumptions above, the necessary and sufficient condition for stability of the procedure is that $\rho \in \mathcal{U}$. In the stable case x and u converge to the values*

$$x_r = \rho_r \left(\frac{1}{w_r} \frac{\partial C(\rho)}{\partial \rho_r} \right)^{1/\alpha}, \quad u_r = \rho_r. \tag{17.8}$$

Proof The condition is plainly necessary: u must have a long-term average equal to ρ if x is not to become infinite, but this must imply infinite cost if $\rho \notin \mathcal{U}$. To deduce sufficiency, note that $\phi(x, u)$ is strictly concave in u, so that if both ρ and u lie in $\mathcal{U}$ we have

$$\phi(x, u) \leq \phi(x, \rho) + \nabla_\rho \phi(x, \rho)(u - \rho),$$

with equality only for $u = \rho$. Here $\nabla_\rho \phi(x, \rho)$ is the gradient of $\phi(x, \rho)$ with respect to ρ, a row vector. We have then

$$\nabla_\rho \phi(x, \rho)(\rho - u) \leq \phi(x, \rho) - \phi(x, u) \leq 0,$$

if u is the value maximising ϕ.

Consider now the function

$$L(x) = \sum_r w_r \frac{x_r^{1+\alpha}}{(1+\alpha)\rho_r^\alpha} - \nabla_\rho C(\rho) x,$$

which will prove to be an effective Liapunov function. We have

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \nabla_x L(x) \dot{x} = \nabla_\rho \phi(x, \rho)(\rho - u) \leq 0.$$

Thus $L(x)$ is nonincreasing on the orbit and must converge to its minimal value. The value of x that yields this minimum is unique and given by (17.8). Relation (17.7) implies that ρ is the only possible equilibrium value of u. $\diamondsuit$

When we simply demanded the constraints (17.2), as in Section 17.1, and these were satisfied by $u = \rho$, then the optimal equilibrium value of x was $x = 0$. That is, there was zero backlog in equilibrium. Now that the constraints are enforced by the penalty function $C(u)$, there is the small equilibrium backlog determined in (17.8). The point is that there is now some cost to holding a value $u = \rho$ near the boundary of $\mathcal{U}$, i.e. to holding packets in the buffer queues. It is then best to split the holding between the queue and the backlog store.

17.3 An adaptive rate control

The maximisation of $\phi(x, u)$ with respect to u cannot be performed, both because this is impractical as an on-line operation and because the penalty function $C(u)$ is not known. However, as in Section 17.1, one can follow a steepest-ascent path by using a relation such as

$$\dot{u}_r = \kappa \frac{\partial \phi(x, u)}{\partial u_r}$$

to determine the admission rate. This now yields the evolution equation

$$\dot{u}_r = \kappa[w_r(x_r/u_r)^\alpha - C^{(r)}(u)], \tag{17.9}$$

where

$$C^{(r)}(u) = \frac{\partial C(u)}{\partial u_r}$$

can be identified with the congestion warning signal fed back to the sender. Relation (17.9) then represents the type of realisable optimising control foreshadowed in Section 17.1. However, the replacement of the explicit ϕ-maximising expression for u by the maximum-seeking relation (17.9) implies a slower dynamic response, and one must then ask whether the system remains stable under this relaxation.

Theorem 17.2 *Under the conditions on $C(u)$ stated above the only possible equilibrium solution*

$$\overline{x}_r = \rho_r[C^{(r)}(\rho)/w_r]^{1/\alpha}, \quad \overline{u}_r = \rho_r$$

of equations (17.7) and (17.9) is at least locally stable.

Proof If ξ and η are perturbations of x and u from these equilibrium values then these obey the linear dynamic equations

$$\dot{\xi}_r = -\eta_r, \quad \dot{\eta}_r = f_r\eta_r - \sum_s g_{rs}\eta_s,$$

where

$$f_r = \kappa\alpha\rho_r^{-1}w_r^{1/\alpha}[C^{(r)}(\rho)]^{(\alpha-1)/\alpha}, \qquad g_{rs} = \kappa[\alpha\delta_{rs}C^{(r)}(\rho)/\rho_r + C^{(rs)}(\rho)]$$

and

$$C^{(rs)}(u) = \frac{\partial^2 C(u)}{\partial u_r \partial u_s}.$$

If we define the positive-definite quadratic form

$$Q(\xi, \eta) = \sum_r (f_r\xi_r^2 + \eta_r^2)$$

then

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\kappa\left(\alpha\sum_r C^{(r)}\rho_r^{-1}\eta_r^2 + \sum_r\sum_s C^{(rs)}\eta_r\eta_s\right) \le 0, \tag{17.10}$$

where the ρ-argument of the C-derivatives is understood. Because C is strictly increasing and strictly convex, the intermediate expression in (17.10) can be zero only if η is zero. By the familiar Liapunov argument we deduce then that the equilibrium at $\xi = \eta = 0$ is stable. $\Diamond$

The fact that stability is so clear under such mild conditions on $C(u)$ indicates that response to the particular form of the congestion warning signal is quite robust. Note, also, that the assertions of the theorem are independent of the value of ρ (provided this lies in the permissible region $\mathcal{U}$), so the control adapts to the value of ρ, and would indeed adapt to changing ρ if the rate of change were not too rapid.

The proof of Theorem 17.2 brings an interesting point to light: that the individual pairs of variables (x_r, u_r) would each perform a simple harmonic oscillation about the equilibrium point (at least when sufficiently close to this point), were it not for the global damping induced by the congestion warning feedback.

There are various caveats that should be entered, however. Because the congestion warning arises from earlier inputs to the system, the value of u in equation (17.9) is actually somewhat lagged relative to the value in (17.7). This lag in response will affect the stability analysis. A more thorough investigation of packet dynamics would require the study of full network dynamics on a finer time-scale; this certainly implying a full consideration of lags. Papers by Vinnicombe (2002a,b) and Paganini *et al.* (2005) represent heroic efforts in this direction.

18

Evolving networks and the Worldwide Web

In Chapters 2, 4 and 7–9 we considered the evolution of a network towards optimality by, essentially, following the steepest-descent path of a prescribed criterion function, even if sometimes this calculation took a simple and intuitive form. However, one generally thinks of 'evolution' as something much less premeditated, in which some simple and plausible response rule induces structural changes in the system, whose effect and whose effective guiding principle (if any) become evident only with time. In the case of Darwinian evolution, the effective guiding principle is that of survival. The degree of sophistication induced by this stark criterion still lies beyond our full comprehension, and is comparable in scale only to the time and tirelessness of trial needed to achieve it.

Interest in these matters is demonstrated by the growing literature on evolving automata, artificial neural networks and the like. However, the Worldwide Web now presents an example, if simple, of spontaneous evolution in a technological context.

18.1 Random graphs and the Web

Real-life response rules will in general be local, in that a change is induced in some part of the system in response to stimuli or information manifest in that part. The execution of the rule and the information on which it is based will also show natural variability, i.e. be 'random'. In the case of networks we are thus considering the generation of a random graph from local rules, and are interested to see whether the network thus generated can be seen as a statistical solution to an implied optimisation problem, and what its character may be.

There is now a large literature on random graphs. This was begun by Erdös and Rényi (1959, 1960), with their studies of a graph on N nodes. They assumed that arcs formed between the $N(N-1)/2$ distinct node pairs independently and with a probability $p = a/N$, so giving the nodes an expected degree of a. (The degree of a node is the number of arcs that meet at that node.) The implications of this specification were developed very fully by Bollobás in a series of publications, summarised in Bollobás (2001). However, in a series beginning in 1965 and summarised in Appendix 3 the author developed the theme in quite a different spirit. He set up a dynamic Markov model of arc formation and severance (bond creation and annihilation) whose equilibrium distribution could, in virtue of a reversibility assumption, be determined. This distribution contained both potential and combinatorial (energy and entropy) factors. A special case that developed very

naturally amounted to a general first-shell model of polymer formation. The stochastic formulation itself implies an extremal principle, in that a reduced description of the configuration will take a value close to one of its locally most probable values if the system is large.

These models do indeed provide quite a realistic treatment of polymer formation, but it has become apparent over the last few years that there are important networks whose mode of formation and statistical character seem to be quite different. The particular feature that has been seized upon as significant is the distribution of degree of a randomly chosen node. The Erdös–Rényi model predicts a Poisson distribution; the Whittle models are considerably more general, but still predict a distribution with an exponentially decreasing tail in all but quite extreme cases. However, a number of authors have observed that, in the case of the Worldwide Web, the degree distribution tails off much more slowly: as a power law. The same has been observed for the Internet and for data on co-authorship, ecological nets, word use, power grids and other nets whose growth one would not have expected to be particularly purposeful.

More specifically, the nodes of the Web are taken as the documents, the webpages, and the arcs are the hyperlinks that relate one document to another. The arcs are thus directed, but this fact can be ignored in a first treatment. If degree is denoted by k then the distribution is found to decay as $k^{-\gamma}$ for a range of large k covering several orders of magnitude. The exponent γ is found to take values between 2 and 3.

One of the first discussions of this phenomenon was that by Albert *et al.* (1999), and it was in Barabási and Albert (1999) that a mechanism was suggested that generated both this power-law decay and a predicted value of the exponent γ. It was supposed that new documents were added to the Web in a steady stream, and that each such new document made reference to m randomly chosen existing documents. The random rule was assumed to show 'preferential attachment', however, in that the probability of attachment to a node of degree k is supposed proportional to k. This reflected the fact that a document already established as influential is more likely to attract attention. The evolutionary rule thus formulated is then marked by two key features: continued growth and preferential attachment.

18.2 Evolution of the scale-free network

Barabási and Albert (1999) used deterministic arguments to derive the expected distribution of degree, in essence appealing to mean field properties of the model. The distribution is of course itself random in a system of finite size. We give the argument in a modified form due to Dorogovtsev *et al.* (2000). A rigorous stochastic treatment was later given by Bollobás *et al.* (2001).

Let $n_k(t)$ be the number of nodes of degree k at time t. It is supposed that a new node enters the system at every moment of discrete time, and makes a link with each of m existing nodes. The probability that the first of these links is with a given node whose degree is k_i is then

$$\frac{k_i}{\sum_k k n_k} = \frac{k_i}{2(m_0 + mt)},$$

since $\sum_k kn_k$ counts the number of arc-ends and so is twice the number of arcs. The rate at which arcs of degree k are converted to arcs of degree $k+1$ in this time step is thus

$$\frac{mkn_k}{2(m_0+mt)} = c_k(t)n_k,$$

say, plus terms of smaller order in t. For large t the probability that the m choices of node are not distinct becomes negligible, as does the change in the denominator during the time step. The evolution equation for the n_k thus becomes

$$n_k(t+1) = \begin{cases} n_k(t) - c_k(t)n_k(t) + c_{k-1}(t)n_{k-1}(t) & \text{if } k > m, \\ n_m(t) - c_m(t)n_m(t) + 1 & \text{if } k = m. \end{cases} \tag{18.1}$$

Now assume that $n_k(t)$ has the form $\rho_k t + \mathrm{o}(t)$ for large t. One then finds from (18.1) that

$$\rho_k = \frac{k-1}{k+2}\rho_{k-1} \quad (k > m)$$

whence it follows that the normalised distribution of degree is

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \quad (k \geq m). \tag{18.2}$$

This is indeed of order k^{-3} for large k.

A network of this character has come to be termed 'scale-free', in that, at least for large k, expression (18.2) is homogeneous in k. That is, a change of k-scale affects p_k only by a change in scale. The Barabási–Albert model has been elaborated and extended by many authors. These variations lead also to variations in the deduced value of the exponent γ, but all falling in the observed range of 2 to 4.

18.3 Graph properties

In Appendix 3 we summarise a theory of random graphs which greatly generalises the Erdös–Rényi model. This is based upon a full dynamic–stochastic model of association and dissociation, and amounts to the so-called first-shell theory of polymerisation. It involves a generating function $H(\xi) = \sum_k H_k \xi^k/k!$ whose coefficients H_k reflect the bonding strength of a node of degree k. It also contains one other significant parameter, ν (denoted q in the statistical–mechanical literature), which is the ratio of the rate of formation of a given type of link *inside* a given component of the graph (molecule of the configuration) to formation of the same link *between* existing components. The case $\nu = 0$ thus reflects the case when components are restricted to being trees. Such models exhibit the familiar transition in which, as soon as some activity parameter (such as the spatial density of nodes) is increased past a critical point, the graph becomes almost completely connected. This is the gel-point in the polymerisation context. There is another possible transition, however, that of Potts criticality. In this there is a discontinuous transition to a regime in which components have a degree of internal bonding which is so high as to be almost irreversible. For ν above a certain value this transition will precede the gel-point (i.e. occur at a lower value of the activity parameter).

It is then of interest to see what the distribution of node degree k would be for such models. For sufficiently regular cases it turns out that the probability generating function $\Pi_0(z) = E(z^k)$ is proportional simply to $H(\overline{\xi}z)$, where the value $\overline{\xi}$ of the activity is determined by the spatial density of nodes. One such case is that in which $\log H(\xi)$ is of less than quadratic growth at infinity and $\nu = 1$; the result then holds above or below the gel-point. Another is that in which one requires merely that $H(\xi)$ have a finite radius of convergence, but that $\nu = 0$; all but tree components are forbidden. This latter case can then produce a power-law degree distribution if $H(\xi)$ has a branch point and if one can take $\overline{\xi}$ right up to the branch point. This is really the only way in which such a distribution can be generated from these models, although they readily demonstrate a power law 'with exponential cut-off' if one chooses $\overline{\xi}$ somewhat short of the branch point.

Newman *et al.* (2001) generated graph components by a branching mechanism which, essentially, specified $\Pi_0(z)$. This is an approach developed by a number of earlier authors; see Watson (1958), Good (1960, 1963), Gordon (1962). It is not clear that such an approach is equivalent to the free association/dissociation of a cloud of 'free nodes'; we demonstrate in Section A3.5 of Appendix 3 that, with one reservation, it is. However, the mechanism will only generate trees; it is restricted to the case $\nu = 0$.

The statistical–mechanical approach leads to and exploits the notion of the *thermodynamic limit*: that one has an infinite 'sea' of nodes of prescribed spatial density (a canonical ensemble), and that bond formation, bond release and movement take place freely in this sea. It would seem that the mechanism behind the Web is intrinsically different.

We have then to consider a closed system of N nodes, perhaps set in a notional volume V. We should also accept it as likely that the generating mechanism of the scale-free graphs differs fundamentally from that of polymers. This then prompts the question: does the Barabási–Albert mechanism produce a graph that is optimal in some sense?

The power law suggests something verging on a hierarchical structure, a point made in Albert and Barabási (2002). The differentiation of levels in the hierarchical model of Chapter 5 can be seen as analogous to the differentiation by age in the Barabási–Albert model, the older nodes being those with the greater richness of connections. However, a more detailed examination weakens the analogy. The aim of a hierarchical model is (in the telephone context) to achieve economical connection between any pair of subscribers between whom a call is likely. The aim of the Web is to set up connections between bodies of knowledge that have a natural affinity, and these are direct connections, not necessarily mediated by connection via a node with a wider remit. In this sense it is a 'reference network', to use the term suggested by Dorogovtsev and Mendes (2000, 2003).

Points generally made of the Web net are that it shows a tendency to clustering (in the sense that the neighbours of a given node are more likely to be mutual neighbours themselves) and that it enjoys the 'small world' property. That is, the average number of steps required to find a connecting path between two randomly chosen nodes grows only slowly with N. If one considers the nodes as the lattice points on a regular cubic lattice in d dimensions then this average path length is of order $N^{1/d}$. However, if one considers a random graph of the Erdös–Rényi type with each node having degree k exactly then the average path length is asymptotic to $\log N/\log(k-1)$, and there is an analogous result for more general degree distributions.

This same logarithmic rate of growth can be achieved by a hierarchical system. Suppose we have a system of $N = k^s$ nodes, these nodes being regarded as analogous to the telephone subscribers of Chapter 5. Suppose that s levels of exchange are allowed, with each exchange linked to k exchanges of the level below it. Then one finds that the average path length of calls between subscribers is $2s = 2\log N/\log k$, the 2 coming from the fact that a call must ascend the hierarchy to some level and then descend. The extreme in this direction would be to take $k = N$, giving a universal path length of 2. That is, of a star network with a single hub. Passage to this extreme case emphasises the factors that have been neglected: the costs implied by the physical length of a link and by the realisation of a link or exchange that will carry a multitude of addressed packets.

The paper by Li *et al.* (2004) makes a number of cogent points. The authors were concerned with the Internet, which is also observed to show scale-free properties. They observed that prescription of degree distribution alone by no means determines the statistical characteristics of the network, and demonstrated this by the construction of five distinct networks with the same plausible degree distribution. Performance was measured by traffic successfully carried. Those nets that were generated randomly (either by preferential attachment or by a branching mechanism) performed abysmally; the heavily connected nodes that should have served as routers did not have the capacity to do so. The net that showed by far the best performance was essentially a hierarchical net with capacities (bandwiths) adapted to load. This echoes the conclusions of Chapter 5.

On less substantial points, the notion that nodes should accumulate indefinitely seems somewhat unrealistic. The Web documents that constitute the nodes do not preserve their interest or relevance for ever, and it would be natural to assume that they effectively make a quiet departure at some time. This a point which has been considered by a number of authors (e.g. Dorogovtsev and Mendes, 2000, who concluded that the power law indeed then has an exponential cut-off). Particularly for a stochastic model, it would be natural to allow node-death as well as node-birth; the system could then reach a condition of stochastic equilibrium, which would be the natural one to study. However, there are difficulties. A node-death implies more than a simple removal; it implies also a change in the states of the neighbours of the late node. The polymerisation model described in Appendix 3 presented the same problem. However, the assumption of time reversibility (and so of detailed balance) made it tractable, leaving only some combinatorial challenges to be faced for reduced descriptions of the system. So, it is the irreversibility of the Web model which complicates analysis. In the next section we consider a very much simpler model than that of the Web, proposed by the author in another context. It shows passage to an equilibrium that balances births and deaths, but in which the analogue of preferential attachment does indeed produce power-law distributions.

