The Traffic Assignment Problem—
Models and Methods

The Traffic Assignment Problem—
Models and Methods

Michael Patriksson

Linköping Institute of Technology, Linköping, Sweden

Contents

Preface

This book is the result of several years of research into the modelling and efficient solution of problems in transportation planning and related areas. A previous version appeared as a long survey in my licentiate thesis ([743]) presented at the Department of Mathematics, Linköping Institute of Technology, and received a positive response from some leading researchers in the field of transportation research. Their positive criticism inspired me to further develop the survey into what has become the present book.

The aim of this book is to provide a unified account of the development of models and methods for the problem of estimating equilibrium traffic flows in urban areas, from the early days of transportation planning heuristics to today's advanced equilibrium models and methods. Also, the aim is to show the scope and—just as important—the limitations of present traffic models. The development is described and analyzed using the powerful instruments of nonlinear optimization and mathematical programming within the field of operations research. The book includes historical references as well as many recent developments, and aims to clarify the close relationships between several lines of development by placing them in a new, unifying framework.

The first part of the book is devoted to mathematical models for the analysis of transportation network equilibria. Chapter 1 describes the traditional transportation planning process of which traffic assignment is a central part. The development of traffic assignment heuristics is described. Chapter 2 analyzes the basic models of traffic assignment, based on the principles of Wardrop. Existence, uniqueness and stability results are given. Extensions of the basic models, including non-deterministic travel cost perceptions and additional flow relationships modelled through the introduction of side constraints, are discussed. Chapter 3 analyzes traffic equilibrium models for general travel cost functions such as variational inequality, nonlinear complementarity, and fixed point problems. The recent development of optimization reformulations of asymmetric variational inequalities is accounted for in detail.

The second part of the book is devoted to methods for traffic equilibrium problems. Chapter 4 gives a uniform description of methods for the basic traffic assignment models and their extensions discussed in Chapter 2. Important concepts, such as partial linearization, decomposition, and column generation, are described in detail for general convex programs, and are subsequently used to describe and interrelate traffic assignment methods. Chapter 5 gives the corresponding treatment of the general traffic equilibrium models described in Chapter 3, based on the concepts of cost approximation, decomposition, and column generation. Optimization reformulations of general traffic equilibrium problems are utilized to derive a new class of traffic equilibrium methods which requires mild assumptions on the models.

An appendix summarizes the definitions of the concepts most frequently used.

The scope of the material is limited to static models of traffic equilibrium; neither

dynamic nor combined traffic models are dealt with in detail. The results obtained in this book can, however, be applied to the analysis and solution of such models also.

In order to economize with the space available, the reader is often directed to other works for more details. The resulting reference list is extensive—it contains more than 1,000 entries—and serves the additional purpose of being a source for anyone interested in acquiring deeper knowledge in the field.

I can envisage two main uses for this book. The first is by researchers in transportation, operations research, and quantitative economics—and those entering these areas of research—who wish to extend their knowledge of equilibrium modelling and analysis, and of the foundations of efficient optimization methods adapted for the solution of large-scale models. The second use is in advanced graduate courses in the areas just mentioned. This book could provide the basic material for a course in transportation research. A course in structured mathematical programming, with application to traffic equilibrium problems, is defined by Chapters 2 and 4, or by Chapters 2–5, the latter including the foundations of variational inequality models and methods. A course in equilibrium modelling is defined by Chapters 2 and 3.

The text assumes some familiarity with nonlinear programming theory and techniques. It would therefore be preferable to combine material from this book with that of a modern textbook in nonlinear programming; I personally recommend using Bazaraa *et al.* [43].

A work of this type would be impossible without the help of many people. I especially thank my former tutor Prof. T. Larsson for guiding me through the optimization landscape, and for his collaboration in research upon which parts of this book is based, and Prof. A. Migdalas for introducing me to the area of transportation research. The assistance given by the library staff over the years in gathering many of the references has been invaluable. Pamela Vang helped in improving the English of the text. The book was sponsored in part by grants from the Swedish Transport and Communications Research Board (KFB), Swedish Institute, and the Royal Swedish Academy of Sciences.

Linköping, June 1994

Michael Patriksson

Some notations

The network

$\mathcal{G} = (\mathcal{N}, \mathcal{A})$	Directed network of nodes and links
$\mathcal{P}$, $\mathcal{Q}$, $\mathcal{C}$	Set of origins, destinations, and origin-destination (O-D) pairs
$\mathcal{W}_i$, $\mathcal{V}_i$	Set of links, initiated and terminating at node i
$\mathcal{R}_{pq}$	Set of simple routes between nodes p and q
$\mathbf{\Delta}^{\mathrm{T}} = (\delta_{pqra})$	Link-route incidence matrix
$\mathbf{A} = (a_{ib})$	Node-link incidence matrix
$\mathbf{\Gamma}^{\mathrm{T}} = (\gamma_{pqr})$	Route-O-D pair incidence matrix

Flows and costs

O_p, D_q	Total demand in origin p and destination q
$\mathbf{d} = (d_{pq})$	Fixed demand
$\mathbf{h} = (h_{pqr})$	Route flow
$\boldsymbol{\lambda} = (\lambda_{pqr})$	Portion of the demand on the route
$\mathbf{f} = (f_a)$, $\mathbf{f} = (f_{ij})$	Total link flow
$\mathbf{f}_{pq} = (f_{apq})$	Commodity link flow
$\mathbf{f} = \sum_{k \in \mathcal{C}} \mathbf{f}_k$	Total link flow
$H = \prod_{(p,q) \in \mathcal{C}} H_{pq}$	Set of feasible route flows
H_d	Set of feasible demand and route flows
F^r, F^n	Set of feasible link flows in link-route and node-link representations
F^n_{pq}	Set of feasible commodity flows
F^r_d	Set of feasible demand and link flows
$\mathbf{g} = (g_{pq})$	Demand function
$\mathbf{c} = (c_{pqr})$	Route travel cost
$\boldsymbol{\pi} = (\pi_{pq})$	Shortest route cost
π_{ik}	Potential at node i in commodity k
$\mathbf{t} = (t_a)$	Link travel cost
$\overline{\mathbf{t}} = (\overline{t}_a)$	Marginal link travel cost
t'_a	Derivative of link travel cost
t_a^{-1}	Inverse travel cost

Equilibrium

$\mathcal{R}^*_{pq}$	Equilibrium routes
$\mathbf{d}^*$	Equilibrium demand
H^*	Set of equilibrium route flows
$\mathbf{f}^*$	Equilibrium link flow
F^*, F^*_{pq}	Optimal face of F^n and F^n_{pq}

Sets

X	Nonempty, closed and convex set in $\Re^n$
$\Re^n_+$, $\Re^n_-$	Nonnegative and nonpositive orthants
$\Re_{++}$	The set of positive real numbers
$\mathtt{dim}\ (X)$	Dimension of the set X
2^X	Set of all subsets of X
$\mathcal{X}$	Index set of extreme points of a polyhedral set
$\dot{\mathcal{X}}$	Subset of $\mathcal{X}$
$\mathtt{conv}\ (\mathcal{X})$	Convex hull of the points in $\mathcal{X}$

$\mathcal{D}$	Index set of extreme rays of a polyhedral set
N_X, T_X	Normal and tangent cone of X
P_X, $P_X^{\mathbf{B}}$	Euclidean and $\mathbf{B}$-norm projection onto X
$[\cdot]_+$	Euclidean projection onto $\Re_+$
$\lvert\mathcal{C}\rvert$	Cardinality of a finite set $\mathcal{C}$
$X \times Y$, $\prod_{i\in\mathcal{C}} X_i$	Cartesian product of sets
$[\mathbf{x}, \mathbf{y}]$	Closed line segment

Vectors and matrices

$\mathbf{e}_i$	Unit vector
$\mathbf{x}_i$	Vector in $\Re^{n_i}$
$\mathbf{x}_{i_-}$, $\mathbf{x}_{i_+}$	Subvectors $(\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_{i-1}^{\mathrm{T}})^{\mathrm{T}}$ and $(\mathbf{x}_{i+1}^{\mathrm{T}}, \ldots, \mathbf{x}_m^{\mathrm{T}})^{\mathrm{T}}$
$\mathbf{I}$	Identity matrix
$\mathbf{B}$, $\widetilde{\mathbf{B}}$	Positive definite matrix and its symmetric part
$\mathtt{diag}\,(b_i)$, $\mathtt{diag}\,(\mathbf{B})$	Diagonal matrix
$\lVert\cdot\rVert$	Euclidean vector norm and induced matrix norm
$\lVert\cdot\rVert_{\mathbf{B}}$	Matrix norm defined by a symmetric and positive definite matrix $\mathbf{B}$
$\lvert\!\lvert\!\lvert\cdot\rvert\!\rvert\!\rvert$	Operator norm induced by the Euclidean vector norm

Functions

$T \in C^p$ on X	T is p times continuously differentiable on an open neighbourhood of X
$\boldsymbol{\xi}_T$	Subgradient of T
∇T	Gradient of T
$\nabla_i T$	Gradient of T with respect to $\mathbf{x}_i$
$\nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{y})$	Gradient of φ with respect to its first component $\mathbf{x}$
$\nabla^2 T$	Hessian of T
∇F	Jacobian of F
$T'(\mathbf{x}; \mathbf{p})$	Directional derivative of T at $\mathbf{x}$ in the direction of $\mathbf{p}$
L_X^T	Level set of T with respect to X
Ω	Set of optimal solutions
d_Ω	Distance to the optimal solution set
T^*	Optimal value of T
u.s.c., l.s.c.	Upper and lower semicontinuity
M_T	Lipschitz continuity constant
m_T	Strong convexity (or monotonicity) constant

Algorithms

T	Objective function
$\mathbf{x}^k$, $\mathbf{f}^k$	Iterate
$\mathbf{y}^k$, $\mathbf{y}(\mathbf{x})$	Auxiliary solution
$\mathbf{p}^k$	Search direction
l_k, γ_k	Step length parameters
φ	Convex function
$\boldsymbol{\Phi}$	Cost approximating mapping
ψ	Merit function
$D_{\mathcal{O}}$, $D_{\mathcal{C}}$	Origin and O-D pair based decomposition algorithm
D^S, D^P	Sequential and parallel decomposition algorithm
$C_{\mathcal{R}}$, $C_{\mathcal{O}}$, $C_{\mathcal{A}}$	Column generation based on route flows, origin flows, and link flows.

Part I

Models

Chapter 1

Urban traffic planning

1.1 Introduction

A significant amount of the activity in an urban area concerns the movement of people and goods between different locations in the transportation infrastructure, and a smooth and efficient transportation system is essential for the economic health and the quality of life within the urban region. When analyzing the present infrastructure for future investments and operating policies, a careful study of the transportation system is therefore among the most important components of the planning process.

The decades following World War II have seen an enormous increase in the demand for transportation. A vast majority of this increase is accounted for by the development of personal transport, which has its roots in the urbanization and the rising standards of living.[1] The increase of mobility has, however, also brought many serious problems into urban regions, such as pollution, increased accident rates, unwanted social effects on urban life due to highway expansion, and an inefficient use of the transportation system because of high congestion.

In *transportation planning studies* alterations of the existing transportation systems are evaluated with the objective of alleviating the above mentioned problems (among others), while also utilizing the full range of transport modes available.

Urban transportation planning has been an evolutionary process. Its beginnings may be traced to the home-interview studies conducted in more than 100 cities in the United States during the decade following the end of World War II. The concept of small sample interviews was then combined with cordon line surveys in order to derive patterns of urban travel. Future traffic usage of urban highway projects was predicted by manually assigning selected origin-destination (O-D) movements to the routes being planned. In the early 1950s there were studies investigating land use and traffic relationships because better estimating methods were needed in order to forecast the travel in the design year. Methods of forecasting future population and its distribution, trip generation analysis relating travel to underlying household characteristics (car ownerships, etc.), and planning for networks instead of single routes were introduced at this time. Improved procedures were facilitated by the growing use of punch card data processing systems and later by the increasing capabilities of electronic computers. The latter permitted greater sophistication in transportation planning because they permitted the examination of more alternatives. The "modelling" of future land-use plans and future highway and transit systems was combined with more elegant methods of evaluation. Criteria for determining if plans

[1] Foulds [371, 372] claims that the growth of vehicle fleets on a world-wide basis is of the order of 15 percent per annum.

met community objectives (a concept itself not generally introduced until the mid-1960s) could be increasingly quantified.

The first transportation studies made concerned only highway traffic, and saw the problem as being that of providing enough capacity for the estimated future demand for personal transport. Since the 1950s, however, it has been realized that transportation is not an isolated activity; indeed, the demand for travel facilities is a function of human land use activity and, conversely, the provision of transport facilities stimulates land use activity. This development can also be seen in the Federal-Aid Highway Act of 1962, which states that federally assisted highway projects must be "... based on a continuing comprehensive transportation planning process carried on cooperatively by states and local communities ...". As a result of these findings, recent transportation studies form integrated parts of the overall planning process, and the so called *3C philosophy* of continuing, comprehensive, and cooperative urban transportation planning characterizes the current status of the process. Transport planners focus more on improving public transport, as an alternative to the auto mode, in order to reduce highway congestion.

The transportation system is very complex, and its performance depends on decisions made on many levels of society (the goals and purposes of which may be in conflict with each other). The process of evaluating, designing and managing such a system can therefore not be carried out without the aid of properly formulated models.

Depending on the purpose of the transportation study, models may concern different components of the transportation system (land use patterns, control policies, trip generation and distribution, etc.), different levels of aggregation of the physical reality (macroscopic or microscopic models), different planning horizons (from the use in real-time traffic management systems up to 20 year forecasts), and be based on different modelling principles (statistical models, optimization models, simulation models).

As the understanding of the transportation system has grown, together with the increase in availability of computational tools for its analysis, the planning problem has become more complex. The costs have also increased, due partly to the increase in costs for the inventory stage, and also because several more alternatives are tested.[2] However, viewing these costs against the scale of the plans they produce, the planning costs are less than one percent of the total ([37]).

1.2 The transportation planning process

The basis of the modelling of transportation problems is a set of assumptions, the most important ones being that travel patterns are tangible, stable, and predictable, and that the demand for transportation is directly related to the distribution and intensity of land uses, which are capable of being accurately determined for some future date ([130]).

Domencich and McFadden [264, Chap. 1] provide one list of criteria which a demand-based transportation planning model should meet in order to be a practical tool for policy analysis: it should be *sensitive* to transportation policy, so that the effects of policy alternatives can be forecast; it should be *causal*, establishing the behavioural link between the attributes of the transportation system and the decisions of the individual. This leads to the investigation of *behavioural* models of individual travel demand. Further, it should

[2]Creighton [187] reports that the costs for obtaining data, prepare and test plans, and produce a final report for a three-year transportation study, amount to a cost of \$1.00–\$1.50 per capita, with larger studies costing less (per capita) than smaller studies. Boyce *et al.* [106] report that the annual cost of seven very large studies totalled about \$750,000 each per year.

be *flexible*, allowing application to a wide variety of planning problems without major data collection and calibration costs; it should be *transferable* from one urban setting to another, allowing reuse without expensive reestimation in each new setting; finally, it should be *efficient*, in terms of providing maximum forecasting accuracy per monetary unit spent on data collection.

The traditional approach to transportation planning is to identify a number of simple submodels of the whole system, which are then analyzed separately, and most often in sequence. This transportation planning process can be divided into the following steps:

Step 1 (*Organization and goal definition*) The first stage of the process includes obtaining agreement on the funding, participation, and organizational form, setting up the committee structure, and arranging for staffing the study. Statements of goals and objectives of the study are also made.

Step 2 (*Base year inventory*) At this stage the data that may be relevant to the analysis of the transportation system is collected. It includes an inventory of existing transportation facilities and their characteristics, existing travel patterns determined through origin-destination surveys and traffic measurements, and planning factors, such as land use, income distribution, neighbourhood structure, and types of employment. It also includes the collection of historical data for trend analyses, such as population growth and car ownership.

Step 3 (*Model analysis*) The purpose of this phase is to establish relations among various quantities measured in **Step 2**, and to calibrate these relations for the base year. The relations are usually determined through the use of the following mathematical models, which are considered in sequence, and where the output from one model is input to the next.

(a) (*Trip generation*) This model is used to determine the number of trips originating and terminating in different zones of the study area. These numbers, which are sometimes called production and attraction numbers, are usually defined as functions of socio-economic, locational and land use characteristics of the zone in question, and are divided into different categories of purpose, such as work and recreational trips.

(b) (*Trip distribution*) At this step, formulas are derived to describe the allocation of trips from a point of origin to the destination zones. These formulas are typically defined as functions of the production and attraction numbers of the different zones, produced in step **(a)**, and of the travel costs between them. In some models, traffic counts are used when determining the trip matrix.

(c) (*Modal split*) This model determines the portion of the total number of trips made between an origin and destination using different transport modes, the two most commonly considered being cars and public transit. The portions of trips in an origin-destination relation is normally derived from relative travel times and costs between modes, and also, in some cases, from the socio-economic and land use characteristics of the origin and destination, respectively.

(d) (*Traffic assignment*) In this model, the origin-destination trips are allocated to routes in the transportation network, in order to estimate the traffic volumes and travel times on the roads as functions of the network characteristics.

The underlying behavioural principle in the choice of route is normally that travellers try to minimize their own travel costs.

Step 4 (*Travel forecast*) Based on the data collected in **Step 1** and trend analyses, future land use, population distribution, etc., are predicted for a design year. The models developed and calibrated in **Step 3** are then used to estimate the generation and distribution of trips on the future transportation network.

Step 5 (*Network evaluation*) If alternative future transportation networks and facilities are proposed, in this step costs and benefits are compared between their predicted flow patterns, in order to provide a basis for an economic evaluation of the proposed new facilities.

In order to achieve a consistent output the steps of the planning process must be repeated. Indeed, the travel costs of the future transportation network given by Step 4 influence the trip distribution, and even the projected land use and trip generation! This inconsistency problem can (at least partially) be alleviated by considering parts of the process simultaneously. Recent research efforts are being made in this direction.

In the sequel, we shall study the different parts of the transportation planning process in more detail, and outline the most common methods employed for their solution. We will here concentrate on the models and methods developed within transportation planning studies, and describe those developed through academic research in subsequent parts of the book.

1.3 Organization and goal definition

It is important for the result of the transportation study to establish goals and objectives early in the process, since these will guide the evaluations towards conformity with the desire of the community ([893, 130]). Traditionally, as already mentioned, the main objective of the transportation study has been to evaluate alternative highway constructions for increased personal transport capacity ([862]). Other goals considered have also mainly been orientated toward traffic functional aspects, such as an increased safety, a saving of travel time, a reduction of operating costs, and an increase in efficiency and mobility. It is only during the last 25 years that environmental aspects and the transit alternative have been considered essential elements of the transportation study. See [984, 37] for a more detailed description of the goal setting.

The topology of the study area, the population distribution and many other socio-economic factors vary from study to study. The form of the study may therefore differ significantly among different countries and regions.

Studies may be of long-range type, in which case the most important questions to be answered deals with the density and configuration of the future transportation system. Short-term plans may include immediate-action programs for arterial improvements. The scale of the study may also differ; some plans include proposals for new facilities, such as parking, terminals, and transit lines, while others may describe highway locations with ramp connections pinpointed, or only deal with single corridors.

The personnel organization of the study can also have several different forms. The Transportation and Traffic Engineering Handbook [37, pp. 517–518] lists the following alternatives: A *centralized state staff* may be an existing agency or a new department incorporating the necessary multidisciplinary talents. The Chicago Area Transportation

Study [170], established as an *ad hoc* joint effort and responsive to a multiagency board, illustrates the use of a *semi-independent organization*. A *council of governments* is a study organization which may be created under a council made up of elected representatives of communities within the region. Established planning bodies for metropolitan regions are sometimes the organizations housing the transportation planning staff. In a *contract study organization* consultants under the supervision and monitoring of either a state representative or local study director perform all or some of the stages in the planning process. The procedure has been used extensively in the U.S. ([882]).

Regardless of the organization structure, an additional organization must be appointed to ensure that the activity of the planning staff agrees with the goals and objectives set up ([489]). This organization could comprise of the following committees ([37, pp. 518–520]):

The *policy committee* includes representatives of agencies participating financially in the study, as well as officials and executives of local and regional planning organizations. The function of this committee is to provide budget control, establish regulations for study personnel, supervise technical matters, establish objectives, assist in the plan development, and recommend a final plan ([692]). The *technical committee* includes technical personnel from agencies represented on the policy committee, and sometimes also from other local agencies. The function of this committee is to review and evaluate study methods, assist in developing alternative plans, perform technical evaluations, coordinate technical service contributions of participating agencies, and enlist the interest of local agencies in the planning process. The composition and function of the *citizens advisory committee* vary with the size of the study area, and the interest in the study objectives of the communications media. The committee provides the policy committee with information on public thinking, and can thus assist in the definition of planning goals and objectives, improve public understanding of the planning process, and build support for plan implementation.

1.4 Base year inventory

The inventory stage can be divided into four categories:

(1) (*Transportation facilities*) Here, the study area is defined, and divided into sectors, districts and zones. The physical network is represented by a graph, with streets and road sections represented by links (or arcs), and intersections and trip origins and destinations by nodes. The boundary of the study area, referred to as the *external cordon*, is chosen to approximate the commuter-shed associated with the urban centre. The zones represent aggregates of trips and socio-economic conditions; the choice of zones is very important, since the number of zones determines the complexity of the study, and the wrong choice of zone size and distribution would obscure a lot of the information in the data collected ([130, 264, 805]).[3] The number of zones ranges typically from 10 to 1000, and their sizes from a few blocks to several square kilometers. Whenever zones are small, their locations may be defined by single points in the network description, the so called *zone centroid nodes*.

Next, the characteristics of the existing transportation network are colleted; data includes measurements of traffic flows, speeds, travel times (or delays), link lengths, capacities, and the quality of transit service. There are many techniques for measuring these performance characteristics; some of the data required is recorded automatically by many traffic control systems, other information can be obtained from census data ([130]).

(2) (*Travel patterns*) Data relating to the present-day movement between zones is collected at this stage. Traffic pattern data is required for all combinations of external and internal movements. The data may be divided into trip mode and purpose. The goal of this data collecting is to estimate the number of trips made between zones within the study area, and the number of trips passing through, into or out of the area.

Movements through the area and external-internal movements are surveyed at the external cordon, and possibly at an internal cordon or screen line; this is done by manual or automatic counts. Internal-external movements are surveyed in the home-interview study and at the external cordon, while internal movements are surveyed by home-interview studies and, sometimes in addition to check, by an internal cordon or screen line survey.

The size of the sample to be interviewed depends on the total population of the area, the degree of accuracy required, and sometimes on the density of the population. The recommended sample sizes for home interviews are between 4 and 25 percent of the total population ([130]). For roadside interviews, the sample can be based on time or volume clustering, or could vary among classes of vehicles ([52, Chap. 4]).

In home-interviews, the information gathered includes address and size of household, job information, income, number of vehicles, and information about all journeys made in a previous time period, usually 24 hours. The interview procedure is outlined in Behr [52, Chap. 4]. Additional information is collected by interviews at commercial premises. For further reading, see, e.g., [467].

Roadside interviews are made on the external cordon to cover trips passing through or into the area. The questions asked depend on the purpose of the study, and the type of vehicle. Alternatives to direct interviews which delay travellers, are to ask drivers to complete and return prepaid postcards, to record registration numbers, or to place coded tags on the vehicles; see [52, Chap. 4] for further details.

More economical methods for estimating the existing origin-destination flow pattern are made possible by the automatic counts provided by many traffic control and signal setting systems, and by optimization models, with which possible origin-destination flow matrices may be derived from the counts (see, e.g., [999] and the references cited therein).

(3) (*Economic activity and population*) This information, together with that of land use, form the basis for developing relationships between the movements of goods and people and the distribution and intensity of land use. Data typically collected includes: historic population patterns (past distribution, migrations, density, and trends in growth), present population (distribution by area, density, average income, car ownership), employment trends and present employment, economic activity (patterns of investments in manufacturing, services, redevelopment, and other real estate), and transportation resources (outlays for regional transportation facilities).

(4) (*Land use*) The inclusion of land use studies into the transportation planning process was made during the 1950s; prior to this, the future demand for transportation was extrapolated, using simple growth factors, from counts on existing flows. Mitchell and Rapkin [681], however, demonstrated the close relationship between traffic flows and land use, and subsequent U.S. studies were orientated more towards the influence of land use activity on the generation of flows. While much of the attention still was focused on the traffic functional aspects, the consideration of future land uses was gradually incorporated during the 1960s. In Great Britain, a similar development

can be traced, following the Traffic in Towns Report [133].

Typical data collected include: historic development trends such as patterns of urbanization, topography and physical constraints on development, classified measures of acres of land vacant or in urban use, location of major travel generators, identification of social neighbourhood and community boundaries, nature of existing land use controls, and identification of redevelopment areas ([650, 93]).

This part of the planning process has historically been the most costly one; as much as 49 percent of the total cost has been reported ([496]). With more advanced techniques for forecasting, and as knowledge of the urban transportation system improves, the availability of data from external sources or from automatic measurements increases and more alternatives are evaluated, and thus, the portion of the total cost decreases ([37, p. 521]).

1.5 Model analysis

In this step, relationships are sought between the land use and traffic characteristics of the present-day situation. These relationships are then used to estimate the future traffic situation, given the future estimated land use and proposed network facilities. An underlying assumption in this process is, of course, that these relationships will not alter significantly in the future.

These relationships are usually derived and calibrated through considering a sequence of models, rather than by a single analysis. The basic qualities of these models and their solution are outlined below.

1.5.1 Trip generation

The purpose of the trip generation step is to estimate the number of trips (typically per day) that originates or terminates in each of the zones previously defined, as a function of land use, socio-economic and locational characteristics of the zones. The most important dependent variables used are trip purpose, family income, vehicle ownership, land use activity at the zones defining the trip origin and destination, length and mode of trip, and time of day (see, e.g., [650, 837, 130, 805]).

The first transportation studies employed simple growth rates to estimate future trip generation ([249]). Subsequent studies analyzed the correlation between the above mentioned variables, using multiple regression analysis ([668, 265, 266, 130, 805]).

If it is assumed that trip generation characteristics remain stable with time, then future estimates can be made using the regression equations obtained. However, some modifications are usually necessary to reflect the estimated future conditions.

There are several sources of error in this use of regression analysis ([130, 264, 805]), and the underlying assumptions of independence and continuity among the variables in the regression analysis are not entirely correct. Because of the difficulties associated with the regression technique, household based disaggregate models, usually referred to as *category analyses* ([993, 754]), have been developed and used. The underlying assumption here is that the household is the fundamental unit in the trip generation process, and that the journeys generated depend on household characteristics and location. The main advantage of this method of analysis is that household categories may be estimated from

[3]It has been reported ([344]) that as much as 80 percent of the activity occurred *within* the zones defined in one instance of a traffic study.

census data using known relationships, such as distributions of income, car ownership and family structure; large scale home interviews can thus be avoided, resulting in a large saving compared to the regression approach. Furthermore, the analysis is computationally cheaper, and the disaggregated information may reflect individual behaviour more realistically than the zonal aggregated information. A disadvantage of this technique is that the distributions used may not be valid in the future planning period.

Domencich and McFadden [264, Chap. 2] argue that since transportation facilities do not enter the trip generation step trip frequency is independent of changes in the transportation system, making the trip generation both non-behavioural and non-causal, and also non-policy orientated.

For further reading on trip generation models and methods, see [922, 265, 93, 266].

1.5.2 Trip distribution

The purpose of this model is to estimate the number of trips performed from an origin zone to a destination zone, given aggregated trip numbers from the previous step.

The traditional techniques used for estimating the future origin-destination (O-D) flows can be divided into two categories: growth factor (or analogy) methods, and inter-area travel (or synthetic) methods.

Growth factor methods

The philosophy behind growth factor methods is that present travel patterns may be projected into the future on the basis of zonal growth rates, which may be obtained from the productions and attractions assessed in the previous stage; the future O-D flows are calculated by simply multiplying the present-day pattern by the growth rates.

In other words, the future number of trips, d_{pq}^{new}, from zone $p \in \mathcal{P}$ to zone $q \in \mathcal{Q}$, is calculated from the present number of trips, d_{pq}^{old}, $(p,q) \in \mathcal{C}$, through the general formula

$$d_{pq}^{\text{new}} = g_{pq}(\mathbf{d}^{\text{old}}, \mathbf{d}^{\text{est}}) \cdot d_{pq}^{\text{old}}, \tag{1.1}$$

where $\mathbf{d}^{\text{est}}$ is the vector of the estimated number of trips, d_p^{est}, generated by zone p, and d_q^{est}, attracted by zone q, given by the trip generation part of the process. The function $\mathbf{g} : \Re_+^{|\mathcal{C}|} \times \Re_+^{|\mathcal{P}|+|\mathcal{Q}|} \mapsto \Re_+^{|\mathcal{C}|}$ defines the growth factors for the O-D pairs. This factor may be a single factor, or a combination of several factors, and it may be the same for all O-D pairs, or vary with the zone. The above formula may give results that are inconsistent with the estimated trip totals, $\mathbf{d}^{\text{est}}$; in this case, an iterative procedure must be adopted, whereby the growth factors are modified so as to achieve balanced equations. (As a result, growth factor methods produce a new trip matrix from the old one by multiplying the rows and columns with factors that satisfy row and column total constraints. The methods here described have therefore become known as *balancing methods.*)

In chronological order, the following methods have been developed. In the following, let $E_p = d_p^{\text{est}} / \sum_{q \in \mathcal{Q}} d_{pq}^{\text{old}}$ denote the ratio of the estimated number of trips generated in zone $p \in \mathcal{P}$ to the present number of trips originating in zone p, and E_q the corresponding ratio for attracted trips. Also, let $E = (\sum_{p \in \mathcal{P}} d_p^{\text{est}}) / (\sum_{(p,q) \in \mathcal{C}} d_{pq}^{\text{old}})$ denote the ratio of the total number of estimated trips generated to the total number of present trips.

(1) (*Uniform factor*) A single factor is calculated for the urban area, and multiplied to the existing flows. Using (1.1), the factor may be written as

$$g_{pq} = E, \qquad \forall (p,q) \in \mathcal{C}.$$

This technique fails to recognize any differential rates of development in different parts of the study area.

(2) (*Average factor*) This is the first attempt to take into account differential growth rates; the growth rate is an average of the growth rates defined for the origin and destination zones, i.e.,

$$g_{pq} = \frac{1}{2}(E_p + E_q), \qquad \forall (p,q) \in \mathcal{C}.$$

The values calculated by this formula will probably not give results that are consistent with the estimated number of trips, i.e., for some $p \in \mathcal{P}$, $\sum_{q \in \mathcal{Q}} d_{pq}^{\text{new}} \neq d_p^{\text{est}}$. An iterative process is then utilized, in which the zonal growth rates are adjusted until a balance is achieved. Such a procedure is outlined in [650]. Neither in this procedure is the differences in growth in different areas well accounted for ([726]).

(3) (*Fratar* [378]) In this method, the distribution of future vehicle trips is proportional to the present trip distribution, modified by the growth factor of the zone to which the trips are attracted. Mathematically, the distribution can be written as:

$$g_{pq} = \frac{E_q d_p^{\text{est}}}{\sum_{i \in \mathcal{Q}} E_i d_{pi}^{\text{old}}}, \qquad \forall (p,q) \in \mathcal{C}.$$

(Other similar formulas have been given.) Also in this case, an iterative balancing procedure is applied to obtain consistent output. From this method, many simplified schemes and extensions have been proposed and used (e.g., [87, 121, 395]).

The advantages of growth factor methods are that they are easy to apply, they are flexible, and they can be used to distribute trips by purpose, mode and time of day, by defining different growth factors for each zone. Furthermore, when applied to areas where conditions are stable over the study period, the results haven been found to be quite satisfactory. However, when applied to a study with significant changes in land use, such as proposals of new transportation facilities, and where travel costs change with time, this technique gives unreliable estimates of future trips.

Synthetic methods

When the shortcomings of the growth factor methods were identified, work concentrated on the development of alternative methods. The most successful alternatives, the *synthetic methods*, were based on the assumptions that before travel patterns can be predicted, the underlying causes of movement must be understood, and that the causal relationship giving rise to movement patterns can best be understood if they are considered to be similar to laws of physics.

Three different synthetic methods can be identified: the gravity model, the opportunities models, and the electrostatic model.

(1) (*Gravity model*) The gravity model is the most widely used synthetic model; it is simple to understand and use, and is well documented. The term gravity stems from the assumption that the number of trips performed between an origin and a destination zone are directly proportional to the relative attraction of each zone and inversely proportional to some function of the spatial separation between the two

zones ([954]). The origin of gravity models can therefore be said to be the works on gravity by Newton [712].

The gravity based distribution formula can be written as

$$d_{pq} = A_p O_p B_q D_q f(\pi_{pq}), \qquad \forall (p,q) \in \mathcal{C}, \tag{1.2}$$

where O_p and D_q denote the total number of trips originating in zone p and terminating in zone q, respectively, f is a deterrence function, monotonically decreasing with a generalized travel cost π_{pq} between the zones, and where A_p and B_q denote proportionality constants, which are determined such that the marginal total constraints of flow,

$$\sum_{q \in \mathcal{Q}} d_{pq} = O_p, \qquad \forall p \in \mathcal{P} \tag{1.3a}$$

$$\sum_{p \in \mathcal{P}} d_{pq} = D_q, \qquad \forall q \in \mathcal{Q}, \tag{1.3b}$$

are satisfied.

The doubly constrained negative exponential gravity model is given by letting $f(\pi_{pq}) = \mathrm{e}^{-\gamma \pi_{pq}}$, where γ is a positive parameter, reflecting the influence of cost on the number of trips made. If we also let $r_p = A_p O_p$ and $s_q = B_q D_q$ denote the balancing factors, then the trip matrix can be written as

$$d_{pq} = r_p s_q \mathrm{e}^{-\gamma \pi_{pq}}, \qquad \forall (p,q) \in \mathcal{C}. \tag{1.4}$$

The application of the gravity model in science has a very long history. The Model (1.4) has a sound theoretical basis, and may be derived from gravity ([712, 143, 613]), maximum likelihood ([980]), entropy maximization ([694, 867, 977, 979]), maximum utility ([51, 660]), minimum discrimination information ([826, 574, 861]), cost minimization ([298, 299]), minimum Minkowski norm ([302]), or efficiency ([853, 854]) arguments. For further reading and applications, see [518, 982, 920, 907, 437, 304, 753, 130, 805, 509, 299, 624, 302].

Methods employed for the solution of the gravity models are based on iterative refinements of the proportionality constants A_p and B_q, with the objective of balancing the Equations (1.3). Balancing methods include those of Kruithof [569], Fratar [378], Furness [395], and Bregman [116, 117, 118], but also Newton-type methods ([775, 21]); for overviews of balancing methods, see [36, 157, 578, 817]. The advanced methods for growth factor models may be seen as special cases, derived from the class of balancing methods.

The gravity model has been criticized for its behavioural implications; the analogue with physical systems can not be taken for granted when dealing with human systems ([967, 433, 49, 459, 460]).

(2) (*The opportunities models*) The opportunities models introduce the theory of probability as the foundation of the distribution ([880]), and were developed within the Chicago, Pittsburgh and Penn-Jersey studies ([170, 982, 903]). The assumption basic to the opportunities models is that all trips will want to remain as short as possible, lengthening only if they fail to find an acceptable destination at a shorter distance. The model may be derived from kinetic gas theory, or from the theory of radioactive decay. Letting $D_{pq} = \sum_{l \in \mathcal{Q}: \pi_{pl} < \pi_{pq}} D_l$ denote the total number of trips that are shorter

in distance from zone p than zone q, and L_p the probability density of destination acceptability, then

$$d_{pq} = O_p(\mathrm{e}^{-L_p D_{pq}} - \mathrm{e}^{-L_p(D_{pq}+D_q)}), \qquad \forall (p,q) \in \mathcal{C}.$$

The models are simple to use, and need less input data than the gravity models. This accuracy is, however, reported to be slightly lower ([471, 597]). For further reading on the opportunities models, see [130].

(3) (*The electrostatic model*) Howe [506, 507] developed this model from Coulomb's law of electrostatic force; considering tripmakers as electrons, the attraction to the (positively charged) destination zones is assumed to be proportional to the number of persons employed in the respective zone.

The simplicity of the model is its major merit; it is similar to the early developed gravity models, and the solution principles proposed are balancing methods of the same type as those used for the gravity models. Included in its disadvantages is however its inability to model external flows. Lawson and Dearinger [597] evaluate it against other trip distribution models; although it was found inexpensive to apply (existing movement data is unnecessary), it was found to be less accurate than the gravity model.

Other models of trip distribution include multiple regression ([729]) and linear programming ([93]).

The trip distribution models have been criticized for their simplicity ([114]), which, for instance, means that the model is non-policy orientated ([264, Chap. 2]).

1.5.3 Modal split

Modal split divides the total number of person trips into different modes of travel, based on relative measures of competitiveness.

Modal split models can be classified into two categories, *trip end* models, which are applied *before* the trip distribution stage, and *trip interchange* models, which are applied *after* trip distribution. (The above description of the transportation planning process is based on the use of trip interchange models.)

The diversity of modes is usually ignored or dealt with by combining all the available modes into two dichotomous modes, transit and auto.

The competitiveness measures are derived from an analysis of three basic sets of factors:

(1) (*Journey characteristics*) Factors included are trip length and journey cost, purpose of the trip, and time of day of tripmaking. The two most important factors in this category are journey length and trip purpose.

Journey length can be measured in many ways, the most simple being the bee-line distance between the origin and destination. A more accurate measure may be derived from the travel time on the route most heavily used, for the different modes of transport. It is important to include all parts of the journey (defining a *door-to-door journey*), i.e., even the parts that do not include the use of a vehicle. (This excess travel time includes walking to and from the vehicle, waiting for a vehicle, and changing from one vehicle to another.) The travel-time ratio between competing modes can also be used as a measure of journey length. This measurement used in isolation may, however, obscure large absolute differences in journey time by

competing modes ([130, p. 171]), and should therefore be used in conjunction with some other measure.

Experience has shown that there is a relationship between the numbers using public transport and the purpose of the journey. Home-based journeys often give rise to more public transport journeys than non-home based journeys, whilst home-based school and work journeys have a higher rate of public transport usage than home-based shopping journeys.

(2) (*Traveller characteristics*) The most significant factors in this category are concerned with the socio-economic characteristics of the households making the journeys, and include variables such as income, car ownership, family size and structure, density of residential development, the type of job undertaken, and the location of workplace. These factors are certainly highly interrelated [e.g., car ownership is a function of income ([977, pp. 119–170])], and can therefore not be analyzed in isolation.

It is very difficult to accurately measure the total income at the zones, and substitutes such as car ownership, density of residential development and type of dwelling unit are instead used to indicate the level of income.

It has been found that as net residential density increases, the demand for public transport decreases. Schwartz [820] found in the Pittsburgh Area Transportation Study that school journeys by public transport are inversely related to net residential density, whilst other journeys by public transport are directly related to it. (The inverse relationship between school journeys by public transport and net residential density was attributed to the greater numbers walking to school in the more densely developed areas.) This relationship may be explained by the fact that it is difficult to provide an adequate and economic public transport service in low density areas. In addition, low density areas tend to be occupied by the middle and higher income groups with the result that levels of car ownership are higher, and consequently the demand for public transport lower. In contrast, high density areas can be economically and adequately served by public transport, largely because they were developed in conjunction with the public transport system.

Other socio-economic factors that are used for determining the modal split include family size, the age-sex structure of the family, the proportion of married women in the labour force, the type of property occupied and the type of employment of the head of the household. For examples of the correlation between public transport use and socio-economic variables, see [650, 495, 446].

(3) (*Transportation system characteristics*) The most significant factors in this category are concerned with the travel time and out-of-the-pocket expenses for the journeys, and qualitative measures of the level of service of the competing modes. These measures of competitiveness are usually given as ratios of measures of competing mode alternatives.

Travel times most often express a time ratio of door-to-door travel time by public transport divided by the door-to-door travel time by private vehicle. Absolute differences have also been used, and have been found to sometimes yield more reliable measures ([771]).

Travel cost measures include out-of-pocket expenses (fares for public transport, and fuel and parking costs for private vehicles) only; private vehicle costs such as road tax and insurance are ignored since studies have found that these costs do not influence the journeys made ([130, p. 177]).

Relative levels of service are affected by a large number of factors, the majority

of which are subjective and difficult to quantify, such as comfort, convenience, and ease of changing between modes. Quantifiable measures include time spent outside the vehicle during a journey, e.g., walking, waiting, and parking delay.

In some trip end modal split models accessibility indices have been used as a measurement of the quality of service provided by the different modes. These indices measure the ease with which activity in one area can be reached from a particular zone; one possible definition of an accessibility index for a given zone is the sum of the trip attractions times the *friction factor* for the zonal interchange, where the friction factor is calculated as the reciprocal of the door-to-door travel time raised to some power which varies with the travel time ([961]). Other accessibility indices are defined by the number of routes serving the zone, the frequence of service, and the area of the zone (see [130, Chapter 6]).

The earliest form of modal split models used public transport diversion curves to relate public transport use to relative travel times. The drawback of the simple diversion curve is that it completely ignores the characteristics of the person making the journey.

In trip end modal split models, the total person trip productions are allocated to public transport for each journey purpose considered in the model (typically, home-based work, shopping, social/recreational, and miscellaneous trips), by considering the attractiveness of the public transport system as measured by the variables considered to influence the modal split in the area under examination. The technique most often used is multiple linear regression (e.g., [7, 977]).

Journeys made by private vehicles are derived by subtracting the estimated public transport trip productions from the total person trip production estimate.

Future trip attractions by public transport are usually estimated by multiple linear regression techniques, using, for example, variables such as the location of the destination zone, the employment level in the zone, and the characteristics associated with the use of public transport in that zone. Category analyses have also been used in some studies.

Private vehicle person productions and attractions are converted to vehicle productions and attractions by introducing vehicle occupancy rates. The distribution of the estimated public and private transport is then calculated using, for instance, a gravity model.

More recently *generalized costs* have been introduced in the modal split models. Generalized costs were originally developed by Wilson [978] for use in gravity models, and are linear functions of travel time, distance, excess travel time, and terminal cost. See Quarmby [771] for an empirical justification of the use of generalized costs in modal split models.

Trip interchange modal split models allocate journeys to different modes after the total person movements between pairs of zones have been distributed. A standard trip interchange model uses multiple linear regression techniques to determine zone to zone public transport travels and private trip interchanges, often in conjunction with a gravity model. Variables used represent zone to zone based characteristics of the persons making the journey, the destination of the journey, and the transportation system, and include the relative door-to-door travel time, the income, the net residential density, and the employment density at the destination. The private vehicle trips are derived by subtracting the public transport trips, and the vehicle interchanges between zones are determined by dividing the total number of personal trips between two zones by appropriate car occupancy factors.

A statistical technique, *discriminant analysis*, has been used to predict modal split ([787]). An underlying assumption of this model is that individuals choose a mode of travel based on the (conscious or subconscious) evaluation and weighting of advantages

and disadvantages of the different modes, the factors being related to aspects such as travel time, cost, comfort, and reliability. In the discriminant analysis, an estimate of the most probable values of the weights of the importance of different factors are derived.

The advantages of the trip end approach when compared with the trip exchange approach, is that the former is capable of making separate distributions between the zones; this is considered desirable because of the frequently differing lengths of journey by car and public transport. In this way, more properties of the transportation system can be taken into account when making the trip distribution. Another important difference between the two approaches is the different level of detail present in the models. In the trip end models, characteristics of the transportation system are area wide averages, while they are more precise in the trip interchange models. The higher precision of the latter models should, however, be weighed against the computational burden associated with the much larger number of splits required to determine modal choice for the area under study.

Both approaches to modal split have been criticized on the grounds that present-day levels of service are used, which have often meant that the private vehicle is favoured to the public transport alternative. They have also been criticized for the primitive estimation techniques used, and the way in which the components of the travel time and cost are aggregated ([264, Chap. 2]).

For further reading on modal split models and methods, see [327, 921, 961, 922, 130].

1.5.4 Traffic assignment

Traffic assignment is the part of the process which allocates a given set of trip interchanges to a specific transport network or system. As input the traffic assignment process requires a complete description of the proposed or existing transportation system and a matrix of interzonal trip movements. The output of the process differs with the sophistication of the assignment procedure, but always includes an estimate of the traffic volumes and the corresponding travel times or costs on each link of the transportation system; some assignment techniques also include directional turning movements at intersections and route flows.

The purposes of traffic assignment as part of the transportation planning process are to assess the deficiencies in the existing transportation system by assigning estimated future trips to the existing system, to evaluate the effects of limited improvements and extensions to the existing transportation system by assigning estimated future trips to the network which includes these improvements, to develop construction priorities by assigning estimated future trips for intermediate years to the transportation system proposed for those years, to test alternative transportation system proposals by systematic and readily repeatable procedures, and to provide design hour volumes and turning movements ([130, pp. 145–146]). Modern uses of traffic assignment extend this list with purposes of much shorter time horizons, even real-time use. Here, however, we shall concentrate on the assignment techniques adopted and developed in transportation studies.

Early heuristics

The basic concepts of traffic assignment evolved in the early and middle 1940s ([863]), and in conjunction with the first origin-destination surveys conducted in over 100 cities in the United States shortly after the end of World War II. (With the development of origin-destination studies, vehicular movements between zones became available for use in

determining traffic loads on proposed new routes.) The early work in assignment consisted primarily of estimating the diversion of traffic from existing roads to new, improved, high-speed arterials or freeways, and was based on travel time and cost savings. The first assignments made assumed that the travel time and cost were independent of the flows on the links (a highly unnatural assumption), and, consequently, the results amounted to the proposed road being used either by *all* vehicles between a pair of origin and destination (in the case where the travel time or cost between the origin and destination was found to be less than that of any alternative route between the origin and destination on the existing transportation system), or by *no* vehicle in the origin-destination pair (e.g., [138]). This technique is commonly known as the *all-or-nothing* technique. At a very early date it was found to give unrealistic results, not only because it fails to recognize that travel times and costs increase with the flows on the links, but mainly because of the fact that all travellers are allocated to routes based on a single average characteristic ([250, p. 80]). Empirical studies were later undertaken in the U.S. in an attempt to relate the choice of route to time and distance factors, and as a result the American Association of State Highway Officials developed a standard traffic *diversion curve* (e.g., [250, 691]) as the recommended policy for determining the future use of urban highways.

The curves employed were based on data obtained from observations at some other location with two similar facilities, and estimated the portion of the flow on the traditional route to be transferred to the hypothetical one. Different parameters have been employed in these curve formulas, such as time and distance saved by using the proposed expressway and the ratio of travel times for the two available routes. The assignment was then usually made in such a way that the proposed freeways and the existing routes were assigned flows in proportion to their travel times. However, this technique was only capable of dealing with a single expressway with existing parallel routes (*corridor studies*). The reason for this is that if more than one expressway-type facility is present then the travel time on these alternatives is highly interdependent. Also, the travel time on a link is assumed to be independent of the volume of traffic on the link, and therefore does not take congestion into account. For further reading on this technique, which is also known as *two-route assignment* and *proportional assignment*, and the early development of traffic assignment techniques, see [127, 908, 250, 691, 170, 667, 650, 983, 919, 649, 663, 733].

At the 31st Annual Meeting of the Highway Research Board, Washington, D. C. in 1952, Campbell [139] summarized the techniques of traffic assignment as follows:

> Traffic assignment is fundamental to the justification of a proposed highway facility and to its structural and geometrical design, to spotting points for access, and for advance planning of traffic regulation and control measures. As yet, traffic assignment is considered to be more of an art than a science...

Accordingly, he stressed the need to place traffic assignment on a scientific foundation.

Consequently, in the early 1950s considerable difficulty was experienced in assessing the driver's choice of route to complete his/her interzonal trip, and route-choice decisions were often made manually and arbitrarily based on the engineer's knowledge and judgement and an assessment of travel time, distance and user cost. Since the detailed analysis of present and future urban area transportation, to be performed by the studies, required more logical and accurate assignment procedures and as the very large number of operations to be performed in the assignment phase necessitated the use of automatic data processing machines, the transportation planning community were in desperate need of a more efficient method for assigning traffic to an urban network.

Around 1957 a major (or possibly *the*) breakthrough occurred in network assignment. In the operations research community efficient algorithms for the *shortest route problem*,

i.e., the problem of finding a route of minimal travel time (or cost or distance) through a network with fixed travel times (or costs or distances) on the links, were discovered (e.g., [367, 223, 685, 56, 258, 224]). (This problem obviously is a main ingredient in an assignment program.) Simultaneously, the staff of the Chicago Area Transportation Study was looking for a computer program to assign traffic to a large urban road network, and contracted the Armour Research Foundation to develop it. Their investigation led to the development of a computer program for finding the minimum time routes through a network, based on Moore's [685] algorithm and average travel times and speeds on the links, and consequently to the first reported fully computer aided assignment ([170]). In 1960, further research (by the General Electric Computer Department in collaboration with the District of Columbia) led to the development of an assignment program capable of prohibiting selected turns in the calculation of the minimum path.

The resulting assignment is an *all-or-nothing*, or *desire assignment* ([128, 137, 151]), since all travellers are assigned to the routes which are the cheapest (and hence the *desired* ones), and the more expensive routes receive a zero flow. The main advantage of this approach is that the calculation is made in one single step, i.e., no iterative assignment is made. Therefore, the assignment procedure is economical, and the result is easy to analyze. There are, however, serious drawbacks to this methodology. Since all traffic between two zones is assigned to a single route, the assignment leads to unrealistic traffic volumes on streets with limited capacities, as was noticed already in the results of the first assignment made ([170]). Furthermore, the technique is unstable in the sense that small changes in travel times used may cause a significant change in the resulting flows. As a consequence of these drawbacks, the all-or-nothing assignment method has been rejected by the analysts.

Accidentally (or maybe not!) some other events which enabled a rapid progress of assignment modelling and methodology [or, in some cases, would have, had they been recognized at the time ([102, 104])] took place around this time. In 1952, J. G. Wardrop of the Road Research Laboratory published a paper ([958]) on two principles of flow distribution in a road network: the *user equilibrium* principle, which is based on the assumption that all travellers are minimizing their own travel cost, and the *system optimum* principle, of which the underlying assumption is that the travellers choose their routes so as to minimize the total travel time in the transportation system. These two principles are by far the most popular behavioural principles in assignment models. Although these two principles had been known and used within the academic theoretical economics community for at least 30 years, these principles are often attributed to Wardrop, and therefore referred to as the two Wardrop principles; see Section 2.1 for further discussions.

In 1956, M. J. Beckmann and colleagues published the seminal book "Studies in the Economics of Transportation" ([47]), in which mathematical models for the traffic assignment problem were analyzed. By using nonlinear optimization theory, the two Wardrop principles were shown to correspond to the solution of convex nonlinear optimization problems with linear (network) constraints. (See Section 2.2.) Similar optimization formulations had, however, been developed earlier for closely related problems in the analysis of electrical networks (e.g., [275]); see Section 2.6.3.

Another important event in 1956 was that M. Frank and P. Wolfe published a paper ([377]) on an iterative algorithm for the solution of convex, quadratic optimization problems. When applied to the traffic assignment models of Beckmann *et al.*, the method alternates between an all-or-nothing assignment, based on the travel times at the present flow, and the minimization of the objective function of the optimization problem on the line segment between the vector of the present flow and the all-or-nothing solution. Today

this algorithm is a standard code in transportation planning packages for the solution of traffic assignment problems, but it was not applied to this use until in the late 1960s.

In December 1959, the First Symposium on the Theory of Traffic Flow was held at the General Motors Research Laboratories, Warren, MI. This symposium was set up with the objective of bringing together active researchers from different fields of science and technology to enable cross-fertilization and stimulate new ideas for future research activities.

A few years later (in 1962), due to the increased academic interest and importance of the transportation research field, the Transportation Science Section of the Operations Research Society of America (ORSA) was formed, and academic journals were soon issued.

After these very important steps had been taken, improved assignment techniques were developed in many transportation studies. The unrealistic results of all-or-nothing assignments naturally lead to algorithms, where travel times were modified within the procedure, thereby taking more account of congestion effects. The need for relating travel times and speeds to traffic volumes when assigning traffic to networks resulted in the development of *link performance functions* (also known as *volume delay formulas* and *travel time functions*).

Link performance functions

As a result of growing traffic volumes, the speed on a link tends to decrease, first slowly but as the queueing effects become more significant, the average speed on the link decreases more rapidly, until the queueing has developed into a jamming situation, where very little flow can be observed on the link. In the analysis of traffic systems, average travel times are therefore usually modelled as positive, nonlinear, and strictly increasing functions of flow. Parameters in the formulas often include practical traffic volume capacities, and sometimes also aggregate measures of factors, other than travel time, that influence tripmakers in the route selection process.

Different transportation studies developed their own travel time formulas. The approaches used to define these functions were of two kinds: in the first approach, mathematical functions were proposed in advance, for the sake of simplicity. Various parameters were then calculated through different measurements of traffic and road conditions and speed-to-volume ratios developed by traffic engineers. In the other approach, the formulas were developed from studies of speed and travel times related to network characteristics, such as queueing at intersections, based on queueing theory. The basic parameters of a link performance function, relating travel time, t_a, on link a, to the flow, f_a, on the link, is the *free-flow travel time*, t_a^0, which is a measure of the travel time at zero flow, and the *practical capacity* of the link, c_a, which is a measure of the flow from which the travel time will increase very rapidly if the flow is further increased. Although these formulas were developed for studies of highway systems, they are still often used today for studies of city streets in urban areas. For a survey of the various link performance functions used in transportation studies, we refer to Branston [115]. In Table 1.1 we give a list of link performance functions developed during the 1960s. In Figure 1.1 a graphical example of a typical link performance function is given. Here, β_a and m_a are positive parameters.

Other formulas are found in [230, 889, 15, 16, 890, 866]. Some empirical and experimental work on the subject is found in [996, 109]. Boyce *et al.* [109] found that travel time functions with asymptotes, such as the last formula of Table 1.1, empirically lead to unrealistically high travel times and devious rerouting of trips; the resulting assignments should therefore be used with extreme caution in any planning application.

Travel time formula	Reference
$t_a^0 \cdot \mathrm{e}^{(f_a/c_a-1)}$	[858]
$t_a^0 \cdot 2^{(f_a/c_a-1)}$	[815]
$t_a^0 \cdot (1+0.15(f_a/c_a)^{m_a})$	[919]
$t_a^0 + \log(c_a) - \log(c_a - f_a)$	[690]
$\beta_a - c_a(t_a^0 - \beta_a)/(f_a - c_a)$	[690]

Table 1.1: Travel time formulas

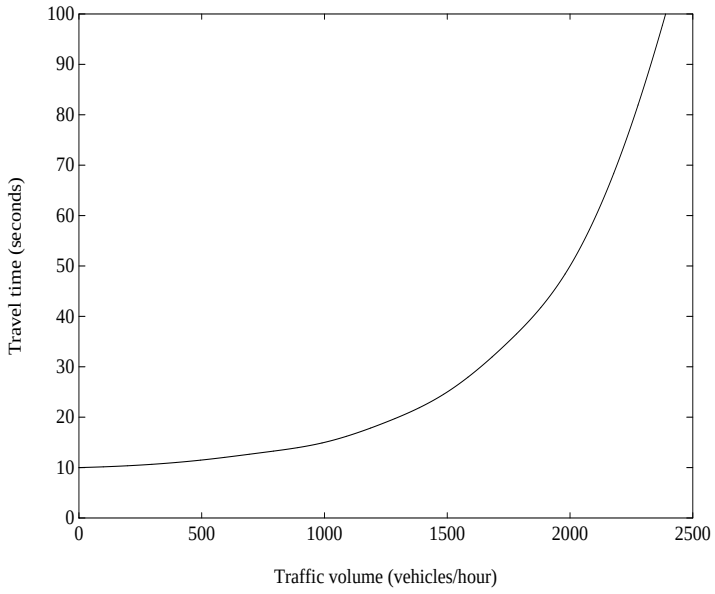

Figure 1.1: A typical link performance function

Many factors other than travel time influence the choice of route for different trip purposes; the use of travel time functions such as those given in Figure 1.1 can therefore only give very rough estimates of the true route choice behaviour of travellers. In empirical studies (e.g., [908, 672, 673, 60, 568, 773, 883, 27, 510, 730, 731, 625, 992, 99, 875]), factors such as distance, frequency of traffic signals, strain, use of petrol, presence of tolls, safety, presence of pleasant scenery, etc., have been reported to have an effect on the drivers' attitudes to the type of route sought. A rather interesting work in this context is that of Jansen [521], whose study of a small area of the San Fransisco urban network shows the remarkable result that the most popular route is the second most expensive one.

Capacity-restraint heuristics

Through the use of the volume-delay formulas, it now became possible to introduce iterative procedures, in which new assignments could be made based on adjusted travel times according to the traffic volumes calculated in an earlier assignment, thereby taking congestion effects and capacities of traffic into account. In a general *capacity-restraint* method, travel times are calculated based on the flow assigned to the network at iteration k. An all-or-nothing assignment is then made based on the fixed costs, and the next assignment is calculated by combining the current flows with the all-or-nothing solution.

The stopping criteria used vary with the different techniques of this kind.

The first capacity-restraint technique to be applied in transportation studies was the *quantal loading* procedure, developed in the Chicago Area Transportation Study [170, p. 108] (see also [151, 815]). The method operates on a zonal basis. One origin is selected randomly, and the shortest routes are calculated to all destinations. All trips originated at the node of current interest are assigned to the routes defined by the shortest route tree, after which the current travel times are updated according to the accumulated flows, and the travel time formulas used. (The Chicago Area Transportation Study applied a simple volume/capacity ratio to adjust travel times.) The process is then repeated with the selection of one of the remaining origins, and the algorithm is terminated when all trips have been assigned to the network. The algorithm differs from the all-or-nothing assignment technique only in the difference in the travel costs used for the assignment of different zonal flows, and therefore requires essentially the same computing time and yields comparable results, i.e., all trips are assigned to a single route for each O-D pair, and the method is subject to the same sensitivity of travel times. Furthermore, the result of the assignment procedure is sensitive to the order in which the origins are chosen, since the last minimum cost routes to be calculated are based on much more congested flows then the first ones ([863]). The fact that the shortest route calculations in the quantal loading technique takes some account of congestion effects has resulted in its use in place of all-or-nothing assignments in subproblems of convergent algorithms based on shortest route calculations ([941, 24]), and in generating starting solutions for use in convergent methods ([24]).

The most simple extension of this technique is the *iterated all-or-nothing assignment* procedure. The first known application of this method is in the Bureau of Public Roads program package ([919]; see also [490, 869, 607]), and it is implemented in the Federal Highway Administration program package PLANPAC/BACKPAC ([318, 319, 320]). The method proceeds as follows: starting with a feasible flow (for instance the flow obtained from an all-or-nothing assignment based on free-flow travel times), link costs are calculated for the flow observed by applying the volume-delay formulas. The whole flow is reassigned to the new routes, and the whole process continues iteratively, until either a specified number of steps have been performed, or the travel times at the beginning of an assignment approximately equal the travel times obtained from the volume-delay formulas. The use of this technique in the search for an equilibrium flow pattern assumes, in effect, that, based on the current travel times, all tripmakers choose the same minimum-cost route to their destination.

Reports on the performance of this method indicate that the flows do not converge. This is not surprising, since the requirement that only one route is used for the interzonal trip volume is unrealistic in congested networks. The shortest route problem is sensitive to input, as discussed above, and the consecutive assignments are therefore subject to wide oscillations ([932, 935]), making it difficult to decide when the process should be stopped. Furthermore, the assumption that the tripmakers all use the same route between two zones makes this method unrealistic in reflecting human behaviour correctly.

To remedy the oscillating behaviour of the repeated all-or-nothing assignment technique, and to allow more than one route to be used in each O-D pair, procedures were next developed in which only (fixed) portions of the total demand are transferred to the new all-or-nothing solutions; given a feasible flow, $\mathbf{f}^k$, in iteration k, and an all-or-nothing solution, $\mathbf{y}^k$, the assignment in iteration $k+1$ is given by

$$\mathbf{f}^{k+1} = (1 - l_k)\mathbf{f}^k + l_k\mathbf{y}^k, \tag{1.5}$$

where $l_k > 0$ denotes the portion of the current solution being reassigned. This process may seem natural from a behavioural viewpoint; when the tripmakers reconsider their route choices according to the prevailing traffic conditions, not all of them will adjust their route choice.

In the first such method presented, the number of iterations (i.e., the number of shortest route calculations) to be performed were fixed *a priori*, and instead of using current travel times, the old travel times were combined using the same weights l_k as for the flows (a process known as *smoothing*). In the PLANPAC/BACKPAC procedure CAPRES ([318, 319, 320]; see also [919, Chapter V]), the number of steps is specified to four, and $l_k = 1/4$, for all k. (The consequence of this choice of weights is, of course, that four routes at most will receive any flow in each O-D pair, and if, in an O-D pair, the four routes are different, the demand will be divided evenly among them.)

Compared to the iterated all-or-nothing assignment, the resulting flow is more accurate since several routes are used between each origin and destination. However, that the travel times are calculated at flows that cannot be observed in the network seems unnatural, and making the proper choice of weighted mean travel times is difficult. Furthermore, the specification in advance of the number of steps to perform makes it impossible to know how good the final solution is.

To yield more reasonable results, the above algorithm was implemented without prespecifying the number of iterations and using actual travel times in the calculations of the all-or-nothing solutions (see Smock [858, 859] and Almond [17, 18, 19]). It was then found that when the number of iterations grew, a fixed weight l_k would yield oscillations in the flows; that is to say, the same phenomenon as in the iterated all-or-nothing assignment technique would occur, although on a smaller scale. This fact had actually been recognized already by Beckmann *et al.* [47, Sec. 3.3], when discussing the stability of an equilibrium state. They analyze the method through a small numerical example, and conclude that the portion of flow transferred to the new routes, termed the *responsive fraction*, should decrease in order to avoid oscillations.

Smock [858, 859] does not present the method mathematically, but refers to the flow adjustment as a process, where the flows are divided *evenly* among the routes hitherto computed, i.e., he uses the responsive fraction $l_k = 1/k$ in iteration k. In his tests (performed on a projected trip table for the proposed 1980 network of Flint, MI) he observes that the number of iterations needed is relatively small for networks with high capacities, and larger for more congested networks. Smock's iterative scheme is also tested by Overgaard [732, 733]. (A similar algorithm, operating in the space of route flows, is given by Fisk [336].)

The basic algorithm of Almond [17, 18, 19] is similar to the one presented by Beckmann *et al.* However, his work can be seen as a development from their method in several respects. Almond demonstrates graphically for a few small examples how oscillation effects evolve when the basic method is employed, and that they tend to be more significant for more congested networks. The natural conclusion is the same as stated in [47, Sec. 3.3.2], namely that the shifting percentage must decrease as the number of iterations increase.

In general, heuristic methods employed for traffic assignment are defined by predetermined steps. What Almond concludes, after studying various congestion levels in connection with the basic method's performance, is that there is a need for flexibility within the algorithmic scheme, i.e., the value of the weighting factor l_k should be determined by the problem being solved, and by the present network condition at each iteration. Almond also presents an extension to the method, where the flow is assigned gradually onto the network (i.e., the weighting Formula (1.5) is replaced by $\mathbf{f}^{k+1} = m_k\mathbf{f}^k + l_k\mathbf{y}^k$,

where $m_k + l_k < 1$ for the first few iterations). His motives for this modification is that in practice, the amount of traffic grows with time. The process simulates this, in the sense that it corresponds to some new traffic entering the network and a portion of the original traffic redistributing itself. Almond finds that the extended version is more efficient. He fails, however, to present a systematic scheme for the choices of l_k and m_k in either of the two methods, without which convergence properties can be established.

Remark 1.1 It is interesting to note the similarity between these latter heuristic approaches and one of the most popular convergent methods for traffic assignment in use today; the Frank–Wolfe algorithm (see Section 4.1 for a detailed description) is obtained by, in (1.5), letting the responsive fraction l_k be chosen so as to minimize a certain objective function. The algorithm of Smock [858, 859] and Overgaard [732, 733] (in which the responsive fraction $l_k = 1/k$ is used) is actually equivalent to the convergent *method of successive averages* (MSA) (see [764]), and thus predates this algorithm by nearly 20 years. Smock is therefore most probably unknowingly responsible for what probably is the first adaptation of a convergent traffic assignment algorithm.

An extension of the previously discussed methods was developed in the Metropolitan Toronto Regional Transportation Study, and presented in [518, 650, 667, 519]. The program package developed was possibly the first attempt to incorporate trip distribution and modal split into a single algorithm. Based on the prevailing traffic conditions, shortest routes are calculated and retained in the memory, with a maximum number of four routes within each O-D pair. (If a fifth route is found that is cheaper than the most expensive one stored to date, then the cheaper route replaces the more expensive one.) Each route is treated separately and trips are assigned to the routes in inverse proportion to their travel times based on a prespecified formula. This approach may be seen as a heuristic column generation algorithm, and is in this sense reminiscent to state-of-the-art codes such as RSD and DSD (see Sections 4.3.4 and 4.3.5). The advantage of this approach is that, since each route is independent, the weighting phase has a greater striving for an equilibrium solution. The convergence of the algorithm can, however, not be guaranteed, since the maximum number of routes is very limited and the weighting process is non-convergent. The work by Nishikawa and Nakahori [723, 724], and Nakahori *et al.* [703] is also similar to this approach. Their trip assignment method is also based on a heuristic choice of route-flow proportion, but no restriction is made *a priori* on the number of routes maintained in the algorithmic process.

A different line of development in assignment methods is the family of *incremental assignment* methods. The common approach is that the trips are gradually entered onto shortest routes based on the prevailing traffic situations, until all the trips have been assigned to the network. Different methods evolve from different choices of increments, and the order in which the trips are assigned. The methods simulate the way in which congestion emerges with growing traffic, and the natural behaviour of tripmakers to make use of different routes when congestion becomes significant.

The first incremental method presented was the *quantal loading* procedure, discussed above. In this algorithm, the increments of flow correspond to total demands from origins, and thus all the flow in an O-D pair utilizes only one route.

One incremental loading scheme that does not suffer from this drawback is implemented in the TRANSET, DODOTRANS I ([797, 639]; see also [932, 936]) and SALMOF ([870, 871]) packages, and is a popular method for creating a starting solution for convergent assignment methods (e.g., [24]). The technique has also been applied to single-commodity network flow problems, such as nonlinear electrical networks (e.g., [508]). The total

number of iterations is fixed *a priori* to, say, K. The corresponding fraction of the demand is loaded onto the shortest route in all O-D pairs given the current flows, and the process is repeated the specified number of steps. The resemblance of the actual situation is clear: as a road gets more congested, people tend to choose different roads. The resulting flow is divided between several routes for each O-D pair (the maximum number of routes used in an OD-pair is, of course, K), making the solution resemble the equilibrium situation better than a single-route assignment. The number of iterations needed is, however, not easy to determine, and the result is sensitive to the choice of increments. Some researchers have reported that the algorithm does not converge to an equilibrium solution (e.g., [708, 325, 288]); others claim that the resulting flows are good enough for practical purposes ([932, 871, 935, 933]).[4]

Extensions of the procedure are presented by Martin and Manheim [649]. Their basic algorithm is based on origin-destination trip increments, where an O-D pair is selected at random, and a fraction of the total trip volume is loaded onto the shortest route. After the travel times on the links defining the route have been adjusted, the process is repeated. The fraction to be loaded is determined by a travel-time dependent function (the *generation rate characteristic*). A consequence is that the number of iterations is not determined in advance. A variant of the basic method constitutes an extension of quantal loading, where a fraction of the demand flows from a subset of the origins is loaded onto the network at each stage. A disadvantage of incremental assignment methods is that, once a route has been assigned a flow, it can never be removed ([359]); Van Vliet [935, 933] presents a heuristic procedure for eliminating the *worst* routes after the completion of the assignment.

The above heuristics may be given a unified description. Below, we give a general algorithm for traffic assignment, based on the iterative solution of shortest route problems.

Given is a network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ of nodes and directed links, a set $(p,q) \in \mathcal{C}$ of O-D pairs with fixed demands d_{pq}, travel time functions $t_a : \Re_+ \mapsto \Re_{++}$, and a linear function v of the iterates.

The general algorithm is as follows.

Step 0 (Starting flow) Calculate an all-or-nothing solution, $\mathbf{y}^0$, based on free-flow travel times. Assign a portion $\alpha \leq 1$ of the total demand to the corresponding routes to give the link flow vector $\mathbf{f}^1$. Set $k = 1$.

Step 1 (Shortest route calculation) Calculate an all-or-nothing solution, $\mathbf{y}^k = (\mathbf{y}^k_{pq})_{(p,q)\in\mathcal{C}}$, based on travel times at the flow $v(\mathbf{f}^k, \mathbf{f}^{k-1}, \ldots, \mathbf{f}^1)$.

Step 2 (Flow update) Determine nonnegative weights m^k_{pq} and l^k_{pq} to yield the new flow $\mathbf{f}^{k+1}_{pq} = m^k_{pq}\mathbf{f}^k_{pq} + l^k_{pq}\alpha\mathbf{y}^k_{pq}$, for all $(p,q) \in \mathcal{C}$.

Step 3 (Convergence check) If $k + 1 > k_{\max}$ or if some convergence criterion holds $\rightarrow$ terminate. Otherwise, go to Step 1 with $k := k + 1$.

In Table 1.2, we describe the methods outlined above within this framework.

In Figure 1.2, some relations between heuristic schemes are given. An arrow pointing between two boxes indicates a development from one method to another, either historically or conceptually.

[4] In the case of single-commodity flows, incremental assignment methods can be made convergent ([325]).

Method	$k_{\max}$	m_{pq}^k	l_{pq}^k	α	v
A-O-N	1	0	1	1	$\mathbf{f}^k$
I A-O-N	$+\infty$	0	1	1	$\mathbf{f}^k$
CAPRES	4	$\frac{3}{4}$	$\frac{1}{4}$	1	$\frac{3}{4}\mathbf{f}^k + \frac{1}{4}\mathbf{y}^k$
F–W	$+\infty$	$1 - l^k$	$\min T(\mathbf{f}^{k+1})$	1	$\mathbf{f}^k$
Q L	1	$1,\ \forall p < k,\ \forall q$	$1,\ \forall (p,q), p = k$	1	$\mathbf{f}^k$
I A	K	1	1	$\frac{1}{K}$	$\mathbf{f}^k$

Table 1.2: Heuristic methods: comparisons

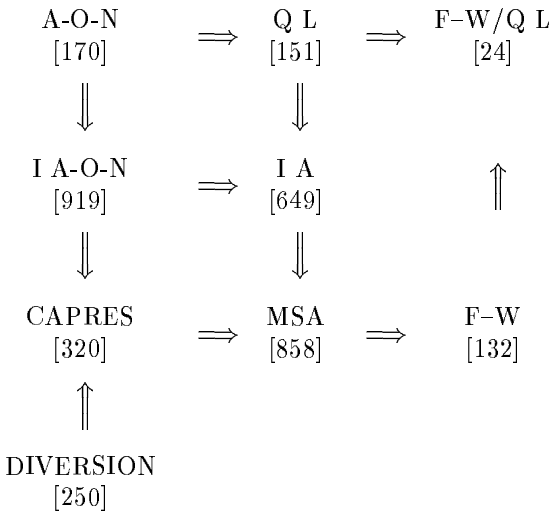

Figure 1.2: Some relations between heuristic schemes

Discussion

The development of heuristic traffic assignment schemes shows that the assignment problem was not clearly defined or understood by the transportation planning staffs. The techniques utilized were based on ad hoc rules, and even if one may trace a growing understanding of the problem to be solved, the development that is outlined above may be described as one of trial-and-error.

Simultaneous to the development of heuristic techniques, the traffic assignment problem was given precise mathematical modelling formulations by Beckmann *et al.* [47] and others (see Section 2.2). The heuristic methods employed for traffic assignment could thus have been evaluated against this model, and their deficiences revealed. The convex optimization algorithms that were available at that time (such as the Frank–Wolfe method) could then have replaced these heuristics, had they been known to the transportation planners. These possibilities were, however, not exploited for more than ten years, and the first convergent algorithms for traffic assignment (excluding that of Smock and Overgaard), were developed around 1968 by operations researchers, and not within transportation planning studies.

Boyce [104] examines the history of assignment modelling and methodology, and states the reason for the above mentioned (unfortunate) lost opportunities as the transportation planners' lack of a *rigorous scientific approach* to problem formulation, derivation of conditions governing the solution of the problem, and design and testing of convergent

algorithms for computing the solution, and also states ([102]):

> It is a useful example of the maxim that the key to the solution of a problem lies in its well-formulated statement.

In the mid-70s, several studies were made to validate the models and convergent algorithms developed in the academic community (e.g., [360, 974, 941, 288]), and planning agencies were beginning to employ them in their activities; by the end of the 70s, several program packages, developed at transportation study groups at North American and European universities and based on scientifically well founded models and methods, were available for practical use.

For further reading on heuristic algorithms for traffic assignment, see [650, 667, 315, 511, 662, 171, 319, 798, 512, 799, 972, 370, 655].

1.6 Travel forecast

At this point in the long-range study the interactions of land use and transportation are explicitly recognized.

Selecting a design (or horizon) year is usually a compromise. It must be far enough in the future for major programs to be initiated and construction staged according to funding availability, yet the design year can not be so far in the future that forecasts of future development and traffic will have a doubtful reliability. Thus, forecasts in the 15 to 25 year range are most common. In certain situations it will however be easier for laymen to understand projected levels of anticipated transportation demand which may actually occur prior to, during, or after the design.

Transportation studies may develop several land use alternatives instead of preparing a single forecast or plan of land use. Alternatives may be developed to challenge or confirm existing recommended plans, discover whether or not one land use form offers particular advantages over another, probe community values and provoke public discussion on key issues, and also to educate the public in the values of planning, and to identify the need for change in financing or government organization to facilitate plan implementation ([106, p. 30]).

To a large extent, the growth in population determines the growth in economic activity, the requirements for additional or new land uses and also the future level of transportation demand. The total amount of population growth expected in an urban area is important since it is basic to the estimation of future trip generation and interzonal travel. Equally important is knowledge of the location of this growth. Population forecasting techniques are described, for instance, in [923, 650].

A forecast of future urban area economic activity, along with the population forecast, provides the basis for estimates of future non-residential land requirements and future trip generation. The depth and scope of an analysis of urban economic activity will vary, depending on the accuracy and detail of the results desired, time and money available, etc. It is essential that reliable and accurate estimates are obtained for pattern changes and productivity in industrial and commercial employment and per capita income ([650, p. 65]).

Based on the above two forecasts, estimates of automobile ownership, future land use, accessibility and the other parameters necessary in the trip forecasting are then projected for the design year ([456, 461, 706, 650, 579]).

1.7 Network evaluation

Proposals to improve or extend existing transportation systems can range from limited improvements, such as the widening of a section of road, or the improvement of a junction, to comprehensive proposals which involve the construction of significant sections of urban motorway, the development of new forms of public transport and the close integration of different transport systems. Before decisions are taken to proceed with any proposals, either small or large, an attempt is normally made to evaluate the efficiency of the proposed investments. There are several grounds on which the proposals must be evaluated: the realism of the numerical results must be judged, the proposed transportation system must be evaluated against the predicted future transportation requirements, the environmental effects of the operation of the proposed system must be considered and the economic consequences of the provision and operation of the system must be estimated. The economic consequences of carrying out transportation schemes have received increasing attention during the last 30 years; the environmental effects, on the other hand, have received increased attention only recently.

Large-scale transportation proposals are normally assessed by means of a cost-benefit analysis, which aims to compare the costs and benefits associated with alternative schemes. For detailed accounts and examples of evaluations of proposed transportation systems, see, e.g., [984, 130, 805, 638].

The transportation planning process emerged during the 1960s, when transportation planners expected private automobiles to continue to be the primary transport mode. The travel forecasts made in the process reflected this bias, and the evaluations in this step were hence orientated toward the expansion of highway capacity. During the 1970s, however, the goals of the planning shifted toward an expansion of the public transportation network.

1.8 Discussion

In this chapter we have presented the classical approach to transportation planning. Historically the transportation planning process has developed using a series of submodels; these models were outlined, together with classical methods for their analysis. The obvious drawback of the sequential planning approach is that it may obscure the fact that the models are integral parts of the whole. The most immediate consequence of the separation into submodels is that some feedback should be introduced into the process, to allow, for instance, the travel times estimated in the traffic assignment part to influence the trip generation and distribution ([981, 106, 264]). This iteration is, however, seldom applied in practical studies; it has also been recognized that even with such a feedback, the process would not necessarily converge to a consistent solution. The conclusion one must draw is that the transportation planning process should recognize the intimate relationships between the different parts of the transportation system, and subsequently develop combined models.

The criticism against the planning studies' assignment models and methods may be extended to the other parts of the process as well; not until in the 1960s did the trip generation and distribution models used become well founded. The different parts of the planning process are also very different in their sophistication. Compared to the development of modal split and assignment models, the land use prediction is a very underdeveloped part of the process ([767]).

Domencich and McFadden [264, Chap. 1] summarize the drawbacks of the prevailing transportation planning study methodologies, particularly with respect to their criteria list:

(1) The models are basically non-behavioural. They replicate the results of conditions existing at the time of the survey and provide little or no guidance to the effects on travel decisions of changes in travellers' circumstances or in terms upon which they are offered competing alternatives in the transportation environment.

(2) Except for the modal split, the models are basically non-policy orientated. The effects of the variables which policy-makers are able to control are excluded from the trip generation and attraction functions and applied mechanically, and to a limited extent at best, in the trip distribution There is essentially no interaction between system performance and the choices of trip frequency or trip destination.

(3) The decision of time of day to travel is seldom, if ever, modelled.

(4) Equilibration is essentially ignored, except to the limited extent that auto route assignment models take account of capacity constraints in assigning routes.

(5) Models are based on data representing zonal aggregates of trips and socio-economic conditions which obscures much of the information in the data and, together with the lack of a behavioural structure, makes the models difficult to generalize from city to city.

One may add here that the process should be a continuing one rather than a once-and-for-all study; the information gathered during the inventory and forecasting stages could be effectively used together with a continuing monitoring of the developments of the plan to correct future estimates; compared with the actual land use and traffic developments, divergences could be recognized, and the models subsequently adjusted. This was recognized at an early stage of the development of planning models (e.g., [650]), but follow-up studies are not always made.

For further discussions on the shortcomings of the conventional transportation planning methods, see, e.g., [878, 465, 114, 130, 264, 711]. As responses to these shortcomings, new approaches to transportation planning have developed from the 1970s. For an account of the latest developments, see [740, pp. 564–574].