18.4 Emulation and the power law

One may say that the Barabási–Albert mechanism replaces optimisation by emulation. That is, the connections that a new node is to make are determined, not by a system optimisation, but by a reaction to the apparent ‘success’ of other nodes. Whittle (2004) considered what amounted to a model of activity allocation, in which the members of a population may choose various economic roles. They may also change roles, such a change being equivalent to death in one role and birth in another (although there is no

difficulty in formulating an 'open' model, if desired). Optimisation of allocation amounts to a simple linear programme, in the equilibrium deterministic case. However, the twist was given that individuals made their choice by emulation rather than by optimisation. That is, by adopting roles that seemed to be successful rather than by making the economic calculation themselves. This leads to smooth dynamics and a plausible 'naive' policy, but leads also to power-law distributions in the stochastic case. The model is simpler than the network model in that a new 'node' is only choosing a role for itself rather than a set of links to other nodes.

We consider initially a deterministic model with continuous variables. Let x_r be the scaled number of individuals in the population who adopt role r. We shall assume birth–death rules that conserve $\sum_r x_r$, so with a normalisation we shall have $\sum_r x_r = 1$, and can identify x_r as the proportion of the population in role r. Suppose that an individual in role r gives an economic yield of w_r per unit time and consumes an amount a_{jr} of resource j per unit time. Then the optimisation problem is simply to choose x to maximise $\sum_r w_r x_r$ subject to $x \geq 0$, $\sum_r x_r = 1$ and

$$\sum_r a_{jr} x_r \leq b_j \quad (j = 1, 2, \ldots, J). \tag{18.3}$$

Here b_j is the scaled rate of supply of resource j per unit time.

This then is just a simple linear programme. One can choose the distribution x to maximise the Lagrangian form

$$\sum_r w_r x_r + \sum_j \lambda_j (b_j - \sum_r a_{jr} x_r) = \sum_j \lambda_j b_j + \sum_r x_r A_r(\lambda),$$

say. Here the Lagrange multiplier λ_j can be regarded as an effective unit price for resource j, and one sees that the distribution x is concentrated on the values of r for which $A_r(\lambda)$ is maximal. The prices λ_j are determined by the conditions (18.3) or, equivalently, as a solution of the dual linear programme: the nonnegative values minimising $\sum_j \lambda_j b_j + \max_r A_r(\lambda)$.

Denote solutions of the primal and dual problems by $\overline{x}$ and $\overline{\lambda}$. If we retain only those values r for which the vectors $(w_r, a_{1r}, a_{2r}, \ldots, a_{Jr})$ are extreme members of the set of such vectors (as can be done without loss) then $\overline{x}$ is unique. If we furthermore assume, for simplicity, that all constraints are active at the optimum, then the distribution $\overline{x}$ satisfies (18.3) with equality, and its support is restricted to a basis set $\mathcal{B}$ of at most $J+1$ values of r.

This conventional analysis would then solve the problem completely if it were realistic to assume that the optimiser could immediately impose a distribution x on the population. However, suppose that the optimiser can control the distribution only by management of λ, which is indeed to be regarded as a vector of prices levied (to individuals) on the resources. Suppose also that the response of the population is not instantaneous (which would lead to discontinuous changes in x) but is subject to its own continuous dynamics.

These response dynamics must be such as to lead x to concentrate on the values of r maximising $A_r(\lambda)$. However, for individuals to immediately choose their role in this way would lead to abrupt swings of policy in the whole population, and would give no guide as to which policy to choose if the maximum of A_r were achieved for several values

of r, as must ultimately be the case. A rule that achieves smooth dynamics, expresses the preference for larger A_r sufficiently strongly and leaves the size of the population invariant is

$$\dot{x_r} = x_r(A_r - \sum_s x_s A_s) \quad (r = 1, 2, \ldots, R) \tag{18.4}$$

if we take the λ argument of A_i as understood. Here the dot indicates the rate of change of x with time t.

The dynamics expressed by (18.4) are exactly those of a laboratory culture of a mixture of organisms subject to continuous dilution, in which strain r grows at rate $A_r x_r$, but the culture is diluted to maintain a constant total organism density of unity, say. The culture will then ultimately contain only those strains of greatest growth rate A_r, if these are constant in time. In the role-allocation context this is analogous to a situation in which individuals do not choose their role on a rational basis, but rather by favouring a role that is already popular. Role numbers are then growing in proportion to their existing sizes, i.e. 'reproducing', but subject to an overall death rate that preserves total numbers. If we accept this model, then the question is whether there is a way of varying the price vector λ so as to ultimately bring x to the desired value $\overline{x}$.

Let us assume that a scaled buffer stock u_j is maintained of resource j, and that maintenance of these stocks incurs a scaled cost of $C(u)$ per unit time. This reflects the cost of both excessive reserves and of insufficient reserves (when stocks may run out and force departures from the optimal pattern $\overline{x}$). We shall assume $C(u)$ convex and differentiable, with a unique minimum at an ideal inventory value $\overline{u}$. The dynamic equations (18.4) are then to be supplemented by the equations

$$\dot{u}_j = b_j - \sum_r x_r a_{rj} = -d_j \quad (j = 1, 2, \ldots, J), \tag{18.5}$$

say. It may be that there are resources whose limitations imply no constraint at the optimal equilibrium, but may do so during passage to that equilibrium. We shall simply assume that sufficient stocks are held to meet these contingencies, and shall leave such resources out of the analysis.

Let $\overline{A}$ be the value of $A_r(\overline{\lambda})$ for r in $\mathcal{B}$, and set $\mu = \lambda - \overline{\lambda}$, so that μ represents the variable perturbation in λ by which we hope to steer the system to its ideal configuration. Modification of A_r to

$$A_r(\lambda) - \overline{A} = -\delta_r - \sum_j a_{jr}\mu_j, \tag{18.6}$$

say, does not affect equation (18.4). At the optimum the constants δ_r are zero for r in $\mathcal{B}$ and strictly positive outside it. Denote the partial differential $\partial C(u)/\partial u_j$ by c_j.

Theorem 18.1 *Consider the system governed by the dynamic equations (18.4) and (18.5). If the constants f and g are strictly positive then the control rule*

$$\mu_j = -fc_j + gd_j \quad (j = 1, 2, \ldots, J) \tag{18.7}$$

takes the system to its optimal configuration.

Proof Define the quantity

$$D = -\sum_r \overline{x}_r \log x_r.$$

As a function of x it has its unique minimum at $\overline{x}$. Consider the nonnegative function

$$L(x, u) = D(x) + fC(u).$$

Then one finds from relations (18.4)–(18.6) that

$$\frac{\mathrm{d}}{\mathrm{d}t}L = -g\sum_j d_j^2 - \sum_r \delta_r x_r.$$

The Liapunov function L is thus nonincreasing with time and ceases to decrease only when all d_j are zero and x is confined to the basis $\mathcal{B}$. This implies indeed that $x = \overline{x}$. Equations (18.4) and (18.7) then imply respectively that $\mu = 0$ and $c = 0$, so that $u = \overline{u}$. Convergence to the target values is thus established. ♢

Suppose now that we stochasticise the model to the extent that the control relation (18.7) is replaced by

$$\mu_j = -fc_j + gd_j + \epsilon_j,$$

where the ϵ_j are independent white noise processes of power v. The assumption is thus that the control can be implemented only imperfectly. Then standard arguments are used in Whittle (2004) to show that in equilibrium the random variables x and u have joint probability density

$$\rho(x, u) \propto \exp[-2(g/v)L(x, u)] = \exp[-2(fg/v)C(u)]\prod_r x_r^{\gamma_r} \tag{18.8}$$

where $\gamma_r = 2g\overline{x}_r/v$. The final product in (18.8) indicates that the proportions x_r obey power-law distributions, and would be independent if they were not subject to the constraint $\sum_r x_r = 1$. This a consequence of the 'preferential growth' expressed by (18.4). There is another consequence, however. If ever x_r falls to zero for some particular r then it will remain there; a role that depends on emulation for its survival can never reappear once it dies out. Validity of (18.8) then requires that there be a mechanism such as immigration or mutation, however feeble, which can revive an extinguished role.

Appendix 1

Spatial integrals for the telephone problem

This appendix supplies the technical background for the discussion of Section 5.7. Physical space is taken as p-dimensional Euclidean space, $\mathbf{R}^p$. For simplicity, we use terms such as 'cube', 'plane', 'volume' and 'surface area' to denote the p-dimensional equivalents of the familiar three-dimensional concepts. The volume and the surface area of the unit ball in $\mathcal{R}^p$ will be denoted by ω_p and σ_p respectively.

We shall not necessarily suppose the basic cell Γ to be cubic; the discussion is in many ways simpler if we consider a general case. However, we shall assume throughout that Γ is bounded and convex with a boundary that is locally planar almost everywhere. We shall continue to use Γ_N to denote the basic cell scaled by a factor of N, and to suppose that call rate $\rho(r)$ over distance r decays at least as fast as r^{-q} for some $q > p+1$.

A1.1 The limit outflow density

Lemma A1.1 *The limit outflow density has the form*

$$z(\infty) = (\omega_p/2) \int_0^\infty r^p \rho(r) \mathrm{d}r. \tag{A1.1}$$

Proof The conditions on Γ ensure that, for large N, the boundary at most given boundary points P of Γ_N appears as an infinite plane separating the infinite half-spaces of the interior and exterior of Γ_N. The outflow density at P can thus be regarded as the contribution from all points in Γ_N that lie on the perpendicular to the boundary at P. The density of contribution to this outflow over a distance r and from a point at depth w will thus be $\rho(r)$ times the surface area of the spherical cap constituted by that part of the sphere of radius r centred on the source at depth w that lies outside Γ_N; see Fig. A1.1.

If we now integrate over positive w then we essentially integrate uniformly over the external half of the sphere of radius r centred on P; see Fig. A1.2. Note that the contours in this figure are not contours of constant radius from P, but contours of the circle of fixed radius r and varying centre on the perpendicular. The integration nevertheless gives equal weight to all points of this external hemisphere.

With this observation the lemma is proved: the coefficient of $\rho(r)$ in expression (A1.1) is just the volume of a hemisphere of radius r. $\Diamond$

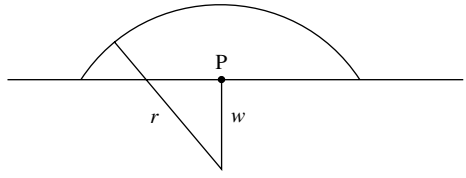


Fig. A1.1 The spherical cap over which one must integrate for calls over a distance r lying at a depth w inside the cell.

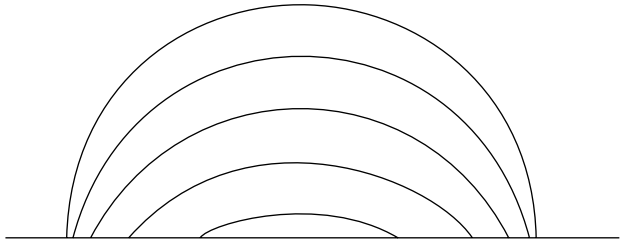


Fig. A1.2 Integration over w then gives the effect of a uniform integration over the hemisphere of radius r.

A1.2 Outflow and withinflow

Let us denote outflow $x(N)$ by $X(\Gamma_N)$, thus exhibiting it as a function of the cell rather than of the scale. We thus have

$$X(\Gamma) = \int_{\xi\in\Gamma}\int_{\eta\notin\Gamma} \rho(|\xi-\eta|)\mathrm{d}\xi\mathrm{d}\eta = \int_0^\infty h(r)r^{p-1}\rho(r)\mathrm{d}r, \tag{A1.2}$$

where $h(r)$ is a function to be determined. We can also define the *withinflow* of Γ

$$Y(\Gamma) = \int_{\xi\in\Gamma}\int_{\eta\in\Gamma} \rho(|\xi-\eta|)\mathrm{d}\xi\mathrm{d}\eta = \int_0^\infty g(r)r^{p-1}\rho(r)\mathrm{d}r, \tag{A1.3}$$

the total flow of calls between points of Γ. The function $g(r)$, also to be determined, is such that $g(r)r^{p-1}$ is proportional to the probability density of a random variable r, the distance between two points distributed uniformly and independently over Γ. The identity

$$\int_\Gamma\int_{\overline{\Gamma}} = \int_\Gamma\int - \int_\Gamma\int_\Gamma,$$

where $\overline{\Gamma}$ is the region outside Γ and the integrand is the same as that of the previous two equations, implies

Lemma A1.2 *The functions $h(r)$ and $g(r)$ are related by*

$$h(r) = V\sigma_p - g(r), \tag{A1.4}$$

where V is the volume of Γ.

One can now make an important sequence of assertions concerning the function g.

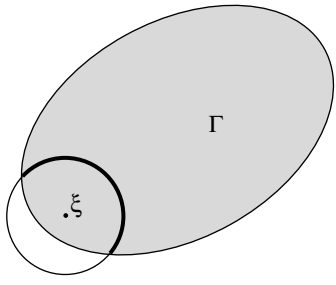


Fig. A1.3 The function $I_\Gamma(\xi, r)$ is the proportion of the surface of the sphere of radius r and centre ξ that lies inside Γ.

Theorem A1.3 *(i) The function $g(r)$ decreases monotonically and convexly to zero as r increases from zero to the maximal diameter D of Γ.*
(ii) One can also assert that

$$g(r) = V\sigma_p - (A\omega_p/2)r + o(r), \tag{A1.5}$$

for small r, where A is the surface area of Γ.
(iii) The g and h functions for a scaled cell Γ_N are $g_N(r) = N^p g(r/N)$ and $h_N(r) = N^p h(r/N)$ respectively.

Proof (i) Consider a sphere of radius r centred on a point ξ of Γ, and let $I_\Gamma(\xi, r)$ denote the proportion of the surface area of this sphere that lies within Γ (see Fig. A1.3). Then, for fixed ξ, I is a function of r that certainly has the properties asserted in (i). Since

$$g(r) = \sigma_p \int_\Gamma I_\Gamma(\xi, r)\mathrm{d}\xi, \tag{A1.6}$$

assertion (i) follows.
(ii) It follows from relation (A1.5) for $r = 0$ that $g(0) = V\sigma_p$. For positive r this value will be diminished by r^{1-p} times the total traffic at call distance r from the interior of Γ to its exterior, and it follows from the argument of Lemma A1.1 that this diminution is $r^{1-p}A\omega_p r^p/2 = (A\omega_p/2)r$, to within terms of smaller order.
(iii) This follows from a change of scale in the scaled versions of relations (A1.2) and (A1.3). ◊

A1.3 Character of the outflow density

Theorem A1.4 *The outflow density function*

$$z(N) = \frac{x(N)}{N^{p-1}A} \tag{A1.7}$$

belongs to $\mathcal{Q}$. More specifically, it increases monotonically and concavely from zero to the limit value $z(\infty)$ given by expression (A1.1) as N increases from zero to ∞.

Proof Let us for simplicity condense an expression such as $\int f(r)r^{p-1}\mathrm{d}r$ to $\int f(r)$, the missing elements under the integral sign to be replaced later. We have then from (A1.2), Theorem A1.3(iii) and (A1.7) that

$$z(N) = \frac{N^p}{n^{p-1}A}\int_0^\infty h(r/N) = (N/A)\int_0^\infty \min[V\sigma_p, h(r/N)]. \tag{A1.8}$$

The rather strange reformulation of the first integral in terms of the second follows from the fact that $g(r)$ changes from a positive value to a zero value as r increases through D, the maximal diameter of Γ, and generally changes analytical form in doing so. The analytical form that held at values of r immediately below D will in general be negative immediately for values immediately above. Correspondingly, we see from (A1.4) that $h(r)$ will increase to the value $V\sigma_p$ at $r = D$. It is in fact frozen at this value as r increases further, but it may well be that the analytical form that held at values of r immediately below D will continue to increase strictly as r passes through D. It is to forbid this infraction that we write expression (A1.8) in the final form. We can then write

$$Az(N) = N\int_0^{ND} h(r/N) + N\int_{ND}^\infty V\sigma_p,$$

since $h(r) = V\sigma_p - g(r)$ is increasing. From this relation we deduce that

$$Az'(N) = \int_0^{ND}[h(r/N) - (r/N)h'(r/N)] + \int_{ND}^\infty V\sigma_p > 0 \tag{A1.9}$$

and

$$Az''(N) = \int_0^{ND}(r/N)^2 h''(r/N) - Dh'(D) < 0. \tag{A1.10}$$

The final inequalities in (A1.9) and (A1.10) appeal to the fact that $h(r)$ is concave increasing, and so $h' \geq 0$, $h'' \leq 0$ and $h(r) - rh'(r) \geq 0$. The inequalities displayed ensure that the function $z(N)$ is likewise increasing and concave. Its limit value for large N has already been evaluated in (A1.1), but follows from (A1.4) and (A1.5) if we let N tend to infinity in (A1.8). $\Diamond$

Actually, if the boundary of Γ is not sufficiently smooth, then the analytic form of $h(r)$ may change at a number of r values, as we shall see in Section A1.5. The derivatives of h may suffer a discontinuity at these points, and the argument just given may have to be modified to take account of this. However, we saw from this argument that the possible switch of analytic form at $r = D$ did not invalidate the conclusion, and may expect that the same would be true for other switch points.

A1.4 Flows between cells

We have to consider the evaluation of what we have termed the 'terminal cost' $K(N)$, the cost per unit volume of termination at scale N. If one allows complete connection at this level then

$$K(N) = N^{-p}\sum_j Nd_j\chi(x_j(N)) = N^{1-p}\sum_j d_j\chi(x_j(N)), \tag{A1.11}$$

where j labels the final-level exchanges, $x_j(N)$ is the flow from a sender cell labelled by 0, say, to a receiver cell j, and Nd_j is the distance between these two exchanges. We thus have to form some estimate of the inter-cell flow $x_j(N)$ at this level.

In Section 5.7 we appeal to the following assertions concerning the character of $K(N)$.

Theorem A1.5 *Suppose that $\chi(x)$ has the form x^ν, and that $\rho(r)$ decays at least as fast as r^{-q} for large r, where*

$$q > \frac{p+1}{\nu}. \tag{A1.12}$$

Then: (i) If $\nu = 1$ the terminal cost function $K(N)$ increases from $2z$ to $2pz$ as N increases from zero to ∞, where $z = z(\infty)$ is the limiting outflow density evaluated in (5.40).
(ii) If $0 < \nu < 1$ and $p > 1$ then $K(N)$ decreases from ∞ to 0 over the same range.

Proof For small N the flow rate from one cell to another distant Nd is $N^{2p}\rho(Nd)$, to first order, so incurring a cost of $Nd[N^{2p}\rho(Nd)]^\nu$. The corresponding $K(N)$ equals the total cost per unit volume of the sender cell, which is

$$\begin{aligned} N^{-p}\sum_d Nd[N^{2p}\rho(Nd)]^\nu &= N^{2p(\nu-1)}\sum_d Nd[\rho(Nd)]^\nu N^{-p} \\ &\sim \omega_p N^{2p(\nu-1)}\int_0^\infty r^p\rho(r)^\nu \mathrm{d}r, \end{aligned} \tag{A1.13}$$

where ω_p is the volume of the unit ball in $\mathcal{R}^p$. Here the sum $\sum_d$ is over the distances d from the origin of the points of a unit cubic lattice in $\mathcal{R}^p$, so the last relationship in this sequence is valid for small N, and demonstrates the assertions of the theorem for small N.

Turning now to the case of large N, one sees that the flow from a sender cell to its immediate neighbour will be of order $N^s z$ if the cells share a face of dimension s ($s = 1, 2, \ldots, p-1$). The cost per unit volume of the sender cell for the flow between the two will then be of order $N^{1-p}N^s$, where the extra power of N comes from the scaling of the path between the two cells. The contribution of highest order thus comes from the $2p$ neighbouring cells sharing a $(p-1)$-dimensional face with the sender cell, which is of order $N^{-(p-1)(1-\nu)}$. This is vanishing for $\nu < 1$, and for $\nu = 1$ gives the known evaluation $2pz$.

However, we have yet to evaluate the contribution to $K(N)$ from non-neighbouring cells. The cost per unit volume of sender cell 0 of flow to cell j will be of order $N^{1-p}d_j[N^{2p}\rho(Nd_j)]^\nu \propto N^{1-p+(2p-q)\nu}d_j^{1-q\nu}$, if we assume the slowest rate of decay for ρ that assumptions permit. The total contribution to $K(N)$ from this source is thus of order of the sum of this latter expression over all nonzero j. Inequality (A1.12) ensures both that this sum is convergent and that N is raised to a negative power, and so that the total contribution to $K(N)$ from this source tends to zero with increasing N. $\Diamond$

A1.5 Cubic examples

In treating cubic examples we can confine attention to the unit cube, for we know from Theorem A1.3(iii) what the effect of a scale change of N would be. In the case $p = 1$ this is then the unit interval. We find readily that

$$\int_0^1 \int_0^1 \rho(|\xi - \eta|)\mathrm{d}\xi\mathrm{d}\eta = 2\int_0^1 (1-r)_+\rho(r)\mathrm{d}r, \tag{A1.14}$$

implying that

$$g(r) = (1-r)_+, \quad h(r) = 2 - g(r) = 2\min(1, r). \tag{A1.15}$$

Note how these expressions change analytic form as r increases through the maximal possible value, of unity.

The case $p = 2$ leads us effectively to evaluate the distribution of the distance between two points chosen randomly on the unit square. This was a problem first tackled by Borel (1925), and several authors have since considered different shapes and dimensions of cell. We shall summarise the calculation for the square.

The value of $g(r)$ for this case is to be determined from

$$4\int_0^1 \int_0^1 (1-r_1)(1-r_2)\rho(\sqrt{r_1^2 + r_2^2})\mathrm{d}r_1\mathrm{d}r_2 = \int_0^{\sqrt{2}} g(r)\rho(r)r\mathrm{d}r, \tag{A1.16}$$

since the absolute values r_1 and r_2 of the Cartesian components of the vector displacement between the two points are 'independently distributed' with density proportional to the g of (A1.13).

Transforming to polar co-ordinates (r, θ) in the double integral of (A1.16) we find that

$$g(r) = 4\int_{\phi}^{\pi/2-\phi} (1 - r\cos\theta)(1 - r\sin\theta)\mathrm{d}\theta \tag{A1.17}$$

where the limits on the θ integral are such as to constrain the co-ordinate point to the unit square. We see from Fig. A1.4 that we must choose $\phi = 0$ for $0 \le r \le 1$ and

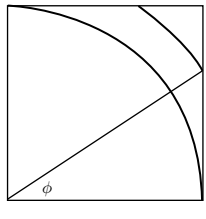


Fig. A1.4 Integration over the square in polar co-ordinates.

$\phi = \cos^{-1}(1/r)$ for $1 \leq r \leq \sqrt{2}$, while $g(r) = 0$ for $r \geq \sqrt{2}$. By determining the integral (A1.14) for these values of ϕ, using the relation

$$h(r) = 2\pi - g(r)$$

and performing the N-scaling, we deduce the weighting asserted at the end of Chapter 5.

Note the two switch points, corresponding to maximal distances along a side of the square or in the full square. In the p-dimensional case there will be switch points at $\sqrt{j}$ for $j = 1, 2, \ldots, p$.

Appendix 2

Bandit and tax processes

The seminal works are those of Gittins (1979, 1989), which presented the key insight and unlocked the multi-armed bandit problem, and Klimov (1974, 1978), which developed a frontal attack on the tax problem. Gittins had in fact perceived the essential step to solution by 1968, and described this at the European Meeting of Statisticians in 1972; see Gittins and Jones (1974). By this time Olivier (1972) had independently hit upon much the same approach. A direct dynamic programming proof of the optimality of the Gittins index policy was given in Whittle (1980), and then generalised to the cases of an open process and of the tax problem in Whittle (1981b, 2005).

For present purposes, we shall not follow the historical sequence of steps that suggested the Gittins index, but shall simply give the dynamic programming arguments that confirm optimality of the index policy and evaluate its performance. We consider only the undiscounted case, which shows particular simplifications and covers the needs of the text.

As mentioned in the text, there are at least two levels of aspiration to optimality in the undiscounted case. One is simply to minimise the average cost, and the second is to minimise also the transient cost: the extra cost incurred in passage from a given initial state to the equilibrium regime. As it turns out, the index policy is optimal at both levels.

A2.1 Bandit processes

Consider a Markov decision process with the same dynamics over states as that supposed for items over nodes in Chapter 14. That is, that projects of state j enter the system with probability intensity λ_j and move from state j to state k with probability intensity $a\mu_{jk}$ if effort a is exerted. We suppose also that reward is earned at rate ar_j if a project in state j is engaged at effort rate a. The state variable of the problem is $n = \{n_j\}$, where n_j is the number of projects currently in state j. It is traditional for the bandit problem that one thinks in terms of rewards rather than costs; a convention easily reversed when we turn to the tax problem.

We shall assume that the available effort is concentrated on a single project at a given time, so that a equals A for that project and zero for all others – distribution of effort is achieved by switching in time. The average reward γ and the transient reward $f(n)$ then obey the dynamic programming equation

$$\gamma = \max_j [r_j + \Omega_j f(n)] \tag{A2.1}$$

where

$$\Omega_j f(n) = \sum_k \lambda_k [f(n+e_k) - f(n)] + \sum_k \mu_{jk} [f(n-e_j+e_k) - f(n)] \tag{A2.2}$$

if we assume that $A = 1$. The general case is easily restored by multiplying r and μ by A. The maximum in (A2.1) is over those j for which $x_j > 0$.

One has new projects entering the system all the time. For balance, these must be lost either by discharge to an explicit 'write-off' state $j = 0$ or by accumulation in a set of states of lower promise, which effectively constitutes a write-off set.

For this problem the Gittins index is determined by the equation system

$$\max[r_j - \nu + \sum_k \lambda_k \psi_k + \sum_k \mu_{jk}(\psi_k - \psi_j), -\psi_j] = 0, \tag{A2.3}$$

where the ψ_j should really be written $\psi_j(\nu)$, to indicate their dependence upon the proffered retirement income ν. The value of ν for which state j is on the decision boundary (of resignation or continuation) is the Gittins index ν_j. Note that $\psi_j(\nu)$ is zero for all states j of index less than or equal to ν. Then the Gittins index policy would be to concentrate effort on one of the projects of currently greatest index. The typical situation would then be that there is some state s that marks the boundary, in that states j of index higher than ν_s are engaged so frequently that the corresponding n_j remain finite, that states of index lower than ν_s are ultimately never engaged, and projects in state s are engaged when nothing better is available. In these latter two cases the corresponding n_j become infinite in time. This sets the pattern for

Theorem A2.1 *The Gittins index policy is optimal. The average and transient costs have the evaluations*

$$\gamma = \nu_s, \quad f(n) = \sum_k \psi_k(\nu_s) n_k, \tag{A2.4}$$

for some state s, determined so that the process runs exactly at full capacity.

Proof It is convenient to order the states so that ν_j is nonincreasing in j. Suppose that the transient cost indeed has the linear form

$$f(x) = \sum_k f_k n_k. \tag{A2.5}$$

If we assume no natural discharge from the system then there must be a set of the states k for which n_k becomes indefinitely large; for these we must have $f_k = 0$. Relation (A2.1) then implies that the equation

$$\gamma = r_j + \sum_k \lambda_k f_k + \sum_k \mu_{jk}(f_k - f_j) \tag{A2.6}$$

holds for all those projects in states j that are fully used, and $f_j = 0$ for those that are not. But this system of equations exactly implies the system of equations (A2.3) if we make the identifications (A2.4). The determination of the cut-off value s by capacity considerations comes from an earlier stage in the argument (Whittle, 1981a): relations (A2.3) hold only if the system is working exactly to capacity. ◊

If one tries to set the cut-off point higher than is dictated by capacity, then one is indeed working with projects restricted to a more profitable level, but one does not realise the full benefit of this, because one then has idle time during which no reward is earned.

The bandit problem shows considerable degeneracy in the undiscounted case. Suppose that capacity exceeds need, so that all projects receive attention sooner or later. Then the order in which projects are engaged will not affect the total reward, and so will not affect the average reward. There is then no question of optimising γ, although the index policy does indeed maximise the transient reward $f(n)$. If the capacity constraint is active, then γ is indeed optimised by the index rule, to the extent that those projects to be operated (those above the optimal cut-off point s) are determined by the index rule. But one could indeed operate these projects in any order without affecting average reward.

A2.2 Tax processes

In the case of tax processes, one is minimising the average cost $cn = \sum c_k n_k$ rather than maximising the average value of reward r. The dynamic programming equation (A2.1) then becomes modified to

$$\gamma = \min_j [cn + \Omega_j f(n)], \tag{A2.7}$$

where the operator Ω_j again has the action (A2.2), if $A = 1$. For the general case the rates μ_{jk} would be replaced by $A\mu_{jk}$ in (A2.2), but the value of the cost vector c in (A2.7) would not be affected.

We shall now speak in terms of items rather than projects. The critical difference between the two cases is that all items present in the system now incur a cost, rather than just the items currently being processed. Items cannot then be lodged in a 'write-off' state; they must be processed to completion, and then discharged to a definite exit state $j = 0$. There is then also a stability condition: that the effort available should *exceed* requirements. So, for the bandit problem the interesting case was that for which work input exceeded capacity; for the tax problem, the only acceptable case is that for which capacity exceeds input.

For all that there are fundamental differences, treatment of the tax process can formally be derived from that of a rather more general version of the bandit process. One must consider the discounted version of the tax process, carry out the operation that converts it to a rather novel discounting of a bandit process, and then proceed to the undiscounted limit. There are also points to be made about the optimal stopping problem which are inherent in the treatment of both processes. This latter point is treated in Whittle (1981) and the passage between the two processes in Whittle (2005).

We shall give direct optimality proofs later, and shall for the moment merely assert that the tax analogue of the index-determining equation system (A2.4) is

$$\max[-\nu + \sum_k \lambda_k \Delta_k + \sum_k \mu_{jk}(\Delta_k - \Delta_j), -\psi_j(\nu)] = 0, \tag{A2.8}$$

where $\Delta_j = \psi_j(\nu) - c_j$. As before, the index value ν_j for items in state j is determined as the value of ν for which the alternative expressions in the square brackets of (A2.8) are equal.

For simplicity, let us consider first the case of the fluid limit, when the differences in the n-arguments of (A2.2) are replaced by differentials, and relation (A2.7) is replaced by

$$0 = \min_j [cn + \Omega_j' f(n)] \tag{A2.9}$$

where Ω_j' has the action

$$\Omega_j' f(x) = \sum_k \lambda_k f_k + \sum_k \mu_{jk}[f_k - f_j] \tag{A2.10}$$

and $f_j = \partial f(x)/\partial x_j$. We have set $\gamma = 0$ in (A2.9) because the system (supposed stable) will ultimately empty itself, and incur no further cost.

Theorem A2.2 *For the fluid limit version of the problem, specified in (A2.9), the index policy determined by (A2.8) is optimal, and the minimal transient cost takes the quadratic form*

$$f(x) = \frac{1}{2} x^\top Q x. \tag{A2.11}$$

Assume the states ordered by nonincreasing index. Then the elements Q_{jk} of Q are determined by the relations

$$c_i + \sum_k \mu_{jk}(Q_{ki} - Q_{ji}) + \sum_k \lambda_k Q_{ki} = 0 \quad (i \geq j). \tag{A2.12}$$

Equivalently, Q has the evaluation

$$Q = \int_0^\infty \frac{\partial \psi(\nu)}{\partial \nu} \frac{\partial \psi(\nu)^\top}{\partial \nu} \mathrm{d}\nu = \left[\int_0^\infty \frac{\partial \psi_j(\nu)}{\partial \nu} \frac{\partial \psi_k(\nu)}{\partial \nu} \mathrm{d}\nu \right], \tag{A2.13}$$

where $\psi(\nu)$ is the column vector with elements $\psi_j(\nu)$.

Proof The reason for the restriction on the set of i-values in (A2.12) is that $x_i = 0$ for $i < j$, and so there is neither need nor basis for such a condition in this range. Let us denote $\psi_j'(\nu)$ by p_j and $\sum_k \mu_{jk}(p_k - p_j) + \sum_k \lambda_k p_k$ by $L_j p$. Then, if evaluation (A2.13) is assumed, condition (A2.12) amounts to

$$c_i + \int_0^\infty (L_j p) p_i \mathrm{d}\nu = 0 \quad (i \geq j). \tag{A2.14}$$

But $L_j \Delta = \nu$ and so $L_j p = 1$ for $\nu < \nu_j$, and $\psi_i = 0$ and so $p_i = 0$ for $\nu > \nu_i$. The expression on the left in (A2.12) thus equals

$$c_i + \int_0^{\nu_i} p_i \mathrm{d}\nu = c_i + \psi_i(\nu_i) - \psi_i(0) = c_i + 0 - c_i = 0.$$

Expression (A2.11) thus satisfies the dynamic programming equation (A2.9). This solution is unique, so expression (A2.11) is indeed verified as the minimal cost. ◊

In an attempt to keep the main line of the argument clear we have not proved relation (A2.13) and its equivalence to (A2.12). For these matters we refer to Whittle (2005). The analogue of Theorem A2.2 for the stochastic case now follows immediately.

Theorem A2.3 *For the Markov version of the problem, determined by (A2.7), the index policy determined by (A2.8) is optimal, and the minimal average cost and transient cost are given by*

$$f(n) = \frac{1}{2} n^\top Q n + R^\top n, \quad \gamma = \sum_k \lambda_k f(e_k), \tag{A2.15}$$

where Q has the same evaluation as in the previous theorem, and the coefficients R_j are determined by

$$\sum_k \mu_{jk}[R_k - R_j + \frac{1}{2}(Q_{jj} - 2Q_{jk} + Q_{kk})] = 0. \tag{A2.16}$$

The proof follows as for the previous theorems. The minimal cost γ is determined by a linear equation accompanying (A2.16). However, the attractive form given in (A2.15), noted by a referee, follows from a consideration of the recurrence cycle between consecutive empty phases (i.e. those phases when for the moment there is no work).

A2.3 Adaptation to fixed work-stations

The dynamic programming (A2.7) is appropriate for the case when there is complete flexibility in the allocation of effort. However, the more realistic case is that in which processing is divided between work-stations, indexed by h. Each work-station can tackle only items that are ready for its particular process, which for station h corresponds to item states j lying in a set $\mathcal{P}_h$, say. We shall suppose for simplicity that work-stations cannot substitute for each other, so that the sets $\mathcal{P}_h$ are disjoint and their union covers all processes.

The processing effort available must also be divided; we shall suppose that work-station h disposes of a proportion q_h of the total unit processing effort available. We shall also suppose that work-stations have been balanced, in that q_h equals the proportion of effort spent on items in $\mathcal{P}_h$ under a free optimisation. An allocation decision now has to be made for every work-station, and equation (A2.7) is modified to

$$\gamma = \sum_h q_h \min_{j \in \mathcal{P}_h} [cn + \Omega_j g(n)]. \tag{A2.17}$$

Relation (A2.17) follows from the facts that μ_{jk} now becomes $q_h \mu_{jk}$ for $j \in \mathcal{P}_h$ (both expressions to be multiplied by A, in general) and $\sum_h q_h = 1$.