The transportation planning process is described in detail in many text books and articles; see, e.g., [814, 152, 650, 726, 467, 52, 670, 93, 762, 130, 879, 37, 805, 638, 255].

Chapter 2

The basic equilibrium model and extensions

In this and the next chapter we introduce and analyze mathematical models for the perhaps most central part of the traffic planning process: the assignment of traffic onto the routes of an existing or proposed transportation network. We also briefly discuss models presented for related network problems (such as electrical networks) and extensions to more complex problems. In this chapter we concentrate on optimization models; in the next chapter, non-optimization models are discussed.

2.1 The Wardrop conditions

Any well founded traffic model must recognize the individual travellers' decision-making with regards to when, where and how to travel. A traffic assignment model, in which one aims at providing a macroscopic description or prediction of the traffic volume resulting from route choices made in the traffic network, must therefore be based on a route-choice behavioural principle. Alternative assumptions about route-choice behaviour naturally lead to alternative model formulations.

In the analysis of traffic assignment models, *congestion* is a fundamental notion. As a result of growing traffic volumes, the average speed on a link tends to decrease, first slowly, but as the interaction among the vehicles and the queueing effects become more and more significant, the average speed decreases more rapidly, until the queueing has developed into a jamming situation in which very little flow can be observed on the link. In the analysis of traffic systems, average travel times are modelled as *link performance* functions (see Section 1.5.4, and Figure 1.1 in particular), relating travel time to the volume of traffic on the link. To account for the congestion effects, these functions are typically nonlinear, positive, and strictly increasing with flow. Parameters in the formulas often include *practical capacities*, which measure the breakpoint at which the travel time starts to grow rapidly with additional flows.

Kohl [564, p. 76] recognized that the routes chosen by the travellers were those that were individually perceived as being the shortest under the prevailing traffic conditions, i.e., travellers minimize their individual travel times. Although it has been observed that many factors other than travel time influence the drivers in their route-choice process (see Section 1.6), travel time is still the main component in the travel cost. (Recognizing this fact, we shall use the terms *travel cost* and *travel time* interchangeably.)

The result from such decisions made by all travellers individually is a situation in which no driver can reduce his/her journey time by unilaterally choosing another route,

and it is therefore known as the *user optimal* situation. The user optimal situation is characterized by the fact that all routes actually used between an origin and destination have the same *average travel time*. This is realized by considering the situation where two utilized routes give rise to different travel times. Then, users on the longer route have an incentive to change to the shorter one, and the present flow is hence not a user optimum flow. The term *user equilibrium*, which is also frequently used, stems from this characterization. The first to use the term *equilibrium* to describe the traffic pattern is perhaps the economist F. H. Knight [562].

The reader should note that the principle here described, assumes implicitly both that each traveller has *complete* and *accurate* information about all the paths available and about their characteristics, and that the pattern of network flows is so *stable over time* that past experience (such as the times over particular routes no longer used by the traveller) is still valid. (Models in which travellers are assumed to have incomplete information are known as *stochastic models*, as opposed to the above *deterministic* model; these are described in Section 2.8.1. Models in which flow is assumed to vary with time are known as *dynamic models*, as opposed to the *static* model above.)

By influencing the travellers' choices of routes society may guide tripmakers towards an optimal utilization of the traffic network, i.e., to minimize the total journey time. The resulting travel pattern from these prescriptive route choices is known as the *system optimal* flow. The system optimal situation is characterized by the fact that all routes used between an origin and destination have equal *marginal travel times*. Indeed, if the marginal travel times were different for two used routes, then it would be possible to shift a portion of the flow from the route with the higher marginal cost to the route with the lower marginal cost, and thereby decrease the total travel time.

The total travel time is generally not minimized by the user optimal travel pattern, as already observed by Pigou [756]. The only situation in which the user and system optimal flows are equal, is in the idealized case when no congestion exists ([533]).[1] In the real urban traffic system, observed flows are likely to be closer to a user than a system optimum ([268]).[2]

There are essentially two alternative means to obtaining a system optimal flow. In the first, route choices are imposed upon the users of the traffic network (*involuntary* system optimum [871, 413]). There are some transportation systems, particularly those in which there is a centralized control over tripmaking decisions, for which this will seem reasonable. For example, in an industrial logistics system, goods shipments from factories to warehouses and distribution centres may well be made in such a way as to minimize total distribution cost or simply total transportation cost, or to maximize profit; a similar situation is to be found in rail and computer communication networks ([637, 871, 413]). Such a *prescriptive* solution is also applicable in certain traffic management systems (see [413, 415], and [740, pp. 413–422]).

---

[1] Actually, it can be shown that the user and system optimal principles coincide precisely when the travel time functions are given by

$$t_a(f_a) = k_a f_a^{\beta}, \qquad \forall a \in \mathcal{A},$$

where k_a is a positive constant which may differ among the links, and β is a nonnegative universal constant of the network ([209]). For a positive value of α, marginal costs do not equal user cost, but they differ only with a universal multiplicative constant $(1+\beta)$, which does not affect the flow distribution. See also [59].

[2] Some report, however, that system optimal flows are observed ([92, 891]), or that the observed flow lies somewhere in between the two principles ([996]). Theoretical work on the difference between user and system optimal flows may be found in [50, 683].

The second alternative is to try to persuade drivers to choose their routes efficiently, by charging them *tolls* equal to their contribution to the total cost. This *voluntary* system optimal strategy is known as the *congestion pricing strategy*; see Section 2.4.

Given functions describing the relationships between traffic volume, travel cost, and demand for transportation, traffic may be assigned onto a transportation network according to either of two main principles: the principle of equal journey times, also called *descriptive assignment* as it is the most likely one to be observed, or the principle of minimal total cost, which is also called *normative assignment*.

The demand for transportation is usually considered as being average *trip rates*, i.e., average frequencies of trips entering the network during a time period (for instance part of the morning peak-hour). The notion of an equilibrium should then be thought of as the steady-state evolving after a number of time periods have passed, and the travellers have adjusted their route choices according to the prevailing conditions.

The two behavioural principles described above are usually attributed to J. G. Wardrop of the Road Research Laboratory, and are therefore referred to as the two *Wardrop conditions*. We state below the two principles as cited from Wardrop [958].

Wardrop's first principle:[3]

> The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Wardrop's second principle:

> The average journey time is a minimum.

The notion of user and system optimality, which are the other common terms, was coined by Dafermos [210, 209], based on the works of Jorgensen [533].

This work of Wardrop is not original however; the two principles were discussed, in the following terms, as early as 1920 ([756, p. 194]):

> Suppose there are two roads ABD and ACD both leading from A to D. If left to itself, traffic would be so distributed that the trouble involved in driving a "representative" cart along each of the two roads would be equal. But, in some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an "artificial" situation superior to the "natural" one. But the measure must be rightly chosen.

The purpose of this statement, which was made by the economist A. C. Pigou, is to give an example of the consequences of total freedom of companies' factory investments. He concludes that they would choose to invest in factories with higher marginal investment costs, and that society, by a correctly chosen taxation, can direct the companies to invest more wisely, from the society's point of view. In this, he has in fact both stated the two above mentioned route-choice principles and also introduced the principle of *congestion pricing*; see Section 2.4 for further discussions on this topic.

Transportation networks where a system optimal flow pattern is imposed, can be represented in economical terms as the *economy of the firm*, i.e., a system where the utilization of the production facilities are centrally controlled. In this context, the principle of equal

[3]This is the correct ordering of the two Wardrop principles.

marginal cost is well known, and states that the market price for a product should be equal to the marginal cost of the item last produced. For more detailed discussions on these relationships, we refer to Beckmann *et al.* [47] and Gartner [413].

The traffic assignment problem may also be interpreted in *game theoretical* terms. Wardrop [958] describes the properties of a user equilibrium state as follows:

> The first criterion is quite a likely one in practice, since it might be assumed that traffic will tend to settle down into an equilibrium situation in which no driver can reduce his journey time by choosing a new route.

From the above, it is natural to believe that the traffic pattern satisfying Wardrop's first principle is a Nash [705] equilibrium of a network game among the tripmakers. To our knowledge, the first to recognize this fact are Charnes and Cooper [160, 162], who identified the players as the different origin-destination pairs. The game theoretical interpretation is further discussed in Section 2.6.1.

2.1.1 The fixed demand case

To formulate the user equilibrium conditions mathematically, we consider a feasible flow pattern and an arbitrary origin-destination pair $(p,q) \in \mathcal{C}$. Let c_{pqr} denote the travel time on a route r from the origin node p to the destination node q resulting from the given flow, and assume, without any loss of generality, that the routes between p and q are so ordered, that the first l are actually used, i.e., carry a positive route flow. Then, the network flow is a user equilibrium if and only if it is true that

$$c_{pq1} = c_{pq2} = \ldots = c_{pql},$$

and the unused routes in the O-D pair (routes $l+1, \ldots$) have travel times that are at least as large as that of the used routes.

Letting $\mathcal{R}_{pq}$ denote the index set of simple routes[4] in origin-destination pair $(p,q) \in \mathcal{C}$, h_{pqr} the flow on route r, and π_{pq} the travel time on the shortest route from p to q, given the flow $\mathbf{h} = (h_{pqr})_{r \in \mathcal{R}_{pq}, (p,q) \in \mathcal{C}}$, the above Wardrop user equilibrium conditions may equivalently be stated as

$$h_{pqr} > 0 \implies c_{pqr} = \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq}, \tag{2.1a}$$

$$h_{pqr} = 0 \implies c_{pqr} \geq \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq}, \tag{2.1b}$$

to hold for all pairs $(p,q) \in \mathcal{C}$. Including the feasibility restrictions for the flow $\mathbf{h}$, the conditions for user equilibrium may be summarized as

$$h_{pqr}(c_{pqr} - \pi_{pq}) = 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.2a}$$

$$c_{pqr} - \pi_{pq} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.2b}$$

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.2c}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.2d}$$

$$\pi_{pq} \geq 0, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.2e}$$

where (2.2a)–(2.2b) is a restatement of (2.1), (2.2c) ensures the feasibility with respect to the (fixed) demands, and (2.2d)–(2.2e) ensure the nonnegativity of the route flows and corresponding travel costs.

[4] A *simple route* is a route without cycles, i.e., a route that does not utilize the same link twice.

2.1.2 The variable demand case

Frequently, the traffic assignment problem is formulated as a problem with *variable* (or *elastic*) demands, where the trip rates in the origin-destination pair are modelled as functions of the least travel cost between the origin and destination. The basic premise behind such a model is that a traveller has a number of choices available and is motivated by economical considerations in his/her decisions; the minimum travel cost is a measure of the perceived benefit to the travellers in an O-D pair, and the incentive to make a trip decreases with an increasing disutility.

Because the demand function influences both the number of trips generated at the zones and the distribution of trips among the destinations, the elastic demand model may be viewed as a simple *combined* trip generation, distribution and assignment model.

In the case of elastic demands, the interpretation of the traffic assignment problem in economical terms is more natural than in the fixed demand case. Viewing the transportation network system as an economic market, the demand side corresponds to the potential travellers, or consumers, of the network, who, in their decisions, are governed by the travel demand functions. The supply side corresponds to the network itself, offering transportation facilities to the consumers, at prices corresponding to travel times. The commodity traded at the market is tripmaking. The *market equilibrium* situation is one, where the number of trips between an origin and destination equals the travel demand given by the market price, i.e., the travel time, for the tripmaking.

To extend the user equilibrium conditions to the case of elastic demands, let the demand for transportation between the nodes p and q be a function of the cheapest route costs $\boldsymbol{\pi}$, i.e., let

$$d_{pq} \stackrel{\text{def}}{=} g_{pq}(\boldsymbol{\pi}), \qquad \forall (p,q) \in \mathcal{C}.$$

(Demand functions are discussed, for instance, in [573, 971].)

Then the Wardrop conditions for route flows and demands state that

$$h_{pqr} > 0 \implies c_{pqr} = \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq}, \tag{2.3a}$$
$$h_{pqr} = 0 \implies c_{pqr} \geq \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq}, \tag{2.3b}$$
$$d_{pq} > 0 \implies d_{pq} = g_{pq}(\boldsymbol{\pi}), \tag{2.3c}$$
$$d_{pq} = 0 \implies g_{pq}(\boldsymbol{\pi}) \leq 0, \tag{2.3d}$$

to hold for all pairs $(p,q) \in \mathcal{C}$.

The Conditions (2.3a)–(2.3b) correspond to the fixed demand Condition (2.1); the Conditions (2.3c)–(2.3d) state that the demand for transportation in an O-D pair equals the value of the demand function at the shortest route cost, and that the demand is zero if the travel cost is too high to induce any O-D flows.

We further assume that the demand function g_{pq} is nonnegative on $\Re_+^{|\mathcal{C}|}$, for all $(p,q) \in \mathcal{C}$. Including the feasibility restrictions for the flow $\mathbf{h}$ and demand $\mathbf{d}$, the conditions for a variable demand user equilibrium may be described as

$$h_{pqr}(c_{pqr} - \pi_{pq}) = 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.4a}$$
$$c_{pqr} - \pi_{pq} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.4b}$$
$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = g_{pq}(\boldsymbol{\pi}), \qquad \forall (p,q) \in \mathcal{C}, \tag{2.4c}$$
$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.4d}$$
$$\pi_{pq} \geq 0, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.4e}$$

which is equivalent to (2.2) with the exception that the demands are given by the demand functions at the shortest route costs.

2.1.3 Discussion

The Systems (2.2) and (2.4) do not include integrality restrictions on the route flows, and thus define a *continuous relaxation* of the actual Wardrop conditions. This approximation is made for the sake of simplicity, and should be quite accurate for any realistically sized network. Discrete formulations of traffic assignment problems are discussed in Section 2.6.2.

The Wardrop conditions given above are very general, since they do not assume any particular properties of the travel costs and demand functions other than nonnegativity. When studying optimization formulations in this chapter we shall, however, consider the classical, simple form of travel cost, in which the travel cost on a route is defined as the sum of the costs of the links defining the route (i.e., the travel costs are assumed *additive*), and the travel cost on a link is assumed to be independent of the flows on any other link in the network (i.e., the travel costs are assumed *separable*). A separability assumption is made also on the demand functions. (These assumptions will subsequently be relaxed in Chapter 3.)

In the next section, we shall derive optimization problems, with (2.2) and (2.4) as respective optimality conditions, through which we investigate the properties of equilibrium solutions. These formulations also enable the development of efficient optimization-based procedures for computing equilibrium solutions.

2.2 The mathematical program for user equilibrium

To our knowledge, the first optimization formulation of a traffic assignment problem, based on the Wardrop principles as the optimality conditions, is due to Prager [765]. His optimization formulation is based on an analogy between flows of traffic and electric currents (see further Section 2.6.3); this optimization problem includes, however, too restrictive assumptions to be useful for studying traffic flows.

The objective functions of the mathematical programs to be derived in this section are based on total link flows. The route and link flows, and their associated travel times, are related according to the following. The commodity link flows, $\mathbf{f}_{pq} = (f_{apq})$, given the route flows $\mathbf{h}$, are given by

$$f_{apq} \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \delta_{pqra} h_{pqr}, \qquad \forall (p,q) \in \mathcal{C},\ \forall a \in \mathcal{A}, \tag{2.5a}$$

where

$$\delta_{pqra} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if route } r \in \mathcal{R}_{pq} \text{ uses link } a, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall a \in \mathcal{A},\ \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C} \tag{2.5b}$$

defines the *link-route incidence matrix*, $\mathbf{\Delta}^{\mathrm{T}} = (\delta_{pqra})$, for the network $\mathcal{G}$. The total link flows, $\mathbf{f} = (f_a)$, are then given by

$$f_a \stackrel{\text{def}}{=} \sum_{(p,q) \in \mathcal{C}} f_{apq}, \qquad \forall a \in \mathcal{A}, \tag{2.5c}$$

or, in compact notation, summarizing (2.5a)–(2.5c),

$$\mathbf{f} = \mathbf{\Delta h}. \tag{2.5d}$$

Due to the additivity assumption on the route costs, the travel costs on the links are related to the route costs by

$$c_{pqr}(\mathbf{h}) = \sum_{a \in \mathcal{A}} \delta_{pqra} t_a(f_a), \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.5e}$$

or, compactly,

$$\mathbf{c}(\mathbf{h}) = \boldsymbol{\Delta}^{\mathrm{T}} \mathbf{t}(\mathbf{f}). \tag{2.5f}$$

Note that (2.5) defines the sequence in which the travel times on the routes in the networks are calculated from given route flows.

2.2.1 The fixed demand case

When deriving the optimization problem corresponding to the Wardrop Conditions (2.1), we shall assume the following properties of the network and the functions associated with it.

Assumption 2.A (Properties of the traffic network)

(1) *The network is strongly connected, i.e., at least one route joins each origin-destination pair* $(p,q) \in \mathcal{C}$ $(|\mathcal{R}_{pq}| \geq 1)$.

(2) *The demand* d_{pq} *is nonnegative for each* $(p,q) \in \mathcal{C}$.

(3) *The travel time function* $t_a : \Re_+ \mapsto \Re_{++}$ *is positive and continuous for each* $a \in \mathcal{A}$.

The following theorem relates the user equilibrium Conditions (2.2) to the optimization problem

[TAP]

$$\min\ T(\mathbf{f}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s)ds, \tag{2.6a}$$

subject to

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.6b}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.6c}$$

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a \in \mathcal{A}. \tag{2.6d}$$

Theorem 2.1 (Equivalent optimization formulation) *Let Assumption 2.A hold. The first-order optimality conditions of* [TAP] *is equivalent to the user equilibrium Conditions (2.2).*

Proof We associate a set of multipliers $\boldsymbol{\pi} = (\pi_{pq})$ with the Constraints (2.6b), and formulate the *Lagrangean function*

$$L(\mathbf{h}, \boldsymbol{\pi}) \stackrel{\text{def}}{=} T(\mathbf{f}(\mathbf{h})) + \sum_{(p,q) \in \mathcal{C}} \pi_{pq} \left(d_{pq} - \sum_{r \in \mathcal{R}_{pq}} h_{pqr} \right), \tag{2.7}$$

where the definitional Constraints (2.6d) are used to formulate T as a function of route flows.

The only remaining constraints are the nonnegativity Restrictions (2.6c) on the route flows, so the stationary point conditions for the Lagrangean (2.7) state that

$$h_{pqr}\frac{\partial L(\mathbf{h},\boldsymbol{\pi})}{\partial h_{pqr}} = 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.8a}$$

$$\frac{\partial L(\mathbf{h},\boldsymbol{\pi})}{\partial h_{pqr}} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.8b}$$

$$\frac{\partial L(\mathbf{h},\boldsymbol{\pi})}{\partial \pi_{pq}} = 0, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.8c}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{2.8d}$$

In order to further develop (2.8), we first note that, from (2.6d),

$$\frac{\partial T(\mathbf{f}(\mathbf{h}))}{\partial h_{pqr}} = \sum_{a\in\mathcal{A}} \frac{\partial T}{\partial f_a}\frac{\partial f_a}{\partial h_{pqr}}(\mathbf{f}(\mathbf{h})) = \sum_{a\in\mathcal{A}} \delta_{pqra} t_a(f_a) = c_{pqr}(\mathbf{h}), \tag{2.9}$$

i.e., the partial derivative of T with respect to the route flow variable h_{pqr} at a given flow equals the cost of travel along route r in O-D pair (p,q).

Using the Expression (2.9), we obtain from (2.8) that

$$h_{pqr}\left(c_{pqr}(\mathbf{h}) - \pi_{pq}\right) = 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.10a}$$

$$c_{pqr}(\mathbf{h}) - \pi_{pq} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.10b}$$

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.10c}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{2.10d}$$

From (2.10a)–(2.10b), we may interpret the multiplier π_{pq} as being the minimum route cost between p and q; from the positiveness assumptions on the travel time functions,

$$\pi_{pq} \geq 0, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.10e}$$

and (2.10) and (2.2) are identical. The stationary point Conditions (2.10) are *necessary* for the optimality of $\mathbf{h}$ in [TAP], since the Constraints (2.6b)–(2.6d) of [TAP] are *linear*, in which case the *Abadie constraint qualification* always holds ([43, Lemma 5.1.4]; see also [436, 42, 752, 44]). □

The problem [TAP] was first formulated by Dafermos and Sparrow [210, 209]. We will refer to this formulation as the *link-route* formulation, and the set of link flow solutions to the System (2.6b)–(2.6d), by F^r; the sets of feasible commodity route flow solutions will be denoted by H_{pq}, and the product set by H.

2.2.2 Network representations

Since the equilibrium conditions are given in terms of route flows and costs, it follows naturally that the optimization problem [TAP] is based on route flows. The link-route representation of the network, $\mathcal{G}$ is, however, not the only one possible. The physical network may also be represented by a set $\mathcal{N}$ of nodes, corresponding to intersections and origin-destination zones, and a set $\mathcal{A}$ of directed links, corresponding to roads joining the intersections.

To simplify the discussion in this section, we redefine the link flow variables as f_{ij}, denoting the flow on the directed link, (i,j), from node i to j. Further, we let $\mathbf{f}_k = (f_{ijk})$ denote the vector of flows for commodity $k \in \mathcal{C}$.

Assuming that demands are fixed, a feasible flow for commodity k then is a vector $\mathbf{f}_k$ satisfying

$$\sum_{j\in\mathcal{W}_i} f_{ijk} - \sum_{j\in\mathcal{V}_i} f_{jik} = d_{ik}, \qquad \forall i \in \mathcal{N}, \tag{2.11a}$$

$$f_{ijk} \geq 0, \qquad \forall (i,j) \in \mathcal{A}, \tag{2.11b}$$

where

$$d_{ik} \stackrel{\text{def}}{=} \begin{cases} d_k, & \text{if node } i \text{ is the origin of commodity } k, \\ -d_k, & \text{if node } i \text{ is the destination of commodity } k, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall i \in \mathcal{N}$$

defines the demand vector, $\mathbf{d}_k$, for commodity k, and

$$\begin{aligned} \mathcal{W}_i &\stackrel{\text{def}}{=} \{\, j \mid (i,j) \in \mathcal{A} \,\}, \\ \mathcal{V}_i &\stackrel{\text{def}}{=} \{\, j \mid (j,i) \in \mathcal{A} \,\} \end{aligned}$$

denotes, respectively, the set of links initiated and terminating at node i.

A compact form of (2.11) is obtained by introducing the *node-link* incidence matrix, $\mathbf{A} = (a_{ib})$, a matrix in $\{-1, 0, 1\}^{|\mathcal{N}| \times |\mathcal{A}|}$, with

$$a_{ib} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } i \text{ is the origin node of link } b, \\ -1, & \text{if } i \text{ is the destination node of link } b, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall b \in \mathcal{A},\ \forall i \in \mathcal{N}. \tag{2.12}$$

With the node-link incidence matrix at hand, (2.11), for all O-D pairs, may be compactly summarized as

$$\mathbf{A}\mathbf{f}_k = \mathbf{d}_k, \qquad \forall k \in \mathcal{C}, \tag{2.13a}$$

$$\mathbf{f}_k \geq \mathbf{0}, \qquad \forall k \in \mathcal{C}, \tag{2.13b}$$

and the total link flows are given by

$$f_a \stackrel{\text{def}}{=} \sum_{k \in \mathcal{C}} f_{ak}, \qquad \forall a \in \mathcal{A}. \tag{2.13c}$$

We let F^n denote the feasible set of total link flows defined by (2.13).

The System (2.13) yields an alternative to (2.6b)–(2.6d) for representing feasible link flows. The problem [TAP], with (2.6b)–(2.6d) replaced by (2.13), is referred to as the *link-node* formulation. The two problems are *not* equivalent, however, as the following theorem will establish.

Theorem 2.2 [14, Th. 3.5] (Flow decomposition theorem) *Every route and cycle flow has a unique representation as nonnegative link flows. Conversely, every nonnegative link flow may be represented as a route and cycle flow (though not necessarily uniquely).*

Since cycle flows are not included in the link-route formulation, we may conclude from the above theorem that

$$F^r \subset F^n,$$

and that the difference of the two sets corresponds to the commodity cycle flows which are included in F^n.[5]

This difference between the two sets may be explained by making use of the connections between polyhedral theory and multicommodity network flows. The *representation theorem* (e.g., [590, Th. 3.2] or [43, Th. 2.6.7]) states that any point $\mathbf{x}$ in a polyhedral set X may be written as a convex combination of the extreme points of X plus a nonnegative linear combination of the extreme rays of X. In other words, letting $\{\, \mathbf{y}^j \mid j \in \mathcal{X} \,\}$ and

[5] Although the set F^r is included in F^n, Aashtiani and Magnanti [4] consider the link-route formulation more general, since it allows for more flexibility in modelling users' perception of available routes.

$\{ \mathbf{d}^i \mid i \in \mathcal{D} \}$ be the sets of extreme points and extreme rays of X, respectively, $\mathbf{x} \in X$ if and only if

$$\mathbf{x} = \sum_{j \in \mathcal{X}} \lambda^j \mathbf{y}^j + \sum_{i \in \mathcal{D}} \mu^i \mathbf{d}^i, \tag{2.14a}$$

$$\sum_{j \in \mathcal{X}} \lambda^j = 1, \tag{2.14b}$$

$$\lambda^j, \mu^i \geq 0, \qquad \forall j \in \mathcal{X},\ \forall i \in \mathcal{D}. \tag{2.14c}$$

It is well known that in the polyhedral feasible set corresponding to an uncapacitated multicommodity network (such as F^r and F^n), there is a one-to-one correspondence between *extreme points* and *simple routes* (or *spanning trees*), while *extreme rays* correspond to *cycles*. Using the representation theorem we may conclude that the polyhedral set F^r is obtained from F^n by letting $\mathcal{D} = \emptyset$.

The fact that the two feasible sets are not equal does, however, not imply that using either the link-node or the link-route formulation results in different sets of equilibria; if travel costs are positive, no traveller would choose to travel in a cycle, since it would increase his/her travel cost, and therefore the set of equilibria coincide. (In the case of separable costs, the fact that cycle flows can not be present in an equilibrium solution with positive travel costs is discussed by Newell [711, p. 154–155].)

The link-route formulation is advantageous, since the constraint structure is very simple; disregarding the link-flow defining constraints, the set H_{pq}, defining the feasible flows for commodity (p, q), is a *simplex*, that is, a set defined by nonnegativity restrictions on the variables and one additional constraint stating that the sum of the variables should equal a constant. Furthermore, the number of equality constraints in the link-route and link-node formulations is $|\mathcal{A}| + |\mathcal{C}|$ and $|\mathcal{A}| + |\mathcal{N}| \cdot |\mathcal{C}|$, respectively.

The advantage of the link-route formulation in terms of the number of constraints is, however, paid for by the enormous number of route flow variables, a number that generally grows exponentially in the size of the network. The number of route variables that are positive at a solution is, on the other hand, very few; if we knew the optimal set of routes, the equilibrium solution would be (relatively) easily obtained. The idea behind *column generation* algorithms for traffic equilibrium problems is to generate route flow variables *as needed*, i.e., routes that potentially will carry a positive flow in an equilibrium solution are generated algorithmically. Column generation algorithms are described in more detail in Sections 4.2.3 and 4.3.5, as well as the two representations of feasible flows.

To further illustrate the relationship between the two formulations, we shall below show that under Assumption 2.A, the optimality conditions of the link-node formulation is equivalent to the user equilibrium Conditions (2.2).

Alternative proof of Theorem 2.1 Consider the problem of minimizing T, as given in (2.6a), subject to the Constraints (2.13).

Similarly to the proof of Theorem 2.1, we introduce multipliers for the flow conservation constraints, in this case the Constraints (2.13a). In other words, let π_{ik} be the multiplier (or node price) corresponding to Constraint (2.11a), and consider the Lagrangean

$$L(\mathbf{f}, \boldsymbol{\pi}) \stackrel{\text{def}}{=} T\left(\sum_{k \in \mathcal{C}} \mathbf{f}_k\right) + \sum_{k \in \mathcal{C}} \sum_{i \in \mathcal{N}} \pi_{ik} \left(\sum_{j \in \mathcal{W}_i} f_{ijk} - \sum_{j \in \mathcal{V}_i} f_{jik} - d_{ik} \right), \tag{2.15}$$

where the definitional Constraints (2.13c) are used to formulate T as a function of commodity link flows.

The only remaining constraints are the nonnegativity Restrictions (2.13b) on the link flows, so the stationary point conditions for the Lagrangean (2.15) state that

$$f_{ijk} \frac{\partial L\left(\sum_k \mathbf{f}_k, \boldsymbol{\pi}\right)}{\partial f_{ijk}} = 0, \qquad \forall (i,j) \in \mathcal{A},\ \forall k \in \mathcal{C}, \tag{2.16a}$$

$$\frac{\partial L\left(\sum_k \mathbf{f}_k, \boldsymbol{\pi}\right)}{\partial f_{ijk}} \geq 0, \qquad \forall (i,j) \in \mathcal{A},\ \forall k \in \mathcal{C}, \tag{2.16b}$$

$$\frac{\partial L\left(\sum_k \mathbf{f}_k, \boldsymbol{\pi}\right)}{\partial \pi_{ik}} = 0, \qquad \forall k \in \mathcal{C}, \tag{2.16c}$$

$$f_{ijk} \geq 0, \qquad \forall (i,j) \in \mathcal{A}.\ \forall k \in \mathcal{C}. \tag{2.16d}$$

While (2.16c), (2.16d) are equivalent to the primal feasibility Constraints (2.13), (2.16a)–(2.16b) yields

$$f_{ijk}\left(t_{ij}(f_{ij}) + \pi_{ik} - \pi_{jk}\right) = 0, \qquad \forall (i,j) \in \mathcal{A},\ \forall k \in \mathcal{C}, \tag{2.17a}$$

$$t_{ij}(f_{ij}) + \pi_{ik} - \pi_{jk} \geq 0, \qquad \forall (i,j) \in \mathcal{A},\ \forall k \in \mathcal{C}. \tag{2.17b}$$

Let r be any utilized route in O-D pair k, and let the origin and destination node be p and q, respectively. Summing the Equation (2.17a) over the links defining route r, we obtain, by also noting that $f_{ijk} > 0$ on these links, that

$$\sum_{(i,j)\in r} \{t_{ij}(f_{ij}) + \pi_{ik} - \pi_{jk}\} = \pi_{pk} - \pi_{qk} + \sum_{(i,j)\in r} t_{ij}(f_{ij}) = 0, \tag{2.18}$$

i.e., the travel cost of route r equals the difference in potential between the origin and destination node. Since this cost is independent of the choice of route in the commodity, we may conclude that the travel cost is equal for *any* utilized route between p and q. An analogous argument using (2.17b) establishes that no unused route in the O-D pair can have a travel cost less than

$$c_k = \pi_{qk} - \pi_{pk}, \tag{2.19}$$

which thus is the minimum travel cost. It is also clear from (2.18) and the positiveness of the link costs that the travel cost is nonnegative. We have thus shown that the link-node formulation of [TAP] is equivalent to the Wardrop conditions of user equilibrium. □

Remark 2.1 From (2.19), it follows that the optimal node potentials are unique only up to an additive constant. This is a well-known property of network flow problems, due to the redundancy present in the System (2.13a). We may, for instance, define the origin potential as zero, $\pi_{pk} = 0$, whence it follows that the travel cost between the origin and destination, at equilibrium, equals the potential at the destination node.

2.2.3 The elastic demand case

When deriving the optimization problem corresponding to the variable demand user equilibrium Conditions (2.3), we shall assume that each demand function g_{pq} is nonnegative on $\Re_+$, and further *strictly decreasing* in shortest route cost, i.e.,

$$\pi^1_{pq} > \pi^2_{pq} \Longrightarrow g_{pq}(\pi^1_{pq}) < g_{pq}(\pi^2_{pq}), \qquad \forall \pi^1_{pq}, \pi^2_{pq} \geq 0.$$

(This is a one-dimensional version of strict negative monotonicity, see Definition A.2.c.) Under this assumption, the demand function is invertible, in which case

$$d_{pq} = g_{pq}(\pi_{pq}) \Longleftrightarrow \pi_{pq} = g^{-1}_{pq}(d_{pq}) \tag{2.20}$$

whenever $d_{pq} > 0$.

We summarize the properties assumed below.

Assumption 2.B (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand function* $g_{pq} : \Re_+ \mapsto \Re_+$ *is nonnegative, continuous and strictly decreasing for each* $(p,q) \in \mathcal{C}$.

(3) *The travel time function* $t_a : \Re_+ \mapsto \Re_{++}$ *is positive and continuous for each* $a \in \mathcal{A}$.

The following problem will be shown below to yield optimality conditions corresponding to the variable-demand user equilibrium conditions.

[TAP-E]

$$\min\ T(\mathbf{f},\mathbf{d}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \int_0^{f_a} t_a(s)ds - \sum_{(p,q)\in\mathcal{C}} \int_0^{d_{pq}} g_{pq}^{-1}(s)ds, \tag{2.21a}$$

subject to

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.21b}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{2.21c}$$

$$\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a\in\mathcal{A}, \tag{2.21d}$$

$$d_{pq} \geq 0, \qquad (p,q)\in\mathcal{C}. \tag{2.21e}$$

Remark 2.2 The Constraints (2.21e) are redundant because of the Constraints (2.21b)–(2.21c), and are included in the formulation of [TAP-E] only to highlight the fact that the demand is variable.

Theorem 2.3 (Equivalent optimization formulation) *Let Assumption 2.B hold. The first-order optimality conditions of* [TAP-E] *is equivalent to the elastic demand user equilibrium Conditions (2.4).*

Proof The proof is similar to that of Theorem 2.1. Introducing multipliers (π_{pq}) corresponding to the Constraints (2.21b), and substituting links flows, the Lagrangean function becomes

$$L(\mathbf{h},\mathbf{d},\boldsymbol{\pi}) \stackrel{\text{def}}{=} T(\mathbf{f}(\mathbf{h}),\mathbf{d}) + \sum_{(p,q)\in\mathcal{C}} \pi_{pq} \left(d_{pq} - \sum_{r\in\mathcal{R}_{pq}} h_{pqr} \right), \tag{2.22}$$

The Lagrangean is subject to nonnegativity restrictions on the route flows and demands. The stationary point conditions yield (2.8), from differentiating with respect to the h_{pqr} and π_{pq} variables, and

$$d_{pq} \frac{\partial L(\mathbf{h},\mathbf{d},\boldsymbol{\pi})}{\partial d_{pq}} = 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.23a}$$

$$\frac{\partial L(\mathbf{h},\mathbf{d},\boldsymbol{\pi})}{\partial d_{pq}} \geq 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.23b}$$

$$d_{pq} \geq 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.23c}$$

from differentiating with respect to the d_{pq} variables.

As for the fixed demand case, (2.8) yields the System (2.10), while (2.23) yields

$$d_{pq}\left(\pi_{pq} - g_{pq}^{-1}(d_{pq})\right) = 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.24a}$$

$$\pi_{pq} - g_{pq}^{-1}(d_{pq}) \geq 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.24b}$$

$$d_{pq} \geq 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.24c}$$

By virtue of the strictly decreasing property of the demand functions [see (2.20)], the System (2.24) states that if, in an O-D pair, the demand is positive, it is given by the value of the demand function, while, if

it is zero, the travel cost is too large to induce a positive demand [see (2.24b)]. But this is precisely the elastic demand equilibrium condition for the origin-destination demand flows. □

The problem [TAP-E] was first formulated by Beckmann *et al.* [47, Sec. 3.1.2], using a link-node formulation, and the then relatively new results on necessary conditions of optimality in nonlinear programming from Kuhn and Tucker [572]. The first to formulate the problem of transportation market equilibrium as a mathematical program is Samuelson [806]. However, the first to formulate the problem taking congestion effects into consideration are Beckmann *et al.*; see Section 2.6.4 for a brief description of market equilibrium problems. We let H_d denote the set of feasible demand and route flows of [TAP-E], and F_d^T the corresponding set of demand and link flow solutions.

2.2.4 Equivalent fixed demand reformulations

It is possible to transform the elastic demand problem to a problem with fixed demands, by a suitable augmentation of the network. By performing such a transformation, it is thus possible to utilize the many efficient algorithms available to [TAP] for the solution of [TAP-E]. The basic idea is to interpret the inverse O-D demand function as the cost function for an auxiliary link, joining the O-D pairs.

The first known transformation is due to Murchland [697]. He transforms [TAP-E] to a problem of finding a minimum-cost circulation flow, by introducing fictitious links (*return generating links*) from the destinations to the origins, with costs equal to the negative of the respective inverse demand function.

Murchland's transformation is further developed by Dantzig *et al.* [226] and Gartner [414]. By slightly modifying the minimum-cost circulation flow problem, a fixed demand problem is obtained. In this transformation, the network is augmented by a dummy destination node for each origin, which are connected by a zero-cost overflow branch. *Generating links* are added between the destinations, corresponding to the origin in question, and the dummy node, with a cost function equal to the negative of the inverse demand function. By defining an overestimate of the optimal demands to be the fixed demands between the origin and the corresponding dummy destination node, all excess demands are transferred to the zero-cost links. This reformulation then is equivalent to a fixed demand assignment problem.

Gartner [414] gives an *excess-demand* transformation, based on the addition, for each O-D pair, of a *forward generating link* between the origin and destination. An overestimate of the demand is defined as the fixed demand of the O-D pair, and the excess demand is carried on the artificial link, whose cost function is the inverse demand function. Since this transformation involves the least amount of additional data of the ones given, it is probably the most efficient computationally, and from some computational experience ([414]) one might expect that, through this transformation, elastic demand problems are solved within 1.25–1.75 times the time needed for solving a fixed demand problem on the same original network and using the same algorithm.

2.2.5 Discussion

After the transformation has been made, there is no distinction between the original (supply) links and the fictitious (demand) ones. The conclusion is thus that the elastic demand problem [TAP-E] is a special case of [TAP], and, since the fixed demand problem corresponds to a special case of [TAP-E] where the demand function is constant, [TAP] and [TAP-E] can be said to be equivalent.

In the case of fixed demands, the Objective (2.6a) has no satisfying physical or economic interpretation ([533, 209, 413]), although artificial interpretations exist in terms of incremental (or cumulative) travel costs (e.g., Newell [711, Sec. 6.3], and Erlander and Stewart [302, Sec. 8.6–7]). More natural interpretations exist in assignment models, where the flow is assumed to be integer valued (see Section 2.6.2).

Beckmann *et al.* [47] considered the Objective (2.21a) of [TAP-E] as a mathematical construction for obtaining the user equilibrium conditions as the optimality conditions; this fact led some researchers ([101, 711]) to argue that there is no economic significance in the objective of [TAP-E]. Some others (e.g., [359, 601, 661]) incorrectly (according to Gartner [413]) assert that the solution of [TAP-E] amounts to maximizing *consumer surplus*; Beckmann *et al.* warn against this interpretation, which is valid only when the travel times are independent of flow ([413]). The Objective (2.21a) is however related to *social welfare functions* ([417, 640, 413]).

Extensions of the elastic demand model arise in the development of cost versus time models; the objective of [TAP-E] is then augmented by terms related to the monetary expenses associated with a trip; see Laurent [592].

So far, we have not established the *existence* of an equilibrium solution. This is the topic of the next section.

2.3 Properties of equilibrium solutions

2.3.1 Existence of equilibrium solutions

Through the equivalence Theorems 2.1 and 2.3, the existence of an equilibrium may be established by ensuring the existence of solutions to [TAP] and [TAP-E] (for both the link-route and link-node formulations). This is the case, since from the above theorems, any flow satisfying the corresponding necessary optimality conditions is a fixed and elastic demand equilibrium flow, respectively.

In the case of elastic demands, the demand functions are required to be *strictly decreasing* in order to formulate the problem [TAP-E]; this property of the demand functions is, however, not necessary for ensuring the existence of an elastic demand equilibrium solution, see Theorem 3.17. When analyzing the existence and uniqueness of elastic demand equilibria, we will therefore not use the formulation [TAP-E].

Theorem 2.4 (Existence of equilibrium solutions) *Let Assumption 2.A hold.*

(a) *There exists an optimal solution to both the link-route and link-node formulations of* [TAP]*, which hence is a fixed demand user equilibrium flow.*

(b) *Assume that each demand function g_{pq} is nonnegative, continuous and bounded from above. Then there exists an elastic demand user equilibrium demand and flow.*

Proof

(a) According to Weierstrass' Theorem, a continuous function attains its minimum on a nonempty, closed and bounded set. From the assumptions, in order to be able to apply this theorem to [TAP], the only thing left to prove is that the feasible set may, without affecting the solution set, be restricted to a bounded set. In the link-route case, by virtue of the absence of cycles in the formulation, the feasible set is itself bounded. In the case of the link-node formulation, the feasible set is unbounded, but from the positiveness of the travel cost functions, one may add the constraints

$$f_{ijk} \leq d_k, \qquad \forall (i,j) \in \mathcal{A},\ \forall k \in \mathcal{C} \tag{2.25}$$

without affecting the equilibrium solution. By Weierstrass' Theorem, there exists a solution to [TAP], and from Theorem 2.1 this is a user equilibrium flow.

(b) The proof is rather involved algebraically, and we therefore refer the interested reader to Aashtiani and Magnanti [4] for a detailed analysis. We only mention here that the proof involves the reformulation of the equilibrium Conditions (2.4) as a *fixed point* problem (see Section 3.1.4), and the restriction of the commodity flows to a bounded set through the additional constraints [cf. (2.25)]

$$f_{ijk} \leq \sup_{\pi \geq 0} g_{pq}(\pi), \qquad \forall k \in \mathcal{C}.$$ □

For further discussions on the existence of traffic equilibria, see Section 3.3.

We denote the set of equilibrium route flows by H^*.

With the above assumptions, there may be more than one equilibrium solution, and the travel costs of different flows i H^*, as well as the demands, may differ. (This follows from the possible non-monotonicity of the travel and demand costs.) To alleviate this unwanted property, we impose further conditions on the travel time and demand functions to ensure that the equilibrium travel times and demands are unique. We also establish the uniqueness of the total link flows.

2.3.2 Uniqueness of equilibrium solutions

Assumption 2.C (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand d_{pq} is positive for each $(p,q) \in \mathcal{C}$ (fixed demand case).*

(3) *The demand function $g_{pq} : \Re_+ \mapsto \Re_+$ is positive, continuous, upper bounded and non-increasing for each $(p,q) \in \mathcal{C}$ (elastic demand case).*

(4) *The travel time function $t_a : \Re_+ \mapsto \Re_{++}$ is positive, continuous and non-decreasing for each $a \in \mathcal{A}$.*

Theorem 2.5 (Uniqueness of equilibria) *Let Assumption 2.C hold.*

(a) *The equilibrium travel times are unique.*

(b) *Assume that each demand function is strictly decreasing. Then the equilibrium demands are unique.*

(c) *Assume that each travel time function is strictly increasing. Then the equilibrium link flows are unique.*

Proof

(a) In the fixed demand case, we will show the result by establishing the convexity of the objective of [TAP] in terms of route flows, and using the result of Mangasarian [636] that the gradient mapping of a convex program is constant on its solution set (H^*). From (2.9), it then follows that all routes have an invariant cost on the set H^* of equilibrium solutions.

We first have that

$$\begin{aligned}
[\mathbf{c}(\mathbf{h}^1) - \mathbf{c}(\mathbf{h}^2)]^{\mathrm{T}}(\mathbf{h}^1 - \mathbf{h}^2) &= [\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{t}(\mathbf{f}^1) - \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{t}(\mathbf{f}^2)]^{\mathrm{T}}(\mathbf{h}^1 - \mathbf{h}^2) && (2.26a)\\
&= [\mathbf{t}(\mathbf{f}^1) - \mathbf{t}(\mathbf{f}^2)]^{\mathrm{T}}(\boldsymbol{\Delta}\mathbf{h}^1 - \boldsymbol{\Delta}\mathbf{h}^2) && (2.26b)\\
&= [\mathbf{t}(\mathbf{f}^1) - \mathbf{t}(\mathbf{f}^2)]^{\mathrm{T}}(\mathbf{f}^1 - \mathbf{f}^2) && (2.26c)\\
&\geq 0, \qquad \forall \mathbf{f}^1, \mathbf{f}^2 \in F^r, && (2.26d)
\end{aligned}$$

where (2.26a) follows from (2.5f), (2.26c) from (2.5d), and (2.26d) from the monotonicity property of the link costs [cf. (A.14)].

The development (2.26) shows that if the link costs are monotone, then so are the route costs. It hence follows that the objective of [TAP] is convex, both in terms of the total link flow variables, $\mathbf{f}$, and in the route flow variables, $\mathbf{h}$ (see also [43, Th. 3.3.4]). The result follows.

The result for the elastic demand case follows from similar arguments, but for details we refer to Aashtiani and Magnanti [4].

(b) From the strictly decreasing property of the demand functions, it follows that the Objective (2.21a) of [TAP-E] is well defined, and also *strictly convex* in the demands. It is a well known property of a strictly convex function that its minimum is unique (see, e.g., [43, Th. 3.4.2]). The result follows.

(c) Analogous to the proof of **(b)**, using the strict convexity of the objective with respect to $\mathbf{f}$. □

Under the assumptions of Theorem 2.5.c the problem [TAP] is strictly convex, and hence the total link flow in equilibrium is unique. We denote this flow vector by $\mathbf{f}^*$.

Viewing the problem in terms of route flow variables only, the objective function is only convex, since a link flow pattern may correspond to several route flow patterns (see Theorem 2.2). An equilibrium route flow is hence not unique in general, even though the link flow is. (Even when the equilibrium link flow solution is known, finding an equilibrium route flow is still not a trivial task; see [270, 14] for algorithms by which link flows may be decomposed into route flows.) An interesting property of [TAP] is that the definitional Constraints (2.6d) [i.e, the System (2.5d)] induce a projection onto a subspace of the feasible set of route flows, defined by (2.6b), (2.6c), where the objective is strictly convex. This property has some interesting consequences for the study of a Lagrangean dual problem associated with [TAP] and [TAP-E]; this is the topic of the next section.

The above established properties of the *primal* formulation of traffic equilibrium problems are covered in many text books on traffic assignment (see, e.g., [762, 871, 638, 711, 831]); *dual* formulations are analyzed in the existing traffic assignment literature to a much lesser degree, despite their richness in interpretations and usefulness for constructing computational schemes.

2.3.3 Further properties of equilibrium solutions

We consider [TAP-E] under the assumptions of Theorem 2.5.

[TAP-E]

$$\min\ T(\mathbf{f},\mathbf{d}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \int_0^{f_a} t_a(s)ds - \sum_{(p,q)\in\mathcal{C}} \int_0^{d_{pq}} g_{pq}^{-1}(s)ds, \tag{2.27a}$$

subject to

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.27b}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{2.27c}$$

$$\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a\in\mathcal{A}, \tag{2.27d}$$

$$d_{pq} \geq 0, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.27e}$$

$$f_a \geq 0, \qquad \forall a\in\mathcal{A}. \tag{2.27f}$$

Note that the Constraints (2.27e)–(2.27f) are redundant in [TAP-E], but not in the Lagrangean dualized problem.

We introduce multipliers $\boldsymbol{\mu} = (\mu_a)_{a\in\mathcal{A}}$ for the link flow definitional Constraints (2.27d), and define the dual traffic equilibrium problem

[DTAP-E]

$$\max\ \theta(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \theta_{SC}(\boldsymbol{\mu}) + \theta_{SR}(\boldsymbol{\mu}), \tag{2.28a}$$

where

[SC]

$$\theta_{SC}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \min_{f_a\geq 0}\left\{\int_0^{f_a} t_a(s)ds - \mu_a f_a\right\} \tag{2.28b}$$

and

[SR]

$$\theta_{SR}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{(p,q)\in\mathcal{C}} \min \left\{\sum_{r\in\mathcal{R}_{pq}}\left(\sum_{a\in\mathcal{A}}\delta_{pqra}\mu_a\right)h_{pqr} - \int_0^{d_{pq}} g_{pq}^{-1}(s)ds\right\}, \tag{2.28c}$$

$$\text{s.t.}\quad (2.27b), (2.27c), (2.27e). \tag{2.28d}$$

Observe that dualizing the Constraints (2.27d) in the link-route formulation of [TAP-E] corresponds to dualizing the Constraints (2.13c) in the link-node formulation.

Due to the decoupling effect of dualizing (2.27d), the dual objective is the sum of two functions; the evaluation of these functions, i.e., the solution of the subproblems [SC] and [SR], further separates into $|\mathcal{A}|$ strictly convex single-variable minimization problems ([SC_a]), and $|\mathcal{C}|$ shortest route problems and strictly convex single-variable minimization problems ([SR_{pq}]), respectively.

We first discuss the solution of [SC]. Assume that each function $t_a : \Re_+ \mapsto \Re_{++}$, $a \in \mathcal{A}$, is *weakly coercive* on $\Re_+$ (see Definition A.3.a) , i.e., that

$$\lim_{f_a\to+\infty} t_a(f_a) = +\infty. \tag{2.29}$$

It is then straightforward to show that [SC_a] is solved by

$$f_a(\mu_a) = \begin{cases} t_a^{-1}(\mu_a), & \text{if } \mu_a \geq t_a(0), \\ 0, & \text{otherwise,} \end{cases} \qquad \forall a \in \mathcal{A}, \tag{2.30}$$

where t_a^{-1} is the continuous inverse of t_a ([794, p. 90]).[6, 7]

Turning to the problem [SR], through the Relation (2.27b) the problem [SR_{pq}], $(p,q) \in \mathcal{C}$, is equivalent to

$$\min_{\substack{r\in\mathcal{R}_{pq}\\ d_{pq}\geq 0}} \left\{d_{pq}\sum_{a\in\mathcal{A}}\delta_{pqra}\mu_a - \int_0^{d_{pq}} g_{pq}^{-1}(s)ds\right\}. \tag{2.31}$$

Its solution is obtained in two steps. First, a shortest route between p and q given link costs $\boldsymbol{\mu}$ is obtained; let $\pi_{pq}(\boldsymbol{\mu})$ denote its cost. The Problem (2.31) then reduces to

$$\min_{d_{pq}\geq 0}\left\{\pi_{pq}(\boldsymbol{\mu})d_{pq} - \int_0^{d_{pq}} g_{pq}^{-1}(s)ds\right\}; \tag{2.32}$$

[6] The travel time formulas given in Table 1.1 satisfy the Condition (2.29), and have explicit inverses t_a^{-1}.

[7] If the redundant link flow nonnegativity restrictions are removed from [TAP-E], then the solution is $f_a(\mu_a) = t_a^{-1}(\mu_a)$, in which case $-f_a = -t_a^{-1}$ is the *conjugate function* ([779, Sec. 12]) of $\int_0^{\cdot} t_a(s)ds$.

the solution to this strictly convex single-variable problem is

$$d_{pq}(\pi_{pq}(\boldsymbol{\mu})) = \max\{0, g_{pq}(\pi_{pq}(\boldsymbol{\mu}))\}, \tag{2.33}$$

i.e., the value of the demand function given the shortest route cost $\pi_{pq}(\boldsymbol{\mu})$, if nonnegative.[8] [Compare with the equilibrium Conditions (2.24).] Note that $\pi_{pq}(\boldsymbol{\mu})$ may not be the equilibrium shortest route cost, unless $\boldsymbol{\mu}$ is optimal in [DTAP-E].

To summarize, the solution set of [SR$_{pq}$] is the convex combination of the simple route flows corresponding to the shortest routes given the travel costs $\boldsymbol{\mu}$, and carrying the flow d_{pq}, given by (2.33).

We introduce the solution sets

$$H(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \{\, \mathbf{h} \mid \mathbf{h} \text{ solves [SR] at } \boldsymbol{\mu} \,\}, \tag{2.34}$$

$$\mathcal{R}_{pq}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \{\, r \mid \text{route } r \text{ solves [SR}_{pq}\text{] at } \boldsymbol{\mu} \,\}, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.35}$$

$$f_a(\mu_a) \stackrel{\text{def}}{=} \{\, f_a \mid f_a \text{ solves [SC}_a\text{] at } \mu_a \,\}, \qquad \forall a \in \mathcal{A}. \tag{2.36}$$

The set $H(\boldsymbol{\mu})$ is a polytope, but not a singleton for all values of $\boldsymbol{\mu}$ (in particular not at an optimal solution).

Some properties of the dual function θ are given below.

Theorem 2.6 [582] (Properties of θ) *The dual objective θ is the sum of a concave, piecewise linear function, and a strictly concave, differentiable function. It is thus everywhere finite, continuous, concave, and subdifferentiable. Its subdifferential is a bounded polyhedron, given by*

$$\partial\theta(\boldsymbol{\mu}) = \mathtt{conv} \left\{ \left(\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}(\mu)} \delta_{pqra} h_{pqr} - f_a \right)_{a\in\mathcal{A}} \;\middle|\; \mathbf{f} = \mathbf{f}(\boldsymbol{\mu}), \mathbf{h} \in H(\boldsymbol{\mu}) \right\}. \tag{2.37}$$

Further, $\theta(\boldsymbol{\mu}) \leq T(\mathbf{f}^*)$, for all $\boldsymbol{\mu} \in \Re^{|\mathcal{A}|}$.

Consider an arbitrary dual solution $\boldsymbol{\mu}$, and let

$$\hat{\boldsymbol{\mu}} \stackrel{\text{def}}{=} \max\{\boldsymbol{\mu}, \mathbf{t}(\mathbf{0})\},$$

where the maximum is taken component-wise. Then, $\mathbf{f}(\hat{\boldsymbol{\mu}}) = \mathbf{f}(\boldsymbol{\mu})$, so that $\theta_{SC}(\hat{\boldsymbol{\mu}}) = \theta_{SC}(\boldsymbol{\mu})$. Further, $\theta_{SR}(\hat{\boldsymbol{\mu}}) \geq \theta_{SR}(\boldsymbol{\mu})$, since $\hat{\boldsymbol{\mu}} \geq \boldsymbol{\mu}$, and it follows that $\theta(\hat{\boldsymbol{\mu}}) \geq \theta(\boldsymbol{\mu})$. Since the dual objective is to be maximized over $\mathcal{R}^{|\mathcal{A}|}$, one can therefore, without any loss of generality, impose the restrictions $\boldsymbol{\mu} \geq \mathbf{t}(\mathbf{0})$ (see also [388, 145, 582]).

The Lagrangean dual problem may now be stated as

[DTAP-E]

$$\max_{\mu \geq t(0)} \; \theta(\boldsymbol{\mu}). \tag{2.38}$$

The (convex) program [DTAP-E] has a nice interpretation; in contrast to the primal problem [TAP-E], in which equilibrium link flows and demands are sought, [DTAP-E] is the problem of determining the equilibrium travel times, $\boldsymbol{\mu}^*$ (see [388, 145, 147]). The optimal solution is unique, since $\mathbf{f}^*$ is unique and t_a^{-1} is strictly increasing on $\mu_a \geq t_a(0)$ ([582]).

We finally relate the optimal solutions to [TAP-E] and [DTAP-E].

[8] If the redundant demand nonnegativity constraints are removed from [TAP-E], then, similarly to the case of [SC], $d_{pq}(\pi_{pq}(\boldsymbol{\mu})) = g_{pq}(\pi_{pq}(\boldsymbol{\mu}))$, and there is a conjugacy relation between d_{pq} and $\int_0^{\cdot} g_{pq}^{-1}(s)ds$.

Theorem 2.7 [582] (Relationships between [TAP-E] and its dual) *Strong duality holds, i.e.,* $\theta(\boldsymbol{\mu}^*) = T(\mathbf{f}^*, \mathbf{d}^*)$. *Furthermore,* $\mathbf{f}^* = \mathbf{f}(\boldsymbol{\mu}^*)$, $\mathbf{d}^* = \mathbf{d}(\boldsymbol{\pi}(\boldsymbol{\mu}^*))$,

$$H^* = \left\{ \mathbf{h} \in H(\boldsymbol{\mu}^*) \;\middle|\; \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}(\mu^*)} \delta_{pqra} h_{pqr} = f_a^*, \qquad \forall a \in \mathcal{A} \right\} \subseteq H(\boldsymbol{\mu}^*), \quad (2.39)$$

and $\mathcal{R}_{pq}^* = \mathcal{R}_{pq}(\boldsymbol{\mu}^*)$, *for all* $(p,q) \in \mathcal{C}$, *where* $\mathcal{R}_{pq}^*$ *is the set of shortest routes at equilibrium for O-D pair* (p,q).

Proof Strong duality follows from the convexity of [TAP-E] in $(\mathbf{f}, \mathbf{d})$ (see [43, Th. 6.2.4]). Further, the set of optimal route flows may be characterized as the Lagrangean subproblem solutions for $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ which also satisfy the dualized Constraints (2.27d) [43, Th 6.5.1].

The uniqueness of $\mathbf{f}(\boldsymbol{\mu}^*)$ gives $\mathbf{f}^* = \mathbf{f}(\boldsymbol{\mu}^*)$ [and similarly for $\mathbf{d}^*$], and the expression for H^* follows. The set $\mathcal{R}_{pq}^*$ is obtained from a linear approximation of [TAP-E] with respect to $\mathbf{f}$ at $\mathbf{f} = \mathbf{f}^*$. Since $\mathbf{t}(\mathbf{f}^*) = \boldsymbol{\mu}^*$, this linearized problem is equivalent to [SR] defined at $\boldsymbol{\mu} = \boldsymbol{\mu}^*$. It follows that $\mathcal{R}_{pq}^* = \mathcal{R}_{pq}(\boldsymbol{\mu}^*)$. □

To conclude, the theorem states that the unique optimal link flows, $\mathbf{f}^*$, and demands, $\mathbf{d}^*$, are obtained from the subproblem [SC] and [SR] defined at the optimal dual solution, $\boldsymbol{\mu}^*$, respectively. Further, the set of routes solving [SR], given $\boldsymbol{\mu}^*$, coincide with those that are the shortest at equilibrium. However, an optimal route flow pattern $\mathbf{h}^* \in H^*$ is, in general, not directly available from the subproblem [SR], even though the optimal dual solution is at hand. This is so because $H(\boldsymbol{\mu}^*)$ is usually not a singleton, or, equivalently, because θ_{SR} is non-differentiable at $\boldsymbol{\mu}^*$. (In the case where $(\mathbf{f}^*, \mathbf{d}^*)$ is known, algorithms for calculating an $\mathbf{h}^* \in H^*$ are given in [270, 14].)

Remark 2.3 The nonuniqueness of the route flows is an unrealistic property among assignment models; route flows are certainly unique in practice. There are essentially two ways of alleviating this unwanted property. One may solve the corresponding model with an algorithm that is guaranteed to converge to one of the many possible solutions; depending on the algorithm chosen, the solution generated has different characteristics. One may also add a submodel after the assignment, in order to generate a unique set of route flows from link flows according to some behavioural principle ([792, 523]).

The reader is asked to verify the simplifications that arise in the above analysis from considering instead the fixed demand problem [TAP] and its dual [DTAP].

The problems [DTAP] and [DTAP-E] are studied by Hall and Peterson [452], using *generalized geometric programming* theory ([276]), and Fukushima [388, 389], using *conjugate duality* ([779]). Carey [145] uses traditional Lagrange duality theory, as we have done here. Duality formulations of problems similar to [TAP] have been studied also in [785, 428, 429, 476].

Algorithmic approaches for traffic assignment problems, utilizing the dualization of the definitional Constraints (2.27d) are described in Section 4.3.7.

The analysis of the optimality conditions of [TAP-E] was based on the relaxation of the network defining Constraints (2.27b). The effect of a combined relaxation of (2.27b) and (2.27d) is discussed below.

Introduce dual variables (or multipliers) $\boldsymbol{\pi} = (\pi_{pq})_{(p,q)\in\mathcal{C}}$ for the Constraints (2.27b). The dual objective is the sum of two functions. The first equals θ_{SC}. The second is

$$\vartheta(\boldsymbol{\mu}, \boldsymbol{\pi}) \stackrel{\text{def}}{=} \sum_{(p,q)\in\mathcal{C}} \inf_{\substack{h_{pqr}\geq 0,\ \forall r \\ d_{pq}\geq 0}} \left\{ \sum_{r\in\mathcal{R}_{pq}} \left(\sum_{a\in\mathcal{A}} \delta_{pqra}\mu_a - \pi_{pq} \right) h_{pqr} + \pi_{pq} d_{pq} - \int_0^{d_{pq}} g_{pq}^{-1}(s)ds \right\}.$$

To ensure the finiteness of ϑ, we must impose the constraints

$$\sum_{a\in\mathcal{A}} \delta_{pqra}\mu_a \geq \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C} \tag{2.40}$$

on $(\boldsymbol{\mu}, \boldsymbol{\pi})$. Due to the monotonicity properties of g_{pq} and the fact that we wish to maximize the dual function, we may without any loss of generality for any given value of $\boldsymbol{\mu} \geq \mathbf{0}$ choose the maximal value of π_{pq} that is dual feasible, i.e., $\pi_{pq} = \pi_{pq}(\boldsymbol{\mu}) = \min_{r\in\mathcal{R}_{pq}}\{\sum_{a\in\mathcal{A}} \delta_{pqra}\mu_a\}$. This choice of $\boldsymbol{\pi}$ results in ϑ reducing to a function only of $\boldsymbol{\mu}$, which is given by

$$\vartheta(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{(p,q)\in\mathcal{C}} \min_{d_{pq}\geq 0} \left\{ \pi_{pq}(\boldsymbol{\mu}) d_{pq} - \int_0^{d_{pq}} g_{pq}^{-1}(s)ds \right\},$$

where $\pi_{pq}(\boldsymbol{\mu})$ is the cost of the shortest route from node p to q given the travel costs $\boldsymbol{\mu}$. In other words, $\vartheta \equiv \theta_{SR}$ [cf. (2.32)], and the relaxation of the Constraints (2.27b) in combination with (2.27d) effectively results in only relaxing (2.27d).

Dual formulations of general traffic equilibrium problems are given in Section 3.3.3.

There is a plentiful of literature on algorithms based on the relaxation of network constraints [i.e., of (2.27b) or the corresponding Constraints (2.13a) in the link-node formulation]. In the context of traffic assignment, such schemes have almost invariably been used for the solution of *single-commodity* network problems that arise as subproblems in assignment algorithms (based on the decomposition of the problem into a sequence of single-commodity assignment problems). Just as methods based on the relaxation of (2.27d), these are highly parallelizable. See Section 4.3.7 for further details.

2.3.4 Stability and sensitivity of equilibrium solutions

In 1968 Braess [112] (see also [695, 876, 341, 375, 884, 873, 455]) presented an example, in which the addition of a route to a network resulted in all travellers being worse off than in the previous equilibrium solution. Known as *Braess' paradox*, this phenomenon is readily explained from the non-cooperative nature of user equilibrium flows; each traveller minimizes his/her travel time, without considering the travel times of others or, for that matter, the total travel time. Since user and system optimal flows are different in general, there is no reason to expect the total travel time to decrease when increasing the capacity of the network. Indeed, Knödel [563] presents a case, where Braess' paradox occurred in practice. (If, however, travellers choose their routes according to Wardrop's principle of minimum total travel time, Braess' paradox can not occur.) The principle underlying Braess' paradox has been recognized for many years by traffic engineers, and is considered fundamental to the understanding of traffic distribution in signal controlled traffic networks ([840, 257, 843, 842, 844]).[9] It is also more directly utilized in some *traffic control policies*, such as ramp metering ([711, Chap. 7.3]); any time one restricts the flow or capacity on a street, one is essentially increasing the travel cost on that link, presumably with the objective of improving the overall travel conditions. In view of this fact, Braess' paradox is obviously very important when evaluating proposed improvements of existing traffic networks, i.e., in *network design* ([600, 108, 876, 632]). Failure to recognize this phenomenon may lead to severe congestion, as in the practical case reported by Knödel [563].

[9] It was recently ([177]) observed that a similar paradoxical behaviour may appear in mechanical and electrical networks.

Interest in the stability and sensitivity of equilibrium demands, flows and travel times grew as a result of Braess' observation.

Conditions under which Braess' paradox can not occur are found in [874, 200, 201]; see further Section 3.3.4. Stability results for traffic equilibria are also given in [47, 209, 205, 840].

Hall [451] shows that, for strictly increasing travel costs t_a, the equilibrium cost π_{pq} is a non-decreasing, continuous (but possibly non-differentiable) function of the corresponding demand d_{pq}.[10] See also Fang [309].

Most of the sensitivity analysis of traffic equilibrium problems have been made in the context of general, and possibly asymmetric, cost functions; see Section 3.3.4.

2.4 User equilibrium versus system optimum

As described in Section 2.1, Pigou [756] had already noted the difference between a user equilibrium and system optimal travel pattern, and mentioned introducing a differential taxation to divert traffic towards a more efficient flow. In a critical article, Knight [562] examines Pigou's example, and explains the reason for this difference more clearly:

> Suppose that between two points there are two highways, one of which is broad enough to accommodate without crowding all the traffic which may care to use it, but is poorly graded and surfaced, while the other is a much better road but narrow and quite limited in capacity. If a large number of trucks operate between the two termini and are free to chose either of the two routes, they will tend to distribute themselves between the roads in such proportions that the cost per unit of transportation, or effective result per unit of investment, will be the same for every truck on both routes. As more trucks use the narrower and better road, congestion develops, until at a certain point it becomes equally profitable to use the broader but poorer highway. The congestion and interference resulting from the addition of any particular truck to the stream of traffic on the narrow but good road affects in the same way the cost and output of all the trucks using that road. It is evident that if, after equilibrium is established, a few trucks should be arbitrarily transferred to the broad road, the reduction in cost, or increase in output, to those remaining on the narrow road would be a clear gain to the traffic as a whole. The trucks so transferred would incur no loss, for any of them on the narrow road is a marginal truck, subject to the same relation between cost and output as any truck using the broad road. Yet whenever there is a difference in the cost, to an additional truck, of using the two roads, the driver of any truck has an incentive to use the narrow road, until the advantage is reduced to zero for all the trucks. Thus, as the author [A. C. Pigou] contends, individual freedom results in a bad distribution of investment between industries of constant and industries of increasing cost.
>
> In such a case social interference seems to be clearly justified. If the government should levy a small tax on each truck using the narrow road, the tax would be considered by the trucker as an element in his cost, and would cause the number of trucks on the narrow road to be reduced to the point where the *ordinary cost, plus the tax*, became equal to the cost on the broad road, assumed to be left tax free. The tax could be so adjusted that the number of trucks on the narrow road would be such as to secure the maximum efficiency in the use of the two roads taken together. The revenue obtained from such a tax would be a clear gain to the society, since no individual truck would incur higher costs than if no tax had been levied.

The idea of pricing economic activities to obtain a system optimum (i.e., equal marginal costs) were introduced to economics literature by Dupuit [285];[11] a further development is

[10] Fisk [335] gives an example in which such a perturbation leads to an increase in travel costs in other O-D pairs, and hence to a paradox related to that of Braess.

[11] The idea of pricing the use of the roads for the benefit of society is, however, much older still. In order to fund for the building and maintenance of the stone walls surrounding and protecting the city of

found in [795, 796, 707, 134]. The difference between user equilibrium and system optimum is accounted for by the individual user's failure to share the cost he/she contributes to the total travel cost, since, as stated by Knight, any additional user of a road is a marginal user. In other words, *private cost* does not equal *social cost.*

The system optimal flow minimizes the total travel cost

$$\sum_{a\in\mathcal{A}} t_a(f_a)f_a = \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} c_{pqr}(\mathbf{h})h_{pqr}. \tag{2.41}$$

The *marginal travel cost* of a link a at the flow f_a is defined as the increase in total travel cost on link a caused by an additional (marginal) tripmaker, i.e.,

$$\overline{t}_a(f_a) \stackrel{\text{def}}{=} \frac{\mathrm{d}}{\mathrm{d}f_a}(t_a(f_a)f_a) = t_a(f_a) + t'_a(f_a)f_a. \tag{2.42}$$

The difference between private and social cost then is $t'_a(f_a)f_a$, which obviously is positive unless $t'_a(f_a) = 0$ for every link a with a non-zero flow, i.e., unless there are no congestion effects in the network ([533]).

In order to achieve economic efficiency, every traveller must be made aware of the the cost he/she imposes on the other travellers. In this way, the traveller is supplied with an incentive to minimize social cost. From (2.42), it is clear that the travellers should perceive the travel costs $\overline{\mathbf{t}}$ instead of $\mathbf{t}$.

What is described here is, in fact, a mathematical description of Pigou's rightly chosen taxation, and constitutes the foundation for a *marginal cost pricing* strategy; by making every tripmaker realize $\overline{\mathbf{t}}$ instead of $\mathbf{t}$ as their travel times, the resulting equilibrium flow in terms of $\overline{\mathbf{t}}$ will be a system optimal flow in terms of the original cost. (This is easily established by, for instance, replacing t_a by $\overline{t}_a$ in the definition of [TAP]. By tracing the proof of Theorem 2.1, it then follows that the equilibrium solution is characterized by an equal marginal travel cost on the utilized routes, i.e., a minimal total cost.)

This of course means that any system optimum problem may be solved as a user equilibrium problem by redefining travel costs as $t_a := \overline{t}_a$ ([47, 533, 209, 515, 762, 941, 711]). [Conversely, any user equilibrium problem could be solved as a system problem with the appropriate redefinition of the marginal travel costs.] By reformulating the system optimization problem as a user equilibrium problem, the results of Sections 2.3.1 and 2.3.2 may be used to establish existence and uniqueness properties of system optima. (For instance, the system optimum problem is convex if the marginal cost function is non-decreasing; from (2.42), this property holds if the travel costs are differentiable, non-decreasing and convex.)

A further theoretical development of the concept of congestion tolls have been made to include bottlenecks, several vehicle types, choices of departure times, etc. (see [47, 955, 947, 682, 839, 793, 888, 948, 828, 949, 210, 829, 950, 956, 209, 823, 205, 671, 709, 710, 204, 207, 50, 488, 598, 841, 413, 425, 29, 64, 422]).

Some authors claim that system optimal solutions are only applicable in the fixed demand case ([226]), while some others (e.g., [53, 697, 709, 973]) incorrectly (according to Gartner [413]) assert that the solution of the system optimal problem amounts to maximizing *consumer surplus.*

York, a tax (known as the *murage*) was levied on all vehicles entering and leaving the city. (Even the mayor of York and the Dean and Chapter of the Minster paid this tax, although reluctantly; Booth [97] describes how, in 1305, the people of York must petition to King Edward I for the recovery of 73 pounds which the mayor had appropriated.)

The flows resulting from congestion pricing strategies are examples of a *voluntary system optimum* ([871, 413]), that is, a system optimal solution is obtained by means which do not limit travellers' freedom of choice.

An *involuntary system optimum* ([637, 871, 413]), on the other hand, is obtained by imposing the route choices on the users without charging tolls. This is only possible in networks where flows are under the complete control of a central authority, such as in rail networks or military transportation, and in some traffic management and route guidance systems ([415, 740]).

These two concepts differ in the elastic demand case, where a user is free to choose not to make the trip. The outcome of the use of these two approaches will be different in this case, since the demand is determined by social costs and average costs, respectively. In the fixed demand case, the two approaches to system optimum are equivalent.

Although the mathematical principles for obtaining a maximal utilization of the network is clear from the above, the practical problems involved in an implementation of a road pricing system, i.e., to make the user sensitive to the social cost, is immense.

Prager [765] was among the first to point out that since, in practice, the tripmakers cannot obtain full information about the status of alternate routes, they cannot be guided by Wardrop's second principle. This criticism can, in turn, be used against Wardrop's first principle; the concept of a Nash equilibrium contains, implicitly, the assumption that, when choosing a strategy, the player has full information about the values of all his/her alternative strategies, based on the current situation.

Netter [709, 710] discusses efficiency tolls in connection with multiclass-user transportation, in which different vehicles can have different values of transportation cost. He concludes that the pricing process, based upon the notion of tolls of marginal cost, needs a revision, since the pricing mechanism need not lead to a system optimal solution. The reason for this is that, in the general case of multiclass-user transportation, the Wardrop condition of a system optimum is not a convex problem. Multiclass-user transportation will be discussed further in the context of general cost models (see Section 2.5).

As discussed earlier, different results have emerged from studies of the validity of the principles of Wardrop. In this context, some researchers conclude that the differences between user and system optimal solutions are so small, that the benefit from obtaining system optimality would not add up to the costs for operating a system for marginal cost pricing (e.g., [941, 602]). Wardrop [958] had already demonstrated that the difference between the corresponding flows may be small, when analyzing a small example network.

2.5 Nonseparable costs and multiclass-user transportation networks

The Wardrop conditions for user equilibrium presented in Section 2.1 allow for very general choices of travel time and demand functions; indeed, the Conditions (2.2) and (2.4) only assume that these functions are continuous and nonnegative. (Even the continuity assumptions may be relaxed, if the definition of an equilibrium is slightly altered; see Section 3.3.1.) Yet, when formulating the problems [TAP] and [TAP-E] in Section 2.2, the travel cost on a link was assumed to be *separable*, that is, independent of the flow of all other links in the network. The demand functions were also assumed to be separable, and furthermore strictly decreasing with least route cost.

These assumptions limit the applicability of the traffic assignment models [TAP] and [TAP-E], since separable costs and demands are not realistic representations of the cost

and demand relationships in real traffic networks.

This limitation was observed quite early in the development of traffic assignment models. Prager [765] argued that the traffic moving in the opposite direction of a two-way link should be taken into account in its travel cost function, and developed an equilibrium model based on this assumption. Similar ideas appear in [162, 19].

In an urban area the delays of traffic streams at an intersection are highly interrelated; consequently, the travel time on a link will depend on the traffic flow on the other intersecting links. This example was used by Dafermos [205, 206] to motivate the development of traffic assignment models with *nonseparable* cost functions.

In the modelling of *multiclass-user* transportation networks, travel costs evolve that crucially depend on the flow on several links in the network. The underlying (natural) assumption is that the flows in a transportation network may be divided into different classes of drivers and vehicles, each of which has its own individual demand and cost-flow relationships, and contributes to the corresponding functions of the other classes. A classification of vehicle types could distinguish trucks and buses from cars, heavy vehicles from light ones, private transport from transit, etc.

Multiclass-user networks are often modelled by associating an individual copy of the original network with each class; all travellers belonging to the same class use only one network (e.g., [160, 162]). Consequently, instead of having several classes of transportation in one network, an enlarged network, in which we may view all traffic as belonging to the same class, is used. Obviously, the travel cost on one link in the enlarged network is dependent of the flow in all the other copies of the same link for the other vehicle classes.

Roth [793] was perhaps the first to state the need for models that take different user classes into account. Sender and Netter [823] study elastic demand problems with different vehicle types and two-way links, and develop existence results for equilibria as well as studying marginal cost pricing in this context. Independently, Dafermos [206] develops a general model for different user classes, and studies the existence and uniqueness of user equilibria and system optimal solutions. Jeevanantham [527] examines the influence of differences in the travel cost perceptions of different user classes on the distribution of equilibrium flows on some small examples.

In the general case, then, the cost functions t_a depend on the whole network flow. Thus, let $t_a : \Re_+^{|\mathcal{A}|} \mapsto \Re_+$ be the travel time on link $a \in \mathcal{A}$, given the whole vector of link flows. (We let $\mathcal{A}$ denote the set of directed links; note that in the case of multiclass-user transportation networks, the same original link may occur several times in the set $\mathcal{A}$, corresponding to different vehicle classes.)

As noted above, the concept of an equilibrium solution is not altered if complex travel time functions are introduced. However, it does affect the presence of equivalent problems [TAP] and [TAP-E], and the conditions for the uniqueness of an equilibrium solution.

To be able to solve the general equilibrium problem as an equivalent optimization problem of the form [TAP] or [TAP-E] the travel cost function $\mathbf{t}$ must be *integrable*, i.e., $\mathbf{t} : \Re_+^{|\mathcal{A}|} \mapsto \Re_+^{|\mathcal{A}|}$ must be a gradient mapping. Otherwise, the function

$$T(\mathbf{f}) \stackrel{\text{def}}{=} \int_{\mathbf{0}}^{\mathbf{f}} \mathbf{t}(\mathbf{s})^{\mathrm{T}} d\mathbf{s}, \tag{2.43}$$

which is to replace (2.6a) as the objective function of [TAP], is not well defined ([727, 925, 144]).

If $\mathbf{t}$ is differentiable, the Integral (2.43) is well defined (i.e., the mapping $\mathbf{t}$ is a gradient mapping) if and only if the *Jacobian* matrix $\nabla \mathbf{t}(\mathbf{f})$ is symmetric everywhere ([727,

Th. 4.1.6]), that is[12]

$$\frac{\partial t_a(\mathbf{f})}{\partial f_b} = \frac{\partial t_b(\mathbf{f})}{\partial f_a}, \qquad \forall a, b \in \mathcal{A},\ \forall \mathbf{f} \in F^n. \tag{2.44}$$

The symmetry Condition (2.44) states that the flows on any two pairs of links have an equal influence on each other's disutility. An analogous result can be stated for the general demand function $\mathbf{g} : \Re_+^{|\mathcal{C}|} \mapsto \Re_+^{|\mathcal{C}|}$, ensuring that the objective of [TAP-E] is well defined. (See the dissertation by Bernstein [63] for a detailed analysis of conditions for the existence of mathematical programs for the solution of traffic equilibrium problems.)

Observe that the above condition is satisfied automatically in the separable model, where the Jacobian matrix is diagonal, and hence symmetric.

The results on the uniqueness of equilibria for nonseparable, symmetric costs and demands, corresponding to Theorem 2.5, are valid as stated, with the only change being that the property of an increasing function is replaced by the more general property of monotonicity (see Definition A.2).

Under a monotonicity assumption on the travel time function, the problem [TAP] is convex. The strict monotonicity assumption (which generalizes the property of a strictly increasing function, and results in unique link flows) is implied by a diagonal dominance condition on the Jacobian matrices, stating, in effect, that the dominant factor determining the cost on a link is the flow on the link itself.

It should be noted that the system optimum problem is well defined also for nonseparable cost functions. With the use of a general cost function, it is, however, unlikely that the total travel cost $\mathbf{t}(\mathbf{f})^{\mathrm{T}}\mathbf{f}$ is a convex function of $\mathbf{f}$ ([823, 709, 710]); one important consequence is that congestion pricing policies based on marginal travel costs may lead to non-optimal solutions.

The nonseparable, symmetric models given above have been criticized by many researchers for being too unrealistic. The symmetry condition was first criticized by Sender and Netter [823] (see also [709, 710]), in the context of marginal cost pricing in multiclass-user transportation networks. They argue that symmetry of the cost-flow relationships is very unnatural when considering several vehicle types; symmetry would, in effect, reduce the problem to one where all vehicles are uniform.

One theme of modelling development that has received a lot of attention during the last 15 years is that of *asymmetric* models, in which the complex relationships that cannot be accounted for by the simple assignment models are captured through the introduction of, usually, asymmetric travel cost functions. The resulting models have been extensively studied from a theoretical and algorithmic viewpoint. These models are studied in Chapter 3. It would seem that the asymmetric models' popularity is a consequence of their mathematical elegance and nice interpretations rather than of their applicability, since real-world applications seem to be lacking. A major reason for this is probably the practical difficulty of calibrating the asymmetric travel time functions. The second possibility for improving the quality of a traffic assignment model is through the introduction of a set of side constraints, modelling the more complex restrictions on possible flow patterns (such as joint capacities on two-way links or bounds on total flows through junctions). We believe this approach to be much more appealing from a practical point of view, since

[12] In the case of additive route costs, i.e., if (2.5f) holds, symmetric link costs imply symmetric route costs. Indeed, using (2.5d),

$$\nabla \mathbf{c}(\mathbf{h}) = \nabla_{\mathbf{h}} \left(\mathbf{\Delta}^{\mathrm{T}} \mathbf{t}(\mathbf{f}) \right) = \mathbf{\Delta} \nabla_{\mathbf{h}} \mathbf{t}(\mathbf{\Delta h}) = \mathbf{\Delta} \nabla \mathbf{t}(\mathbf{f}) \mathbf{\Delta}^{\mathrm{T}},$$

and hence $\nabla \mathbf{c}(\mathbf{h})$ is symmetric whenever $\nabla \mathbf{t}(\mathbf{f})$ is.

it is certainly easier for the traffic engineer to identify a suitable set of side constraints (which may have immediate physical interpretations), than to estimate proper values of parameters in complex travel time functions. This alternative approach has, however, received much less attention than asymmetric models; an account of the development made, and motivations for its use, is given in Section 2.8.2.

2.6 Related network problems

2.6.1 Traffic equilibria and network games

Knight [562] describes the steady-state situation where each traveller minimizes his/her travel cost as an equilibrium state, and later Wardrop [958] gives a similar characterization of the user optimal flow. Charnes and Cooper [160, 162] describe the user equilibrium flow as a non-cooperative Nash [705] equilibrium, in which the players are defined by the O-D pairs, competing to minimize the travel times of their respective commodity flows. Dafermos [209, 205] further discussed along these lines; these first investigations of the relationships between a Wardrop equilibrium and a network game is rather intuitive, and no formal relationships are derived.

A non-cooperative N-person game is formally given by a set of *penalty functions* $\varphi_i : \prod_{i=1}^{N} X_i \mapsto \Re$, defined on the joint strategy space $X \stackrel{\text{def}}{=} \prod_{i=1}^{N} X_i$, and assumed convex on the individual *strategy space* X_i, $i \in \{1, \ldots, N\}$. A point $\mathbf{x}^* \in X$ is a non-cooperative Nash equilibrium if and only if, for each $i \in \{1, \ldots, N\}$,

$$\varphi_i \left(\mathbf{x}_{i-}^*, \mathbf{x}_i^*, \mathbf{x}_{i+}^* \right) = \min_{\mathbf{x}_i \in X_i} \varphi_i \left(\mathbf{x}_{i-}^*, \mathbf{x}_i, \mathbf{x}_{i+}^* \right), \tag{2.45}$$

i.e., if all the players' strategies are optimal with respect to their individual penalty functions, based on the strategies of the other players.

The theory of non-cooperative N-person games was first addressed by Nash [704, 705]; results of the existence and uniqueness of Nash equilibria are given in [790, 466, 615, 544, 380, 398, 435].

Rosenthal [791] studies a discrete version of the user equilibrium traffic assignment problem. In the game defined on the traffic network, players are defined as the individual travellers, with strategy spaces equal to their respective sets of routes available. Travellers choose pure strategies by taking a route to their respective destination, thereby seeking to minimize their payoff function, i.e., their individual travel time. The game is shown to be equivalent to a non-cooperative, pure-strategy Nash game in the traffic network.

This result is extended to the continuous case by Devarajan [251]. He, as do Charnes and Cooper, defines the O-D pairs as the players, and defines a continuum of pure strategies, consisting of the feasible route flows for the fixed demands, for all commodities. The game is restricted to separable travel time functions, and is defined through payoff functions equal to sums of integrated travel times over links used by an O-D pair,

$$\varphi_{pq}(\mathbf{f}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}_{pq}} \int_0^{f_a} t_a(s) ds. \tag{2.46}$$

The Nash game thus defined is equivalent to a Wardrop equilibrium. The same definition of the game is given by Garcia and Zangwill [410, 1005].

More general non-cooperative game formulations of traffic equilibria are given by Fisk [342] and Haurie and Marcotte [468, 469]; see also Colony [179].

Haurie and Marcotte divide the travel made in an O-D pair into a number of players, sharing the same penalty, namely the cost of the routes chosen. The game obtained in the limiting case, when the number of players in each O-D pair tends to infinity, while sharing the same strategy, is shown to be equivalent to a Wardrop equilibrium. The players can not be associated with individual travellers, since a player, as defined, may use several routes simultaneously; in equilibrium, all players divide their flow on all routes used in the O-D pair. This discrepancy from the intuitive game among tripmakers is not surprising; it is inherent in the continuous formulation of the Wardrop conditions that the travellers are infinitesimal.

2.6.2 Discrete traffic equilibrium models

To derive a discrete traffic assignment model, let d_{pq} be the number of tripmakers travelling from an origin p to a destination q, with $D \stackrel{\text{def}}{=} \sum_{(p,q)\in\mathcal{C}} d_{pq}$ being the total number of trips performed. Denote by r_k the route chosen by traveller $k \in \{1,\ldots,D\}$. The *trip pattern* then is the vector $\mathbf{u} \stackrel{\text{def}}{=} (r_1,\ldots,r_D)$. Further let $h_{pqr} = h_{pqr}(\mathbf{u})$ denote the number of trips made on route $r \in \mathcal{R}_{pq}$ in the trip pattern $\mathbf{u}$, and $t_a(f_a)$ the travel time for a vehicle on link a at the volume

$$f_a(\mathbf{u}) = \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr}(\mathbf{u}).$$

Let $c(r_k \mid r_1,\ldots,r_{k-1})$ be the cost of tripmaker k if he/she chooses route r_k and enters the network when there are $k-1$ tripmakers present, using routes $(r_1,\ldots,r_{k-1})$, and assume that $c(r_k \mid r_1,\ldots,r_{k-1}) = \sum_{a\in r_k} t_a(f_a^k)$, where f_a^k denotes the volume on link a when the k first trips have been allocated, i.e., the cost of performing the trip on route r_k is the sum of the costs of the links defining it. Finally, let $c(\mathbf{u})$ denote the *cumulative* user-cost function,

$$c(\mathbf{u}) \stackrel{\text{def}}{=} c(r_1,\ldots,r_D) = \sum_{k=1}^{D} c(r_k \mid r_1,\ldots,r_{k-1}).$$

The ordering of the trips in the definition of the trip pattern is immaterial ([855]). Hence,

$$c(\mathbf{u}) = \sum_{a\in\mathcal{A}} \sum_{k=1}^{f_a} t_a(k). \tag{2.47}$$

The *discrete user equilibrium condition* then is defined as a trip pattern $\mathbf{u}^*$, for which, for each k and r_k with the same origin and destination as r_k^*, we have

$$c\left(r_k^* \mid r_1^*,\ldots,r_{k-1}^*,r_{k+1}^*,\ldots,r_D^*\right) \le c\left(r_k \mid r_1^*,\ldots,r_{k-1}^*,r_{k+1}^*,\ldots,r_D^*\right). \tag{2.48}$$

The definition implies that if $\mathbf{u}^*$ is a user equilibrium then it is not possible for any one of the D tripmakers, given his/her O-D pair, to take an alternate route that is less costly than the route r_k^* already chosen. This also means that it is a Nash equilibrium.

Of course, the costs of the routes used may be different in this case, since trips are necessarily integer valued. The behavioural model, however, is the same as in the continuous Wardrop equilibrium definition.

The model derived above is due to Smith [855] (see also [853, 854, 300, 301], and [63] for a discussion on relationships between discrete and continuous traffic models). The Objective (2.47) is also derived by Rosenthal [791] (see Section 2.6.1) as the objective of a

discrete assignment model, which is shown to correspond to a Nash equilibrium model. A continuous formulation in which the cost functions are appropriate step functions yields the same solutions; it is also clear that for large flows, the Objective (2.47), defined as the sum of travel time functions, is closely approximated by the Objective (2.6a), defined by sums of Riemann integrals. Sufficient conditions on the cost-flow functions for a continuous approximation of the discrete assignment models to be reasonable are found in [963].

The above model may be derived from the *efficiency principle* ([853, 855]). Assume that the trip pattern $\mathbf{u}$ can be described by a probability distribution, which is *efficient* in the sense that a more costly trip pattern is less probable than a less costly one. Also assuming that $c(\mathbf{u})$ is rational, the equilibrium solutions [see (2.48)] are the most probable trip patterns (see also [301, 857, 302]).

In the single-commodity case, a discrete equilibrium flow is obtained by successively assigning tripmakers to the cheapest route. The continuous (single-commodity) model is solved in the limit as the increment of flow tends to zero; this solution principle corresponds to applying, in the limit, an infinite number of steps of incremental assignment ([871, Sec. 5.4.2]).

2.6.3 Traffic equilibria and electrical networks

It has long been recognized that the problem of finding the currents and voltages in a resistive electrical network is an equilibrium problem in a single-commodity network ([181]). An electrical network is composed of a set of links (branches), corresponding to various electrical devices, and nodes, corresponding to connection points between these.

In a steady-state, the voltages and currents in the network are governed by Kirchoff's [552] (equilibrium) laws, which state that:

(1) (Kirchoff's current law) *The current flows are balanced*, i.e.,

$$\sum_{j \in \mathcal{W}_i} f_{ij} - \sum_{j \in \mathcal{V}_i} f_{ji} = d_i, \qquad \forall i \in \mathcal{N},$$

where f_{ij} is the flow of current (in amperes) on branch $(i,j) \in \mathcal{A}$, and where d_i is positive (negative) at current sources (sinks), and zero otherwise.

(2) (Kirchoff's voltage law) *The potential difference between adjacent nodes equals the voltage over the branch connecting the nodes*, i.e.,

$$\pi_i - \pi_j = \Theta_{ij}(f_{ij}), \qquad \forall (i,j) \in \mathcal{A},$$

where π_i is the potential (in volts) at node i, and Θ_{ij} is an increasing function relating the current flow to the voltage on branch (i,j). (If the electrical device on branch (i,j) is a linear resistor, then $\Theta_{ij} = R_{ij} f_{ij}$, i.e., the voltage is given by the product of the resistance R_{ij} (in ohms) and the current on the branch, according to Ohms law.)

Note the similarity between Kirchoff's current law and the flow conservation constraints in the traffic equilibrium model [cf. (2.11)], and between the voltage law and the equilibrium Conditions (2.17). (The major difference between the two models is the single-commodity nature of electrical networks.)

Kirchoff's laws, and other relations reflecting the electrical properties of the devices on the branches, may hence be used to formulate an equilibrium model for currents and

voltages in the electrical network. One may also show that Kirchoff's equilibrium laws may be obtained through the solution of a nonlinear network flow problem, in which the energy loss of the devices is minimized. The objective is composed by the sum of integrals of the voltage-current relations Θ_{ij}, which thereby play the same role as the flow-cost relations t_{ij} in the transportation network. There is, however, no energy interpretation of the integral objective in the traffic model.

It is interesting to note that mathematical programming formulations of the problem of finding equilibrium currents and voltages in electrical networks had already been developed in the middle of the 1940s by Duffin [274, 275] (see also [229, 89, 161, 248, 162, 88, 508, 181]), and hence preceded the (similar) traffic models of Prager [765] and Beckmann *et al.* [47] by almost a decade. Further electrical network analogies of traffic equilibrium problems are presented in [804, 809].

Similar single-commodity flow models arise in other applications, such as in pipe networks ([189, 450, 180, 181, 178]); surveys of applications of nonlinear network flows are found in [32, 591].

2.6.4 Spatial price equilibria

Let $\mathcal{M}$ and $\mathcal{N}$ denote the set of supply and demand markets involved in the production and distribution, respectively, of a commodity. In supply market $i \in \mathcal{M}$, let s_i denote the supply of the commodity, and π_i its supply price. Correspondingly, in demand market $j \in \mathcal{N}$, d_j denotes the demand of the commodity, and ρ_j its demand price. Finally, let f_{ij} denote the shipment between the pair (i,j) of supply and demand markets, and c_{ij} the unit transaction cost which includes transportation costs and in some applications also taxes and/or subsidies.

The *market equilibrium* conditions state that if there is any trade between the market pair (i,j), the supply price at i plus the transaction costs c_{ij} equals the demand price at j; if, however, the supply price plus the transaction costs is larger than the demand price, there is no trading between the pair (i,j). That is, for all $(i,j) \in \mathcal{M} \times \mathcal{N}$,

$$f_{ij} > 0 \implies \pi_i + c_{ij} = \rho_j, \tag{2.49a}$$

$$f_{ij} = 0 \implies \pi_i + c_{ij} \geq \rho_j. \tag{2.49b}$$

To formulate a mathematical model for spatial price equilibrium, feasibility constraints are introduced, stating that supplies are cleared and demands are satisfied:

$$\sum_{j \in \mathcal{N}} f_{ij} = s_i, \qquad \forall i \in \mathcal{M}, \tag{2.50a}$$

$$\sum_{i \in \mathcal{M}} f_{ij} = d_j, \qquad \forall j \in \mathcal{N}, \tag{2.50b}$$

$$f_{ij} \geq 0, \qquad \forall (i,j) \in \mathcal{M} \times \mathcal{N}. \tag{2.50c}$$

Further, we assume that the supply price is dependent on the supply of the commodity,

$$\boldsymbol{\pi} = \boldsymbol{\pi}(\mathbf{s}),$$

the demand price on the demand of the commodity,

$$\boldsymbol{\rho} = \boldsymbol{\rho}(\mathbf{d}),$$

and the transaction costs on the shipments being made,

$$\mathbf{c} = \mathbf{c}(\mathbf{f}).$$

Note that the problem here defined can be made more general by introducing a general network, so that a transaction may involve choosing several routes for transportation, and by introducing more than one commodity being traded.

If the supply price function $\boldsymbol{\pi}$, the demand function $\boldsymbol{\rho}$, and the transaction cost function $\mathbf{c}$ are separable, then the equilibrium conditions may be obtained by minimizing

$$\sum_{i\in\mathcal{M}}\int_0^{s_i}\pi_i(x)dx+\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{N}}\int_0^{f_{ij}}c_{ij}(x)dx-\sum_{j\in\mathcal{N}}\int_0^{d_j}\rho_j(x)dx \tag{2.51}$$

over the set of Constraints (2.50). In the general case, where cost and demand functions are asymmetric, the problem may be formulated as a variational inequality ([702, Chap. 3]).

Network-based models of spatial price equilibrium problems were already known to Cournot [186], and have developed rapidly since the pioneering works of Enke [297], Koopmans [566], Samuelson [806, 807], and Takayama and Judge [886, 887]. For more reading on models of economical problems on networks, and their relationships to traffic equilibrium problems, see [199, 197, 532, 464, 702].

It is interesting to note that although traffic and economic network equilibrium problems are highly interrelated, the development of models and methods for these problems have to a large extent taken parallel paths.

2.6.5 Optimal message routing in computer communication networks

A problem with a structure similar to [TAP] is that of finding the optimal routing of data messages in a computer communications network. Let d_{pq} denote the rate of messages (in bits per second) entering the computer network from an origin node p and exiting the network at a destination node q. The routing problem is to segment the messages into *packets*, and send them along routes to their destinations so as to minimize the average delay of the data packets. At intermediate nodes, the packets are stored in *queues* for each outgoing channel, and sent forward when the channel becomes free. The queueing delays of messages at the links are not easily quantified, but there are simplified models that represent queueing delay at a link as a function of the packet arrival rate at the link. A commonly used delay formula is

$$t_{ij}(f_{ij})\stackrel{\text{def}}{=}\frac{f_{ij}}{c_{ij}-f_{ij}}+p_{ij}f_{ij},\qquad\forall(i,j)\in\mathcal{A}, \tag{2.52}$$

where f_{ij} is the arrival rate of packets at link (i,j), c_{ij} the transmission capacity of the link, and p_{ij} the processing and propagation delay on the link.

Such simplified models give rise to multicommodity flow problems, in which a function of the form (2.52) is to be minimized over flow conservation constraints ([374]). Many of the algorithms presented for the problem [TAP] in Chapter 4 have originally been proposed for the solution of routing problems. For further reading, see [556, 997, 998, 141, 421, 654, 65].

2.7 Discussion

The mathematical models presented in this chapter are based on Wardrop's behavioural principles. They are intuitively very appealing, and have had successful applications

in a wide variety of social sciences. They may, however, be criticized for being too simplified and based on unrealistic assumptions on traffic network characteristics and traveller behaviour.

The simplicity of a model has its merits, of course, and the more complex models presented the last 15 years have had very little practical use. As Dow [941] puts it:

> In fact it often appears that in practice a model which is easy to analyze but not particularly accurate is preferred to a more accurate model which is difficult to analyze.

With the exception of few studies (e.g., [360, 361, 974, 941, 288, 99, 602]) no published articles on traffic equilibrium models and methods are devoted to empirical testing to validate the models' practical use. An important reason is the expense and difficulty of collecting data for performing such a task. Another reason is the lack of publications from practitioners of transportation planning methods.

As a consequence, the support for network equilibrium models is almost exclusively a result of the underlying theory. Not even the theoretical foundations of the models are well developed, however. The use of Wardrop's equilibrium principles as formulations of problems to be solved with efficient mathematical programming techniques rather than—as was the original purpose—as behavioural principles to be analyzed and, if appropriate, be utilized in the building of proper mathematical models, has had a tremendous impact on the directions research in the area has taken. The vast amount of research being performed at universities across the world on traffic equilibrium models today is actually based on a very weak scientific foundation—this is quite worrying.

The separability of the travel time and demand functions were early recognized as too restrictive in certain applications; in particular, such simple formulas can be used to model realistically neither different vehicle types nor the interaction among vehicles at intersections. The development of asymmetric assignment models arose from such observations. (These models are described and analyzed in Chapter 3.)

Unfortunately, this development began before the symmetric models were fully understood, and most of the research that has been made on the asymmetric models is motivated more by their elegance and by the challenge involved in devising efficient methods for their numerical solution rather than by their appropriateness. The difficulty in describing the complex flow-cost relationships in these general models is much more pronounced than in the separable models, and the asymmetric models are not used in practice. One may in fact argue that most of the models that have been developed since the pioneering days of Wardrop and Beckmann are models that should be utilized for describing an idealized equilibrium state rather than for the numerical solution of practical problems.

The fundamental principles underlying the assignment models were stated some 40 years ago. The traffic flows in the then relatively uncongested urban networks were probably suitable for approximation by steady-state flows, as Wardrop did. Since those days, the traffic networks have become much more complex and the demand for transportation have become orders of magnitude higher, and the approximation of present traffic flows by steady-state flows is far less realistic. During the last ten years, *dynamic assignment models*, which take the time-varying character of traffic flows into account, have received increased attention since their first appearance in the pioneering work of Merchant and Nemhauser [665, 666], particularly since the advent of systems for vehicle guidance and dynamic flow control and management ([105, 107, 943]). So far, no well-founded dynamic models free from any serious anomaly such as instant propagation of some travellers, infinite cycling, failure to recognize the first-in-first-out principle, etc., have appeared, and their numerical solution most often rely on a time-discretization which brings the dynamic model into a (typically very large) static one. In this book, we concentrate on the

static models, and refer the reader to the following articles and their respective references: [995, 146, 148, 969, 383, 970, 851, 852, 111, 149, 272, 28, 271].

Even in dynamic models, an equilibrium assumption is present. Researchers have begun to question if this fundamental basis of assignment models is realistic; see, e.g., Bell and Bennett [54].

An implicit assumption in the user equilibrium principle is that every traveller has both full and accurate information about all the alternatives and their characteristics, and also a uniform travel cost perception and route-choice behaviour. In stochastic models, differences in perceived costs and route-choice characteristics are modelled by introducing random components in the travel time formulas. These models are introduced in Section 2.8.1.

The assignment problem is only part of the transportation planning process; as discussed in Section 1.8, to be able to accurately predict future traffic distribution, traffic models should be integrated rather than analyzed in sequence. During the last 20 years, transportation planning research has focused much of the research efforts on improving predictive modelling. Recognizing the strong interrelation between users' decision making and the performance of transportation systems, the trend has been towards integrated modelling approaches. In *combined traffic assignment models*, parts of the planning process are analyzed simultaneously; this is a major advantage to the sequential approach in the traditional transportation planning models, since they guarantee internal consistency. For an introduction to this area, see [694, 863, 904, 905, 978, 979, 304, 208, 305, 306, 108, 619, 299, 620, 150, 104, 632, 381, 154, 803, 302, 166, 58, 382, 577, 725, 881], and the relevant references cited therein.

The assignment models discussed so far include constraints that ensure that the demand is satisfied, but not normally the constraints that ensure that traffic capacity, speed limits, etc., are met. Examples of traffic assignment problems with side constraints are given in Section 2.8.2. We also show that formulating traffic models with side constraints may be an interesting alternative to the asymmetric models.

The basic assignment model and its properties is discussed in several expository articles; see, e.g., [48, 131, 530, 713, 717, 774, 216, 413, 326, 103, 104, 351, 630, 352, 110, 675, 358].

2.8 Some extensions

2.8.1 Stochastic assignment models

Introduction

The notion of a user equilibrium is intimately associated with each traveller having accurate information about travel costs, and all travellers being uniform and rational in their decision-making. The user equilibrium model is therefore known as a *deterministic* model of traffic assignment.

In reality, travellers' perception of travel times are subject to variations, and routes are chosen based on perceived travel times rather than the actual travel times. The routes utilized are therefore not necessarily only those that are the shortest, nor do they necessarily have the same actual travel time. To reflect the variations in travellers' travel cost perception, in a *non-deterministic* (or *stochastic*) assignment model, a random component is added to the travel cost function.

Let $T_a(\mathbf{f})$ denote the perceived travel time on link $a \in \mathcal{A}$ for an arbitrary user, and associate with each route $r \in \mathcal{R}_{pq}$ a perceived travel time C_{pqr}, distributed across the

population of drivers. Assuming travel time additivity,

$$C_{pqr}(\mathbf{h}) = \sum_{a \in \mathcal{A}} \delta_{pqra} T_a(\mathbf{f}).$$

The function C_{pqr} expresses the probability that a randomly chosen tripmaker associates a given travel time with the route, or, equivalently, the probability with which he/she will choose to make the trip along the route in question.

Let $P_{pqr} = P_{pqr}(\mathbf{c})$ be the probability that route $r \in \mathcal{R}_{pq}$ is perceived as the shortest, given actual travel times, $\mathbf{c}$, i.e.,

$$P_{pqr}(\mathbf{c}) \stackrel{\text{def}}{=} \Pr\left(C_{pqr} \leq C_{pql},\ \forall l \neq r,\ l \in \mathcal{R}_{pq} \mid \mathbf{c}\right).$$

The random variable C_{pqr} is assumed to be given by the sum of the actual travel time, c_{pqr}, and a (flow-dependent) random error term, ξ_{pqr}, which may vary from traveller to traveller, with mean value $\mathrm{E}(\xi_{pqr}) = 0$. This means that

$$\mathrm{E}\left(C_{pqr}\right) = \mathrm{E}\left(c_{pqr} + \xi_{pqr}\right) = c_{pqr},$$

so that, on average, perceived travel times equal actual travel times. (This relation also follows from $\mathrm{E}(T_a) = t_a$, and travel time additivity.)

The difference in perceived travel costs is accounted for by the variance of the stochastic variables $\boldsymbol{\xi}$. In mildly congested networks, the variations in perceived costs can be expected to be more significant compared to the actual costs than in heavily congested networks; the consequence is that stochastic models are more applicable to lightly congested networks (see [934]).

The distribution of flows in a stochastic assignment model depends on the probability distribution of the stochastic variables $\boldsymbol{\xi}$. The two flow distributions most frequently applied in stochastic assignment are the *logit* and *probit* models, which correspond to choosing an independent Weibull–Gumbel and a normal probability distribution, respectively.

In the following, we introduce the definition of stochastic user equilibrium, study the two most important approaches employed to disutility perception modelling, and present a general optimization formulation for obtaining a stochastic user equilibrium flow.

Stochastic user equilibrium

The natural extension of Wardrop's principle of user equilibrium to the non-deterministic case is to define a stochastic user equilibrium situation as one in which no user *believes* he/she can improve the travel time by unilaterally changing routes ([220]). In other words, *perceived* travel times are equal on the utilized routes within an O-D pair.

The stochastic user equilibrium conditions can be characterized by the equations[13]

$$d_{pq} P_{pqr}(\mathbf{c}) = h_{pqr}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{2.53}$$

[13]These conditions are natural, considering the weak law of large numbers; if d_{pq} is large, and if the travellers act independently, then

$$P_{pqr}(\mathbf{c}) \approx \frac{h_{pqr}}{d_{pq}}.$$

Also, the network flow constraints

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.54a}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{2.54b}$$

must hold.

Note that link and route flows are random variables, based on the perceived travel time distribution ([214]). The variables $\mathbf{f}$ and $\mathbf{h}$ should therefore be recognized as the means of these random variables. Note also that if route flows are given by (2.53), then (2.54a) is satisfied automatically.

To show that the stochastic user equilibrium conditions generalize the (deterministic) user equilibrium conditions, we redefine the route choice probability as

$$P_{pqr}\in\left[\Pr\left(C_{pqr}<C_{pql},\ \forall l\neq r,\ l\in\mathcal{R}_{pq}\mid\mathbf{c}\right),\Pr\left(C_{pqr}\leq C_{pql},\ \forall l\neq r,\ l\in\mathcal{R}_{pq}\mid\mathbf{c}\right)\right],$$

for all $r\in\mathcal{R}_{pq}$, $(p,q)\in\mathcal{C}$, by which the stochastic user equilibrium condition may be written as ([831, Chap. 12])

$$\lambda_{pqr}\in\left[\Pr\left(C_{pqr}<C_{pql},\ \forall l\neq r,\ l\in\mathcal{R}_{pq}\mid\mathbf{c}\right),\Pr\left(C_{pqr}\leq C_{pql},\ \forall l\neq r,\ l\in\mathcal{R}_{pq}\mid\mathbf{c}\right)\right], \tag{2.55}$$

for all $r\in\mathcal{R}_{pq}$, $(p,q)\in\mathcal{C}$, where $\lambda_{pqr}=h_{pqr}/d_{pq}$ is the portion of the demand that will utilize route $r\in\mathcal{R}_{pq}$ in the transportation.

The stochastic user equilibrium Condition (2.55) is applicable to both continuous and discrete random variables (C_{pqr}) of perceived travel times. In the continuous case, the interval defined in (2.55) degenerates into a point, and (2.55) reduces to (2.53). In the discrete case, then, (2.55) generalizes (2.53).

To see that (2.55) extend the deterministic equilibrium conditions, we note that in the deterministic case, the probability statements in (2.55) are either one or zero. If $h_{pqr}>0$, the route-choice probability is one, and route r is therefore a shortest route. If $h_{pqr}=0$, the route-choice probability must be zero, in which case route r can not be a shortest route. These conditions are equivalent to the (deterministic) user equilibrium conditions.

A mathematical program for stochastic user equilibrium

Sheffi and Powell [833] present an unconstrained mathematical program whose optimality conditions coincide with the stochastic user equilibrium conditions. The travel costs $\mathbf{t}$ are assumed positive, strictly increasing and twice continuously differentiable, and the distribution of perceived travel costs are translationary invariant.

The program is as follows.

[TAP-SUE]

$$\min\ T(\mathbf{f},\mathbf{h}) \stackrel{\text{def}}{=} -\sum_{(p,q)\in\mathcal{C}} d_{pq}E\left(\min_{r\in\mathcal{R}_{pq}}\{C_{pqr}\}\,\middle|\,\mathbf{c}(\mathbf{h})\right)+\sum_{a\in\mathcal{A}}t_a(f_a)f_a-\sum_{a\in\mathcal{A}}\int_0^{f_a}t_a(s)ds.$$

To show that the unconstrained minima of T coincide with stochastic user equilibrium solutions, we characterize the stationary points of T. We first use a result by Williams [976] that

$$\frac{\partial}{\partial c_{pqr}}E\left(\min_{r\in\mathcal{R}_{pq}}\{C_{pqr}\}\,\middle|\,\mathbf{c}(\mathbf{h})\right)=P_{pqr}(\mathbf{c}), \tag{2.56}$$

to obtain that ([833])

$$\frac{\partial}{\partial h_{pqr}}E\left(\min_{r\in\mathcal{R}_{pq}}\{C_{pqr}\}\,\middle|\,\mathbf{c}(\mathbf{h})\right)=P_{pqr}(\mathbf{c})\sum_{a\in\mathcal{A}}\delta_{pqra}t_a'(f_a).$$

We thus obtain that

$$\frac{\partial T(\mathbf{f},\mathbf{h})}{\partial h_{pqr}}=(-d_{pq}P_{pqr}(\mathbf{c})+h_{pqr})\sum_{a\in\mathcal{A}}\delta_{pqra}t_a'(f_a),$$

which must be zero at a stationary point. From the assumptions on $\mathbf{t}$, we must then have

$$h_{pqr} = d_{pq} P_{pqr}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.57}$$

i.e., the stochastic user equilibrium Condition (2.53). By summing (2.57) over the routes in each O-D pair, we obtain the demand feasibility constraints, which thus are fulfilled automatically at a stationary point of [TAP-SUE].

The objective of [TAP-SUE] is non-convex in general, but it is shown in [833, 831] that there is only one stationary point, and that the objective is strictly convex in the link flow variables in the neighbourhood of this point, which hence is the unique stochastic equilibrium flow. (There may, however, be more than one stochastic equilibrium *route* flow solution.)

Although there have been many suggestions of models of error perception, (see, e.g., [663, 951, 135]), the most popular are the logit and probit distribution models.

The logit-based stochastic model

In the logit assignment model, trips are assumed to be distributed according to the formula

$$h_{pqr} = d_{pq} \frac{\mathrm{e}^{-\Theta c_{pqr}}}{\sum_{l \in \mathcal{R}_{pq}} \mathrm{e}^{-\Theta c_{pql}}}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.58}$$

where Θ is a positive parameter associated with the random cost component.

The logit model is derived from the assumption that the random components in the travel costs are identically and independently distributed *Weibull–Gumbel variates* ([441, 264, 831]). Under this assumption, the perceived travel time is

$$C_{pqr} = c_{pqr} - \frac{1}{\Theta} \varepsilon_{pq}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.59}$$

where ε_{pq} is a Gumbel variate, and the parameter Θ is used to calibrate the variance in the cost perception, sometimes referred to as the *perception error*.

It follows from (2.59) that if the value of Θ is large, the perception error is small, and travellers will tend to choose minimum-cost routes. Indeed, the cost perception tends towards being accurate as $\Theta \to +\infty$. A small value of Θ indicates a large variance in the perception of travel cost, with travellers choosing routes with considerably larger actual costs than those of the least-cost ones. It is also clear from the logit Formula (2.58) that for all values of Θ, *all* routes receive flow, regardless of their travel times; the flow on a route, however, monotonically decreases with an increasing actual cost. In the limit, when $\Theta \to 0$, all routes within an O-D pair receive an equal share of the O-D flow.

The logit model of assignment was first studied for the case of constant travel times by Dial [254], who also presented algorithms for assigning logit-based flows on the links of a traffic network (see Section 4.5.2). The model is further developed in [442, 898, 772]; Fisk [334, 336] discusses the calibration of the parameter Θ.

A logit model for flow-dependent travel times is presented by Fisk [336]. Consider the mathematical program

[TAP-SUE-L]

$$\min\ T(\mathbf{f}, \mathbf{h}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s) ds + \frac{1}{\Theta} \sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} h_{pqr} \log h_{pqr}, \tag{2.60a}$$

subject to

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{2.60b}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{2.60c}$$

$$\sum_{(p,q)\in\mathcal{C}}\sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a\in\mathcal{A}. \tag{2.60d}$$

The problem [TAP-SUE-L] is, for every positive value of Θ, a strictly convex program in both the link and route flow variables, and thus has the advantage over the deterministic assignment problem of providing *unique route flows*. (Here, $0\log 0$ is defined as zero.) Note that in the limit of $\Theta\to+\infty$, the problem [TAP] is obtained.

To show that the unique optimal solution of [TAP-SUE-L] satisfies (2.58), we associate a set of multipliers (π_{pq}) with the Constraints (2.60b), and formulate the Lagrangean function

$$L(\mathbf{f},\mathbf{h},\boldsymbol{\pi}) \stackrel{\text{def}}{=} T(\mathbf{f},\mathbf{h}) + \sum_{(p,q)\in\mathcal{C}} \pi_{pq}\left(d_{pq} - \sum_{r\in\mathcal{R}_{pq}} h_{pqr}\right).$$

In an optimal solution, $h_{pqr}>0$ must hold for every $r\in\mathcal{R}_{pq}$, $(p,q)\in\mathcal{C}$.

Solving the equations

$$\frac{\partial L(\mathbf{f},\mathbf{h},\boldsymbol{\pi})}{\partial h_{pqr}} = \frac{1}{\Theta}(\log h_{pqr}+1) + c_{pqr} - \pi_{pq} = 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C},$$

yields

$$h_{pqr}(\pi_{pq}) = \mathrm{e}^{\Theta(\pi_{pq}-c_{pqr})-1}, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}. \tag{2.61}$$

Inserting the Expression (2.61) into the relaxed constraint finally yields that $\mathbf{h}$ satisfies (2.58), that is, that the unique solution to [TAP-SUE-L] is a logit flow distribution. It may also be shown that the solution of [TAP-SUE] yields the logit distribution of flows, given the above choice of perception error ([833]).

The problem [TAP-SUE-L] is intractable for any non-trivial network because of the property that every route is utilized. It is, however, possible to obtain optimal *link flows* by iterative methods, and for known subsets $\hat{\mathcal{R}}_{pq}$ of the set of routes $\mathcal{R}_{pq}$, $(p,q)\in\mathcal{C}$, optimal route flows may be obtained. See Section 4.5.2 for further details.

The use of the deterministic model assumes complete information about the alternatives available to each traveller, i.e., each traveller accurately perceives the travel costs on each route. This model results from the choice $\Theta=+\infty$. If, on the other hand, *no* information is available, then the most probable macrostate is that which maximizes the entropy, i.e., maximizes $\{-\sum_{(p,q)\in\mathcal{C}}\sum_{r\in\mathcal{R}_{pq}} h_{pqr}\log h_{pqr}\}$ subject to the demand feasibility constraints. This objective is obtained from T by choosing $\Theta=0$. By choosing the value of Θ within the interval $[0,+\infty]$, the corresponding route-choice problem is one where the amount of information available to the travellers varies from no information to full information. (An equivalent way of expressing this is, by assuming that the traveller makes use of all the information available, that the traveller ranges from being completely insensitive to the travel costs to being a cost-minimizer.) In terms of travel perception, by letting the value of Θ range from 0 to $+\infty$ the perception tends from totally random (no correlation with the travel times) to the case of variance zero. The value of Θ may be viewed as an aggregated measure of the amount of (or the accuracy of the) information that is available among the tripmakers about the actual travel costs.

The many nice properties of the entropy function makes the logit approach the natural choice in the modelling of stochastic user equilibria. The logit model is, however, associated with some serious drawbacks. In the extreme case where $\Theta=0$, the route-choice

probabilities are $\lambda_{pqr} = 1/|\mathcal{R}_{pq}|$ for all $r \in \mathcal{R}_{pq}$, i.e., only the number of routes in an O-D pair affects the flow distribution. This dependency on the network topology is significant also for relatively large values of Θ, and results in the solution to [TAP-SUE-L] being dependent on the representation of the network; this property is very unwanted in any model.

One of the consequences is that overlapping routes receive overestimates of flows, and that the larger the number of routes passing a particular link, the more flow it receives, regardless of the travel cost on it. The solution to the stochastic user equilibrium problem based on the logit model suffers from these drawbacks even for moderate values of Θ, and may, as a consequence, provide very inaccurate flows in certain applications. However, for networks where the error in the perceived travel costs are mildly correlated and have a similar probability distribution, the logit model is applicable.

The route-choice anomaly results from the inability of the logit model to account for the correlation between the cost of the different alternatives, and stems from the independence assumption on the random cost terms (termed the axiom of independence from irrelevant alternatives; see [622]). Daganzo and Sheffi [220] argue that this must be the case for any model assuming an independence between the random variables. (This would include for instance the model proposed by Von Falkenhausen [951].)

Another deficiency results from the fact that the random components are *identically* distributed with the same variances; the route choice probabilities are thus solely based on absolute travel time differences, and do not take into account the magnitude of costs.

Further discussions of the properties of the logit model are given in [659, 816, 136, 356, 220, 57, 831].

The probit-based stochastic model

In a probit route choice model, the cost perception errors are assumed to be distributed according to a multivariate normal law. As with the logit model, the mean perception error is zero. The variance of C_{pqr} equals βc_{pqr}, where β is a proportionality constant; the covariance between the perceived travel costs on overlapping routes is given by

$$\mathrm{cov}(C_{pqr}, C_{pql}) = \beta \sum_{a \in \mathcal{A}} \delta_{pqra} \delta_{pqla} t_a .$$

Hence, overlapping routes are correlated, and the variance depends on the travel costs; this is a major improvement over the logit model.

The probit-based assignment model can, in general, not be given an analytical formulation of an optimization problem because of the non-analytical formulation of the route-choice probabilities.[14] Because of the large number of alternatives (routes) available, analytical approximations such as Clark's [172] method are also less useful; see also [580]. Methods for probit-based route choice models invariably use simulation techniques for the numerical calculation of perceived travel times; see Section 4.5.2 for further details.

Another line of development is that of Soroush and Mirchandani [864]; in their stochastic model, the network itself is stochastic.

[14] From (2.56), the formulation of the probit-based assignment model in the framework of the program [TAP-SUE] is possible only if the variance of perceived travel costs is independent of the mean perceived travel costs ([831, Sec. 12.1]); the variance can instead be defined by the free-flow travel times or link lengths.

2.8.2 Side constrained assignment models

The inherent simplicity of the traffic assignment problem makes it inapplicable to more complex real-world traffic problems (e.g., [823]). For instance, it does not capture the interactions between the flow on intersecting links, or between vehicles of different types. An illustrative example of the inapplicability of the basic model and its possible consequences is provided by Hearn [473], who comments on its property of allowing every road to carry arbitrarily large volumes of traffic. This deficiency in the model causes that

> the predicted flow on some links will be far lower or far greater than the traffic engineer knows they should be *if all assumptions of the model are correct.* In practice, the result is that the model predictions are ignored, or, more often, the user will perturb the components of the model (trip table, volume delay formulas, etc.) in an attempt to bring the model output more in line with the anticipated results.

In order to improve the model's ability to accurately describe, reproduce, or predict a real-world traffic situation, two fundamentally different approaches may be utilized.

The traditional approach is to capture additional flow relationships through the introduction of nonseparable, and typically also asymmetric, travel cost functions. A solution to the Wardrop conditions can then, however, not be obtained through the solution of an optimization model of the form [TAP], due to the non-integrability of the resulting travel cost function. Instead, the Wardrop conditions are formulated as non-optimization models, such as variational inequalities. The resulting class of models has been extensively studied both from a theoretical and an algorithmical point of view (see Chapter 3). Due mainly to the practical difficulty of calibrating asymmetric travel cost functions, real-world applications are scarce however.

The alternative—which has so far been surprisingly little studied—approach to improve the quality of the basic equilibrium model is to introduce a set of side constraints to model additional restrictions on possible flow patterns. Such side constraints could be used to describe, for instance, the interaction of vehicles in a junction, joint capacities on two-way streets and links in intersections and roundabouts, requirements that observed flows on some links should be reproduced in the calculated solution, a traffic control policy, or dynamic aspects. We believe this approach is appealing from a practical point of view, since it is certainly easier for the traffic engineer to identify a suitable set of side constraints—which may have immediate physical interpretations—than to estimate proper values of parameters in complex travel cost functions. (In the example provided by Hearn [473], the proper improvement of the basic model is the introduction of link capacity constraints corresponding to the engineer's anticipation of reasonable levels of traffic flow.) The approach of introducing side constraints in traffic equilibrium models was first discussed by Larsson and Patriksson [587, 589].

Although this alternative approach seems to be more useful than that based on asymmetric cost functions, it has been given comparatively very limited attention. We present a general side constrained assignment model and investigate its optimality conditions; these may be interpreted as a generalization of Wardrop's equilibrium Principle (2.1) in the respect that an equilibrium holds for a well defined *generalized* cost function. Moreover, we show that the side constrained assignment problem may be equivalently solved as a standard equilibrium model using this travel cost function. This result leads to an interesting relationship between side constrained and asymmetric models of traffic equilibria, which motivates the further study of side constrained models.

A side constrained assignment model

Let $g_k : \Re_+^{|\mathcal{A}|} \mapsto \Re$, $k \in \mathcal{K}$, be convex and continuously differentiable, and define the side constraints

$$g_k(\mathbf{f}) \leq 0, \qquad \forall k \in \mathcal{K}.$$

Here, the index set $\mathcal{K}$ may, for instance, consist of the index set of the links, nodes, routes, or O-D pairs, or any combination of subsets of them. The constraints are, without any loss of generality, given as inequalities.

Consider the general side constrained traffic equilibrium problem

[TAP-SC]

$$\min\ T(\mathbf{f}) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s)ds, \tag{2.62a}$$

subject to

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{2.62b}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.62c}$$

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a \in \mathcal{A}, \tag{2.62d}$$

$$g_k(\mathbf{f}) \leq 0, \qquad \forall k \in \mathcal{K}. \tag{2.62e}$$

We assume that the feasible set of [TAP-SC] is non-empty; in cases where some functions g_k are nonlinear, we also assume that a constraint qualification (e.g., [43, Chap. 5]) holds. The convexity of [TAP-SC] then ensures the existence of an optimal solution, which is unique in the total link flows and characterized by the first-order optimality conditions. We next show that the optimality conditions of [TAP-SC] give rise to a Wardrop equilibrium principle in terms of *generalized route travel costs.*

Theorem 2.8 (A generalization of the Wardrop principle) *Let $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ and $\boldsymbol{\beta} \in \Re^{|\mathcal{K}|}$ be vectors of optimal Lagrange multipliers for the Constraints (2.62b) and (2.62e), respectively. If $(\mathbf{h}, \mathbf{f})$ solves the problem* [TAP-SC]*, then*

$$h_{pqr} > 0 \implies \overline{c}_{pqr} = \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq}, \tag{2.63a}$$

$$h_{pqr} = 0 \implies \overline{c}_{pqr} \geq \pi_{pq}, \qquad \forall r \in \mathcal{R}_{pq} \tag{2.63b}$$

holds for all O-D pairs $(p,q) \in \mathcal{C}$, where

$$\overline{c}_{pqr} \stackrel{\text{def}}{=} c_{pqr}(\mathbf{h}) + \sum_{a \in \mathcal{A}} \delta_{pqra} \left(\sum_{k \in \mathcal{K}} \beta_k \frac{\partial g_k(\mathbf{f})}{\partial f_a} \right), \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{2.64}$$

Proof Stating the stationary point conditions for the Lagrangean function

$$L(\mathbf{f}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} T(\mathbf{f}) + \sum_{k \in \mathcal{K}} \beta_k g_k(\mathbf{f}) \tag{2.65}$$

subject to (2.62b)–(2.62d) we obtain, from the convexity of [TAP-SC], that $(\mathbf{h}, \mathbf{f})$ is a solution if and only if

$$h_{pqr}\,(\overline{c}_{pqr} - \pi_{pq}) = 0, \quad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.66a}$$

$$\overline{c}_{pqr} - \pi_{pq} \geq 0, \quad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.66b}$$

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \quad \forall (p,q) \in \mathcal{C}, \tag{2.66c}$$

$$h_{pqr} \geq 0, \quad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{2.66d}$$

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \quad \forall a \in \mathcal{A}, \tag{2.66e}$$

$$\beta_k g_k(\mathbf{f}) = 0, \quad \forall k \in \mathcal{K}, \tag{2.66f}$$

$$g_k(\mathbf{f}) \leq 0, \quad \forall k \in \mathcal{K}, \tag{2.66g}$$

$$\beta_k \geq 0, \quad \forall k \in \mathcal{K}, \tag{2.66h}$$

where $\overline{c}_{pqr}$ is given by (2.64).

The Condition (2.66b), together with (2.66a) and (2.66c), implies that the multiplier π_{pq} is the minimum generalized travel cost $\overline{c}_{pqr}$ in O-D pair (p,q), and (2.66a) further states that these costs are equal for all routes utilized in the O-D pair. Hence, the Conditions (2.66a)–(2.66b) imply (2.63), and the theorem is proved. □

Solutions to [TAP-SC] thus are flows satisfying a generalization of the Wardrop equilibrium conditions, based on the generalized travel Costs (2.64) [as opposed to the *actual* travel costs in the Wardrop equilibrium Condition (2.1)]. (If, for some route, no constraint feasibility in (2.62e) is affected by the flows on the links defining the route, then its generalized route travel cost equals the actual route cost. The same conclusion holds if the Constraints (2.62e) that are affected by these flows are satisfied with strict inequality [cf. (2.66f)].)

The interpretations of the optimal Lagrange multipliers and the Conditions (2.63) depend on the form of the constraint functions. For example, in the case of simple upper bounds on the link flows ($\mathcal{K} = \mathcal{A}$ and $g_a(\mathbf{f}) = f_a - u_a$, $u_a \in [0, +\infty]$, for each $a \in \mathcal{A}$), (2.64) reduces to $\overline{c}_{pqr} = \sum_{a \in \mathcal{A}} \delta_{pqra}(t_a(f_a) + \beta_a)$, and the multipliers $\boldsymbol{\beta}_a$ may be associated with equilibrium queueing delays on saturated links, and $\boldsymbol{\pi}_{pq}$ with the (minimal) sum of total travel cost and queueing delay in each O-D pair; see below. The reader should note that the optimal multipliers $\boldsymbol{\beta}$ are not necessarily unique.

An equivalent standard assignment problem

The side constrained assignment model [TAP-SC] may be solved as an equivalent, convex, standard traffic equilibrium problem, with an appropriately chosen adjustment of the travel costs, referred to as [TAP-A].

Theorem 2.9 (An equivalent standard assignment problem) *The solution set of* [TAP-SC] *equals that of a standard traffic assignment model with travel cost mapping*

$$\mathbf{t}(\cdot) + \nabla \mathbf{g}(\cdot)\boldsymbol{\beta}, \tag{2.67}$$

where $\boldsymbol{\beta}$ is an arbitrary vector of optimal Lagrange multipliers for the Constraints (2.62e).

Proof Consider the Lagrangean Function (2.65). It follows from the strict convexity of T that the optimal solution to [TAP-SC] is obtained from the solution to the Lagrangean subproblem

$$\begin{aligned} \min \quad & L(\mathbf{f}, \boldsymbol{\beta}), \\ \text{s.t.} \quad & (2.62b)\text{–}(2.62d), \end{aligned}$$

defined for optimal multipliers $\boldsymbol{\beta}$ (see the discussions following [43, Th. 6.5.1]). But this is a standard traffic equilibrium problem with objective $L(\cdot, \boldsymbol{\beta})$ and link cost mapping $\nabla L(\cdot, \boldsymbol{\beta}) = \mathbf{t}(\cdot) + \nabla \mathbf{g}(\cdot)\boldsymbol{\beta}$. □

Hence, the link travel cost Mapping (2.67) provides a precise statement of the influence of the side constraints on the travel cost perception of the users of the traffic network, and therefore on their route choice behaviour. Note that the link travel cost Mapping (2.67) is, through (2.62d), equivalent to the generalized route travel cost mapping defined by (2.64).

The variational inequality problem corresponding to the first-order optimality conditions of the problem [TAP-A] is to find an $\mathbf{f}^* \in F^r$ such that

$$[\mathbf{t}(\mathbf{f}^*) + \nabla \mathbf{g}(\mathbf{f}^*)\boldsymbol{\beta}]^{\mathrm{T}}(\mathbf{f} - \mathbf{f}^*) \geq 0, \qquad \forall \mathbf{f} \in F^r.$$

In contrast to the variational inequality formulation [VIP] this problem is symmetric, since its cost Mapping (2.67) is integrable.

Using non-optimal Lagrange multipliers $\boldsymbol{\beta}$ in the travel cost Mapping (2.67) corresponds to solving a perturbed version of [TAP-SC]. The below corollary is immediate from Theorem 2.9 and Everett's Theorem (e.g., [590, Th. 8.3, p. 402]).

Corollary 2.1 (An Everett-type result for [TAP-SC]) *If the Lagrange multipliers $\boldsymbol{\beta}$ employed in the travel cost Mapping (2.67) are non-optimal, then the solution $\mathbf{f}(\boldsymbol{\beta})$ to the resulting standard assignment problem solves*

$$\min \; T(\mathbf{f}),$$

subject to

$$\begin{aligned}
\sum_{r \in \mathcal{R}_{pq}} h_{pqr} &= d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \\
h_{pqr} &\geq 0, \qquad \forall r \in \mathcal{R}_{pq}, \; \forall (p,q) \in \mathcal{C}, \\
\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} &= f_a, \qquad \forall a \in \mathcal{A}, \\
g_k(\mathbf{f}) &\leq \overline{g}_k, \qquad \forall k \in \mathcal{K},
\end{aligned}$$

where

$$\overline{g}_k \overset{\text{def}}{=} \begin{cases} g_k(\mathbf{f}(\boldsymbol{\beta})), & \text{if } \beta_k > 0, \\ \max\{0, g_k(\mathbf{f}(\boldsymbol{\beta}))\}, & \text{if } \beta_k = 0. \end{cases}$$

Note that when the multipliers $\boldsymbol{\beta}$ tend to optimal ones, the perturbed problem tends to [TAP-SC], since $\overline{g}_k$ then tends to zero for all $k \in \mathcal{K}$.

In order to find (near-)optimal values of $\boldsymbol{\beta}$ one may solve the Lagrangean dual problem

$$\max_{\beta \geq 0} \; L(\boldsymbol{\beta}),$$

where

$$L(\boldsymbol{\beta}) = \min_{\mathbf{f} \in F^r} L(\mathbf{f}, \boldsymbol{\beta}).$$

For a link capacity side constrained equilibrium model, Larsson and Patriksson [587] investigate and evaluate an augmented Lagrangean dualization technique for finding optimal values of $\boldsymbol{\beta}$ and show that it is actually more efficient than traditional Lagrange

dualization; moreover, for certain instances of augmented Lagrangean schemes, the dual sequence generated can be shown to converge (at least linearly) although the set of dual solutions is not a singleton in general. (See Section 4.6.1 for further details.)

To summarize, if explicit side constraints are utilized in a refinement of the basic assignment model, the solution of the resulting model [TAP-SC] automatically produces the travel cost mapping of an equivalent standard traffic equilibrium model. Hence, through a process in which one or more models [TAP-SC] are solved, one may derive (i.e., determine $\mathbf{g}$) and calibrate (i.e., find proper coefficients $\boldsymbol{\beta}$) adjusted travel cost functions for use in existing transportation analysis tools based on traditional equilibrium models, in order to (indirectly through the cost functions) take into account the additional model components which are described by the side constraints. The solution of an augmented Lagrangean dual problem may then be viewed as a means for calibrating these travel cost functions.

Next, we discuss the only well-studied case of side constrained assignment problems.

Capacitated traffic assignment

There have been many suggestions for choices of classes of functions to be used to model the congestion effects, i.e., to describe the dependencies between traffic flows and travel times (see Table 1.1). In practice, the most frequently used functions are polynomials whose degrees and coefficients are determined from real-world data through statistical methods. These travel time functions are however unrealistic in the sense that the resulting travel times will become finite whenever the link flows are finite; this means that the links are actually assumed to be able to carry arbitrarily large volumes of traffic flows although the links will in practice undoubtedly have some finite limits on traffic flows, because, for instance, of congestion, speed limits, or cycle-times for traffic-signals.

An obvious and very simple way of improving the quality of a traffic assignment model would thus be to include link flow capacities. This can be done either explicitly through the introduction of explicit upper bounds on the link flows, or implicitly through the use of travel time functions which tend to infinity when the link flows approach their respective capacities ([211, 212]), but neither of these two techniques have been studied to any greater extent.

From a modelling point of view, explicit upper bounds have the advantage of allowing link flows to attain the capacity values, whereas the use of travel time functions with an asymptote at the capacity level will force all link flows to be strictly less than the capacities. Moreover, the result of using travel time functions with asymptotes is, according to the empirical findings of Boyce *et al.* [109], that the estimates of equilibrium travel times become unrealistically high. One disadvantage of imposing explicit capacities is that they destroy the profitable Cartesian product structure which is inherent in the uncapacitated problem. (The loss of the product structure in the feasible set is perhaps the main reason why traffic assignment problems with explicit capacities have been so little studied.)

In addition, a solution to an explicitly capacitated traffic assignment problem will, in the user equilibrium case, no longer comply with Wardrop's first principle ([473]), which has over the years been established and accepted as the fundamental behavioural assumption; however, the solution will satisfy Wardrop's first principle if the usual travel costs are replaced by certain well-defined generalized travel costs [cf. (2.64)].

The capacitated traffic assignment problem ([TAP-C]) is a special case of the general side constrained assignment problem [TAP-SC], with side constraints defined by

$$g_a(\mathbf{f}) \stackrel{\text{def}}{=} f_a - u_a, \qquad \forall a \in \mathcal{A},$$

where $\mathbf{u} \in [0, +\infty]^{|\mathcal{A}|}$ is the vector of upper bounds on the link flows.

The optimality conditions of [TAP-C] yield, from (2.63)–(2.64), that the utilized routes in an O-D pair (p, q) have equal generalized costs

$$\overline{c}_{pqr} \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \delta_{pqra}(t_a(f_a) + \beta_a).$$

It is also possible to derive an equilibrium condition in terms of *actual* costs. Assume, without any loss of generality, that, in O-D pair (p, q), routes are numbered in the order of increasing costs, that the first l routes are utilized, and among these the first m are *saturated*, i.e., contain at least one link which carries flow on its capacity level. Then, the network flow is a capacitated user equilibrium if and only if it is true that

$$c_{pq1} \leq \ldots \leq c_{pqm} \leq c_{pq,m+1} = \ldots = c_{pql},$$

and the unused routes in the O-D pair have generalized route costs that are at least as large as those of the used routes.

Such a characterization of the solution of [TAP-C] was first made by Jorgenson [533], in the special case of constant travel times. (In this case, [TAP-C] reduces to a linear multicommodity flow problem, in which commodities are defined by the O-D pairs.) For flow-dependent travel costs, similar characterizations are given in [473, 484, 516].

Here, it is important to note that one can, in general, not relate the actual travel costs of the unused routes to those of the used ones; it may for instance happen that the cheapest route in an origin-destination pair is not used because its generalized cost is too high. Furthermore, one can for the capacitated problem not formulate a simple optimality condition, similar to Wardrop's first principle, in terms of actual travel costs only. (The extensions of Wardrop's first principle stated by Anantharamaiah [20] and Stefek [872] are incorrect or, possibly, poorly formulated.) This is due to the fact that the Wardrop principles are intimately associated with the Cartesian product structure of the feasible set of [TAP].

The Lagrange multipliers for the capacity constraints "measure the time gained by users of routes filled to capacity compared to the fastest route still available" ([533]), but can also be given other interesting interpretations. First of all, in a network where oversaturated links have queues at their exits, they may be interpreted as the equilibrium time delays caused by the queueing ([678, 750, 895, 850]); this result is extended to the case of non-constant link travel times in [678, 516]. Secondly, they can be seen as the link tolls that drivers on saturated routes are willing to pay for letting them continue to use routes that are faster than the non-saturated ones ([50]). [One may note that although the equilibrium generalized route travel costs are unique, this is not necessarily true for the multipliers $\boldsymbol{\beta}$.]

It is important to note that the capacitated equilibrium link flow pattern found by solving [TAP-C] may also be found by solving the corresponding uncapacitated problem [TAP] with travel time functions adjusted to $t_a(f_a) + \beta_a$ for all $a \in \mathcal{A}$; this is a special case of the adjusted travel time Formula (2.67). Solving a capacitated problem can therefore be used as a tool for guiding the traffic engineer how to correct the travel time functions in order to bring the flow pattern into agreement with the anticipated results ([473]). As compared to heuristic adjustments of the travel time functions, the described strategy has the advantage that it is certainly easier for the engineer to give reasonable estimates of link capacities than to estimate how an adjustment of the travel time functions will affect the uncapacitated equilibrium flow pattern. Lagrangean (and augmented Lagrangean) dual methods for [TAP-C] can actually be interpreted as an automatized process of adjusting

the travel time functions towards the correct ones, which are reached in the limit; see Section 4.6.1.

An interesting subject for future research is to develop and formalize this technique into a means for constructing travel time functions which take link interactions into account. Such a procedure would involve formulating and solving a traffic assignment problem [TAP-SC] which includes a set of side constrains describing the link interactions, and then to utilize the Lagrange multipliers for the side constraints to derive adjusted travel time functions which directly reflect the link interactions. This way of deriving complex travel time functions may be preferable to a calibration of parameters in the travel cost functions, since it may be comparably easy to identify an appropriate set of side constraints and estimate proper values of their coefficients, since these may have very tangible physical interpretations.

Conclusions

The lack of practical applications of asymmetric models of traffic equilibria may, at least partially, be explained by the following. Following the hypothesis that the additional flow relationships modelled through the introduction of asymmetric travel cost functions are actually better represented by a set of side constraints, we observe that:

(1) The interactions and restrictions on the traffic flows captured through the introduction of nonseparable and asymmetric travel cost functions are not described in terms of the physical relationships that they actually represent; in a side constrained assignment model, these physical relationships are modelled explicitly, and they should therefore be easier to derive and calibrate.

(2) The asymmetric model is equivalent to a symmetric one with a travel cost function of the form (2.67), whose parameters, i.e., the multipliers $\boldsymbol{\beta}$, are unknown.

The alternative strategy of extending the basic model with side constraints gives a large flexibility in the construction of the model, since the side constraints can be nonlinear as well as nonseparable. Moreover, it provides a means for the construction of proper adjustments of tentative travel cost functions.

The many possibilities for realistically modelling traffic interactions with explicit side constraints, and the strong relationships to equilibrium models with asymmetric or properly adjusted symmetric travel costs, motivate the further exploration of this modelling strategy for traffic equilibrium problems. The successful outcome of this exploration relies on cooperation between operations researchers and users of today's transportation planning systems.

Applications of network flow models with side constraints also arise in network flows with conversions, losses and gains, and shared resources (e.g., [188, 960, 991, 454, 124, 534, 10, 238, 153, 98, 173, 426, 14, 394, 1007]); usual terms to describe these models are *embedded* and *generalized* networks.

Chapter 3

General traffic equilibrium models

3.1 Introduction

3.1.1 Alternative definitions of equilibria

In the case of separable costs, various definitions of equilibria are known to be equivalent. For more general cost structures, solutions to these equilibrium conditions may, however, differ.

Let r and s denote two arbitrary, but different routes in O-D pair (p,q), where $h_{pqr} > 0$. Further, let $\mathbf{D}_{rs}$ be a vector of the same dimension as $\mathbf{h}$, which is zero in every position except those corresponding to the routes r and s, where the elements of the vector are -1 and 1, respectively.

Definition 3.1 (Alternative equilibrium definitions)

(a) (*Wardrop equilibrium* [958]) The vector $\mathbf{h}$ is a Wardrop equilibrium if and only if

$$c_{pqr}(\mathbf{h}) \leq c_{pqs}(\mathbf{h}).$$

The flow is a Wardrop equilibrium if for each driver, the present cost on any alternative route is at least as great as the cost on his/her present route.

(b) (*User-optimized* [209]) The vector $\mathbf{h}$ is user-optimized if and only if[1]

$$c_{pqr}(\mathbf{h}) \leq c_{pqs}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}), \qquad \forall \varepsilon \in [0, h_{pqr}].$$

The flow is user-optimized if any driver who changes to an alternative route will experience a cost that is at least as great as the old one on his/her old route.

(c) (*Equilibrated* [494]) The vector $\mathbf{h}$ is equilibrated if and only if

$$c_{pqr}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}) \leq c_{pqs}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}), \qquad \forall \varepsilon \in [0, h_{pqr}].$$

The flow is equilibrated if any driver who changes to an alternative route will experience a cost that is at least as great as the new one on his/her old route.

---

[1] Dafermos [205] subsequently generalizes the concept of user-optimized flows, to state that $\mathbf{h}$ is a user-optimized flow if and only if no "sufficiently small" portion of the users of any utilized route can reduce their travel costs by simultaneously changing to any other route, i.e., that there is an $\alpha > 0$ such that

$$c_{pqr}(\mathbf{h}) \leq c_{pqs}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}), \qquad \forall \varepsilon \in [0, \min\{\alpha, h_{pqr}\}].$$

Note that this definition reduces to the one of Dafermos and Sparrow [209] if $\alpha = h_{pqr}$.

Smith [849] demonstrates that Wardrop equilibria may not be user-equilibrated for general cost functions, although the two concepts are equivalent in the separable case. Heydecker [494] shows that, under a monotonicity assumption on the route costs $\mathbf{c}$, Wardrop equilibria coincide with equilibrated flows. The concept of the Wardrop equilibrium, which is the classical one, is hence also the most general of those given above, and will therefore be considered in the following.

The difference between user-optimized and equilibrated flows is characterized by a difference in the assumption on drivers' cost perception ([494]):

> The adoption of the equilibrated condition ... is equivalent to supposing that each driver is instantaneously aware of the cost of travel on all of the routes which he could use. If a driver does try an alternative route, then he will decide whether or not to habituate it by comparing the cost he experiences on the new route with the new cost experienced by other travellers who remain on his old route. By contrast, the adoption of the user-optimized condition is equivalent to supposing that each driver, in attempting to minimize his travel costs, acts only on the basis of his own experience. If he tries an alternative route, he will habituate it if the cost he experiences is less than the cost he experienced on his old route.

The properties of equilibria have been studied through reformulations of the Wardrop conditions as variational inequality problems (see Section 3.2.1), nonlinear complementarity problems (see Section 3.2.2), and fixed point problems (see Section 3.2.3). Below, we state the general problems, and provide a list of their most important properties. These results are subsequently specialized to traffic equilibrium problems.

3.1.2 Variational inequality problems

Let $X \subseteq \Re^n$ be a nonempty, closed and convex set, and $F : X \mapsto \Re^n$ a continuous mapping on X. (These properties are assumed to hold throughout this section.) The *Variational Inequality Problem* then is to find an $\mathbf{x}^* \in X$ such that[2]

[VIP]

$$F(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) \geq 0, \qquad \forall \mathbf{x} \in X. \tag{3.1}$$

(This problem is also known under the names *generalized equations* and as *stationary point* problems.) We let Ω denote the set of solutions to [VIP]. Conditions under which this set is guaranteed to be nonempty are given below. (For statements of the properties of F mentioned, consult Appendix A.)

Theorem 3.1 (Existence of solutions to [VIP]) *A solution to* [VIP] *exists under either one of the following additional properties of F or X.*[3]

(a) *The set X is bounded ([466, 119]).*[4]

(b) *The mapping F is coercive ([466, 686]).*

(c) *The mapping F is strongly monotone ([868]).*

[2]The problem [VIP] may be interpreted as the problem of finding a point $\mathbf{x}^* \in X$ at which the vector field F is an inward normal to X.

[3]If a certain Slater constraint qualification holds, the conditions on the mapping F may be replaced by only pseudomonotonicity ([464]).

[4]The boundedness condition can always be relaxed if the feasible set can be restricted to a bounded set without affecting the optimal solution set.

The solution set is convex if F is pseudomonotone, and also bounded under a Slater constraint qualification or coercivity of F ([464]).

The following theorem gives a sufficient condition for a solution to [VIP] to be unique.

Theorem 3.2 [868] (Uniqueness of solutions to [VIP]) *The solution set of* [VIP], *if nonempty, is a singleton if F is strictly monotone on X.*

The relationships between [VIP] and optimization problems are next studied. The following result is a well-known optimality condition for the mathematical program

[P]

$$\min_{\mathbf{x}\in X} T(\mathbf{x}). \tag{3.2a}$$

Theorem 3.3 [43] (Optimality condition) *Let $T : X \mapsto \Re$ be in C^1 on X. If $\mathbf{x}^* \in X$ is an optimal solution to* [P], *then $\mathbf{x}^*$ is a solution to* [VIP], *with $F \equiv \nabla T$. The converse holds whenever T is pseudoconvex.*

The results of the theorem imply that whenever F is the gradient of a real-valued function T, a solution to [VIP] may be found through the solution of the optimization problem [P].

Conditions for F to be a gradient mapping may be found in [778, 925, 144, 464].

Theorem 3.4 [727, Th. 4.1.6] (Sufficient condition for F to be a gradient) *Let $F : X \mapsto \Re^n$ be in C^1 on an open convex set $X_0 \subset X$. Then F is a gradient mapping on X_0 if and only if $\nabla F(\mathbf{x})$ is symmetric for all $\mathbf{x} \in X_0$.*

Under this symmetry condition, the line integral

$$T(\mathbf{x}) \stackrel{\text{def}}{=} \int_{\mathbf{0}}^{\mathbf{x}} F(\mathbf{s})^{\mathrm{T}} d\mathbf{s} \tag{3.2b}$$

is path independent according to Green's theorem, and F is integrable. The problem [VIP] can hence be put as an equivalent mathematical program, with an objective of the form (3.2b).[5]

Properties of T relative to F are given below.

Theorem 3.5 (Relationships between monotonicity properties of T and F) *Let $F \equiv \nabla T$. Then,*

(a) *F is monotone on $X \iff T$ is convex on X.*

(b) *F is strictly monotone on $X \iff T$ is strictly convex on X.*

(c) *F is strongly monotone on $X \iff T$ is strongly convex on X.*

The properties of T can be further studied through Definition A.2.

[5] An alternative formulation is

$$T(\mathbf{x}) \stackrel{\text{def}}{=} \int_0^1 \sum_{j=1}^{n} F_j(\mathbf{x}^0 + s(\mathbf{x} - \mathbf{x}^0))(x_j - x_j^0) ds,$$

where $\mathbf{x}^0$ is an arbitrary point in X.

3.1.3 Nonlinear complementarity problems

Let $F : \Re^n_+ \mapsto \Re^n$ be continuous. The *Nonlinear Complementarity Problem* is to find an $\mathbf{x}^* \in \Re^n$ such that

[NCP]

$$F(\mathbf{x}^*)^{\mathrm{T}}\,\mathbf{x}^* = 0, \tag{3.3a}$$
$$F(\mathbf{x}^*) \geq \mathbf{0}, \tag{3.3b}$$
$$\mathbf{x}^* \geq \mathbf{0}. \tag{3.3c}$$

The problem [NCP] is equivalent to a variational inequality defined on $X = \Re^n_+$ ([544, 551]). Existence and uniqueness results for [NCP] therefore follow from those for [VIP] (Theorems 3.1 and 3.2); see also [541, 542, 543, 544, 565, 464, 184].

3.1.4 Fixed point problems

Let $F : X \mapsto \Re^n$ be continuous. The *Fixed Point Problem* is to find an $\mathbf{x}^* \in X$ such that

[FPP]

$$F(\mathbf{x}^*) = \mathbf{x}^*. \tag{3.4}$$

In the field of traffic equilibrium, fixed point problems have mostly been applied as an instrument of establishing the existence of solutions to variational inequality or complementarity models. The general proof is based on the definition of an appropriate continuous mapping, which transforms the original model into an equivalent fixed point problem, for which existence is then established by imposing strong enough properties onto the original problem data. There are several classical existence results for fixed point problems (e.g., [122, 539, 277, 546]).

Theorem 3.6 [539] (Existence of a fixed point) *Let X be bounded, and F be a mapping from X to X. Then there exists a solution to* [FPP].

Theorem 3.7 [727, Th. 5.1.3] (Existence of a unique fixed point) *Let F be contractive on X. Then there exists a unique solution to* [FPP]. *Furthermore, the sequence $\{\mathbf{x}^k\}$, defined by $\mathbf{x}^0 \in X$,*

$$\mathbf{x}^{k+1} = F(\mathbf{x}^k), \qquad k = 0, 1, \ldots, \tag{3.5}$$

converges to the unique fixed point.

Theorem 3.8 [289] (A fixed point characterization of Ω) *Let $\mathbf{B}$ be a symmetric and positive matrix in $\Re^{n\times n}$. Then,*

$$\mathbf{x}^* \in \Omega \Longleftrightarrow \mathbf{x}^* = P_X^{\mathbf{B}}\left(\mathbf{x}^* - \mathbf{B}^{-1}F(\mathbf{x}^*)\right), \tag{3.6a}$$

where

$$P_X^{\mathbf{B}}(\mathbf{x}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{y}\in X} \|\mathbf{x}-\mathbf{y}\|_{\mathbf{B}} \tag{3.6b}$$

is the projection of $\mathbf{x}$ onto X with respect to the norm

$$\|\mathbf{x}\|_{\mathbf{B}} \stackrel{\text{def}}{=} (\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x})^{1/2}. \tag{3.6c}$$

Applying Theorem 3.8, with $\mathbf{B}$ equal to the identity matrix $\mathbf{I}$, to [NCP], yields the equivalent fixed point problem of finding an $\mathbf{x}^* \geq \mathbf{0}$ such that ([901])

$$H(\mathbf{x}) = \mathbf{x}, \tag{3.7a}$$

where

$$H_j(\mathbf{x}) = \max\{0, x_j - F_j(\mathbf{x})\}, \qquad \forall j \in \{1, \ldots, n\}. \tag{3.7b}$$

Theorems 3.7 and 3.8 provide the basis for many algorithmic procedures proposed for the solution of [VIP] and [NCP], through the solution of an equivalent fixed point problem, in particular for the class of projection algorithms; see Section 5.3.

Another fixed point characterization of solutions to [VIP] is given next.

Theorem 3.9 (A fixed point characterization of Ω)

$$\mathbf{x}^* \in \Omega \Longleftrightarrow \mathbf{x}^* \in H(\mathbf{x}^*), \tag{3.8a}$$

where

$$H(\mathbf{x}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{y} \in X} F(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}). \tag{3.8b}$$

Theorem 3.9 was first stated by Zuhovickiĭ *et al.* [1018] (see also [246, 307, 641]). The optimization problem defining H is equivalent to that solved as a search direction finding subproblem in the Frank–Wolfe algorithm. The mapping H is not contractive, and can therefore not be used in conjunction with the result of Theorem 3.7 in the construction of a convergent algorithm.[6] The result of Theorem 3.9 is, however, utilized in methods based on the minimization of the primal gap function and has a nice interpretation in terms of the equilibrium conditions (see Sections 3.1.5 and 5.3).

In the fixed point problem defined in (3.8a), H is a point-to-set mapping. The existence of a solution to (3.8a) is ensured by a generalization of Theorem 3.6 from continuous point-to-point mappings to upper semicontinuous point-to-set mappings.

Both Theorems 3.8 and 3.9 are special cases of a general fixed point result for variational inequality problems, which is one property of a class of optimization reformulations of [VIP]. This is the topic of the next section.

3.1.5 Mathematical programming reformulations

Whenever F is not the gradient of any function, the integral in (3.2b) is not unambiguously defined, and [VIP] can in this case not be converted into an equivalent optimization problem of the form (3.2). (In analogy with the symmetry characterization of Theorem 3.4, we then say that the problem [VIP] is *asymmetric*.)

The objective functions (or merit functions) of the mathematical programming reformulations of [VIP] studied in this section are of the type given by the following definition.

Definition 3.2 (Gap function) *Let Ω be the set of solutions to* [VIP]. *A function $\psi : X \mapsto \Re \cup \{-\infty, +\infty\}$ is a gap function for* [VIP] *if*

(1) *ψ is restricted in sign on X, and*

(2) $\psi(\mathbf{x}) = 0 \Longleftrightarrow \mathbf{x} \in \Omega$.

A gap function provides a measure of the violation of [VIP] at any point $\mathbf{x} \in X$, and by minimizing ψ over X (assuming that ψ is nonnegative on X), a point in Ω is obtained. It may therefore be used as a merit function for variational inequalities.

[6] Utilizing the mapping H in a method of the form (3.5) would, in the context of traffic assignment, correspond to the iterated all-or-nothing heuristic; see Section 1.5.4.

The primal and dual gap functions

Variational inequalities are highly related to equilibrium problems in non-cooperative game theory (e.g., [615, 84]).[7] The saddle point problem characterizing the solutions to the game problem is to find $(\mathbf{x}^*, \mathbf{y}^*) \in X \times Y$ such that ([779])

$$L(\mathbf{x}^*, \mathbf{y}) \leq L(\mathbf{x}^*, \mathbf{y}^*) \leq L(\mathbf{x}, \mathbf{y}^*), \qquad \forall (\mathbf{x}, \mathbf{y}) \in X \times Y. \tag{3.9}$$

This problem is said to be *convex-concave* if $L : X \times Y \mapsto \Re$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$. The Problem (3.9) may be reformulated as the minimax problem ([779, Le. 36.2]; see also [222, 243])

$$\sup_{\mathbf{y} \in Y} \inf_{\mathbf{x} \in X} L(\mathbf{x}, \mathbf{y}) = L(\mathbf{x}^*, \mathbf{y}^*) = \inf_{\mathbf{x} \in X} \sup_{\mathbf{y} \in Y} L(\mathbf{x}, \mathbf{y}). \tag{3.10}$$

In the study of N-person non-cooperative games, Zuhovickiĭ *et al.* [1017, 1018, 1019, 1020] observe that under a monotonicity assumption on F, $\mathbf{x}^* \in X$ is an equilibrium point of the Game (2.45) if and only if $(\mathbf{x}^*, \mathbf{x}^*)$ is a saddle point of $L(\mathbf{x}, \mathbf{y}) = F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y})$, i.e.,

$$\max_{\mathbf{y} \in X} \min_{\mathbf{x} \in X} F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) = 0 = \min_{\mathbf{x} \in X} \max_{\mathbf{y} \in X} F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}). \tag{3.11}$$

Exploring the rightmost equality of (3.11) we obtain the *primal gap function* $G : X \mapsto \Re_+ \cup \{+\infty\}$ defined by

$$G(\mathbf{x}) = \sup_{\mathbf{y} \in X} F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}). \tag{3.12}$$

The function G has been studied extensively in various contexts; below, we summarize its most important properties.

Theorem 3.10 (Properties of G) *For any $\mathbf{x} \in X$, let $Y(\mathbf{x})$ denote the (possibly empty) set of optimal solutions to the problem defined in (3.12).*

(a) *G is a gap function.*

(b) *G is l.s.c. on X.*

(c) *If X is bounded and $F \in C^1$ on X, then G is Lipschitz continuous on X.*

(d) *If $F \in C^1$ on X, then G is differentiable at $\mathbf{x} \in X$ if $Y(\mathbf{x}) = \{\mathbf{y}(\mathbf{x})\}$, with*

$$\nabla G(\mathbf{x}) = F(\mathbf{x}) + \nabla F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}(\mathbf{x})).$$

(e) *If $F \in C^1$ on X, and F is monotone on X, then if $\mathbf{x} \notin \Omega$ and $Y(\mathbf{x}) = \{\mathbf{y}(\mathbf{x})\}$, $\mathbf{p} = \mathbf{y}(\mathbf{x}) - \mathbf{x}$ defines a feasible direction of descent with respect to G at $\mathbf{x}$, and the directional derivative satisfies*

$$G'(\mathbf{x}; \mathbf{p}) = \nabla G(\mathbf{x})^{\mathrm{T}} \mathbf{p} \leq -G(\mathbf{x}).$$

(f) *G is convex on X if $\mathbf{x} \mapsto F(\mathbf{x})^{\mathrm{T}}\mathbf{x}$ is convex on X and each component of F is concave on X.*

[7] Consider the non-cooperative N-person game (2.45). If each penalty function φ_i is differentiable, then by defining $F(\mathbf{x})^{\mathrm{T}} = (\nabla_1 \varphi_1(\mathbf{x})^{\mathrm{T}}, \ldots, \nabla_N \varphi_N(\mathbf{x})^{\mathrm{T}})$, we obtain that $\mathbf{x}^* \in X$ is an equilibrium point of the game if and only if

$$F_i(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_i^*) \geq 0, \qquad \forall \mathbf{x}_i \in X_i,\ i \in \{1, \ldots, N\},$$

i.e., if and only if $\mathbf{x}^* \in X$ solves [VIP] defined on $\prod_{i=1}^N X_i$.

(g) (A fixed point characterization of Ω)

$$\mathbf{x} \in \Omega \Longleftrightarrow \mathbf{x} \in Y(\mathbf{x}).$$

(h) (A stationary point characterization of Ω) *Under the conditions on F in* **(e)**,

$$\mathbf{x} \in \Omega \Longleftrightarrow G'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \geq 0, \qquad \forall \mathbf{y} \in X.$$

Proof

(a) See [1017, 1018, 1019, 1020]. See also [246, Sec. 3.1] and [472, 474].
(b) See [280].
(c) See [641].
(d) See [244]. See also [1020, 34, 472, 474, 478].
(e) See [244]. See also [474] for the first result, and [641].
(f) See [472, 474, 478].
(g) See [1018]. See also [246, 307, 641]. (This is Theorem 3.9.)
(h) See [641]. □

Thus, if [VIP] has a solution, the set of solutions to

[P_G]

$$\inf_{\mathbf{x} \in X} \; G(\mathbf{x}) \tag{3.13}$$

is the set Ω.

When X is polyhedral, (3.12) reduces to a linear program, which, in the symmetric case, is equivalent to that solved in the Frank–Wolfe algorithm.

In the context of traffic equilibrium, the primal gap function G provides a measure of the violation of the Wardrop user equilibrium conditions; indeed, letting $\mathbf{f} \in F^r$,

$$G(\mathbf{f}) = \max_{\mathbf{y} \in F^r} \; \mathbf{t}(\mathbf{f})^{\mathrm{T}}(\mathbf{f} - \mathbf{y}) \tag{3.14a}$$

$$= \mathbf{t}(\mathbf{f})^{\mathrm{T}}\mathbf{f} - \min_{\mathbf{y} \in F^r} \; \mathbf{t}(\mathbf{f})^{\mathrm{T}}\mathbf{y}, \tag{3.14b}$$

i.e., $G(\mathbf{f})$ is the difference in total travel costs between that of the flow $\mathbf{f}$ and that of the corresponding shortest route solution. A positive value of the gap function hence corresponds to a situation in which there is a potential benefit for some travellers in adjusting their route choices, and the value is zero exactly when no traveller has an incentive to change route, i.e., when the flow satisfies the Wardrop conditions of user equilibrium. This measure of the violation of the Wardrop conditions is a standard output of many traffic assignment packages ([472]). [For instance, the lower bound on the optimal value provided by the Frank–Wolfe subproblem in the separable model is related to this measure by $G(\mathbf{f}) = T(\mathbf{f}) - \underline{T}(\mathbf{y}(\mathbf{f}))$, where $\mathbf{y}(\mathbf{f})$ is any shortest route pattern given the travel costs $\mathbf{t}(\mathbf{f})$; cf. (4.8).]