One might expect that each work-station should concentrate its effort on one of the items of highest index in its buffer (queue). However, to take this view is to ignore the effect that choice has on the progress of items of higher index.

Consider the effect of the operator Ω_j on the transient cost $f(n)$ for the freely optimised policy. We find that

$$cn + \Omega_j f(n) = \sum_{i<j} \omega_{ij} n_i, \tag{A2.18}$$

where ω_{ij} is the expression in the left-hand member of equation (A2.12), known to be zero for $i \geq j$. If we chose j in $\mathcal{P}_h$ to minimise expression (A2.18) we would be

determining the policy that would be optimal if one had to operate under the work-station constraints for an instant before reverting to optimal operation of the unconstrained case. It is plausible that this represents at least the direction in which policy should move if one had to operate under the work-station constraints indefinitely.

By the same argument as was used in Theorem A1.2 to evaluate ω_{ij} for $i \geq j$ we find that

$$\omega_{ij} = \int_{\nu_j}^{\nu_i} (L_j p - 1) p_i \mathrm{d}\nu \quad (i < j), \tag{A2.19}$$

where $p_j = \psi_j'(\nu)$. This is positive, since both factors in the integrand are negative in the ν-interval indicated.

Evaluation of ω_{ij} from (A2.19) is much easier than from the left-hand expression in (A2.12), but an easier way still is to use the reduction from the ν-formulation to the α-formulation employed in Section 14.2: essentially a way of normalising out the effects of immigration. One finds that

$$\omega_{ij} = \int_{\alpha_j}^{\alpha_i} (L_j^* p^* - 1) p_i^* \frac{\mathrm{d}\alpha}{\mathrm{d}\nu} \mathrm{d}\alpha \quad (i < j). \tag{A2.20}$$

Here $p_j^* = \partial\psi_j(\alpha)/\partial\alpha$, where $\psi(\alpha)$ is determined by the equation system

$$\max[\sum_k \mu_{jk}(\Delta_k - \Delta_j) - \alpha, -\psi_j(\alpha)] = 0,$$

which we write as

$$\max[L_j^* \Delta - \alpha, -\psi_j(\alpha)] = 0.$$

The term $\mathrm{d}\alpha/\mathrm{d}\nu$ in (A2.20) is the reciprocal of the expression for $\mathrm{d}\nu/\mathrm{d}\alpha$ given in (14.13).

Appendix 3

Random graphs and polymer models

The size of this appendix may seem excessive, in view of the limited mention of random graphs in the main body of the text. However, given that a review as meticulous and wide-ranging as that by Albert and Barabási (2002) regards certain central questions as open for which a full analysis exists in the literature, it seems appropriate to give a brief and self-contained account of that analysis. It cannot be long concealed that much of the analysis referred to is that developed by the author over some 30 years. Detailed references are given in the literature survey of Section A3.8; the principal advances are treatment of the so-called first-shell model, leading to the integral represention (A3.13), (A3.31) and (A3.35) of various generating functions, in which a raft of conclusions is implicit.

The random graph model can be expressed either in terms of graphs or in terms of polymers if we identify nodes with atoms, arcs with bonds, components (connected subgraphs) with polymer molecules and node 'colours' with types of atom. We shall for preference use the term 'unit' rather than atom, since the unit does occasionally have a compound (but fixed) form. The situation in which a 'giant' component forms (i.e. in which all but an infinitesimal proportion of the nodes form a single component) is identified with the phenomenon of gelation in the polymer context.

The advantage of the more physical view is that it brings in aspects rarely considered in the graph context, although very relevant to it: the weighting of distributions reflecting the binding energy of a configuration, the idea of an open process in which molecules (components) may enter and leave a defined region of space. As in graph theory, a configuration is prescribed in terms purely of connections rather than of physical position, although spatial relationships emerge in an embryonic form when one considers configurations of N units distributed in a region of volume V, or even the joint distributions of configurations in several such regions.

A3.1 The zeroth-order model

Suppose that N indistinguishable units are distributed in a physical region of volume V. The configuration $\mathcal{C}$ of the assembly is determined by the numbers of bonds s_{ab} between units a and b $(a, b = 1, 2, \ldots, N)$. In the zeroth-order model the probability of this configuration is assumed to be

$$P(\mathcal{C}) \propto Q_{NV}(\mathcal{C}) = \prod_a \prod_b \frac{(h/2)^{s_{ab}}}{s_{ab}!}, \tag{A3.1}$$

where $h = 1/(\kappa V)$. The assumption is then that the numbers of bonds between unit pairs (in a specified direction) are independent Poisson variables with expectation $1/(2\kappa V)$. This can be seen as the equilibrum distribution of a model in which bonds form from one prescribed unit a to another b with probability intensity $h = 1/(2\kappa V)$, and break independently with unit intensity.

Let us suppose that N and V are large, but with the ratio N/V held as near as possible to a prescribed value of 'density' ρ. The passage $V \to \infty$ is then spoken of as the 'thermodynamic limit'. The probability that a given unit forms a multiple bond with some other unit or forms a self-bond is of order V^{-1}, and so negligible for large V. This model is then effectively the Erdös–Rényi model, described in Section 18.2. However, the assumption that the number of bonds has a Poisson distribution rather than a (0, 1) distribution is mathematically more natural, and possibly even physically so.

In the next section we shall deduce something of the statistical character of units and polymers for a more general model, but we can make a simple calculation now which proves enlightening. Define the quantity

$$S_{NV}(\mathcal{C}) = \frac{V^N}{N!} Q_{NV}(\mathcal{C}). \tag{A3.2}$$

The factor V^N reflects an integration over the possible positions of the N units; the factor $(N!)^{-1}$ reflects a summation over permutations of these statistically equivalent units. Consider now the distribution of N units over two disjoint regions of volumes V_1 and V_2, with $V_1 + V_2 = V$ and $N/V \approx \rho$. The total configuration $\mathcal{C}$ is the combination $\{\mathcal{C}_1, \mathcal{C}_2\}$ of the configurations in the two compartments. We shall then suppose that

$$S_{N,V_1+V_2}(\mathcal{C}) = S_{N_1,V_1}(\mathcal{C}_1) S_{N_2,V_2}(\mathcal{C}_2) \quad (N_1 + N_2 = N). \tag{A3.3}$$

The fact that we write $\mathcal{C}$ as $(\mathcal{C}_1, \mathcal{C}_2)$ implies that we allow no bonding between different compartments; $\mathcal{C}_1$ and $\mathcal{C}_2$ are separated graphs. The fact that we allow N_1 and N_2 to vary subject to $N_1 + N_2 = N$ indicates that we are allowing free migration between the two compartments. The migrations must be of complete polymers, however, since only within-compartment bonding is permitted. If the terms in the right-hand member were configuration-independent for given values of the N_i then relation (A3.3) would indicate a binomial distribution of units between the two compartments; relation (A3.3) as it stands generalises this appropriately. Model (A3.3) is an extension of model (A3.1), with the most modest intimation of spatial structure. It turns out to be consistent, however, in that the determination of the gelation point we shall now derive agrees with that determined directly in Sections A3.4 and A3.5.

Let us define

$$Q_{NV} = \sum_{\mathcal{C}} Q_{NV}(\mathcal{C}), \tag{A3.4}$$

where the sum is over all configurations of N units. For the particular case (A3.1) we have

$$Q_{NV} = \mathrm{e}^{N^2/2\kappa V}. \tag{A3.5}$$

Consider the particular case of a distribution of $2N$ units over two communicating regions each of volume V. The probability that the $2N$ units distribute themselves in a given way between the two compartments, regardless of configuration, is then

$$P(N_1, N_2) \propto \frac{Q_{N_1,V} Q_{N_2,V}}{N_1! N_2!} \quad (N_1 + N_2 = 2N), \tag{A3.6}$$

where $P(\cdot)$ is used to represent the equilibrium distribution of whichever quantity is indicated. We see now that the occurrence of the factorials indeed comes about by a summation over permutations of the individual nodes for given values of the N_i. Let us write (N_1, N_2) as $(N+n, N-n)$, so that n represents the imbalance in distribution between the two compartments. By (A3.5), (A3.6) this has the distribution

$$P(n) \propto \frac{\exp(n^2/\kappa V)}{(N+n)!(N-n)!} \quad (-N \le n \le N). \tag{A3.7}$$

One finds by direct arguments that, for $\rho < \kappa$, expression (A3.6), symmetric in n, has a single maximum, located at $n = 0$. For $\rho > \kappa$ it has two maxima, displaced from $n = 0$ by the same amount each way.

The interpretation is that $\rho_g = \kappa$ is a critical value of density. Below this density there is statistical equidistribution of matter between the two compartments, but above it matter tends to 'lump' in one compartment or the other. This lumping becomes total in the thermodynamic limit, in that all but a vanishing fraction concentrates in one of the compartments.

One construes this 'lumping' as the gel state: matter is in a single complex that can only be in one compartment or the other. The value ρ_g is the critical density of gelation.

A3.2 The first-shell model: unit statistics

Model (A3.1) can be generalised to

$$P(\mathcal{C}) \propto Q_{NV}(\mathcal{C}) = \prod_a \prod_b \frac{(h/2)^{s_{ab}}}{s_{ab}!} \Phi(s), \tag{A3.8}$$

where $\Phi(s)$ is a function of all the bonds and so of configuration. We think of $\Phi(s)$ as a weighting of configurations reflecting their different binding energies. The factor before it represents the essential combinatorial term of the full description (that in which nodes are identified and given a notional 'position'). A natural first such generalisation of the zeroth-order model (A3.1) is

$$P(\mathcal{C}) \propto Q_{NV}(\mathcal{C}) = \prod_a \prod_b \frac{(h/2)^{s_{ab}}}{s_{ab}!} \prod_k H_k^{n_k}. \tag{A3.9}$$

Here n_k is the number of nodes of degree k, a function of s in that the degree of node a is $\sum_b (s_{ab} + s_{ba})$ if we take no account of bond direction. The distribution of degree for a randomly chosen node is of interest in itself, as we have seen in Chapter 18. Note that

$$N = \sum_k n_k, \quad B = (1/2) \sum_k k n_k, \tag{A3.10}$$

where B is total number of bonds (arcs). The quantity H_k expresses the intrinsic attractiveness of the k-degree state of a unit; it will usually be an exponential function of the binding energy of that state. It is a parameter of the model but, as is common in statistical–mechanical contexts, it also serves as marker variable for the random variable n_k. More specifically, regard the sum Q_{NV} defined in (A3.4) as a function $Q_{NV}[H] = Q_{NV}[\{H_k\}]$ of the parameters H_k. Then $P(\{n_k\})$ is proportional to the term in $\prod_k H_k^{n_k}$ in the expansion of $Q_{NV}[H]$. Otherwise expressed

$$\mathrm{E}\left[\prod_k z_k^{n_k}\right] = \frac{Q_{NV}[\{H_k z_k\}]}{Q_{NV}[\{H_k\}]}, \tag{A3.11}$$

so that, for example,

$$\mathrm{E}(n_k) = H_k \frac{\partial \log Q_{NV}}{\partial H_k}. \tag{A3.12}$$

Theorem A3.1 *Assume the form (A3.9) for the unnormalised equilibrium distribution $Q_{NV}(\mathcal{C})$. Then the sum Q_{NV} has the evaluation*

$$Q_{NV} = \sqrt{\frac{\kappa V}{2\pi}} \int_{-\infty}^{+\infty} H(\xi)^N \mathrm{e}^{-\kappa V \xi^2/2} \mathrm{d}\xi, \tag{A3.13}$$

where

$$H(\xi) = \sum_{k=0}^{\infty} \frac{H_k \xi^k}{k!}. \tag{A3.14}$$

Proof Suppose that

$$H_k = \int x^k m(\mathrm{d}x) \tag{A3.15}$$

for some signed measure m on the real line. Then

$$\begin{aligned} Q_{NV} &= \int \cdots \int \sum_s \left[\prod_a \prod_b (h x_a x_b/2)^{s_{ab}}/s_{ab}!\right] \prod_a m(\mathrm{d}x_a) \\ &= \int \cdots \int \exp\left[(h/2)\left(\sum_a x_a\right)^2\right] \prod_a m(\mathrm{d}x_a). \end{aligned} \tag{A3.16}$$

Appeal now to the identity

$$\exp(h\sigma^2/2) = \frac{1}{\sqrt{2\pi h}} \int_{-\infty}^{\infty} \exp[\sigma\xi - \xi^2/2h] \mathrm{d}\xi. \tag{A3.17}$$

Setting $\sigma = \sum_a x_a$ in (A3.17) and substituting from (A3.17) into (A3.16) we deduce relation (A3.13) with

$$H(\xi) = \int \mathrm{e}^{x\xi} m(\mathrm{d}x) = \sum_0^{\infty} \frac{H_k \xi^k}{k!}. \ \Diamond$$

Relation (A3.13) expresses an identity under all circumstances, in that if we expand both sides in powers of the H_k then the evaluation of coefficients on the two sides agrees. However, a representation (A3.15) may not be valid and the integral of (A3.13) is certainly divergent if H_k increases sufficiently rapidly with k. A sufficient condition would be that $\ln H(\xi)$ grows more slowly than $|\xi|^2$ with increasing $|\xi|$, which will be so if H_k grows no faster than $k^{\alpha k}$ for some $\alpha < 1/2$.

Corollary A3.2 *The degree abundances n_k have the joint distribution*

$$P(\{n_k\}) \propto \frac{(2B)!}{B!(2\kappa B)^B} \prod_k \frac{(H_k/k!)^{n_k}}{n_k!} \quad (\sum_k n_k = N). \tag{A3.18}$$

This follows by evaluation of the term in $\prod_k H_k^{n_k}$ in the expansion of expression (A3.13) in such powers. Note that expression (A3.18) may not be summable if the H_k grow too rapidly with k. Even though the n_k are all finite for prescribed N, the distribution may assign positive weight to infinite k-values if nodes of high degree are favoured excessively.

Define now $\mathcal{H}_1$ as the class of generating functions $H(\xi)$ for which $\ln H(\xi)$ is of less than quadratic growth at infinity, so that the function $J(\xi)$ defined in equation (A3.19) below has a unique maximum, attained at a finite positive value of ξ.

Corollary A3.3 *Suppose $H \in \mathcal{H}_1$, and let $\overline{\xi}$ be the value of ξ maximising*

$$J(\xi) = \rho \ln H(\xi) - \kappa\xi^2/2. \tag{A3.19}$$

Then the probability generating function of the degree of a randomly chosen unit is

$$E(z^k) = \frac{H(\overline{\xi}z)}{H(\overline{\xi})} \tag{A3.20}$$

in the thermodynamic limit.

Proof It follows from (A3.12), (A3.13) that

$$\begin{aligned} \mathrm{E}(n_k/N) \propto \int \frac{H_k\xi^k}{k!} H(\xi)^{N-1} \mathrm{e}^{-\kappa V\xi^2/2} \mathrm{d}\xi \sim \int \frac{H_k\xi^k}{k!} \exp[VJ(\xi)] \mathrm{d}\xi \\ \sim \text{const.} \frac{H_k\overline{\xi}^k}{k!}, \end{aligned} \tag{A3.21}$$

whence the assertion follows. ◊

Relation (A3.20) is obviously relevant to the discussion of an observed degree distribution in Chapter 18. We shall collect some observations in Section A3.6

Now that we have relation (A3.13), essentially just

$$Q_{NV} \propto \int \mathrm{e}^{VJ(\xi)} \mathrm{d}\xi, \tag{A3.22}$$

we can reproduce the argument of the last section for the location of the gelation point. It leads to the conclusion that gelation occurs as soon as the inequality

$$\rho \frac{H''(\bar{\xi})}{H(\bar{\xi})} < \kappa \tag{A3.23}$$

is transgressed. Here $\bar{\xi}$ is again the value maximising $J(\xi)$, and so is determined by

$$\rho \frac{H'(\xi)}{H(\xi)} = \kappa \xi. \tag{A3.24}$$

This condition is given a more explicit characterisation in Section A3.4. In fact, as pointed out in Whittle (1994), there is a much more general and immediate characterisation of the gelation point. Suppose that we can write

$$\frac{V^N}{N!} Q_{NV} = \exp[VM(\rho) + o(V)]$$

for large V and some function M. Then as ρ increases $M(\rho)$ will be concave initially (reflecting entropy effects) and then become convex (reflecting energy effects), and the transition point is just the gelation point. These statements hold for the first-shell model with H in $\mathcal{H}_1$, but also much more generally.

A3.3 Polymer statistics

Recall that polymer molecules are to be identified with the components of the graph. These can be classified in varying levels of detail: perhaps purely by size, perhaps by numbers of nodes of the different degrees, perhaps by a much more detailed topological characterisation. There is a classical lemma (see e.g. Percus, 1971 p. 52) which gives a key purchase on polymer statistics.

Lemma A3.4 *Suppose that to each graph $\mathcal{C}$ on an arbitrary number of labelled nodes can be attached a weight $W(\mathcal{C})$ with the properties (i) that $W(\mathcal{C})$ is invariant under permutation of the nodes, and (ii) that $W(\mathcal{C}) = W(\mathcal{C}_1)W(\mathcal{C}_2)$ if $\mathcal{C}$ can be decomposed into mutually unconnected graphs $\mathcal{C}_1$, $\mathcal{C}_2$. Then*

$$\sum_{\mathcal{C}} W(\mathcal{C})/N(\mathcal{C})! = exp[\sum_r W_r], \tag{A3.25}$$

where the sum over $\mathcal{C}$ covers all graphs, $N(\mathcal{C})$ is the number of nodes in $\mathcal{C}$, the term for the empty graph is taken as unity and W_r is the W-value for the rth distinct connected graph (an 'r-mer').

Proof It follows from (ii) that

$$\sum_{\mathcal{C}}^{(N)} W(\mathcal{C}) = \sum_{\mathcal{C}}^{(N)} \prod_r W_r^{m_r(\mathcal{C})} = N! \sum_m^{(N)} \prod_r (W_r^{m_r}/m_r!), \tag{A3.26}$$

where $m_r(\mathcal{C})$ is the number of r-components (r-mers) in configuration $\mathcal{C}$ and the (N) over the summation sign denotes a sum consistent with a total of N nodes. The final equality in (A3.26) follows from a summation over all permutations of nodes consistent with $m_r(\mathcal{C}) = m_r$ for all r. Assertion (A3.25) then follows from (A3.26). $\diamondsuit$

The lemma implies an immediate statistical characterisation of polymer statistics.

Theorem A3.5 *Suppose that*

$$P_N(\mathcal{C}) \propto Q_N(\mathcal{C}) = \prod_r \gamma_r^{m_r(\mathcal{C})}. \tag{A3.27}$$

Then the probability generating function $\Pi_N(z) = \Pi(\{z_r\})$ *of the numbers of r-mers in an assembly of N units is proportional to the coefficient of* w^N *in the expansion in powers of w of the expression*

$$\sum_N \frac{(wV)^N}{N!} \sum_{\mathcal{C}}^{(N)} \prod_r (\gamma_r z_r)^{m_r(\mathcal{C})} = exp[\sum_r \gamma_r z_r (wV)^R]. \tag{A3.28}$$

The assertion follows immediately from the lemma with the choice $W_r = \gamma_r z_r (wV)^R$, where R is the number of units in an r-mer. The interpretation is that in an 'open' process with unconstrained N the numbers N_r of polymers of the recognised types are independent Poisson variables with expectations $\gamma_r (wV)^R$. The constraint of prescribed N is applied by confining the distribution to those terms for which w occurs to power N and renormalising. If one coarsens one's description of a polymer then all polymers of a given class in the reduced description will presumably have the same size, and one deduces the new γ value simply by summing γ_r over all r falling in that class.

The distribution will indeed take the form (A3.27) if one assumes that the binding energy is a property of the polymer, and that there is no energy interaction between polymers. However, even so, the γ_r will contain combinatorial factors whose determination is one of the main problems. In the case of the first-shell model we have the relation (A3.13), which goes part of the way to solving the problem, at least if the integral (A3.13) is convergent. In this case the natural level of specification of an r-mer is by $r = (r_0, r_1, r_2, \dots)$, where r_k is the number of units of degree k in the molecule. Define the function

$$\Gamma(H, w) = \sum_r \gamma_r (wV)^R. \tag{A3.29}$$

The quantities H_k themselves serve as marker variables for the random variables N_k; in fact we see from (A3.9) that the expectation $\mathrm{E}(N_r)$ will have the form

$$\gamma_r V^R = \omega_r V^{R-L_r} (2\kappa)^{-L_r} \prod_k H_k^{r_k}. \tag{A3.30}$$

Here ω is a purely combinatorial term, $R = \sum_k r_k$ is the number of units in an r-mer and $L_r = (1/2)\sum_r k r_k$ is the number of bonds. The determination of ω_r is latent in relation (A3.31) below.

Theorem A3.6 *Consider the first-shell model (A3.9) and suppose that $H \in \mathcal{H}_1$. Then the generating function $\Gamma(H, w)$ has the evaluation*

$$\exp[\Gamma(H, w)] = \sqrt{\frac{\kappa V}{2\pi}} \int_{-\infty}^{\infty} \exp[V(wH(\xi) - \kappa\xi^2/2)]\mathrm{d}\xi. \tag{A3.31}$$

The formal assertion follows immediately by substitution of the expression (A3.13) for Q_{NV} in the sum $\sum_N (wV)^N Q_{NV}/N!$. However, whereas the integral of (A3.13) will be convergent under the conditions stated, the integral of (A3.31) cannot be if H_k is nonzero for some $k > 2$, i.e. if branching is possible in the polymers. Nevertheless, relation (A3.31) correctly defines $\Gamma(H, w)$ as a power series, term by term, and this, by Theorem A3.5, determines the polymer distribution. It must then be that relation (A3.31) correctly equates two divergent power series. To take the simplest example, consider the zero-order model, for which Q_{NV} is given by (A3.5), and so (A3.31) becomes

$$\exp[\Gamma(H, w)] = \sum_N \frac{(wV)^N}{N!} \mathrm{e}^{N^2/(2\kappa V)}. \tag{A3.32}$$

The series on the right is certainly divergent, and so then is the consequent determination of the series for $\Gamma(H, w)$. One can say that the right-hand member diverges because the summation over N for fixed V takes one into supercritical cases. One can say that $\Gamma(H, w)$ diverges as a power series in w because the number of possible polymer structures increases so rapidly with R, the size of the molecule.

Nevertheless, these formal relations do express structural relations, and sometimes transparently so. For an example, suppose the first-shell model (A3.9) modified to

$$P(\mathcal{C}) \propto Q_{NV}(\mathcal{C}) = \prod_a \prod_b \frac{(h/2)^{s_{ab}}}{s_{ab}!} \prod_k H_k^{N_k} \nu^{B+C-N}, \tag{A3.33}$$

where B and C are respectively the number of bonds (arcs) and the number of polymers (components) in an assembly of N units. Now, a polymer of n units must contain at least $n-1$ bonds. It cannot be connected if it contains fewer, and if it contains exactly this number it must be a tree. Adding this inequality over all polymers we see that $B \geq N - C$, with equality if and only if all the polymers are trees. The quantity $B + C - N$ thus represents the number of bonds over and above what is needed to hold the polymer molecules together: bonds that contribute to cycles of one kind or another. The larger the parameter ν the more such excess bonding is encouraged, and in fact the parameter occurs if in the dynamic model one increases all bond-formation rates by a factor of ν if they add a new internal bond to an existing polymer.

We see from (A3.33) that the effect of the new factor is to modify relation (A3.30) to

$$\gamma_r V^R = \omega_r V(\nu/V)^{L_r+1-R}(2\kappa)^{-L_r} \prod_k H_k^{r_k}. \tag{A3.34}$$

We thus deduce

Corollary A3.7 *The expected abundance (A3.34) is, in its dependence upon V, proportional to V if $L_r + 1 - R = 0$ (i.e. if the r-mer is a tree). For fixed ν it is of smaller order in V for all other polymers.*

In fact, the order in V becomes ever smaller as the r-mers develop more cycles. However, there are so many of these nontree r-mers that there is a cumulative effect nevertheless.

We see from (A3.34) that the effect of the introduction of the parameter ν is to replace V by V/ν and to multiply γ by ν. This implies the modification of relation (A3.31) to

$$\exp[\Gamma(H, w)] = \left[\sqrt{\frac{\kappa V}{2\pi\nu}} \int_{-\infty}^{\infty} \exp[(V/\nu)(wH(\xi) - \kappa\xi^2/2)]\mathrm{d}\xi\right]^{\nu}. \tag{A3.35}$$

If ν is a positive integer then this modification has a clear implication. Instead of considering an assembly of units in a single region of volume V one is effectively considering ν replica regions each of volume V/ν between which migration of polymers is free. The 'compartmental' model that gave us our first evidence of possible gelation thus seems inherent in the model. This generalisation is both realistic and interesting; we shall pursue some of its implications in the following sections.

A3.4 Nothing but trees

The case $\nu = 0$, when only tree molecules are permitted, is interesting. In this case all the molecules with a degree of internal cyclisation, whose infinite variety caused the essential divergences of expressions such as (A3.31), (A3.35) and possibly (A3.13), have been stripped out. Also, this is the case that was found historically easier, for reasons we shall see. It follows from (A3.34) that in this case we can write

$$\Gamma(H, w) = V \sum_{r\in\mathcal{T}} \gamma_r w^R = VG(w), \tag{A3.36}$$

say, where $\mathcal{T}$ is the set of possible tree molecules. The generating function $G(w)$ is then independent of V; its continuing dependence upon the parameters H_k is taken for granted and not indicated in the notation. Whittle (1965a, 1986) derived the evaluation

$$G(w) = \max_{\xi}[wH(\xi) - \kappa\xi^2/2] = \max_{\xi} F(w, \xi), \tag{A3.37}$$

say. The form F will in general show a first maximum in ξ and then a minimum, before a growth to infinity, and it is this first maximum that the 'max' operator is intended to locate. Relation (A3.37) looks like some kind of limit deduction from the identity (A3.35). However, the impropriety of the integral in (A3.35) puts any such direct calculation out of the question, and a surprisingly circuitous and sophisticated argument is needed. For interest, we indicate its essentials.

Denote expression (A3.35) by $M(H, V)$ and define the degree-multiplying and bond-creating operators

$$T = \sum_k kH_k \frac{\partial}{\partial H_k},$$

$$U = \sum_k H_{k+1} \frac{\partial}{\partial H_k}.$$

Then it can be verified that M obeys the equation

$$\kappa VTM = U^2 M. \tag{A3.38}$$

For the case of general ν we know from (A3.35) that

$$\mathrm{e}^{\Gamma} = M(H, V/\nu)^{\nu}, \tag{A3.39}$$

and hence that

$$\kappa V\, T\Gamma = (U\Gamma)^2 + \nu U^2 \Gamma. \tag{A3.40}$$

Setting $\Gamma = VG$ and going to the limit of zero ν we find from (A3.40) that

$$\kappa TG = (UG)^2. \tag{A3.41}$$

One can verify that (A3.41) has expression (A3.37) as solution.

All these calculations are of course formal in high degree, since it is first when we come to $G(w)$ that we have a generating function that is possibly convergent. However, all the above operations are legitimate in their implied manipulation of the coefficients of power series, and it is by appeal to the fact that we require the power series expansion of $G(w)$ to begin with the term wH_0 that we fix on the particular solution (A3.37).

We shall now have the evaluation

$$\begin{aligned} V^N Q_{NV}/N! &\propto \text{coefficient of } w^N \text{ in } \mathrm{e}^{VG(w)} \\ &\sim \min_w \exp\{\max_\xi [F(w,\xi]\} w^{-N} \propto \max_\xi H(\xi)^N \mathrm{e}^{-\kappa V \xi^2/2}, \end{aligned} \tag{A3.42}$$

where the evaluation in the second line follows from the steepest-descent evaluation of a contour integral. The minimisation with respect to w implies then the stationarity condition

$$\rho = w\frac{\partial G(w)}{\partial w} = wH(\overline{\xi}),$$

where $\overline{\xi}$ is the locally maximising value in the final expression of (A3.42). It follows from relation (A3.42) that assertions (A3.20), (A3.21) for the distribution of degree hold also in this case. Let the value of w determined by the last equation be denoted $\overline{w}$.

Theorem A3.8 *(i) The values $\overline{w}$ and $\overline{\xi}$ are determined by the pair of equations*

$$wH' = \kappa\xi, \quad wH = \rho. \tag{A3.43}$$

(ii) Suppose that if $H_1 > 0$ (so that single units are viable) and $H_k > 0$ for some $k > 2$ (so that branching is possible). Then the inequality

$$\rho H'' < \kappa H, \tag{A3.44}$$

equivalent to

$$\xi H'' < H', \tag{A3.45}$$

holds at $(\overline{w}, \overline{\xi})$ *for small enough* ρ. *The smallest value of* ρ *for which it first ceases to do so marks the gelation point.*

Proof Relations (A3.43) follow from the stationarity conditions in (A3.42). Under the conditions stated $H'(\xi)$ is increasing from a positive value at $\xi = 0$ and is strictly convex. The ξ-roots of the first equation of (A3.43) for a given value of w are then the crossing points in the graph of Fig A3.1. For sufficiently small w there will be just two of these, the smaller and the larger corresponding respectively to a local maximum and a local minimum of the form F of (A3.37). The smaller root is the relevant one, and the direction of crossing implies that $wH'' - \kappa$ is negative at this point, which is equivalent to the inequalities (A3.44) and (A3.45).

Plainly some kind of critical behaviour occurs when w becomes so large that the form F no longer possesses a local maximum in ξ. This occurs when the line in the figure just touches the curve as illustrated; at this point equality holds in (A3.44), (A3.45). We shall demonstrate shortly that this critical point is to be identified with the gel point.

The parameter w is important, as follows from the relation $\mathrm{E}(N_r) \approx V\gamma_r \overline{w}^R$. We have now to relate the value of w to that of ρ (the overlinings being understood). With some calculation we find from equations (A3.43) that the rate of change $w_\rho = \partial w/\partial \rho$ is determined by

$$[\xi(H')^2 + (H' - \xi H'')H]w_\rho = H' - \xi H''. \tag{A3.46}$$

From this it follows that w decreases as ρ increases from zero until the gel point is reached, and then decreases as ρ continues to increase further by some positive amount. This confirms then that inequalities (A3.44), (A3.45) hold for small enough ρ and then reverse as ρ increases through the gel point. Criticality of some kind will be reached

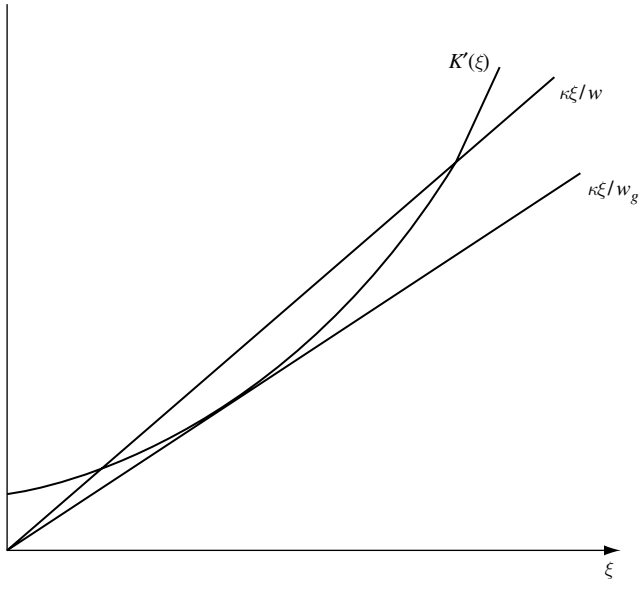


Fig. A3.1 An illustration of the solution of equations (A3.43) and the critical case.

when the local maximum of the bracket fails to exist. This occurs when a maximum and a minimum coalesce, and the second derivative becomes zero. That is, when the inequality

$$wH'' - \kappa < 0 \tag{A3.47}$$

is first violated. Substituting for w from either of the relations (A3.43) we obtain the equivalent characterisations (A3.44) and (A3.45). ♢

The following assertion identifies the critical point determined above as the gel point.

Theorem A3.9 *In the subcritical regime, the expected size of the molecule within which a randomly chosen unit finds itself is*

$$\mathrm{E}(R) = 1 + \frac{w(H')^2}{H(\kappa - wH'')} = \frac{\rho W(H')^2}{H(\kappa H - \rho H'')}, \tag{A3.48}$$

where all evaluations are at $\overline{w}$, $\overline{\xi}$.

Proof The probability that a randomly chosen molecule is an r-mer of size R is proportional to $\gamma_r \overline{w}^R$, so the probability that the molecule containing a randomly chosen unit is such an r-mer is proportional to $R\gamma_r \overline{w}^R$. Hence

$$\mathrm{E}(R) = \frac{\sum_r R^2 \gamma_r \overline{w}^R}{\sum_r R\gamma_r \overline{w}^R} = 1 + \frac{w^2 G_{ww}}{wG_w}, \tag{A3.49}$$

where the subscripts indicate differentials with respect to w, evaluated at $\overline{w}$. Now, it follows from (A3.37) by standard arguments that

$$G_w = F_w, \quad G_{ww} = F_{ww} - F_{w\xi}^2 / F_{\xi\xi}. \tag{A3.50}$$

Substituting the evaluations of these expressions into (A3.49) we verify relation (A3.48). ♢

One sees then that $\mathrm{E}(R)$ approaches infinity as ρ approaches the critical value ρ_g from below.

It would appear from (A3.44), (A3.45) that the critical value of density is the same in the cases $\nu = 0$ and $\nu = 1$. In fact, there are heavy provisos, and this is a point we shall return to in Section A3.5. A further elegant fact which emerges is that, if ρ_- and ρ_+ are the lesser and greater values of ρ corresponding to a given value of w, then the statistics of the system at the subcritical density ρ_- are just the statistics of the sol-fraction at the supercritical density ρ_+.

If the coefficient of w_ρ in (A3.46) is positive for all positive ξ then w continues to decrease with ρ as ρ increases indefinitely past the gel point, which points to some kind of regularity. This condition on the coefficient amounts to

$$L' - \xi L'' > 0, \tag{A3.51}$$

where $L(\xi) = \ln H(\xi)$. A double integration of (A3.51) yields

$$\ln H(\xi) < a + b\xi^2/2 \tag{A3.52}$$

for some constants a and b. But this is almost equivalent to the condition required before: that $\ln H(\xi)$ be of less than quadratic growth at infinity, and so $H \in \mathcal{H}_1$. As an example, for the zeroth-order model we have $H(\xi) = \mathrm{e}^{\xi}$, and deduce easily that $\overline{w}$ shows the dependence upon ρ

$$\overline{w} = \rho \mathrm{e}^{-\rho/\kappa}, \tag{A3.53}$$

graphed in Fig. A3.2. This increases with ρ up to the gel point at $\rho_g = \kappa$ and decreases thereafter.

As an example of a case for which the condition is not satisfied we can assume that $H(\xi)$ has a finite radius of convergence: the choice

$$H(\xi) = (1-\xi)^{-1} \tag{A3.54}$$

is the simplest. Conditions (A3.43) then yield the relations

$$\rho = \kappa\xi(1-\xi), \quad w = \kappa\xi(1-\xi)^2, \tag{A3.55}$$

graphed in Fig. A3.3; these determine $w(\rho)$ parametrically. The quantities ρ and w are then maximal at ξ equal to 1/2 and 1/3 respectively. The relation $w(\rho)$ determined by (A3.54) is graphed in Fig. A3.4. The function rises to its maximum of $w_g = 4\kappa/27$ at criticality, when $\rho = \rho_g = 2\kappa/9$, but then moves on to a lower branch of the function at $\rho = \kappa/4$. This is the maximal value of density that the model admits, seeming to indicate a second critical threshold. We return to the point in the next section.