Murchland [697] was the first to study the primal gap Function (3.12) in the context of traffic equilibria; he derived it as the gap between the value of the objective of [TAP] and the value of a conjugate dual formulation of [TAP]. (The interpretation of the general merit Function (3.18) as the gap in Fenchel's [323] inequality is made in [33, 744, 747].) Hearn [472, 474] suggests using this measure as a merit function in the solution of the separable traffic equilibrium model.

The primal gap function G has been used as a merit function in several algorithms for [VIP] ([1019, 1020, 244, 766, 594, 641, 644, 642, 645, 647, 286, 646]), some of which are applied to traffic equilibrium problems; see Sections 5.2.5, 5.3.4, and 5.3.5. (For the special case of [NCP], it is also studied in [182, 543, 185].)

Turning to the leftmost equality of (3.11) we obtain the *dual gap function* $g : X \mapsto \Re_- \cup \{-\infty\}$ defined by

$$g(\mathbf{y}) = \inf_{\mathbf{x}\in X} F(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}). \tag{3.15}$$

Theorem 3.11 (Properties of g) *Let $F : X \mapsto \Re^n$ be pseudomonotone on X. For any $\mathbf{y} \in X$, let $X(\mathbf{y})$ denote the (possibly empty) set of optimal solutions to the problem defined in (3.15).*

(a) *g is a gap function.*

(b) *g is concave on X.*

(c) *If $F \in C^1$ on X, then g is differentiable at $\mathbf{y} \in X$ if $X(\mathbf{y}) = \{\mathbf{x}(\mathbf{y})\}$, with*

$$\nabla g(\mathbf{y}) = -\nabla F(\mathbf{x}(\mathbf{y})).$$

Proof

(a) See [1017, 1018]. See also [244] and [34, Sec. VII.5].

(b) See [483, 478, 719].

(c) See [483, 478]. □

As a consequence of the gap function property of g, the set of solutions to

[P_g]

$$\sup_{\mathbf{y}\in X} g(\mathbf{y}) \tag{3.16}$$

is the set Ω. See Sections 5.2.5 and 5.3.6 for methods applied to [P_g].

Summarizing the properties of the saddle function L, although it is not convex in $\mathbf{x}$ we have, for pseudomonotone F,

$$g(\mathbf{y}) \le 0 \le G(\mathbf{x}), \qquad \forall (\mathbf{x},\mathbf{y}) \in X \times X,$$
$$\sup_{\mathbf{y}\in X} g(\mathbf{y}) = 0 = \inf_{\mathbf{x}\in X} G(\mathbf{x}),$$

and $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point to L if and only if $\mathbf{x}^*, \mathbf{y}^* \in \Omega$.

Finally, one may note the nice symmetry in the properties of G and g. While the evaluation of $G(\mathbf{x})$ is a convex problem, the minimization of G is in general both a nonconvex and nondifferentiable problem; on the other hand, while the evaluation of $g(\mathbf{y})$ is in general a nonconvex problem, the maximization of g is always a convex problem. In addition, convexity of G holds for affine and monotone maps F, and under the same conditions the evaluation of $g(\mathbf{y})$ is a convex problem.

Smith's class of gap functions

Smith [845, 846, 847] develops a family of gap-type merit functions for variational inequalities on bounded polyhedral sets. Letting

$$[x]_+ \stackrel{\text{def}}{=} \max\{0, x\}, \qquad \forall x \in \Re,$$

the family of merit functions is defined by

$$G^p(\mathbf{x}) = \sum_{j \in \mathcal{X}} \left([F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}^j)]_+ \right)^p, \qquad \forall \mathbf{x} \in X, \tag{3.17}$$

where $p > 0$. Smith shows that for all $p \geq 1$, G^p is a gap function for [VIP]. Hearn *et al.* [478] relate this gap function to the primal gap function G; since

$$\lim_{p \to +\infty} (G^p(\mathbf{x}))^{1/p} = \max_{j \in \mathcal{X}} F(\mathbf{x})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}^j) = G(\mathbf{x}), \qquad \forall \mathbf{x} \in X,$$

the primal gap function is a limiting case of Smith's class of gap functions.

They proceed to show that G^p is differentiable on X for $2 \leq p < +\infty$, and convex for $p \geq 2$ and $p = +\infty$ for instance if F is affine and monotone. Note that the extreme points $\mathbf{y}^j$ of X must be known explicitly in order to evaluate $G^p(\mathbf{x})$, for all finite values of p. The practical use of these gap functions are therefore limited to polyhedral sets, in combination with column generation techniques ([845, 846]).

A class of merit functions for [VIP]

Auchmuty [33] (see also [744, 746, 747, 588, 748, 749]) introduces a class of merit functions for [VIP], which includes the primal gap function, and some other known merit functions, as special cases. Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X, and define

$$\psi(\mathbf{x}) \stackrel{\text{def}}{=} \sup_{\mathbf{y} \in X} L(\mathbf{y}, \mathbf{x}), \tag{3.18a}$$

where

$$L(\mathbf{y}, \mathbf{x}) \stackrel{\text{def}}{=} \varphi(\mathbf{x}) - \varphi(\mathbf{y}) + [F(\mathbf{x}) - \nabla\varphi(\mathbf{x})]^{\mathrm{T}}(\mathbf{x} - \mathbf{y}). \tag{3.18b}$$

Theorem 3.12 [33, 746, 747, 588] (ψ is a gap function) *Let $\varphi : X \mapsto \Re$ be a convex function in C^1 on X. Then $\psi : X \mapsto \Re_+ \cup \{+\infty\}$ is a gap function for* [VIP].

Proof For any $\mathbf{x} \in X$, $\psi(\mathbf{x}) \geq 0$ holds, since $L(\mathbf{x}, \mathbf{x}) = 0$. Next, assume that $\mathbf{x}^* \in \Omega$. Then

$$\begin{aligned} L(\mathbf{y}, \mathbf{x}^*) &\leq F(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x}^* - \mathbf{y}) \quad (\varphi \text{ convex}) \\ &\leq 0, \ \forall \mathbf{y} \in X. \qquad (\mathbf{x}^* \in \Omega) \end{aligned}$$

Since $L(\mathbf{x}^*, \mathbf{x}^*) = 0$ holds, we have that $\psi(\mathbf{x}^*) = 0$. Conversely, let $\psi(\mathbf{x}^*) = 0$, for some $\mathbf{x}^* \in X$. Since, from (3.18a), $L(\mathbf{y}, \mathbf{x}^*) \leq 0$ must hold, for all $\mathbf{y} \in X$, $\mathbf{x}^*$ must be a solution to the problem defining $\psi(\mathbf{x}^*)$. Replacing $\mathbf{y}(\mathbf{x}^*)$ with $\mathbf{x}^*$ in the necessary optimality conditions for (3.18a),

$$[\nabla\varphi(\mathbf{y}(\mathbf{x}^*)) + F(\mathbf{x}^*) - \nabla\varphi(\mathbf{x}^*)]^{\mathrm{T}}(\mathbf{y} - \mathbf{y}(\mathbf{x}^*)) \geq 0, \qquad \forall \mathbf{y} \in X,$$

then yields that $\mathbf{x}^*$ solves [VIP]. □

The inner problem of (3.18) in the definition of ψ is defined as follows. Consider replacing the mapping F in [VIP] by the symmetric and monotone mapping $\nabla\varphi$. The error thus introduced, $\nabla\varphi - F$, is, at the point $\mathbf{x} \in X$, approximated by the fixed cost

term $\nabla\varphi(\mathbf{x})-F(\mathbf{x})$, which is added to $\nabla\varphi$. The approximate variational inequality (which we denote by $[\mathrm{VIP}_{\nabla\varphi}]$) then has the mapping

$$\nabla\varphi + F(\mathbf{x}) - \nabla\varphi(\mathbf{x}),$$

which is integrable, and $[\mathrm{VIP}_{\nabla\varphi}]$ is equivalent to solving the convex program of minimizing

$$\varphi(\mathbf{y}) + [F(\mathbf{x}) - \nabla\varphi(\mathbf{x})]^{\mathrm{T}}\,(\mathbf{y}-\mathbf{x})$$

over $\mathbf{y} \in X$. This problem is obviously the same as the inner problem of (3.18).

Remark 3.1 The better $\nabla\varphi$ approximates F, the closer will the solution, $\mathbf{y}(\mathbf{x})$, to $[\mathrm{VIP}_{\nabla\varphi}]$ be to Ω. If F is symmetric and monotone, then $\nabla\varphi = F$ is a possible choice; in this case, the approximate problem $[\mathrm{VIP}_{\nabla\varphi}]$ is equivalent to [VIP]. If the mapping $\nabla\varphi$ is replaced by a more general, possibly asymmetric, mapping Φ, then $[\mathrm{VIP}_{\Phi}]$ may yield even better approximations to Ω. In this case, though, $[\mathrm{VIP}_{\Phi}]$ is not equivalent to a convex optimization problem, or to the calculation of a merit function ψ.

The approximate problem $[\mathrm{VIP}_{\Phi}]$ may also be constructed as follows. Rewrite the cost mapping F equivalently as

$$\Phi + [F - \Phi]. \tag{3.19}$$

The subproblem $[\mathrm{VIP}_{\Phi}]$ is then, at $\mathbf{x} \in X$, constructed by fixing the term within the brackets to its value at $\mathbf{x}$. The process leading to $[\mathrm{VIP}_{\Phi}]$ is referred to as a *cost approximation* ([747]); see Section 5.2.1 for a detailed description of this algorithm concept.

The merit function ψ corresponding to $\varphi \equiv 0$ is the primal gap function discussed above. As shown in [588], the generalization of the dual formulation (3.15) does not define a gap function for [VIP] in general.

Theorem 3.13 [747, 588] (Properties of ψ) *For any $\mathbf{x} \in X$, let $Y(\mathbf{x})$ denote the (possibly empty) set of optimal solutions to the problem defined in (3.18).*

(a) [33] *ψ is l.s.c. on X.*

(b) *If X is bounded or if φ is strongly convex on X, then ψ is continuous on X.*

(c) *If X is bounded, $F \in C^1$ on X and $\varphi \in C^2$ on X, then ψ is Lipschitz continuous on X.*

(d) *Let $F \in C^1$ on X and $\varphi \in C^2$ on X. If X is bounded and φ strictly convex on X, or if φ is strongly convex on X, then $\psi \in C^1$ on X, with*

$$\nabla\psi(\mathbf{x}) = F(\mathbf{x}) + [\nabla F(\mathbf{x})^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x})](\mathbf{x} - \mathbf{y}(\mathbf{x})).$$

(e) *Let $F \in C^1$ on X and $\varphi \in C^2$ on X. If X is bounded, $\mathbf{x} \notin \Omega$, and $\nabla F(\mathbf{x})^{\mathrm{T}} - \nabla^2\varphi(\mathbf{x})$ is positive definite, then there is a $\mathbf{y} \in Y(\mathbf{x})$ such that $\mathbf{p} = \mathbf{y} - \mathbf{x}$ defines a feasible direction of descent with respect to ψ at $\mathbf{x}$. If furthermore F is strongly monotone on X, φ strongly convex on X, and $\nabla\varphi$ Lipschitz continuous on X, then for any $\mathbf{x} \in X$,*

$$\nabla\psi(\mathbf{x})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbf{x}) \leq -(m_F + m_\varphi - M_{\nabla\varphi})\|\mathbf{y}(\mathbf{x}) - \mathbf{x}\|^2.$$

(f) *ψ is convex on X if F is affine, with $F(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, φ is quadratic, with $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, and $\mathbf{A} + \mathbf{A}^{\mathrm{T}} - \mathbf{B}$ is positive semidefinite.*

(g) (A fixed point characterization of Ω)

$$\mathbf{x} \in \Omega \Longleftrightarrow \mathbf{x} \in Y(\mathbf{x}).$$

(h) (A stationary point characterization of Ω) *Under the conditions on F and φ in* **(e)**,

$$\mathbf{x} \in \Omega \Longleftrightarrow \psi'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \geq 0, \qquad \forall \mathbf{y} \in X.$$

The fixed point result **(g)** generalizes Theorems 3.8 and 3.9, for the choices $\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$ and $\varphi(\mathbf{x}) \equiv 0$, respectively, where $\mathbf{B} \in \Re^{n\times n}$ is a symmetric and positive definite matrix.

The function φ may, instead of being chosen as a fixed function, be chosen dependently (and adaptively) of the point $\mathbf{x}$ at which the approximation is made. From the viewpoint of making good symmetric approximations of [VIP], and the construction of efficient numerical methods, this may indeed be of great advantage. For a sequence $\{\mathbf{x}^k\}$ of points, we may therefore introduce a sequence $\{\varphi^k\}$ of convex functions $\varphi^k : X \mapsto \Re$ in C^1 on X (or, in more generality, a sequence $\{\Phi^k\}$ of monotone cost approximating mappings).

The corresponding sequence $\{\psi^k\}$ of gap functions include, as special cases, the primal gap function ($\varphi^k \equiv 0$), the gap functions of Fukushima [391, 885] ($\varphi^k(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, $\mathbf{B}$ symmetric and positive definite), the gap functions of Wu *et al.* [994] ($\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}(\mathbf{y})\mathbf{x}$), the gap function of Auchmuty [33, 747, 588] ($\varphi^k(\mathbf{x}) = \varphi(\mathbf{x})$ for all k), and the gap functions of Zhu and Marcotte [1013, 1014] ($\varphi^k(\mathbf{x}) = \gamma_k\varphi(\mathbf{x}, \mathbf{x}^k)$, $\gamma_k > 0$). See Larsson and Patriksson [746, 747, 588, 748, 749] for further details on the relationships between these merit functions, and algorithms based on their minimization.

In Section 5.2.1, we introduce a general iterative scheme for [VIP], in which the solution to a subproblem of the form [VIP$_{\Phi^k}$] is utilized either as a search direction with respect to a merit function ψ^k or as the definition of a successive approximation scheme. A large number of existing algorithms for asymmetric traffic equilibria may be put into this framework.

3.2 Traffic equilibrium models

In this section we shall derive equivalent reformulations of the Wardrop conditions as variational inequality, nonlinear complementarity, and fixed point problems, which will enable us to establish existence and uniqueness results for traffic equilibria.

3.2.1 Variational inequality models

The fixed demand case

Assumption 3.A (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand d_{pq} is nonnegative for each $(p, q) \in \mathcal{C}$.*

(3) *The route cost $c_{pqr} : \Re_+^{|\mathcal{R}|} \mapsto \Re_{++}$ is positive and continuous for each $r \in \mathcal{R}_{pq}$ and $(p, q) \in \mathcal{C}$.*

(4) *If the route costs are additive, i.e., if (2.5f) holds, then the travel time function $t_a : \Re_+^{|\mathcal{A}|} \mapsto \Re_{++}$ is positive and continuous for each $a \in \mathcal{A}$.*

Theorem 3.14 [630] (Variational inequality formulations of the fixed demand Wardrop conditions) *Let Assumption 3.A hold.*

(a) *The fixed demand Wardrop Conditions (2.2) are equivalent to the variational inequality problem of finding an* $\mathbf{h}^* \in H$ *such that*

[TAP-VIP-H]

$$\mathbf{c}\,(\mathbf{h}^*)^{\mathrm{T}}\,(\mathbf{h}-\mathbf{h}^*) \geq 0, \qquad \forall \mathbf{h} \in H. \tag{3.20}$$

(b) *Assume that the route costs are additive. Then the fixed demand Wardrop Conditions (2.2) are equivalent to the variational inequality problem of finding an* $\mathbf{f}^* \in F^r$ *such that*

[TAP-VIP-F^r]

$$\mathbf{t}\,(\mathbf{f}^*)^{\mathrm{T}}\,(\mathbf{f}-\mathbf{f}^*) \geq 0, \qquad \forall \mathbf{f} \in F^r. \tag{3.21}$$

Proof

(a) The flow $\mathbf{h}^*$ solves [TAP-VIP-H] if and only if it solves the linear program

$$\min_{\mathbf{h}\in H} \mathbf{c}\,(\mathbf{h}^*)^{\mathrm{T}}\,\mathbf{h}. \tag{3.22}$$

The primal-dual optimality conditions of (3.22) are (2.2a)–(2.2d). The positivity assumption on c_{pqr} implies that $\pi_{pq} \geq 0$.

(b) The theorem is proved by establishing the equivalence of [TAP-VIP-H] and [TAP-VIP-F^r] under additivity. Let $\mathbf{h}^* \in H$ solve [TAP-VIP-H]. The flow $\mathbf{h}^*$ corresponds to a (unique) link flow solution, $\mathbf{f}^* \in F^r$, through (2.6d). By virtue of the additivity assumption, for any pair $(\mathbf{h}, \mathbf{f})$ satisfying (2.6b)–(2.6d),

$$\mathbf{c}\,(\mathbf{h}^*)^{\mathrm{T}}\,(\mathbf{h}-\mathbf{h}^*) = \mathbf{t}\,(\mathbf{f}^*)^{\mathrm{T}}\,(\mathbf{f}-\mathbf{f}^*)\,, \tag{3.23}$$

and therefore (3.20) implies (3.21). Thus, $\mathbf{f}^*$ solves [TAP-VIP-F^r]. Conversely, let $\mathbf{f}^* \in F^r$ solve [TAP-VIP-F^r], and construct a route flow solution, $\mathbf{h}^* \in H$, satisfying (2.6b)–(2.6d) together with $\mathbf{f}^*$. As above, we may conclude that, for any pair $(\mathbf{h}, \mathbf{f})$ satisfying (2.6b)–(2.6d), (3.23) holds, and therefore (3.21) implies (3.20). Thus, $\mathbf{h}^*$ solves [TAP-VIP-H]. □

The formulations [TAP-VIP-H] and [TAP-VIP-F^r] are due to Smith [840], although Dickson [256] is, perhaps, the first to state a variational inequality formulation corresponding to the Wardrop conditions; his formulation, [TAP-VIP-F^r], is, however, based on separable travel costs. (In this case, [TAP-VIP-F^r] represents the optimality conditions of the link-route version of [TAP].)

The elastic demand case

Assumption 3.B (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand function* $g_{pq} : \Re_+^{|\mathcal{C}|} \mapsto \Re_+$ *is nonnegative and continuous for each* $(p,q) \in \mathcal{C}$.

(3) *The route cost* $c_{pqr} : \Re_+^{|\mathcal{R}|} \mapsto \Re_{++}$ *is positive and continuous for each* $r \in \mathcal{R}_{pq}$ *and* $(p,q) \in \mathcal{C}$.

(4) *If the route costs are additive, i.e., if (2.5f) holds, then the travel time function* $t_a : \Re_+^{|\mathcal{A}|} \mapsto \Re_{++}$ *is positive and continuous for each* $a \in \mathcal{A}$.

Introduce the *route-O-D pair incidence matrix*, $\mathbf{\Gamma}^{\mathrm{T}} = (\gamma_{pqr})$, defined by

$$\gamma_{pqr} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if route } r \text{ joins O-D pair } (p,q), \\ 0, & \text{otherwise,} \end{cases} \qquad \forall r \in \mathcal{R},\ \forall (p,q) \in \mathcal{C}. \tag{3.24}$$

Then the Wardrop Conditions (2.4) may be compactly written as

$$\mathbf{h}^{\mathrm{T}}(\mathbf{c} - \mathbf{\Gamma}\boldsymbol{\pi}) = 0, \tag{3.25a}$$
$$\mathbf{c} - \mathbf{\Gamma}\boldsymbol{\pi} \geq \mathbf{0}, \tag{3.25b}$$
$$\mathbf{\Gamma}^{\mathrm{T}}\mathbf{h} - \mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}, \tag{3.25c}$$
$$\mathbf{h} \geq \mathbf{0}, \tag{3.25d}$$
$$\boldsymbol{\pi} \geq \mathbf{0}. \tag{3.25e}$$

Theorem 3.15 (Variational inequality formulations of the elastic demand Wardrop conditions) *Let Assumption 3.B hold.*

(a) *The elastic demand Wardrop Conditions (2.4) are equivalent to the variational inequality problem of finding* $(\mathbf{h}^*, \boldsymbol{\pi}^*) \in \Re_+^{|\mathcal{R}|+|\mathcal{C}|}$ *such that*

[TAP-E-VIP]

$$\begin{bmatrix} \mathbf{c}(\mathbf{h}^*) - \mathbf{\Gamma}\boldsymbol{\pi}^* \\ \mathbf{\Gamma}^{\mathrm{T}}\mathbf{h}^* - \mathbf{g}(\boldsymbol{\pi}) \end{bmatrix}^{\mathrm{T}} \left(\begin{pmatrix} \mathbf{h} \\ \boldsymbol{\pi} \end{pmatrix} - \begin{pmatrix} \mathbf{h}^* \\ \boldsymbol{\pi}^* \end{pmatrix} \right) \geq 0, \qquad \forall (\mathbf{h}, \boldsymbol{\pi}) \in \Re_+^{|\mathcal{R}|+|\mathcal{C}|}. \tag{3.26}$$

(b) *Assume that* $\mathbf{g}$ *is invertible. Then the elastic demand Wardrop Conditions (2.4) are equivalent to the variational inequality problem of finding* $(\mathbf{h}^*, \mathbf{d}^*) \in H_d$ *such that*

[TAP-E-VIP-H_d]

$$\mathbf{c}(\mathbf{h}^*)^{\mathrm{T}}(\mathbf{h} - \mathbf{h}^*) - \mathbf{g}^{-1}(\mathbf{d}^*)^{\mathrm{T}}(\mathbf{d} - \mathbf{d}^*) \geq 0, \qquad \forall (\mathbf{h}, \mathbf{d}) \in H_d. \tag{3.27}$$

Proof

(a) See, e.g., [343, 630].

(b) See, e.g., [193, 630]. □

The formulation [TAP-E-VIP] is due to Fisk and Boyce [343]. Florian [349] and Dafermos [193] present link flow based versions of [TAP-E-VIP-H_d] (which we denote by [TAP-E-VIP-F_d^r], cf. [TAP-VIP-F^r]).

Note that in the fixed demand case ($\mathbf{g} \equiv \mathbf{d}$), [TAP-E-VIP-$H_d$] reduces to [TAP-VIP-$H$] (and, similarly, [TAP-E-VIP-F_d^r] reduces to [TAP-VIP-F^r]).

3.2.2 Nonlinear complementarity models

Consider the formulation [TAP-E-VIP] of the elastic demand Wardrop Conditions (2.4). This formulation is equivalent to a nonlinear complementarity problem (since the feasible set is the nonnegative orthant, see Section 3.1.3), with $\mathbf{x} = (\mathbf{h}, \boldsymbol{\pi})$ and

$$F(\mathbf{x}) = \begin{pmatrix} \mathbf{c}(\mathbf{h}) - \mathbf{\Gamma}\boldsymbol{\pi} \\ \mathbf{\Gamma}^{\mathrm{T}}\mathbf{h} - \mathbf{g}(\boldsymbol{\pi}) \end{pmatrix}.$$

We will refer to this formulation as [TAP-E-NCP].

Theorem 3.16 (Nonlinear complementarity formulation of the elastic demand Wardrop conditions) *Let Assumption 3.B hold. The elastic demand Wardrop Conditions (2.4) are equivalent to the nonlinear complementarity problem* [TAP-E-NCP].

Proof Follows immediately from the equivalence of [TAP-E-NCP] and [TAP-E-VIP], and Theorem 3.15.a. □

The model [TAP-E-NCP] is given by Aashtiani and Magnanti [1, 2, 4].

3.2.3 Fixed point models

Fixed point formulations of the equilibrium conditions arise in two different ways. A range of fixed point formulations may be obtained from a reformulation of a variational inequality or nonlinear complementarity formulation, through the general fixed point Theorem 3.13.g. Such fixed point formulations are primarily used for establishing quantitative properties of the underlying traffic model, or the convergence properties of an iterative algorithm for its solution.

Sender and Netter [823] formulate the first known fixed point formulations of the Wardrop conditions, both for elastic and fixed demands, and for multi-modal networks. The formulations may, in fact, be derived from the variational inequality formulation [TAP-E-VIP-H_d] and the fixed point Theorem 3.8 using $\mathbf{B} = \mathbf{I}$.

In the elastic demand case, it is also possible to derive a fixed point model based directly on the demand-travel cost relationship. For a given demand $\mathbf{d} \geq \mathbf{0}$, let $\boldsymbol{\pi}(\mathbf{d})$ denote the vector of minimum travel costs (assumed unique) obtained when assigning the demand onto the network according to the principle of user equilibrium. Introducing the demand function $\mathbf{g}$, these travel times yield a demand $\mathbf{g}(\boldsymbol{\pi}(\mathbf{d}))$ [which may differ from $\mathbf{d}$]. Then, the equilibrium conditions may be written as $\mathbf{d}^*$ solving the fixed point problem

[TAP-E-FPP]

$$\mathbf{g}\left(\boldsymbol{\pi}\left(\mathbf{d}^*\right)\right) = \mathbf{d}^*. \tag{3.28}$$

The formulation [TAP-E-FPP] is due to Fisk and Nguyen [338]; it is further studied in [339, 337] for multi-modal and multi-class user transportation networks.

3.3 Properties of equilibrium solutions

3.3.1 Existence of equilibrium solutions

Based on the reformulations of the Wardrop conditions presented above, several results for the existence of equilibria have been established. Below, we shall give the most general of these.

Assumption 3.C (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand function* $g_{pq} : \Re_+^{|\mathcal{C}|} \mapsto \Re_+$ *is nonnegative, upper bounded and continuous for each* $(p,q) \in \mathcal{C}$.

(3) *The route cost* $c_{pqr} : \Re_+^{|\mathcal{R}|} \mapsto \Re_{++}$ *is positive and continuous for each* $r \in \mathcal{R}_{pq}$ *and* $(p,q) \in \mathcal{C}$.

(4) *If the route costs are additive, i.e., if (2.5f) holds, then the travel time function* $t_a : \Re_+^{|\mathcal{A}|} \mapsto \Re_{++}$ *is positive and continuous for each* $a \in \mathcal{A}$.

Theorem 3.17 (Existence of equilibrium solutions) *Let Assumption 3.C hold.*

(a) *There exists an equilibrium solution* $(\mathbf{d}^*, \mathbf{h}^*)$ *to the elastic demand Wardrop Conditions (2.4).*

(b) *Further, assume that each demand function* g_{pq} *is fixed. Then there exists an equilibrium solution* $\mathbf{h}^*$ *to the fixed demand Wardrop Conditions (2.2).*

Proof

(a) See Aashtiani and Magnanti [4].

(b) See Aashtiani and Magnanti [4]. For the case of additivity, see Smith [840]. □

Remark 3.2 Smith [840, 846] establishes existence results for the elastic demand case (both for additive and non-additive travel costs), in which the boundedness assumption on the demand functions is replaced by a certain boundary condition.

Remark 3.3 The results in [4] are based on the formulation [TAP-E-NCP], while that in [840] utilizes [TAP-VIP-F^r]. In both cases, reformulations into equivalent fixed point problems are used, in combination with fixed point theorems, such as Theorem 3.6. Braess and Koch [113] consider [TAP-VIP-F^r], and establish existence under an additional assumption of monotonicity of $\mathbf{t}$. Harker [462] utilizes [TAP-E-NCP-F_d^r], and an existence result for [NCP] due to Smith [856], to prove a result similar to Theorem 3.17.a for an additive model. Florian [349] utilizes [TAP-E-VIP-F_d^r] to establish the existence of an equilibrium under monotonicity assumptions on $\mathbf{t}$ and $-\mathbf{g}$ ($-\mathbf{g}$ is assumed strictly monotone). The assumptions used by Sender and Netter [823] are similar to those in [349]. Further existence results are found in [1, 2, 339, 310, 364], some of which are established for multiclass-user equilibria. Equilibria on multiclass-user networks may be studied on single-mode networks, where each mode defines its own network copy ([206, 94, 95, 1, 2]). A summary of some of the above mentioned results are found in [326, 462].

Remark 3.4 In the modelling of additive traffic equilibrium problems, a node-link formulation of the feasible set of link flows may also be used. In such a formulation, the set F^r is replaced by the set F^n, defined by the flow conservation Constraints (2.13). The set F^n is not bounded, due to the presence of cycles (see Section 2.2.2); however, under the assumption that travel costs are positive everywhere, no traveller can reduce his/her travel cost by travelling in a cycle, and the existence of an equilibrium solution using a link-node formulation may hence be established by using the same arguments as for the link-route formulation used in this section.

With the above assumptions, there may be more than one equilibrium solution, and the travel costs of different flows i H^*, as well as the demands, may be different. (This follows from the non-monotonicity of the travel costs and demands.) To alleviate this unwanted property, we will in the next section impose further conditions on the travel time and demand functions to ensure that the equilibrium travel times and demands are unique. We will also establish the uniqueness of the total link flows.

Discontinuous cost functions and equilibria

The existence results of Theorem 3.17 depend heavily on the continuity of the travel cost and demand functions.

Asmuth [30, 31] assumes that $\mathbf{t} : \Re_+^{|\mathcal{A}|} \mapsto 2^{\Re_{++}^{|\mathcal{A}|}}$ and $\mathbf{g} : \Re_+^{|\mathcal{C}|} \mapsto 2^{\Re_+^{|\mathcal{C}|}}$ are closed (see Definition A.7.a) and convex valued point-to-set mappings, i.e., that given a flow $\mathbf{f}$, the travel cost is an element of the closed and convex set $\mathbf{t}(\mathbf{f})$.[8]

With the additional assumption that $\mathbf{g}$ is upper bounded, Asmuth establishes the existence of an equilibrium solution (cf. Theorem 3.17.a). Similar results are obtained by Fang and Peterson [308, 311, 312, 313, 314].

Motivated by the modelling of congestion pricing schemes using discontinuous step functions, Bernstein and Smith [64] extend the notion of user equilibrium to incorporate discontinuous cost functions.

Let each cost function $c_{pqr} : \Re_+^{|\mathcal{R}|} \mapsto \Re_{++}$ be positive, bounded and lower semicontinuous (see Definition A.6.a). As in Section 3.1.1, we let r and s denote two arbitrary, but different, routes in O-D pair (p, q), where $h_{pqr} > 0$. Further, let $\mathbf{D}_{rs}$ be a vector of the same dimension as $\mathbf{h}$, which is zero in every position except those corresponding to the routes r and s, where the elements of the vector are -1 and 1, respectively.

Definition 3.3 (Discontinuous user equilibrium) *The vector* $\mathbf{h}$ *is a discontinuous user equilibrium if and only if*

$$c_{pqr}(\mathbf{h}) \leq \liminf_{\varepsilon \downarrow 0} c_{pqs}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}).$$

It is clear that if each cost function c_{pqr} is continuous, then discontinuous user equilibria coincide with Wardrop equilibria ([64, Th. 2.1.iii]).

To ensure the existence of a solution to the discontinuous user equilibrium conditions, Bernstein and Smith introduce a notion of *regularity* of the cost functions.

Definition 3.4 (Regular cost) *A lower semicontinuous cost structure is regular if and only if, for every* $\mathbf{h} \in H$ *and* $r, s \in \mathcal{R}_{pq}$, *with* $r \neq s$ *and* $h_{pqs} > 0$, $(p, q) \in \mathcal{C}$,

$$\liminf_{\varepsilon \downarrow 0} c_{pqr}(\mathbf{h} + \varepsilon \mathbf{D}_{rs}) = \sum_{a \in r \cap s} t_a(\mathbf{f}) + \sum_{a \in r - s} \tilde{t}_a(\mathbf{f}),$$

where $\tilde{t}_a(\mathbf{f}) \stackrel{\text{def}}{=} \lim_{\varepsilon \downarrow 0} \sup \{t_a(\mathbf{x}) \mid \mathbf{x} \in F^r, \|\mathbf{x} - \mathbf{f}\| < \varepsilon\}$ *is the upper hull of* t_a.

Observe that additive and continuous costs c_{pqr} are always regular. The regularity condition states that a flow shift can only create cost discontinuities on those links where link flows actually change.

We now have the following result, which extends that of Theorem 3.17.b to discontinuous regular costs.

Theorem 3.18 [64] (Existence of discontinuous equilibrium solutions) *Let the network be strongly connected, the demand* $\mathbf{d}$ *be nonnegative, and each cost function* $c_{pqr} : \Re_+^{|\mathcal{R}|} \mapsto \Re_{++}$ *be positive, bounded, additive, lower semicontinuous and regular. Then there exists a discontinuous user equilibrium solution.*

[8] The motivation behind the use of point-to-set demand functions is their possible derivation from utility maximization behaviour; no clear motivation exists for the choice of the travel costs as point-to-set mappings.

3.3.2 Uniqueness of equilibrium solutions

To establish the uniqueness of the demands, travel costs and link flows, the below assumptions will be used.

Assumption 3.D (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand function* $g_{pq} : \Re_+^{|\mathcal{C}|} \mapsto \Re_+$ *is nonnegative, upper bounded and continuous for each* $(p,q) \in \mathcal{C}$. *Further,* $-\mathbf{g}$ *is monotone on* $\Re_+^{|\mathcal{C}|}$.

(3) *The travel time function* $t_a : \Re_+^{|\mathcal{A}|} \mapsto \Re_{++}$ *is positive and continuous for each* $a \in \mathcal{A}$. *Further,* $\mathbf{t}$ *is monotone on* $\Re_+^{|\mathcal{A}|}$.

Theorem 3.19 (Uniqueness of equilibria) *Let Assumption 3.D hold.*

(a) *Let either* $-\mathbf{g}$ *or* $\mathbf{t}$ *be strictly monotone. Then the equilibrium travel times are unique.*

(b) *Let* $-\mathbf{g}$ *be strictly monotone. Then the equilibrium demands are unique.*

(c) *Let* $\mathbf{g}$ *be positive on* $\Re_+^{|\mathcal{C}|}$, *and* $\mathbf{t}$ *strictly monotone. Then the equilibrium travel times and link flows are unique.*

Proof

(a) See Aashtiani and Magnanti [4].

(b) See Florian [349].

(c) See Aashtiani and Magnanti [4]. □

Remark 3.5 Asmuth [30, 31] establishes the uniqueness of the demands, travel times and link flows under the assumption that $-\mathbf{g}$ and $\mathbf{t}$ are strictly monotone set-valued mappings. Smith [840] establishes the uniqueness of the link flows under a strict monotonicity assumption on $\mathbf{t}$ in the fixed demand case. Dafermos [193] assumes strong monotonicity. (This is one example of the use of overly restrictive assumptions in the analysis of a traffic model; the main reason for this is that the assumptions are simultaneously used to ensure the convergence of an iterative algorithm for its solution.) The uniqueness results stated are based on a uniqueness theorem of the form of Theorem 3.2.

Remark 3.6 Observe that the uniqueness of the equilibrium travel costs may be established under milder monotonicity assumptions on $\mathbf{t}$ in the case of separable costs; see Theorem 2.5.a.

3.3.3 Further properties of equilibrium solutions

We here extend the results of Section 2.3.3 to asymmetric models.

Consider the elastic demand problem of finding $(\mathbf{f}^*, \mathbf{d}^*) \in F_d^r$ such that

[TAP-E-VIP-F_d^r]

$$\mathbf{t}\,(\mathbf{f}^*)^{\mathrm{T}}\,(\mathbf{f}-\mathbf{f}^*) - \mathbf{g}^{-1}\,(\mathbf{d}^*)^{\mathrm{T}}\,(\mathbf{d}-\mathbf{d}^*) \geq 0, \qquad \forall (\mathbf{f},\mathbf{d}) \in F_d^r, \tag{3.29}$$

where $F_d^r \stackrel{\text{def}}{=} \{ (\mathbf{f}, \mathbf{d}) \in \Re^{|\mathcal{A}|+|\mathcal{C}|} \mid (\mathbf{f}, \mathbf{d})$ satisfies (2.27b)–(2.27d) $\}$.

We introduce multipliers $\boldsymbol{\mu} \in \Re^{|\mathcal{A}|}$ and $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ for the link flow definitional Constraints (2.27d) and the demand feasibility Constraints (2.27b), respectively, and define the dual variational inequality ([778, 688, 784, 62, 755, 736]) formulation of the traffic equilibrium problem, where $(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) \in \Pi_\mu$ is sought such that

[DTAP-E-VIP-F_d^r]

$$\mathbf{f}\,(\boldsymbol{\mu}^*)^{\mathrm{T}}\,(\boldsymbol{\mu} - \boldsymbol{\mu}^*) - \mathbf{g}\,(\boldsymbol{\pi}^*)^{\mathrm{T}}\,(\boldsymbol{\pi} - \boldsymbol{\pi}^*) \geq 0, \qquad \forall (\boldsymbol{\mu}, \boldsymbol{\pi}) \in \Pi_\mu, \tag{3.30}$$

where $\Pi_\mu \stackrel{\text{def}}{=} \{ (\boldsymbol{\mu}, \boldsymbol{\pi}) \in \Re^{|\mathcal{A}|+|\mathcal{C}|} \mid \sum_{a \in \mathcal{A}} \delta_{pqra} \mu_a \geq \pi_{pq},\ \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C} \}$.

In contrast to the primal problem [TAP-E-VIP-F_d^r], equilibrium travel costs and demands are here sought directly. The dual formulation [DTAP-E-VIP-F_d^r] is well defined if $\mathbf{t}$ and $-\mathbf{g}^{-1}$ are strictly monotone and coercive; these conditions ensures that

$$\mathbf{t}(\mathbf{f}) = \boldsymbol{\mu} \iff \mathbf{f} = \mathbf{f}(\boldsymbol{\mu}) = \mathbf{t}^{-1}(\boldsymbol{\mu}), \tag{3.31a}$$

$$\mathbf{g}^{-1}(\mathbf{d}) = \boldsymbol{\pi} \iff \mathbf{d} = \mathbf{d}(\boldsymbol{\pi}) = \mathbf{g}(\boldsymbol{\pi}) \tag{3.31b}$$

for all values of $(\boldsymbol{\mu}, \boldsymbol{\pi})$. (See, e.g., [1006] for results on the existence and boundedness of inverses of monotone operators.)

Fukushima and Itoh [393] establish some relationships between the solutions to [DTAP-E-VIP-F_d^r] and [TAP-E-VIP-F_d^r] under strong monotonicity assumptions on $\mathbf{t}$ and $-\mathbf{g}^{-1}$; in particular, they prove a generalization of the first part of Theorem 2.7, i.e., that the equilibrium link flows and demands are obtained through (3.31) from the solution $(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*)$ to [DTAP-E-VIP-F_d^r].

The explicit statement of [DTAP-E-VIP-F_d^r] is intractable due to the large number of constraints defining Π_μ. One may, however, without any loss of generality, for any solution $\boldsymbol{\mu}$ let the multipliers $\boldsymbol{\pi}$ be defined by

$$\pi_{pq} = \pi_{pq}(\boldsymbol{\mu}) = \min_{r \in \mathcal{R}_{pq}} \left\{ \sum_{a \in \mathcal{A}} \delta_{pqra} \mu_a \right\}$$

(cf. Section 2.3.3), i.e., as the shortest route costs given the fixed link costs $\boldsymbol{\mu}$; this choice of $\boldsymbol{\pi}$ results in [DTAP-E-VIP-F_d^r] being reduced to a problem only in the travel cost variables. (This problem may be viewed as the fixed point problem of finding a travel cost $\boldsymbol{\mu}^*$ such that $\mathbf{f}(\boldsymbol{\mu}^*)$ is among the possible link flow allocations of the demand $\mathbf{d}(\boldsymbol{\pi}(\boldsymbol{\mu}^*))$. This problem is a generalization of the dual problem [DTAP-E], where the corresponding formulation is that of finding a zero subgradient of the dual function θ [cf. (2.37) and (2.39)]. Compare with the fixed point model [TAP-E-FPP], which is a primal model based on the demand variables.)

3.3.4 Stability and sensitivity of equilibrium solutions

Nonparametric sensitivity analysis

First motivated by Braess' paradox (see Section 2.3.4), the stability and sensitivity of traffic equilibria have been studied with regards to changes in the network topology, demands and costs.

Most of the results presented may be derived from the following general results.

Theorem 3.20 [201, 202] (Sensitivity and stability of solutions to [VIP])

(a) (Sensitivity) *Let* $\mathbf{x}^*$ *be a solution to* [VIP]. *Let* $\widetilde{F}$ *be a perturbed function, and* $\widetilde{\mathbf{x}}^*$ *a solution to the corresponding variational inequality. Then*

$$\left[\widetilde{F}(\widetilde{\mathbf{x}}^*) - F(\mathbf{x}^*)\right]^{\mathrm{T}} (\widetilde{\mathbf{x}}^* - \mathbf{x}^*) \leq 0. \tag{3.32}$$

(b) (Sensitivity) *Let* F *be monotone on* X. *Then*

$$\left[\widetilde{F}(\widetilde{\mathbf{x}}^*) - F(\widetilde{\mathbf{x}}^*)\right]^{\mathrm{T}} (\widetilde{\mathbf{x}}^* - \mathbf{x}^*) \leq 0. \tag{3.33}$$

(c) (Stability) *Let* F *be strongly monotone on* X. *Then*

$$\|\widetilde{\mathbf{x}}^* - \mathbf{x}^*\| \leq \frac{1}{m_F} \|\widetilde{F}(\widetilde{\mathbf{x}}^*) - F(\widetilde{\mathbf{x}}^*)\|. \tag{3.34}$$

In Braess' [112] example, the addition of a route resulted in the increase in travel costs for every user of the O-D pair joined by the route.

When applied to the model [TAP-VIP-H] or [TAP-VIP-F^r], Theorems 3.20.a and 3.20.b state that, although the travel cost may increase for some travellers, improving the travel cost functions results in decreases in the incurred travel costs on average. Under strong assumptions on the travel costs, Braess' paradox may be guaranteed not to occur.

Theorem 3.21 [201] (Braess' paradox does not occur) *Assume that* $\mathbf{t} : \Re_+^{|\mathcal{A}|} \mapsto \Re_{++}^{|\mathcal{A}|}$ *is strongly monotone and in* C^1 *on* $\Re_+^{|\mathcal{A}|}$. *Let* $\mathbf{f}^*$ *solve* [TAP-VIP-F^r]. *Let link* $a \in \mathcal{A}$ *be improved while the rest remain unchanged, i.e.,*

$$\tilde{t}_a(\mathbf{f}) < t_a(\mathbf{f}), \qquad \forall \mathbf{f} \in F^r,\ \tilde{f}_b = f_b,\ \forall b \neq a, \tag{3.35}$$

and that

$$\frac{\partial t_b(\mathbf{f})}{\partial f_a} = 0, \qquad \forall \mathbf{f} \in F^r,\ \forall b \neq a.$$

Let $\tilde{\mathbf{f}}$ *solve the corresponding problem* [TAP-VIP-F^r]. *Then,*

$$\tilde{f}_a^* \geq f_a^*, \qquad \tilde{t}_a(\tilde{\mathbf{f}}^*) \leq t_a(\mathbf{f}^*).$$

Note that the Requirement (3.35) is fulfilled in the separable case, i.e., in the model [TAP]; the result thus extends that of Hall [451].

Theorem 3.20.c states that a small change in the function F results in a small change in the incurred solution. Applied to the model [TAP-VIP-F^r], it yields a continuity result for the equilibrium link flows as functions of the travel cost functions.

Theorem 3.20 may also be applied to the elastic demand models, e.g., to [TAP-E-VIP-F_d^r], in which case simultaneous changes in the travel demand and cost functions are made. Dafermos and Nagurney [203] state Theorem 3.20 for [TAP-E-VIP], and extend Theorem 3.21 to show that an improvement in the demand function for an O-D pair yields a decrease in the equilibrium travel cost and an increase in the demand; these results extend those in [451, 309]. Similar results are obtained for the fixed demand model [TAP-VIP-F^r] under perturbations of the demand vector, and for spatial economic equilibrium problems ([201, 200, 202]).

Steinberg and Zangwill [874] and Dafermos and Nagurney [200] derive formulas for the cost change that is induced by a change in the demand, or the addition of a route, in the model [TAP-VIP-F^r]. (Sheer *et al.* [830] show how to calculate them using graph theory techniques.) Dafermos and Nagurney also establish conditions under which Braess' paradox will not occur when a route is added to the network (see also [376, 232, 233]).

Remark 3.7 The stability and sensitivity results for changes in the network topology, and cost and demand functions, have not been described within a general framework, although they are very similar. The similarity should not be surprising, though, for two main reasons. Firstly, the addition of a route (or link) to an existing network may be thought of as an improvement to an existing route (or link) by reducing its cost from a level at which it is not used by any traveller; consequently, network improvements may be studied as special cases of travel cost improvements. Secondly, perturbations in demands and travel costs are essentially equivalent. Indeed, consider the perturbation of the (fixed) demand d_k to αd_k, where $\alpha > 0$. By redefining the route and link flows as $(1/\alpha)h_{kr}$ and $(1/\alpha)f_{ak}$, respectively, it is straightforward to show that the demand perturbation is equivalent to perturbing the travel cost function instead to

$$\tilde{\mathbf{t}}(\mathbf{f}) = \mathbf{t}\left(\alpha \mathbf{f}_k + \sum_{i \neq k} \mathbf{f}_i\right),$$

i.e., a scaling by α of the contribution of the flow in commodity k to the travel cost. A cost improvement is hence equivalent to a demand decrease.

Parametric sensitivity analysis

In parametric sensitivity analyses of traffic equilibrium models, a perturbed problem of the below form is studied:

[TAP-VIP-F_ε^r]

$$\mathbf{t}(\mathbf{f}^*, \boldsymbol{\varepsilon})^{\mathrm{T}} (\mathbf{f} - \mathbf{f}^*) \geq 0, \qquad \forall \mathbf{f} \in F_\varepsilon^r, \tag{3.36}$$

where

$$F_\varepsilon^r = \{ \mathbf{f} \in \Re^{|\mathcal{A}|} \mid \mathbf{f} = \boldsymbol{\Delta}\mathbf{h},\ \boldsymbol{\Gamma}\mathbf{h} = \mathbf{d}(\boldsymbol{\varepsilon}),\ \mathbf{h} \geq \mathbf{0} \},$$

and where $\boldsymbol{\varepsilon}$ is a vector of perturbation parameters.

Results studied are the stability of the solution set of the perturbed problem with respect to the perturbation parameters, in terms of (Lipschitz) continuity and directional differentiability, and is highly related to the stability results of solutions to the first-order optimality conditions of nonlinear programs (i.e., the Karush–Kuhn–Tucker conditions; see, e.g., [43, Chap. 4]) under appropriate regularity assumptions. The interest of such results to the field of traffic equilibrium is, for example, the possibility of calculating the sensitivity of the solution to the input data, and the development of solution methods for complex traffic models, in which traffic equilibrium problems arise as subproblems. Examples of the latter are equilibrium network design and O-D matrix estimation models (both of which are continuous bilevel programs, where the traffic equilibrium problem is at the lower level). For theoretical results of parametric sensitivity analysis and applications to traffic equilibrium models, see [777, 328, 899, 576, 198, 900, 769, 384, 273, 382, 702, 770, 881].

Part II

Methods

Chapter 4

Algorithms for the basic model and its extensions

In this chapter we will review and develop convergent algorithms for separable traffic equilibrium models. The algorithms are given a unified presentation based on algorithm concepts such as decomposition, column generation, and partial linearization. This form of presentation serves two purposes: firstly, it facilitates comparisons among methods proposed; secondly, it becomes possible to introduce new algorithms through suitable combinations of the above concepts.

For the reader's convenience, we again state the program [TAP].

[TAP]

$$\min\ T(\mathbf{f}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \int_0^{f_a} t_a(s)ds, \tag{4.1a}$$

subject to

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{4.1b}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{4.1c}$$

$$\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a\in\mathcal{A}. \tag{4.1d}$$

Throughout this chapter, the network is assumed to satisfy the following.

Assumption 4.A (Properties of the traffic network)

(1) *The network is strongly connected.*

(2) *The demand d_{pq} is positive for each $(p,q)\in\mathcal{C}$ (fixed demand case).*

(3) *The demand function $g_{pq}:\Re_+\mapsto\Re_+$ is positive, continuous, upper bounded and strictly decreasing for each $(p,q)\in\mathcal{C}$ (elastic demand case).*

(4) *The travel time function $t_a:\Re_+\mapsto\Re_{++}$ is positive, continuous and strictly increasing for each $a\in\mathcal{A}$.*

The first statement of a convergent algorithm for the solution of the fixed demand traffic assignment model [TAP] is that of Smock [858, 859] (see Section 1.5.4). This algorithm,

which is equivalent to what is known today as the method of successive averages (MSA), may be viewed as a simplified version of the still most popular method for the solution of [TAP]—the Frank–Wolfe [377] method. (The first application of this method to [TAP] is that of Bruynooghe *et al.* [132].)

Previous to the development of algorithms of the above type, however, algorithms based on duality were established for nonlinear single-commodity network flow problems ([160, 45]). These algorithms were available to the community since, a few years earlier, Beckmann *et al.* [47] and others had formulated the mathematical program [TAP] to be solved in order to yield equilibrium flows, and procedures transforming this multicommodity flow problem into a sequence of single-commodity flow problems had been known for a long time. [One such procedure is the cyclic decomposition scheme, see Section 4.2.2. This was not utilized for the solution of [TAP] until in the late 1960s ([210, 423, 424, 209]).]

There may be many explanations for the postponement of the application of efficient methods for traffic assignment problems; the most obvious one is that transportation analysts and practitioners, and the operations researchers at universities in Europe and the United States, were not aware of each others' work.

An efficient algorithm for [TAP], and its extensions, must take the problem structure into account, because of the size of real-world problems. The most important structures inherent in [TAP] are:

(1) (*Network constraints*) The main constraints of traffic assignment problems are network defining ones; this is true also in side constrained models.

(2) (*Independent constraints*) The constraints of [TAP] define a Cartesian product of feasible sets for the different O-D pairs; in side constrained models, this structure is lost, but may be regained by using relaxation strategies for handling the side constraints.

(3) (*Nonlinear and convex objective*) The objective of [TAP] is a nonlinear, differentiable and convex function. It is also separable with respect to the links of the network.

4.1 The Frank–Wolfe algorithm and its extensions

The method of Frank and Wolfe [377] was originally proposed for the solution of convex quadratic programs, but is in fact applicable to any optimization problem with a pseudoconvex and continuously differentiable objective and a nonempty, compact and convex feasible set (e.g., [653, Chap. 14] and [768, Sec. III.3]). When applied to a problem defined on a bounded polyhedral feasible set, the algorithm alternates between the solution of a linear program defined by a tangential approximation of the objective, and a line search, minimizing the original objective over the line segment defined by the current iterate and the solution to the linear program. For convex problems, the linear subproblem defines a lower bound on the optimal value, which may be be used in termination criteria.

As applied to [TAP], the algorithm may be described as follows.

4.1.1 The Frank–Wolfe algorithm

Step 0 (Initialization) Let $\mathbf{f}^0$ be a feasible solution to [TAP], $LBD = 0$, $\varepsilon > 0$, and $k = 0$.

Step 1 (Search direction generation) v Let

$$\underline{T}(\mathbf{f}) \stackrel{\text{def}}{=} T(\mathbf{f}^k) + \nabla T(\mathbf{f}^k)^{\mathrm{T}}(\mathbf{f} - \mathbf{f}^k). \tag{4.2}$$

Solve the linear programming subproblem

$$\min \underline{T}(\mathbf{f}), \tag{4.3a}$$

subject to

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{4.3b}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \tag{4.3c}$$

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a \in \mathcal{A}. \tag{4.3d}$$

Let $\mathbf{y}^k$ be its solution, and $\mathbf{p}^k = \mathbf{y}^k - \mathbf{f}^k$ the resulting search direction.

Step 2 (Convergence check) Let $LBD := \max\{LBD, \underline{T}(\mathbf{y}^k)\}$. If

$$\frac{T(\mathbf{f}^k) - LBD}{LBD} < \varepsilon, \tag{4.4}$$

then terminate, with $\mathbf{f}^k$ as the approximate solution. Otherwise, continue.

Step 3 (Line search) Find a step length, l_k, which solves the one-dimensional problem

$$\min \{\, T(\mathbf{f}^k + l\mathbf{p}^k) \mid 0 \leq l \leq 1 \,\}. \tag{4.5}$$

Step 4 (Update) Let $\mathbf{f}^{k+1} = \mathbf{f}^k + l_k \mathbf{p}^k$.

Step 5 (Convergence check) If

$$\frac{T(\mathbf{f}^{k+1}) - LBD}{LBD} < \varepsilon, \tag{4.6}$$

then terminate, with $\mathbf{f}^{k+1}$ as the approximate solution. Otherwise, let $k := k+1$, and go to Step 1.

The initial solution $\mathbf{f}^0$ may be obtained as an all-or-nothing assignment given free-flow travel times (see Section 1.5.4).

Due to the separability of the constraints of [TAP] and the absence of flow capacity constraints, the linear Subproblem (4.3) separates into $|\mathcal{C}|$ independent problems, in which the shortest route is sought between each origin and destination, based on fixed travel costs $\partial / \partial f_a T(\mathbf{f}^k) = t_a(f_a^k)$. The reader should note that although [TAP] is based on the (intractable) route flow based formulation, it is not necessary to enumerate the routes within the Frank–Wolfe algorithm.

4.1.2 Termination criteria

The program [TAP] is a special case of the program

[P]

$$\min_{\mathbf{x}\in X} T(\mathbf{x}), \tag{4.7}$$

where T is convex and continuously differentiable on the polyhedral set X [which in [TAP] is defined by the network Constraints (4.1b)–(4.1d)]. From the facts that $\mathbf{y}^k$ solves the linearly approximated problem and that T is convex, respectively, it follows that

$$\begin{aligned} \underline{T}(\mathbf{y}^k) &= T(\mathbf{x}^k) + \nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{y}^k - \mathbf{x}^k) && (4.8a)\\ &\leq T(\mathbf{x}^k) + \nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x}^* - \mathbf{x}^k) && (4.8b)\\ &\leq T^*, && (4.8c) \end{aligned}$$

where $\mathbf{x}^*$ is an arbitrary solution of [P], i.e., that $\underline{T}(\mathbf{y}^k)$ is a lower bound on the optimal value of [P].

Moreover, if $\underline{T}(\mathbf{y}^k) = T(\mathbf{x}^k)$ for some k, then $\mathbf{x}^k$ solves [P]; this explains the use of the termination Criteria (4.4) and (4.6).

There is, however, a hesitance among transportation researchers to utilize the (artificial) objective function T in termination criteria due to its lack of interpretation. One alternative is to utilize Wardrop's equilibrium Conditions (2.1) directly, by measuring the difference in the travel costs of the routes used within an O-D pair. The error could then be defined as

$$\max_{(p,q)\in\mathcal{C}} \left\{ \max_{r,s\in\mathcal{R}_{pq}} \left\{ c_{pqr}(\mathbf{h}^k) - c_{pqs}(\mathbf{h}^k) \;\middle|\; h^k_{pqr}, h^k_{pqs} > 0 \right\} \right\}, \tag{4.9}$$

where $\mathbf{h}^k$ is the route flow solution in iteration k. The use of such a termination criterion of course requires the knowledge of which routes are being used in the network, and the Frank–Wolfe algorithm normally does not provide this information. There are, however, situations in which route flow information is desirable when analyzing the equilibrium solution (e.g., [934, 941, 414]); see Section 4.3.5 for a thorough discussion on route-flow based algorithms.

It has been observed (e.g., [586]) that the lower bound provided by the shortest route solutions in the Frank–Wolfe algorithm gives poor estimates of the quality of the current flow solution. This is not surprising, considering the global use of the tangential approximation; in particular, when the network is heavily congested, there is a considerable difference between the equilibrium flow, in which several routes are used, and an all-or-nothing solution, in which only one route is used in every O-D pair.

Hearn [472] suggests improving this bound, by solving the linear program

$$\begin{aligned} \min_{\mathbf{w}\in X} \quad & z, \\ \text{s.t.} \quad & z \geq T(\mathbf{x}^k) + \nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{w} - \mathbf{x}^k), \\ & z \geq T(\mathbf{y}^k) + \nabla T(\mathbf{y}^k)^{\mathrm{T}}(\mathbf{w} - \mathbf{x}^k), \end{aligned} \tag{4.10}$$

obtaining the *minimax* bound, $\underline{z}$. The problem (4.10) corresponds to making a tangential approximation of T also at the optimal solution of the linear subproblem. The unstructured linear Program (4.10) may, however, be too expensive to solve repeatedly.

Various convergence tests, and the effects of the choice of stopping criteria on the error of predicted equilibrium flows are discussed by Rose *et al.* [788].

4.1.3 The use of the Frank–Wolfe approach for the solution of [TAP]

The first known application of the Frank–Wolfe algorithm to [TAP] is that of Bruynooghe *et al.* [132]; another early reference is Murchland [697]. It has since then been rediscovered by several researchers, for instance by LeBlanc [599, 606, 607], Nguyen [713], Steenbrink [871] and Golden [430] in transportation analysis contexts, and by Yaged [997], Gerla [420], Fratta *et al.* [379], and Klessig [557] in the context of computer communication networks (see Section 2.6.5).

Depending on the interpretations of the algorithm, it has also become known under different names:

(1) (*Linearization*) A possible interpretation of the Frank–Wolfe algorithm is that of a simple outer approximation scheme, in which only one supporting hyperplane is used to approximate the epigraph of T.

(2) (*Conditional gradient*) Another possible interpretation of the Frank–Wolfe algorithm is that of a constrained steepest descent algorithm, the reason being that the subproblem corresponds to finding the direction of most negative directional derivative of T among all feasible directions, i.e.,

$$\min \left\{ \nabla T(\mathbf{x}^k)^{\mathrm{T}}\mathbf{p} \;\middle|\; \mathbf{p} \in T_X(\mathbf{x}^k), \mathbf{x}^k + \mathbf{p} \in X \right\}, \tag{4.11}$$

where $T_X(\mathbf{x}^k)$ is the tangent cone of X at $\mathbf{x}^k$ (the cone of feasible directions).[1]

(3) (*Flow deviation*) This name may stem from the interpretation of the Frank–Wolfe method in behavioural terms as an iterative process, in which, in each iteration, some travellers adjust their choice of routes to less congested ones, according to the current traffic conditions.[2]

(4) (*Convex combination*) This name stems from the adjustment process in **Step 3**, in which the next iterate is chosen as the optimal convex combination of the current iterate and the subproblem solution.

The Frank–Wolfe algorithm was made popular by the work of LeBlanc *et al.* [607], Nguyen [717], and the validation studies of Florian and Nguyen [360] and Van Vliet and Dow [941]. It is now the most common approach to solving equilibrium problems in transportation planning studies; some program packages available are ATIM ([697]), UROAD-UTPS ([799]), TRAFFIC ([720]), and EMME/2 ([35]). It has also been extended to elastic demands (e.g., [423, 715, 717]); see Section 4.4.

The relative ease of implementing it, along with its popularity among practitioners despite its drawbacks (see below), have resulted in the development of many research-based packages (e.g., [633]), and its practical performance is therefore well known also to scientists.

---

[1] Note that the condition of $\mathbf{p}$ being a feasible direction is redundant because of the presence of the feasibility constraints ($\mathbf{x}^k + \mathbf{p} \in X$). Since there is no normalization made on the direction $\mathbf{p}$, the solution to (4.11) is an extreme point of the set X. Replacing the condition $\mathbf{x}^k + \mathbf{p} \in X$ with a normalization, such as $-1 \leq p_j \leq 1$ for all j, or $\mathbf{p}^{\mathrm{T}}\mathbf{p} \leq 1$, in the feasible set of (4.11), yields feasible direction methods of the Zoutendijk type, and successive linear programming methods (e.g., [43, Chap. 10]). Another way in which one may avoid generating extreme points is to add a nonlinear normalization term to the objective of (4.11); such methods are studied in Section 4.2.1.

[2] The reader is warned to draw too many conclusions from this interpretation; the model is inherently static, and the travellers make their route choices before leaving their origins.

Among its merits one may count its utilization of the network constraints, and the very limited core storage needed;[3] these are important properties of any method if it is to be effective in solving large-scale structured problems. Moreover, since rough approximations of the optimal solution are often acceptable in practice, and since the algorithm has been found to be efficient in the first few iterations, the algorithm is considered sufficiently good for practical use.

Remark 4.1 The development of more efficient algorithms for the traffic assignment problem has mostly been of academic concern, and few developments during the last two decades have been applied in practice (e.g., Florian [354]). The main reason for traffic equilibrium problems drawing the attention of academic researchers is not so much because of the necessity to solve real-life traffic planning problems to facilitate the improvement of the traffic system; rather, the reasons are their rich modelling possibilities, their complexity and special structure, and their size, which make them a challenge for academic research both in mathematical modelling and in the design of efficient algorithms.

One reason for the popularity of the Frank–Wolfe method may be its similarity to many of the heuristics used in transportation planning during the 50s and 60s. According to the description given in Section 1.5.4, many heuristics can be viewed as simplified versions of the Frank–Wolfe method, where the line search **Step 3** is replaced by a predetermined step length; see Table 1.2 and Figure 1.2.

4.1.4 Shortest route algorithms

In the Frank–Wolfe algorithm the vast majority of the calculations—well over 90% in large-scale applications—are spent on solving the shortest route problems in **Step 1**. The importance of choosing and implementing the most efficient shortest route algorithm is therefore obvious, and a large number of articles are devoted to this subject.

The Problem (4.3) is the problem of finding the shortest route between each pair of origin and destination in a directed graph with positive link costs ($t_a(f_a^k)$). Most algorithms used in practice for the solution of (4.3) are shortest route algorithms that yield, for a given origin node, a tree of shortest routes to all destination nodes (*tree-building* methods); in (4.3), such an algorithm would be applied $|\mathcal{O}|$ times. Due to the nonnegativity of the link costs, it is possible to apply *label-setting* shortest route methods, which have the best worst-case time complexity.

The basic label-setting algorithm is known as Dijkstra's [258] algorithm, although others discovered it independently; the first computerized assignment was made using Moore's [685] version of this method (see Section 1.5.4). The main differences among label-setting algorithms lie in the way in which a temporarily labelled node is selected, and which data structure is used to implement it ([408, 409, 14]); both the theoretical and practical time complexity vary with the implementation, which hence must be chosen carefully together with the coding of the network. Shortest route algorithms in this class for use in transportation planning are presented in [968, 253, 741, 938]; comparisons between shortest route algorithms in applications to transportation planning are made in [269, 871, 937, 252, 407, 684].

When searching for the shortest route trees for all origins in the network, the naive approach (and, in fact, the one used in practice) is to apply Dijkstra's algorithm (from scratch) $|\mathcal{O}|$ times. One alternative is to use a matrix-based algorithm for the solution

[3] Only total link flows need to be stored and the network constraints need not be considered explicitly.

of the all-pairs shortest route problem ([366, 959]); because of the high storage requirements, and the fact that usually not all pairs of nodes define O-D pairs, such algorithms are considered impractical. The search for a shortest route tree for an origin may, however, utilize the shortest route tree for another origin as a (primal infeasible but dual feasible) starting solution ([405, 406, 407]). Another possible improvement is to build two trees simultaneously, from both the origin and a destination node, instead of only one as described above ([224, 721, 487]).

As the main iterations of the Frank–Wolfe algorithm proceed, the costs $(t_a(f_a^k))$ defining (4.3) will vary little from one iteration to the next; a very natural approach is then to utilize the solution of (4.3) from iteration k as a (primal feasible, and near-optimal) starting solution in iteration $k+1$. This constitutes a potential improvement of any algorithm for [TAP] and its extensions, that utilizes shortest route calculations. This reoptimization approach is addressed in [990, 621, 786, 449, 696, 259, 405, 406, 408] (see also [871, Sec. 7.7] and [675, Sec. 2.1.5]), and then studied theoretically mostly for simple cost changes. Its practical consequences have, however, been studied very little (e.g., [260, 404]); no modern textbook on network flows includes a discussion on this important topic. One reason for this may be the need to store the shortest route trees, which previously may have been considered impractical.

The framework of *auction* algorithms for network problems (e.g., [75]) may be applied to shortest route problems ([73, 74]). It has, however, yet to be applied in a transportation planning context, although its performance is very encouraging.

4.1.5 Convergence characteristics of the Frank–Wolfe method

The convergence properties of the Frank–Wolfe algorithm have been studied extensively for applications to general nonlinear programs ([377, 242, 612, 245, 140, 1004, 246, 987, 500, 768, 279, 280]). Applied to [TAP], under Assumption 4.A, it may be shown that the sequence $\{\mathbf{f}^k\}$ converges to the unique link flow solution $\mathbf{f}^*$. In terms of commodity link flows, $\mathbf{f}_{pq}$, the optimal solution is not unique, and the Frank–Wolfe algorithm converges to the set of optimal commodity flows, i.e.,

$$\left\{ \inf_{\mathbf{f}_{pq} \in \mathbf{f}_{pq}^*} \|\mathbf{f}_{pq}^k - \mathbf{f}_{pq}\| \right\} \to 0, \qquad \forall (p,q) \in \mathcal{C}.$$

Furthermore, $\{T(\mathbf{f}^k) - \underline{T}(\mathbf{f}^k)\} \to 0$. If, in addition, the cost functions are differentiable, then ∇T is Lipschitz continuous on X (see Definition A.4), in which case the exact line search **Step 3** may be replaced by inexact line searches, such as Armijo or Goldstein step length rules (see **Rules A** and **G** in Appendix A, and [280, 747]).

The unsatisfactory performance of the Frank–Wolfe algorithm, in particular at the vicinity of an optimal solution, was observed quite early. Its convergence rate is discussed in [140, 1004, 246, 987] for applications to general convex programs [P] defined on polyhedral sets. The conclusion is that the theoretical convergence rate is *arithmetic* (or *sublinear*), i.e., that $T(\mathbf{x}^k) - T^* = O(1/k)$.[4, 5] Examples of the poor performance are given in [1003, 987].

The reason for its poor performance lies in the search direction finding phase (**Step 1**); the linear approximation is valid only locally, but is used globally when solving (4.11),

[4] A colleague of mine, who wishes to remain anonymous, says that this convergence result amounts to the Frank–Wolfe method being *convergent, but just almost.*

[5] There are actually stronger convergence results for the Frank–Wolfe algorithm (e.g., [246, 987, 279, 280]). The conditions for these results are, however, very unlikely to hold for traffic equilibrium problems.

since its variables are not normalized. The extreme point solution to this problem, and thereby the search direction generated, will therefore depend more on the structure of the feasible set than on the objective. The implication is that when the optimal solution is approached, the search directions will tend to become orthogonal to the steepest descent direction, i.e., the directional derivatives will tend to zero ([987]). Lupi [628] makes a very interesting observation in the context of traffic assignment; while the steepest descent algorithm in unconstrained optimization yields angles of 90° between successive search directions, he observes that the corresponding angle in the Frank–Wolfe algorithm is around 120°. This phenomenon agrees well with observations that both methods suffer from rapidly decreasing step lengths, and explains the term *zig-zagging* used to describe the behaviour of the Frank–Wolfe algorithm.

An interesting empirical observation is made by Janson and Zozaya-Gorostiza [524], who claim that, although an equilibrium solution can not include any cycle flows (see Section 2.2.2), cycles may be generated in the Frank–Wolfe algorithm, especially in the first iterations. The cyclic flows are very unlikely to be removed, and thus degrade the efficiency of the algorithm.[6]

This empirical result is very natural, considering that all the routes that are generated in (4.3) will retain some amount of flow in any finitely generated solution, since the step length in the line search is positive and since it is unlikely that unit steps will ever be taken. The consequence of this is of course, that also non-equilibrium routes receive positive flows in any finitely generated solution. (Individual routes for different O-D pairs are, in the Frank–Wolfe algorithm, aggregated into all-or-nothing solutions. Since equilibrium and non-equilibrium routes (in different O-D pairs) that are generated simultaneously are given equal weights, to eliminate non-equilibrium routes is very difficult.)

4.1.6 Improvements and extensions

The poor convergence of the Frank–Wolfe method observed led to the development of modifications and extensions of the original scheme. The modifications are of three fundamentally different types; either the line search is modified in order to take longer steps, the search direction is improved by combining it with previous ones, or the linear subproblem is modified in order to avoid generating extreme point solutions. Below, we describe some of these improvements.

The simplest modification is the use of predetermined step lengths in place of the line search. Although, strictly speaking, it is not appropriate to say that these methods were proposed as improvements to the Frank–Wolfe algorithm—they were actually applied even before it—some predetermined step length formulas have been observed to sometimes yield better convergence than line searches ([936]), the most important reason being that the predetermined step lengths may not tend to zero as rapidly as those determined by the line searches. Powell and Sheffi [764] prove the convergence of a Frank–Wolfe algorithm in which the line search **Step 3** is replaced by a sequence $\{l_k\}$ of step lengths in the interval $[0, 1]$. Under Assumption 4.A and the additional assumption that each t_a is differentiable, convergence of the sequence $\{\mathbf{f}^k\}$ towards the optimal link flows is ensured for sequences

[6] This property is, in fact, inherent in most algorithms for traffic assignment based on total link flows. This is due to the fact that convex combinations of cycle-free solutions may not be cycle-free, unless very restrictive assumptions hold ([402]).

of step lengths satisfying

$$\sum_{k=1}^{+\infty} l_k = +\infty, \qquad \sum_{k=1}^{+\infty} l_k^2 < +\infty. \tag{4.12}$$

The step length formula $l_k = 1/k$, $k = 0, 1, \ldots$, is the largest possible step length choice which satisfies the Condition (4.12); the sequence $\{\mathbf{f}^k\}$ is defined by the average of the previously generated all-or-nothing solutions and the resulting algorithm is therefore known as the *method of successive averages* (MSA).[7]

Weintraub *et al.* [965] propose taking a larger step than that indicated by the line search. Convergence is ensured if each step leads to a feasible solution with a lower objective value; the step given by the line search is used whenever necessary to enforce these conditions. Numerical tests are performed on randomly generated networks, with encouraging results.

Van Vliet and Dow [941] and Arezki and Van Vliet [24, 22] propose replacing (4.3) with a quantal loading procedure (see Section 1.5.4), in which the travel cost is updated between applications of the shortest route algorithm for each successive origin (or, more generally, subset of the origins); the all-or-nothing solution generated thus takes into account that some links in the network may become heavily loaded. Although convergence is not ensured (since the all-or-nothing solution is not guaranteed to yield a descent direction), convergence is observed in practice, and improves that of the original approach.

One line of development of improved Frank–Wolfe type algorithms is based on the storing and utilizing of (a few) previously generated all-or-nothing solutions in the definition of the search direction, with the objective of reducing the zig-zagging phenomenon.

The *parallel tangents* (PARTAN) approach originates in conjugate direction methods for unconstrained quadratic programs ([825, 623]). It extends the Frank–Wolfe algorithm by combining the solution obtained from each line search step with the previous solution, $\mathbf{f}^{k-1}$, thus introducing an additional line search. The PARTAN algorithm was first applied to nonlinear networks by Collins *et al.* [178] and to [TAP] by LeBlanc *et al.* [605] (see also [365, 357]). The maximal step in the second line search was first calculated explicitly using previous steps taken ([439, 365, 605]); Arezki [22, 25] subsequently derived an analytical formula which alleviated the need to store previous step lengths. The approach has been observed to reduce the zig-zagging inherent in the Frank–Wolfe algorithm; Janson and Zozaya-Gorostiza [524] however show that cycle flows may be generated in this algorithm also (see above), and propose a modification of the basic scheme and the PARTAN version in which cycles of length two are eliminated.

Fukushima [387] proposes storing a number of previously generated all-or-nothing solutions, and performing the line search **Step 3** towards a convex combination of these. Fukushima reports behaviour similar to that of PARTAN, using relatively few all-or-nothing solutions and a fairly crude choice of convex combination. This algorithm is highly related to the conceptual algorithm of Meyer [669], in which the number of line searches within one main iteration (to be applied on combinations of the extreme currently stored points) is user-specified. With the choice of only one line search, a method like Fukushima's is obtained, while, if the number is very large, the problem [TAP] is (arbitrarily accurately) solved over the convex hull of the stored all-or-nothing solutions, thus defining a simplicial decomposition algorithm for [TAP] (see Section 4.3.4).

[7] The method of Powell and Sheffi is, however, not the first application of predetermined step lengths in a Frank–Wolfe type scheme (see, e.g., [278, 279]); moreover, the step length Rule (4.12) is not limited to use in the Frank–Wolfe algorithm only.

An algorithm similar to the PARTAN approach and that of Fukushima is given by Lupi [628]. In the algorithm, the Frank–Wolfe direction, $\mathbf{y}^k - \mathbf{f}^k$, is combined with the previous direction, $\mathbf{p}^{k-1}$, such that the direction obtained, $\mathbf{p}^k$, is feasible and close to orthogonal to the direction of $\mathbf{p}^{k-1}$. Arezki [22] shows that both Lupi's and Fukushima's algorithms, when keeping only the previous all-or-nothing solution, can be optimized in terms of directional derivatives, and that they then are very similar (in fact, the first three iterates are always identical).

We next consider a subproblem of the form (4.11), with a normalization of its variables. Given a feasible solution $\mathbf{f}^k$, the cone of feasible directions for commodity (p,q) in the link-node formulation of [TAP] (see Section 2.2.2) is

$$T_{F^n_{pq}}(\mathbf{f}^k) = \left\{ \mathbf{p}_{pq} \in \Re^{|\mathcal{A}|} \;\middle|\; \mathbf{A}\mathbf{p}_{pq} = \mathbf{0},\ p_{apq} \geq 0 \text{ if } f^k_{apq} = 0,\ \forall a \in \mathcal{A} \right\}, \tag{4.13}$$

where $\mathbf{A}$ is the node-link incidence matrix, and Zoutendijk's [1016] method corresponds to solving $|\mathcal{C}|$ independent linear flow circulation subproblems

$$\min_{\mathbf{p}_{pq}} \left\{ \mathbf{t}(\mathbf{f}^k)^{\mathrm{T}}\mathbf{p}_{pq} \;\middle|\; \mathbf{p}_{pq} \in T_{F^n_{pq}}(\mathbf{f}^k),\ -1 \leq p_{apq} \leq 1,\ \forall a \in \mathcal{A} \right\}. \tag{4.14}$$

The bounds $|p_{apq}| \leq 1$ present in the constraints of (4.14) define a trust region ([345]) for the linear approximation. The algorithm has mainly been used for single-commodity flows ([45, 664, 508, 555, 962, 8, 635, 548]), but is proposed for use in [TAP] and [TAP-E] in [324, 964].

4.2 Algorithm concepts

In this section we describe three important concepts in the formulation of algorithms for the solution of [TAP]. The first is a general iterative descent algorithm based on the solution of auxiliary convex direction-finding problems, the second an approach for the decomposition of a problem defined over a Cartesian product of feasible sets into a sequence of smaller problems, and the third a scheme for algorithmically generating profitable variables in a large-scale problem. They are as follows:

(1) (*Partial linearization*) The class of partial linearization methods is a framework of descent algorithms for continuous optimization problems. A search direction is obtained from the solution of a convex auxiliary problem, defined by an approximation of the original objective through a first-order approximation of an additive part of an equivalent reformulation; alternately, a line search is made in the direction obtained with respect to the original objective. The algorithm may be applied to a variety of representations of a convex problem, such as Lagrangean dual formulations and projections of a linearly constrained problem onto the space of non-basic variables. The Frank–Wolfe algorithm is an instance of this class of methods, as is the Newton method, gradient projection and reduced gradient methods.

(2) (*Decomposition algorithms*) Disregarding the link flow defining Constraints (4.1d), the feasible set of the problem [TAP] is a Cartesian product with respect to the different commodities. The objective, however, is a function of the total link flows, and is therefore not separable in the O-D pairs. In the Frank–Wolfe algorithm, separable subproblems are obtained from linearizing the original objective. A separable subproblem is, however, obtained from any choice of a separable approximation in the partial linearization method. The resulting subproblems may be solved sequentially

or in parallel, the proper choice depending on the computer facilities available. Decomposition schemes may, in this manner, enable the solution of large-scale problems through the solution of a sequence of problems in much smaller dimensions. The classical Jacobi and Gauss–Seidel approaches for the solution of systems of nonlinear equations are instances of this general decomposition scheme.

(3) (*Column generation*) This is an algorithm principle for the solution of a mathematical programming problem where the variables (or *columns*) of the problem as they are needed. The algorithm principle consists of two main steps: in the first, the original problem is solved over the set of known variables (the so called *restricted master problem*), and in the second, the solution to this problem is the basis for the formulation of a *subproblem*, which is solved to generate variables that may improve the restricted master problem solution. A particular column generation algorithm is the result of the choice **(a)** of variable definition, **(b)** the formulation of the subproblem, and **(c)** the methodologies by which the restricted master and subproblems are solved. The choice of variable definition determines the number of constraints in the restricted master problems and the number of variables in the complete master problem (i.e., the level of aggregation), while the choice **(b)** determines the difficulty of the subproblem and the characteristics of the variables that it generates (e.g., if they are extreme points of a polyhedron). The structure of [TAP] naturally leads to certain choices of variables and types of subproblems, and the full generality of column generation has therefore not been explored. The most common approaches are to define route flows or all-or-nothing solutions as variables and generate profitable variables by solving Frank–Wolfe subproblems. The first approach is what is usually termed column generation, and the second is usually referred to as simplicial decomposition; the difference between the two, however, lies only in the level of aggregation of the variable definitions.

Nearly all the methods that have been proposed for the solution of [TAP] may be described by making proper combinations of the above three concepts. The purpose of describing algorithms in this rather unorthodox way is to make comparisons among algorithms simple, to identify the merits and drawbacks of existing methods, and to be able to bring out and justify proposals of new ones.

Below we describe these concepts in more detail.

4.2.1 Partial linearization algorithms

The general algorithm

The class of partial linearization methods to be presented in this section was introduced in [745] to characterize and interrelate a number of iterative algorithms for continuous optimization problems. The discussions here are for applications to a convex program of the form

[P]

$$\min_{\mathbf{x}\in X} T(\mathbf{x}), \tag{4.15}$$

where $T : X \mapsto \Re$ is convex and continuously differentiable on the set X, which is assumed nonempty, closed and convex. We also let Ω denote the set of optimal solutions to [P].

One iteration of the algorithm consists of the following two main steps:

(1) Given a feasible point, a feasible search direction is defined through the (possibly inexact) solution of an approximation of the original problem, in which the original objective is approximated by a convex function.

(2) The direction defined by the solution to the above described subproblem is a feasible direction of descent with respect to the original objective. A (possibly inexact) line search is made with respect to this function in the direction obtained, and the resulting step length defines a new point with a reduced value of the original objective function.