For a fuller understanding one should determine the nature of the function $G(w)$ and its implications for criticality. Consider again the zeroth-order case $H(\xi) = \mathrm{e}^{\xi}$, for which the gel point is located at $\rho_g = \kappa$ in both the cases ν equal to 0 and 1.

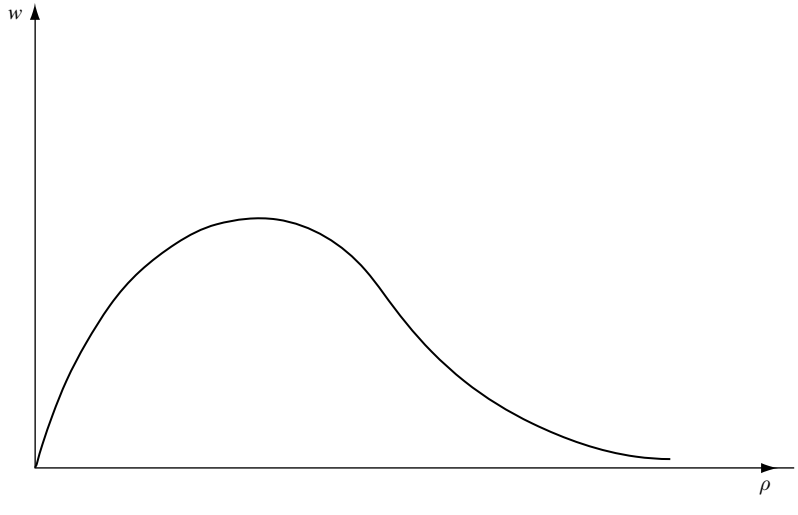


Fig. A3.2 A graph of the dependence (A3.53) of the activity parameter $\overline{w}$ upon the spatial density of nodes ρ.

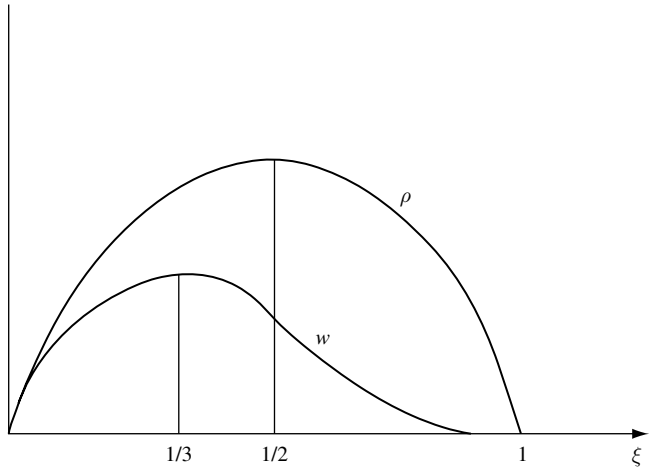


Fig. A3.3 The graphs of ρ and w against ξ in the case (A3.54) when $H(\xi)$ has a finite radius of convergence. Values of these variables outside a certain range simply cannot occur – once w reaches the branchpoint the parameters thenceforth describe the sol-fraction rather than the total.

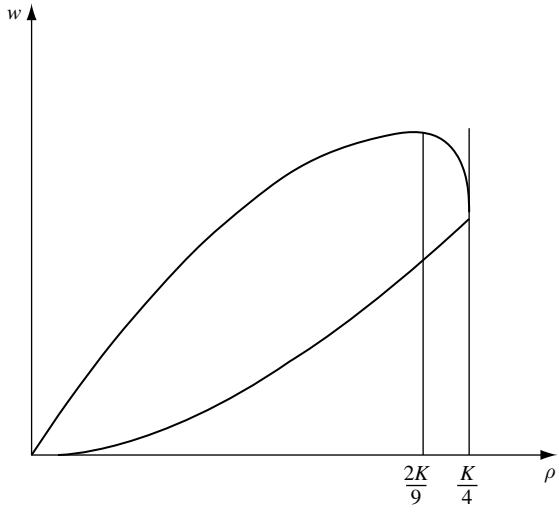


Fig. A3.4 The relation between ρ and w implied by relations (A3.55).

Theorem A3.10 *Consider the zeroth-order model in the case $\nu = 0$. The natural level of description of a molecule with this parameterisation is simply its size R, and the coefficient γ_R of w^R in the expansion of $G(w)$ is*

$$\gamma_R = \frac{R^{R-4}}{\kappa^{R-1} R!}, \tag{A3.56}$$

for $R > 2$.

Proof The determination of the coefficient follows from the expression

$$G(w) = \frac{1}{2\pi i} \oint \frac{F(\xi, w)(\kappa - wH'')}{\kappa\xi - wH'} \mathrm{d}\xi \tag{A3.57}$$

for the implicitly determined function. Here the contour of integration is a circle in the complex plane around the value of ξ making the denominator of the integrand zero; it is the evaluation of the residue at this sole singularity that enforces the stationarity condition defining G. If one expands the integrand in powers of w the singularity shifts to $\xi = 0$, at which there are multiple poles and the evaluation becomes

$$\gamma_R = \text{the coefficient of } w^N \xi^0 \text{ in } F(\xi, w) \left[1 - \frac{w\mathrm{e}^{\xi}}{\kappa}\right] \sum_{s=0}^{\infty} \left[\frac{w\mathrm{e}^{\xi}}{\kappa\xi}\right]^s. \tag{A3.58}$$

With some reduction, this yields the evaluation (A3.56); the evaluation is modified for R equal to 1 or 2. $\Diamond$

We see from (A3.56) that γ_R is of order $R^{-9/2}(\mathrm{e}/\kappa)^R$ for large R, which agrees exactly with the fact that $G(w)$ has a singularity at $w = w_g = \kappa/\mathrm{e}$. This critical value is a branch point of the function G rather than a pole of integral order. Note that G and its first three derivatives exist right up to the branch point. Appealing to (A3.53) we find that the density of R-mers for a unit density of ρ is

$$\gamma_r \overline{w}^R = \kappa \frac{R^{R-4}}{R!} \left(\frac{\rho}{\kappa} \mathrm{e}^{-\rho/\kappa}\right)^R. \tag{A3.59}$$

As ρ increases from zero then $\overline{w}$ also increases until it reaches its maximum value at the critical point $\rho = \kappa$. At this point the sum diverges; $\overline{w}$ has reached the value κ/e that is a branch point of the function $G(w)$. As ρ increases further then $\overline{w}$ decreases, and expression (A3.59) describes the statistics of those units that are not incorporated in the gel fraction: the sol fraction.

If we perform the same calculation for the case

$$H(\xi) = (1 - \xi)^{-\alpha} \tag{A3.60}$$

we find that

$$\gamma_R \propto \frac{(\alpha R + 2R - 3)!}{(\alpha R + R)!(R - 1)!} \left(\frac{\alpha}{\kappa}\right)^R \sim R^{-5/2} \left(\frac{\alpha(\alpha + 2)^{\alpha+2}}{\kappa(\alpha + 1)^{\alpha+1}}\right)^R, \tag{A3.61}$$

where $x!$ for nonintegral x is taken as the Gamma function $\Gamma(x + 1)$. The less regular character of H in this case is reflected in the fact that the second derivative of G does not exist at the branch point. Evaluation (A3.61) agrees with the radius of convergence $4\kappa/27$ deduced previously for the case $\alpha = 1$.

A3.5 The branching analogue; Potts criticality

We have seen that for models for which $H(\xi)$ falls in the class $\mathcal{H}_1$ the gel points are the same whether ν takes the value 0 or 1. However, if H is in the class $\mathcal{H}_2$, having finite radius of convergence, then one can determine a gel point in the first case, as we saw from Theorem A3.8, but not in the second, because of the divergence of the integral (A3.13). The explanation is that there is another kind of criticality, Potts criticality, which sets in at a ν-dependent value $\rho_P(\nu)$ of density. At the gel point a unit suddenly has high probability of being connected to most other units at some remove. This is a continuous transition, in that the unit would notice little difference in the pattern of immediate connections it makes as density increases through the gel point. At the Potts transition there is suddenly a rich system of intramolecular connections, so rich as to be almost irreversible. The transition point $\rho_P(\nu)$ is a decreasing function of ν, and for large enough ν will be smaller than ρ_g, so that as density is increased from zero Potts criticality sets in before gelation. Let ν_P denote the value of ν at which the two transition points coincide. For H in $\mathcal{H}_1$ one knows that ν_P is greater than one, and it is in fact equal to two in most cases. For H in $\mathcal{H}_2$ one has $0 < \nu_P < 1$, explaining why one has a well-defined gel point at one extreme but not at the other.

The statistics of the first-shell model for variable ν are examined in Whittle (1994), where improper integral characterisations such as (A3.35) are converted to proper equations involving extremal operators. The calculation is interesting, in that the Fenchel transform of a convex function that found such natural application in Part I has now to be generalised to the Legendre transform, multi-valued, of a function that shows mixed convex/concave behaviour.

The material of the last section raises another point. Some of the early models of polymer formation (Watson, 1958; Gordon, 1962; Good, 1963) were formulated as a branching process. This is a mechanism suggested first in the context of population growth. A man has a random number of sons, they themselves do the same independently, following the same probability distribution, and so on. These individuals are then the nodes of a tree, which has a positive probability of becoming infinite if the expected number of sons exceeds unity. The authors mentioned carried this analogy to the case of polymerisation, one of the units to which a given unit is attached being regarded as his 'father' and the others as his 'sons'. The cases of sub- and supercriticality are those in which the consequent tree is infinite with zero or positive probability respectively. The approach was rediscovered by Newman *et al.* (2001) in their attempt to generate a random graph with prescribed degree distribution.

It is not at all clear that this mechanism represents the physical realities of polymerisation. However, one can demonstrate that the mathematics of the branching model agree with those of the first-shell polymerisation model in the case $\nu = 0$, when cycles are forbidden. This was established in Whittle (1986). We repeat the analysis, since this relates to matters much discussed in the current literature: the apparent power-law distribution of degree in nets such as the Web.

Consider first the branching process analysis. Let $\Pi_0(z) = \sum_k P_k z^k$ be the probability generating function (p.g.f.) of the total number k of bonds a unit forms. Let $\Pi_1(z)$ be

the p.g.f. of the further number of bonds formed by a unit at one end of a randomly chosen bond. Then this is proportional to $\sum_k kP_k z^{k-1}$, so that

$$\Pi_1(z) = \frac{\Pi_0'(z)}{\Pi_0'(1)}. \tag{A3.62}$$

It is $\Pi_1(z)$ that is the p.g.f. of the number of sons, in the population version of the model. Then the total progeny X of a single individual will be infinite with positive probability if the expected number of sons $\mu = \Pi_1'(1)$ exceeds unity. The p.g.f. $\Psi(z)$ of X (the initial individual being included) satisfies the functional equation

$$\Psi(z) = z\Pi_1[\Psi(z)]. \tag{A3.63}$$

Equation (A3.63) has two solutions, corresponding to the two solutions of the equation

$$\beta = \Pi_1(\beta), \tag{A3.64}$$

to which (A3.63) reduces when $z = 1$. If $\mu < 1$, so that total progeny is finite, then the smaller solution of (A3.64) is unity, and the solution of (A3.63) for which $\Psi(1) = 1$ is indeed the required and proper p.g.f. If $\mu > 1$ then the smaller solution β of (A3.64) is the probability of 'extinction', i.e. of the event $X < \infty$. If $\Psi^*(z)$ is the p.g.f. of X conditional on $X < \infty$ then $\beta\Psi^*(z)$ satisfies (A3.63), and is to be identified with the solution of (A3.63) for which $\Psi(1) = \beta$. This is the equivalent of the assertion that the smaller ξ-root of the system (A3.43) determines the statistics of the sol fraction, whether the system is sub- or supercritical.

If one chooses an individual at random to constitute the root of the tree then the total population R with which he is connected includes himself and relations through his father as well as his actual offspring. It has p.g.f.

$$\Phi(z) = z\Pi_0[\Psi(z)]. \tag{A3.65}$$

Turning now to the actual tree-polymer process, we know from (A3.20) that we must have the identification

$$\Pi_0(z) = \frac{H(\bar{\xi}z)}{H(\bar{\xi})}, \tag{A3.66}$$

implying further that

$$\Pi_1(z) = \frac{H'(\bar{\xi}z)}{H'(\bar{\xi})}. \tag{A3.67}$$

If we are to prove consistency of the polymer formulation and the branching process formulation then we must demonstrate that the Φ and Ψ deduced from equations (A3.63) and (A3.65) and specifications (A3.66) and (A3.67) are consistent with the equivalent quantities derived from the polymer model. In fact, we shall proceed in the opposite direction.

Let R denote the size of the polymer within which a randomly chosen unit lies and X the number of units connected by some path to a given end of a randomly chosen bond.

These are the quantities whose respective probability generating functions should be $\Phi(z)$ and $\Psi(z)$. We shall determine their evaluations from the polymer model and then verify that they do indeed obey relations (A3.63) and (A3.66).

Theorem A3.11 *Consider polymer statistics of the first-shell model in the case $\nu = 0$ for a fixed value $\overline{w}$ of w (corresponding to either an open process or a closed process in the thermodynamic limit). Let $\overline{\xi}$ be the solution of*

$$\overline{w}H'(\overline{\xi}) = \kappa\overline{\xi} \tag{A3.68}$$

and ξ^* the solution of

$$\overline{w}zH'(\xi^*) = \kappa\xi^*. \tag{A3.69}$$

Then

$$\Phi(z) = \frac{zH(\xi^*)}{H(\overline{\xi})} \tag{A3.70}$$

and

$$\Psi(z) = \xi^*/\overline{\xi} \tag{A3.71}$$

and these functions satisfy (A3.63), (A3.65).

Proof By the usual reasoning we have

$$\mathrm{E}(z^R) = \frac{\sum_r R\gamma_r(\overline{w}z)^R}{\sum_r R\gamma_r\overline{w}^R} = \frac{(G_w)_{w=\overline{w}z}}{(G_w)_{w=\overline{w}}}. \tag{A3.72}$$

Since $G_w = wH$ then (A3.70) follows.

We have now

$$\Psi(z)^2 = \frac{\sum_r (R-1)\gamma_r(\overline{w}z)^R}{\sum_r (R-1)\gamma_r\overline{w}^R}, \tag{A3.73}$$

since both sides evaluate the p.g.f. of the size of a polymer within which a randomly chosen bond lies. Now, for corresponding w and ξ we have

$$\sum_r (R-1)\gamma_r w^R = w\xi H' = \kappa\xi^2. \tag{A3.74}$$

Thus (A3.73) reduces to

$$\Psi(z)^2 = (\xi^*/\overline{\xi})^2,$$

whence (A3.71) follows.

Appealing now to relations (A3.66)–(A3.69) one quickly verifies that the functions defined by (A3.70) and (A3.71) satisfy the relations (A3.63) and (A3.65). $\Diamond$

The conclusion is then that the branching analysis is consistent with the full polymer analysis in the case $\nu = 0$. One specifies the degree distribution by specifying $\Pi_0(z)$, otherwise expressed in (A3.66). The variable $\overline{\xi}$ appears merely as a parameter of the

model; one can take any positive value for it up to (but possibly not including) the radius of convergence of H. One could interpret it as an 'activity' parameter, encouraging bond formation. On this basis, relation (A3.66) is a proper specification for any value of $\overline{\xi}$ in this range, whether the regime be sub- or supercritical. The only difficulty comes if we wish to relate the activity ξ to the density ρ, the latter having a meaning only in the polymer version. The relation is specified by requiring that $\overline{\xi}$ be the value of ξ for which $H(\xi)^\rho \exp(-\kappa\xi^2/2)$ has a local maximum. This value of ξ will exist and increase monotonically with ρ for H in $\mathcal{H}_1$ and ν is sufficiently small, but may fail to do so otherwise.

A3.6 The distribution of degree

The observation in Chapter 18 of the ubiquity of power-law degree distributions naturally poses the question of whether such a distribution could be produced within the class of first-shell models. One would think that it is only a question of appropriately specifying the function $H(\xi)$, which determines the degree distribution as having p.g.f.

$$\Pi_0(z) = \sum_k P_k z^k = \frac{H(\overline{\xi}z)}{H(\overline{\xi})} \tag{A3.75}$$

under a wide range of conditions. Here $\overline{\xi}$ can be interpreted as an activity parameter, related to the unit density ρ by the condition just stated at the end of Section A3.5. Relation (A3.75) certainly holds for H in $\mathcal{H}_1$ with the cyclisation-favouring parameter ν set equal to unity, both in the sub- and supercritical regimes. However, such H have infinite radius of convergence, and so can never generate a power-law distribution, however large $\overline{\xi}$ may be.

If one sets $\nu = 0$, thus forbidding all but tree-molecules, then relation (A3.75) holds for H with a finite radius of convergence $\hat{\xi}$. As mentioned in the last section, Newman *et al.* revived the approach of earlier authors, explaining the development of tree-molecules by a branching model, with the p.g.f. $\Pi_0(z)$ arbitrarily specified. On this view there is nothing to stop the activity parameter $\overline{\xi}$ being taken right up to $\hat{\xi}$. In the last section we demonstrated the equivalence of the branching model and the dynamic polymerisation model as far as molecular statistics are concerned in the tree-restricted case. However, the polymer model does show a second critical transition (at a greater density than the gel point) for some H not belonging to $\mathcal{H}_1$, which can set an upper limit on $\overline{\xi}$ short of $\hat{\xi}$.