Formally, in iteration k we introduce a function $\varphi^k : X \mapsto \Re$, convex and continuously differentiable on X. Expressing the original objective in the form

$$T(\mathbf{x}) = \varphi^k(\mathbf{x}) + [T(\mathbf{x}) - \varphi^k(\mathbf{x})], \tag{4.16}$$

the second term expresses the error obtained in the objective of [P] when replacing T with φ^k. The idea of a partial linearization method is to take this error into account by a linearization of the error term, i.e., a first-order Taylor expansion of $T - \varphi^k$ at the iterate $\mathbf{x}^k$. The subproblem objective obtained equals

$$T^k_{\varphi^k}(\mathbf{x}) = \varphi^k(\mathbf{x}) + T(\mathbf{x}^k) - \varphi^k(\mathbf{x}^k) + [\nabla T(\mathbf{x}^k) - \varphi^k(\mathbf{x}^k)]^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k), \tag{4.17}$$

and the subproblem becomes

$[\mathrm{P}^k_{\varphi^k}]$

$$\min_{\mathbf{x}\in X} \; T^k_{\varphi^k}(\mathbf{x}). \tag{4.18}$$

The subproblem objective $T^k_{\varphi^k}$ is convex and continuously differentiable on X; the subproblem $[\mathrm{P}^k_{\varphi^k}]$ is hence a convex program.

The complete algorithm is described below. A sequence $\{\varphi^k\}$ of convex functions is assumed to be given. (Note that each function may be chosen adaptively, given $\mathbf{x}^k$.)

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in X$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a $\mathbf{y}^k \in X$ that solves $[\mathrm{P}^k_{\varphi^k}]$. The resulting search direction is $\mathbf{p}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ solves $[\mathrm{P}^k_{\varphi^k}] \rightarrow$ Stop ($\mathbf{x}^k$ solves [P]). Otherwise, continue.

Step 3 (*Line search*) Find a step length, l_k, which solves the one-dimensional problem

$$\min \{T(\mathbf{x}^k + l\mathbf{p}^k) \mid \mathbf{x}^k + l\mathbf{p}^k \in X,\; l \geq 0\}.$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + l_k\mathbf{p}^k$, and $k := k + 1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution $\rightarrow$ Stop. Otherwise, go to Step 1.

In most algorithms that will be identified as special cases from the class of partial linearization methods, the sequence $\{\varphi^k\}$ is given by a function of the form $\varphi^k(\mathbf{x}) = \varphi(\mathbf{x}, \mathbf{x}^k)$, i.e., a function $\varphi : X \times X \mapsto \Re$ of the form $\varphi(\mathbf{x}, \mathbf{y})$, convex and in C^1 on X with respect to $\mathbf{x}$ and continuous with respect to $\mathbf{y}$.

The algorithm here described is generalized to variational inequalities in [744, 746, 747, 588, 748]; see Sections 3.1.5 and 5.2.1 for a discussion.

Before illustrating the general algorithm by providing some well-known instances, we give the most important convergence properties of the above algorithm.

Convergence properties of partial linearization methods

Theorem 4.1 [745, 746, 747] (Properties of the search directions) *For any* $\mathbf{x} \in X$, *let* $Y(\mathbf{x})$ *denote the (possibly empty) set of optimal solutions to* $[\mathrm{P}^k_{\varphi^k}]$, $\mathbf{y} \in Y(\mathbf{x})$ *and* $\mathbf{p} = \mathbf{y} - \mathbf{x}$.

(a) $\mathbf{x} \in \Omega \iff \mathbf{x} \in Y(\mathbf{x}) \iff T_\varphi(\mathbf{y}) = T_\varphi(\mathbf{x})$.

(b) *Let* $\overline{\mathbf{x}} \in X$ *be any point such that* $T_\varphi(\overline{\mathbf{x}}) < T_\varphi(\mathbf{x})$. *Then* $\nabla T(\mathbf{x})^{\mathrm{T}}(\overline{\mathbf{x}} - \mathbf{x}) < 0$. *Especially, for* $\overline{\mathbf{x}} = \mathbf{y}$, *if* $\mathbf{x} \notin \Omega$, *then* $\nabla T(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x}) < 0$.

(c) *Let* φ *be strictly convex on* X. *If* $Y(\mathbf{x})$ *is nonempty, then it is singleton.*

(d) *Let* φ *be strongly convex on* X *(with modulus* m_φ*). Then* $Y(\mathbf{x})$ *is nonempty and singleton, and*

$$\nabla T(\mathbf{x})^{\mathrm{T}}\mathbf{p} \leq -m_\varphi \|\mathbf{p}\|^2. \tag{4.19}$$

(e) *Let* $\varphi : X \times X \mapsto \Re$ *be a continuous function on* $X \times X$ *of the form* $\varphi(\mathbf{x}, \mathbf{y})$, *convex and in* C^1 *on* X *with respect to* $\mathbf{x}$. *Then the direction finding mapping* $\mathbf{x} \mapsto D(\mathbf{x}) \stackrel{\text{def}}{=} Y(\mathbf{x}) - \mathbf{x}$ *is closed on* X.

Property **(a)** validates the termination criterion of **Step 2**, and provides a fixed point characterization of the solution set of [P].[8] The descent properties **(b)** and **(d)** are especially important in exact implementations of partial linearization algorithms. While **(b)** enables inexact solutions of $[\mathrm{P}^k_{\varphi^k}]$, **(d)** enables inexact line search strategies, such as Armijo-type rules and predetermined step lengths.

The property **(e)** implies that the class of partial linearization algorithms is globally convergent when it is supplied with a line search rule with a closed algorithmic map (such as the exact line search Rule M); this result follows from the well-known convergence theorem of Zangwill [1004, Sec. 4.5].

Theorem 4.2 [745, 746, 747] (Basic convergence of partial linearization algorithms) *Let* $\varphi : X \times X \mapsto \Re$ *be a continuous function on* $X \times X$ *of the form* $\varphi(\mathbf{x}, \mathbf{y})$, *convex and in* C^1 *on* X *with respect to* $\mathbf{x}$. *Assume that the point* $\mathbf{x}^0 \in X$ *is chosen so that the level set* $L_X^T(\mathbf{x}^0)$ *is bounded, and further that* $[\mathrm{P}_\varphi]$ *is well defined, in the sense that* $Y(\mathbf{x})$ *is nonempty and bounded for every* $\mathbf{x} \in X$. *Then, under Rule M,* $\{f(\mathbf{x}^k)\} \to f(\overline{\mathbf{x}})$ *for some* $\overline{\mathbf{x}} \in \Omega$, *any accumulation point* $\mathbf{x}^\infty$ *of the sequence* $\{\mathbf{x}^k\}$ *(at least one such point exists) lies in* Ω, *and*

$$\left\{ \inf_{\mathbf{x} \in \Omega \cap L_X^T(\mathbf{x}^0)} \|\mathbf{x}^k - \mathbf{x}\| \right\} \to 0. \tag{4.20}$$

Assuming further that X is bounded and that ∇T is Lipschitz continuous, the exact line search Rule M may be replaced by the Armijo Rule A (see Appendix A), with the same conclusions. If, in addition, each function φ^k is strongly convex on X (with modulus m_{φ^k}) and ∇T is Lipschitz continuous on X (with modulus $M_{\nabla T}$), then it may also be replaced by the predetermined step length Rule P, in which the step length l_k is chosen in the interval $(0, 2m_{\varphi^k}/M_{\nabla T})$.

[8] It also indicates the advantage of choosing the current iterate as the starting point in the solution of $[\mathrm{P}^k_{\varphi^k}]$; another desirable property of a partial linearization algorithm is the possibility to reoptimize $[\mathrm{P}^{k+1}_{\varphi^{k+1}}]$ from the solution of $[\mathrm{P}^k_{\varphi^k}]$.

The subproblem $[P^k_{\varphi^k}]$, for a general choice of function φ^k, is potentially as difficult to solve as [P]. The idea behind the *truncated* partial linearization algorithm ([746]) is to reduce the work performed on $[P^k_{\varphi^k}]$ by limiting the number of iterations performed with a descent algorithm for solving it. This strategy introduces a trade-off between the computational effort spent on solving the subproblem and the quality of the search direction obtained. The result of Theorem 4.1.b ensures that if *any* improvement is made over $T^k_{\varphi^k}(\mathbf{x}^k)$, then the point obtained will define a feasible descent direction with respect to T. Under the assumptions of Theorem 4.2, the truncated partial linearization algorithm converges if at least one iteration is performed on each subproblem $[P^k_{\varphi^k}]$ using a descent algorithm with a closed algorithmic map ([746]); under an additional Lipschitz continuity assumption on ∇T, it is possible to apply Armijo-type line searches ([747]). The criteria used for terminating the solution of $[P^k_{\varphi^k}]$ of course has a major influence on the overall convergence rate of the algorithm. If the Frank–Wolfe algorithm is used for the approximate solution of the subproblem, then its termination may be based on the (local) relative objective error (see [677]);[9] the convergence of the overall algorithm is ensured by requiring the relative error to tend to zero. If the termination criterion is defined such that each subproblem is solved sufficiently accurately, then the theoretical convergence rate of the exact algorithm may also be kept in the truncated version. (See, e.g., [239, 240, 241] for such results in the special case of Newton's method.)

Instances and interpretations

Choosing $\varphi^k \equiv 0$ for all k yields the Frank–Wolfe algorithm. The class of partial linearization methods thus generalizes the Frank–Wolfe algorithm, by allowing an additive part of the objective to be linearized, thereby retaining more of the original objective in the subproblems. A nice property of the Frank–Wolfe method is the availability of a lower bound on the optimal value of [P] (see (4.8)) obtained from the Subproblem (4.11). This property is inherited by a partial linearization algorithm only if the error function $T - \varphi^k$ is convex ([675, 583, 747]); if, however, the subproblem $[P^k_{\varphi^k}]$ is solved using a truncated Frank–Wolfe algorithm, then the first iteration in each subproblem yields the same lower bound as the Frank–Wolfe method, since $\nabla T^k_{\varphi^k}(\mathbf{x}^k) = \nabla T(\mathbf{x}^k)$. (This is an interesting special case of a truncated algorithm for the solution of $[P^k_{\varphi^k}]$, since it is easily implemented by slightly modifying an existing Frank–Wolfe code; see below for further discussions on truncated algorithms for $[P^k_{\varphi^k}]$.)

If φ^k is strictly convex, then the subproblem in iteration k has a unique solution (if one exists) [cf. Theorem 4.1.c]. An optimal solution to [P] is then also obtained from the sequence $\{\mathbf{y}^k\}$ of subproblem solutions; this is not the case in the Frank–Wolfe algorithm. Moreover, the choice of a strictly convex subproblem facilitates the use of dual methods for $[P^k_{\varphi^k}]$.

An interesting interpretation of the subproblem $[P^k_{\varphi^k}]$ is given next. Let φ^k be chosen such that $\nabla\varphi^k(\mathbf{x}^k) = \mathbf{0}$. Then the subproblem $[P^k_{\varphi^k}]$ is equivalent to

$$\min_{\mathbf{x}\in X} \{\nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k) + \varphi^k(\mathbf{x})\}. \tag{4.22}$$

[9] Given some $\varepsilon_k > 0$, a point $\mathbf{y}^k_{\varepsilon_k} \in X$ is then found, satisfying

$$\nabla T^k_{\varphi^k}(\mathbf{y}^k_{\varepsilon_k})^{\mathrm{T}}(\mathbf{y} - \mathbf{y}^k_{\varepsilon_k}) \geq -\varepsilon_k, \qquad \forall \mathbf{y} \in X, \tag{4.21}$$

i.e., $\mathbf{y}^k_{\varepsilon_k}$ is an ε_k-optimal solution to $[P^k_{\varphi^k}]$ [247, Le. 1.8.2]. The validity of (4.21) is checked automatically in the Frank–Wolfe algorithm, since the left hand side of (4.21) is minimized in the subproblem phase.

In the Frank–Wolfe algorithm, the first-order approximation, which is valid only locally around the iterate $\mathbf{x}^k$, is used globally in the Subproblem (4.11). When φ^k is chosen nonlinear, as shown in (4.22), the subproblem in the partial linearization algorithm introduces a regularization term in the objective function of the Frank–Wolfe subproblem, restricting (indirectly through the penalization term) the distance between the current iterate and the corresponding subproblem solution. A particularly illustrative example is the choice $\varphi^k = 1/(2\gamma_k)\|\cdot\|^2$, $\gamma_k > 0$, in which case the objective of (4.22) becomes

$$\nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2\gamma_k}\|\mathbf{x} - \mathbf{x}^k\|^2; \tag{4.23}$$

this subproblem (which is the subproblem of the gradient projection method of Goldstein, Levitin and Polyak [432, 612]) may be viewed as a relaxation of the subproblem of Zoutendijk-type methods [see, e.g., (4.14)] in the sense that, instead of introducing explicit trust regions (or normalization bounds, such as $\|\mathbf{x} - \mathbf{x}^k\|^2 \leq 1$) on the subproblem variables, trust regions are introduced implicitly through the regularization term (see also [675, Sec. 4.5.2]).

By retaining the nonlinearity of the original objective function, partial linearization methods may therefore avoid the zig-zagging phenomenon inherent in the Frank–Wolfe algorithm caused by the generation of extreme point subproblem solutions.

Subproblems of the form (4.22) are inherent in the regularized Frank–Wolfe algorithm of Migdalas [677] and the nonlinear proximal descent method of Tseng [912]; the class of partial linearization algorithms has a strong relationship also to the auxiliary problem principle of Cohen [174, 175] (see [747, 588, 749]).

The choice of $\varphi^k(\mathbf{x}) = 1/(2\gamma_k)\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x}$, where $\gamma_k > 0$ and $\mathbf{B}_k$ is symmetric and positive semidefinite, yields a subproblem objective

$$\nabla T(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2\gamma_k}(\mathbf{x} - \mathbf{x}^k)^{\mathrm{T}}\mathbf{B}_k(\mathbf{x} - \mathbf{x}^k) \tag{4.24}$$

of the *deflected gradient projection algorithm*; if $\mathbf{B}_k$ is positive definite, then $T^k_{\varphi^k}$ is strongly convex, and the solution to $[\mathrm{P}^k_{\varphi^k}]$ is given by

$$\mathbf{y}^k = P_X^{\mathbf{B}_k}(\mathbf{x}^k - \gamma_k\mathbf{B}_k^{-1}\nabla T(\mathbf{x}^k)), \tag{4.25}$$

where $P_X^{\mathbf{B}_k}(\mathbf{x})$ denotes the projection of $\mathbf{x}$ onto X with respect to the norm $\|\mathbf{x}\|_{\mathbf{B}_k} = (\mathbf{x}^{\mathrm{T}}\mathbf{B}_k\mathbf{x})^{1/2}$. For the choice $\mathbf{B}_k = \mathbf{0}$, for all k, we recover the Frank–Wolfe algorithm, while the gradient projection algorithm is obtained from choosing $\mathbf{B}_k = \mathbf{I}$. *Newton's method* ([612, 768, 281]) is obtained by the choices $\mathbf{B}_k = \nabla^2 T(\mathbf{x}^k), \gamma_k = 1$; *quasi-Newton* methods follow from choices of $\mathbf{B}_k$ approximating $\nabla^2 T(\mathbf{x}^k)$.

Hence, the class of partial linearization methods includes algorithms with convergence rates ranging from sublinear (the Frank–Wolfe algorithm) to quadratic (Newton's method).

The class of partial linearization methods also includes a variety of *regularization* methods, where, in the subproblem, a strictly convex objective is added to the original objective function. Let $\varphi^k(\mathbf{x}) = T(\mathbf{x}) + r^k(\mathbf{x})$. Then

$$T^k_{\varphi^k}(\mathbf{x}) = T(\mathbf{x}) + r^k(\mathbf{x}) - r^k(\mathbf{x}^k) - \nabla r^k(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k);$$

if $r^k \equiv r$, then $D_r(\mathbf{x}) = r(\mathbf{x}) - r(\mathbf{x}^k) - \nabla r(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^k)$ is the *D-function* of Bregman [116, 117, 118], also studied in [157, 234, 158, 892, 164, 293]. If $r^k(\mathbf{x}) = (1/\gamma_k)r(\mathbf{x}, \mathbf{x}^k)$, $\gamma_k > 0$, with $\nabla r^k(\mathbf{x}^k) = \mathbf{0}$, then the traditional form, $T^k_{\varphi^k}(\mathbf{x}) = T(\mathbf{x}) + (1/\gamma_k)r(\mathbf{x}, \mathbf{x}^k)$, of the

subproblem objective of regularization methods ([896, 612, 897, 760]) is obtained. The main reason for considering regularization methods is the wish to strictly convexify a non-strictly convex function. An extensively studied special case is obtained from the choice $r^k(\mathbf{x}) = 1/(2\gamma_k)\|\mathbf{x}\|^2$, which yields $T^k_{\varphi^k}(\mathbf{x}) = T(\mathbf{x}) + 1/(2\gamma_k)\|\mathbf{x}-\mathbf{x}^k\|^2$, i.e., the subproblem objective of the *proximal point method* ([680, 687, 651, 652, 761, 781, 782, 783, 70, 629, 84]). (This algorithm, in turn, includes a number of *splitting algorithms* (e.g., [120, 294]); for an objective of the form $T = g + h$, φ^k is given by $g + 1/(2\gamma_k)\|\cdot\|^2$.)

The reader should note that a given choice of partial linearization, when applied to different representations of the problem [P], yields different algorithms. For example, gradient projection type partial linearizations, when applied to a problem which is projected onto the space of non-basic variables, yield *reduced gradient* methods (e.g., [43, Sec. 10.6]), and to steepest descent methods when applied to a Lagrangean dual formulation; similarly, Newton methods turn into second-order reduced gradient methods when applied to the former representation, and further to *sequential quadratic programming* methods (e.g., [43, Sec. 10.4]) when applied to the Karush–Kuhn–Tucker optimality conditions of [P]. A final example is provided by the proximal point algorithm; when applied to a Lagrangean dual formulations of [P], it yields the class of *augmented Lagrangean methods* (e.g., [491, 763, 445, 780, 781, 782, 70, 84]), also known as the *method of multipliers*.

Utilization of problem structures

The large freedom of choice of function φ^k enables the partial linearization methods to adapt to problem structures; the most important ones in the context of traffic assignment are listed on page 96.

A partial linearization algorithm that may be said to utilize the *network structure* must execute the vast majority of the operations involved in the solution of $[\mathrm{P}^k_{\varphi^k}]$ directly on the network; in order to define an efficient algorithm for large-scale problems, the problem $[\mathrm{P}^k_{\varphi^k}]$ should not be very difficult to solve repeatedly, nor should the storage requirements for carrying out the operations increase very rapidly with the network size. Several algorithms that may be placed within the framework of partial linearization have been successfully specialized to nonlinear networks, including convex simplex (e.g., [801, 714, 178, 486]), reduced gradient (e.g., [46]), scaled reduced gradient (e.g., [237]), Newton (e.g., [559, 10, 236, 560, 902, 514, 1011]), and gradient projection methods (e.g., [400]).

In network problems with additional constraints, the network components must be identified (e.g., [90, 126, 125, 91]), and the non-network components treated separately. The most common strategy has been to develop specializations of simplex-type algorithms, where basis partitioning techniques are used to separate the network basis from the non-network components of the model (e.g., [548, Chap. 7]). An approach less investigated (but one which is more general, since it may be applied to nonlinear side constraints) is the handling of the non-network constraints through dualization/penalization techniques. In such an approach, the side constraints are included in an extended objective function by means of a penalty function, or a (augmented) Lagrangean function with parameters including Lagrange multipliers for the dualized constraints ([484, 587, 757]); the subproblems are then pure network problems.

The utilization of network structures can be enforced on a partial linearization algorithm through the proper choice of solution procedure for the corresponding subproblems $[\mathrm{P}^k_{\varphi^k}]$. One particularly interesting choice of algorithm for the solution of $[\mathrm{P}^k_{\varphi^k}]$ is the truncated Frank–Wolfe algorithm, which may be easily implemented on the basis of an existing Frank–Wolfe code.

The objective of [TAP] is separable and strictly convex in the total link flows (see Section 2.3.2). In order to define a rapidly convergent partial linearization algorithm, the choice of the sequence $\{\varphi^k\}$ must be made such that the *nonlinearity* of the objective is preserved; one should, however, note that each subproblem should be effectively solvable. (In the extreme case, with $\varphi^k \equiv T$, the subproblem is equivalent to the original problem.) In addition, choosing strictly convex functions φ^k ensures that the optimal solution may be identified from the sequence of subproblem solutions, and that methods based on duality may be utilized for their solution (e.g., [43, Sec. 6.5]).

The standard traffic assignment problem has a favourable constraint structure; the requirements on each commodity flow is independent of the requirements on the other commodity flows, i.e., the constraints define a Cartesian product of feasible sets. The objective, however, is defined by the sum of the independent commodity flows, and hence it is nonseparable. If the objective is replaced by a separable approximation, then the corresponding traffic assignment problem decomposes into as many single-commodity problems as there are commodities in the network; these problems may either be solved in parallel or sequentially in a manner similar to Gauss–Seidel methods.

In the next section, we study convex problems defined on Cartesian products of feasible sets, and show how partial linearization algorithms may define sequential and parallel decomposition methods for such problems.

4.2.2 Decomposition algorithms

Let the constraints defining the feasible set X of [P] in (4.15) be given by

$$\begin{pmatrix} \boxed{\mathbf{B}_1} & & & \\ & \boxed{\mathbf{B}_2} & & \\ & & \ddots & \\ & & & \boxed{\mathbf{B}_m} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_m \end{pmatrix}, \tag{4.26a}$$

$$\mathbf{x}_i \geq \mathbf{0}, \qquad i = 1, \ldots, m, \tag{4.26b}$$

where $m \geq 1$, $\mathbf{x} = \prod_{i=1}^m \mathbf{x}_i$, $\mathbf{x}_i \in \Re^{n_i}$, $\sum_{i=1}^m n_i = n$, and the matrices $\mathbf{B}_i$ and vectors $\mathbf{d}_i, i = 1, \ldots, m$, have appropriate dimensions. Define the Cartesian product $X = \prod_{i=1}^m X_i$, where

$$X_i = \left\{ \mathbf{x}_i \in \Re_+^{n_i} \;\middle|\; \mathbf{B}_i \mathbf{x}_i = \mathbf{d}_i \right\}, \qquad i = 1, \ldots, m.$$

Let $\varphi^k : X \mapsto \Re$ be of the form $\varphi^k(\mathbf{x}) = \sum_{i=1}^m \varphi_i^k(\mathbf{x}_i)$, where $\varphi_i^k : X_i \mapsto \Re$ is a convex function in C^1 on X_i, i.e., a function of the variables $\mathbf{x}_i$ only. Then the subproblem objective becomes

$$T_{\varphi^k}^k(\mathbf{x}) = \sum_{i=1}^m T_{\varphi_i^k}^k(\mathbf{x}_i) = \sum_{i=1}^m \{\varphi_i^k(\mathbf{x}_i) + [\nabla_i T(\mathbf{x}^k) - \nabla \varphi_i^k(\mathbf{x}_i^k)]^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_i^k)\},$$

where $\nabla_i T$ denotes the partial derivative of T with respect to $\mathbf{x}_i$. The subproblem $[\mathrm{P}_{\varphi^k}^k]$ thus separates into m independent problems

$[\mathrm{P}_{\varphi_i^k}^k]$

$$\min_{\mathbf{x}_i \in X_i} \; T_{\varphi_i^k}^k(\mathbf{x}_i). \tag{4.27}$$

Based on the expositions in [746, 747] we shall below relate sequential and parallel partial linearization methods for the solution of [P] to the classical Jacobi and Gauss–Seidel algorithms for the solution of systems of nonlinear equations ([945, 727]), and give convergence results for implementations with different degrees of parallelism and asynchronism.

Sequential decomposition algorithms

In the sequential (Gauss–Seidel) version of the decomposition algorithm, we choose in iteration k the index $i_k \in \{1, \ldots, m\}$, and solve $[\mathrm{P}^k_{\varphi^k_{i_k}}]$, with the solution $\mathbf{y}^k_{i_k}$, and then let

$$\mathbf{x}_i^{k+1} = \begin{cases} \mathbf{x}_i^k + l_k(\mathbf{y}_i^k - \mathbf{x}_i^k), & i = i_k, \\ \mathbf{x}_i^k, & \text{otherwise}, \end{cases}$$

where l_k minimizes $T(\mathbf{x}^k_{i_k-}, \mathbf{x}^k_{i_k} + l(\mathbf{y}^k_{i_k} - \mathbf{x}^k_{i_k}), \mathbf{x}^k_{i_k+})$ over $l \in \{\, l \geq 0 \mid \mathbf{x}^k_{i_k} + l(\mathbf{y}^k_{i_k} - \mathbf{x}^k_{i_k}) \in X_i \,\}$ (if the exact line search Rule M is used).

Convergence of this method is guaranteed under conditions similar to those of Theorem 4.2 when indices i_k are chosen according to the *cyclic rule*,

$$i_k = k \pmod m + 1. \tag{4.28}$$

Truncation strategies for the solution of the subproblems are also valid under conditions similar to those mentioned above for the original partial linearization scheme; if ∇T is Lipschitz continuous and the functions φ_i are chosen strictly convex, the algorithm may also be supplied with inexact line searches, and the cyclic rule may be replaced by the less restrictive *essentially cyclic rule* ([610, 155, 909, 912, 626]), in which every index $i \in \{1, \ldots, m\}$ is assumed to be chosen at least once every B successive iterations, i.e., there is a $B \geq m$ such that[10]

$$\{1, \ldots, m\} \subseteq \{i_k, \ldots, i_{k+B-1}\}, \qquad \forall k \geq 1. \tag{4.29}$$

Under the additional assumption that each function φ_i is strongly convex, predetermined step lengths may be used; these may be chosen individually for the different components, with upper limits $2m_{\varphi_i^k}/M_{\nabla T}$.

The above algorithm is a sequential partial linearization method in which the information obtained from the updating of the variable component $\mathbf{x}_i$ is utilized immediately

[10]The possibility of executing the algorithm with the essentially cyclic rule brings forward possibilities for devising rules for choosing indices that may speed up the practical convergence, compared to the cyclic rule. The basis for this is the fact that the bound B can be made arbitrarily large; within this bound the indices may be chosen according to any specified rule. A good strategy may be to generalize the Gauss–Southwell (remotest) order ([155, 623]), in which the index chosen in a particular iteration is the one corresponding to the variable block being, in a certain sense, farthest from the set of optimal solutions. The essentially cyclic rule may, in turn, be replaced by the *free-steering* order ([811, 812, 296, 727, 674, 331]), in which each index is only assumed to be chosen an infinite number of times in the sequence $\{i_k\}$. If the sequence $\{\mathbf{x}^k\}$ is generated by the decomposition version of the partial linearization algorithm, then it is possible to show that

$$\nabla_i T(\mathbf{x}^\infty)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_i^\infty) \geq 0, \qquad \forall \mathbf{x}_i \in X_i$$

holds for those indices that occur infinitely many times in building the subsequence converging to $\mathbf{x}^\infty$. This result is analogous to those obtained earlier for general fixed point problems in [330, 331], and can probably not be strengthened without introducing restrictive convexity conditions on T. (Convergence under the free-steering rule has been established for the Gauss–Seidel method applied to specially structured strictly convex problems [735, 83, 84] and strongly convex unconstrained programs [811, 812, 296, 727].)

in the update of $\mathbf{x}_{i+1}$, as opposed to the parallel algorithms to be described below; this approach enables the solution of large-scale problems through a decomposition into sequences of problems of smaller dimensions.

We next show that the above cyclic version of the partial linearization algorithm includes a block version of the classical Gauss–Seidel algorithm (or the *method of successive replacements*) as a special case. To this end, we will write the objective as $T(\mathbf{x}_{i-}, \mathbf{x}_i, \mathbf{x}_{i+})$. One iteration of the block Gauss–Seidel algorithm for the solution of [P] is defined through the following m subproblems, solved sequentially:

[G–S^k]

$$\min_{\mathbf{x}_i \in X_i} \; T(\mathbf{x}_{i-}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+}^{k}), \qquad i = 1, 2, \ldots, m. \tag{4.30}$$

Let

$$\varphi_i^k(\mathbf{x}_i) = T(\mathbf{x}_{i-}^{k}, \mathbf{x}_i, \mathbf{x}_{i+}^{k}). \tag{4.31}$$

Then $\nabla \varphi_i^k(\mathbf{x}_i^k) = \nabla_i T(\mathbf{x}^k)$, and hence the objective of [P$^k_{\varphi_i^k}$] equals that of [G–S$_i^k$]. Since T is minimized in each step during iteration k we have that $\mathbf{x}_i^{k+1} = \mathbf{y}_i^k$, and the sequence of problems [P$^k_{\varphi_i^k}$] over $i \in \{1, \ldots, m\}$ is equivalent to the sequence of problems [G–S^k]. (See [811, 812, 1001, 727, 84] for convergence results for the Gauss–Seidel algorithm.)

Parallel decomposition algorithms

The above mentioned methods are inherently sequential; Jacobi-type versions (or *methods of simultaneous replacements*) of the partial linearization method, in which subproblems are allowed to be solved in parallel, are therefore introduced in this section.

Assume that we have access to a parallel computer with m independent processors, each one responsible for a component $\mathbf{x}_i$ of $\mathbf{x}$. The above decomposition scheme can then be alternatively implemented in such a way that the subproblems [P$^k_{\varphi_i^k}$], $i \in \{1, \ldots, m\}$, are solved simultaneously, followed by a global (possibly approximate) line search. The resulting parallel partial linearization algorithm has the same convergence properties as the original partial linearization algorithm (with the exception that individual step lengths may be used in Rule P).

The block Jacobi algorithm is the result of choosing the functions φ_i^k according to (4.31), in which case the independent subproblems (solved in parallel) become

[J$_i^k$]

$$\min_{\mathbf{x}_i \in X_i} \; T(\mathbf{x}_{i-}^{k}, \mathbf{x}_i, \mathbf{x}_{i+}^{k}). \tag{4.32}$$

Note that the Jacobi algorithm is traditionally not supplied with a line search ([84, Sec. 3.3]).

To make sure that the sequence of iterates generated by the parallel partial linearization algorithm agrees with that given by the original (sequential) method, the implementation requires a *synchronization mechanism* [513, 499, 84, 438], by which the processors are coordinated to operate on the correct data and in the correct sequence. Although the parallel algorithm may speed-up the practical convergence rate, the need for a synchronizing step in the algorithm (the exchange of subproblem solutions among processors, and possibly a global line search) may still deteriorate the performance, since faster processors (subproblems) must wait for slower ones before the information exchange can be

made. The efficiency can be further degraded by memory conflicts or slow communication channels ([575, 84]). In addition, if some (inexact) line search is used, the efficiency is degraded due to an increase in serial computations. (The possible inefficiencies resulting from the need for a synchronizing step are further discussed in [191, 84].) We therefore also consider asynchronous versions of the Jacobi-type methods, in which processors do not wait to receive the latest information available. The advantage of such an approach is a minimal delay in communication, which may speed-up the convergence compared to synchronous algorithms. Convergence still holds, provided that the information that any particular processor utilizes is not arbitrarily outdated.

Removing the synchronization of the processors enables the faster processors to execute more iterations since they are not required to wait for the most recent results to become available. Because the speed of computations and communications can vary among the processors, and communication delays can be substantial, the processors will, as a result, perform the calculations out of phase with each other. Thus, the advantage of a reduced synchronization penalty is paid for by an increase in interprocessor communications, a use of outdated information that may be counterproductive if certain conditions are not met, and a more difficult convergence detection, see [84]. (Certainly, the convergence analysis also becomes much more complicated.) Recent numerical experiments indicate, however, that the introduction of *asynchronous* computations can substantially enhance the efficiency of parallel iterative methods ([295, 76, 159]).

In a *partially asynchronous* parallel algorithm, there is an assumed upper bound on the communication delays and differences in the frequency of computation of different processors.[11]

In the partially asynchronous partial linearization algorithm, each processor i calculates the subproblem solution $\mathbf{y}_i^k$ based on the latest information available on the components $\mathbf{x}_j$, $j \in \{1, \ldots, m\}$, updates its own variable component by a predetermined step length, and communicates the result to all the other processors. It then resumes calculations on a new subproblem, based on the new information about its own component and possibly new information received about other variable components.

Denoting the upper bound on the communication delays and processor idle times by B, global convergence for the partially asynchronous partial linearization algorithm is guaranteed for Lipschitz continuous functions ∇T and $\nabla \varphi_i$ and strongly convex functions φ_i, under the condition that the step lengths used in the update of the variables are bounded above by $m_\varphi / (M_{\nabla T}[1/2 + (m+1)B])$, where $m_\varphi = \min_{i \in \{1,\ldots,m\}} m_{\varphi_i}$ ([746, 747]).[12]

4.2.3 Column generation algorithms

The general algorithm

Consider the linearly constrained convex program

[11] Partially asynchronous versions of partial linearization algorithms have only been studied to a limited extent; deflected gradient methods in unconstrained programming [84, Sec. 7.5], gradient projection methods [916, 917, 918, 84, 913], and coordinate ascent methods for strictly convex network flows [915] are the only examples to date.

[12] Interesting observations can be made regarding the relationships among the maximal step length, the allowed amount of asynchronism, the number of processors involved in the calculations, and the properties of the given problem, that are brought forward in this expression. Compare the maximal step length allowed with that of the original algorithm, and the sequential decomposition version. The decreasing intervals of allowed step lengths is a consequence of the decreasing quality of the step directions, resulting from the usage of more outdated information in the update of the variables.

[P]

$$\min\ T(\mathbf{x}), \tag{4.33a}$$

subject to

$$\begin{pmatrix} \mathbf{A} & & & \\ \mathbf{B}_1 & & & \\ & \mathbf{B}_2 & & \\ & & \ddots & \\ & & & \mathbf{B}_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_m \end{pmatrix}, \tag{4.33b}$$

$$\mathbf{x}_i \geq \mathbf{0}, \qquad i = 1, \ldots, m, \tag{4.33c}$$

where $m \geq 1$, $\mathbf{x} = \prod_{i=1}^{m} \mathbf{x}_i$, $\mathbf{x}_i \in \Re^{n_i}$, $\sum_{i=1}^{m} n_i = n$, and the matrices $\mathbf{A}$, $\mathbf{B}_i$ and vectors $\mathbf{b}$, $\mathbf{d}_i, i = 1, \ldots, m$, have appropriate dimensions. [Constraints of the form (4.33b) are termed *block angular* (e.g., [554, 800, 818, 819]).] Define $X = \prod_{i=1}^{m} X_i$, where

$$X_i = \left\{ \mathbf{x}_i \in \Re_+^{n_i} \;\middle|\; \mathbf{B}_i \mathbf{x}_i = \mathbf{d}_i \right\}, \qquad i = 1, \ldots, m,$$

and

$$Z = \{ \mathbf{x} \in \Re^n \mid \mathbf{A}\mathbf{x} = \mathbf{b} \};$$

the problem [P] may then be written as

$$\min_{\mathbf{x} \in Z \cap X}\ T(\mathbf{x}).$$

We assume that $Z \cap X$ is nonempty and that X is bounded. According to the Representation Theorem [cf. (2.14)], each set X_i may then be given an internal representation as the convex hull of its extreme points $\mathbf{y}_i^j$, $j \in \mathcal{X}_i$,

$$X_i = \left\{ \mathbf{x}_i \in \Re^{n_i} \;\middle|\; \mathbf{x}_i = \sum_{j \in \mathcal{X}_i} \lambda_i^j \mathbf{y}_i^j,\ \sum_{j \in \mathcal{X}_i} \lambda_i^j = 1,\ \lambda_i^j \geq 0,\ \forall j \in \mathcal{X}_i \right\}, \tag{4.34}$$

or, compactly, $X_i = \texttt{conv}\ (\mathcal{X}_i)$. (Disregarding the constraints defining $\mathbf{x}_i$ in (4.34), this set is, in the variables λ_i^j, a $(|\mathcal{X}_i| - 1)$-*simplex*.)

The problem [P] may hence be equivalently formulated as the *complete master problem*

[CMP]

$$\min\ T(\mathbf{x}), \tag{4.35a}$$

subject to

$$\mathbf{x} \in Z \cap \prod_{i=1}^{m} \texttt{conv}\ (\mathcal{X}_i). \tag{4.35b}$$

The original problem [P] has n variables, while the complete master problem [CMP] has $\sum_{i=1}^{m} |\mathcal{X}_i|$ variables. The number of linear constraints in the two equivalent problems differ in the representation of the sets X_i. The complete master problem has m convexity constraints, representing these sets; this number is, in general, much lower than the number of constraints representing the sets X_i in the original formulation.

In general, the number of variables in [CMP] is much larger than in the original problem; if, however, the extreme points of the sets X_i that are necessary to express the components $\mathbf{x}_i$ of an optimal solution of [P] are *known* (the number of these variables is in general much less than the total number of variables), then due to the simple structure of the constraints of the master problem, the optimal solution to [P] may be obtained efficiently.

The idea behind a *column generation* scheme is to algorithmically generate variables (or columns) in the sets X_i that potentially may be used to span convex hulls that include the components $\mathbf{x}_i$ of an optimal solution.

Let $\hat{\mathcal{X}}_i$ be a set of known points—not necessarily extreme points—in X_i, and consider the *restricted master problem*

[RMP]

$$\min\ T(\mathbf{x}), \tag{4.36a}$$

subject to

$$\mathbf{x} \in Z \cap \prod_{i=1}^{m} \texttt{conv}\ (\hat{\mathcal{X}}_i), \tag{4.36b}$$

i.e., a restriction of the problem [P] to the subsets $\texttt{conv}\ (\hat{\mathcal{X}}_i)$ of X_i, $i = 1, \ldots, m$.

Alternately to solving a restricted master problem, in the column generation scheme a vector (column) in $X_i \setminus \texttt{conv}\ (\hat{\mathcal{X}}_i)$, $i = 1, \ldots, m$, is generated through the solution of a *subproblem* in order to enable the solution to [RMP] to be improved. The form of the subproblem and the techniques employed for its solution vary with the application; see Lasdon [590] for an overview. Column generation methods also usually include schemes for dropping previously generated columns that are no longer believed to be necessary in order to express an optimal solution.

Instances and definitions of columns

Dantzig–Wolfe decomposition ([227, 228, 225, 822, 418, 590, 261]) is a well-known special case from the column generation principle, in which columns are generated as solutions of approximations of [P], where the (coupling) constraints defining Z are Lagrangean dualized; the objective of the restricted master problem is traditionally defined by an inner linearization of T at the generated points. Extensions of this approach include the use of nonlinear penalizations of the dualized constraints (known as nonlinear pricing [528, 529, 322])—which are related to augmented Lagrangean methods for [P] (e.g., [70])—, and approximations of T in the subproblem (e.g., [503]).

The convergence of column generation methods is based on Carathéodory's Theorem ([142]; see also [43, Th. 2.1.6]), which states that any point in X_i—and in particular the corresponding components of an optimal solution of [P]—may be represented as a convex combination of at most $n_i + 1$ points in X_i. (Thus, in total $\sum_{i=1}^{m}(n_i + 1) = n + m$ points suffice.) [This result may be sharpened to requiring only $\texttt{dim}\ (X_i) + 1$ points, where $\texttt{dim}\ (X_i)$ is the dimension of the affine hull of X_i (e.g., [877, Th. 2.2.12]).] Note that for bounded polyhedra, the Representation Theorem is a special case of Carathéodory's Theorem in which the columns that are used to express an arbitrary point in X_i are extreme points of X_i.

Although it is not necessary to consider extreme points only, the traditional algorithmic approach for the solution of [P] through the solution of a sequence of restricted master problems with increasing sizes, is based on the generation of extreme points of the sets X_i,

through the solution of linear programming subproblems. (In the terminology of Geoffrion [418], such column generation algorithms are algorithms based on *inner linearization followed by restriction.*)

In traditional presentations of column generation methods, the restricted master problem contains only one convexity constraint. In the context of the problem [P], it would mean that the points $\mathbf{y}_i^j \in \Re^{n_i}$, $i = 1, \ldots, m$, that are generated simultaneously, would be concatenated into *one* column, $\mathbf{y}^j \in \Re^n$ for all $j \in \hat{\mathcal{X}}$; the restricted master problem then becomes

[RMP]

$$\min\ T(\mathbf{x}), \tag{4.37a}$$

subject to

$$\mathbf{x} \in Z \cap \mathtt{conv}\ (\hat{\mathcal{X}}). \tag{4.37b}$$

We shall refer to this as an *aggregated* master problem, as opposed to the *disaggregated* master Problem (4.36).

The complete master problem, in which $\hat{\mathcal{X}}$ in (4.37) is replaced by the set $\mathcal{X}$ of all extreme points of X, contains $|\mathcal{X}| = \prod_{i=1}^m |\mathcal{X}_i|$ columns. Contrary to what one might expect, aggregating the columns results in a much larger complete master problem, since $\prod_{i=1}^m |\mathcal{X}_i| \gg \sum_{i=1}^m |\mathcal{X}_i|$. (The maximum number of points needed, however, reduces from $n + m$ to $n + 1$, according to Carathéodory's Theorem.)

For some special cases of sets Z, the points $\mathbf{y}_i^j$ may be aggregated such that the dimensions of the columns $\mathbf{y}^j$ are reduced drastically. Assume that

$$T(\mathbf{x}) = T\left(\sum_{i=1}^m \mathbf{x}_i\right),$$

and that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{I} & \ldots & \mathbf{I} \\ \mathtt{diag}\ (a_2) & \mathtt{diag}\ (a_2) & \ldots & \mathtt{diag}\ (a_2) \end{pmatrix},$$

where $\mathbf{I} \in \Re^{n \times n}$ is an identity matrix, and $\mathtt{diag}\ (a_2) \in \Re^{n \times n}$ is a diagonal matrix. We may then redefine the variable vector $\mathbf{x}$ as

$$\mathbf{x} = \sum_{i=1}^m \mathbf{x}_i, \tag{4.38}$$

and view the problem [P] is terms of variables $\mathbf{x}$. Instead of concatenating the vectors $\mathbf{y}_i^j$ when defining a column $\mathbf{y}^j$ of [RMP], we let

$$\mathbf{y}^j = \sum_{i=1}^m \mathbf{y}_i^j, \qquad \forall j \in \hat{\mathcal{X}}; \tag{4.39}$$

the dimension of these columns is m times lower than that of the concatenated ones.

Letting $Z_2 = \{\, \mathbf{x} \in \Re^n \mid \mathbf{A}_2\mathbf{x} = \mathbf{b}_2 \,\}$, the restricted master problem becomes

[RMP]

$$\min\ T(\mathbf{x}), \tag{4.40a}$$

subject to

$$\mathbf{x} \in Z_2 \cap \texttt{conv}\,(\hat{\mathcal{X}}), \tag{4.40b}$$

where $\hat{\mathcal{X}}$ denotes the set of points defined by (4.39). Note that the total number of variables in a complete master problem of the form (4.40) equals that of the complete master problem of the form (4.37), although the columns generated are of much lower dimension. (The maximum number of columns needed changes drastically, from $m(n+1)$ in the disaggregated formulation of [RMP] to $n+1$ in the aggregated one.)

When performing the aggregation defined by (4.39), information is lost; in order to recover a solution to the disaggregated formulation, some additional work is needed.

Simplicial decomposition

Simplicial decomposition ([952, 953]) is a special case of column generation in which the column generating subproblem is the same as the direction-finding subproblem of the Frank–Wolfe algorithm,[13] and the original objective is used in the solution of the restricted master problem. The basic decomposition theory outlined above, thus applies to this methodology. The notion of simplicial decomposition (or SD for short) is due to von Hohenbalken [952, 953]. Techniques of this type have a longer history however:

(1) (*Improvement of Frank–Wolfe*) Simplicial decomposition type methods have been studied as improvements of the Frank–Wolfe algorithm ever since its zig-zagging behaviour became apparent (e.g., [1004, 697, 987, 504, 669]). Most of these discussions are intuitive, but a formalization naturally leads to a simplicial decomposition approach.

(2) (*Quadratic programming*) Simplicial decomposition is highly related to finite methods in quadratic programming. Relationships among the capacity method ([505]), Wolfe's least-distance programming method ([988, 989]), and some pivoting methods ([985, 225, 927, 926, 931]) are given in [928, 929, 930, 86]. Cottle and Djang [183] show that the least-distance method produces the same sequence of iterates as simplicial decomposition when applied to this special quadratic program. Pang [734] extends this result to hold for any convex quadratic program, when the symmetric programming algorithm ([931]) is applied to its (implicit) inner representation. See Djang [262] for a thorough study on this topic, and also [802, 835].

(3) (*Network flows*) One of the first methods to be implemented for use in traffic assignment was developed in the Metropolitan Toronto Regional Transportation Study; the algorithm may be viewed as a version of simplicial decomposition, in which shortest routes define (disaggregated) restricted master problems of *a priori* bounded size which are heuristically solved (see Section 1.5.4 for more details). Among the first convergent methods to be proposed and tested for traffic assignment we also find column generation methods of the simplicial decomposition type (e.g., [424]), and simplicial decomposition was early applied to computer communication networks ([420, 141]). For more details of applications of column generation to traffic assignment, see Sections 4.3.4 and 4.3.5.

[13] In primal applications dual variables are not involved in simplicial decomposition algorithms.

To simplify the discussion, let us consider the linearly constrained convex program [P] of the form (4.7), where the feasible set X is a bounded polyhedron.[14] Recall that the Frank–Wolfe Subproblem (4.11) yields an extreme point, say $\mathbf{y}^k$, of X, given a point $\mathbf{x}^k \in X$.[15] The point $\mathbf{y}^k$ is added to the set $\hat{\mathcal{X}}$ of known extreme points of X, and the restricted master problem

[RMP]

$$\min\ T(\mathbf{x}), \tag{4.41a}$$

subject to

$$\mathbf{x} = \lambda^0 \mathbf{x}^k + \sum_{j \in \hat{\mathcal{X}}} \lambda^j \mathbf{y}^j, \tag{4.41b}$$

$$\lambda^0 + \sum_{j \in \hat{\mathcal{X}}} \lambda^j = 1, \tag{4.41c}$$

$$\lambda^0, \lambda^j \geq 0, \qquad \forall j \in \hat{\mathcal{X}} \tag{4.41d}$$

is solved to yield $\mathbf{x}^{k+1}$.

Note that the feasible set of [RMP] is the convex hull of the solution $\mathbf{x}^k$ and $\texttt{conv}\ (\hat{\mathcal{X}})$. If $\mathbf{x}^k \in \texttt{conv}\ (\hat{\mathcal{X}})$—which indeed is the case if no column has been dropped from $\hat{\mathcal{X}}$—then the column $\mathbf{x}^k$ is redundant; when column dropping is applied, it however becomes necessary to include $\mathbf{x}^k$ as a column in [RMP] in order to ensure convergence.

The simplicial decomposition principle is closely related to that of Holloway [504], who develops an extension of the Frank–Wolfe algorithm by means of an inner linearization (i.e., inner representation) of the feasible set, followed by restriction. The Holloway technique belongs to the family of inner linearization/restriction type algorithms defined by Geoffrion [418]. As applied to [P], simplicial decomposition is, in fact, an instance of the algorithm class of Holloway.[16]

Convergence rate results, and conditions under which convergence is finite, are scarce for general column generation methods. For the special case of simplicial decomposition, however, this is a well studied topic. Holloway shows that the convergence rate of the extended Frank–Wolfe algorithm is linear, and subsequently von Hohenbalken [953] shows that for the special case of simplicial decomposition convergence is finite. This result allows for the removal of columns which receive zero weights in the solution to a restricted master problem.

According to Carathéodory's Theorem, the maximal number of columns in [RMP] needed to express an optimal solution to [P] is $n + 1$ (or $\texttt{dim}\ (X) + 1$ from its sharpened version). This number is too large to be useful as a limit in practice, however, and the convergence result of von Hohenbalken enables the use of *column dropping* rules, by which columns with small weights are removed from [RMP]. This possibility is also important for another computational reason; in order to be able to apply efficient second-order methods to [RMP], the number of variables must be kept small.

There is, however, another side to this. Letting the maximum number of points retained in [RMP], $r \geq 1$ say, be smaller, implies the need to solve a larger number of master

[14] If X is unbounded, then extreme directions may be obtained from the solution of the Subproblems (4.11); these may be included in a restricted master problem with a feasible set of the form (2.14).

[15] The starting solution, $\mathbf{x}^0 = \mathbf{y}^0$, may be obtained from the solution to (4.11) given $\mathbf{x} = \mathbf{0}$.

[16] An interesting relationship exists between simplicial decomposition and Dantzig–Wolfe decomposition. As shown in [585], Dantzig–Wolfe decomposition for linear programs is the result of applying simplicial decomposition to a saddle point problem formulation of the linear program.

problems in order to obtain the right subset of columns and, in fact, the *finite* convergence result for simplicial decomposition is lost if the value of r is chosen too small. Indeed, letting $r = 1$ in (4.41) yields the Frank–Wolfe algorithm, which is only asymptotically convergent.

It is then interesting to be able to estimate the smallest possible value of r that implies finite convergence. Although it is problem dependent, it is possible to give it a precise value. The value sought is obviously equal to the number of extreme points that may be necessary to express an optimal solution to [P], and is obtained by applying Carathéodory's Theorem to the face of X of smallest dimension that includes the set of optimal solutions; this set is known as the *optimal face* of X (e.g., [987]),

$$X^* = \left\{ \mathbf{x} \in X \;\middle|\; \nabla T(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) = 0 \right\}, \qquad \mathbf{x}^* \in \Omega.$$

Hence, finite convergence is ensured under the condition that $r \geq \mathtt{dim}\,(X^*) + 1$ ([479]).

4.2.4 Discussion

The three algorithm concepts for the solution of traffic equilibrium problems presented in the above sections are sufficient to describe nearly every method that has been proposed in the literature. (For example, the classical Frank–Wolfe algorithm is obtained from the choice $\varphi^k \equiv 0$ for all k in the partial linearization algorithm. It is a decomposition algorithm since the direction-finding subproblems separate into $|\mathcal{C}|$ shortest route problems (although traditionally they are solved sequentially over origins). Finally, it may be viewed as a very special column generation algorithm, in which a very crude column dropping rule is applied.)

A very large class of algorithms is derived from the combination of the three algorithm concepts. It also includes several interesting algorithmic approaches that have not been presented previously, but which deserve further study. We provide such an example below, which is based on combining second-order partial linearization algorithms with simplicial decomposition.

It is well known that due to the linearity of the subproblem the Frank–Wolfe algorithm often exhibits a very slow convergence. Simplicial decomposition algorithms utilize Frank–Wolfe subproblems in the column generation phase, and therefore, to some extent, inherit the drawbacks of this method. [This has been observed in particular when column dropping rules are applied on problems with an optimal face of high dimension (see [481] for examples).] Consider replacing the Frank–Wolfe subproblem in the simplicial decomposition algorithm with a nonlinear partial linearization subproblem. (Such a method would extend both the partial linearization algorithm (in the sense that its line search step is replaced by a multidimensional search, whenever more than one column is retained in the restricted master problem) and the traditional simplicial decomposition algorithm (which is obtained from the choice $\varphi^k \equiv 0$ for all k).[17] The use of a nonlinear partial linearization subproblem in the column generation phase may undoubtedly enhance the performance of a simplicial decomposition scheme, since the column generating subproblem retains more information from the original problem than the Frank–Wolfe subproblem. (The increase in subproblem computations can be partly circumvented by choosing φ^k to exploit the problem structure, and by applying truncation strategies to $[\mathrm{P}^k_{\varphi^k}]$.) If this scheme is supplied with the proper column dropping rules, we obtain a

[17] The idea of using non-vertex points in restricted master problems of column generation methods is suggested in [550, 585].

partial linearization algorithm which utilizes restricted master problems only when it is necessary for making sufficient progress. To illustrate this feature, we consider using a Newton subproblem, i.e., let $\varphi^k(\mathbf{x}) = (1/2)\mathbf{x}^{\mathrm{T}}\nabla T(\mathbf{x}^k)\mathbf{x}$ for all k. It is well know that locally, around a solution, Newton's method converges when using unit step lengths. Using Newton-based column generating subproblems within the simplicial decomposition scheme, the sequence of restricted master problems would thus increase in dimension in the first few iterations, but as the solution is approached, unit steps begin to be taken towards the latest subproblem solution, and the column dropping rules would thus eventually reduce the simplex to a line segment, i.e., to the original Newton algorithm. This scheme can be viewed as a *dynamic* partial linearization algorithm, where master problems are used only when necessary. The theoretical and practical properties of this scheme are subjects of ongoing research efforts ([585]).

The concepts of partial linearization, decomposition, and column generation, will be used in the subsequent description of algorithmic approaches for the solution of the traffic assignment problem [TAP] and its extensions.

4.2.5 A taxonomy of algorithms for [TAP]

In order to be able to distinguish between and interrelate algorithms easily, we introduce a taxonomy for describing an algorithm within the framework.

We first introduce a notation for the problem being solved; we only need to distinguish between the two possible representations of the feasible set of [TAP]. We use the notation F^r to denote the link-route representation, i.e., the set of feasible link flows defined by the System (2.6b)–(2.6d), and the notation F^n for the link-node representation, i.e., the feasible set defined by (2.13).

To describe a *decomposition* method for [TAP], we need to distinguish **(a)** between a sequential and a parallel decomposition scheme, and **(b)** between a decomposition over either origins or O-D pairs. To this end, we introduce the notation D^S (sequential) and D^P (parallel), and $D_{\mathcal{O}}$ (origin-based) and $D_{\mathcal{C}}$ (O-D-based). (A sequential decomposition scheme over origins, for example, is denoted by $D^S_{\mathcal{O}}$.)

Column generation schemes proposed for [TAP] are invariably based on columns defined by various degrees of aggregations of shortest route flows (and are hence simplicial decomposition algorithms). To denote the level of aggregation in a column generation scheme, we shall use the notation $C_{\mathcal{R}}$, $C_{\mathcal{O}}$, and $C_{\mathcal{A}}$ for, respectively, columns defined by individual route flows, flows from separate origins, and link flows. (The number of convexity constraints present in a restricted master problem is, respectively, $|\mathcal{C}|$, $|\mathcal{O}|$, and one.)

To describe the hierarchy in which these two concepts are applied, we shall write them within brackets, according to the following rule: if an instance A of an algorithm concept is embedded in an instance B of another algorithm concept, then it is denoted by $B[A]$. To give an example, suppose that an algorithm for [TAP] is a column generation scheme over individual route flows, where the restricted master problem is solved by a parallel decomposition scheme over origins. In the taxonomy introduced, this algorithm would be described by

$$F^r(C_{\mathcal{R}}[D^P_{\mathcal{O}}]).$$

A *partial linearization* algorithm may be applied to a variety of representations of [TAP] (e.g., directly to the link-node formulation, to the solution of a projection of [TAP] to a non-basic variable space, or to a Lagrangean dual formulation, see Section 4.2.1)

and to its restrictions (e.g., to a restricted master problem, or to single-commodity flow problems within a decomposition scheme), as well as in a column generating subproblem phase of a column generation scheme (although this alternative is not well studied, since column generations are always made using Frank–Wolfe subproblems). We will denote a partial linearization algorithm by the corresponding choice of sequence $\{\varphi^k\}$, since it may (in essence) be identified by the choice of subproblem.

4.3 Algorithms for the basic model

In this section we shall provide a review of the methods proposed for the solution of [TAP], following the development of the Frank–Wolfe algorithm. In order to place each algorithm within the framework of the combination of the three algorithm concepts, we shall use the taxonomy introduced above.

We begin by studying decomposition methods.

4.3.1 Decomposition algorithms

Historically, sequential decomposition algorithms of the Gauss–Seidel type were the first alternatives to the Frank–Wolfe algorithm to be developed; a main reason for this development is that efficient codes for single-commodity network flows were available. Given a feasible flow, $\mathbf{f}^k$, in iteration k a subproblem of the form

$[\mathrm{TAP}^k_{i_k}]$

$$\min\ T(\mathbf{f}^k_{i_k-}, \mathbf{f}_{i_k}, \mathbf{f}^k_{i_k+}) = \sum_{a\in\mathcal{A}} \int_0^{f_{ai_k}+\sum_{i\neq i_k} f^k_{ai}} t_a(s)ds, \tag{4.42a}$$

subject to

$$\mathbf{A}\mathbf{f}_{i_k} = \mathbf{d}_{i_k}, \tag{4.42b}$$

$$\mathbf{f}_{i_k} \geq \mathbf{0} \tag{4.42c}$$

is solved. (Here, i_k denotes one O-D pair or a larger subset of the O-D pairs, such as an origin; the node-link formulation of [TAP] is used without any loss of generality.) In the sequential approach, the solution to $[\mathrm{TAP}^k_{i_k}]$ defines $\mathbf{f}^{k+1}_{i_k}$.

If the flows in all commodities but the commodity i_k are held fixed at their current values, then the corresponding restriction of T, $\mathbf{f}_{i_k} \mapsto T(\mathbf{f}^k_{i_k-}, \mathbf{f}_{i_k}, \mathbf{f}^k_{i_k+})$, is strictly convex, and the single-commodity flow problem $[\mathrm{TAP}^k_{i_k}]$ is uniquely solvable. Note, however, that the equilibrium solution is not unique in the commodity flows (see the discussion following Theorem 2.5, and Theorem 2.7); a commodity may therefore receive flows that oscillate wildly in the sequence $\{\mathbf{f}^k\}$ (e.g., [424]), and the flows obtained in the limit of this sequence depend on the ordering of the commodities made in the sequence $\{i_k\}$.

The cyclic decomposition (or block Gauss–Seidel) version, where i_k is chosen according to the cyclic Rule (4.28),[18] is convergent, in the sense that every accumulation point of the sequence $\{\mathbf{f}^k\}$ is optimal in [TAP] (e.g., [34, Th. VI.1.3] and [84, Prop. 3.3.9];[19] a

[18] Viewing the traffic assignment problem as a non-cooperative game among the commodities (see Section 2.6.1), this algorithm may be interpreted as a process in which each player, in turn, chooses an optimal strategy based on the other players' current choices of strategy; the process terminates when no player can improve his/her conditions, and this then defines an equilibrium state.

[19] In the result of [824], additional monotonicity assumptions on the link performance functions are made.

parallel (or block Jacobi) version requires an additional global line search step to yield convergence without additional assumptions on the model (see Section 4.2.2).

The number of iterations needed to obtain a desired accuracy is expected to be lower for the sequential approach, since the information obtained from the solution to [TAP$_i$] is utilized when solving [TAP$_{i+1}$]. In the context of traffic assignment, this may be explained by the fact that a decomposition results in the problems [TAP$_i$] neglecting the interactions between the flows of different commodities, and a faster transfer of information about adjusted commodity flows may reduce the potentially detrimental effects of these neglected interactions. In a parallel computing environment, the Jacobi-type approach could be more efficient, however, since more iterations are carried out during a given time period.

We first study the development of sequential decomposition algorithms, and concentrate on the different approaches presented for the iterative solution of the single-commodity problems [TAP$^k_{i_k}$]. For the simplicity of presentation, we shall drop the iteration counter, and refer to [TAP$_i$].

4.3.2 Sequential decomposition algorithms

Equilibration operator type approaches

The first sequential decomposition algorithms were independently developed in 1968 by Dafermos [210, 209, 205, 206] and Bruynooghe *et al.* [132]. The two methods are essentially equivalent, although they are presented for different representations of [TAP].

Dafermos' algorithm ($F^r(D^S_{\mathcal{C}})$) is based on determining the most expensive (simple) route used and the least expensive (simple) route in an O-D pair, and a following transfer of flow between the two routes towards the least expensive one using a simple line search. This flow transfer, which is termed the *equilibration operator*, results either in the two routes receiving flows with equal costs, or the most expensive route receiving a zero flow. The next O-D pair is considered when all routes are equilibrated, i.e., when [TAP$^k_{i_k}$] is solved. It is to be noted that the method is impractical for the solution of large-scale problems, since all the routes in the network must be enumerated. In [210, 209] quadratic networks are considered, while in [209, 205, 206] more general costs are treated, and in [701] the method is extended to elastic demand problems.

The method of Bruynooghe *et al.* ($F^n(D^S_{\mathcal{O}})$) is the same, with the exceptions that it is based on the link-node formulation and a definition of a commodity as an origin. The flow transfers thus correspond to the transfer of flows from the most expensive tree of used routes towards the least expensive tree. This method is also impractical, since the calculation of the most expensive tree of (simple) routes is an *NP*-complete problem ([411, p. 213]), due to the presence of cycles. Both methods may, however, be applicable to large-scale problems, when embedded in a column generation scheme.

From the experiments conducted with this algorithm, two important observations were made ([424]): firstly, the number of equilibrium routes is very limited compared to the total number of routes in the network; secondly, these routes are identified early by the algorithm. Gibert [424] proposes therefore storing the shortest routes obtained, and applying the algorithm of Bruynooghe *et al.* to this subset. One advantage is immediate: the determination of both the shortest and the longest route in each O-D pair becomes a very simple comparison operation. Gibert's algorithm includes a rule for dropping routes with zero flows, a truncation strategy for the solution of the restriction of [TAP$^k_{i_k}$], and a termination criterion of the form (4.9). Global convergence is established for two different versions of the algorithm; the basic version ($F^r(D^S_{\mathcal{O}}[C_{\mathcal{R}}])$) is a direct extension

of the algorithm of Bruynooghe *et al.*, where routes are generated within the solution of $[\mathrm{TAP}^k_{i_k}]$, while in the modified algorithm the flow updates are made simultaneously over all the origins. (Gibert's algorithm is the first convergent column generation algorithm for traffic assignment presented; the column generation algorithm given by Martin *et al.* [650] (see Section 1.5.4) is a heuristic.) The algorithm is extended to elastic demand problems in [423]. An algorithm essentially the same as the basic one above was proposed much later by Schittenhelm [813], who also extends the algorithm to the solution of a combined trip distribution and assignment problem, and by Lee [608].

The method of Leventhal *et al.* [611]—sometimes acknowledged as the first column generation method for the nonlinear traffic assignment problem—is similar to the one proposed earlier by Gibert. The basic algorithm is of the form $F^r(C_{\mathcal{R}}[D^S_{\mathcal{C}}])$; they choose to augment the subset $\cup_{(p,q)\in\mathcal{C}}\hat{\mathcal{R}}_{pq}$ of the routes by one new route only, and propose to use the equilibration operator approach of Dafermos and Sparrow [209] on the single-commodity networks, modified to equilibrate only two routes per O-D pair and iteration. Finite convergence is established for the column generation method, including the possibility of dropping routes with zero flows. For quadratic networks, they also propose a finitely convergent quadratic programming method ([225, 926]). Experiments on small quadratic networks with these algorithms indicate the drawback of the original algorithm of Dafermos and Sparrow, due to the need to enumerate all the routes.

In the algorithms described so far, relatively few links (in the most disaggregated formulation only the links defining two routes) receive an update of their flows in each iteration, and one should expect that a large number of iterations is needed to obtain an accurate solution. One is therefore led to consider algorithms, where more than two routes are involved simultaneously in the flow updates.

Reduced gradient type approaches

To this end, we first consider extensions of the simplex method to nonlinear network flows, and investigate their relationships to the above type of methods.

In the *convex simplex* algorithm ([1002]), variables are partitioned into basic and non-basic variables, as in the simplex method for linear programs, and in an iteration, only one non-basic variable may change its value. Main differences to the case of linear programs are that non-basic variables may assume non-zero values and that a basis change is performed only if the resulting change forces a basic variable to a bound.

Introducing a partitioning of $\mathbf{f}_i^{\mathrm{T}}$ into $(\mathbf{f}_{Bi}^{\mathrm{T}}, \mathbf{f}_{Ni}^{\mathrm{T}})$, where B and N denote the basic and non-basic links, respectively, and, correspondingly, of the node-link incidence matrix $\mathbf{A}$ into $(\mathbf{B}, \mathbf{N})$ [$\mathbf{B}$ is assumed non-singular],[20] suppressing the index k, and defining $g(\mathbf{f}_i) = T(\mathbf{f}^k_{i_k-}, \mathbf{f}_{i_k}, \mathbf{f}^k_{i_k+})$, the single-commodity problem may be written in terms of non-basic variables in the form

$$\min\ g(\mathbf{B}^{-1}\mathbf{d}_i - \mathbf{B}^{-1}\mathbf{N}\mathbf{f}_{Ni}, \mathbf{f}_{Ni}), \tag{4.43a}$$

subject to

$$\mathbf{B}^{-1}\mathbf{d}_i - \mathbf{B}^{-1}\mathbf{N}\mathbf{f}_{Ni} \geq \mathbf{0}, \tag{4.43b}$$

$$\mathbf{f}_{Ni} \geq \mathbf{0}. \tag{4.43c}$$

(Normally, the components $\mathbf{f}_{Ni}$ are assumed to be the variables with the largest values.)

[20] One equation in the Constraints (4.42b) is deleted to ensure full row rank.

The basis $\mathbf{f}_{Bi}$ defines a rooted spanning tree (if the index i corresponds to an origin), or a simple route (if i denotes an O-D pair).

The *reduced gradient* (the gradient of g in the space of non-basic variables) is

$$\mathbf{r}_{Ni}^{\mathrm{T}} = \nabla_N g(\mathbf{f}_i)^{\mathrm{T}} - \nabla_B g(\mathbf{f}_i)^{\mathrm{T}} \mathbf{B}^{-1}\mathbf{N}. \tag{4.44}$$

If $\mathbf{r}_{Ni} = \mathbf{0}$, then $\mathbf{f}_i$ solves [TAP$_i$]; otherwise, a direction $\mathbf{p}_{Ni}$ defines a feasible descent direction with respect to g if $p_{ai} \geq 0$ when $f_{ai} = 0$, $a \in N$, and $\mathbf{r}_{Ni}^{\mathrm{T}}\mathbf{p}_{Ni} < 0$. (Note that in the space of basic variables the direction becomes $-\mathbf{B}^{-1}\mathbf{N}\mathbf{p}_{Ni}$.) In particular, if r_{ai}, $a \in N$, is negative (positive), then $\mathbf{e}_a$ ($-\mathbf{e}_a$) is an improving direction. The determination of the largest value of $|r_{ai}|$ over the non-basic variables corresponds to finding the minimum reduced cost of cycles that are formed by the tree (or route) and a non-basic link. A line search then determines the amount of flow to send in the cycle formed.[21]

The convex simplex method is specialized and applied to nonlinear networks in [801, 163, 178, 486, 548], and applied to [TAP$_i$] in a decomposition method of the form $F^n(D_{\mathcal{O}}^S)$ by Nguyen [713, 714, 715, 717], who also extends it to elastic demands, in [715, 717]. Similar algorithms are discussed in [226, 742]. Nguyen's [714] algorithm is applied to a network model of the city of Winnipeg in [360, 361], with encouraging results; the algorithm compares favourably with the Frank–Wolfe algorithm with respect to its convergence rate, but is considered equal when including memory costs.

To make the connection with the equilibration operator approach clear, let us apply the convex simplex algorithm to the link-route formulation of [TAP$_i$], where a subset $\hat{\mathcal{R}}_i \subseteq \mathcal{R}_i$ of the total set of routes in O-D pair $i \in \mathcal{C}$ is known. This restriction is given by [cf. (2.6), where O-D pairs are denoted by (p, q)]

[TAP$_i$]

$$\min\ g(\mathbf{f}_i), \tag{4.45a}$$

subject to

$$\sum_{r \in \hat{\mathcal{R}}_i} h_{ir} = d_i, \tag{4.45b}$$

$$h_{ir} \geq 0, \qquad \forall r \in \hat{\mathcal{R}}_i, \tag{4.45c}$$

$$\sum_{r \in \hat{\mathcal{R}}_i} \delta_{ira} h_{ir} = f_{ai}, \qquad \forall a \in \mathcal{A}, \tag{4.45d}$$

where we use the definition of the objective as in (4.43).

We first observe from the Constraints (4.45b)–(4.45c) that a basis is defined by one route variable, h_{ir_B} say. The transformation corresponding to the one made above leads to the equivalent problem

$$\min\ g(d_i - \sum_{r \neq r_B} h_{ir}, \mathbf{h}_{ir_N}), \tag{4.46a}$$

subject to

$$d_i - \sum_{r \neq r_B} h_{ir} \geq 0, \tag{4.46b}$$

$$\mathbf{h}_{ir_N} \geq \mathbf{0}; \tag{4.46c}$$

[21] It is to be noted that an iteration in *any* feasible-direction method for [TAP] corresponds to determining and sending flows in cycles of the networks, since

$$\mathbf{f}^k \in F^n,\ \mathbf{f}^{k+1} = \mathbf{f}^k + l_k \mathbf{p}^k \in F^n \Longrightarrow \mathbf{A}\mathbf{p}^k = \mathbf{0}.$$

Explicit determinations of cycles akin to Zoutendijk methods, are proposed for [TAP$_i$] in [964].

the reduced gradient equals

$$\mathbf{r}_{ir_N} = \nabla_N g(\mathbf{h}_i) - \nabla_B g(\mathbf{h}_i)\mathbf{1} \tag{4.47a}$$
$$= \mathbf{c}_{ir_N}(\mathbf{h}_i) - c_{ir_B}(\mathbf{h}_i)\mathbf{1}, \tag{4.47b}$$

where $\mathbf{1}$ is a $(|\hat{\mathcal{R}}_i| - 1)$-dimensional vector of ones, i.e., the reduced gradient equals the difference in travel cost between the non-basic routes and the basic one.

Now, assume that given the current flows the basic route chosen is the cheapest in $\hat{\mathcal{R}}_i$. Clearly, $\mathbf{r}_{ir_N} \geq \mathbf{0}$, and its maximal element is defined by the most expensive route. The direction defined by the convex simplex algorithm then is $p_{ir_B} = +1$, $p_{ir_p} = -1$, and $p_{ir} = 0$, for all $r \neq r_B, r_p$, where r_p is the most expensive route among those with a positive flow, i.e., the transfer of flow from the most expensive route used to the least expensive route. This is the search direction of the equilibration operator approach, which can therefore be interpreted as a convex simplex method, where a basis is defined by a cheapest route. Note that the shortest route can be a different one in the next iteration; this corresponds to a basis change.

In *reduced gradient* methods ([986, 317, 698, 623, 43]) more than one non-basic variable is allowed to change in each iteration, and this allows more rapid changes both in the variables and in the basis. (The convex simplex algorithm is therefore a special case of the reduced gradient algorithm.) As in the convex simplex algorithm, the search direction is based on the negative of the reduced Gradient (4.44); in order to ensure the feasibility of the direction, it is modified to

$$p_{ai} = \begin{cases} -r_{ai}, & \text{if } r_{ai} < 0 \text{ or } f_{ai} > 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall a \in N. \tag{4.48}$$

The search Direction (4.48) is the steepest descent direction in the space of non-basic variables, projected onto the nonnegativity constraints; the search direction of the convex simplex algorithm can analogously be viewed as a steepest coordinate descent direction (e.g., [623, p. 359]). In order to establish global convergence of the reduced gradient algorithm, the search direction must be modified to yield a closed algorithmic map. The modification consists of a deflection from the nonnegativity constraints according to

$$p_{ai} = \begin{cases} -r_{ai}, & \text{if } r_{ai} \leq 0, \\ -f_{ai} r_{ai}, & \text{if } r_{ai} > 0, \end{cases} \qquad \forall a \in N. \tag{4.49}$$

The reduced gradient algorithm is specialized to nonlinear networks in [237, 46]. Applications to [TAP$_i$] are considered by Nguyen [713, 715, 717] and Florian [352]. As already mentioned, using the least-cost route to define the basis for the Constraints (4.45b)–(4.45c), the components of the reduced gradient defined by (4.47) are the differences in route travel costs between the non-basic routes and the basic one. The search direction defined by (4.49) becomes

$$p_{ir} = h_{ir}(c_{ir_B} - c_{ir}), \qquad \forall r \in N, \tag{4.50}$$

and $p_{ir_B} = -\sum_{r \in N} p_{ir}$.

Remark 4.2 The Expression (4.50) states that the *swapping rate* p_{ir} from route $r \in N$ to route r_B is proportional to the cost and to the flow of route r. This is reasonable from a behavioural point of view: the more expensive a route is and the more drivers there are on that route, the more inclined a driver is to swap to the less expensive route. It also follows naturally from the Wardrop conditions, which may equivalently be formulated as

$$h_{ir}[c_{ir}(\mathbf{h}) - c_{is}(\mathbf{h})]_+ = 0, \qquad \forall r, s \in \mathcal{R}_i, \ \forall i \in \mathcal{C}. \tag{4.51}$$

To solve the dynamic system associated with (4.51), Smith [848] (see also [63, Chap. III]) suggests adjusting a non-equilibrium flow using the swapping rate $h_{ir}[c_{ir}(\mathbf{h}) - c_{is}(\mathbf{h})]_+$ from route r to s; the swapping rate defined by (4.50) is indeed of this form.

The advantage over the convex simplex approach is immediate: the flows on all the routes are simultaneously adjusted towards an equilibrium. The swapping rate is also very reasonable. (It is important to note that the swapping rate defined by the search direction of the Frank–Wolfe algorithm is proportional neither to the cost nor to the flow on the routes.) This algorithm is proposed for the solution of $[\mathrm{TAP}_{i_k}^k]$ by Nguyen, in decomposition algorithms of the form $F^r(C_{\mathcal{R}}[D_{\mathcal{C}}^S])$ ([713]) and $F^r(D_{\mathcal{C}}^S[C_{\mathcal{R}}])$ ([359, 717]); Florian's [352] method is the same as that given in [359, 717], with the exception that the basic variable is defined by the route with the largest flow. Extensions to elastic demands are discussed in [713, 717].

Reduced gradient algorithms have also been developed from intuitive ideas such as those discussed in the above remark. Van Vliet (see [942, 957]) describes the cost differences $c_{ir} - c_{ir_B}$ as a *social pressure* on the route flows that forces the routes towards an equilibrium; this interpretation naturally leads to an adjustment process based on the search Direction (4.50). The basic algorithm—as described in [957]—is of the form $F^r(D_{\mathcal{C}}^P[C_{\mathcal{R}}])$. In order to enhance the practical efficiency of the algorithm, Van Vliet also considers different normalizations of the social pressure. These normalizations actually define special choices of scaled (or deflected) reduced gradients (e.g., [698]), i.e., the search directions are defined by a premultiplication of the reduced gradient by a positive definite square matrix. With the choice of a diagonal scaling matrix, $\mathbf{S} = \mathtt{diag}\,(s_{ir})$, where $s_{ir} > 0$ for all $r \in N$, and with r_B being the basic route, the search direction becomes

$$p_{ir} = \begin{cases} s_{ir}(c_{ir_B}(\mathbf{h}_i) - c_{ir}(\mathbf{h}_i)), & \text{if } c_{ir}(\mathbf{h}_i) \le c_{ir_B}(\mathbf{h}_i), \\ s_{ir}h_{ir}(c_{ir_B}(\mathbf{h}_i) - c_{ir}(\mathbf{h}_i)), & \text{if } c_{ir}(\mathbf{h}_i) > c_{ir_B}(\mathbf{h}_i), \end{cases} \qquad \forall r \in N. \tag{4.52}$$

Gradient projection type approaches

Consider the problem $[\mathrm{TAP}_i]$ of the form (4.45). Applying the deflected gradient projection algorithm amounts to finding, given $\mathbf{h}_i^k \in H_i$,

$$\mathbf{y}_i^k = P_{H_i}^{\mathbf{B}_{ik}}(\mathbf{h}_i^k - \gamma_k \mathbf{B}_{ik}^{-1} \nabla g(\mathbf{h}_i^k)), \tag{4.53}$$

defining the search direction $\mathbf{p}_i^k = \mathbf{y}_i^k - \mathbf{h}_i^k$ [cf. (4.25)]. If $\mathbf{B}_{ik}$ is a diagonal matrix, then the projection can be performed in linear time by dualizing the Constraint (4.45b) and performing a line search with respect to its Lagrange multiplier (e.g., [129]). (Note that it is much easier to perform projections onto the feasible set of the link-route formulation of $[\mathrm{TAP}_i]$ than onto that of the link-node formulation, since the former is defined by a simplex.)

Let $\mathbf{B}_{ik} = \mathtt{diag}\,(b_{ir}^k)$, where $b_{ir}^k > 0$ for all $r \in \hat{\mathcal{R}}_i$. Then, from (4.24), the Projection (4.53) is equivalent to minimizing

$$\sum_{r \in \hat{\mathcal{R}}_i} \left(c_{ir}(\mathbf{h}_i^k)(h_{ir} - h_{ir}^k) + \frac{b_{ir}^k}{2\gamma_k}(h_{ir} - h_{ir}^k)^2 \right) \tag{4.54}$$

over H_i. The convergence of the sequence $\{\mathbf{h}_i^k\}$ to the solution of $[\mathrm{TAP}_i]$ requires the sequence $\{b_{ir}^k/(2\gamma_k)\}$ to be bounded away from zero and infinity. (When a line search is

performed in the direction $\mathbf{p}_i^k = \mathbf{y}_i^k - \mathbf{h}_i^k$ this condition suffices, but a stronger assumption on the lower bound on $b_{ir}^k/(2\gamma_k)$ is required when unit steps ($\mathbf{h}_i^{k+1} = \mathbf{y}_i^k$) are used [cf. Section 4.2.1].) An interesting choice of b_{ir}^k is

$$\frac{\partial^2 g(\mathbf{h}_i^k)}{\partial h_{ir}^2} = \sum_{a \in \mathcal{A}} \delta_{ira} t_a'(f_a(\mathbf{h}_i^k)),$$

i.e., $\mathbf{B}_{ik}$ is a diagonal approximation of the Hessian of g at $\mathbf{h}_i^k$.

Bertsekas [69] proposes approximate Newton methods of this type to optimal routing problems in computer communication networks (see Section 2.6.5). Bertsekas argues that with the above choice of the scaling matrix, the values of the parameters γ_k can preferably be chosen around unity, and he gives a method of the form $F^r(D_{\mathcal{C}}^S[C_{\mathcal{R}}])$. He also discusses the possibility of updating several O-D pairs simultaneously; see Section 4.3.3. (See also [403], and [68, 81, 399, 80], which also include discussions on the choice of the parameter γ_k based on line search strategies such as those in [67].) This algorithm is extended to asymmetric assignment models ([78, 594]); see Section 5.3.5.

This algorithm is (obviously) highly related to scaled reduced gradient algorithms. Let r_B be a basic route, and consider the transformed Problem (4.46). Let $\gamma_k > 0$ and $\mathbf{B}_{ik}$ be a diagonal matrix in the non-basic variables. Then (4.53) reduces to

$$y_{ir}^k = \max\{0, h_{ir}^k - \frac{\gamma_k}{b_{ir}^k}(c_{ir}(\mathbf{h}_i^k) - c_{ir_b}(\mathbf{h}_i^k))\}, \qquad \forall r \in N. \tag{4.55}$$

A diagonal approximation of a reduced Newton method is obtained from choosing

$$b_{ir}^k = \sum_{a \in \mathcal{A}} (\delta_{ira} - \delta_{ir_B a})^2 t_a'(f_a(\mathbf{h}_i^k)).$$

Observe that if a route has a higher cost than the basic one, then its flow is reduced, proportionally to the cost difference, and that if such a route has a zero flow, then its flow will stay at zero. (For this reason, these routes need not be included in the updating step.) Note, however, that the swapping rate is not proportional to the flow, as is normal for reduced gradient methods. The result is that in practice, the algorithm may more quickly identify the routes that should receive a zero flow, especially if the basic route is chosen to be the cheapest one. (Using line search strategies for choosing γ_k, the line search can actually continue beyond points at which some route receives a zero flow.)

Bertsekas [71, 82, 65] proposes this algorithm for a restriction of the optimal routing problem in a decomposition scheme of the form $F^r(D_{\mathcal{C}}^S[C_{\mathcal{R}}])$.

Although the algorithms discussed above are much more efficient than the Frank–Wolfe algorithm, their convergence rates are still only linear; in order to obtain rapid convergence near the optimal solution, it is necessary to include off-diagonal terms of the Hessian matrix to take into account the interactions among the routes within each commodity and among the different commodities, and to introduce some form of line search procedure. We are hence led to consider Newton type approaches.

Newton type approaches

Bertsekas and Gafni [79] present a superlinearly convergent projected Newton ([67, 70, 72]) algorithm for the solution of the link-route formulation of [TAP] embedded in a column generation scheme. They propose solving each quadratic subproblem approximately with a conjugate gradient algorithm, and show that computations can be performed directly on the network without the need to store Hessian information explicitly; this is a crucial

property for its applicability to large-scale problems. An application to a small example is presented in [400].

To illustrate the possibility of improving the convergence rate of an algorithm by embedding it in an algorithm with a higher convergence rate with only minor modifications to the existing algorithm and increase in storage requirements, Dembo and Tulowitzki [241] apply a truncated Newton algorithm to the link-node formulation of [TAP], in which the quadratic subproblems are solved approximately using the Frank–Wolfe algorithm or its PARTAN modification. (This is a special case of a truncated partial linearization algorithm.) In the link-node formulation, the Hessian matrix is diagonal and easily computed (provided of course that the link travel cost functions are differentiable). Applications to the Hull and Winnipeg networks demonstrate a speedup of around 60 % or more compared to the original algorithm; the best results were obtained when at most four iterations of the Frank–Wolfe algorithm were applied on each quadratic subproblem. (Note that if only one iteration of the Frank–Wolfe algorithm is used in each main iteration, then the overall algorithm reduces to the Frank–Wolfe algorithm.) In order to obtain better asymptotic behaviour, Dembo and Tulowitzki suggest replacing the Frank–Wolfe algorithm with one more rapidly convergent.

Newton type methods have also been applied to link-node formulations of single-commodity problems; such algorithms can be applied to solve each single-commodity problem in a decomposition scheme over commodities. Dembo and Klincewicz [237] present a scaled reduced gradient approach, while Klincewicz [558, 559] (see also [545]) presents an exact Newton algorithm, in which conjugate gradient techniques are used in the solution of the quadratic subproblems. The algorithm exhibits a quadratic convergence rate, and (similar to the algorithms of Bertsekas and Gafni [79, 400], applied to link-route formulations) the computations may be performed using graph operations. Truncated Newton methods are discussed in [235, 303, 236, 1011].

One common property of the methods for the solution of [TAP$_i$] discussed so far is their *primal* nature, that is, they are based on generating a sequence of primal feasible solutions and improving search directions. This property, together with the fact that the problem data for a certain commodity varies only slightly from one iteration to the next (and then only in the objective function), facilitates and motivates the use of truncated primal algorithms for each single-commodity problem, reoptimized from the solution to the previous problem. When a single-commodity problem solution is terminated, a primal feasible and near-optimal solution is then at hand.

In contrast, a *dual* algorithm generates a sequence of primal infeasible solutions, and a primal feasible (and simultaneously optimal) solution is obtained only in the limit of this sequence. On the other hand, dual algorithms are very easy to implement and use, and they are well suited for parallel and distributed computations. In the next section, we shall outline the basic algorithm approaches for the solution of [TAP$_i$] using dual techniques; we shall also discuss the possibility of finitely generating primal feasible solutions within a dual algorithm, for use in criteria for finite termination.

Dual approaches

We consider the single-commodity Problem (4.42), where for simplicity of presentation we drop both the iteration counter and the commodity index; further we let $g(\mathbf{f})$ denote the restriction of T to the commodity in question, and thus arrive at the problem (cf.

Section 2.2.2)

$$\min\ g(\mathbf{f}) = \sum_{(i,j)\in\mathcal{A}} g_{ij}(f_{ij}), \tag{4.56a}$$

subject to

$$\sum_{j\in\mathcal{W}_i} f_{ij} - \sum_{j\in\mathcal{V}_i} f_{ji} = d_i, \qquad \forall i \in \mathcal{N}, \tag{4.56b}$$

$$f_{ij} \geq 0, \qquad \forall (i,j) \in \mathcal{A}. \tag{4.56c}$$

Let π_i be the multiplier for the flow conservation Constraint (4.56b) corresponding to node i, the Lagrange function be defined by

$$L(\mathbf{f},\boldsymbol{\pi}) \stackrel{\text{def}}{=} g(\mathbf{f}) + \boldsymbol{\pi}^{\mathrm{T}}(\mathbf{A}\mathbf{f} - \mathbf{d}) = g(\mathbf{f}) + \sum_{i\in\mathcal{N}} \pi_i \left(\sum_{j\in\mathcal{W}_i} f_{ij} - \sum_{j\in\mathcal{V}_i} f_{ji} - d_i \right), \tag{4.57}$$

and the dual problem by

$$\max_{\pi\in\Re^{|\mathcal{N}|}}\ \theta(\boldsymbol{\pi}), \tag{4.58a}$$

where

$$\theta(\boldsymbol{\pi}) = \min_{\mathbf{f}\geq\mathbf{0}}\ L(\mathbf{f},\boldsymbol{\pi}) \tag{4.58b}$$

$$= -\mathbf{d}^{\mathrm{T}}\boldsymbol{\pi} + \sum_{(i,j)\in\mathcal{A}} \min_{f_{ij}\geq 0} \{g_{ij}(f_{ij}) + \pi_i - \pi_j\}. \tag{4.58c}$$

We assume throughout this section that each function g_{ij} and g'_{ij} is coercive on $\Re_+$ [cf. (2.29)]; this ensures the existence of a (unique) solution to (4.58c) for any values of $\boldsymbol{\pi}$, and ultimately that the dual Problem (4.58) has an optimal solution.

The solution to the Subproblem (4.58c) is denoted by $\mathbf{f}(\boldsymbol{\pi})$; the solution to each strictly convex single-link problem is given by

$$f_{ij}(\boldsymbol{\pi}) = f_{ij}(\pi_j - \pi_i) = \max\ \{0, g'_{ij}(\pi_j - \pi_i)^{-1}\}, \qquad \forall (i,j) \in \mathcal{A},$$

where $g'_{ij}(\cdot)^{-1}$ denotes the inverse function of the derivative of g_{ij}. Some properties of the dual problem are given below. (For proofs, we refer to [785, 83, 84].)

Theorem 4.3 (Properties of θ) *The dual objective θ is finite, continuous, concave and differentiable, with*

$$\nabla\theta(\boldsymbol{\pi}) = \mathbf{A}\mathbf{f}(\boldsymbol{\pi}) - \mathbf{d}. \tag{4.59}$$

Further, $\theta(\boldsymbol{\pi}) \leq g(\mathbf{f}^*)$, for all $\boldsymbol{\pi} \in \Re^{|\mathcal{N}|}$.

Theorem 4.4 (Relationships between (4.56) and its dual) *Strong duality holds, i.e., for any dual optimal solution $\boldsymbol{\pi}^*$, $\theta(\boldsymbol{\pi}^*) = g(\mathbf{f}^*)$. Furthermore, $\mathbf{f}^* = \mathbf{f}(\boldsymbol{\pi}^*)$.*

The dual optimal solution is not unique, due to the linear dependence of the Constraints (4.56b); the potentials $\pi^*_j - \pi^*_i$, $(i,j) \in \mathcal{A}$, however, are unique.

The dual problem may be solved by any technique for unconstrained, differentiable convex programs, but its special structure motivates the use of network-based methods.

The set of optimal solutions to (4.58) is the set of solutions to the system of nonlinear equations

$$\nabla\theta(\boldsymbol{\pi}) = \mathbf{0}, \tag{4.60}$$

which, from (4.59), is equivalent to determining node prices such that the *divergence* (or *imbalance*) of each node,

$$\delta_i(\boldsymbol{\pi}) \stackrel{\text{def}}{=} \frac{\partial\theta(\boldsymbol{\pi})}{\partial\pi_i} = \sum_{j\in\mathcal{W}_i} f_{ij}(\pi_j - \pi_i) - \sum_{j\in\mathcal{V}_i} f_{ji}(\pi_i - \pi_j) - d_i,$$

is zero. A natural approach then is to iteratively choose an unbalanced node, and balance it by adjusting its node price. In iteration k, let i_k be the node chosen. Assume without any loss of generality that $\delta_{i_k}(\boldsymbol{\pi}^k) > 0$. The function δ_{i_k} is non-decreasing in π_{i_k}; let $\pi_i^{k+1} = \pi_i^k$ for all $i \neq i_k$, and $\pi_{i_k}^{k+1} = \pi_{i_k}^k - l_k$ for some $l_k > 0$ such that $\delta_{i_k}(\boldsymbol{\pi}^{k+1}) \approx 0$. The solution of this equation amounts to adjusting the flows of the links initiated or terminating at node i_k, so that the node receives a (approximately) balanced flow. [The proper value of l_k is found for instance by defining an interval of l in which δ_{i_k} changes sign, and performing a (inexact) line search within this interval (e.g., [1009, 584]).] A search for a solution to $\delta_{i_k}(\boldsymbol{\pi}^{k+1}) = 0$ by an adjustment of π_{i_k} is equivalent to performing a line search along the coordinate π_{i_k} with respect to θ, and therefore this rather intuitive approach is a Gauss–Seidel method for the solution of the dual problem. Since it is unconstrained, the node prices may also be updated simultaneously, thus defining a Jacobi algorithm.

This simple scheme has a very long history in the solution of nonlinear flow problems, such as trip distribution (see Section 1.5.2), constrained matrix problems (e.g., [155, 817, 702]), and single-commodity flows (e.g., [1009, 77, 83, 1010]). (Surveys of such methods are given in [735, 156, 614, 84, 909, 1007].) Under standard assumptions, the sequence $\{\mathbf{f}(\boldsymbol{\pi}^k)\}$ is known to converge linearly, provided that the indices are chosen according to either the essentially cyclic or the Gauss–Southwell node ordering ([627]). Under additional assumptions on the dual optimal solution, the dual sequence $\{\boldsymbol{\pi}^k\}$ converges, and the ordering can be made arbitrarily, as long as each node is iterated upon an infinite number of times ([77, 83, 909, 915, 914]); this enables the use of distributed computations, whereby the node prices are simultaneously updated by several processors, and asynchronous computations (see Section 4.2.2), for which major speedups have been reported in applications to single-commodity flows ([77, 1010, 915, 159, 1008]).

It is of little advantage to balance a node very accurately at any given step of the algorithm; indeed, its balance will later be destroyed when iterating on an adjacent node, and, partly for this reason, each node must be iterated upon a large number of times. (It was indeed observed in [584] that for well-conditioned problems the best results were obtained from choosing the values of $\pi_{i_k}^{k+1}$ at the midpoint of the bracketing interval discussed above.)

The efficiency of coordinate ascent approaches is known to be high initially, and slow as an optimal solution is approached; its behaviour is also very sensitive to the conditioning of the problem (e.g., [83, 584] and [43, Sec. 8.5]). In order to obtain rapid convergence near an optimal solution, a Newton-type approach must be used (to which the coordinate ascent procedure provides an advanced starting solution). In conjugate gradient approaches, the network structure may be utilized efficiently (e.g., [21, 614, 514, 946, 447]).

In light of the fact that in the decomposition scheme a sequence of single-commodity flow problems are to be solved, it is crucial for the efficiency of the overall algorithm that the computational effort spent on solving an individual single-commodity problem is kept low, and thus its solution must be truncated prior to finding its optimal solution. Even

though it is not necessary to obtain a primal feasible solution from a single-commodity problem in order to start solving the next one,[22] it is of advantage computationally, since with feasible solutions at hand, it is possible to devise effective termination criteria for the solution of each single-commodity flow problem as well as the overall scheme based on upper and lower bounds on the corresponding optimal value. A disadvantage of a dual scheme then is that a primal feasible solution is obtained only in the limit of the sequence $\{\mathbf{f}(\boldsymbol{\pi}^k)\}$, i.e., simultaneously to obtaining the optimal primal solution. To obtain a near-optimal and primal feasible solution finitely, it is therefore necessary to use a *primal feasibility heuristic*, whereby a subproblem solution is converted into a primal feasible one. Such a heuristic is embedded in the dual scheme, and applied with regular intervals. (This idea is very well known in applications of Lagrangean relaxation to combinatorial optimization (e.g., [332, 333]), but has been studied and used surprisingly little for continuous optimization problems.)

If $\boldsymbol{\pi}^k$ is near-optimal, then $\mathbf{f}(\boldsymbol{\pi}^k)$ is both near-optimal and near-feasible to the primal problem (e.g., [590, Sec. 8.3] and [43, Sec. 6.5]), and the manipulation needed in order to obtain feasibility tends to zero. A heuristic procedure for converting a subproblem solution into a feasible solution to (4.56) should fulfill two requirements. Firstly, in order to obtain the optimal solution in the limit, the heuristic alteration of the subproblem solution should be conservative in the following sense. Let $\overline{\mathbf{f}}^k = \overline{P}(\mathbf{f}^k)$ be a heuristic projection of the subproblem solution $\mathbf{f}^k = \mathbf{f}(\boldsymbol{\pi}^k)$ onto the feasible set F^n of (4.56). If the mapping $\overline{P} : \Re_+^{|\mathcal{A}|} \mapsto F^n$ satisfies

$$\{\|\overline{P}(\mathbf{f}) - \mathbf{f}\|\} \to 0 \text{ when } \{\|P_{F^n}(\mathbf{f}) - \mathbf{f}\|\} \to 0, \tag{4.61}$$

then, from

$$\|\overline{\mathbf{f}}^k - \mathbf{f}^*\| \leq \|\overline{\mathbf{f}}^k - \mathbf{f}^k\| + \|\mathbf{f}^k - \mathbf{f}^*\| = \|\overline{P}(\mathbf{f}^k) - \mathbf{f}^k\| + \|\mathbf{f}^k - \mathbf{f}^*\|,$$

$\{\overline{\mathbf{f}}^k\} \to \mathbf{f}^*$ follows ([581, 617]). Hence, under the Requirement (4.61) the sequence of solutions generated by the feasibility heuristic tends to the optimal solution. Secondly, in order to make the heuristic procedure computationally cheap, the structure of the feasible set must be exploited in its construction.

The first heuristic of this kind for nonlinear flows is given by Curet [190].[23] The heuristic adjusts an infeasible flow through the solution of a capacitated linear flow problem, which essentially minimizes the maximal flow change in any link necessary in order to obtain feasibility; it can be shown to fulfill (4.61). A recent study of primal feasibility heuristics has been made by Marklund [648]. Embedded in a dual gradient method for quadratic transportation problems, three heuristics are applied; the first is similar to Curet's but minimizes the total flow rerouted, the second uses a breadth-first search to reroute the flow in a residual graph formed by the non-balanced nodes, and the third further takes the travel cost into account by rerouting the flows along cheapest routes in the same residual graph. All three heuristics are more efficient than Curet's in computational comparisons, and provide better primal solutions.

Non-cyclic decomposition algorithms

Choosing the sequence $\{i_k\}$ of indices according to the present conditions instead of to an *a priori* defined ordering of the commodities [such as (4.28)] could lead to a significant

[22]This is clear if one views the decomposition scheme dually as a decomposition scheme over commodities for the solution of the Lagrangean dual formulation of the multicommodity problem [TAP].

[23]The heuristic applied by Ventura [946] to generate upper bounds in dual algorithms for quadratic network problems does not fulfill (4.61).

improvement in the practical performance of a decomposition scheme. Preferably, such an ordering could be defined by the choice of an index i_k in iteration k corresponding to the commodity which is, in some measure, farthest from an optimal solution. (Such orderings define generalizations of the Gauss–Southwell (remotest) ordering ([155, 623]), and may be viewed as block coordinate-wise steepest descent methods.) A number of such measures are defined by the solution to the shortest route Subproblem (4.3) based on the travel costs at the flow $\mathbf{f}^k$.

Petersen [751] defines a measure of the violation of the Wardrop conditions for each origin as a quadratic function of the node prices associated with the shortest route tree, and chooses the index i_k corresponding to the origin with the largest such measure. In order to reduce these additional computations in the decomposition scheme, the shortest route trees are reoptimized. The single-commodity flow problems are solved using a piecewise linear approximation of the objective with few linear segments (e.g., [548, Sec. 8.1] and [540]), and the resulting capacitated, linear single-commodity flow problems are solved by an out-of-kilter method (e.g., [369, 548, 14]). In one of the algorithms proposed by Migdalas [676], the index i_k is chosen corresponding to the O-D pair with the maximal contribution to the value of the gap function,

$$i_k \in \arg\max_{i \in \mathcal{C}} \left\{ \max_{\mathbf{y}_i \in F_i'} \mathbf{t}(\mathbf{f}^k)^{\mathrm{T}} (\mathbf{f}_i^k - \mathbf{y}_i) \right\},$$

or, in other words, the maximal difference between the total transportation cost in a commodity and the total cost of transportation along a shortest route [cf. (3.14)]. Migdalas suggests and validates the application of the Frank–Wolfe algorithm to the single-commodity problems $[\mathrm{TAP}_{i_k}^k]$, supplied with truncation strategies of the form (4.21); this algorithm is a particular example of a truncated sequential partial linearization algorithm.

4.3.3 Parallel decomposition algorithms

If the necessary computer facilities are available, then the sequential decomposition algorithms discussed previously can be implemented such that each single-commodity problem is solved simultaneously, followed by a global update of the commodity flows. Each such problem, or a larger subset of the single-commodity problems, would then be solved by an independent processor. (Since the single-commodity problems can be very large, these processors need of course to be quite powerful.)

Jacobi type approaches

The natural parallel decomposition scheme is the block Jacobi method, in which the structure of the original problem is kept, but the interactions between the commodities are ignored; in iteration k, $|\mathcal{C}|$ problems $[\mathrm{TAP}_i^k]$ of the form (4.42) are solved in parallel (at least conceptually). Note that the Jacobi algorithm is the result of choosing $\varphi^k(\mathbf{f}) = \sum_{i \in \mathcal{C}} \varphi_i^k(\mathbf{f}_i)$, with [cf. (4.31)]

$$\varphi_i^k(\mathbf{f}_i) = T(\mathbf{f}_{i-}^k, \mathbf{f}_i, \mathbf{f}_{i+}^k) = \sum_{a \in \mathcal{A}} \int_0^{f_{ai} + \sum_{j \neq i} f_{aj}^k} t_a(s) ds.$$

The algorithm of Feijoo and Meyer [321] is a Jacobi algorithm for the link-node formulation. The subproblems are solved using piecewise linear approximation techniques.

Numerical examples indicate an increased efficiency of the parallel algorithm with an increase in problem size; this is natural, since the fraction of the total computational time that is spent on the serial part of the algorithm (e.g., the line search) then decreases.

The algorithms of Chen and Meyer [168, 167, 169] resemble the Jacobi algorithm. Their methods are based on a scaled separable approximation of T, which may be derived from choosing ([746])

$$\varphi_i^k(\mathbf{f}_i) = \sum_{a \in \mathcal{A}} \int_0^{\sigma f_{ai} + \sum_{j \neq i} f_{aj}^k} t_a(s) ds, \qquad \sigma > 0,$$

i.e., the subproblem is an extension of that of Jacobi (which follows from choosing $\sigma = 1$) in which a diagonal dominance may be introduced ($\sigma > 1$). The subproblems are solved using a method similar to that of Feijoo and Meyer, but also include a trust region approach for limiting the number of linear segments. Sequential and parallel versions of the algorithm are considered, as well as a combination of the two, for which speedups over serial implementations indicate a relatively small overhead of communication and idle time of the processors.

The algorithm of Larsson *et al.* [584] results from choosing ([746])

$$\varphi_i(\mathbf{f}_i) = \sum_{a \in \mathcal{A}} \int_0^{f_{ai}} t_a(s) ds;$$

the subproblems are solved using a dual coordinate ascent method. The above function φ is not iteration dependent, i.e., in the subproblem of one commodity the flows of the other commodities are removed from the network; the performance of the algorithm should therefore be expected to be less efficient than of the Jacobi approach.

The algorithms given above are based on solving subproblems with an objective of the same form as the original one. This is of advantage in the sense that the number of main iterations that is needed in order to obtain an accurate solution is low; on the other hand, each subproblem is difficult to solve, since the objective is highly nonlinear. Moreover, the complexity of the subproblems corresponding to the different processors may be very different, and hence the efficiency of the parallel algorithm may be degraded due to some processors frequently becoming inactive.

One possibility for improving the efficiency is to introduce asynchronous computations (see Section 4.2.2). Another possibility is to modify the Jacobi subproblems so that the individual subproblems are of the same complexity. The approaches discussed next also lead to much easier subproblems.

Gradient projection type approaches

Consider the link-route formulation of [TAP], and the corresponding Jacobi subproblem [TAP$_i$] of the form (4.45). A second-order approximation of the Jacobi subproblem corresponds to choosing each of the functions φ_i^k as

$$\varphi_i^k(\mathbf{h}_i) = \sum_{r \in \hat{\mathcal{R}}_i} \left(c_{ir}(\mathbf{h}^k)(h_{ir} - h_{ir}^k) + \frac{1}{2} \frac{\partial^2 T(\mathbf{h}^k)}{\partial h_{ir}^2} (h_{ir} - h_{ir}^k)^2 \right); \tag{4.62}$$

each resulting subproblem objective is of the form (4.62), and therefore this *linearized Jacobi* algorithm is equivalent to a *diagonalized* Newton algorithm [cf. (4.54)].

Bertsekas [69] suggests using such an approach for the solution of the optimal routing problem in a decomposition algorithm of the form $F^r(D_C^P[C_{\mathcal{R}}])$ (see also Section 4.3.2 for

sequential versions); the possibilities for operating the algorithm in a distributed manner are discussed particularly.[24]

A method related to the gradient projection methods is the projection method of Rosen [789]. Schwartz and Cheung [821] apply it to the link-node formulation of the optimal routing problem; they observe that the projection decomposes into independent single-commodity problems, and find that the algorithm compares favourably with the Frank–Wolfe algorithm for small networks. For larger networks, however, it is probably prohibitively expensive since the projection operation does not utilize the network structure. Rosen's projection method adapts much better to restrictions of the link-route formulation. Such an approach is applied by Soumis [865] in a column generation algorithm of the form $F^r(D_{\mathcal{C}}^S[C_{\mathcal{R}}])$.

4.3.4 Aggregate simplicial decomposition algorithms

What is traditionally referred to as a simplicial decomposition algorithm for the traffic assignment problem is, according to what has been said in Section 4.2.3, a column generation algorithm based on the Frank–Wolfe subproblem, and normally also based on an aggregated representation of feasible link flows. Applied to traffic assignment, the basic algorithm consists of the same steps as the Frank–Wolfe algorithm (see Section 4.1.1), with the exception of **Step 3**, which is replaced by the addition of the all-or-nothing solution, $\mathbf{y}^k$, obtained from **Step 2**, to the set of previously generated all-or-nothing solutions (i.e., $\hat{\mathcal{X}} := \hat{\mathcal{X}} \cup \{\mathbf{y}^k\}$), and the solution of the restricted master problem

[RMP]

$$\min\ T(\mathbf{f}), \tag{4.63a}$$

subject to

$$\mathbf{f} = \lambda^0 \mathbf{f}^k + \sum_{j \in \hat{\mathcal{X}}} \lambda^j \mathbf{y}^j, \tag{4.63b}$$

$$\lambda^0 + \sum_{j \in \hat{\mathcal{X}}} \lambda^j = 1, \tag{4.63c}$$

$$\lambda^0, \lambda^j \geq 0, \qquad \forall j \in \hat{\mathcal{X}} \tag{4.63d}$$

to yield $\mathbf{f}^{k+1}$.

The feasible set of [RMP] is the convex hull of the previous solution $\mathbf{f}^k$ and the convex hull of the known subset $\hat{\mathcal{X}}$ of the set of extreme points of F^n, i.e., the set of link flows that can be described as convex combinations of the known all-or-nothing solutions. (Disregarding the definitional Constraints (4.63b), there is hence only one linear constraint in [RMP].) We note here that the maximum number of extreme points needed to describe any feasible link flow solution to [TAP] is $|\mathcal{A}| + 1$, while the number of columns in the complete master problem is $\prod_{(p,q)\in\mathcal{C}} |\mathcal{R}_{pq}|$.

The distinction between two aggregate simplicial decomposition algorithms for [TAP] is defined by the respective method used for solving [RMP], and the rules used for dropping columns from $\hat{\mathcal{X}}$. In terms of the taxonomy introduced in Section 4.2.5, all these methods can be described as column generation methods of the form $F^n(C_{\mathcal{A}})$. Below, we outline the development of aggregate simplicial decomposition (ASD) methods for [TAP].

[24] The algorithm can also be executed in a partially asynchronous fashion (see Section 4.2.2); convergence results for this algorithm may be found in [916, 917, 918, 84, 913].

The first ASD scheme is due to Cantor and Gerla [420, 141], who develop the *extremal flows method* for the optimal routing problem. They propose solving [RMP] by using Rosen's [789] gradient projection method, which is easily applied because of the simple structure of the constraints of [RMP]. They employ Carathéodory's Theorem explicitly in their column dropping rule, i.e., they allow a maximum of $|\mathcal{A}| + 1$ extreme points in $\hat{\mathcal{X}}$. Best [85] proposes to use a conjugate direction method for [RMP].

Florian [347] applies the *away step* procedure of Wolfe [987] to [RMP]. The away step direction is similar to the Frank–Wolfe direction, but is directed from the worst extreme point, and actually leads to a better convergence rate (see also [440]). (The away step approach is not directly applicable to [TAP], since it involves calculating the direction from the longest simple route pattern towards the current flow, and this is an *NP*-complete problem (e.g., [411, p. 213]), due to the presence of cycles.) When comparing directional derivatives to that of the Frank–Wolfe direction, the away step direction was always chosen. It is interesting to note the resemblance between the away step procedure and the class of equilibration operator approaches (which was interpreted as convex simplex methods in Section 4.3.2). Let $\mathbf{p}_{FW}^k = \mathbf{y}^k - \mathbf{f}^k$ be the Frank–Wolfe direction and $\mathbf{p}_A^k = \mathbf{f}^k - \mathbf{z}^k$ the away step direction, where $\mathbf{y}^k$ and $\mathbf{z}^k$ are, respectively, the shortest and longest route patterns given $\mathbf{f}^k$. Then $\mathbf{p}_{FW}^k + \mathbf{p}_A^k = \mathbf{y}^k - \mathbf{z}^k$, i.e., the sum of the Frank–Wolfe and away step directions, yields the direction from the longest route pattern towards the cheapest one, and hence the direction of the equilibration operator approach.

Dow [267, 941] applies two algorithms for [RMP], a Frank–Wolfe approach and a method based on second derivatives. The latter produced better results, but in a limited comparison with a heuristic quantal loading procedure (see Section 1.5.4), the results were discouraging. He concludes, however, that the simplicial decomposition approach is preferable from the viewpoint of solution analysis, since more information about the optimal solution becomes available.

Guélat [439] presents a reduced gradient approach for [RMP]; the overall algorithm is restarted at regular intervals by the dropping of all the extremal flows stored.

The fact that [RMP] may have relatively few variables enables the use of second-order methods for its solution; von Hohenbalken [953] was the first to suggest such methods. Pang and Yu [739] approximate each [RMP] by a quadratic program (as suggested in [953]), for which they apply the algorithm in [931]. The algorithm shows good performance, and is also extended to solve asymmetric models (see Section 5.3.5).

Hearn *et al.* [595, 480, 481] use the projected Newton method of Bertsekas [67, 70, 72] to solve [RMP], and also consider a quadratic approximation as in [739]. They investigate the result of [479], which states that finite convergence is obtained if the maximum number of extreme points stored satisfies $r \geq \mathtt{dim}\,(F^*) + 1$, i.e., if the maximum number of points retained is higher than the dimension of the optimal face of F^n (see Section 4.2.3), by varying the value of the parameter r. Results taken from experiments on traffic assignment and other nonlinear network flow problems are very promising; the number of shortest route calculations is small compared to most other ASD schemes and the overall algorithm is more efficient, especially when the value of r is large.

In order to enable the use of second-order methods, the parameter value must be kept small, but the lower it becomes, the larger number of restricted master problems will need to be solved. It is therefore important to give the parameter a proper value; the difficulty of estimating the dimension of the optimal face of F^n *a priori*, however, makes this choice very difficult. The so called *restricted simplicial decomposition* (RSD) algorithm is still considered as one of the state-of-the-art codes for traffic assignment.

Montero [684] investigates many aspects of the implementation of an ASD scheme, including the proper choices of the starting solution, shortest route algorithm, algorithms and stopping criteria for each restricted master problem, and criteria for dropping columns with small weights. Experiments are performed on most of the known test networks in the literature and large networks modelling the cities of Barcelona and Madrid. In comparisons with different projection algorithms, projected Newton methods become more efficient than fixed metric projection methods when the dimensions of the restricted master problems are in the order of ten and above. Fixed projection methods are also shown to be very sensitive to the choice of step length parameter. For larger networks, it is clear that column dropping is necessary; moreover, a column with a zero weight in the optimal solution of a restricted master problem can always be removed without affecting the following main iterates.

4.3.5 Disaggregate simplicial decomposition algorithms

Relations between column generation and simplicial decomposition revisited

By combining the analysis made in Section 2.2.2 on the relations between the link-node and link-route formulations of [TAP] and the analysis made in Section 4.2.3 on the connections between column generation and simplicial decomposition, we here conclude the analysis by providing the connections between column generation and simplicial decomposition for traffic assignment problems.

The link-node formulation of [TAP] is a special case of the general Problem (4.33) of Section 4.2.3; the Constraints (2.13) defining F^n are easily verified to be of the form (4.33b)–(4.33c), where $\mathbf{x}_i = \mathbf{f}_i$ and $n_i = |\mathcal{A}|$ for all $i \in \mathcal{C}$ (i.e., $m = |\mathcal{C}|$). The feasible sets X_i, $i \in \mathcal{C}$, correspond to the commodity flow conservation Constraints (2.13a)–(2.13b), and the set Z to the link flow definitional Constraints (2.13c). Each extreme point of X_i is a feasible single-commodity flow solution defined by the demand d_i of flow carried on one simple route. The internal Representation (4.34) of the set X_i thus corresponds to the set of commodity link flows defined by all convex combinations of such route flows, and hence all feasible commodity route flow solutions. The complete master Problem (4.35) is hence equivalent to an inner representation of the link-node formulation of [TAP]. This relationship is well known in applications of decomposition methods to multicommodity network flows ([368, 369, 904, 905, 906, 32, 548, 785, 586, 14, 316, 531]). (Note that the cycle flows that are present in the link-node formulation are eliminated.)

The restricted master Problem (4.36) corresponds to the case where the sets $\mathcal{X}_i$ of the routes in commodity i is replaced by a known subset $\hat{\mathcal{X}}_i$, and is the *disaggregated* formulation of a restricted master problem in a simplicial decomposition scheme for [TAP]. An *aggregated* restricted master problem is obtained by defining the columns according to the Aggregation (4.39) of the extreme points of the individual sets X_i; this is equivalent to aggregating individual route flows into all-or-nothing solutions, and results in the aggregated restricted master Problem (4.63).

To make the relationships to the link-route formulation clear, we substitute the index i for the O-D pair (p, q), the sets $\hat{\mathcal{X}}_i$ for the subsets $\hat{\mathcal{R}}_{pq} \subset \mathcal{R}_{pq}$ and the index j for r, and introduce the route flow variables h_{pqr} through ([586])

$$h_{pqr} = \lambda_{pqr} d_{pq}, \qquad \forall r \in \hat{\mathcal{R}}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{4.64}$$

From this substitution, it follows that the internal representation of X_i given by (4.34) is equivalent to the demand feasibility Constraints (2.6b)–(2.6c) of the link-route formulation

(2.6) of [TAP], and that the disaggregated restricted master problem is equivalent to a restriction of this formulation to the known subsets $\hat{\mathcal{R}}_{pq}$ of the routes. This is the familiar formulation of the restricted master problem of the column generation methods for [TAP], and we have thus shown that the aggregate simplicial decomposition scheme for [TAP] is nothing but an aggregate version of the general column generation scheme for the link-route formulation.

The link between the two algorithm classes is the *disaggregate simplicial decomposition* (DSD) algorithm, i.e., a simplicial decomposition scheme where extreme points are stored individually for the different feasible sets in the Cartesian product $\prod_{i\in\mathcal{C}} X_i$; each such set is hence given an internal representation, which results in using one convexity constraint for each set X_i in [RMP].

Disaggregate simplicial decomposition

The term disaggregate simplicial decomposition is due to Larsson and Patriksson [586], who also establish its relationships to column generation methods. In their algorithm, each restricted master problem is solved using a combination of two decomposition methods over O-D pairs; a reduced gradient algorithm is employed to reach a near-optimal solution, after which a diagonalized Newton method is used. Shortest route columns are generated after the solution of each restricted master problem, and thus the overall method is of the form $F^r(C_{\mathcal{R}}[D^P_{\mathcal{C}}])$. Numerical tests performed on most known test networks indicate a very rapid convergence, and in particular that the number of shortest route calculations are minimized in this approach. (This is due to the fact that routes are stored individually, and that shortest routes are not calculated until a restricted master problem has been solved sufficiently accurately.)

We next discuss some important consequences of the choice of aggregation in simplicial decomposition.

4.3.6 Comparisons between aggregated and disaggregated representations

Size of master problems and finite convergence

We first note that a complete aggregated master problem has only one linear (convexity) constraint, while the number of columns, which is the number of possible all-or-nothing solutions, is $\prod_{(p,q)\in\mathcal{C}} |\mathcal{R}_{pq}|$; the maximum number of columns needed to express any feasible solution and an optimal one is $|\mathcal{A}| + 1$ and $\mathtt{dim}\ (F^*) + 1$, respectively, where $\mathtt{dim}\ (F^*)$ is the dimension of the optimal face of F^n. Correspondingly, a complete disaggregated master problem has $|\mathcal{C}|$ linear (convexity) constraints, while the number of variables is $\sum_{(p,q)\in\mathcal{C}} |\mathcal{R}_{pq}|$, which is much smaller than in the aggregated version; the maximum number of columns needed to express any feasible commodity flow solution and an optimal one in particular is $|\mathcal{C}|(|\mathcal{A}| + 1)$ and $\sum_{(p,q)\in C}(\mathtt{dim}\ (F^*_{pq}) + 1)$, respectively.

As was noted in Section 4.3.4, it is very difficult to estimate *a priori* the dimension of the optimal face of F^n, and consequently the maximum number of extreme points of F^n to be retained in the restricted master problems. In the DSD scheme, let r_{pq} be the maximum number of routes retained for commodity (p, q). The minimum value of this parameter that ensures the finite convergence of the algorithm is, from the above, $\mathtt{dim}\ (F^*_{pq}) + 1$. The dimension of the optimal face of the commodity link flow polyhedron F_{pq} is the number of routes actually used within the commodity in an equilibrium solution, minus one. It follows that in order to yield finite convergence, the value of the parameter

r_{pq} must not be less than the number of routes used. This is perhaps an obvious result, but it illustrates that in a disaggregated version of the simplicial decomposition scheme, the storage requirements needed can be estimated from, for instance, knowledge of the level of congestion of the traffic network being studied.

Levels of aggregation

From the tests performed with the DSD algorithm in [586], it is reported that the number of main iterations needed is, approximately, bounded by the maximal number of routes utilized in any O-D pair. This implies that the total number of shortest route calculations needed is very limited, and that the DSD algorithm seems to be optimal with respect to the number of shortest route calculations. (It is well known that in ASD schemes for large networks, the vast majority (around 80%–90% according to [481]) of the calculations is spent in the shortest route subproblem phase. In the disaggregated version, these portions are essentially reversed.) Among the optimal route flow solutions, the DSD algorithm seems to provide that which utilizes the minimum number of routes. Larsson and Patriksson [586] conclude that in the disaggregated approach, column dropping is unnecessary. (The advantages of using a disaggregated representation in decomposition methods is therefore supported both by applications to nonlinear ([586]) and linear ([531]) network flow problems.)

Because each route flow may be adjusted independently in the disaggregated version, an optimal solution may be reached much faster in terms of numbers of iterations than in the aggregated version, where one route may be present in several extreme points. (One may say that the routes have a greater striving for the optimum.) The difference becomes even more pronounced with a smaller value of the r parameter in the aggregated version.

This is illustrated below, where we also compare the generation of the iterate $\mathbf{f}^{k+1}$ to that defined by the Frank–Wolfe algorithm. For simplicity of presentation, we assume that no column dropping has been used, and that the initial solution is given by an all-or-nothing solution.

In the Frank–Wolfe algorithm, $\mathbf{f}^{k+1}$ is given by [cf. (4.5)]

$$\mathbf{f}^{k+1} = (1-\lambda^k)\left(\sum_{j=0}^{k-1} \overline{\lambda}^j \mathbf{y}^j\right) + \lambda^k \mathbf{y}^k, \qquad \lambda^k \in [0,1],$$

where

$$\overline{\lambda}^j = \begin{cases} \prod_{l=0}^{k-1}(1-\lambda^l), & j=0, \\ \lambda^j \prod_{l=j+1}^{k-1}(1-\lambda^l), & j=1,\ldots,k-1. \end{cases}$$

In the ASD algorithm,

$$\mathbf{f}^{k+1} = \sum_{j=0}^{k} \lambda^j \mathbf{y}^j,$$

where

$$\sum_{j=0}^{k} \lambda^j = 1, \qquad \lambda^j \geq 0,\ \forall j \in \{0,\ldots,k\}.$$

Finally, in the DSD algorithm,

$$\mathbf{f}_{pq}^{k+1} = \sum_{j=0}^{k_{pq}} \lambda_{pq}^j \mathbf{y}_{pq}^j = \sum_{r \in \hat{\mathcal{R}}_{pq}} \delta_{pqra} \lambda_{pqr} d_{pq}, \qquad \forall (p,q) \in \mathcal{C},$$

where

$$\sum_{j=0}^{k_{pq}} \lambda_{pq}^j = 1, \qquad \lambda_{pq}^j \geq 0, \ \forall j \in \{0, \ldots, k_{pq}\}, \ \forall (p,q) \in \mathcal{C},$$

and where we let $k_{pq} = |\hat{\mathcal{R}}_{pq}| - 1$ and use the Substitution (4.64).

The DSD algorithm reduces to the ASD algorithm when $\lambda_{pq}^j = \lambda^j$ for all $(p,q) \in \mathcal{C}$, which further reduces to the Frank–Wolfe algorithm by letting $\lambda^j = \overline{\lambda}^j$ for all $j \leq k-1$. (The Frank–Wolfe algorithm is therefore an ASD algorithm where the previously generated all-or-nothing solutions are given weights according to earlier line searches.) It is clearly seen how an aggregation imposes interactions among the individual routes.

The reader should note that, although search directions in an iterative algorithm for [TAP] may be described in terms of individual commodity flows, (inexact) line searches should always be performed with respect to total link flows, since the evaluations of the objective then involves fewer operations; it is always possible to translate a commodity-based search direction to one in the space of total link flows, by the use of the Relations (2.6d) and (2.13c). (The step length obtained is then used in the original direction.) Special care may need to be taken in order to guarantee that the nonnegativity constraints are satisfied in the space of commodity flows, since a feasible step in the space of total link flows may not correspond to a feasible step in the space of commodity flows.[25]

Effective termination criteria based on lower bounds on the optimal objective value are not only available in the Frank–Wolfe algorithm. A lower bound on the optimal value of each restricted master problem of a simplicial decomposition scheme is determined automatically when we evaluate a value of the gradient, from the lowest value among its components; this is in fact the lower bound that the Frank–Wolfe subproblem, applied to the restricted master problem, would produce at the given point.

Reoptimization facilities

It is necessary for an assignment algorithm to have a good reoptimization capability when traffic assignment problems arise as subproblems in, for instance, the solution of a time-sliced traffic assignment problem or an equilibrium network design problem, and also when intercity freight flows or origin-destination matrices are to be estimated.

Larsson and Patriksson discuss the reoptimization capabilities of the DSD algorithm with respect to changes in link performance functions, travel demands and network topology. The disaggregated representation enables these perturbations to be much more efficiently handled than in an aggregated formulation. Consider, for instance, a change in the demand vector; because the solution is described in terms of individual route flows, through a simple scaling of these flows, the optimal solution to the unperturbed problem is a feasible solution to the perturbed one. In an aggregated formulation, however, a perturbation of the demand vector destroys the feasibility of the previous solution. Topology

[25] Assume that [TAP] is solved with a truncated partial linearization algorithm in the space of total link flows. In iteration k, let the subproblem $[\mathrm{TAP}_{\varphi^k}^k]$ be solved approximately with a_k iterations of a descent algorithm, and assume that the search direction generation step of this algorithm yields auxiliary flows $\mathbf{y}_i^k \in F^n$, $i = 1, \ldots, a_k$, that all correspond to commodity-feasible flows. (If, for instance, $[\mathrm{TAP}_{\varphi^k}^k]$ is solved using a truncated Frank–Wolfe algorithm, then the solutions $\mathbf{y}_i^k$ correspond to all-or-nothing solutions, which are clearly commodity-feasible.) The maximum step in the obtained search direction, $\mathbf{y}^k - \mathbf{f}^k$, such that the commodity flows are nonnegative is the maximum step l such that $\mathbf{f}^k + l(\mathbf{y}^k - \mathbf{f}^k)$ is in the convex hull of the points $\mathbf{y}_i^k \in F^n$, $i = 1, \ldots, a_k$, and can easily be calculated by only keeping track of the step lengths taken in the subproblem phase. Indeed, if the step lengths used in the subproblem phase are α_i, $i = 1, \ldots, a_k$, then the maximum step is $1/(1 - \prod_{i=1}^{a_k}(1-\alpha_i))$.

changes are also more easily handled because of the fact that more information is stored; the removal of links only affects the individual routes that utilize these links, while in an aggregated representation of feasible flows, all columns must be removed. The excellent reoptimization capabilities of the DSD algorithm have been utilized in extensions to capacitated traffic assignment (see Section 4.6.1) and stochastic user equilibrium problems (see Section 4.5.2), as well as to solve O-D matrix estimation problems ([273]).

With modern computer technology at our disposal, route-flow based algorithms have become practical tools for the analysis of traffic networks. This is important for two main reasons. Firstly, it is increasingly important to obtain route flow information, and link-flow based algorithms do not provide this information automatically; route flow information is crucial for the estimation of pollutant emissions and induced origin-destination flows in subareas, and of the potential uses of route guidance strategies. Secondly, the excellent reoptimization capabilities of route-flow based algorithms provide the means for analyzing different scenarios quickly.

A unified description of column generation methods for [TAP]

We conclude our discussions on column generation algorithms for the solution of traffic assignment problems by providing a list of references for such methods.

The list is given in an increasing order of the number of shortest route calculations that can be anticipated to be needed to yield an optimal solution. (A majority of these methods can be viewed as methods in which the solution of each restricted master problem is truncated very early; the total number of routes generated, and possibly also the total number of routes used in the equilibrium solution obtained, then increases significantly.)

$F^r(C_{\mathcal{R}}[D^P_{\mathcal{C}}])$: [586]

$F^r(C_{\mathcal{R}}[D^C_{\mathcal{C}}])$: [611, 713]

$F^r(D^S_{\mathcal{O}}[C_{\mathcal{R}}])$: [424]

$F^r(D^S_{\mathcal{C}}[C_{\mathcal{R}}])$: [359, 717, 865, 69, 71, 82, 352, 813, 65, 608]

$F^r(D^P_{\mathcal{C}}[C_{\mathcal{R}}])$: [69, 957]

$F^n(C_{\mathcal{A}})$: [420, 141, 85, 347, 267, 941, 439, 739, 595, 480, 481]

4.3.7 Dual algorithms

The fundamental property of a traffic system in equilibrium, as described by Wardrop's first principle, is characterized in terms of travel times. The natural basis for mathematical models for the analysis of traffic equilibria would therefore be travel time variables, instead of flow variables which is the predominantly utilized modelling basis. In Section 2.3.3 we developed a mathematical model for finding the equilibrium travel times as an inverse (or dual) problem to the standard traffic assignment problem. In this section, we outline the methods proposed for the solution of this inverse problem, and show how it can be used to indirectly solve the traffic assignment problem.

Consider the program [DTAP-E], that is, the Problem (2.38). According to Theorem 2.6, it is a convex program with a subdifferentiable objective, and can therefore be solved using any algorithm for nondifferentiable convex programs. Its nature as a dual

program, however, would seem to suggest our using dual ascent or subgradient optimization approaches; the latter applies to [DTAP-E] as follows.

Given a tentative travel cost, $\boldsymbol{\mu}^k \geq \mathbf{t}(\mathbf{0})$, a subgradient of θ at $\boldsymbol{\mu}^k$ is computed. According to Theorem 2.6, the subdifferential of θ at $\boldsymbol{\mu}^k$ is the convex hull of the values of the function defined by the link flow definitional Constraint (2.27d) at the solution set of [SR] and [SC]. Hence, a subgradient is given by

$$\boldsymbol{\xi}_\theta^k = \mathbf{y}^k - \mathbf{f}(\boldsymbol{\mu}^k);$$

the vector $\mathbf{y}^k$ is an all-or-nothing solution which is formed by the shortest routes given the link travel cost vector $\boldsymbol{\mu}^k$ and carrying the demand $\mathbf{d}^k = \mathbf{d}(\boldsymbol{\pi}(\boldsymbol{\mu}^k))$ given by (2.33), and $\mathbf{f}(\boldsymbol{\mu}^k)$ is the (unique) solution to [SC]. The new solution is defined by

$$\mu_a^{k+1} = \max\{t_a(0), \mu_a^k + l_k \xi_{\theta a}^k\}, \qquad \forall a \in \mathcal{A}, \tag{4.65}$$

where l_k is a step length which guarantees convergence (e.g., [247, 836]). If, for instance, the sequence $\{l_k\}$ satisfies the divergent series condition

$$\lim_{k\to\infty} l_k = 0, \qquad \sum_{k=1}^{\infty} l_k = +\infty, \qquad \text{and} \qquad \sum_{k=1}^{\infty} l_k^2 < +\infty, \tag{4.66}$$

then the sequence $\{\boldsymbol{\mu}^k\}$ is bounded and converges to the unique solution $\boldsymbol{\mu}^*$ of [DTAP-E] ([582]). An alternative is to use the well-known modification of Polyak's [759] step length formula

$$l_k = \gamma_k \frac{\theta^* - \theta(\boldsymbol{\mu}^k)}{\|\boldsymbol{\xi}_\theta^k\|^2}, \qquad 0 < \varepsilon_1 \leq \gamma_k \leq 2 - \varepsilon_2 < 2, \tag{4.67}$$

where the (unknown) optimal value θ^* of [DTAP-E] (and [TAP-E], from the strong duality result of Theorem 2.7) usually is replaced by an upper bound. To obtain an upper bound on θ^*, a feasible solution to [TAP-E] must be generated from the subproblem solution; in applications of Lagrangean relaxation in general, this is a difficult problem, but in the application to [TAP-E], the relaxed constraints are definitional, and a feasible solution is obtained directly from the all-or-nothing solution $\mathbf{y}^k$. Hence, with θ^* replaced by $T(\mathbf{y}^k, \mathbf{d}^k)$, the Formula (4.67) may be used. The use of this formula is questionable, however, since this upper bound does not converge to θ^* when $\{\boldsymbol{\mu}^k\} \to \boldsymbol{\mu}^*$.

Fukushima [389] applies the Method (4.65), (4.67), with θ^* replaced by upper bounds given by all-or-nothing solutions, to [DTAP-E] and shows in limited experiments that it is at least comparable to the Frank–Wolfe algorithm. In [388], the subgradient optimization procedure is replaced by a proximal point type dual ascent algorithm, where each subproblem is solved using a cutting plane method (see [386, 553, 497, 263]); the basic algorithm component still is the calculation of shortest routes, and a computational test performed on the Sioux Falls network shows a behaviour similar to that of the Frank–Wolfe algorithm.

Goffin [428, 429] applies an ellipsoid algorithm and also subgradient optimization to a problem of the form [DTAP] which arises from a dualization of an optimal routing problem. He concludes that both algorithms are sensitive to the choice of tuning parameters (in the case of subgradient optimization: the step length parameters), but are efficient when parameters are properly chosen.

The above approaches will provide equilibrium travel costs in the limit of $\{\boldsymbol{\mu}^k\}$; the sequence of upper bounds will not, however, converge to the optimal value (and in general not even come close), and therefore primal optimality is obtained only when $\mathbf{f}(\boldsymbol{\mu}^k)$ becomes

feasible, which occurs in the limit only. In order to yield a (near-)optimal primal feasible solution to [TAP-E], the dual algorithm must be supplied with an additional scheme for calculating primal feasible solutions that tend to optimal ones.

Larsson *et al.* [582] develop a simple scheme for generating primal feasible flows which optimize in the limit. [This is based on primal convergence results for linear programs ([836, 581]).] In one version of the algorithm the primal feasible solutions are given by

$$\mathbf{f}(l) = \frac{1}{l}\sum_{k=1}^{l} \mathbf{y}^k, \qquad \forall l \geq 1, \tag{4.68}$$

i.e., as simple averages of the all-or-nothing solutions obtained when calculating the subgradients of θ; the sequence of demand flows is constructed analogously. Convergence of $\{\mathbf{f}(l), \mathbf{d}(l)\}$ to the primal optimal solution is guaranteed when the step lengths in (4.66) are chosen according to $l_k = a/(b + ck)$, where $a, c > 0$ and $b \geq 0$.

Note that the sequence $\{\mathbf{f}(l)\}$ given by (4.68) is similar to that given by the Frank–Wolfe algorithm, and in particular to that given by the MSA algorithm (see Section 4.1.6). The main difference is that the travel times upon which the generation of the all-or-nothing solutions are based, in those algorithms are given by $\mathbf{t}(\mathbf{f}^k)$ for some feasible flow $\mathbf{f}^k$, while in the dual algorithm they are given by the values of the dual variables. The experiments conducted in [582] indicate that $\{\boldsymbol{\mu}^k\}$ converges more rapidly to the optimal travel times than $\{\mathbf{t}(\mathbf{f}^k)\}$ does in the Frank–Wolfe or MSA algorithms.

The dual Procedure (4.65), (4.66) has the nice property that after a finite number of iterations, the routes solving [SR] are among the equilibrium routes ($\mathcal{R}_{pq}(\boldsymbol{\mu}^k) \subseteq \mathcal{R}^*_{pq}$). Since the dual method is memory-less, the averaging Process (4.68) may be postponed until some iteration L, and if the value of L is chosen sufficiently large, then all the routes defining $\mathbf{f}(l)$ will be equilibrium routes. This may accelerate the convergence of $\{\mathbf{f}(l)\}$ towards $\mathbf{f}^*$ significantly, and is a major improvement over the Frank–Wolfe algorithm, in the sense that the weights of the non-equilibrium routes generated early in the Frank–Wolfe algorithm tend to zero very slowly (see Section 4.1.5).

In contrast to most primal methods for traffic assignment, such as the Frank–Wolfe algorithm, dual algorithms are quite easily extended to more complex models. (Larsson *et al.* [582] show how their method easily extends to traffic assignment models with link flow observations, capacitated assignment, combined distribution and assignment, and stochastic models.) Furthermore, an estimate of the equilibrium travel times (which may be available from either travel time measurements or flow observations) can be used as an advanced start in dual algorithms, and thus facilitate the generation of a near-optimal primal feasible solution in a few iterations only; such information can not, however, be as easily utilized in primal methods, since the information may very well be inconsistent with respect to the flow conservation constraints.

The cutting planes generated in the methods of [388, 428, 429] are supporting hyperplanes of the epigraph of θ; these can be used to obtain a sequence of primal feasible solutions which optimize in the limit. Hearn and Lawphongpanich [476, 477] apply a cutting plane algorithm (which may be interpreted as a nonlinear Dantzig–Wolfe algorithm) to the dual program of a capacitated traffic assignment problem. (The upper bound constraints on f_a are appended to [CS].) The evaluation of $\theta(\boldsymbol{\mu}^k)$ through the solution of [SR] and [SC] yields a cut which is included in a linear restricted master problem of the form

$$\max_{\boldsymbol{\mu}, w} \; w, \tag{4.69a}$$

subject to

$$w \leq \sum_{a \in \mathcal{A}} \left\{ \int_0^{f_a(\mu_a^k)} t_a(s)ds + \mu_a(y_a^k - f_a(\mu_a^k)) \right\}, \qquad \forall k, \tag{4.69b}$$

$$\mu_a \geq 0, \qquad \forall a \in \mathcal{A}. \tag{4.69c}$$

It is well known that a primal feasible solution is obtained from the linear programming dual of this problem, and that this solution optimizes in the limit (e.g., [43, Sec. 6.5]). In the method of [476, 477], an approximate line search is made in the dual space towards the solution of (4.69) in order to obtain dual ascent. The algorithm is also extended to asymmetric models; see Section 5.3.6. Convergence towards the optimal primal solution can be expected to be faster in terms of numbers of iterations than in the simple averaging technique of [582]; whether the smaller number of iterations required amortizes the need to solve a linear restricted master problem in each iteration is uncertain, however.

4.3.8 Network aggregation algorithms

The high computational cost for solving large-scale traffic equilibrium problems has prompted the use of techniques, in which a network is aggregated in some way. An aggregated network may be defined through the combination of nodes and links into supernodes and superlinks, but it is more common to extract an interesting subnetwork from the original one for separate study; the subnetwork usually consists of the main arterials and centroids.

To compensate for the loss of detail, so that the resulting flows and costs that are obtained in the subnetwork are (approximately) consistent with those of the larger network, it is essential that both the level of aggregation and the characteristics of the nodes and links of the subnetwork (i.e., the O-D matrix and the link travel cost functions) are chosen appropriately. It is, however, neither possible to define a general aggregation policy that retains the characteristics of the original network, due to the congestion effects, nor to derive measures of accuracy of the output from the aggregated model. [This is in contrast to linear flow models, where such measures are available ([1015]).] Empirical results ([618, 472, 522, 99, 100, 287]) also confirm that the quality of the results deteriorate with the level of aggregation.

Although aggregated models are interesting in themselves, iterative aggregation can also be used as a means for solving large problems. Decomposition algorithms, highly related to generalized Benders decomposition techniques ([419]), have been proposed by Dantzig *et al.* [226] and Hearn *et al.* [472, 40, 475, 41], and extended to asymmetric models in [593, 596]. In these algorithms, a network is divided into, for example, two subnetworks, which are analyzed alternately; through a set of artificial links, the solution to one subnetwork problem defines an O-D matrix for the other subnetwork, and in the limit the flows on the subnetworks are consistent with the equilibrium flows of the original network. These algorithms do not, however, result in a true aggregation of the original network.

In the heuristic of Haghani and Daskin [448], links where small amounts of flow are anticipated are iteratively removed from the network. To compensate for the resulting loss of capacity, the O-D matrix is also updated. Bovy and Jansen [99, 100] consider aggregating the network while retaining primary links. The network obtained is then analyzed with the travel cost functions unaltered. As pointed out by Hazelton [470], the smaller network is required to carry nearly as much flow as the original one, but with a drastically reduced capacity; this results in an overestimation of the flows and travel

times on the links. Hazelton investigates the changes of the travel cost functions required in order to obtain consistent results, for networks of simple structure.

The problem of network aggregation has also been studied from the viewpoints of applied statistics (e.g., [525, 526]), and of the continuous representation of networks (e.g., [192, 215, 217, 834, 808]). See Hearn [472, 475] for comprehensive reviews of the aggregation problem, and an account of aggregation practices.

4.3.9 Other algorithms

Kuhn [570, 571] applies a fixed point technique to the link-route formulation of [TAP]. The algorithm is based on Scarf's [810] labeling method, which was originally proposed for economic equilibrium problems. The algorithm is found to work well for small examples; however the need to enumerate routes makes it inapplicable to larger networks. Similar fixed point methods are proposed in [827, 410, 1005].

A few attempts have been made to solve the traffic equilibrium problem by addressing the Wardrop conditions directly. Although this is of course possible, the solution of the primal-dual system of optimality conditions by any method for systems of nonlinear equations would probably not utilize the network structure and thus be applicable only to very small networks. A Newton-Raphson method is applied to this system by Wilkie and Stefanek [975]. Kulash's [573] method is applicable only to linear cost functions, and involves the inversion of network-based matrices.

The methods of Snell *et al.* [1000, 860] are based on Pontryagin's maximum principle.

We finally mention that algorithmic approaches for the basic model and its extensions have been surveyed earlier in [359, 717, 69, 71, 82, 352, 813, 65, 608, 713, 798, 799, 412, 416, 774, 972, 32, 631, 941, 370, 351, 630, 381, 352, 656, 110, 675, 894, 358].

We now turn to study some algorithms specialized for elastic demand problems.

4.4 Algorithms for elastic demand problems

The development of algorithms for the elastic demand problem has, to a large extent, been parallel to that of algorithms for the fixed demand model; this is true also for the heuristics first suggested for use in traffic assignment. In addition, through the fixed demand reformulations outlined in Section 2.2.4, any method applicable to a fixed demand problem can be used for the solution of elastic demand problems.

The first convergent algorithms specialized for use in elastic demand models are due to Gibert [423] and Bruynooghe *et al.* [132]; the algorithms are essentially adaptations of the equilibration operator approach. The convex simplex method (embedded in a cyclic decomposition scheme over origins) is applied by Nguyen [715, 716, 717], and the reduced gradient algorithm is considered in [713, 717].

In Wigan's [971] elastic demand algorithm, fixed demand problems are solved iteratively using the Frank–Wolfe algorithm. The Frank–Wolfe algorithm may, however, be extended to solve elastic demand problems directly ([423, 697, 715, 362, 226, 716, 717, 412, 414, 603]).

Assuming that each demand function g_{pq} is lower and upper bounded (by l_{pq} and u_{pq}, respectively), given a tentative flow and demand, $(\mathbf{f}^k, \mathbf{d}^k)$, the Frank–Wolfe subproblem

yields an auxiliary demand

$$z_{pq}^k = \begin{cases} l_{pq}, & \text{if } \pi_{pq}^k > g_{pq}^{-1}(d_{pq}^k), \\ d_{pq}^k, & \text{if } \pi_{pq}^k = g_{pq}^{-1}(d_{pq}^k), \\ u_{pq}, & \text{if } \pi_{pq}^k < g_{pq}^{-1}(d_{pq}^k), \end{cases} \qquad \forall (p,q) \in \mathcal{C}, \tag{4.70}$$

where π_{pq}^k is the shortest route cost at the flow $\mathbf{f}^k$. The flow demands are assigned to the shortest routes, and a line search is made simultaneously in the spaces of flows and demands.

Essentially the same algorithm would result from an application of the Frank–Wolfe algorithm to the fixed demand reformulations of Section 2.2.4; such applications are discussed in [697, 226, 414]. Ferland [324] applies a Zoutendijk-type algorithm ([555]) to a minimum-cost flow circulation reformulation similar to that of Murchland [697].

The elastic demand version of the Frank–Wolfe algorithm thus consists of the same steps as those in the fixed demand case, and has roughly the same convergence behaviour. Nguyen [716] compares the algorithm to the convex simplex approach of [715, 717], and shows that it is inferior with respect to solution time.

It is apparent from the form of the subproblem Solution (4.70) that the demand updates are unstable, since the subproblem solution oscillates between the lower and upper bounds, depending on the *circulation cost* $\pi_{pq}^k - g_{pq}^{-1}(d_{pq}^k)$. This oscillating behaviour is observed in [443, 604]; to reduce the oscillating behaviour, LeBlanc and Farhangian [604] suggest an iterative updating of the lower and upper bounds on the demand function.

The oscillating behaviour originates in the linear approximation of the demand function. A possible means to avoid this unwanted property would therefore be to linearize only the part of the objective that corresponds to the original link flow variables. In other words, the linear Frank–Wolfe subproblem objective is replaced by

$$T^k(\mathbf{f}, \mathbf{d}) = T(\mathbf{f}^k, \mathbf{d}^k) + \sum_{a \in \mathcal{A}} t_a(f_a^k)(f_a - f_a^k) - \sum_{(p,q) \in \mathcal{C}} \int_0^{d_{pq}} g_{pq}^{-1}(s) ds.$$

The function T^k is obtained from a partial linearization of T, and may be derived from the general partial linearization algorithm by the choice $\varphi^k(\mathbf{d}) = -\sum_{(p,q) \in \mathcal{C}} \int_0^{d_{pq}} g_{pq}^{-1}(s) ds$ for all k. This special choice of partial linearization algorithm is given by Evans [304, 305, 306] (who actually uses the term partial linearization to describe it). The algorithm is applied to a combined trip distribution and assignment model, which may be viewed as a special elastic demand problem with additional marginal total constraints.

Each subproblem separates into $|\mathcal{C}|$ shortest route and strictly convex single-variable problems. The shortest route subproblems are solved given the link costs $\mathbf{t}(\mathbf{f}^k)$; the auxiliary demands are then calculated from the demand function given the shortest route costs π_{pq}^k, $(p,q) \in \mathcal{C}$, i.e., they are given by $\max\{0, g_{pq}(\pi_{pq}^k)\}$, $(p,q) \in \mathcal{C}$. The auxiliary link flows are obtained by assigning this demand to the shortest routes, and a simultaneous line search in the link flow and demand space then yields the next iterate. (Note that the calculations involved in Evans' algorithm are the same as those in the dual algorithm of Section 4.3.7 as applied to [TAP-E], with the only exceptions being that the travel costs are given by dual variable values and that the line search is replaced by a simple flow and demand averaging.) Since the demand functions are not linearized, the auxiliary demands will not oscillate as in the Frank–Wolfe approach; since each iteration requires essentially the same amount of work, the overall efficiency should therefore be much better. This is indeed observed (e.g., [373, 603]). In addition, the solution to each subproblem provides a lower bound on the optimal value, as does in the Frank–Wolfe algorithm ([304, 305, 306]); it can in fact, be proved from the strict convexity of the demand part

of the objective of [TAP-E], that given a feasible flow, the lower bound provided by the partially linearized subproblem is strictly better than that provided by the Frank–Wolfe subproblem (e.g., [583]). [The convergence of the link flows may however suffer from the zig-zagging behaviour of the Frank–Wolfe algorithm, and alternatives to the simple linearization of the link flow part of the objective should be considered.]

Florian and Nguyen [359] apply a generalized Benders decomposition scheme, embedded in a cyclic decomposition scheme over O-D pairs, to [TAP-E]; a basis for this scheme is the recognition that the demand variables are the complicating ones. The fixed demand Benders subproblems are solved using the column generation/reduced gradient method of Nguyen [713] (see Section 4.3.2). Here, the reoptimization capabilities, enabled by the disaggregated representation, are utilized in the sense that the solution to a previous Benders subproblem is scaled to yield a demand-feasible (and eventually near-optimal) solution to the next one. Their computational results are encouraging, but the comparative experiments with the convex simplex and Frank–Wolfe algorithms made by Nguyen [716] show that it may be quite slow, probably due to the need to solve many fixed demand subproblems.

4.5 Algorithms for stochastic assignment models

4.5.1 Stochastic network loading

The fundamental subproblem in the basic, deterministic, assignment model is that of finding an all-or-nothing solution given fixed travel costs; this problem may be solved using any method for solving shortest route problems. In the field of non-deterministic cost models this problem is known as the *stochastic network loading problem*, which is to find the probabilities with which a given route is chosen. The main difficulty with stochastic network loading is the fact that calculating perceived travel costs on competitive routes and the corresponding probabilities are prohibitive because of the vast number of possible routes in large-scale networks. To avoid route enumeration, stochastic network loading procedures are therefore based on link flows, the calculation of which often involves either a simulation process or a procedure for (implicitly) reducing the number of routes that may contribute to these link flows.

In order to calculate the route choice probabilities, the probability function of the perceived travel times on each route must be specified. The two models most frequently applied are the logit and probit models.

The logit model

Given actual route costs $\mathbf{c} = \mathbf{c}(\mathbf{h})$, the route choice probability in the logit model is

$$P_{pqr} = \frac{\mathrm{e}^{-\Theta c_{pqr}}}{\sum_{l \in \mathcal{R}_{pq}} \mathrm{e}^{-\Theta c_{pql}}}, \qquad \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}. \tag{4.71}$$

The corresponding route flow solution, obtained by letting $h_{pqr} = d_{pq} P_{pqr}$, $r \in \mathcal{R}_{pq}$, $(p,q) \in \mathcal{C}$, is characterized as the minimizer of

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \frac{1}{\Theta} h_{pqr} \log h_{pqr} + c_{pqr} h_{pqr} \tag{4.72}$$

over the Constraints (2.60b)–(2.60c) (cf. [336]). This objective is the result of partially linearizing the objective (2.60a) of [TAP-SUE-L], in the sense that the link flow part

is linearized ([221]); in the context of partial linearization, it corresponds to choosing $\varphi(\mathbf{h}) = (1/\Theta)\sum_{(p,q)\in\mathcal{C}}\sum_{r\in\mathcal{R}_{pq}} h_{pqr}\log h_{pqr}$. It is also easy to show that the optimal solution to this subproblem defines a lower bound on the optimal value of [TAP-SUE-L] and that the solution to the subproblem defines a descent direction with respect to the objective of [TAP-SUE-L].

The fact that the probabilities can be given a closed expression singles the logit model out among the probability models used. The constant Θ is a positive scaling factor used to remove the dependency of the units of travel time measurements, and should be calibrated for different applications. We see from the travel cost perception Formula (2.59) that a smaller value of Θ corresponds to a larger perception error, and, by (4.71), the distribution of flow among competitive routes will be equal in the limit of $\Theta \to 0$ regardless of the actual travel times.

From (4.71) we can also see that all routes will receive a positive flow in the stochastic network loading, regardless of their actual travel costs. If it is of interest to obtain the route flows resulting from the stochastic network loading, it becomes necessary to enumerate all the routes of the network or to extract a number of routes, $\hat{\mathcal{R}}_{pq} \subseteq \mathcal{R}_{pq}$, $(p,q) \in \mathcal{C}$, according to some specified criterion, and to apply Formula (4.71) to the given subsets. The need to enumerate the routes is one main difference from the all-or-nothing assignment problem in the deterministic case.

If a rough link flow solution to the stochastic network loading problem is sufficient, then by (implicitly) restricting the number of routes used to fulfill some regularity conditions it is possible to perform the network loading directly on the links, and the resulting algorithm is not much more complicated than performing an all-or-nothing assignment. If, however, all routes are to be included, the resulting link flows from the stochastic network loading may also be calculated (arbitrarily accurately) by performing simulations.

The most well known method for stochastic network loading is Dial's [254] link-flow based method. The method implicitly defines the subsets $\hat{\mathcal{R}}_{pq}$, $(p,q) \in \mathcal{C}$, of the routes of the network, and assigns link flows according to Formula (4.71) to them. The restricted route sets considered by the algorithm are given by the following definition.

Definition 4.1 (Efficient route) *Let $(p,q) \in \mathcal{C}$ and for each node $i \in \mathcal{N}$ let p_i denote the shortest route cost from p to i, and q_i the shortest route cost from i to q. Then a route $r \in \mathcal{R}_{pq}$ is efficient if and only if for all links (i,j) belonging to route r, $p_i < p_j$ and $q_i > q_j$ holds.*

In other words, a route is efficient if by traversing any link of the route a traveller comes further away from the origin and closer to the destination. Note that the set of efficient routes is dependent on the prevailing actual travel costs.

Dial [254] shows that a stochastic network loading can be performed in terms of link flows on a network defined by this implicit restriction of the routes with a computational effort equivalent to two all-or-nothing assignments, and devises the STOCH algorithm for its solution. By slightly redefining the concept of efficiency, he shows how such an assignment can be performed in a computational time equivalent to one all-or-nothing assignment only. The assignments obtained by either one of these algorithms will of course not be the link flow solution of a true stochastic network loading of the original network; Dial argues, however, that the number of travellers that would choose non-efficient routes is small enough to justify the use of the model.

The STOCH algorithm is incorporated into the UMTA Transportation Planning System ([924]).

Sheffi [831, Sec. 11.1] remarks that the first version of the STOCH algorithm is not efficiently implementable together with tree-building shortest route methods, since each O-D pair must be considered separately. The second version (the single-pass algorithm) is, however, implementable together with such algorithms.

The probit model

In the probit model, the perceived route travel costs are normally distributed. An important consideration for the practical use of probit models is the fact that it is possible to derive the route travel cost perception distribution from the probability distribution of the perceived link travel costs, and therefore a Monte Carlo simulation of the perceived travel costs can be based on link flows.

The simulation technique, which is due to Sheffi and Powell [832], is quite simple. The density function of the perceived link travel cost of each link is sampled once. This results in a set of realizations of perceived travel costs which is then used in an all-or-nothing assignment. The process of sampling and assignment is repeated, and the tentative link flow solution is defined as the average of the individual assignments. The process is terminated either after a fixed number of iterations or when the variance of the average link flows is small enough.

The above simulation process is applicable to the stochastic network loading problem phase in any stochastic user equilibrium problem of the form [TAP-SUE] ([832, 831]).

4.5.2 Stochastic user equilibrium

A Frank–Wolfe type algorithm for [TAP-SUE] would involve **(a)** a stochastic network loading based on the given flow and the corresponding actual travel costs and the given perception distribution, and **(b)** an update of the current flow towards the flow resulting from the network loading. Step **(b)** of this algorithm can not be performed through a line search in the non-deterministic case, for two reasons. Firstly, in general stochastic network loading can not be done exactly, and the resulting direction, which is a random variable, is only a descent direction on the average. Secondly, it is in general very difficult to evaluate the objective of [TAP-SUE] since it requires a route enumeration.

Standard methods applied to [TAP-SUE] and its special cases are therefore mainly based on taking predetermined steps in the directions defined by the stochastic network loading.

The logit model

The first algorithm proposed for the solution of [TAP-SUE-L] is the incremental assignment type method of Dial [254]. Portions of the demand are assigned iteratively to the network based on the current travel costs, according to the output of the STOCH algorithm. Florian [346] suggests updating the costs in the process of Dial's algorithm to take congestion into account.

The method of successive averages (MSA) (see Sections 1.5.4 and 4.1.6) has been applied to [TAP-SUE] in several versions. In the MSA algorithm, a search direction is obtained through a stochastic network loading, and the step taken towards that solution corresponds to taking the average of all the previously generated solutions, i.e., the step length in iteration k is $1/k$. The MSA algorithm is discussed in [218, 764, 219]; Tobin [898] suggests applying Dial's algorithm in place of a stochastic network loading. A dual algorithm for [TAP] (see Section 4.3.7) is extended to the solution of [TAP-SUE-L] by Larsson

et al. [582]; the main difference of the above algorithms lies in the usage of dual variable values as the actual travel costs.

Chen and Alfa [165] suggest improvements to the MSA algorithm, where the predetermined weighting is replaced by a line search with respect to either the deterministic part of T or to a restricted form of the entire objective. This latter line search is performed in terms of link flows through the use of a pseudo-inverse of the link-route incidence matrix; this approach is not applicable to large networks, since it requires the enumeration of the routes, and, furthermore, may result in inconsistent flows ([55]).

The above algorithms all operate in the space of link flows. It can be argued that since all routes are utilized in the stochastic user equilibrium solution, it is impractical to (approximately) solve the problem in terms of route flows. There are, however, motives for developing such algorithms. Firstly, it is difficult to investigate the amount of overlap present in an equilibrium solution if the routes are not available explicitly. It is also of interest to obtain explicit route flow information, for instance in the evaluation of route guidance systems ([940, 96, 221]). Secondly, if subsets of the total set of routes in the network are generated algorithmically by some column generation approach, the user can limit the number of routes that are generated within the method, and what is more, control the amount of overlap present in the network in a rather straightforward manner ([221]) by dropping routes that overlap previously generated routes more than a maximal allowed measure.

A descent algorithm based on a column generation approach is given by Damberg *et al.* [221]. Routes are generated through shortest route calculations based either on actual costs or on a single drawing of perceived travel costs. The algorithm may be thought of as an extension of the MSA approaches, where instead of averaging the all-or-nothing solutions, the routes generated are stored and utilized in a restricted master problem. Each restricted master problem is solved using the restriction of the stochastic network loading Formula (4.71) to the subsets $\hat{\mathcal{R}}_{pq}$ in order to generate descent directions and lower bounds on the optimal value of the restricted master problem; a line search is then made in the direction obtained. The resulting route flow solution solves the restriction of [TAP-SUE-L] to the known subsets of the routes. Damberg *et al.* also show that the problem of overlap can be managed effectively within the method; this is done by introducing column dropping rules based on different measures of overlap. They present computational results for the network of Winnipeg.

Bell *et al.* [55] improve on the algorithms of Chen and Alfa [165] by introducing both a route generation process and a balancing scheme in order to ensure that link and route flows are consistent.

The probit model

The only practical approach to the solution of the probit model is the application of an MSA type approach ([213, 832, 679, 831]). The question of the proper number of drawings in the inner loop in each main iteration is addressed by Sheffi [831, Sec. 12.3] (see also Mimis [679]); the conclusion is that the overall best result is obtained by a streamlined approach where a single drawing of the perceived travel costs is made in each main iteration. This approach is extended to an asymmetric stochastic user equilibrium model in [831]. For further reading on stochastic user equilibrium models and methods, see, e.g., [214, 831, 728, 894].

4.6 Algorithms for side constrained assignment models

The side constrained assignment model [TAP-SC] introduced in Section 2.8.2 constitutes a generalization and an improvement over the basic assignment model. As was demonstrated in that section, very little study has been devoted to side constrained models however, either from the modelling point of view or computationally. The only side constrained assignment model computationally studied is the capacity side constrained model, which we discuss below.

4.6.1 Algorithms for capacity side constrained assignment models

Consider the link capacity side constrained assignment model

[TAP-C]

$$\min\ T(\mathbf{f}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \int_0^{f_a} t_a(s)ds, \tag{4.73a}$$

subject to

$$\sum_{r\in\mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q)\in\mathcal{C}, \tag{4.73b}$$

$$h_{pqr} \geq 0, \qquad \forall r\in\mathcal{R}_{pq},\ \forall (p,q)\in\mathcal{C}, \tag{4.73c}$$

$$\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a\in\mathcal{A}, \tag{4.73d}$$

$$f_a \leq u_a, \qquad \forall a\in\mathcal{A}, \tag{4.73e}$$

where $u_a \in [0,+\infty]$ is the upper bound on the flow of link $a\in\mathcal{A}$.

From a modelling point of view, it is preferable to use explicit upper bounds than to use link travel cost functions with asymptotes at their respective bounds. A disadvantage of imposing explicit link capacities is that they destroy the Cartesian product structure of the uncapacitated problem, and thus make the problem more demanding computationally. In particular, the linear subproblems of the Frank–Wolfe and simplicial decomposition type methods will become linear multicommodity minimum cost network flow problems ([557]), which are computationally burdensome. Under strong assumptions on the travel time functions and the choice of the initial solution, the multicommodity flow subproblem of the Frank–Wolfe method may be relaxed into shortest route subproblems while maintaining convergence to an optimal flow pattern ([211, 212, 485]).

Computationally, the asymptotic travel time functions have the disadvantage that they may result in numerical difficulties. In addition, whenever the problem is solved by a feasible-direction algorithm (e.g., the Frank–Wolfe method), these travel time functions make it necessary to initialize the algorithm through the calculation of a flow pattern which is strictly feasible with respect to the implicit upper bounds on the link flows ([212]); this task is however non-trivial (e.g., [516]).

Solution methods proposed for [TAP-C] are often based on the recognition of the fact that linear multicommodity flow subproblems ([557]) are prohibitively expensive to solve repeatedly, and may be divided into two categories. In the first, attempts are made to use shortest route subproblems to generate search directions. In the second approach,

the capacitated problem is converted into a sequence of uncapacitated problems through a penalization/dualization of the capacity Constraints (4.73e), so that efficient methods for [TAP] may be applied for the solution of [TAP-C]. We outline below these two lines of development.

Frank–Wolfe type algorithms

In the first approach to [TAP-C], attempts are made to use shortest route subproblems to generate search directions. The algorithm is initialized at an inner point with respect to the link capacities, and to ensure convergence the travel cost functions must satisfy the *coercivity* condition

$$\lim_{f_a \to u_a} \int_0^{f_a} t_a(s)ds = +\infty, \qquad \forall a \in \mathcal{A}, \tag{4.74}$$

which effectively reduces the problem to an uncapacitated problem with asymptotic cost functions ([211, 212]). Hearn and Ribera [485] instead assume that the sequence $\{l_k\}$ of step lengths is bounded from below by some positive number. One sufficient condition for this assumption to be fulfilled is that the initial point is strictly better (in terms of the objective value) than any feasible solution at which some capacity constraint is active; the existence of such an initial point is not guaranteed for the travel time formulas most often used, but is however implied by Condition (4.74). One possible way to ensure convergence when using general travel time formulas is to invoke a Frank–Wolfe subproblem (a linear multicommodity flow problem) whenever the shortest route solution does not yield a sufficient progress (i.e., when a step length l_k falls below some prespecified parameter $\underline{l} > 0$); see the dissertation by Stefek [872].

Stefek's main theme is the development of simplicial decomposition type algorithms for the capacitated problem. In these algorithms, the line search step of his Frank–Wolfe type method is replaced by a multi-dimensional search over the intersection of the convex hull of the hitherto generated subproblem solutions and the set defined by the capacity constraints. The safe-guarding strategy of the Frank–Wolfe type algorithm is also used in these methods; whenever the extreme points corresponding to the shortest route patterns do not provide sufficient descent in the master problem, a linear multicommodity flow subproblem is invoked. (In a direct application of simplicial decomposition, subproblems would always be multicommodity flow problems and the master problem would not include the capacity constraints; such a scheme would not, however, be efficient, because of the high computational cost of the subproblems.) Stefek also presents a variation in which Lagrange multipliers for the capacity constraints of the master problem are used to price-out those constraints in the subproblem, thereby reducing the number of iterations in which a multicommodity flow subproblem has to be invoked. Computational experiments with three medium- and large-scale problems show that for lightly capacitated problems these extensions of the simplicial decomposition principle are superior to a straightforward application of this principle (where the multicommodity flow subproblems are solved by a Dantzig–Wolfe decomposition, i.e., a column generation, approach), but that they are inferior for heavily capacitated ones.