To summarise, if by a 'power-law' distribution one means that

$$P_k \sim \text{const. } k^{-\gamma}\theta^k \tag{A3.76}$$

for large k, where $\theta < 1$ and $\gamma > 0$, then this is easily achieved over a variety of polymerisation models, although it does imply that $H(\xi)$ has a branch point at its radius of convergence $1/\theta$. If one requires that $\theta = 1$ (i.e. that there be no 'exponential cut-off') then one is certainly asking for something more special: that H have a branch point at $\xi = 1$ for which $\gamma > 2$, with the possibility that $\overline{\xi} = 1$ might not correspond to a permissible value of ρ in the polymerisation model.

A3.7 Literature and further directions

We restrict attention to the random graph literature before the proposal of mechanisms such as 'preferential attachment' to explain power-law degree distributions. The interest has been precisely to see whether any of the classical generating mechanisms could explain such distributions.

As already mentioned, the pure mathematical approach to random graphs began with the discussion of the zeroth-order model by Erdös and Rényi (1959, 1960). This line has been greatly elaborated, notably by Bollobás (see Bollobás, 2001 for a review of the now extensive literature). However, this approach misses the rich structure of more physically based models and the analytic development which these permit. Bollobás does refer to the 'generating function approach' as having some merit, but not yielding the detailed conclusions that nonanalytic methods can (painfully!). However, he is referring to the tree-restricted techniques associated with a purely branching model.

On the more physical formulations, one may say that an explicit statistical and graph-theoretical treatment of polymerisation began with the classic work of Stockmayer (1943, 1944) and Flory (1953). Both these assumed a model of pure association whose parameters evolved deterministically in time, and whose statistics were assumed to adapt immediately to the corresponding Gibbs distribution. This hybrid approach captured a good deal. In particular, Stockmayer was able to calculate formulae of the type of (A3.61) for molecular abundances and determine the gel point. However, the approach was incomplete as a model. Watson (1958), Gordon (1962) and Good (1963) then introduced the branching process analysis described in Section A3.5. Again, this seemed to be capturing something, but did not amount to a full physical model. It was limited to tree polymers, but does show insight, in that the analysis does turn out to have a clear correspondence with that for a more transparently physical model, as demonstrated in Section A3.5. In particular, Good (1960) deduced Stockmayer's distributional results by a systematic appeal to the Lagrange expansion of an implicitly determined function, reproduced here in the passage from (A3.57) to (A3.58). Newman *et al.* (2001) proposed the branching mechanism again, apparently unaware of its previous appearance in this context, but with the particular motivation of developing a class of models with specified degree distribution.

Whittle (1965a,b) sought for a more explicit physical model, and set up a dynamic stochastic model of association and dissociation which, being time-reversible, showed detailed balance. In fact, one could write down the equilibrium distribution of a complete description immediately. The problem comes when one considers reduced descriptions and runs into heavy combinatorics. Attention focused principally on the so-called first-shell model, one step up from the zeroth-order model. However, simplifying insights emerged in Whittle (1980b, 1981) and led to a fairly full theory, treated briefly in Whittle (1985) and, much more extensively, as one part of Whittle (1986). It says something for the intuitions of earlier workers that their conclusions survived passage to a more comprehensive model: more comprehensive in that it proposed a more fundamental physical model as well as allowing immediate generalisation. The model and its analysis generalise immediately to the case of several types of unit, either fixed or mutable, and were later extended to the case of directed bonds (Whittle, 1990a) and other variations (Whittle, 1990b, 1992).

The particular role of the parameter ν, interpretable both as indicating a favouring of intra- over interpolymer bonds in one model and as the number of alternative locations or states that molecules can adopt in a related model, was perceived in Whittle (1986). However, it had already been remarked by Kasteleyn and Fortuin (1969) for what was essentially a zeroth-order model. In the physical literature the parameter is denoted by q, and the equivalence mentioned had in fact been known to mathematicians; see Wu (1982). The discussion of dependence of behaviour upon a nonintegral ν in Whittle (1994) marks a shift from what are essentially Fourier-transform methods to Legendre-transform methods. This type of shift is of course familiar in the passage from wave optics to geometric optics, and the like.

References

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993) *Network Flows: Theory, Algorithms and Applications.* Englewood Cliffs, NJ: Prentice-Hall.

Albert, R., Jeong, H. and Barabási, A. L. (1999) Diameter of the world-wide web. *Nature*, **401**, 130–31.

Albert, R. and Barabási, A. L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47–97.

Aldous, D. J. (1987) Ultimate instability of exponential back-off protocol for acknowledgement-based transmission control of random access communication channels. *IEEE Trans. Inf. Th.*, **IT-33**, 219–23.

Amari, S. (1998) Natural gradient works efficiently in learning. *Neural Computat.*, **10**, 251–76.

Amari, S. and Nagaoka, H. (2000) *Methods of Information Geometry*. New York: Oxford University Press and American Mathematical Society.

Arthurs, A. M. (1970, republished 1980) *Complementary Variational Principles.* Oxford: Clarendon Press.

Bagge, M. (2000) A model of bone adaptation as an optimization process. *J. Biomechs.*, **33**, 1349–57.

Ball, M. O., Magnanti, T. L., Monma, C. L. and Nemhauser, G. L. (eds) (1995a) *Network Models*. Volume 7, *Handbooks in Operational Research and Management Science*. Amsterdam: North-Holland.

Ball, M. O., Magnanti, T. L., Monma, C. L. and Nemhauser, G. L. (eds) (1995b) *Network Routing*. Volume 8, *Handbooks in Operational Research and Management Science*. Amsterdam: North-Holland.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A. and Sugiyama, Y. (1995) Dynamical model of traffic congestion and numerical simulation. *Phys. Rev.* E, **51**, 1035–42.

Barabási, A. L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–12.

Beckmann, M., McGuire, C. B. and Winsten, C. B. (1956) *Studies in the Economics of Transportation*. New Haven, CT: Yale University Press.

Bendsøe, M. P. (1986) Generalized plate models and optimal design. In Erickson, J. L. *et al.* (eds), *Homogenization and Effective Moduli of Materials and Media*. Berlin: Springer-Verlag; pp. 1–26.

Bendsøe, M. P. and Kikuchi, N. (1988) Generating optimal topologies in structural design using a homogenization method. *Comput. Meth. Appl. Mech. Eng.*, **71**, 197–224.

Bendsøe, M. P. and Sigmund, O. (2003) *Topology Optimization*. Heidelberg: Springer-Verlag.

Bertsekas, D. P. (1987) *Dynamic Programming; Deterministic and Stochastic Models.* Englewood Cliffs, NJ: Prentice-Hall.

Bertsekas, D. P. (1998) *Network Optimization: Continuous and Discrete Models*. Nashua, NH: Athena Scientific.

Bertsekas, D. P., Nedic, A. and Asuman, E. O. (2003) *Convex Analysis and Optimization.* Nashua, NH: Athena Scientific.

Bollobás, B. (2001) *Random Graphs*, 2nd edn. Cambridge: Cambridge University Press.

Bollobás, B., Riordan, O., Spencer, J. and Tusnády, G. (2001) The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, **18**, 279–90.

Bonald, T. and Massoulié, L. (2001) Impact of fairness on Internet performance. *Proc. ACM Sigmetrics 2001*.

Borel, E. (1925) *Principes et formules classiques du Calcul des Probabilités. Traité du Calcul des Probabilités et de ses Applications*. Paris: Gauthier-Villars.

Braess, D. (1968) Über ein paradoxon der verkehrsplanung. *Unternehmensforschung*, **12**, 258–68.

Bramson, M. (1994a) Instability of FIFO queueing networks. *Ann. Appl. Prob.*, **4**, 414–31.

Bramson, M. (1994b) Instability of FIFO queueing networks with quick service times. *Ann. Appl. Prob.*, **4**, 693–718.

Bramson, M. (1996) Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems*, **22**, 5–45.

Bramson, M. (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, **30**, 89–148.

Bramson, M. and Dai, J. G. (2001) Heavy traffic limits for some queueing networks. *Ann. Appl. Prob.*, **11**, 49–90.

Brockmeyer, E., Halstrom, H. L. and Jensen, A. (1948) *The Life and Works of A. K. Erlang*. Copenhagen: Academy of Technical Sciences.

Burman, D. Y., Lehoczky, J. P. and Lim, Y. (1984) Insensitivity of blocking probabilities in a circuit-switching network. *J. Appl. Prob.*, **21**, 850–59.

Carter, D. R. and Beaupré, G. S. (2001) *Skeletal Function and Form*. Cambridge: Cambridge University Press.

Carter, D. R., Orr, T. E. and Fyhrie, D. P. (1989) Relationships between loading history and femoral cancellous bone architecture. *J. Biomech.*, **22**, 231–44.

Cheng, G. D. and Olhoff, N. (1982) Regularized formulation for optimal design of axisymmetric plates. *Int. J. Solids Struct.*, **18**, 153–69.

Cheng, G. D. and Pedersen, P. (1997) On sufficiency conditions for optimal design based on extremal principles of mechanics. *J. Mech. Phys. Solids*, **45**, 135–50.

Dai, J. G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Prob.*, **5**, 49–77.

de Veciana, G., Lee, T. -J. and Konstantopoulos, T. (2001) Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Network.*, **9**, 2–14.

Dorogovtsev, S .N. and Mendes, J. F. F. (2000) Evolution of networks with aging of sites. *Phys. Rev. E*, **62**, 1842–1845.

Dorogovtsev, S. N. and Mendes, J. F. F. (2003) *Evolution of Networks*. Oxford: Oxford University Press.

Dorogovtsev, S. N., Mendes, J. F. F. and Samukhin, A. N. (2000) Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, **85**, 4633–4636.

Doyle, P. G. and Snell, J. L. A. (1984) *Random Walks and Electrical Networks*. Carus Mathematical Monographs. Mathematical Association of America.

Erdös, P. and Rényi, A. (1959) On random graphs I. *Publ. Math. Debrecen*, **5**, 290–97.

Erdös, P. and Rényi, A. (1960) On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, 17–61.

Fletcher, R. (1987) *Practical Methods of Optimization*. New York: John Wiley & Sons.

Flory, P. J. (1953) *Principles of Polymer Chemistry*. Ithaca, NY: Cornell University Press.

Freeman, W. J. (1975) *Mass Action in the Nervous System.* New York: Academic Press.

Freeman, W. J. (1987) Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biol. Cybernet.*, **56**, 139–50.

Freeman, W. J. (1992) Tutorial on neurobiology: from single neurons to brain chaos. *Int. J. Bifurcat. Chaos*, **2**, 451–82.

Freeman, W. J. (1994) Role of chaotic dynamics in neural plasticity. In van Pelt, J., Corner, M. A., Uylings, H. B. M. and Lopez da Silva, F. H. (eds), *Progress in Brain Research*, Vol. 102. Amsterdan: Elsevier; pp. 319–33.

Freeman, W. J., Yao, Y. and Burke, B. (1988) Central pattern generating and recognizing in the olfactory bulb: a correlation learning rule. *Neural Networks*, **1**, 277–88.

Frégnac, Y. (1999) A tale of two spikes. *Nature Neurosci.*, **2**, 299–301.

Gallager, R. G. (1977) A minimum delay routing algorithm using distributed computation. *IEEE Trans. Commun.*, **25**, 73–85.

Gerla, M. and Kleinrock, L. (1977) On the topological nature of distributed computer networks. *IEEE Trans. Commun.*, **25**, 48–60.

Gibbens, R. J. (1988) Dynamic routing in circuit-switched networks: the dynamic alternative routing strategy. Ph.D. thesis, University of Cambridge.

Gibbens, R. J., Kelly, F. P. and Key, P. B. (1988) Dynamic alternative routing – modelling and behaviour. In Bonatti, M. (ed.), *Proc. 12th Int. Teletraffic Congress, Turin*. Amsterdam: Elsevier.

Gibbens, R. J. and Kelly, F. P. (1990) Dynamic routing in fully connected networks. *IMA J. Math. Cont. Informat.*, **7**, 77–111.

Gibbens, R. J. and Kelly, F. P. (1999) Resource pricing and the evolution of congestion control. *Automatica*, **35**, 196–198.

Girard, A. and Ouimet, Y. (1983) End-to-end blocking for circuit-switched networks: polynomial algorithms for some special cases. *IEEE Trans. Commun.*, **31**, 1269-73.

Gittins, J. C. (1979) Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B*, **41**, 148–77.

Gittins, J. C. (1989) *Multi-armed Bandit Allocation Indices.* Chichester: John Wiley & Sons.

Gittins, J. C. and Jones, D. M. (1974) A dynamic allocation index for the sequential design of experiments. In Gani, J. (ed.), *Progress in Statistics*. Amsterdam: North-Holland; pp. 241–66.

Good, I. J. (1960) Generalisations to several variables of Lagrange's expansion, with applications to stochastic processes. *Proc. Camb. Phil. Soc.*, **56**, 366–80.

Good, I. J. (1963) Cascade theory and the molecular weight averages of the sol fraction. *Proc. Roy. Soc. A*, **272**, 54–9.

Gordon, M. (1962) Good's theory of cascade processes applied to the statistics of polymer distributions. *Proc. Roy. Soc. A*, **268**, 240–59.

Grossberg, S. (1988) Nonlinear neural networks: principles, mechanisms and architectures. *Neural Networks*, **1**, 17–62.

Harrison, J. M. (1985) *Brownian Motion and Stochastic Flow Systems*. New York: John Wiley & Sons.

Harrison, J. M. and Nguyen, V. (1993) Brownian models of multiclass queueing networks. *Queueing Systems*, **13**, 5–40.

Harrison, J. M. and Wein, L. M. (1990) Scheduling networks of queues: heavy traffic analysis of a simple open network. *Oper. Res.*, **38**, 1052–64.

Harrison, J. M. (2000) Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Prob.*, **10**, 75–103.

Harrison, J. M. (2003) A broader view of Brownian networks. *Ann. Appl. Prob.*, **13**, 1119–50.

Hassoun, M. H. (1995) *Fundamentals of Artificial Neural Networks*. Cambridge MA, London: MIT Press.

Hebb, D. O. (1949) *The Organisation of Behaviour.* New York: John Wiley & Sons.

Hemp, W. S. (1973) *Optimal Structures.* Oxford: Clarendon Press.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci.*, **79**, 2254–8.

Huiskes, R. (2000) If bone is the answer, what is the question? *J. Anat.*, **197**, 145–56.

Intriligator, M. D. (republished 2002) *Mathematical Optimization and Economic Theory.* SIAM.

Jackson, J. R. (1963) Jobshop-like queueing systems. *Mgmt. Sci.*, **10**, 131–42.

Jackson, R. R. P. (1954) Queueing systems with phase-type service. *Operat. Res. Quart.*, **5**, 109–20.

Jacobson, V. (1988) Congestion avoidance and control. In *Proc. ACM SIG-COMM '88*, pp. 314–29.

Kasteleyn, P. W. and Fortuin, C. M. (1969) Phase transitions in lattice systems with random local properties. *J. Phys. Soc. Japan*, **26** (Suppl.), 11–14.

Kay, L. M., Shimoide, K. and Freeman, W. J. (1995) Comparison with EEG time series from rat olfactory system with model composed of nonlinear coupled oscillators. *Int. J. Bifurcat. Chaos*, **5**, 849–58.

Kelly, F. P. (1979) *Reversibility and Stochastic Networks*. Chichester: John Wiley & Sons.

Kelly, F. P. (1986) Blocking probabilities in large circuit-switched networks. *Adv. Appl. Prob.*, **18**, 473–505.

Kelly, F. P. (1986) Blocking and routing in circuit-switched networks. In Boxma, O. J., Cohen, J. W. and Syski, R. (eds), *Teletraffic Analysis and Computer Performance Evaluation*. Amsterdam: Elsevier; pp. 37–45.

Kelly, F. P. (1988a) Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Adv. Appl. Prob.*, **20**, 112–44.

Kelly, F. P. (1988b) The optimisation of queueing and loss networks. In Boxma, O. J. and Syski, R. (eds), *Queueing Theory and its Applications*. Amsterdam: Elsevier; pp. 375–92.

Kelly, F. P. (1990) Routing and capacity allocation in networks with trunk reservation. *Maths. Op. Res.*, **15**, 771–93.

Kelly, F. P. (1991a) Loss networks. *Ann. Appl. Prob.*, **1**, 317–78.

Kelly, F. P. (1991b) Network routing. *Phil. Trans. Roy. Soc. London A*, **337**, 343–67.

Kelly, F. P. (1995) Modelling communications networks, present and future. *Proc. Roy. Soc. A*, **444**, 1–20.

Kelly, F. P. (2000) Models for a self-managed Internet. *Phil. Trans. Roy. Soc. A*, **358**, 2335–48.

Kelly, F. P. (2001) Mathematical modelling of the Internet. In Engquist, B. and Schmid, W. (eds), *Mathematics Unlimited – 2001 and Beyond*. Berlin: Springer-Verlag; pp. 685–702.

Kelly, F. P. (2003) Fairness and stability of end-to-end congestion control. *Europ. J. Control*, **9**, 159–76.

Kelly, F. P. and Gibbens, R. J. (1990) Dynamic routing in fully connected networks. *IMA J. Math. Control. Inf.*, **7**, 77–111.

Kelly, F. P., Gibbens, R. J. and Key, P. B. (1995) Dynamic alternative routing. In Steenstrup, M. E. (ed.), *Routing in Communication Networks*. Englewood Cliffs, NJ: Prentice-Hall; pp. 13–47.

Kelly, F. P. and Laws, C. N. (1993) Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems*, **13**, 47–86.

Kelly, F. P. and MacPhee, I. M. (1987) The number of packets transmitted by collision detect random access schemes. *Ann. Prob.*, **15**, 1557–68.