Dualization/penalization algorithms

In the second approach, the capacitated problem is converted into a sequence of uncapacitated problems through a penalization/dualization of the capacity Constraints (4.73e), so that efficient methods for [TAP] may be applied for the solution of [TAP-C]. (Of course,

[TAP-C] may be relaxed in alternative ways; in [476, 582], the definitional constraints (4.73d) are Lagrangean dualized, and in [516], all constraints but (4.73d) are augmented Lagrangean dualized.)

For the case of constant travel times, Jorgensen [533] suggests applying the Dantzig–Wolfe decomposition method (which may be interpreted as a cutting plane method applied to a dual problem), but does not give any computational results. For the case of flow-dependent travel times, he suggests using approximating piecewise constant travel time functions; the approximate problem may then be restated as a problem with constant travel times in an enlarged network. Miller *et al.* [678] present a column generation approach for the case of constant travel costs, in which the restricted master problems are solved using a generalized upper bounding technique (e.g., [590]).

Letting $U = \{\, \mathbf{f} \in \Re^{|\mathcal{A}|} \mid g_a(f_a) \stackrel{\text{def}}{=} f_a - u_a \leq 0,\ \forall a \in \mathcal{A} \,\}$, the feasible set of [TAP-C] is $F^r \cap U$. In an *exterior penalty method* (e.g., [329]) for [TAP-C], the Constraints (4.73e) are included in an extended objective function by means of a penalty function $P : \Re^{|\mathcal{A}|} \mapsto \Re$ satisfying

(1) $P(\mathbf{f}) \geq 0$ for all $\mathbf{f} \in F^r$,

(2) $P(\mathbf{f}) = 0$ if and only if $\mathbf{f} \in F^r \cap U$,

(3) P is continuous on F^r.

An example of such a penalty function is

$$P(\mathbf{f}) = \sum_{a \in \mathcal{A}} p_a(f_a), \tag{4.75a}$$

where

$$p_a(f_a) = r_a [g_a(f_a)]_+^{m_a} = r_a \max\,\{0, g_a(f_a)\}^{m_a},\ r_a > 0,\ m_a \geq 2. \tag{4.75b}$$

By introducing a penalty parameter $c > 0$, the penalized objective

$$P_c(\mathbf{f}) = T(\mathbf{f}) + cP(\mathbf{f}),$$

the penalty subproblem

$$P_c = \min_{\mathbf{f} \in F^r} P_c(\mathbf{f}),$$

which amounts to solving an uncapacitated traffic assignment problem, and its solution

$$\mathbf{f}(c) = \arg\min_{\mathbf{f} \in F^r} P_c(\mathbf{f}),$$

one may show that

(1) $P_c \leq T(\mathbf{f}^*)$, for all $c > 0$, and

(2) $\lim_{c \to +\infty} \mathbf{f}(c) = \mathbf{f}^*$.

For a differentiable and separable penalty function, like (4.75), optimal Lagrange multipliers for the penalized constraints may be estimated using the result (e.g., [473])

$$\lim_{c \to +\infty} c \left. \frac{\mathrm{d}p_a}{\mathrm{d}f_a} \right|_{f_a = f_a(c)} = \beta_a, \qquad \forall a \in \mathcal{A}.$$

Hearn [473] proposes to include the explicit link flow capacities in an extended objective function by means of an exterior penalty function of the form (4.75), thereby obtaining an uncapacitated traffic assignment subproblem (which is solved by the Frank–Wolfe method). The behaviour of the overall penalty method is illustrated through small-size numerical examples. Inouye [516] applies an interior penalty method in which the subproblems are solved using the Frank–Wolfe method, and presents results for a small example.

In order to avoid the ill-conditioning inherent in the penalty approach, one may introduce a Lagrangean term in the extended objective, thus creating an *augmented Lagrangean function* ([491, 763, 781, 66, 70]). Letting $\boldsymbol{\mu}$ denote the vector of Lagrange multipliers for the dualized constraints and using the penalty Function (4.75) with $r_a = 1/2$ and $m_a = 2$ for all $a \in \mathcal{A}$, the augmented Lagrangean function becomes ([780])

$$L_c(\mathbf{f}, \boldsymbol{\mu}) = T(\mathbf{f}) + \sum_{a \in \mathcal{A}} \overline{p}_a(f_a, \mu_a, c),$$

where

$$\overline{p}_a(f_a, \mu_a, c) = \frac{1}{2c}([\mu_a + c g_a(f_a)]_+^2 - \mu_a^2).$$

Defining the augmented Lagrangean dual objective function through the solution of the uncapacitated traffic assignment subproblem

$$L_c(\boldsymbol{\mu}) \;=\; \min_{\mathbf{f} \in F^r} L_c(\mathbf{f}, \boldsymbol{\mu})$$

and denoting the subproblem solution with

$$\mathbf{f}(\boldsymbol{\mu}, c) \;=\; \arg\min_{\mathbf{f} \in F^r} L_c(\mathbf{f}, \boldsymbol{\mu}),$$

we have that for any $c \geq 0$ ([780]),

(1) $L_c(\boldsymbol{\mu}) \leq L_c(\boldsymbol{\beta}) = T(\mathbf{f}^*)$ for all $\boldsymbol{\mu} \geq \mathbf{0}$,

(2) $\lim_{\boldsymbol{\mu} \to \boldsymbol{\beta}} \mathbf{f}(\boldsymbol{\mu}, c) = \mathbf{f}(\boldsymbol{\beta}, c) = \mathbf{f}^*$.

Hence, the augmented Lagrangean dual objective function is, for any $c \geq 0$, maximized by arbitrary optimal values of the Lagrangean multipliers, and the optimal flow pattern may be obtained for finite values of the penalty parameter. Moreover, although the flow pattern $\mathbf{f}(\boldsymbol{\mu}, c)$ is in general infeasible in [TAP-C] unless $\boldsymbol{\mu} = \boldsymbol{\beta}$, it will become near-feasible for near-optimal values of the multipliers.

The choice $c = 0$, which gives the ordinary Lagrangean dualization scheme, is feasible because of the strict convexity of T; see, e.g., the discussion following Theorem 6.5.1 in [43]. In general, however, the augmented Lagrangean schemes have superior convergence characteristics, and from now on, we thus presume that $c > 0$.

Optimal multipliers may be found by solving the augmented Lagrangean dual problem

$$\max_{\boldsymbol{\mu}} \; L_c(\boldsymbol{\mu}),$$

where L_c is concave and differentiable, with

$$\frac{\partial L_c(\boldsymbol{\mu})}{\partial \mu_a} = \max\{g_a(f_a(\boldsymbol{\mu}, c)), -\frac{\mu_a}{c}\}, \qquad \forall a \in \mathcal{A}.$$

A steepest-ascent multiplier update with step length c yields (see [70, p. 162])

$$\mu_a := [\mu_a + c g_a(f_a(\boldsymbol{\mu}, c))]_+, \qquad \forall a \in \mathcal{A};$$

if c is sufficiently small, then the value of L_c will ascend. (One may also show ([70, Prop. 5.8]) that if $\boldsymbol{\mu}$ is sufficiently close to an optimal dual solution, the value of the Lagrangean dual function, L_0, will also ascend.)

Although convergence is ensured for any positive value of c, a good practical performance demands for a careful choice (e.g., [492, 70]). In particular, there is a trade-off between a high rate of convergence in the multiplier space and the degree of ill-conditioning of the Lagrangean subproblem; see [623, Chap. 13]). Usually, the parameter c is initially given a low value, and then increased whenever a measure of the total infeasibility in the dualized constraints does not improve sufficiently rapidly (e.g., [763]). We thus introduce a non-decreasing sequence $\{c_k\}$ of positive penalty parameters, and define a sequence of primal-dual iterates through the formulas

$$\mathbf{f}^k = \mathbf{f}(\boldsymbol{\mu}^k, c_k), \tag{4.76a}$$

$$\mu_a^{k+1} = [\mu_a^k + c_k g_a(f_a^k)]_+, \qquad \forall a \in \mathcal{A}, \tag{4.76b}$$

where $k = 1, 2, \ldots$, and with $\boldsymbol{\mu}^1$ being some initial guess.

Vanderstraeten-Tilquin [944], Hearn and Ribera [484], Polak [758], and Larsson and Patriksson [587] all employ iterative augmented Lagrangean schemes. In Vanderstraeten-Tilquin's scheme, the uncapacitated subproblems are solved by the application of a nonlinear version of the out-of-kilter method to single-commodity problems obtained in a cyclic decomposition manner. In the scheme of Hearn and Ribera, the subproblems are solved by the Frank–Wolfe method. They consider two types of augmented Lagrangean functions and apply one of them to a small numerical example. Vanderstraeten-Tilquin also gives two other solution principles for the capacitated problem. The first is a subgradient optimization procedure for finding optimal allocations of the total link capacities to the separate commodities; this is essentially the same algorithm as the one for linear multicommodity network flows proposed by Kennington and Shalaby [547]. The second involves the solution of a sequence of lower-dimensional subproblems obtained through partitionings of variables and relaxations of nonnegativity constraints (see also, e.g., [590, Chap. 5]). From some experimentation with small-scale test problems, Vanderstraeten-Tilquin concludes that the latter method is unfeasible for larger problems, and that the augmented Lagrangean scheme is the most viable of the two others, at least in the absence of an *a priori* knowledge of a good estimate of the optimal objective value. Larsson and Patriksson [587] apply the disaggregate simplicial decomposition (DSD) algorithm (see [586] and Section 4.3.5) to each traffic assignment subproblem; in comparisons with ordinary Lagrangean dualization and a penalty approach, the augmented Lagrangean approach (which may be viewed as a combination of them) is clearly superior. The efficiency and very good reoptimization capabilities of the DSD algorithm motivated its choice for use in solving the sequence of uncapacitated traffic assignment subproblems; indeed, the computational effort needed for solving the subproblems was observed to decrease significantly for every iteration of the augmented Lagrangean method.

Because of the dual character of augmented Lagrangean schemes, feasible solutions to the original problem will generally be found in the limit only, even though the primal solutions' infeasibilities will in later iterations be small. Larsson and Patriksson [587] therefore introduce a procedure, of the type described in Section 4.3.2, which heuristically constructs feasible solutions by carefully manipulating the (slightly) infeasible solutions to the augmented Lagrangean subproblems; this is done by repeatedly shifting flow from

a route in an origin-destination pair utilizing over-saturated links to routes within the same pair that are strictly feasible with respect to the capacities. They also construct an advanced starting solution for the dual algorithm. Together with the efficiency of the DSD algorithm and the good performance of the feasibility heuristic, they are able to conclude that the introduction of link capacities increased the computing times by no more than a factor of four.

As stated in Section 2.8.2, the optimal Lagrange multipliers for the capacity constraints may be seen as link tolls which, when imposed upon the travellers, yield an uncapacitated user equilibrium traffic flow pattern that fulfills the link capacities. The iterative search Procedure (4.76) may thus be interpreted as a mathematical simulation of a real-life process in which a traffic engineer attempts to limit link flows by introducing link tolls and modifying them until the travellers' behavioural response is the intended one. Moreover, the traffic engineer employs the very natural strategy of modifying the link tolls in proportion to the violations of the link flow limitations that he/she is trying to impose. (Of course, this strategy for finding suitable link tolls can not be implemented in the real-life traffic system.) It is also possible to show (see [587]) that under some additional assumptions on the way in which this dual search procedure is carried out, the sequence $\{\boldsymbol{\mu}^k\}$ converges to the vector $\boldsymbol{\beta}$ of multipliers of minimum Euclidean norm. The simulation of the traffic engineer's strategy thus automatically yields the minimal link tolls. A similar nice interpretation is obtained by instead viewing the multipliers $\boldsymbol{\beta}$ as equilibrium queueing delays.

We finally note that the above augmented Lagrangean approach is applicable to the general side constrained model [TAP-SC].

4.7 Discussion

In this chapter we have given a unified description of methods proposed for the solution of the basic user equilibrium problem and some of its extensions. With the Frank–Wolfe algorithm as the starting point, we outlined the development made based on the concepts of partial linearization, decomposition and column generation.

The decomposition and column generation methods presented are not only very similar (nearly all of them are based on a block Gauss–Seidel iteration) but some of them have been rediscovered several years after their first publication. That the methods were rediscovered is not surprising, since the methods are intuitively natural; another reason is that the methods were developed in two different areas of nonlinear network optimization (equilibrium in traffic networks and optimal routing in computer communication networks), between which little communication took place for a number of years. It is still surprising though, considering the availability of nonlinear programming methods in the literature, that the development of methods for traffic assignment exhibits such a one-sidedness.

The comparative efficiency of algorithms for traffic assignment have unfortunately not been thoroughly examined. The few studies which exist (e.g., [716, 717, 498, 439]) include only limited comparisons of a few algorithms, and due to the enormous development of computer technology and the increasing size of traffic problems routinely being solved, the conclusions drawn are not necessarily valid any more. This is particularly true with regard to the conclusions made about methods based on route generation, which were considered impractical until only a few years ago.

We do not believe that there is one best algorithm for traffic assignment. The algorithm

to recommend depends on many factors:

(1) (*Computer facilities available*) The internal memory capacity, and also the speed of the processor(s), determine which algorithms it is possible to apply to an assignment model of a given size. Considering multiprocessor systems, very little experience of the use of such algorithms for traffic equilibrium problems is available in the open literature; we believe, however, that such implementations should be considered closely, since the inherent structure of the model makes it possible to utilize different levels of parallelism, and the growing need for real-time solutions of traffic assignment problems, for instance in the use of traffic management and route guidance systems, makes it necessary to explore high speed computations of equilibrium solutions.

The availability of subroutines for the efficient solution of shortest route problems, quadratic network flow problems, etc., could also naturally lead to the consideration of certain types of algorithms.

(2) (*Size of problems*) The algorithms that it is possible to use are also determined by the size of the problem to be solved. Several algorithms proposed for traffic assignment in the past were believed to be efficient because of results based on tests performed on very small networks; for the solution of a large-scale model, however, the recommendable algorithm must utilize the problem structure.

(3) (*A priori information available*) Any *a priori* information available about the solution to a traffic assignment problem should of course be utilized. Most types of information, such as estimates of travel times and link flows, naturally lead to the consideration of dual algorithms, in particular if the information contains measurement errors and are inconsistent, since they make it difficult to immediately utilize the information in a primal algorithm.

(4) (*Information required*)

(a) (*Travel costs*) If equilibrium travel costs are sought, then the inverse (dual) problem should be addressed, for which simple algorithms are available.

(b) (*Link flows*) If equilibrium link flows are sought, then any of the algorithms presented in the chapter are applicable. If commodity link flows are sought, however, some decomposition scheme is to be preferred. The equilibrium commodity link flows are not unique, and the result is therefore greatly influenced by the algorithm chosen.

(c) (*Route flows*) If an equilibrium route flow solution is sought, then a disaggregate simplicial decomposition/column generation scheme should be used. As in the case of commodity link flows, the solution obtained depends on the choice of algorithm.

(5) (*Accuracy required*) If a rough solution is sufficient, then there is little to gain in using second-order methods, since, for instance, the Frank–Wolfe algorithm may reach a sufficiently accurate solution just as quickly. If, however, for some reason, a very accurate solution is required, then a second-order method should replace a method with a lower convergence rate when approaching the solution. It should, however, be noted that a very high accuracy is of little value if—which is often the case—the input data is inaccurate.

(6) (*Reoptimization capabilities*) If there is a need to repeatedly solve a traffic assignment problem with slightly varying data, for instance when traffic assignment problems arise as subproblems in more complex models, then the algorithm used must be able

to utilize the optimal solution to a previous problem as an advanced start when a perturbed one is to be solved; this becomes a crucial point when real-time applications are considered. From the discussions in Section 4.3.6, we conclude that the more information that an algorithm keeps about the problem, the more effective it can be for reoptimization purposes.

(7) (*Extension capabilities*) At first sight, it seems a good idea to make an implementation amenable to an easy extension to more complex models. There is, however, the danger that the algorithm may be used on models that it is not meant to solve or, even worse, in theory can not possibly solve; the user of a system must be made very much aware of the limitations of the given model and method.

Since the circumstances under which assignments are to be carried out may vary, mobility may also become important. (This does not favour very sophisticated implementations.)

A few algorithms are available for public use; the RSDNET code ([482]) is perhaps the most recent example. For the benefit of the field, a library of both state-of-the-art codes and test networks should be set up in the near future.

We conclude by supplying a list of references to known test networks.

City	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{C}\rvert$	# Centroids	Reference
	4	10	12		[2]
	9	13	4	4	[719]
	9	18	4	4	[40]
	9	24	10		[724]
	9	36	12		[871]
	20	28	2		[699]
	14	22	23		[428]
	61	148	122		[428]
Sioux Falls	24	76	528	24	[607]
Hull	155	376	690	27	[714]
Hull	501	798	142	23	[353]
Dallas	584	1462		55	[605]
Winnipeg	1035	2789		140	[360]
Winnipeg	1052	2836	4344	147	[353]
Barcelona	1020	2522	7922		[39]
Barcelona	930	2522	7922	110	[684]
Leeds	1352	3756		589	[941]
Du Page County	9400	29000		999	[288]
Madrid	3201	8659	26037	556	[684]

Table 4.1: Test networks

Chapter 5

Algorithms for general traffic equilibria

5.1 Introduction

In the general formulation of traffic equilibrium models, the assumption of the separability of the travel cost and demand functions is relaxed. As a result, a solution to the Wardrop user equilibrium conditions can not be found by solving a convex program of the form [TAP] or [TAP-E]. Instead, they are transformed into a variational inequality, a nonlinear complementarity or a fixed point problem, for which standard algorithms are applied.

A main difference between this and algorithms for separable traffic equilibrium problems is that here no merit function is directly available for monitoring the convergence. Convergence is instead often based on a guaranteed monotone decrease of an *artificial* merit function, such as the (unknown) Euclidean distance to the equilibrium solution, which it is not possible to evaluate and utilize in termination criteria or in line search procedures for the acceleration of the convergence.

In general, in order to establish convergence theoretically, these algorithms also require stronger monotonicity assumptions on the travel cost and demand functions than they do in the separable case, and frequently knowledge of the values of certain parameters of the model which are difficult to estimate.

The need for these strong assumptions to hold in order to guarantee the convergence of algorithms is very unfortunate, in view of the fact that a traffic equilibrium model may fail to satisfy them, as shown by Heydecker [493].

Recently, reformulations of variational inequalities such as, in general nonconvex, nonlinear programs have been shown to naturally lead to convergent descent algorithms; in such methods, an asymmetric model is supplied with a specially constructed merit function which is utilized both in line searches and for monitoring the convergence. (These merit functions are outlined in Section 3.1.5.) Some of these algorithms are theoretically convergent under weaker conditions on the problem data than the traditional ones, and the introduction of the line search may also lead to a higher practical convergence rate. We will show that these algorithms may be viewed as simple modifications of the traditional approaches where a predetermined step length (usually a unit step) is replaced by a line search with respect to a merit function.

We shall concentrate our discussions on variational inequality formulations of the Wardrop conditions, since they are the predominant modelling basis.

Algorithms applied to general traffic equilibrium problems are extensions of the iterative algorithms for separable models that were presented in the previous chapter, and may

be given the same uniform description in terms of partial linearization, decomposition and column generation. In this chapter, we shall make such a unified presentation, and also show how, through a simple modification, traditional algorithms may be enforced to yield convergence under mild monotonicity assumptions.

5.2 Algorithm concepts

The three algorithm concepts for the solution of nonlinear programs introduced in Section 4.2 are here extended to the solution of monotone variational inequality problems. The concept of partial linearization is generalized to that of cost approximation, which was introduced by Patriksson [747], while those of decomposition and column generation are immediately applicable to variational inequality problems.

5.2.1 Cost approximation algorithms

The general algorithm

Consider the variational inequality problem of finding an $\mathbf{x}^* \in X$ such that

[VIP]

$$F(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x}-\mathbf{x}^*) \geq 0, \qquad \forall \mathbf{x} \in X, \tag{5.1}$$

where $X \subseteq \Re^n$ is a nonempty, closed and convex set, and $F : X \mapsto \Re^n$ is a continuous and monotone mapping on X. We let Ω denote the set of solutions to [VIP].

One iteration of the cost approximation algorithm consists of the following two main steps:

(1) Given a feasible point, a feasible search direction is defined through the (possibly inexact) solution of an approximation of the original problem, in which the mapping F is approximated by a monotone mapping.

(2) The direction defined by the solution to the above described subproblem is a feasible direction of descent with respect to a merit function whose minima coincide with the set of solutions to [VIP]. A (possibly inexact) line search is made with respect to this merit function in the direction obtained, and the resulting step length defines a new point with a reduced value of the merit function.

Formally, in iteration k we introduce a monotone *cost approximating mapping* $\Phi^k : X \mapsto \Re^n$. If, at $\mathbf{x}^k \in X$, the mapping F is replaced by the mapping Φ^k in [VIP], then the error made in the approximation obviously is $F - \Phi^k$. This error is taken into account by adding to Φ^k the fixed error term $F(\mathbf{x}^k) - \Phi^k(\mathbf{x}^k)$. [Alternatively, the approximation made can be seen as a fixation of the second term of an equivalent reformulation, $\Phi^k + [F - \Phi^k]$, of the original cost mapping, at $\mathbf{x}^k$.] Thus, we arrive at the variational inequality subproblem in which a point $\mathbf{y}^k \in X$ is sought such that

[$\mathrm{VIP}^k_{\Phi^k}$]

$$[\Phi^k(\mathbf{y}^k) + F(\mathbf{x}^k) - \Phi^k(\mathbf{x}^k)]^{\mathrm{T}}(\mathbf{y}-\mathbf{y}^k) \geq 0, \qquad \forall \mathbf{y} \in X, \tag{5.2}$$

We let $Y(\mathbf{x}^k)$ denote the set of solutions to [$\mathrm{VIP}^k_{\Phi^k}$].

By its construction, the value of the problem defining mapping of $[\mathrm{VIP}^k_{\Phi^k}]$ coincides with that of the original cost map at the point of approximation. This fact is important, since it provides a termination criterion for the algorithm: if $\mathbf{x}^k$ solves the subproblem $[\mathrm{VIP}^k_{\Phi^k}]$, defined at $\mathbf{x}^k$, then it immediately follows that $\mathbf{x}^k$ also solves the original problem [VIP]. (The converse is also true.)

If Φ^k is chosen as the gradient mapping of a continuously differentiable convex function $\varphi^k : X \mapsto \Re$, then the subproblem $[\mathrm{VIP}^k_{\Phi^k}]$ amounts to solving the convex minimization problem

$$\min_{\mathbf{y}\in X} \left\{\varphi^k(\mathbf{y}) + [F(\mathbf{x}^k) - \nabla\varphi^k(\mathbf{x}^k)]^{\mathrm{T}}\mathbf{y}\right\}. \tag{5.3}$$

Thus, by applying the cost approximation concept, [VIP] can be solved as a sequence of optimization problems. Indeed, in most of the iterative methods for the solution of [VIP] that we will identify as special cases from the class of cost approximation algorithms, subproblems of the form (5.3) are solved.

We then make two important observations. Firstly, the symmetric Subproblem (5.3) is equivalent to the inner problem of (3.18) which defines one class of merit functions for [VIP] (see Section 3.1.5), and hence symmetric cost approximations form the building block of descent algorithms for variational inequalities. Secondly, if F is the gradient mapping of a continuously differentiable convex function $T : X \mapsto \Re$, then the objective of the Subproblem (5.3) is equivalent to an approximation of T obtained by linearizing the second term of $\varphi^k + [T - \varphi^k]$ at $\mathbf{x}^k$, i.e., a partial linearization of T; hence, the concept of cost approximation is a direct extension of that of partial linearization from nonlinear programs to variational inequality problems.

We can not generally expect the original problem to be solved by $\mathbf{y}^k$. A new iteration point is therefore defined by taking a step in the direction of $\mathbf{y}^k - \mathbf{x}^k$ such that a merit function, $\psi : X \mapsto \Re \cup \{+\infty\}$, whose minima coincide with the set of solutions to [VIP], is decreased sufficiently. In symmetric models of traffic equilibria, the problem [VIP] corresponds to a mathematical program, i.e., the map F is the gradient of a function T; it is then natural to choose $\psi = T$ as the merit function. In asymmetric models, this is not the case, and another merit function must be identified; one natural merit function for [VIP] is inherent in the subproblem corresponding to (5.3) and can, under certain conditions, be utilized for the solution of this problem.

At the new point, the original mapping is again approximated—now perhaps with a different mapping Φ^{k+1}—and the algorithm proceeds until some stopping criterion is fulfilled.

Below, we summarize the different steps of the general algorithm.

A sequence $\{\Phi^k\}$ of monotone cost approximating mappings and a merit function ψ are assumed to be given. (Note that each mapping may be chosen adaptively, given $\mathbf{x}^k$.)

Step 0 (*Initial guess*) Choose an initial point $\mathbf{x}^0 \in X$, and let $k = 0$.

Step 1 (*Search direction generation*) Find a $\mathbf{y}^k \in X$ that solves $[\mathrm{VIP}^k_{\Phi^k}]$. The resulting search direction is $\mathbf{p}^k = \mathbf{y}^k - \mathbf{x}^k$.

Step 2 (*Convergence check*) If $\mathbf{x}^k$ solves $[\mathrm{VIP}^k_{\Phi^k}]$ $\rightarrow$ Stop ($\mathbf{x}^k$ solves [VIP]). Otherwise, continue.

Step 3 (*Line search*) Find a step length, l_k, which solves the one-dimensional problem

$$\min\{\psi(\mathbf{x}^k + l\mathbf{p}^k) \mid \mathbf{x}^k + l\mathbf{p}^k \in X,\ l \geq 0\}.$$

Step 4 (*Update*) Let $\mathbf{x}^{k+1} = \mathbf{x}^k + l_k \mathbf{p}^k$, and $k := k + 1$.

Step 5 (*Convergence check*) If $\mathbf{x}^k$ is acceptable as a solution $\rightarrow$ Stop. Otherwise, go to Step 1.

Although both the subproblem solution (Step 1) and the line search (Step 3) are performed exactly in the above description, under certain assumptions they can both be performed inexactly, while still ensuring convergence of the algorithm.

By appropriately choosing $\{\Phi^k\}$ and ψ, many well known algorithms for [VIP] can be identified as special cases from the class of cost approximation methods. A vast majority of the algorithms proposed for asymmetric traffic equilibria are included in this framework, and may therefore be given a unified description.

There is no obvious choice of merit function to be used in the line search in Step 3, due to the nonexistence of an unambiguous integral of the form (3.2b). In the traditional variational inequality methods that can be identified as special cases from the class of cost approximation algorithms, Step 3 is normally executed by taking a predetermined step in the direction of $\mathbf{p}^k$ in order to yield a decrease in an artificial merit function, i.e., a merit function which is known to exist but which it is not possible to evaluate. The convergence often relies on the monotone decrease of the Euclidean distance to the set Ω of solutions to [VIP], in which case the merit function is

$$\psi(\mathbf{x}) = d_\Omega(\mathbf{x}) \stackrel{\text{def}}{=} \inf_{\mathbf{y} \in \Omega} \|\mathbf{x} - \mathbf{y}\|, \tag{5.4}$$

where $\|\cdot\|$ is some appropriate vector norm. Clearly, a line search can not be made with respect to this merit function, since the set Ω is unknown. Convergence is ensured, however, by showing that the underlying algorithmic mapping, $\mathbf{x}^{k+1} \in A(\mathbf{x}^k)$, satisfies the two conditions:

(1) The algorithmic map is a fixed point map, i.e.,

$$\mathbf{x} \in \Omega \Longleftrightarrow \mathbf{x} \in A(\mathbf{x}), \tag{5.5}$$

(2) A is contractive with respect to the norm defined in (5.4) [see Definition A.2.c].

It will subsequently be shown that the majority of the methods applied to asymmetric traffic equilibrium problems can be identified by $A \equiv Y$ and $l_k \equiv 1$, where $\mathbf{x} \mapsto Y(\mathbf{x})$ is the map that defines the set of optimal solutions to [VIP$^k_{\Phi^k}$], that is, they can be seen as methods where the algorithmic map corresponds to a cost approximation subproblem followed by a unit step.

The validity of the fixed point Property (5.5) is immediate for the mapping $A \equiv Y$. (The corresponding result in the case where Y denotes a convex programming subproblem is given in Theorem 3.13.g.)

Algorithms in which $l_k \equiv 1$ is chosen are usually called *successive approximation algorithms*; the sequence of iterates in such algorithms are defined by a sequence of solutions to approximations of the original problem, where the original mapping F is replaced by monotone mappings F^k, $k = 0, 1, \ldots$. [In the case of cost approximation, the mapping F^k has the form $\Phi^k + F(\mathbf{x}^k) - \Phi^k(\mathbf{x}^k)$.] As discussed above, there are two main motives for considering cost approximation algorithms with step lengths chosen through line searches instead. First, convergence can generally be established under weaker conditions; second, introducing a line search may enhance both the theoretical and practical convergence rate of the algorithm.

The convergence of most of these algorithms have been analyzed both locally (in which case one assumes that the initial solution is chosen in a neighbourhood of the solution) and globally (when such an assumption is not made). Although the global convergence results are the most interesting, local convergence results are of interest in the convergence rate analysis and in the context of reoptimization.

In the next section, we present instances of the successive approximation version of the class of cost approximation algorithms.

Instances of successive cost approximation algorithms

The fact that cost approximation generalizes partial linearization implies that the extension to [VIP] of all the algorithms discussed in Section 4.2.1 are instances of the general cost approximation scheme; these extensions are obtained simply by replacing ∇T with the more general mapping F.

Traditionally, successive approximation algorithms for [VIP] are divided into linear and nonlinear approximation algorithms, distinguishing between affine and non-affine approximating mappings F^k.

Linear approximation algorithms

In linear approximation algorithms, given an iterate $\mathbf{x}^k \in X$, the next iterate is defined as the solution to an approximation of [VIP], where the original mapping F is replaced by an *affine* mapping F^k, i.e.,

$$F^k(\mathbf{x}) = F(\mathbf{x}^k) + (1/\gamma_k)\mathbf{B}_k(\mathbf{x} - \mathbf{x}^k), \qquad \forall \mathbf{x} \in X, \tag{5.6}$$

where $\gamma_k > 0$ and $\mathbf{B}_k \in \Re^{n\times n}$ is a positive semidefinite matrix. This class of subproblem is obtained from the general cost approximation subproblem [$\mathrm{VIP}^k_{\Phi^k}$] by choosing $\Phi^k(\mathbf{x}) = \mathbf{B}_k\mathbf{x}$, $\mathbf{x} \in X$.

Whenever $\mathbf{B}_k$ is symmetric, the resulting variational inequality reduces to the convex quadratic minimization problem

$$\min_{\mathbf{y}\in X} \left\{ F(\mathbf{x}^k)^{\mathrm{T}}(\mathbf{y} - \mathbf{x}^k) + \frac{1}{2\gamma_k}(\mathbf{y} - \mathbf{x}^k)^{\mathrm{T}}\mathbf{B}_k(\mathbf{y} - \mathbf{x}^k) \right\}, \tag{5.7}$$

which further reduces to the well-known scaled projection problem [cf. (3.6) and (4.25)]

$$\mathbf{y}^k = P_X^{\mathbf{B}_k}(\mathbf{x}^k - \gamma_k\mathbf{B}_k^{-1}F(\mathbf{x}^k)) \tag{5.8}$$

whenever $\mathbf{B}_k$ is positive definite. Note that $\mathbf{x}^{k+1} = \mathbf{y}^k$ for all k.

In Table 5.1 we list some well known instances of linear approximation algorithms. Assuming that $F \in C^1$ on X, we let its Jacobian, ∇F, at $\mathbf{x}^k \in X$ be decomposed as $\nabla F(\mathbf{x}^k) = \mathbf{L}_k + \mathbf{D}_k + \mathbf{U}_k$, where $\mathbf{D}_k = \mathtt{diag}\,(\nabla F(\mathbf{x}^k))$ and $\mathbf{L}_k$ and $\mathbf{U}_k$ are the lower and the upper triangular part of $\nabla F(\mathbf{x}^k)$, respectively. Further, we let $\gamma_k \equiv 1$, $0 < \omega < 2$, and $\mathbf{B}$ be a symmetric and positive definite matrix in $\Re^{n\times n}$.

The convergence of linearization algorithms has been studied extensively. A general global convergence result is given by Pang and Chan [738] for continuous mappings $\mathbf{B}$ [i.e., $\mathbf{B}_k = \mathbf{B}(\mathbf{x}^k)$]: letting $\tilde{\mathbf{G}}$ denote the symmetric part of $\mathbf{G}$, i.e., $\tilde{\mathbf{G}} = (1/2)[\mathbf{G} + \mathbf{G}^{\mathrm{T}}]$, if there exists a positive definite matrix $\mathbf{G} \in \Re^{n\times n}$ and a scalar $b < 1$ such that $\mathbf{B}(\mathbf{x}) - \mathbf{G}$ is positive semidefinite on X and

$$\|\tilde{\mathbf{G}}^{-1}[F(\mathbf{x}) - F(\mathbf{y}) - \mathbf{B}(\mathbf{y})(\mathbf{x} - \mathbf{y})]\|_{\tilde{\mathbf{G}}} \le b\|\mathbf{x} - \mathbf{y}\|_{\tilde{\mathbf{G}}}, \qquad \forall \mathbf{x}, \mathbf{y} \in X,$$

Choice of $\mathbf{B}_k$	Resulting method	Basic references
$\nabla F(\mathbf{x}^k)$	Newton	[290, 537]
$\approx \nabla F(\mathbf{x}^k)$	Quasi-Newton	[538]
$\frac{1}{2}[\nabla F(\mathbf{x}^k) + \nabla F(\mathbf{x}^k)^{\mathrm{T}}]$	Symmetrized Newton	[192, 458]
$\mathbf{L}_k + \mathbf{D}_k/\omega$ or $\mathbf{U}_k + \mathbf{D}_k/\omega$	SOR; Linearized Gauss-Seidel ($\omega = 1$)	[727, 738]
$\mathbf{D}_k$	Linearized Jacobi	[738]
$\mathbf{B}$	Projection	[838, 192]

Table 5.1: Examples of linear approximation methods

then $\{\mathbf{x}^k\}$ converges to a solution to [VIP].

Local convergence results for Newton's method are given in [776, 291, 537, 538, 536, 535], for quasi-Newton methods in [538] and for the symmetrized Newton method in [457, 458]. The Newton algorithm can be made globally convergent by replacing the unit step by a line search with respect to the primal gap function; see Section 5.2.5.

The projection algorithm was one of the first algorithms studied for the solution of variational inequalities; see, e.g., [444, 401, 561, 838]. Convergence results are given in [838, 244, 34, 9, 192, 310, 78, 284, 738, 196]; convergence is ensured if F is strongly monotone on X (with modulus m_F) and Lipschitz continuous on X (with modulus M_F), and the smallest eigenvalue, $\mu_{\mathbf{B}}$, of $\mathbf{B}$ is larger than $M_F^2/2m_F$. (This result is developed from the results of Cohen [176]; see also [747, 588].) In this result, convergence is based on the merit Function (5.4); convergence proofs, based on contractive arguments, are given in [192, 738, 196]. (In the latter results, the above condition is replaced by the stronger condition that $\mu_{\mathbf{B}}^2/\|\mathbf{B}\| > M_F^2/2m_F$, which, on the other hand, implies a linear rate of convergence.)

The projection algorithm of Korpelevich [567] completely obviates the strong monotonicity assumption on F, to require only monotonicity. This *extragradient* algorithm is given by

$$\mathbf{x}^{k+\frac{1}{2}} = P_X(\mathbf{x}^k - \gamma F(\mathbf{x}^k)), \tag{5.9a}$$

$$\mathbf{x}^{k+1} = P_X(\mathbf{x}^k - \gamma F(\mathbf{x}^{k+\frac{1}{2}})), \qquad k = 0, 1, \ldots, \tag{5.9b}$$

where $\gamma \in (0, 1/M_F)$. A modification of this method, where the parameter γ is allowed to be chosen adaptively, is given by Khobotov [549] (see also Marcotte [643], who applies the algorithm to [TAP-VIP-F^r]).

Nonlinear approximation algorithms

The class of regularization algorithms ([123, 38, 434]) is obtained by letting $\Phi^k = F + 1/(2\gamma_k)R^k$, where $\gamma_k > 0$ and $R^k : X \mapsto \Re^n$ is a strongly monotone mapping. Special cases of regularization algorithms are the proximal point algorithm ([651, 783]), which is obtained by choosing R^k as the identity mapping, and splitting algorithms ([396, 397, 910, 911, 294]) [see also Section 4.2.1].

5.2.2 Decomposition algorithms

Let the constraints defining the feasible set X of [VIP] be defined by (4.26), i.e., define a Cartesian product of sets X_i. As in the case of nonlinear programming (see Section 4.2.2)

sequential and parallel decomposition algorithms that utilize this problem structure can be devised for the solution of [VIP].

Let the ith block component of Φ^k be of the form $\Phi_i^k : X_i \mapsto \Re^{n_i}$, where Φ_i^k is a monotone mapping which depends only on the variable block component $\mathbf{x}_i$. The resulting variational inequality subproblem [VIP$^k_{\Phi^k}$] then decomposes into m independent problems

[VIP$^k_{\Phi_i^k}$]

$$[\Phi_i^k(\mathbf{y}_i^k) + F_i(\mathbf{x}^k) - \Phi_i^k(\mathbf{x}_i^k)]^{\mathrm{T}}(\mathbf{y}_i - \mathbf{y}_i^k) \geq 0, \qquad \forall \mathbf{y}_i \in X_i. \tag{5.10}$$

Sequential decomposition algorithms

In the sequential version of the decomposition algorithm, in iteration k the index $i_k \in \{1, \ldots, m\}$ is chosen (for rules for selecting these indices, see Section 4.2.2), and the corresponding subproblem [VIP$^k_{\Phi^k_{i_k}}$] is solved, with the solution $\mathbf{y}^k_{i_k}$. We then let

$$\mathbf{x}_i^{k+1} = \begin{cases} \mathbf{x}_i^k + l_k(\mathbf{y}_i^k - \mathbf{x}_i^k), & i = i_k, \\ \mathbf{x}_i^k, & \text{otherwise}, \end{cases}$$

where the value of l_k is chosen such that convergence is guaranteed.

The cyclic version of the cost approximation algorithm includes a block variant of the Gauss–Seidel method. One iteration of this method is defined through the following m subproblems, solved in sequence:

[G–S^k]

$$F_i(\mathbf{x}_{i-}^{k+1}, \mathbf{x}_i^{k+1}, \mathbf{x}_{i+}^k)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{x}_i^{k+1}) \geq 0, \qquad \forall \mathbf{y}_i \in X_i,\ i = 1, 2, \ldots, m. \tag{5.11}$$

This algorithm is also known as the *relaxation method*, and as the *diagonalization method*.

To show that the Gauss–Seidel algorithm is a special case of cyclic cost approximation, let

$$\Phi_i^k(\mathbf{x}_i) = F_i(\mathbf{x}_{i-}^k, \mathbf{x}_i, \mathbf{x}_{i+}^k). \tag{5.12}$$

Then $\Phi_i^k(\mathbf{x}_i^k) = F_i(\mathbf{x}_i^k)$ holds, and it follows, with the choice of $l_k = 1$ for all k, that the solution to [VIP$^k_{\Phi_i^k}$] is the same as the solution to [G–S$_i^k$].

Convergence results for the Gauss–Seidel scheme are given in [12, 737, 738, 736, 84]. Convergence results for essentially cyclic cost approximation algorithms, where the step lengths l_k are chosen through line searches with respect to a merit function, are given in Patriksson [747]. Convergence results for cyclic linear and nonlinear approximation algorithm are given in [736] and [84, Sec. 3.5].

Parallel decomposition algorithms

If the necessary computer facilities are available, then the independent problems [VIP$^k_{\Phi_i^k}$] can be solved simultaneously, thus defining a parallel decomposition algorithm; the iterates are defined in exactly the same manner as in the basic cost approximation algorithm. If asynchronous computations are introduced, then the updating is made individually for the different variable blocks; see Section 4.2.2 for descriptions of different parallel implementations of partial linearization algorithms.

The Jacobi algorithm is a special case of the parallel cost approximation approach, in which the cost approximating mapping is given by (5.12), and unit steps l_k are chosen. Among the first applications of the Jacobi algorithm to variational inequalities we find non-cooperative games ([427, 398]), the PIES model ([502, 517, 11]), and multiclass-user traffic equilibria ([823, 348, 6]).

Convergence properties of the Jacobi algorithm are found in [13, 737, 738, 194, 196, 736, 84]; the convergence rate depends on the amount of interaction in F among the independent variables, and increases with a lesser dependency.

5.2.3 Column generation algorithms

The principle of column generation described for nonlinear programming problems in Section 4.2.3 extends immediately to variational inequality problems, since it involves only inner representations of the feasible set. Care must be taken, however, when developing column generation algorithms supplied with column dropping rules; for instance, the Frank–Wolfe algorithm—which is obtained by choosing the parameter value $r = 1$ in a restricted simplicial decomposition scheme—is not convergent when applied to variational inequalities (e.g., [630]).

A column dropping rule based on the value of the primal gap function is given by Lawphongpanich and Hearn [594].

5.2.4 Algorithmic equivalence results

Below, we show that two well-known classes of successive approximation algorithms are included in the framework of cost approximation.

The algorithm class of Cohen

In our notation, the *auxiliary problem principle* of Cohen [176] generates a sequence $\{\mathbf{x}^k\}$ where, given $\mathbf{x}^k \in X$, $\mathbf{x}^{k+1}$ is the solution to

$$\min_{\mathbf{y}\in X} \left\{\varphi(\mathbf{y}) + [\varepsilon F(\mathbf{x}^k) - \nabla\varphi(\mathbf{x}^k)]^{\mathrm{T}}\mathbf{y}\right\}, \tag{5.13}$$

where $\varphi : X \mapsto \Re$ is a strongly convex function in C^1 on X, and $\varepsilon > 0$.

The algorithm class of Dafermos

Let $\Gamma : X \times X \mapsto \Re^n$ be a continuous mapping satisfying

(1) $\Gamma(\mathbf{x},\mathbf{x}) = F(\mathbf{x})$ for all $\mathbf{x} \in X$,

(2) $\nabla_{\mathbf{x}}\Gamma(\mathbf{x},\mathbf{y})$ is symmetric and positive definite for any fixed $\mathbf{x},\mathbf{y} \in X$.

In the algorithm framework of Dafermos [196], given $\mathbf{x}^k \in X$, the point $\mathbf{x}^{k+1}$ is obtained by solving the variational inequality subproblem

$$\Gamma(\mathbf{x}^{k+1},\mathbf{x}^k)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^{k+1}) \geq 0, \qquad \forall \mathbf{x} \in X. \tag{5.14}$$

Theorem 5.1 (Algorithmic equivalence results)

(a) *Given $\mathbf{x}^k \in X$, let Φ^k be given by $\Phi^k(\mathbf{x}) = (1/\varepsilon)\nabla\varphi(\mathbf{x})$, where $\varphi : X \mapsto \Re$ is a convex function in C^1 on X and $\varepsilon > 0$. Then the resulting subproblem* $[\mathrm{VIP}^k_{\Phi^k}]$ *is equivalent to the Subproblem (5.13) of [176].*

(b) *Given* $\mathbf{x}^k \in X$, *let* Φ^k *be given by* $\Phi^k(\mathbf{x}) = \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}^k)$, *where* $\varphi : X \times X \mapsto \Re$ *is a continuous function on* $X \times X$, *and convex and in* C^1 *on* X *with respect to its first argument. Then the resulting subproblem* $[\mathrm{VIP}^k_{\Phi^k}]$ *is equivalent to the Subproblem (5.14) of [196].*

Proof

(a) See [747, 588].

(b) See [746, 747, 588]. □

Based on these algorithmic equivalence results, we are now in the position to establish the convergence of instances of the *successive* cost approximation algorithm, i.e., the cost approximation algorithm using unit step lengths.

Theorem 5.2 (Convergence of successive cost approximation algorithms)

(a) *Let* F *be strongly monotone and Lipschitz continuous on* X, *and let* $\varphi : X \mapsto \Re$ *be a strongly convex function in* C^1 *on* X. *Let the sequence* $\{\mathbf{x}^k\}$ *be generated by the successive cost approximation algorithm, where* $\Phi^k = \nabla\varphi$. *If* φ *is chosen such that*

$$2m_\varphi > \frac{M_F^2}{m_F}, \tag{5.15}$$

then $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *where* $\mathbf{x}^*$ *is the unique solution to* [VIP].

(b) *Let* X *be bounded and* F *in* C^1 *on* X, *and let* $\nabla\varphi : X \times X \mapsto \Re^n$ *be in* C^1 *on* $X \times X$. *Further, let* $\nabla^2_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{y})$ *be positive definite for any fixed* $\mathbf{x}, \mathbf{y} \in X$. *Let the sequence* $\{\mathbf{x}^k\}$ *be generated by the successive cost approximation algorithm, where* $\Phi^k(\mathbf{x}) = \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}^k)$. *If*

$$\Big\| \nabla^2_{\mathbf{x}}\varphi(\mathbf{x}^1, \mathbf{y}^1)^{-1/2} [\nabla^2_{\mathbf{xy}}\varphi(\mathbf{x}^2, \mathbf{y}^2) + \nabla F(\mathbf{y}^2) - \nabla^2_{\mathbf{xy}}\varphi(\mathbf{y}^2, \mathbf{y}^2)] \cdot \nabla^2_{\mathbf{x}}\varphi(\mathbf{x}^3, \mathbf{y}^3)^{-1/2} \Big\| < 1 \tag{5.16}$$

for all $\mathbf{x}^1, \mathbf{y}^1, \mathbf{x}^2, \mathbf{y}^2, \mathbf{x}^3, \mathbf{y}^3 \in X$, *then* $\{\mathbf{x}^k\} \to \mathbf{x}^*$, *where* $\mathbf{x}^*$ *is the unique solution to* [VIP].

Proof

(a) Follows from [176, Th. 2.2] and Theorem 5.1.a, where we choose $\varepsilon = 1$.

(b) Follows from [196, Th. 2.1], Theorem 5.1.b, and by identifying $\Gamma(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{y}) + F(\mathbf{y}) - \nabla_{\mathbf{x}}\varphi(\mathbf{y}, \mathbf{y})$. □

The above results require the mapping F to be at least strictly monotone, and it is difficult to verify the conditions for convergence in practice. This is in contrast to the mild (and in many cases easily checked) conditions that guarantee the convergence of this type of algorithm in the case of nonlinear programming.

Remark 5.1 It is well known that the route cost function $\mathbf{c}$ is not strictly monotone even when the corresponding link travel cost function $\mathbf{t}$ is strongly monotone. The consequence is, of course, that algorithms which require strong monotonicity in order to be convergent can not be applied to traffic equilibrium problems in the space of route flows. Recently, however, Zhu and Marcotte [1012] showed that the auxiliary problem principle

of Cohen [176] is convergent under conditions similar to that of Theorem 5.2.a, but where F is assumed to be *co-coercive* only (see Definition A.2.d). As opposed to the strong monotonicity property, the co-coercivity property (which is implied by strong monotonicity) is preserved under affine transformations; as a result, this class of successive cost approximation algorithms may be utilized for solving traffic equilibrium problems in the space of route flows, provided that the link travel cost functions are strongly monotone.

In the next section, we introduce descent algorithms based on specially constructed merit functions. Subsequently, we will show that a modification of the above algorithms yields convergence both under mild monotonicity assumptions and under conditions that are much easier to verify.

5.2.5 Descent algorithms for variational inequalities

The direct extension of the line search Step 3 of the general descent algorithm of Section 4.2.1 to variational inequalities amounts to solving the one-dimensional variational inequality

$$F(\mathbf{x}^k + l_k\mathbf{p}^k)^{\mathrm{T}}\mathbf{p}^k(l - l_k) \geq 0, \qquad \forall l \in [0, l_k^{\max}], \tag{5.17}$$

where $l_k^{\max}$ is the maximum feasible step length in the direction of $\mathbf{p}^k$.

This direct extension of nonlinear programming methods does not define convergent algorithms in general (see, e.g., [630] for a counter-example using the extension of the Frank–Wolfe algorithm.) It can be used, however, at regular intervals in successive approximation algorithms for [VIP], as shown by Harker [463]; he introduces this strategy in Step 3 in every second iteration of Dafermos' [196] scheme, establishes convergence under some additional technical assumptions, and shows through tests performed on the special cases of the Jacobi and projection algorithms that it may enhance the efficiency of the scheme significantly.

Algorithms based on the primal and dual gap function

Algorithms for [VIP] based on the minimization of (calculable) merit functions have been considered by Russian scientists since at least the late 1960s.

The primal gap function, and the corresponding optimization Formulation (3.13), was first studied in an algorithmic context by Zuhovickiĭ *et al.* [1019, 1020] (see also [244, 766]). In an extension of the Frank–Wolfe algorithm to [VIP], the step lengths are chosen according to an inexact line search with respect to the primal gap function. The convergence, however, relies on properties of the feasible set which precludes applications to polyhedral feasible sets, and therefore to traffic equilibrium problems.

Marcotte [641] shows how a descent direction with respect to the primal gap function can be obtained from the set of solutions $Y(\mathbf{x})$ to (3.12) by using a polyhedral approximation technique similar to those used in bundle methods in nondifferentiable convex optimization.

In [644, 645, 286, 646] the Newton subproblem is shown to yield a descent direction with respect to the primal gap function whenever F is monotone; under additional monotonicity and regularity assumptions, this descent algorithm is shown to act locally like the original Newton algorithm, thereby obtaining quadratic convergence. (Demyanov and Pevnyi [244] discuss a descent algorithm based on the Newton algorithm, but do not provide any convergence results.)

Hearn [472, 474] adopts Polyak's [759] subgradient algorithm to the non-differentiable Program (3.13), in the case of nonlinear programming. The use of Polyak's method [see (4.67)] is advantageous, since the optimal value is known.

Turning to the dual gap function, methods proposed for the Program (3.16) are mostly applied to the equivalent semi-infinite linear program

$$\begin{aligned} \max \quad & z, && (5.18a)\\ \text{s.t.} \quad & z \le F(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \qquad \forall \mathbf{x} \in X, && (5.18b)\\ & \mathbf{y} \in X. && (5.18c) \end{aligned}$$

Zuhovickiĭ *et al.* [1017, 1018, 1020] apply a cutting plane approach in which (5.18b) is replaced by constraints $z \le F(\mathbf{x}^j)^{\mathrm{T}}(\mathbf{x}^j - \mathbf{y})$, $j = 0, 1, \ldots, k$; in iteration k, the solution, $\mathbf{y}^k$, to the corresponding restriction of (5.18) defines $\mathbf{x}^{k+1}$ and a new cutting plane. Dem'yanov and Pevnyi [244] derive $\mathbf{x}^{k+1}$ from the evaluation of g at $\mathbf{y}^k$. Auslender [34, Sec. VII.5] considers applying other cutting plane approaches to (5.18). Nguyen and Dupuis [718, 719] and Hearn and Lawphongpanich [476] suggest performing a line search [i.e., the one-dimensional variational Inequality (5.17)] in the direction of $\mathbf{p}^k = \mathbf{y}^k - \mathbf{x}^k$, and show that it yields a descent step with respect to the objective z of (5.18). Hearn and Lawphongpanich investigate the convergence of the algorithm under different step length rules.

As in the case of the primal gap function, subgradient optimization approaches have been suggested for the solution of the non-differentiable Program (3.16); see [284, 282, 458].

Algorithms based on differentiable gap functions

Smith [845, 846, 847] develops a descent algorithm for [VIP] based on the differentiable gap function G^2 [see (3.17)]; convergence is ensured under a monotonicity and differentiability condition on F; the algorithm, however, requires the knowledge of the extreme points of the feasible set X and must therefore be embedded in a simplicial decomposition scheme.

A number of descent algorithms based on differentiable gap functions of the form defined in (3.18a) and special cases from the class of cost approximation algorithms have been proposed for the solution of [VIP].

The algorithm of Fukushima [391] is based on solving projection subproblems, which are obtained from choosing $\Phi^k(\mathbf{x}) = \mathbf{B}\mathbf{x}$ for all k, where $\mathbf{B}$ is a positive definite and symmetric matrix. Line searches are performed in the resulting directions with respect to the gap function defined by the corresponding choice of $\varphi(\mathbf{x}) = (1/2)\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$ in (3.18a). Convergence is guaranteed under the assumptions that X is bounded and that the Jacobian of F is everywhere positive definite; under an additional strong monotonicity assumption, line searches can be made with an Armijo-type inexact rule. Under strong monotonicity assumptions, the Newton subproblem is shown in [885] to yield descent directions with respect to the same merit function.

Larsson and Patriksson [588] develop a descent algorithm based on the gap function defined by (3.18a) [cf. Theorem 3.13.e]; convergence is established under conditions that imply that F is at least strictly monotone; they show that the algorithm also converges when the subproblems are solved inexactly. This approach is further studied in [747], and applied to variational inequality problems over Cartesian product sets; convergence is established for both parallel and essentially cyclic decomposition versions of the descent algorithm, and for predetermined step length rules.

Wu *et al.* [994] solve symmetric subproblems that can be identified as special cases of cost approximation subproblems with $\Phi^k(\mathbf{x}) = \nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}^k)$ for all k; this subproblem is

equivalent to an extension of the regularized Frank–Wolfe Subproblem (4.22) to [VIP]. The corresponding merit function defined in (3.18a) is utilized in exact and inexact line searches, and convergence is established under conditions that include strong monotonicity on F, but also includes technical assumptions that are very difficult to verify.

Zhu and Marcotte [1014] consider descent algorithms which may be identified by the choices $\Phi^k(\mathbf{x}) = \Phi(\mathbf{x}, \mathbf{x}^k)$ for all k, i.e., by the choice of the sequence $\{\Phi^k\}$ as a fixed, continuous mapping, and two different merit functions: the one proposed by Fukushima [391], and the one given by Wu *et al.* [994]. Convergence is ensured for a version employing an inexact line search rule, under strong monotonicity assumptions and additional nonstandard technical assumptions.

We conclude the discussions on descent algorithms for [VIP] by providing an instance of the above general scheme which is convergent under very mild assumptions on the problem and does not require any knowledge of constants of the problem (such as the Lipschitz constant).

The subproblem of the algorithm of Zhu and Marcotte [1013] is shown in [748] to be a cost approximation subproblem, given the choice $\Phi^k(\mathbf{x}) = (1/\alpha_k)\nabla_{\mathbf{x}}\varphi(\mathbf{x}, \mathbf{x}^k)$ for all k, where $\alpha_k > 0$. In iteration k, the corresponding merit function, ψ_{α_k}, is evaluated at $\mathbf{x}^k$. If

$$\psi_{\alpha_k}(\mathbf{x}^k) \leq \frac{1}{(1-\gamma)\alpha_k} \left\{ \nabla_{\mathbf{x}}\varphi(\mathbf{x}^k, \mathbf{y}^k)^{\mathrm{T}}(\mathbf{y}^k - \mathbf{x}^k) - \varphi(\mathbf{x}^k, \mathbf{y}^k) \right\},$$

where $\gamma \in (0,1)$, then the value of α_k is increased by a fixed amount, $\mathbf{x}^{k+1} = \mathbf{x}^k$ is set, and $\psi_{\alpha_{k+1}}(\mathbf{x}^k)$ is calculated; otherwise, $\alpha_{k+1} = \alpha_k$ is set, and either the point $\mathbf{y}^k$ defines $\mathbf{x}^{k+1}$ or an Armijo-type line search is made in the direction of $\mathbf{y}^k - \mathbf{x}^k$ with respect to ψ_{α_k}.

We next make some interesting observations. Firstly, the symmetric cost approximation subproblems defined in this algorithm are simple modifications of those that define, for instance, the general framework of Dafermos [196]; indeed, the sequence of cost approximating mappings is $\{\Phi^k\} = \{(1/\alpha_k)\nabla_{\mathbf{x}}\varphi(\cdot, \mathbf{x}^k)\}$ (cf. Theorem 5.1.b), which differs from Dafermos' scheme only in the introduction of the constants $(1/\alpha_k)$. Secondly, disregarding the line search step, which is only considered when it is necessary in order to guarantee global convergence, the algorithm contains the same steps as the successive cost approximation algorithms, and therefore amounts to minor changes of most existing codes for variational inequality problems.

The interest in this algorithm lies in the facts that it is convergent under both mild and simple assumptions and that it may be implemented by slightly modifying an existing (successive) cost approximation algorithm. Moreover, it is not necessary to have estimates of problem parameters (typically strong monotonicity and Lipschitz constants), which are normally needed when step length rules are implemented in variational inequality algorithms.

The convergence of this algorithm is given by the below theorem.

Theorem 5.3 [1013] (Convergence of a descent algorithm) *Let X be bounded and F be in C^1 on X, monotone and Lipschitz continuous on X. Let the sequence $\{\mathbf{x}^k\}$ be generated by the above algorithm. Then, any accumulation point of this sequence is a solution to* [VIP].

5.3 Algorithms for general traffic equilibria

The non-existence of an equivalent convex problem of the form [TAP] and [TAP-E] to the equilibrium Conditions (2.1) and (2.3), respectively, is due to the asymmetry of the travel cost and demand functions. The asymmetries arise from the modelling of interaction among vehicles at intersections, and of different user classes by using multiple copies of the traffic network (see Section 2.5).

The most natural algorithmic approach to an asymmetric user equilibrium problem perhaps is to iteratively replace the asymmetric cost and demand functions with symmetric (and preferably separable) ones; the resulting problem can then be solved using any of the algorithms in Chapter 4. Nearly all of the algorithms proposed for the solution of asymmetric traffic models are of this type; the most popular one is an adaptation of the Jacobi approach, in which the asymmetry is iteratively removed by ignoring the cost and demand interactions between the link and O-D flows, thus obtaining separable subproblems. (This algorithm is known also as the *diagonalization* algorithm, as well as the *relaxation* algorithm.)

Below, we outline the development of algorithms for asymmetric user equilibria.

5.3.1 Linear approximation algorithms

The projection algorithm was first considered by Dafermos [192] for [TAP-VIP-F^r], and for a two-mode model in [193]. Convergence results are given in [9, 192, 310, 340, 195]; the iteration

$$\mathbf{f}^{k+1} = P^{\mathbf{B}}_{F^r}(\mathbf{f}^k - \gamma \mathbf{B}^{-1}\mathbf{t}(\mathbf{f}^k)), \qquad k = 0, 1, \ldots, \tag{5.19}$$

defines a contraction whenever $\gamma \in (0, 2m_{\mathbf{t}}/\nu)$, where

$$\nu \stackrel{\text{def}}{=} \max_{\mathbf{f} \in F^r} \rho(\nabla \mathbf{t}(\mathbf{f})^{\mathrm{T}} \mathbf{B}^{-1} \nabla \mathbf{t}(\mathbf{f}))$$

is the maximum eigenvalue of $\nabla \mathbf{t}(\mathbf{f})^{\mathrm{T}} \mathbf{B}^{-1} \nabla \mathbf{t}(\mathbf{f})$ over all feasible link flows.

This condition implies that the method is allowed to take very small steps only, and the convergence, although it is linear in theory, is often quite slow. Furthermore, the projection onto F^r (which is equivalent to a quadratic network flow problem) is time consuming. (It is, of course, possible to partially reduce the computational burden through efficient reoptimizations of previous subproblem solutions.) The accelerating line search step of Harker [463] (see Section 5.2.5) may enhance the convergence in practice.

Dafermos [192] also proposes using a quasi-Newton approach for the solution of [TAP-VIP-F^r]. This amounts to replacing the fixed matrix $\mathbf{B}$ in (5.19) with a sequence $\{\mathbf{B}_k\}$ of symmetric and positive definite matrices defined, for instance, by the symmetric part of the Jacobian of $\mathbf{t}$ at the points $\mathbf{f}^k$. This algorithm is later investigated by Fisk and Nguyen [340], who also discuss the proper choices of values of the step length parameter γ for both the projection and quasi-Newton algorithms. Limited tests, where the quadratic network flow problems are solved using the Frank–Wolfe algorithm, show that the projection algorithm is slow for most choices of the matrix $\mathbf{B}$; this is particularly the case when the Jacobian $\nabla \mathbf{t}$ varies significantly on F^r, i.e., when $\mathbf{t}$ is highly nonlinear, since no fixed matrix then can be a good approximation of the Jacobian of $\mathbf{t}$ on large subsets of F^r (see also [283]). Both the projection and quasi-Newton algorithms were found to generate negative cycles in the subproblem phase.

As in the case of separable costs (see Section 4.3.1), projection algorithms are more easily applied to the link-route formulation; see Section 5.3.5 for examples.

5.3.2 Sequential decomposition algorithms

Surprisingly, considering the popularity of sequential decomposition schemes for the separable model (see Section 4.3.2), such algorithms are discussed relatively scarcely for asymmetric models. The reasons for using a sequential decomposition approach as compared to a parallel decomposition algorithm for example are just as valid here, and the interactions in the costs and demands are even more pronounced.

A point in favour of decomposition algorithms in general is that it may be much more difficult to evaluate the cost and demand functions in the nonseparable case; in a decomposition algorithm, the number of evaluations of the original cost and demand mappings are kept down to a minimum.

Fisk and Nguyen [338] discuss the application of the Gauss–Seidel approach to a multiclass-user network equilibrium model formulated as a system of nonlinear equations.

Dafermos [194] establishes the convergence of a cyclic decomposition scheme over modes, and in [195] over O-D pairs.

5.3.3 Parallel decomposition algorithms

Parallel decomposition algorithms of the Jacobi type were the first methods proposed for the solution of multi-modal networks ([823, 348, 6, 94, 355, 453, 939]). After the introduction of variational inequality formulations of traffic equilibrium problems, nonlinear Jacobi methods were again studied.

Florian [349, 350] extends the algorithm of Florian [348]; local convergence results, and a global result for the affine cost case, are given in [363].

The first general convergence results are due to Dafermos [194, 195, 196], who presents results for both the single-mode and multi-mode cases with elastic demands. In the single-mode case, the cost and demand functions are diagonalized simultaneously over O-D pairs. In the multi-mode case, a diagonalization is made also with respect to the different modes. The intuitive convergence condition is that the interactions are small enough; the weaker the interaction is, the faster the algorithm converges.

The subproblem of a Jacobi type approach is a standard traffic assignment problem. Since many such subproblems may have to be solved before the algorithm stabilizes in the vicinity of a solution, it is important not to spend too much effort on each subproblem, and therefore to terminate their solutions prior to reaching a solution. (The use of truncated subproblem algorithms is discussed in more detail in Section 4.2.1.) Sheffi [831, Sec. 8.2] presents a Jacobi algorithm, in which each separable subproblem is solved using one iteration of the Frank–Wolfe algorithm. The computational effort is similar to that of the Frank–Wolfe algorithm for the separable model, and the algorithm is easily implemented based on such a scheme. Mahmassani and Mouskos [634] perform numerical experiments with a Jacobi algorithm using the truncated Frank–Wolfe algorithm for the solution of the subproblems. Their conclusion is that not more than four Frank–Wolfe iterations should be performed in each subproblem.

Harker [463] applies the accelerating line search Procedure (5.17) to the Jacobi algorithm, and shows that it may yield a substantially lower computational cost when applied to some small examples.

Computational comparisons between nonlinear Jacobi approaches and linear approximation algorithms are made by Fisk and Nguyen [340] and Nagurney [699, 700]. Fisk and Nguyen found that a linearized Jacobi algorithm was superior to the projection algorithm for all choices of fixed matrices tried in the latter method. [Note that the linearized

Jacobi algorithm is a linear approximation algorithm where the matrices $\mathbf{B}_k$ are chosen as diagonal approximations of the Jacobian of $\mathbf{t}$, and therefore may be viewed as a single iteration Newton approach to the nonlinear Jacobi algorithm; cf. (4.62).] The nonlinear Jacobi approach was found to be even more efficient. In all these experiments, the symmetric subproblems were solved using a few steps of the Frank–Wolfe algorithm. Nagurney studies several small networks with different travel cost and demand functions. The overall best performance of the linear approximation algorithms tested were obtained from the linearized Jacobi approach. (Although choosing non-diagonal matrices resulted in convergence in fewer iterations, the total computing time was found to be higher.) The comparative performance of the linear and nonlinear approximation algorithms was found to vary with the nonlinearities of the cost and demand functions. In these tests, both the Frank–Wolfe algorithm and the equilibration operator approach were used for the solution of the symmetric subproblems; of the two, the latter was always found to be more efficient.

A parallel algorithm based on Douglas-Rachford splitting ([616, 294]) is described in [392].

5.3.4 Algorithms based on the primal and dual gap functions

In the algorithms for the Problem (3.13) presented by Zuhovickiĭ *et al.* [1019, 1020] (see Section 5.2.5) step lengths are taken in the direction defined by the solution to the Frank–Wolfe subproblem. Hearn [472, 474] discusses the use of Polyak's [759] method, where the step length is given by (4.67), and the optimal value is zero. Fisk and Nguyen [340] use an averaging procedure, which corresponds to using step lengths $1/k$, and hence defines a direct extension of the MSA algorithm (see Section 4.1.6) to variational inequalities. Results obtained from limited tests are promising.

Marcotte [641] discusses the application to [TAP-VIP-F^r] of the bundle-type algorithm described in Section 5.2.5. In this application, all-or-nothing solutions are retained and combined in order to yield a descent direction with respect to the primal gap function; this idea of combining all-or-nothing solution to obtain a good search direction is similar to that of Fukushima [387] (see Section 4.1.6). The algorithm of Arezki [22, 23] is based on a similar idea.

5.3.5 Column generation algorithms

In this section we trace the development of column generation algorithms for general traffic equilibria. As for the separable models, we divide the presentation between algorithms based on aggregated and disaggregated representations of the feasible set.

For a background to the principles of column generation and simplicial decomposition, see Section 4.2.3; column generation and simplicial decomposition algorithms are discussed for the separable model in the Sections 4.3.4 and 4.3.5.

Aggregate simplicial decomposition algorithms

An aggregate simplicial decomposition algorithm for [TAP-VIP-F^r] is the extension of that for the separable model, where the restricted master Problem (4.63) is replaced by the variational inequality problem of finding $\mathbf{f}^{k+1} \in \hat{F}^n$ such that

$$\mathbf{t}(\mathbf{f}^{k+1})^{\mathrm{T}}(\mathbf{f} - \mathbf{f}^{k+1}) \geq 0, \qquad \forall \mathbf{f} \in \hat{F}^n, \tag{5.20}$$

where $\hat{F}^n$ is given by (4.63b)–(4.63d).

Using the taxonomy of Section 4.2.5, the algorithms may be described in the form $F^n(C_{\mathcal{A}})$.

Smith [845] presents an ASD scheme for monotone [TAP-VIP-F^r], where each restricted master Problem (5.20) is solved using a descent algorithm for the merit function G^2 (see Section 3.1.5). He shows that convergence is ensured also when the Problems (5.20) are solved inexactly, but he does not present any computational results.

Hearn *et al.* [593, 594, 480] apply both a linearized Jacobi and a projection algorithm to (5.20). (This latter algorithm is a special case of the former, since the matrices chosen are always diagonal; this choice of matrices ensures that the quadratic subproblems are analytically solvable.) They show that convergence is guaranteed under the conditions that **(a)** the restricted master problems are solved accurately in the limit,[1] and **(b)** that columns with zero weights are only dropped when the value of the primal gap function has decreased sufficiently. A disaggregated version of the algorithm ($F^n(D^P_{\mathcal{O}}[C_{\mathcal{O}}])$), similar to that given by Bertsekas and Gafni [78], is also tested. Experiments on small-scale networks indicate that the aggregated version is superior to the disaggregated one; each restricted master problem of the latter algorithm is, however, solved very roughly, which might explain this conclusion.

Pang and Yu [739] approximate each restricted master problem by replacing the original cost mapping by the diagonal part of its Jacobian evaluated at the solution to the previous restricted master problem; that is, each restricted master problem is solved using one iteration of a linearized Jacobi algorithm. (The algorithm is therefore an aggregated version of the algorithm of Bertsekas and Gafni [78].) Each separable quadratic network flow problem is solved using a pivoting algorithm ([734]). Tests on various small networks show promising results; for larger networks, the need to solve a large number of shortest route problems due to the crude approximation used suggests that a more accurate solution of each restricted master problem is preferable.

Marcotte and Guélat [647] apply the modified Newton algorithm of Marcotte and Dussault [644, 645, 286, 646] (see Section 5.2.5) to each restricted master problem. Comparisons with a cutting plane approach ([718, 719]; see Section 5.3.6) and a Jacobi approach (in which the separable problems are solved using the PARTAN algorithm) are performed on networks with varying degrees of cost asymmetry. For nearly symmetric problems, the Jacobi approach was found to be the most efficient for the accuracy required, but both the cutting plane and simplicial decomposition/Newton approaches were found to be much more robust when solving highly asymmetric problems.

Montero [684] investigates the performance of ASD algorithms with respect to the choices of starting solution, shortest route algorithm, algorithms and stopping criteria for each restricted master problem, and criteria for column dropping. Experiments are performed on the small networks reported in the literature as well as on some large bimodal networks resulting from the modelling of route guidance systems. (The two modes correspond to guided and unguided vehicles, respectively.) Variable metric projection algorithms are the most efficient among the projection methods tested; projection algorithms with fixed matrices are sensitive to the choice of step length parameter, and in some applications fail to converge.

---

[1] The condition is that

$$\mathbf{t}(\mathbf{f}^{k+1})^{\mathrm{T}}(\mathbf{f}-\mathbf{f}^{k+1}) \geq -\varepsilon_k, \qquad \forall \mathbf{f} \in \hat{F}^n,$$

where $\{\varepsilon_k\} \downarrow 0$ [cf. (4.21)].

Disaggregate simplicial decomposition algorithms

A disaggregated representation of the feasible set leads to a restricted master problem of the form [TAP-VIP-F^r] where the sets $\mathcal{R}_{pq}$ are replaced by subsets $\hat{\mathcal{R}}_{pq}$, $(p,q) \in \mathcal{C}$.

Aashtiani and Magnanti [2, 3, 5] apply an algorithm of the form $F^r(D^S[C_{\mathcal{R}}])$ to the nonlinear complementarity model [TAP-E-NCP]. The level of decomposition they use depends on the size of the problem and the nature of the demand function. Each subproblem of the sequential scheme involves the generation of new routes (and the deletion of routes with zero flows) and the solution of, for example, a single-commodity nonlinear complementarity problem; each such problem is solved with a Newton algorithm, where the linear complementarity subproblems are solved using Lemke's [609] algorithm. In [5], only tests performed on separable models are reported.

Bertsekas and Gafni [78] extend the scaled gradient projection algorithms of Section 4.2.2 to the solution of [TAP-VIP-F^r]. The overall scheme is of the form $F^r(D_{\mathcal{C}}[C_{\mathcal{R}}])$, where the decomposition is either parallel (or *all-at-once*) or cyclic (or *one-at-a-time*). New routes are generated after each iteration of the projection scheme. Although the restricted master problems are not strongly monotone in the space of route flows, they are able to establish convergence of the projection algorithm under strong monotonicity assumptions on the link flows, and also establish the convergence of a more general projection algorithm, where the matrix $\mathbf{B}$ is allowed to change from one iteration to the next whenever sufficient progress has been made. (This convergence result is a special case of that for the auxiliary problem principle under co-coercivity; see Remark 5.1.) In their numerical experiments, the condition for allowing the matrix to change is fulfilled in every iteration of both the parallel and the sequential implementations. In this case, their algorithm turns into a linearized Jacobi approach, which in the case of sequential decomposition may be viewed as a simplified version of that of Aashtiani and Magnanti, in the sense that each restricted master problem is solved less accurately before new routes are generated. The sequential approach can not be shown to converge, but is found to be more efficient than the parallel implementation in experiments on small networks. An implementation is described in [82].

Smith [846] describes a column generation algorithm for [TAP-VIP-F^r] of the form $F^r(C_{\mathcal{R}}[D^P_{\mathcal{C}}])$, which extends the aggregated version in [845].

5.3.6 Dual algorithms

Fukushima and Itoh [393] apply the projection algorithm of Fukushima [390] to [DTAP-E-VIP-F^r_d]. Given an iterate $(\boldsymbol{\mu}^k, \boldsymbol{\pi}^k) \in \Re^{|\mathcal{A}|+|\mathcal{C}|}$, the following iterate is

$$\begin{pmatrix} \boldsymbol{\mu}^{k+1} \\ \boldsymbol{\pi}^{k+1} \end{pmatrix} = P^{\mathbf{B}}_{H^k} \left(\begin{pmatrix} \boldsymbol{\mu}^k \\ \boldsymbol{\pi}^k \end{pmatrix} - \gamma_k \mathbf{B}^{-1} \begin{pmatrix} \mathbf{f}(\boldsymbol{\mu}^k) \\ -\mathbf{g}(\boldsymbol{\pi}^k) \end{pmatrix} \right),$$

where $H^k \subset \Pi_\mu$ is a halfspace defined by the most violated constraint at $(\boldsymbol{\mu}^k, \boldsymbol{\pi}^k)$ among those defining the feasible set Π_μ of [DTAP-E-VIP-F^r_d]. This most violated constraint is identified by a shortest route at the given cost $\boldsymbol{\mu}^k$. This algorithm is similar to the dual algorithms presented in Section 4.3.7 for the separable model, and has the same inherent property of providing a feasible flow in the limit only. Moreover, the calculation of the demand and inverse travel cost functions is very time consuming; in numerical experiments, Fukushima and Itoh find that the portion of the total computation that is spent on these calculations is as high as around 95 %.

The inherent dual character of the algorithm can be circumvented by introducing a simple scheme for generating primal feasible flows from the shortest route subproblem

solutions, as suggested by Larsson *et al.* [582]. Such a scheme is described for the separable case in Section 4.3.7. An alternative is to algorithmically generate the constraints of Π_μ in a cutting plane algorithm and solve master problems over those subsets of constraints; the optimal multipliers of these constraints define a feasible route flow solution, and such an approach is therefore a dual representation of a disaggregate column generation/simplicial decomposition algorithm (see Section 5.3.5). Itoh *et al.* [520] propose such an algorithm for [DTAP-E-VIP-F_d^r].

The Reformulation (5.18) of [VIP] is utilized in the development of cutting plane algorithms for [TAP-VIP-F^r] by Nguyen and Dupuis [718, 719]. Given cutting planes $z + \mathbf{t}(\mathbf{f}^j)^{\mathrm{T}}\mathbf{y} \leq \mathbf{t}(\mathbf{f}^j)^{\mathrm{T}}\mathbf{f}^j$, $j \in \{0, 1, \ldots, k\}$, the auxiliary solution $\mathbf{y}^k$ is obtained by solving the corresponding restriction to (5.18); this solution is obtained from a special simplex algorithm. The new iterate—which also defines the new cut to be added—is obtained by approximately solving the one-dimensional variational Inequality (5.17) over the interval $[\mathbf{f}^k, \mathbf{y}^k]$. When formulating the next linear master problem the currently inactive constraints are dropped, and the problem is reoptimized from the solution of the previous master problem. Numerical experiments performed on medium-scale separable models as well as on small affine asymmetric ones indicate that the number of cuts needed is very limited and that the reoptimization of the restricted master problems is efficient. The algorithm compares favourably with the Frank–Wolfe algorithm on the separable models, and with a Jacobi approach where the separable subproblems are solved using the Frank–Wolfe algorithm on the asymmetric models.

5.3.7 Other algorithms

Among the first algorithms considered for the solution of general traffic equilibria were general algorithms for computing fixed points (e.g., [901, 410]). For example, Asmuth [30, 31] applies the Eaves–Saigal [292] algorithm to a fixed point model of the Wardrop conditions. Such pivoting algorithms can only solve very small problems efficiently, since they do not utilize the network structure. The same conclusion is drawn by Aashtiani [2] from experiments on a pivoting algorithm for a nonlinear complementarity formulation.

Maugeri [657, 658] develops an algorithm which identifies the optimal face of the polyhedron H of feasible route flows by successively reducing its dimension. In [231] it is extended to elastic demand problems through a reformulation of the Wardrop conditions into a quasi-variational inequality problem ([61, 689]).

5.4 Discussion

The development of algorithms for separable models of traffic equilibria followed that for general nonlinear programs and network optimization. The development of algorithms for asymmetric models was, however, a driving force for the development of the whole field of iterative methods for variational inequality problems.

For two main reasons, knowledge about the most efficient algorithms for the solution of traffic equilibrium problems is relatively limited. Firstly, the lack of applications means that most networks applied are small ones constructed for illustration purposes only. The result is that the algorithms have not been tested or compared with regard to their ability to solve large-scale models.

Secondly, the types of algorithms tested are very limited. Decomposition algorithms of the nonlinear Jacobi type, along with projection algorithms, are the two predominant

classes of algorithms proposed and tested. Moreover, several algorithms have not been implemented in the most efficient manner possible. The implementations of disaggregate simplicial decomposition (DSD) schemes is a good example: in the implementations, which are commonly of the form $F^r(D_{\mathcal{C}}[C_{\mathcal{R}}])$, new routes are generated after only a few steps of an iterative algorithm for a given restricted master problem, which implies that the total number of shortest route calculations is much higher than necessary.

As discussed in Section 2.7, research into the efficient solution of asymmetric traffic equilibrium problems has been motivated more by the scientific challenge than the appropriateness of the models; general network equilibrium models should perhaps be seen more as a basis for idealized descriptions of equilibrium states than as models for actually computing equilibrium flows. Algorithms are most often proposed without any investigations as to whether the restrictions imposed upon the network data by the conditions for convergence or the algorithms are realistic or not, or even if it is possible to collect or estimate the data required.

Interesting to note is that in some experiments, the network data has been shown not to satisfy the conditions for convergence of the algorithm used, especially in the applications of diagonalization approaches (e.g., [6, 340, 966, 385, 634]), although the algorithm has been successful in obtaining an equilibrium solution. One conclusion may be that in some cases, the theoretical convergence conditions are too strong, and that it may be possible to weaken them significantly. Some progress has been made in the development of algorithms for variational inequalities which require weak monotonicity assumptions; an especially interesting algorithm class is the modified descent algorithm of Zhu and Marcotte [1013] (see Theorem 5.3), which is obtained from a minor adjustment of the well known algorithm class of Dafermos [196] and therefore includes simple modifications of a majority of the iterative algorithms applied to traffic equilibria as special cases.