Kelly, F. P., Maulloo, A.K. and Tan, D. K. H. (1998) Rate control in communication networks: shadow prices, proportional fairness and stability. *J. Op. Res. Soc.*, **49**, 237–52.

Kelly, F. P. and Williams, R. J. (2004) Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Prob.*, **14**, 1055–83.

Kerner, B. S. and Konhäuser, P. (1994) Structure and parameters of clusters in traffic flow. *Phys. Rev. E*, **50**, 54–83.

Key, P., Massoulié, L., Bain, A. and Kelly, F. P. (2004) Fair Internet traffic integration: network flow models and analysis. *Ann. des Télécomm.*, **59**, 1338–52.

Kleinrock, K. (1964) *Communication Nets: Stochastic Message Flow and Delay.* New York: McGraw-Hill.

Kleinrock, L. (1976) *Queueing Systems, Volume 2: Computer Applications*. New York: John Wiley & Sons.

Klimov, G. P. (1974) Time-sharing service systems I. *Theory. Prob. Appl.*, **19**, 532–51.

Klimov, G. P. (1978) Time-sharing service systems II. *Theory. Prob. Appl.*, **23**, 314–21.

Kumar, S., Saini, S. and Prakash, P. (1996) Alien attractors and memory annihilation of structured sets in Hopfield networks. *IEEE Trans. Neural Networks*, **7**, 1305–9.

Levy, R. (1991) Fixed point theory and structural optimization. *Eng. Opt.*, **17**, 251–261.

Li, L., Alderson, D., Willinger, W. and Doyle, J. (2004) A first-principles approach to understanding the Internet's router-level topology. *Proc. 2004 SIGCOMM Conference on Applications, Technologies, Architectures and Protocols for Computer Communications*. New York: ACM Press; pp. 3–14. Available at http//portal.acm.org.

Lighthill, M. J. and Whitham, G. B. (1955) On kinematic waves. I. Flood movement in long rivers. II: A theory of traffic flow on long, crowded roads. *Proc. Roy. Soc. A*, **229**, 281–345.

Lin, X., Shroff, N. B. and Srikant, R. (2006) On the connection-level stability of congestion-controlled cummunication networks. Working paper, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. Available at http://comm.csl.uinc.edu/∼srikant.

Lippmann, R. P. (1987) An introduction to computing with neural nets. *IEEE Mag. Acoust., Signals Speech Process.*, **4**, 4–22.

Luenberger, D. G. (1989) *Linear and Nonlinear Programming.* Addison-Wesley.

Massoulié, L. and Roberts, J. W. (2000) Bandwidth sharing and admission control for elastic traffic. *Telecomm. Systems*, **15**, 185–201.

Masur, E. F. (1970) Optimum stiffness and strength of elastic structures. *J. Eng. Mechs.*, **96**, 621–40.

Michell, A. G. M. (1904) The limits of economy of material in frame structures. *Phil. Mag.*, **8**, 589–97.

Mo, J. and Walrand, J. (2000) Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Network.*, **8**, 556–67.

Nagel, K., Wagner, P. and Woesler, R. (2003) Still flowing: approaches to traffic flow and traffic jam modelling. *Operat. Res.*, **51**, 681–710.

Nagurney, A. (1993) *Network Economics*. Dordrecht: Kluwer.

Newman, M. E. J., Strogatz, S. H. and Watts, D. J. (2001) Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E*, **64**, 026118–1–026118–17.

Oja, E. (1982) A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, **15**, 267–73.

Olivier, von G. (1972) Cost-minimum priorities in queueing systems of type M/G/1. *Elektron. Rechenanl.*, **14**, 262–71.

Paganini, F., Wang, Z., Doyle, J. C. and Low, S. H. (2005) Congestion control for high performance, stability and fairness in general networks. *IEEE/ACM Trans. Networking*, **13**, 43–56.

Percus, J. K. (1971) *Combinatorial Methods*. Berlin: Springer.

Prager, W. and Taylor, G. I. N. (1968) Problems of optimal structural design. *J. Appl. Mech.* **35**, 102–6.

Querin, Q. M. (1997) Evolutionary structural optimization, formulation and implementation. Ph.D. thesis, Department of Aeronautical Engineering, University of Sydney, Australia.

Rattray, M. and Saad, D. (1999) Analysis of natural gradient descent for multilayer neural networks. *Phys. Rev. E*, **59**, 4523–32.

Rougham, M., Greenberg, A., Kalmanek, C., Rumsewicz, M., Yates, J. and Zhang, Y. (2003) Experience in measuring backbone traffic variability: models, metrics, measurements and meaning. *International Teletraffic Congress (ITC)*, **18**, 2003.

Rozvany, G. I. N. and Zhou, M. (1991) The COC algorithm, part I: Cross-section optimization or sizing. *Comput. Meth. Appl. Mech. Eng.,* **89**, 281–308.

Rozvany, G. I. N., Zhou, M. and Sigmund, O. (1994) Topology optimization in structural design. In Adeli, H. (ed.), *Advances in Design Optimization*. London: Chapman and Hall; pp. 340–99.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L. and PDP Research Group (eds), *Parallel Distributed Programming: Explorations in the Microstructure of Cognition*, Vol. 1, *Foundations*. Cambridge, MA: MIT Press; pp. 318–62.

Rybko, A. N. and Stolyar, A. L. (1992) Ergodicity of stochastic processes that describe the functioning of open queueing networks. *Prob. Inform. Trans.*, **28**, 3–26.

Scott, J. W., McBride, R. L. and Schneider, S. P. (1980) The organization of projections from the olfactory bulb to the pyriform cortex and olfactory tubercle in the rat. *J. Comp. Neurol.*, **194**, 519–34.

Sewell, M. J. (1987) *Maximum and Minimum Principles*. Cambridge, UK. Cambridge University Press.

Steenstrup, M. (ed.) (1995) *Routing in Communications Networks*. Englewood Cliffs, NJ: Prentice-Hall.

Stockmayer, W. H. (1943) Theory of molecular size distribution and gel formation in branched chain polymers. *J. Chem. Phys.*, **11**, 45–55.

Stockmayer, W. H. (1944) Theory of molecular size distribution and gel formation in branched polymers. II. General cross-linking. *J. Chem. Phys.*, **12**, 125–31.

Stolyar, A. L. (2005) Maximizing queueing network utility subject to stability. *Queueing Systems*, **50**, 401–57.

Svanberg, K. (1994) Global convergence of the stress ratio method for truss sizing. *Struct. Optim.*, **8**, 60–68.

Taylor, J. E. (1969) Maximum strength elastic structural design. *Proc. ASCE*, **95**, 653–63.

Taylor, J. E and Rossow, M. P. (1977) Optimal truss design based on an algorithm using optimality criteria. *Int. J. Solid Structs.*, **13**, 913–23.

Thompson, J. M. T. and Hunt, G.W. (1973) *A General Theory of Elastic Stability*. New York: John Wiley & Sons.

Toader, A. M. (1997) Convergence of an algorithm in optimal design. *Struct. Optim.*, **13**, 195–198.

Turing, A. M. (1952) The chemical basis of morphogenesis. *Phil. Trans. Roy. Soc. London B*, **237**, 37–72.

Vinnicombe, G. (2002a) On the stability of networks operating TCP-like congestion control. *Proc. 15th World Congress on Automatic Control, 2002*. Available at http://www.control.eng.com.ac.uk/gv/internet/ifac.pdf.

Vinnicombe, G. (2002b) Robust congestion control for the internet. Working paper, Department of Engineering, University of Cambridge. Available at http://www.control.eng.com.ac.uk/gv/internet/sigcomm.pdf.

von der Malsburg, C. (1973) Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetik*, **14**, 85–100.

Vvedenskaya, N. D., Dobrushin, R. I. and Karpelevich, F. I. (1996) A queueing system with selection of the shortest of two queues. *Prob. Inform. Trans.*, **32**, 15–27.

Wang, X. -J. and Rinzel, J. (2003) Oscillatory and bursting properties of neurons. In Arbib, M. A. (ed.), *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press; pp. 835–40.

Wasiutynski, Z. (1960) On the congruency of the forming according to the minimal potential energy with that according to equal strength. *Bull. de l'Académie Polonaise des Sciences, Série des Sciences Techniques*, **8**, 259–68.

Watson, G. S. (1958) On Goldberg's theory of the precipitin reaction. *J. Immunol.*, **80**, 182–5.

Whittle, P. (1965a) Statistical processes of aggregation and polymerisation. *Proc. Camb. Phil. Soc.*, **61**, 475–95.

Whittle, P. (1965b) The equilibrium statistics of a clustering process in the uncondensed phase. *Proc. Roy. Soc. A*, **285**, 501–19.

Whittle, P. (1980a) Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. B*, **42**, 143–9.

Whittle, P. (1980b) Polymerisation processes with intra-polymer bonding. I. One type of unit. II. Stratified processes. III. Several types of unit. *Adv. Appl. Prob.*, **12**, 94–115, 116–34, 135–53.

Whittle, P. (1981a) Arm-acquiring bandits. *Ann. Prob.*, **9**, 284–92.

Whittle, P. (1981b) A direct derivation of the equilibrium distribution for a polymerisation process. *Teoriya Veroyatnostei*, **26**, 350–61.

Whittle, P. (1982) *Optimisation over Time*, Vol. 1. Chichester: John Wiley & Sons.

Whittle, P. (1985) Random graphs and polymerisation processes. *Ann. Discrete Math.*, **28**, 337–48.

Whittle, P. (1986) *Systems in Stochastic Equilibrium*. Chichester: John Wiley & Sons.

Whittle, P. (1988) Approximation in large-scale circuit-switched networks. *Prob. Eng. Inf. Sci.*, **2**, 279–91.

Whittle, P. (1990a) The statistics of random directed graphs. *J. Stat. Phys.*, **56**, 499–516.

Whittle, P. (1990b) Fields and flows on random graphs. In Grimmett, G. R. and Welsh, D. J. A. (eds), *Disorder in Physical Systems*. Oxford: Oxford University Press; pp. 337–48.

Whittle, P. (1992) Random fields on random graphs. *Adv. Appl. Prob.*, **24**, 455–73.

Whittle, P. (1994) Polymer models and generalised Potts–Kasteleyn models. *J. Stat. Phys.*, **75**, 1063–92.

Whittle, P. (1998) *Neural Nets and Chaotic Carriers.* Chichester: John Wiley & Sons.

Whittle, P. (2004) The distribution of species numbers in a controlled environment. *Austr. NZ J. Statistics*, **46**, 23–7.

Whittle, P. (2005) Tax problems in the undiscounted case. *J. Appl. Prob.*, **42**, 754–65.

Williams, R. J. (1998) Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, **30**, 27–88.

Wilson, H. R. and Cowan, J. D. (1972) Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, **12**, 1–24.

Wolff, J. (1986) *The Law of Bone Remodelling.* Berlin: Springer-Verlag. (A translation of the original 1892 text *Das Gesetz der Transformationen der Knochen*, Hirschwald, Berlin.)

Wu, F. Y. (1982) The Potts model. *Rev. Mod. Phys.*, **54**, 235–68. *Phil. Mag.*, **8**, 589–97.

Xie, Y. M. and Steven, G. P. (1997) *Evolutionary Structural Optimization*. Berlin: Springer-Verlag.

Yao, Y. and Freeman, W. J. (1990) Model of biological pattern recognition with spatially chaotic dynamics. *Neural Networks*, **3**, 483–501.

Zhou, M. and Rozvany, G. I. N. (2001) On the stability of ESO type methods in topology optimization. *Struct. Multidisc. Opt.*, **21**, 80–83.

Index

The topics of the four parts of the text are sufficiently disjoint that it is helpful to compile a separate index for each part, cross-referencing where appropriate.

Part I: Distributional networks
anatomical terms, implicitly defined 113–15

balance relation 10, 42, 100
'black and white' structures 37, 118
Bendsøe–Sigmund material 3, 37–8, 116–25, 130–32
bone structure 38, 113–15, 132–3
braced frameworks 99–101
Braess paradox 93
buckling 101, 133–4

cancellous bone 38, 124, 132
cantilever 112–13, 120–21
coated sphere assemblage 124
'coat-hook' design 110–11, 127–30
complementary potential 96
compliance 117
congestion 86, 87–92
conjugate indices 13, 98
cooling problem 30–39
continuum formulations 24–41
  for structures 103–5
cost functions
  concave 66–70
  convex 10
  see also seminvariant cost functions
crank design 109

demand functions 22–3
design space 24
design optimisation: *see entries beginning with 'primal', 'dual' or 'reduction'*
destination-specific flows 42–6
directed links 42
displacement 95, 103
dissipation 13
dual form for design optimisation 19–20, 26–7, 44–6
  of structures 102–5
  under variable loading 54–6
dual form for flow optimisation 12, 26–7
dynamic programming equations 45, 78

electrical networks (resistive) 12, 13
environmental costs 18, 44, 67, 70–72
evolutionary optimisation algorithms 28–9
  ESO package 119
  for structural design 116–25
  SIMP package 118
  under variable load 61–4
exchange optimisation 74–84
external nodes 14

Fenchel transform 11, 14
flow optimisation: *see entries beginning with 'primal' or 'dual'*
flows, simple 1–25
foundation 99
fundamental diagram 89–90

geodesics 40
glial cells 39, 142
'grey' structures 37, 118, 122–5
grid element 86

Hencky–Prandtl nets 105–8
  examples 108–9
hierarchical structures 4, 72–4
  for exchange optimisation 74–84

homogenisation 123
Hooke's law 96, 97, 98

inequality
  Minkowski 51–2, 65
  triangular 15
interflow 230–31
internal nodes 14
  as locations for a store (depot, exchange) 23, 52
  optimisation of 48–54, 62–4, 71

Lagrange multipliers 11, 21–2, 103, 195
Lagrangian forms 11, 103, 139, 195
load splitting 53–4

MBB beam 120, 122
material costs 14
Michell structures 3, 95–115
multi-commodity flows 42–6

nodes 10
  See internal nodes
nonuniform spaces 39–41

Ohm's law 12
one-step look-ahead rule (OSLA) 79, 80
outflow function 74, 76–7, 83–4, 227–8
outflow density 229–30, 232–3

potential 12, 13, 17, 27, 96
primal form for design optimisation 14, 15, 24–6, 43, 45
  under variable loading 48–9
  and environmental constraint 70, 72
  of structures 101–2
primal form for flow optimisation 11, 24–6
proxies 29

queues 88

reduction of design solution 15, 27, 43
  for structures 101–2
relative flow density 25
road networks 3, 85–94
route choice 46

seminvariant cost functions 1, 12–14, 67
  for structures 97
shear lines (slip lines) 105
simple flows 1–23
star-shaped sets 16
stiffness 98
  stiffness tensor 117
strain 95, 96
  pure strain tensor 105, 117
  strain potential 96
stress 95
structural optimisation 95–134
  by evolutionary methods 116–25
  literature survey 125
  under variable load 126–33
  see also Michell structures
supply sets 16–18

termination of a net 77–83
tolls 45–6, 94
traffic intensity 88
traffic models 89–92
  Braess paradox for traffic 93
transportation problem 15, 16, 43
trunking 2, 49–54, 56–60, 66–8
  for the exchange problem 77–84
  under concave costs 66–70
  under variable load 49–54, 56–60

variable loading 47–65
  for structures 126–34
  the Bendsøe–Sigmund examples 130–32
  with environmental effects 70–72

withinflow 228–9

Part II: Artificial neural networks
activation function 137, 138, 159
adaptive scaling 151
Amari's accelerated learning algorithm 142
autoassociative operator 147

back propagation 139–41

canonical regression 143–5

escapement oscillation 163–4

feedback 151–2
feedforward net 138

Freeman model of neuron 158–9
Freeman oscillator 4, 160–61

granule call 154

Hamming net 148–9
Hebb's rule 140
Hopfield net 149–50

learning rules 141
linear least squares approximation 142–3

McCulloch–Pitts net 137
memory 147
mitral cell 154

neural mass 159
neural nets 159–60
neural oscillation 4, 160–65
  'chaotic' 164–5
  escapement 163–4
  gamma band 160
neuron 137
  Freeman model 160–61
neuronal bursting 4, 163–4

olfactory system 154–5
  analogies with PMA 155–7
oscillatory operation 158–65

probability maximising algorithm (PNA) 150–52
  analogies with olfactory system 155–7

synaptic matrix 138
synaptic weights 138

trace recognition 146–9

Part III: Processing networks
bandit processes 234–6

capacity optimisation 174–5
congestion 173, 175, 213

deterministic (fluid) models

fixed work-stations 187–9, 238–9

Gittins index 176, 234–6

Jackson networks 172–4

Klimov index 176, 181–3, 236–9
  examples 183–5
  performance 185–7, 237–8

priority rules 175, 213

queue populations 177–8
queue, simple 169–71
queueing networks 169–78

routing matrix 172

tax processes 182, 236–9
time and resource sharing 179–89
traffic equations 172–3

Part IV: Communication networks
emulation and power laws 223–6
Erlang's formula 210
Ethernet 212

fairness criterion 213

Internet 211–18

loss networks, deterministic 193–8
  admission and routing rules 193–5
  design rules 195–7
  free node optimisation 197–8
  reliability 198
loss networks, stochastic 199–210
  dynamic alternative routing 208–10
  stochastics of blocking 203–8
  trunk reservation 4, 201–3, 208–10

packet switching 211–12
performance of random and hierarchical nets 223
preferential attachment 220

random graphs 219–223, 240–60
random graphs and polymer models
  degree distribution 243–4, 258
  equivalence of the tree restriction and branching models 255–8
  first shell models, unit statistics and criticality 242–5

polymer statistics 245–8
Potts criticality 255, 259
tree molecules and criticality 248–54, 258
zeroth-order models 240–42

scale-free nets 4, 219. *See also* emulation and power laws

transmission control protocol (TCP) 212
more recent protocol derivations 213–18

Worldwide Web 4, 219–26