We conclude by supplying a list of references to known asymmetric test networks.

City	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{C}\rvert$	# Centroids	Reference
	9	13	4	4	[719]
	7	24	6	6	[393]
	20	28	8		[699]
	22	36	12		[340]
	25	40	5	5	[78]
	25	37	6		[699]
	40	66	6		[699]
Hull	501	798	138	23	[647]
Barcelona	930	2522	7922	110	[684]
Barcelona	2199	5022	7286	90	[684]
Winnipeg	1017	2976	4345	154	[684]

Table 5.2: Asymmetric test networks

Appendix A

Definitions

In this appendix we collect the definitions and abbreviations of various concepts used in the book. In all definitions, X is a nonempty, closed and convex subset of $\Re^n$, T a function from X to $\Re$ and F a function from X to $\Re^n$.

Properties of functions

Definition A.1 (Convexity)

(a) *$T \in C^1$ on X is pseudoconvex on X if*

$$\nabla T(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0 \Rightarrow T(\mathbf{x}) \geq T(\mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.1}$$

(b) *T is convex on X if*

$$T(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda T(\mathbf{x}) + (1-\lambda)T(\mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X,\ \forall \lambda \in [0,1]\,. \tag{A.2}$$

If $T \in C^1$ on X, then (A.2) is equivalent to both the following statements (see [43, Th. 3.3.3] and [43, Th. 3.3.4], respectively).

$$T(\mathbf{x}) \geq T(\mathbf{y}) + \nabla T(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X \tag{A.3}$$

$$\left[\nabla T(\mathbf{x}) - \nabla T(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0, \qquad \forall \mathbf{x}, \mathbf{y} \in X \tag{A.4}$$

If $T \in C^2$ on X, then (A.2) is equivalent to the following [43, Th. 3.3.7].

$$\mathbf{y}^{\mathrm{T}}\nabla^2 T(\mathbf{x})\mathbf{y} \geq 0, \qquad \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in \Re^n \tag{A.5}$$

(c) *T is strictly convex on X if*

$$T(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) < \lambda T(\mathbf{x}) + (1-\lambda)T(\mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y},\ \forall \lambda \in (0,1)\,. \tag{A.6}$$

If $T \in C^1$ on X, then (A.6) is equivalent to both the following statements (see [43, Th. 3.3.3] and [43, Th. 3.3.4], respectively).

$$T(\mathbf{x}) > T(\mathbf{y}) + \nabla T(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y} \tag{A.7}$$

$$\left[\nabla T(\mathbf{x}) - \nabla T(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) > 0, \qquad \forall \mathbf{x}, \mathbf{y} \in X, \mathbf{x} \neq \mathbf{y} \tag{A.8}$$

(d) *T is strongly (uniformly) convex (with modulus m_T) on X if there exists a positive constant m_T such that*

$$T(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda T(\mathbf{x}) + (1-\lambda)T(\mathbf{y}) - \frac{m_T}{2}\lambda(1-\lambda)\left\|\mathbf{x}-\mathbf{y}\right\|^2, \tag{A.9}$$

$\forall \mathbf{x}, \mathbf{y} \in X,\ \forall \lambda \in [0,1]\,.$

If $T \in C^1$ on X, then (A.9) is equivalent to both the following statements (see [760, Sec. 1.1.4, Le. 3] and [612], respectively).

$$T(\mathbf{x}) \geq T(\mathbf{y}) + \nabla T(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) + \frac{m_T}{2}\|\mathbf{x}-\mathbf{y}\|^2, \qquad \forall \mathbf{x},\mathbf{y} \in X \tag{A.10}$$

$$\left[\nabla T(\mathbf{x}) - \nabla T(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq m_T\|\mathbf{x}-\mathbf{y}\|^2, \qquad \forall \mathbf{x},\mathbf{y} \in X \tag{A.11}$$

If $T \in C^2$ on X, then (A.9) is equivalent to the following [760, Sec. 1.1.4].

$$\mathbf{y}^{\mathrm{T}}\left(\nabla^2 T(\mathbf{x}) - m_T\mathbf{I}\right)\mathbf{y} \geq 0, \qquad \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in \Re^n \tag{A.12}$$

Definition A.2 (Monotonicity)

(a) *F is pseudomonotone on X if*

$$F(\mathbf{y})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0 \Rightarrow F(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0, \qquad \forall \mathbf{x},\mathbf{y} \in X. \tag{A.13}$$

(b) *F is monotone on X if*

$$\left[F(\mathbf{x}) - F(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq 0, \qquad \forall \mathbf{x},\mathbf{y} \in X. \tag{A.14}$$

If F is in C^1 on X, then (A.14) is equivalent to the following [727, Th. 5.4.3].

$$\mathbf{y}^{\mathrm{T}}\nabla F(\mathbf{x})\mathbf{y} \geq 0, \qquad \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in \Re^n \tag{A.15}$$

(c) *F is strictly monotone on X if*

$$\left[F(\mathbf{x}) - F(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) > 0, \qquad \forall \mathbf{x},\mathbf{y} \in X, \mathbf{x} \neq \mathbf{y}. \tag{A.16}$$

(d) *F is co-coercive on X if there exists a positive constant α_F such that*

$$\|F(\mathbf{x}) - F(\mathbf{y})\|^2 \leq \alpha_F\left[F(\mathbf{x}) - F(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}), \qquad \forall \mathbf{x},\mathbf{y} \in X. \tag{A.17}$$

(e) *F is strongly (uniformly) monotone on X if there exists a positive constant m_F such that*

$$\left[F(\mathbf{x}) - F(\mathbf{y})\right]^{\mathrm{T}}(\mathbf{x}-\mathbf{y}) \geq m_F\|\mathbf{x}-\mathbf{y}\|^2, \qquad \forall \mathbf{x},\mathbf{y} \in X. \tag{A.18}$$

If F is in C^1 on X, then (A.18) is equivalent to the following [727, Th. 5.4.3].

$$\mathbf{y}^{\mathrm{T}}\left(\nabla F(\mathbf{x}) - m_F\mathbf{I}\right)\mathbf{y} \geq 0, \qquad \forall \mathbf{x} \in X,\ \forall \mathbf{y} \in \Re^n \tag{A.19}$$

Definition A.3 (Coercivity)

(a) *T is weakly coercive on X if*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} T(\mathbf{x}) = +\infty. \tag{A.20}$$

(b) *T is coercive on X if*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} \frac{T(\mathbf{x})}{\|\mathbf{x}\|} = +\infty. \tag{A.21}$$

(c) *F is coercive on X if there exists a vector $\mathbf{x}^0 \in X$ such that*

$$\lim_{\substack{\mathbf{x}\in X \\ \|\mathbf{x}\|\to+\infty}} \frac{F(\mathbf{x})^{\mathrm{T}}(\mathbf{x}-\mathbf{x}^0)}{\|\mathbf{x}\|} = +\infty. \tag{A.22}$$

Definition A.4 (Lipschitz continuity) *F is Lipschitz continuous (with modulus M_F) on X if there exists a nonnegative constant M_F such that*

$$\|F(\mathbf{x}) - F(\mathbf{y})\| \leq M_F \|\mathbf{x} - \mathbf{y}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.23}$$

Definition A.5 (Nonexpansiveness) *Let F be a mapping from X to X, and let $\mathbf{x}^*$ be a fixed point of F.*

(a) *F is nonexpansive if*

$$\|F(\mathbf{x}) - F(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.24}$$

(b) *F is firmly nonexpansive if*

$$\|F(\mathbf{x}) - F(\mathbf{y})\|^2 \leq [F(\mathbf{x}) - F(\mathbf{y})]^{\mathrm{T}} (\mathbf{x} - \mathbf{y}), \qquad \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.25}$$

(c) *F is pseudocontractive with modulus $\alpha \in [0, 1)$ if*

$$\|F(\mathbf{x}) - \mathbf{x}^*\| \leq \alpha \|\mathbf{x} - \mathbf{x}^*\|, \qquad \forall \mathbf{x} \in X. \tag{A.26}$$

(d) *F is contractive with modulus $\alpha \in [0, 1)$ if*

$$\|F(\mathbf{x}) - F(\mathbf{y})\| \leq \alpha \|\mathbf{x} - \mathbf{y}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in X. \tag{A.27}$$

Definition A.6 [779] (Semicontinuity) *Let $T : \Re^n \mapsto \Re \cup \{-\infty, +\infty\}$ be a given function.*

(a) *T is lower semicontinuous (l.s.c.) on $\Re^n$ if*

$$T(\mathbf{x}) = \liminf_{\mathbf{y} \to \mathbf{x}} T(\mathbf{y}), \qquad \forall \mathbf{x} \in \Re^n.$$

(b) *T is upper semicontinuous (u.s.c.) on $\Re^n$ if*

$$T(\mathbf{x}) = \limsup_{\mathbf{y} \to \mathbf{x}} T(\mathbf{y}), \qquad \forall \mathbf{x} \in \Re^n.$$

Definition A.7 [722, 501] (Closedness) *Let $F : X \mapsto 2^Y$ be a point-to-set map.*

(a) *F is closed at $\mathbf{x} \in X$ if*

$$\left.\begin{array}{r} \{\mathbf{x}^k\} \to \mathbf{x} \\ \mathbf{y}^k \in F(\mathbf{x}^k), \{\mathbf{y}^k\} \to \mathbf{y} \end{array}\right\} \Longrightarrow \mathbf{y} \in F(\mathbf{x}).$$

(b) *F is upper semicontinuous (u.s.c.) at $\mathbf{x} \in X$ if for any neighbourhood $N(F(\mathbf{x}))$ there is a neighbourhood $N(\mathbf{x})$ with*

$$\mathbf{z} \in N(\mathbf{x}) \Longrightarrow F(\mathbf{z}) \subset N(F(\mathbf{x})).$$

Algorithmic definitions

The following line search rules are discussed in the text (listed in the order of increasing computational simplicity). Assume that $\mathbf{p}^k$ is the search direction in iteration k, and that l_k is the chosen step length. The line search rules are given for the case where the feasible set $X = \Re^n$. When the set X is bounded in the direction $\mathbf{p}^k$ (assumed locally feasible), an upper bound on l_k must be introduced.

Rule M (Exact minimization) Choose l_k such that

$$T(\mathbf{x}^k + l_k \mathbf{p}^k) = \min_{l \geq 0} T(\mathbf{x}^k + l \mathbf{p}^k).$$

Rule G (Goldstein [431]) Let $0 < \mu_1 \leq \mu_2 < 1$ and choose l_k such that

$$\mu_1 \leq \frac{T(\mathbf{x}^k + l_k \mathbf{p}^k) - T(\mathbf{x}^k)}{l_k \nabla T(\mathbf{x}^k)^{\mathrm{T}} \mathbf{p}^k} \leq \mu_2.$$

Rule A (Armijo [26]) Let $\alpha \in (0,1)$ and $l_k = \varepsilon \beta^{\bar{\imath}}$, where $\beta \in (0,1), \varepsilon > 0$ and $\bar{\imath}$ is the smallest nonnegative integer i such that

$$\alpha \leq \frac{T(\mathbf{x}^k + \varepsilon \beta^i \mathbf{p}^k) - T(\mathbf{x}^k)}{\varepsilon \beta^i \nabla T(\mathbf{x}^k)^{\mathrm{T}} \mathbf{p}^k}. \tag{A.28}$$

Rule P (Predetermined steps) Let $0 < \omega < \Omega$. Choose l_k arbitrary in the interval $[\omega, \Omega - \omega]$. (A particular example is the fixed step length formula, $l_k = l \in (0, \Omega), \forall k$.)

Definition A.8 [623, 43] (Convergence rate) *Let the sequence* $\{\mathbf{x}^k\}$ *converge to* $\mathbf{x}^*$. *The rate of convergence of the sequence is the supremum of the nonnegative numbers* p *satisfying*

$$\limsup_{k \to \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^k - \mathbf{x}^* \right\|^p} = q < +\infty. \tag{A.29}$$

(a) *If* $p = 1$ *and* $q < 1$, *the sequence has a linear rate of convergence (or the rate of geometrical progression) with ratio* q.

(b) *If* $p > 1$, *or if* $p = 1$ *and* $q = 0$, *the sequence has a superlinear rate of convergence (or faster than any geometric progression).*

(c) *If* $p = 2$ *the sequence has a quadratic rate of convergence.*

References

[1] H. Z. Aashtiani, *The multi-modal traffic assignment problem*, Technical Summary 77-1, U. S. Department of Transportation, Washington, D.C., 1977.

[2] ——, *The multi-modal traffic assignment problem*, PhD thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1979.

[3] H. Z. Aashtiani and T. L. Magnanti, *Modeling and computing extended urban traffic equilibria*, draft report, U. S. Department of Transportation, Washington, D.C., 1977.

[4] ——, *Equilibria on a congested transportation network*, SIAM Journal on Algebraic and Discrete Methods, 2 (1981), pp. 213–226.

[5] ——, *A linearization and decomposition algorithm for computing urban traffic equilibria*, in Proceedings of the 1982 IEEE International Large Scale Systems Symposium, Virginia Beach, VA, 1982, pp. 8–19.

[6] M. Abdulaal and L. J. LeBlanc, *Methods for combining modal split and equilibrium assignment models*, Transportation Science, 13 (1979), pp. 292–314.

[7] W. T. Adams, *Factors influencing transit and automobile use in urban areas*, Highway Research Board Bulletin, 230 (1959), pp. 101–111.

[8] D. L. Adolphson, *A note on minimum cost convex flows*, Naval Research Logistics Quarterly, 23 (1976), pp. 713–714.

[9] M. Aganagic, *Variational inequalities and generalized complementarity problems*, Technical Report SOL 78-11, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[10] D. P. Ahlfeld, J. M. Mulvey, R. S. Dembo, and S. A. Zenios, *Nonlinear programming on generalized networks*, ACM Transactions on Mathematical Software, 13 (1987), pp. 350–367.

[11] B.-H. Ahn, *Computation of Market Equilibria for Policy Analysis: The Project Independence Evaluation Study (PIES) Approach*, Garland, New York, NY, 1979.

[12] ——, *A Gauss–Seidel iteration method for nonlinear variational inequality problems over rectangles*, Operations Research Letters, 1 (1982), pp. 117–120.

[13] B.-H. Ahn and W. W. Hogan, *On convergence of the PIES algorithm for computing equilibria*, Operations Research, 30 (1982), pp. 281–300.

[14] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[15] R. Akcelik, *On Davidson's flow rate/travel time relationship*, Journal of the Australian Road Research Board, 8 (1978), pp. 41–44.

[16] ——, *A graphical explanation of the two principles and two techniques of traffic assignment*, Transportation Research, 13A (1979), pp. 179–184.

[17] J. Almond, *Traffic assignment to a road network with journey-time/flow relations*, Laboratory Note LN 570, Road Research Laboratory, Crowthorne, Berkshire, England, 1964.

[18] ——, *Traffic assignment to a road network*, Traffic Engineering & Control, 6 (1965), pp. 616–617, 622.

[19] ——, *Traffic assignment with flow-dependent journey times*, in Vehicular Traffic Science, Proceedings of the 3rd International Symposium on the Theory of Traffic Flow, New York, June 1965, L. C. Edie, R. Herman, and R. Rothery, eds., American Elsevier, New York, NY, 1967, pp. 222–234.

[20] K. M. Anantharamaiah, *Equilibrium conditions in traffic assignment*, in Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney, August 26–28, 1974, D. J. Buckley, ed., Elsevier, New York, NY, 1974, pp. 483–493.

[21] P.-A. Andersson, *On the convergence of iterative methods for the distribution balancing problem*, Transportation Research, 15B (1981), pp. 173–201.

[22] Y. Arezki, *Algorithms for the traffic assignment problem with fixed demand*, PhD thesis, Institute for Transport Studies, University of Leeds, Leeds, 1987.

[23] Y. Arezki, *The SPI (stepwise path increment) algorithm applied to solve the traffic assignment problem with fixed demand*, in Mathematics in Transport Planning and Control, Based on the Proceedings of a Conference on Mathematics in Transport Planning and Control Organized by The Institute of Mathematics and its Applications and Held at the University of Wales College of Cardiff in September 1989, J. D. Griffiths, ed., Clarendon Press, Oxford, 1992, pp. 133–143.

[24] Y. Arezki and D. Van Vliet, *The use of quantal loading in equilibrium traffic assignment*, Transportation Research, 19B (1985), pp. 521–525.

[25] ——, *A full analytical implementation of the PARTAN/Frank–Wolfe algorithm for equilibrium assignment*, Transportation Science, 24 (1990), pp. 58–62.

[26] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific Journal of Mathematics, 16 (1966), pp. 1–3.

[27] B. D. Armstrong, *The need for route guidance*, Supplementary Report SR330, Department of Transport, Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1977.

[28] R. Arnott, A. de Palma, and R. Lindsey, *Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering*, Transportation Science, 27 (1993), pp. 148–160.

[29] R. Arnott, R. de Palma, and R. Lindsey, *Economics of a bottleneck*, Journal of Urban Economics, 27 (1990), pp. 111–130.

[30] R. L. Asmuth, *Traffic network equilibria*, PhD thesis, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[31] ——, *Traffic network equilibria*, Technical Report SOL-78-2, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[32] A. A. Assad, *Multicommodity network flows: a survey*, Networks, 8 (1978), pp. 37–91.

[33] G. Auchmuty, *Variational principles for variational inequalities*, Numerical Functional Analysis and Optimization, 10 (1989), pp. 863–874.

[34] A. Auslender, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[35] A. Babin, M. Florian, L. James-Lefevbre, and H. Spiess, *EMME/2: interactive graphic method for road and transit planning*, Highway Research Record, 866 (1982), pp. 1–9.

[36] M. Bacharach, *Biproportional Matrices and Input-Output Change*, Cambridge University Press, Cambridge, 1970.

[37] J. E. Baerwald, ed., *Transportation and Traffic Engineering Handbook*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[38] A. B. Bakušinskiĭ and B. T. Poljak, *On the solution of variational inequalities*, Soviet Mathematics Doklady, 15 (1974), pp. 1705–1710.

[39] J. Barceló. Private communication, 1990.

[40] R. R. Barton and D. W. Hearn, *Network aggregation in transportation planning models*, Report DOT-TSC-RSPA-79-18, Mathtec, Inc., Princeton, NJ, 1979.

[41] R. R. Barton, D. W. Hearn, and S. Lawphongpanich, *The equivalence of transfer and generalized Benders decomposition methods for traffic assignment*, Transportation Research, 23B (1989), pp. 61–73.

[42] M. S. Bazaraa, J. J. Goode, and C. M. Shetty, *Constraint qualifications revisited*, Management Science, 18 (1972), pp. 567–573.

[43] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY, second ed., 1993.

[44] M. S. Bazaraa and C. M. Shetty, *Foundations of Optimization*, vol. 122 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976.

[45] E. M. L. Beale, *An algorithm for solving the transportation problem when the shipping cost over each route is convex*, Naval Research Logistics Quarterly, 6 (1959), pp. 43–56.

[46] P. Beck, L. Lasdon, and M. Engquist, *A reduced gradient algorithm for nonlinear network problems*, ACM Transactions on Mathematical Software, 9 (1983), pp. 57–70.

[47] M. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*, Yale University Press, New Haven, CT, 1956.

[48] M. J. Beckmann, *On the theory of traffic flow in networks*, Traffic Quarterly, 21 (1967), pp. 109–117.

[49] M. J. Beckmann and T. F. Golob, *A critique of entropy and gravity in travel forecasting*, in Traffic Flow and Transportation, Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation, G. F. Newell, ed., American Elsevier, New York, NY, 1972, pp. 109–117.

[50] ——, *Traveler decisions and traffic flows: a behavioral theory of network equilibrium*, in Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney, August 26–28, 1974, D. J. Buckley, ed., Elsevier, New York, NY, 1974, pp. 453–482.

[51] M. J. Beckmann and J. P. Wallace, III, *Evaluation of user benefits arising from changes in transportation systems*, Transportation Science, 3 (1969), pp. 344–351.

[52] J. B. Behr, ed., *Research on Road Traffic*, Her Majesty's Stationery Office, London, 1965.
[53] M. H. Beilby, *Economics and Operational Research*, Academic Press, New York, NY, 1976.
[54] M. C. Bell and L. D. Bennett, *Investigating states of equilibrium in traffic assignment*, in Proceedings of the 2nd Meeting of the EURO Working Group on Urban Traffic and Transportation, Paris, France, September 15–17, 1993, F. Boillot, N. Bhouri, and F. Laurent, eds., vol. 38 of Actes INRETS, Institut National de Recherche sur les Transport et leur Sécurité (INRETS), Arcueil, France, 1993, pp. 113–124.
[55] M. G. H. Bell, W. H. K. Lam, G. Ploss, and D. Inaudi, *Stochastic user equilibrium assignment and iterative balancing*, in Transportation and Traffic Theory, Proceedings of the 12th International Symposium on the Theory of Traffic Flow and Transportation, Berkeley, CA, July 21–23, 1993, C. F. Daganzo, ed., Elsevier, Amsterdam, 1993, pp. 427–439.
[56] R. Bellman, *On a routing problem*, Quarterly of Applied Mathematics, 16 (1958), pp. 87–90.
[57] M. Ben-Akiva, M. J. Bergman, A. J. Daly, and R. Ramaswamy, *Modelling inter urban route choice behaviour*, in Proceedings of the 9th International Symposium on Transportation and Traffic Theory, Delft, The Netherlands, 11–13 July 1984, J. Volmuller and R. Hamerslag, eds., VNU Science Press, Utrecht, The Netherlands, 1984, pp. 299–330.
[58] O. Ben-Ayed, C. E. Blair, D. E. Boyce, and L. J. LeBlanc, *Construction of a real-world bilevel linear programming model of the highway network design problem*, in Hierarchical Optimization, G. Anandalingam and T. L. Friesz, eds., vol. 34 of Annals of Operations Research, J. C. Baltzer AG, Basel, Switzerland, 1992, pp. 219–254.
[59] L. D. Bennett, *The existence of equivalent mathematical programs for certain mixed equilibrium traffic assignment problems*, European Journal of Operational Research, 71 (1993), pp. 177–187.
[60] J. A. Benshoof, *Characteristics of drivers' route selection behaviour*, Traffic Engineering & Control, 11 (1970), pp. 604–606.
[61] A. Bensoussan, M. Goursat, and J. L. Lions, *Contrôle impulsionnel et inéquations quasi-variationnelles stationnaires*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 276 (1973), pp. 1279–1284.
[62] A. Bermúdez and C. Moreno, *Duality methods for solving variational inequalities*, Computers and Mathematics with Applications, 7 (1981), pp. 43–58.
[63] D. Bernstein, *Programmability of continuous and discrete network equilibria*, PhD thesis, University of Pennsylvania, 1990.
[64] D. Bernstein and T. E. Smith, *Equilibria for networks with lower semicontinuous costs: with an application to congestion pricing*, technical report, Massachusetts Institute of Technology, Cambridge, MA, 1992.
[65] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, second ed., 1992.
[66] D. P. Bertsekas, *Combined primal-dual and penalty methods for constrained optimization*, SIAM Journal on Control, 13 (1975), pp. 521–544.
[67] ——, *On the Goldstein–Levitin–Polyak gradient projection method*, IEEE Transactions on Automatic Control, AC-21 (1976), pp. 174–184.
[68] ——, *Algorithms for nonlinear multicommodity network flow problems*, in Proceedings of the International Symposium on Systems Optimization and Analysis, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, NY, 1979, pp. 210–224.
[69] ——, *A class of optimal routing algorithms for communication networks*, in Proceedings of the 5th International Conference on Computer Communications, Atlanta, GA, 1980, pp. 71–76.
[70] ——, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, San Diego, CA, 1982.
[71] ——, *Optimal routing and flow control methods for communication networks*, in Proceedings of the International Symposium on Systems Optimization and Analysis, Versailles, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, NY, 1982, pp. 615–643.
[72] ——, *Projected Newton methods for optimization problems with simple constraints*, SIAM Journal on Control and Optimization, 20 (1982), pp. 221–246.
[73] ——, *An auction algorithm for shortest paths*, SIAM Journal on Optimization, 1 (1991), pp. 425–447.
[74] ——, *Linear Network Optimization: Algorithms and Codes*, MIT Press, Cambridge, MA, 1991.
[75] ——, *Auction algorithms for network flow problems: a tutorial introduction*, Computational Optimization and Applications, 1 (1992), pp. 7–66.
[76] D. P. Bertsekas and D. A. Castañon, *Parallel synchronous and asynchronous implementations of the auction algorithm*, Parallel Computing, 17 (1991), pp. 707–732.
[77] D. P. Bertsekas and D. El Baz, *Distributed asynchronous relaxation methods for convex network flow problems*, SIAM Journal on Control and Optimization, 25 (1987), pp. 74–85.
[78] D. P. Bertsekas and E. M. Gafni, *Projection methods for variational inequalities with application to the traffic assignment problem*, Mathematical Programming Study, 17 (1982), pp. 139–159.

[79] D. P. Bertsekas and E. M. Gafni, *Projected Newton methods and optimization of multicommodity flows*, IEEE Transactions on Automatic Control, AC-28 (1983), pp. 1090–1096.

[80] D. P. Bertsekas, E. M. Gafni, and R. G. Gallager, *Second derivative algorithms for minimum delay distributed routing in networks*, IEEE Transactions on Communications, COM-32 (1984), pp. 911–919.

[81] D. P. Bertsekas, E. M. Gafni, and K. S. Vastola, *Validation of algorithms for optimal routing of flow in networks*, in Proceedings of the 1979 IEEE Conference on Decision and Control, San Diego, CA, January 10–12, 1979, pp. 220–227.

[82] D. P. Bertsekas, B. Gendron, and W. K. Tsai, *Implementation of an optimal multicommodity network flow algorithm based on gradient projection and a path flow formulation*, Technical Report LIDS-P-1364, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[83] D. P. Bertsekas, P. A. Hosein, and P. Tseng, *Relaxation methods for network flow problems with convex arc costs*, SIAM Journal on Control and Optimization, 25 (1987), pp. 1219–1243.

[84] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, London, 1989.

[85] M. J. Best, *Optimization of nonlinear performance criteria subject to flow constraints*, in Proceedings of the 1975 Midwest Symposium on Circuits and Systems, Kansas City, KS, 1975, pp. 438–443.

[86] ——, *Equivalence of some quadratic programming algorithms*, Mathematical Programming, 30 (1984), pp. 71–87.

[87] H. W. Bevis, *Forecasting zonal traffic volumes*, Traffic Quarterly, (1956), pp. 207–222.

[88] G. Birkhoff, *A variational principle for nonlinear networks*, Quarterly of Applied Mathematics, 21 (1963/64), pp. 160–162.

[89] G. Birkhoff and J. B. Diaz, *Non-linear network problems*, Quarterly of Applied Mathematics, 13 (1956), pp. 431–443.

[90] R. E. Bixby and W. H. Cunningham, *Converting linear programs to network problems*, Mathematics of Operations Research, 5 (1980), pp. 321–357.

[91] R. E. Bixby and R. Fourer, *Finding embedded network rows in linear programs, I: extraction heuristics*, Management Science, 34 (1988), pp. 342–376.

[92] W. R. Blunden, *Some applications of linear programming to transportation and traffic problems*, tech. report, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA, 1956.

[93] ——, *The Land-Use/Transportation System*, Pergamon Press, Oxford, 1971.

[94] J. D. Bolland, M. D. Hall, and D. Van Vliet, *SATURN: a model for the evaluation of traffic management schemes*, Working Paper 106, Institute for Transport Studies, University of Leeds, Leeds, 1979.

[95] J. D. Bolland, M. D. Hall, D. Van Vliet, and L. G. Willumsen, *SATURN: simulation and assignment of traffic in urban road networks*, in Proceedings of the International Symposium on Traffic Control Systems, Berkeley, CA, 1979, pp. 99–115.

[96] P. Bonsall, *The influence of route guidance advice on route choice in urban networks*, Transportation, 19 (1992), pp. 1–23.

[97] R. K. Booth, *York: the History and Heritage of a City*, Barrie & Jenkins, London, 1990.

[98] P. Boulos and T. Altman, *A graph-theoretic approach to explicit nonlinear pipe network optimization*, Applied Mathematical Modelling, 15 (1991), pp. 459–466.

[99] P. H. L. Bovy and G. R. M. Jansen, *Network aggregation effects upon equilibrium assignment outcomes: an empirical investigation*, Transportation Science, 17 (1983), pp. 240–262.

[100] ——, *Spatial aggregation effects in equilibrium and all-or-nothing assignments*, Transportation Research Record, 931 (1983), pp. 98–106.

[101] D. E. Boyce, *A framework for constructing network equilibrium models of urban location*, Transportation Science, 14 (1980), pp. 77–96.

[102] ——, *Editorial*, Environment and Planning, 13A (1981), pp. 395–397.

[103] ——, *Network models in transportation/land use planning*, in Transportation Planning Models, Proceedings of the Course Given at The International Center for Transportation Studies (ICTS), Amalfi, Italy, October 11–16, 1982, M. Florian, ed., North-Holland, Amsterdam, 1984, pp. 475–498.

[104] ——, *Urban transportation network-equilibrium and design models: recent achievements and future prospects*, Environment and Planning, 16A (1984), pp. 1445–1474.

[105] ——, *Towards a research program for modeling the performance of highway networks utilized by guided and unguided vehicles*, Advance Working Paper Series 6, Task 4c, Urban Transportation Center, University of Illinois at Chicago, Chicago, IL, 1991.

[106] D. E. Boyce, N. D. Day, and C. McDonald, *Metropolitan Plan Making*, Regional Science Research Institute, Philadelphia, PA, 1970.

[107] D. E. Boyce, J. Hicks, and A. Sen, *In-vehicle navigation requirements for monitoring link travel times in a dynamic route guidance system*, Advance Working Paper Series 1 and 2, Task 1b, Urban Transportation Center, University of Illinois at Chicago, Chicago, IL, 1991. Paper presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C., January 16, 1991, and published in Operations Review, Chicago Area Transportation Study, 1991.

[108] D. E. Boyce and B. N. Janson, *Experiments with a network design algorithm incorporating a combined trip distribution and assignment model*, in Proceedings of the International Symposium on Travel Supply Models, Montreal, November, 1977.

[109] D. E. Boyce, B. N. Janson, and R. W. Eash, *The effect on equilibrium trip assignment of different link congestion functions*, Transportation Research, 15A (1981), pp. 223–232.

[110] D. E. Boyce, L. J. LeBlanc, and K. S. Chon, *Network equilibrium models of urban location and travel choices: a retrospective survey*, Journal of Regional Science, 28 (1988), pp. 159–183.

[111] D. E. Boyce, B. Ran, and L. J. LeBlanc, *Dynamic user-optimal traffic assignment: a new model and solution technique*, Advance Working Paper Series 3, Task 4c, Urban Transportation Center, University of Illinois at Chicago, Chicago, IL, 1991. Paper presented at the 1st Triennial Symposium on Transportation Analysis, Montreal, June 6–11, 1991.

[112] D. Braess, *Über ein Paradox der Verkehrsplannung*, Unternehmenstorchung, 12 (1968), pp. 258–268.

[113] D. Braess and G. Koch, *On the existence of equilibria in asymmetrical multiclass-user transportation networks*, Transportation Science, 13 (1979), pp. 56–63.

[114] D. Brand, *The state of the art of travel demand forecasting: a critical review*, tech. report, Graduate School of Design, Harvard University, Cambridge, MA, 1972.

[115] D. Branston, *Link capacity functions: a review*, Transportation Research, 10 (1976), pp. 223–236.

[116] L. M. Bregman, *A relaxation method of finding a common point of convex sets and its application to problems of optimization*, Soviet Mathematics Doklady, 7 (1966), pp. 1578–1581.

[117] ——, *Proof of the convergence of Sheliekhovskii's method for a problem with transportation constraints*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 191–204.

[118] ——, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[119] H. Brézis, *équations et inéquations non linéaires dans les espaces vectoriels en dualité*, Annales de l'Institut Fourier, 18 (1968), pp. 115–175.

[120] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[121] G. E. Brokke and W. L. Mertz, *Evaluating trip forecasting methods with an electronic computer*, Highway Research Board Bulletin, 203 (1958), pp. 52–75.

[122] L. E. J. Brouwer, *Über Abbildung von Mannigfaltigkeiten*, Mathematische Annalen, 71 (1912), pp. 97–115.

[123] F. E. Browder, *Existence and approximation of solutions of nonlinear variational inequalities*, Proceedings of the National Academy of Sciences of the United States of America, 56 (1966), pp. 1080–1086.

[124] G. G. Brown and R. D. McBride, *Solving generalized networks*, Management Science, 30 (1984), pp. 1497–1523.

[125] G. G. Brown, R. D. McBride, and R. K. Wood, *Extracting embedded generalized networks from linear programming problems*, Mathematical Programming, 32 (1985), pp. 11–31.

[126] G. G. Brown and W. G. Wright, *Automatic identification of embedded network rows in large-scale optimization models*, Mathematical Programming, 29 (1984), pp. 41–56.

[127] R. M. Brown, *Expressway route selection and vehicular usage*, Highway Research Board Bulletin, 16 (1947), pp. 12–21.

[128] R. M. Brown and H. H. Weaver, *Traffic assignment using IBM computations and summation*, Highway Research Board Bulletin, 130 (1956), pp. 47–58.

[129] P. Brucker, *An $O(n)$ algorithm for quadratic knapsack problems*, Operations Research Letters, 3 (1984), pp. 163–166.

[130] M. J. Bruton, *Introduction to Transportation Planning*, Hutchinson of London, London, second ed., 1975.

[131] M. Bruynooghe, *Affectation du trafic sur un multi-réseau*, tech. report, Institut de Recherche des Transports, Arcueil, France, 1967.

[132] M. Bruynooghe, A. Gibert, and M. Sakarovitch, *Une méthode d'affectation du trafic*, in Proceedings of the 4th International Symposium on the Theory of Road Traffic Flow, Karlsruhe, 1968, W. Leutzbach and P. Baron, eds., Beiträge zur Theorie des Verkehrsflusses Strassenbau und Strassenverkehrstechnik, Heft 86, Herausgegeben von Bundesminister für Verkehr, Abteilung Strassenbau, Bonn, 1969, pp. 198–204.

[133] C. Buchanan, G. H. C. Cooper, A. MacEwen, D. H. Crompton, G. Crow, G. Michell, D. Dallimore, P. J. Hills, D. Burton, A. H. Penfold, M. L. Haxworth, A. G. Richardson, C. Woodward, N. Lichfield, and K. Browne, *Traffic in towns: a study of the long term problems of traffic in urban areas*, report of the working group appointed by the minister of transport, Her Majesty's Stationery Office, London, 1963. Chapter 2 reprinted in [693, pp. 153–183].

[134] J. M. Buchanan, *Peak loads and efficient pricing: comment*, Quarterly Journal of Economics, 80 (1966), pp. 463–480.

[135] J. E. Burrell, *Multiple route assignment and its application to capacity restraint*, in Proceedings of the 4th International Symposium on the Theory of Road Traffic Flow, Karlsruhe, 1968, W. Leutzbach and P. Baron, eds., Beiträge zur Theorie des Verkehrsflusses Strassenbau und Strassenverkehrstechnik, Heft 86, Herausgegeben von Bundesminister für Verkehr, Abteilung Strassenbau, Bonn, 1969, pp. 210–219.

[136] ——, *Multiple route assignment: a comparison of two methods*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 229–239.

[137] E. W. Campbell, *A mechanical method for assigning traffic to expressways*, Highway Research Board Bulletin, 130 (1956), pp. 27–46.

[138] M. E. Campbell, *Route selection and traffic assignment*, tech. report, Highway Research Board Correlation Service, 1950.

[139] ——, *Foreword*, Highway Research Board Bulletin, 61 (1952), pp. iii–iv.

[140] M. D. Canon and C. D. Cullum, *A tight upper bound on the rate of convergence of the Frank–Wolfe algorithm*, SIAM Journal on Control, 6 (1968), pp. 509–516.

[141] D. G. Cantor and M. Gerla, *Optimal routing in a packet-switched computer network*, IEEE Transactions on Computers, c-23 (1974), pp. 1062–1069.

[142] C. Carathéodory, *Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen*, Rendiconti del Circolo Matematico di Palermo, 32 (1911), pp. 193–217.

[143] H. C. Carey, *Principles of Social Science*, J. B. Lippincott & Co., Philadelphia, PA, 1858/59.

[144] M. Carey, *Integrability and mathematical programming models: a survey and a parametric approach*, Econometrica, 45 (1977), pp. 1957–1976.

[145] ——, *The dual of the traffic assignment problem with elastic demands*, Transportation Research, 19B (1985), pp. 227–237.

[146] ——, *A constraint qualification for a dynamic traffic assignment model*, Transportation Science, 20 (1986), pp. 55–58.

[147] ——, *Network equilibrium: optimization formulations with both quantities and prices as variables*, Transportation Research, 21B (1987), pp. 69–77.

[148] ——, *Optimal time-varying flows on congested networks*, Operations Research, 35 (1987), pp. 58–69.

[149] ——, *Nonconvexity of the dynamic traffic assignment problem*, Transportation Research, 26B (1992), pp. 127–133.

[150] M. Carey and K. Siddharthan, *Simultaneous estimation of demand function parameters, trip distribution and equilibrium assignment*, in Modeling and Simulation, the 13th Annual Conference Papers, Pittsburgh, PA, 1982, pp. 1481–1489.

[151] J. D. Carroll, Jr., *A method of traffic assignment to an urban network*, Highway Research Board Bulletin, 224 (1959), pp. 64–71.

[152] J. D. Carroll, Jr. and H. W. Bevis, *Predicting local travel in urban regions*, Papers and Proceedings of the Regional Science Association, 3 (1957), pp. 183–197.

[153] M. F. Carvalho, S. Soares, and P. F. Cuervo, *Convex-cost network flow algorithm with additional linear constraints applied to electrical power generation and transmission systems*, in Modelling, Simulation and Optimization, Proceedings of the IASTED International Symposium, Montreal, Canada, May 22–24, 1990, pp. 87–90.

[154] E. Cascetta and S. Nguyen, *A unified framework for estimating or updating origin/destination matrices from traffic counts*, Transportation Research, 22B (1988), pp. 437–455.

[155] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, SIAM Review, 23 (1981), pp. 444–466.

[156] Y. Censor and G. T. Herman, *On some optimization techniques in image reconstruction from projections*, Applied Numerical Mathematics, 3 (1987), pp. 365–391.

[157] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, Journal of Optimization Theory and Applications, 34 (1981), pp. 321–353.

[158] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with D-functions*, Journal of Optimization Theory and Applications, 73 (1992), pp. 451–464.

[159] E. D. Chajakis and S. A. Zenios, *Synchronous and asynchronous implementations of relaxation algorithms for nonlinear network optimization*, Parallel Computing, 17 (1991), pp. 873–894.

[160] A. Charnes and W. W. Cooper, *Extremal principles for simulating traffic flow in a network*, Proceedings of the National Academy of Sciences of the United States of America, 44 (1958), pp. 201–204.

[161] ——, *Nonlinear network flows and convex programming over incidence matrices*, Naval Research Logistics Quarterly, 5 (1958), pp. 231–240.

[162] ——, *Multicopy traffic network models*, in Theory of Traffic Flow, Proceedings of the Symposium on the Theory of Traffic Flow Held at the General Motors Research Laboratories, Warren, MI, December 7–8, 1959, R. Herman, ed., Elsevier, Amsterdam, 1961, pp. 85–96.

[163] R. S. Chech and A. D. Smith, *The transshipment problem with quadratic costs*, group research project, Queen's University, School of Business, Kingston, Canada, 1971.

[164] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, 3 (1993), pp. 538–543.

[165] M. Chen and A. S. Alfa, *Algorithms for solving Fisk's stochastic traffic assignment model*, Transportation Research, 25B (1991), pp. 405–412.

[166] ——, *A network design algorithm using a stochastic incremental traffic assignment approach*, Transportation Science, 25 (1991), pp. 215–224.

[167] R.-J. Chen, *Parallel algorithms for a class of convex optimization problems*, PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, 1987.

[168] R.-J. Chen and R. R. Meyer, *A scaled trust region method for a class of convex optimization problems*, Technical Report 675, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, 1986.

[169] ——, *Parallel optimization for traffic assignment*, Mathematical Programming, 42 (1988), pp. 327–345.

[170] Chicago Area Transportation Study, *Final Report, Volume II: Data Projections*, Illinois Department of Public Works and Buildings, Chicago, IL, 1960.

[171] C. Chu, *A review of the development and theoretical concepts of traffic assignment techniques and their practical applications to an urban road network*, Traffic Engineering & Control, 13 (1971), pp. 136–141.

[172] C. E. Clark, *The greatest of a finite set of random variables*, Operations Research, 9 (1961), pp. 145–162.

[173] R. H. Clark, J. L. Kennington, R. R. Meyer, and M. Ramamurti, *Generalized networks: parallel algorithms and an empirical analysis*, ORSA Journal on Computing, 4 (1992), pp. 132–145.

[174] G. Cohen, *Optimization by decomposition and coordination: a unified approach*, IEEE Transactions on Automatic Control, AC-23 (1978), pp. 222–232.

[175] ——, *Auxiliary problem principle and decomposition of optimization problems*, Journal of Optimization Theory and Applications, 32 (1980), pp. 277–305.

[176] ——, *Auxiliary problem principle extended to variational inequalities*, Journal of Optimization Theory and Applications, 59 (1988), pp. 325–333.

[177] J. E. Cohen and P. Horowitz, *Paradoxical behaviour of mechanical and electrical networks*, Nature, 352 (1991), pp. 699–701.

[178] M. Collins, L. Cooper, R. Helgason, J. Kennington, and L. LeBlanc, *Solving the pipe network analysis problem using optimization techniques*, Management Science, 24 (1978), pp. 747–760.

[179] D. C. Colony, *An application of game theory to route selection*, Highway Research Record, 334 (1970), pp. 39–47.

[180] L. Cooper and J. Kennington, *Steady-state analysis of nonlinear resistive electrical networks using optimization techniques*, Technical Report IEOR 77012, Department of Operations Research and Engineering Management, School of Engineering and Applied Science, Southern Methodist University, Dallas, TX, 1977.

[181] L. Cooper and L. J. LeBlanc, *Stochastic transportation problems and other network related convex problems*, Naval Research Logistics Quarterly, 24 (1977), pp. 327–337.

[182] R. W. Cottle, *Nonlinear programs with positively bounded Jacobians*, SIAM Journal on Applied Mathematics, 14 (1966), pp. 147–158.

[183] R. W. Cottle and A. Djang, *Algorithmic equivalence in quadratic programming, I: a least-distance programming problem*, Journal of Optimization Theory and Applications, 28 (1979), pp. 275–301.

[184] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, NY, 1992.

[185] R. W. Cottle, J.-S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra and Its Applications, 114/115 (1989), pp. 231–249.

[186] A. Cournot, *Researches into the Mathematical Principles of the Theory of Wealth*, Macmillan, London, 1838.

[187] R. L. Creighton, *Urban Transportation Planning*, University of Illinois Press, Urbana, IL, 1970.

[188] J. E. Cremeans, R. A. Smith, and G. R. Tyndall, *Optimal multicommodity network flows with resource allocation*, Naval Research Logistics Quarterly, 17 (1970), pp. 269–280.

[189] H. Cross, *Analysis of flow in networks of conduits or conductors*, Bulletin 286, Engineering Experiment Station, University of Illinois, Urbana, IL, 1936.

[190] N. D. Curet, *On the dual coordinate ascent approach for nonlinear networks*, Computers and Operations Research, 20 (1992), pp. 133–140.

[191] G. Cybenko, *Dynamic load balancing for distributed memory multiprocessors*, Technical Report 87-1, Department of Computer Science, Tufts University, Medford, MA, 1987.

[192] S. Dafermos, *Traffic equilibrium and variational inequalities*, Transportation Science, 14 (1980), pp. 42–54.

[193] ——, *The general multimodal network equilibrium problem with elastic demand*, Networks, 12 (1982), pp. 57–72.

[194] ——, *Relaxation algorithms for the general asymmetric traffic equilibrium problem*, Transportation Science, 16 (1982), pp. 231–240.

[195] ——, *Convergence of a network decomposition algorithm for the traffic equilibrium model*, in Proceedings of the 8th International Symposium on Transportation and Traffic Theory, Toronto, June 24–26, 1981, V. F. Hurdle, E. Hauer, and G. N. Steuart, eds., University of Toronto Press, Toronto, 1983, pp. 143–156.

[196] ——, *An iterative scheme for variational inequalities*, Mathematical Programming, 26 (1983), pp. 40–47.

[197] ——, *Isomorphic multiclass spatial price and multimodal traffic network equilibrium models*, Regional Science and Urban Economics, 16 (1986), pp. 197–209.

[198] ——, *Sensitivity analysis in variational inequalities*, Mathematics of Operations Research, 13 (1988), pp. 421–434.

[199] S. Dafermos and A. Nagurney, *A network formulation of market equilibrium problems and variational inequalities*, Operations Research Letters, 3 (1984), pp. 247–250.

[200] ——, *On some traffic equilibrium theory paradoxes*, Transportation Research, 18B (1984), pp. 101–110.

[201] ——, *Sensitivity analysis for the asymmetric network equilibrium problem*, Mathematical Programming, 28 (1984), pp. 174–184.

[202] ——, *Sensitivity analysis for the general spatial economic equilibrium problem*, Operations Research, 32 (1984), pp. 1069–1086.

[203] ——, *Stability and sensitivity analysis for the general network equilibrium-travel choice model*, in Proceedings of the 9th International Symposium on Transportation and Traffic Theory, Delft, The Netherlands, 11–13 July 1984, J. Volmuller and R. Hamerslag, eds., VNU Science Press, Utrecht, The Netherlands, 1984, pp. 217–231.

[204] S. Dafermos and F. T. Sparrow, *Optimal resource allocation and toll patterns in user-optimised transport networks*, Journal of Transportation Economy and Policy, 5 (1971), pp. 184–200.

[205] S. C. Dafermos, *An extended traffic assignment model with applications to two-way traffic*, Transportation Science, 5 (1971), pp. 366–389.

[206] ——, *The traffic assignment problem for multiclass-user transportation networks*, Transportation Science, 6 (1972), pp. 73–87.

[207] ——, *Toll patterns for multiclass-user transportation networks*, Transportation Science, 7 (1973), pp. 211–223.

[208] ——, *Integrated equilibrium flow models for transportation planning*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 106–118.

[209] S. C. Dafermos and F. T. Sparrow, *The traffic assignment problem for a general network*, Journal of Research of the National Bureau of Standards, 73B (1969), pp. 91–118.

[210] S.-S. C. Dafermos, *Traffic assignment and resource allocation in transportation networks*, PhD thesis, Johns Hopkins University, Baltimore, MD, 1968.

[211] C. F. Daganzo, *On the traffic assignment problem with flow dependent costs—I*, Transportation Research, 11 (1977), pp. 433–437.

[212] ——, *On the traffic assignment problem with flow dependent costs—II*, Transportation Research, 11 (1977), pp. 439–441.

[213] ——, *Some research on traffic assignment methodology selection*, Working Paper 7703, Institute of Transportation Studies, University of California, Berkeley, CA, 1977.

[214] ——, *Multinomial Probit: The Theory and Its Application to Demand Forecasting*, Academic Press, New York, NY, 1979.

[215] ——, *An equilibrium algorithm for the spatial aggregation problem of traffic assignment*, Transportation Research, 14B (1980), pp. 221–228.

[216] C. F. Daganzo, *Equilibrium analysis in transportation: the state-of-the-art and some new results*, Working Paper UCB-ITS-RR-80-1, Institute of Transportation Studies, University of California, Berkeley, CA, 1980.

[217] ——, *Network representation, continuum approximations and a solution to the spatial aggregation problem of traffic assignment*, Transportation Research, 14B (1980), pp. 229–239.

[218] ——, *Unconstrained extremal formulation of some transportation equilibrium problems*, Transportation Science, 16 (1982), pp. 332–360.

[219] ——, *Stochastic network equilibrium with multiple vehicle types and asymmetric, indefinite link cost Jacobians*, Transportation Science, 17 (1983), pp. 282–300.

[220] C. F. Daganzo and Y. Sheffi, *On stochastic models of traffic assignment*, Transportation Science, 11 (1977), pp. 253–274.

[221] O. Damberg, J. T. Lundgren, and M. Patriksson, *An algorithm for the stochastic user equilibrium problem*, in Proceedings of the First Meeting of the EURO Working Group on Urban Traffic and Transportation, Landshut, Germany, October 1–3, 1992, J. C. M. Baños, B. Friedrich, M. Papageorgiou, and H. Keller, eds., Technical University of Munich, Munich, Germany, 1992. Also as Report LiTH-MAT-R-92-48, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1992.

[222] J. M. Danskin, *The Theory of Max-Min*, Springer-Verlag, Berlin, 1967.

[223] G. B. Dantzig, *Discrete-variable extremum problems*, Operations Research, 5 (1957), pp. 266–277.

[224] ——, *On the shortest route through a network*, Management Science, 6 (1960), pp. 187–190.

[225] ——, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[226] G. B. Dantzig, S. F. Maier, and Z. F. Lansdowne, *The application of decomposition to transportation network analysis*, Interim Report DOT-TSC-OST-76-26, Control Analysis Corporation, Palo Alto, CA, 1976.

[227] G. B. Dantzig and P. Wolfe, *Decomposition principle for linear programs*, Operations Research, 8 (1960), pp. 101–111.

[228] ——, *The decomposition algorithm for linear programming*, Econometrica, 29 (1961), pp. 767–778.

[229] A. D'Auriac, *A propos de l'unicité de solution dans les problèmes de réseaux maillés*, La Houille Blanche, 2 (1947), pp. 209–211.

[230] K. B. Davidson, *A flow travel time relationship for use in transportation planning*, Proceedings of the Australian Road Research Board, 3 (1966), pp. 183–194.

[231] M. De Luca and A. Maugeri, *Quasi-variational inequalities and applications to equilibrium problems with elastic demand*, in Nonsmooth Optimization and Related Topics, Proceedings of the 4th Course of the International School of Mathematics on Nonsmooth Optimization and Related Topics, Erice, Sicily, June 20–July 1, 1988, F. H. Clarke, V. F. Dem'yanov, and F. Giannessi, eds., Plenum Press, New York, NY, 1989, pp. 61–77.

[232] ——, *Quasi-variational inequalities and applications to the traffic equilibrium problem: discussion of a paradox*, Journal of Computational and Applied Mathematics, 28 (1989), pp. 163–171.

[233] ——, *Variational inequalities applied to the study of paradoxes in equilibrium problems*, Optimization, 25 (1992), pp. 249–259.

[234] A. R. De Pierro and A. N. Iusem, *A relaxed version of Bregman's method for convex programming*, Journal of Optimization Theory and Applications, 51 (1986), pp. 421–440.

[235] R. S. Dembo, *The performance of NLPNET, a large-scale nonlinear network optimizer*, Mathematical Programming Study, 26 (1986), pp. 245–248.

[236] ——, *A primal truncated Newton algorithm with application to large-scale nonlinear network optimization*, Mathematical Programming Study, 31 (1987), pp. 43–71.

[237] R. S. Dembo and J. G. Klincewicz, *A scaled reduced gradient algorithm for network flow problems with convex separable costs*, Mathematical Programming Study, 15 (1981), pp. 125–147.

[238] R. S. Dembo, J. M. Mulvey, and S. A. Zenios, *Large-scale nonlinear network models and their application*, Operations Research, 37 (1989), pp. 353–372.

[239] R. S. Dembo and U. Tulowitzki, *Local convergence analysis for successive inexact quadratic programming methods*, Working Paper Series B # 78, School of Organization and Management, Yale University, New Haven, CT, 1984.

[240] ——, *Sequential truncated quadratic programming methods*, in Numerical Optimization 1984, Proceedings of the SIAM Conference on Numerical Optimization, Boulder, CO, June 12–14, 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Society of Industrial and Applied Mathematics, Philadelphia, PA, 1985, pp. 83–101.

[241] ——, *Computing equilibria on large multicommodity networks: an application of truncated quadratic programming algorithms*, Networks, 18 (1988), pp. 273–284.

[242] V. F. Dem'janov and A. M. Rubinov, *On the problem of minimization of a smooth functional with convex constraints*, Soviet Mathematics Doklady, 6 (1965), pp. 9–11.

[243] V. F. Dem'yanov and V. N. Malozemov, *Introduction to Minimax*, John Wiley & Sons, New York, NY, 1974.

[244] V. F. Dem'yanov and A. B. Pevnyi, *Numerical methods for finding saddle points*, USSR Computational Mathematics and Mathematical Physics, 12 (1972), pp. 11–52.
[245] V. F. Dem'yanov and A. M. Rubinov, *The minimization of a smooth convex functional on a convex set*, SIAM Journal on Control, 5 (1967), pp. 280–294.
[246] ——, *Approximate Methods in Optimization Problems*, American Elsevier, New York, NY, 1970.
[247] V. F. Dem'yanov and L. V. Vasil'ev, *Nondifferentiable Optimization*, Optimization Software, New York, NY, 1985.
[248] J. B. Dennis, *Mathematical Programming and Electrical Networks*, John Wiley & Sons, New York, NY, 1959.
[249] Detroit Metropolitan Area Traffic Study, *Part I: Data Summary and Interpretation*, Michigan State Highway Department, Detroit, MI, 1955.
[250] ——, *Part II: Future Traffic and a Long Range Expressway Plan*, Michigan State Highway Department, Detroit, MI, 1956.
[251] S. Devarajan, *A note on network equilibrium and noncooperative games*, Transportation Research, 15B (1981), pp. 421–426.
[252] R. Dial, F. Glover, D. Karney, and D. Klingman, *A computational analysis of alternative algorithms and labeling techniques for finding shortest path trees*, Networks, 9 (1979), pp. 215–248.
[253] R. B. Dial, *Algorithm 360: shortest-path forest with topological ordering*, Communications of the Association of Computing Machinery, 12 (1969), pp. 632–633.
[254] ——, *A probabilistic multipath traffic assignment model which obviates path enumeration*, Transportation Research, 5 (1971), pp. 83–111.
[255] J. W. Dickey, *Metropolitan Transportation Planning*, McGraw-Hill, New York, NY, second ed., 1983.
[256] T. J. Dickson, *Traffic assignment.* Unpublished note, Transport Studies Group, University College London, London, 1977.
[257] ——, *A note on traffic assignment and signal timings in a signal-controlled road network*, Transportation Research, 15B (1981), pp. 267–271.
[258] E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numerische Mathematik, 1 (1959), pp. 269–271.
[259] R. Dionne, *étude et extension d'un algorithme de Murchland*, INFOR, 16 (1978), pp. 132–146.
[260] R. Dionne and M. Florian, *Exact and approximate algorithms for optimal network design*, Networks, 9 (1979), pp. 37–59.
[261] Y. M. I. Dirickx and L. P. Jennergren, *Systems Analysis by Multilevel Methods*, John Wiley & Sons, Chichester, U. K., 1979.
[262] A. Djang, *Algorithmic equivalence in quadratic programming*, PhD thesis, Department of Operations Research, Stanford University, Stanford, CA, 1980.
[263] J. C. Dodu, T. Eve, and M. Minoux, *Implementation of a proximal algorithm for linearly constrained nonsmooth optimization problems and computational results*, Numerical Algorithms, 6 (1994), pp. 245–273.
[264] T. A. Domencich and D. McFadden, *Urban Travel Demand: A Behavioral Analysis*, vol. 93 of Contributions to Economic Analysis, North-Holland, Amsterdam, 1975.
[265] A. A. Douglas and R. J. Lewis, *Trip generation techniques*, Traffic Engineering & Control, 12 (1970), pp. 362–365.
[266] ——, *Trip generation techniques*, Traffic Engineering & Control, 12 (1971), pp. 532–535.
[267] P. D. C. Dow, *Models for strategic road assignment*, PhD thesis, Institute for Transport Studies, University of Leeds, Leeds, 1979.
[268] A. Downs, *The law of peak-hour expressway congestion*, Traffic Quarterly, 16 (1962), p. 393.
[269] S. E. Dreyfus, *An appraisal of some shortest-path algorithms*, Operations Research, 17 (1969), pp. 395–412.
[270] O. Drissi-Kaïtouni, *An algorithm for the decomposition of arc flows into path flows for the general spatial price equilibrium problem*, INFOR, 28 (1990), pp. 403–411.
[271] ——, *A variational inequality formulation of the dynamic traffic assignment problem*, European Journal of Operational Research, 71 (1993), pp. 188–204.
[272] O. Drissi-Kaïtouni and M. Gendreau, *A new dynamic traffic assignment model*, in Proceedings of the First Meeting of the EURO Working Group on Urban Traffic and Transportation, Landshut, Germany, October 1–3, 1992, J. C. M. Baños, B. Friedrich, M. Papageorgiou, and H. Keller, eds., Technical University of Munich, Munich, Germany, 1992.
[273] O. Drissi-Kaïtouni and J. T. Lundgren, *Bilevel origin-destination matrix estimation using a descent approach*, Report LiTH-MAT-R-1992-49, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1992.
[274] R. J. Duffin, *Nonlinear networks, I*, Bulletin of the American Mathematical Society, 52 (1946), pp. 833–838.
[275] ——, *Nonlinear networks, IIa*, Bulletin of the American Mathematical Society, 53 (1947), pp. 963–971.

[276] R. J. Duffin, E. L. Peterson, and C. Zener, *Geometric Programming: Theory and Application*, John Wiley & Sons, New York, NY, 1967.

[277] N. Dunford and J. T. Schwartz, *Linear Operators*, John Wiley & Sons, New York, NY, 1963.

[278] J. C. Dunn, *A simple averaging process for approximating the solutions of certain optimal control problems*, Journal of Mathematical Analysis and Applications, 48 (1974), pp. 875–894.

[279] ——, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM Journal on Control and Optimization, 17 (1979), pp. 187–211.

[280] ——, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM Journal on Control and Optimization, 18 (1980), pp. 473–487.

[281] ——, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM Journal on Control and Optimization, 18 (1980), pp. 659–674.

[282] C. Dupuis, *Le problème de la recherche d'un flot d'équilibre sur un réseau routier*, PhD thesis, Centre de recherche sur les transports, Université de Montréal, Montréal, 1983.

[283] C. Dupuis and J.-M. Darveau, *The convergence conditions of diagonalization and projection methods for fixed demand asymmetric network equilibrium problems*, Operations Research Letters, 5 (1986), pp. 149–155.

[284] C. Dupuis and S. Nguyen, *Sur le calcul d'un point d'équilibre dans le cas des coûts nonsymétriques: revue et nouveaux développements*, Publication 271, Centre de recherche sur les transports, Université de Montréal, Montréal, 1982.

[285] J. Dupuit, *De la mesure de l'utilité des travaux publics*, Annales des Ponts et Chaussées, 8 (1844), pp. 332–375. Translated as *On the measurement of the utility of public works* from the French by R. H. Barback for International Economic Papers 2, A. T. Peacock, R. Turvey, F. A. Lutz and E. Henderson, eds., Macmillan, London, 1952, pp. 83–110; translation reprinted in [693, pp. 19–57].

[286] J.-P. Dussault and P. Marcotte, *Conditions de régularité géométrique pour les inéquations variationnelles*, Recherche opérationnelle, 23 (1989), pp. 1–16.

[287] R. W. Eash, K. S. Chon, Y. J. Lee, and D. E. Boyce, *Equilibrium traffic assignment on an aggregated highway network for sketch planning*, Transportation Research Record, 944 (1983), pp. 30–37.

[288] R. W. Eash, B. N. Janson, and D. E. Boyce, *Equilibrium trip assignment: advantages and implications for practice*, Transportation Research Record, 728 (1979), pp. 1–8.

[289] B. C. Eaves, *On the basic theorem of complementarity*, Mathematical Programming, 1 (1971), pp. 68–75.

[290] ——, *Computing stationary points*, Mathematical Programming Study, 7 (1978), pp. 1–14.

[291] ——, *A locally quadratically convergent algorithm for computing stationary points*, Technical Report SOL-78-13, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[292] B. C. Eaves and R. Saigal, *Homotopies for computation of fixed points on unbounded regions*, Mathematical Programming, 3 (1972), pp. 225–237.

[293] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Mathematics of Operations Research, 18 (1993), pp. 202–226.

[294] J. Eckstein and D. P. Bertsekas, *On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming, 55 (1992), pp. 293–318.

[295] D. El Baz, *A computational experience with distributed asynchronous iterative methods for convex network flow problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 590–591.

[296] R. Elkin, *Convergence theorems for Gauss–Seidel and other minimization algorithms*, PhD thesis, University of Maryland, College Park, MD, 1968.

[297] S. Enke, *Equilibrium among spatially separated markets: solution by electronic analogue*, Econometrica, 10 (1951), pp. 40–47.

[298] S. Erlander, *Accessibility, entropy and the distribution and assignment of traffic*, Transportation Research, 11 (1977), pp. 149–153.

[299] ——, *Optimal Spatial Interaction and the Gravity Model*, vol. 173 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1980.

[300] ——, *On the principle of monotone likelihood and log-linear models*, Mathematical Programming Study, 25 (1985), pp. 108–123.

[301] ——, *On the relationship between the discrete and continuous models for combined distribution and assignment*, Transportation Research, 22B (1988), pp. 371–382.

[302] S. Erlander and N. F. Stewart, *The Gravity Model in Transportation Analysis: Theory and Extensions*, vol. 3 of Topics in Transportation, VSP, Utrecht, The Netherlands, 1990.

[303] L. F. Escudero, *A motivation for using the truncated Newton approach in a very large scale nonlinear network problem*, Mathematical Programming Study, 26 (1986), pp. 240–244.

[304] S. P. Evans, *Some applications of optimisation theory in transport planning*, PhD thesis, Research Group in Traffic Studies, University College London, London, 1973.

[305] S. P. Evans, *Derivation and analysis of some models for combining trip distribution and assignment*, Transportation Research, 10 (1976), pp. 37–57.

[306] ——, *Some models for combining the trip distribution and traffic assignment stages in the transport planning process*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 201–228.

[307] Yu. G. Evtushenko, *Numerical Optimization Techniques*, Optimization Software, New York, NY, 1985.

[308] S.-C. Fang, *Generalized variational inequality, complementarity, and fixed point problems: theory and applications*, PhD thesis, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 1979.

[309] ——, *Fixed point models for equilibrium problems on transportation networks*, Report 80-7, Department of Mathematics, University of Maryland, College Park, MA, 1980.

[310] ——, *Traffic equilibria on multiclass-user transportation networks analysed via variational inequalities*, Tamkang Journal of Mathematics, 13 (1982), pp. 1–9.

[311] S.-C. Fang and E. L. Peterson, *Economic equilibria on networks*, Mathematics Research Report 80-13, Department of Mathematics, University of Maryland, Baltimore County, MD, 1980.

[312] ——, *Generalized variational inequalities*, Journal of Optimization Theory and Applications, 38 (1982), pp. 363–383.

[313] ——, *General network equilibrium analysis*, International Journal of Systems Sciences, 14 (1983), pp. 1249–1257.

[314] ——, *An economic equilibrium model on a multicommodity network*, International Journal of Systems Sciences, 16 (1985), pp. 479–490.

[315] P. G. Farrington, *On the application of current theories of traffic assignment to "real" road systems*, master's thesis, School of Traffic Engineering, University of New South Wales, 1967.

[316] J. M. Farvolden, W. B. Powell, and I. J. Lustig, *A primal partitioning solution for the arc-chain formulation of a multicommodity network flow problem*, Operations Research, 41 (1993), pp. 669–693.

[317] P. Faure and P. Huard, *Résolution de programmes mathématiques à fonction non linéaire par la méthode du gradient réduit*, Revue Française de Recherche Opérationnelle, 9 (1965), pp. 167–205.

[318] Federal Highway Administration, *Traffic assignment: methods, applications, products*, U. S. Department of Transportation, Federal Highway Administration, U. S. Government Printing Office, Washington, D.C., 1972.

[319] ——, *Traffic assignment*, U. S. Department of Transportation, Federal Highway Administration, U. S. Government Printing Office, Washington, D.C., 1973.

[320] ——, *Computer Programs for Urban Transportation Planning: PLANPAC/BACKPAC General Information Manual*, U. S. Department of Transportation, Federal Highway Administration, U. S. Government Printing Office, Washington, D.C., 1977.

[321] B. Feijoo and R. R. Meyer, *Piecewise-linear approximation methods for nonseparable convex optimization*, Management Science, 34 (1988), pp. 411–419.

[322] B. Feinberg, *Coercion functions and decentralized linear programming*, Mathematics of Operations Research, 14 (1989), pp. 177–187.

[323] W. Fenchel, *On conjugate convex functions*, Canadian Journal of Mathematics, 1 (1949), pp. 73–77.

[324] J. A. Ferland, *Minimum cost multicommodity circulation problem with convex arc-costs*, Transportation Science, 8 (1974), pp. 355–360.

[325] J. A. Ferland, M. Florian, and C. Achim, *On incremental methods for traffic assignment*, Transportation Research, 9 (1975), pp. 237–239.

[326] J. E. Fernandez and T. L. Friesz, *Equilibrium predictions in transportation markets: the state of the art*, Transportation Research, 17B (1983), pp. 155–172.

[327] M. Fertal, E. Weiner, A. Balek, and A. Sevin, *Modal split*, tech. report, U. S. Bureau of Public Roads, Washington, D.C., 1966.

[328] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, NY, 1983.

[329] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, NY, 1968.

[330] J.-Ch. Fiorot and P. Huard, *Composition et réunion d'algorithmes généraux d'optimisation*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 280 (1975), pp. 1455–1458.

[331] ——, *Composition and union of general algorithms of optimization*, Mathematical Programming Study, 10 (1979), pp. 69–85.

[332] M. L. Fisher, *The Lagrangian relaxation method for solving integer programming problems*, Management Science, 27 (1981), pp. 1–18.

[333] M. L. Fisher, *An applications oriented guide to Lagrangian relaxation*, Interfaces, 15 (1985), pp. 10–21.
[334] C. Fisk, *Note on the maximum likelihood calibration on Dial's assignment method*, Transportation Research, 11 (1977), pp. 67–68.
[335] ——, *More paradoxes in the equilibrium assignment problem*, Transportation Research, 13B (1979), pp. 305–309.
[336] ——, *Some developments in equilibrium traffic assignment*, Transportation Research, 14B (1980), pp. 243–255.
[337] ——, *A nonlinear equation framework for solving network equilibrium problems*, Environment and Planning, 16A (1984), pp. 67–80.
[338] C. Fisk and S. Nguyen, *A unified approach for the solution of network equilibrium problems*, Publication 169, Centre de recherche sur les transports, Université de Montréal, Montréal, 1980.
[339] ——, *Existence and uniqueness properties of an asymmetric two-mode equilibrium model*, Transportation Science, 15 (1981), pp. 318–328.
[340] ——, *Solution algorithms for network equilibrium models with asymmetric user costs*, Transportation Science, 16 (1982), pp. 361–381.
[341] C. Fisk and S. Pallottino, *Empirical evidence for equilibrium paradoxes with implications for optimal planning strategies*, Transportation Research, 15A (1981), pp. 245–248.
[342] C. S. Fisk, *Game theory and transportation systems modelling*, Transportation Research, 18B (1984), pp. 301–313.
[343] C. S. Fisk and D. E. Boyce, *Alternative variational inequality formulations of the network equilibrium-travel choice problem*, Transportation Science, 17 (1983), pp. 454–463.
[344] C. R. Fleet and S. R. Robertson, *Trip generation in the transportation planning process*, Highway Research Record, 240 (1968), pp. 11–31.
[345] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, Chichester, U. K., second ed., 1987.
[346] M. Florian, *On modelling congestion in Dial's probabilistic assignment model*, Transportation Research, 8 (1974), pp. 85–86.
[347] ——, *An improved linear approximation algorithm for the network equilibrium (packet switching) problem*, in Proceedings of the 10th IEEE Conference on Decision and Control, New Orleans, TX, 1977, pp. 812–818.
[348] ——, *A traffic equilibrium model of travel by car and public transit modes*, Transportation Science, 11 (1977), pp. 166–179.
[349] ——, *Asymmetrical variable demand multi-mode traffic equilibrium problems: existence and uniqueness of solutions and a solution algorithm*, Publication 347, Département d'Informatique et de Recherche Opérationelle, Université de Montréal, Montréal, 1979.
[350] ——, *On the two mode logit mode choice/assignment model*, Publication 177, Centre de recherche sur les transports, Université de Montréal, Montréal, 1980.
[351] ——, *An introduction to network models used in transportation planning*, in Transportation Planning Models, Proceedings of the Course Given at The International Center for Transportation Studies (ICTS), Amalfi, Italy, October 11–16, 1982, M. Florian, ed., North-Holland, Amsterdam, 1984, pp. 137–152.
[352] ——, *Nonlinear cost network models in transportation analysis*, Mathematical Programming Study, 26 (1986), pp. 167–196.
[353] ——. Private communication, 1990.
[354] ——, *Network equilibrium models: from theory to practice.* Paper presented at the 30th ORSA/TIMS Joint National Meeting, Philadelphia, PA, 1990.
[355] M. Florian, R. Chapleau, S. Nguyen, C. Achim, L. James-Lefebvre, S. Galarneau, J. Lefebvre, and C. Fisk, *Validation and application of an equilibrium-based two-mode urban transportation planning method (EMME)*, Transportation Research Record, 728 (1979), pp. 14–23.
[356] M. Florian and B. Fox, *On the probabilistic origin of Dial's multipath traffic assignment model*, Transportation Research, 10 (1976), pp. 339–341.
[357] M. Florian, J. Guélat, and H. Spiess, *An efficient implementation of the "PARTAN" variant of the linear approximation method for the network equilibrium problem*, Networks, 17 (1987), pp. 319–339.
[358] M. Florian and D. Hearn, *Network equilibrium models and algorithms.* 1992.
[359] M. Florian and S. Nguyen, *A method for computing network equilibrium with elastic demands*, Transportation Science, 8 (1974), pp. 321–332.
[360] ——, *An application and validation of equilibrium trip assignment methods*, Transportation Science, 10 (1976), pp. 374–390.
[361] ——, *Recent experience with equilibrium methods for the study of a congested urban area*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 382–395.

[362] M. Florian, S. Nguyen, and J. Ferland, *On the combined distribution-assignment of traffic*, Transportation Science, 9 (1975), pp. 43–53.
[363] M. Florian and H. Spiess, *The convergence of diagonalization algorithms for asymmetric network equilibrium problems*, Transportation Research, 16B (1982), pp. 477–483.
[364] ——, *On binary mode choice/assignment models*, Transportation Science, 17 (1983), pp. 32–47.
[365] ——, *Transport networks in practice*, in Proceedings of the Conference of the Operations Research Society of Italy, Napoli, September 26–28, 1983, 1983, pp. 29–52.
[366] R. W. Floyd, *Algorithm 97: shortest path*, Communications of the Association of Computing Machinery, 5 (1962), p. 345.
[367] L. R. Ford, Jr., *Network flow theory*, Report P-923, Rand Corporation, Santa Monica, CA, 1956.
[368] L. R. Ford, Jr. and D. R. Fulkerson, *A suggested computation for maximal multi-commodity network flows*, Management Science, 5 (1958), pp. 97–101.
[369] ——, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.
[370] L. R. Foulds, *Techniques for predicting flow in a traffic network*, New Zealand Operational Research, 11 (1983), pp. 165–178.
[371] ——, *Traffic network arc elimination by branch and bound enumeration*, The Arabian Journal for Science and Engineering, 10 (1985), pp. 149–157.
[372] ——, *Graph Theory Applications*, Springer-Verlag, New York, NY, 1992.
[373] C. Frank, *A study of alternative approaches to combined trip distribution-assignment modeling*, PhD thesis, Department of Regional Science, University of Pennsylvania, Philadelphia, PA, 1978.
[374] H. Frank and W. Chou, *Routing in computer networks*, Networks, 1 (1971), pp. 99–122.
[375] M. Frank, *The Braess paradox*, Mathematical Programming, 20 (1981), pp. 283–302.
[376] ——, *Cost-deceptive links on ladder networks*, Methods of Operations Research, 45 (1984), pp. 75–86.
[377] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110.
[378] T. J. Fratar, *Vehicular trip distribution by successive approximations*, Traffic Quarterly, 8 (1954), pp. 53–65.
[379] L. Fratta, M. Gerla, and L. Kleinrock, *The flow-deviation method: an approach to store-and-forward computer communication network design*, Networks, 3 (1973), pp. 97–133.
[380] J. W. Friedman, *Oligopoly and the Theory of Games*, North-Holland, Amsterdam, 1977.
[381] T. L. Friesz, *Transportation network equilibrium, design and aggregation: key developments and research opportunities*, Transportation Research, 19A (1985), pp. 413–427.
[382] T. L. Friesz, H.-J. Cho, N. J. Metha, R. L. Tobin, and G. Anandalingam, *A simulated annealing approach to the network design problem with variational inequality constraints*, Transportation Science, 26 (1992), pp. 18–26.
[383] T. L. Friesz, J. Luque, R. L. Tobin, and B.-W. Wie, *Dynamic network traffic assignment considered as a continuous time optimal control problem*, Operations Research, 37 (1989), pp. 893–901.
[384] T. L. Friesz, R. L. Tobin, H.-J. Cho, and N. J. Metha, *Sensitivity analysis based heuristic algorithms for mathematical programs with variational inequality constraints*, Mathematical Programming, 48 (1990), pp. 265–284.
[385] T. L. Friesz, J. Weiss, and J. Gottfried, *Numerical experience with diagonalization algorithms for asymmetric demand traffic assignment*, Civil Engineering Systems, 1 (1983), pp. 63–68.
[386] M. Fukushima, *A descent algorithm for nonsmooth convex optimization*, Mathematical Programming, 30 (1984), pp. 163–175.
[387] ——, *A modified Frank–Wolfe algorithm for solving the traffic assignment problem*, Transportation Research, 18B (1984), pp. 169–177.
[388] ——, *A nonsmooth optimization approach to nonlinear multicommodity network flow problems*, Journal of the Operations Research Society of Japan, 27 (1984), pp. 151–177.
[389] ——, *On the dual approach to the traffic assignment problem*, Transportation Research, 18B (1984), pp. 235–245.
[390] ——, *A relaxed projection method for variational inequalities*, Mathematical Programming, 35 (1986), pp. 58–70.
[391] ——, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Mathematical Programming, 53 (1992), pp. 99–110.
[392] ——, *Splitting algorithms for a class of monotone mappings with application to the traffic equilibrium problem*, Technical Report NAIST-IS-TR93008, Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara, Japan, 1993.
[393] M. Fukushima and T. Itoh, *A dual approach to asymmetric traffic equilibrium problems*, Mathematica Japonica, 32 (1987), pp. 701–721.
[394] J. D. Fuller and B. Lan, *A fast algorithm for bounded generalized processing networks*, Networks, 24 (1994), pp. 57–67.

[395] K. P. Furness, *Time function iteration*, Traffic Engineering & Control, 7 (1965), pp. 458–460.

[396] D. Gabay, *Méthodes numériques pour l'optimisation non linéaire*, PhD thesis, Université Pierre et Marie Curie, Paris, 1979.

[397] ——, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983, pp. 299–331.

[398] D. Gabay and H. Moulin, *On the uniqueness and stability of Nash-equilibria in noncooperative games*, in Applied Stochastic Control in Econometrics and Management Science, A. Bensoussan, P. Kleindorfer, and C. S. Tapiero, eds., North-Holland, Amsterdam, 1980, pp. 271–293.

[399] E. M. Gafni, *Convergence of a routing algorithm*, Report LIDS-TH-907, Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1979.

[400] E. M. Gafni and D. P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM Journal on Control and Optimization, 22 (1984), pp. 936–964.

[401] H. Gajewski and R. Kluge, *Projektionsverfahren bei nichtlinearen Variationsungleichungen*, Mathematische Nachrichten, 46 (1970), pp. 363–373.

[402] R. G. Gallager, *Loops in multicommodity flows*, in Proceedings of the 10th IEEE Conference on Decision and Control, New Orleans, TX, 1977, pp. 819–825.

[403] ——, *A minimum delay routing algorithm using distributed computation*, IEEE Transactions on Communications, COM-25 (1977), pp. 73–85.

[404] G. Gallo, *Updating shortest paths in large scale networks*. Paper presented at the International Workshop on Advances in Linear Optimization Algorithms and Software, Pisa, Italy, 1980.

[405] ——, *Reoptimization procedures in shortest path problems*, Rivista di Matematica e di Scienze Economiche e Sociali, 3 (1980), pp. 3–13.

[406] G. Gallo and S. Pallottino, *A new algorithm to find the shortest paths between all pairs of nodes*, Discrete Applied Mathematics, 4 (1982), pp. 23–35.

[407] ——, *Shortest path methods in transportation models*, in Transportation Planning Models, Proceedings of the Course Given at The International Center for Transportation Studies (ICTS), Amalfi, Italy, October 11–16, 1982, M. Florian, ed., North-Holland, Amsterdam, 1984, pp. 227–256.

[408] ——, *Shortest path methods: a unifying approach*, Mathematical Programming Study, 26 (1986), pp. 38–64.

[409] ——, *Shortest path algorithms*, Annals of Operations Research, 13 (1988), pp. 3–79.

[410] C. B. Garcia and W. I. Zangwill, *Pathways to Solutions, Fixed Pints, and Equilibria*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[411] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, New York, NY, 1979.

[412] N. H. Gartner, *Analysis and control of transportation networks by Frank–Wolfe decomposition*, in Proceedings of the 7th International Symposium on Transportation and Traffic Theory, Kyoto, August 14–17, 1977, T. Sasaki and T. Yamaoka, eds., The Institute of Systems Science Research, Kyoto, Japan, 1977, pp. 591–623.

[413] ——, *Optimal traffic assignment with elastic demands: a review. Part I: analysis framework*, Transportation Science, 14 (1980), pp. 174–191.

[414] ——, *Optimal traffic assignment with elastic demands: a review. Part II: algorithmic approaches*, Transportation Science, 14 (1980), pp. 192–208.

[415] N. H. Gartner, S. B. Gershwin, J. D. C. Little, and P. Ross, *Pilot study of computer-based urban traffic management*, Transportation Research, 14B (1980), pp. 203–217.

[416] N. H. Gartner, B. L. Golden, and R. T. Wong, *Modeling and optimization for transportation systems planning and operations*, in Large Engineering Systems, Proceedings of an International Symposium, University of Manitoba, August 9–12, 1976, A. Wexler, ed., Pergamon Press, 1977, pp. 198–213.

[417] M. Gaudry, *A note on the economic interpretation of delay functions in assignment problems*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 368–381.

[418] A. M. Geoffrion, *Elements of large-scale mathematical programming*, Management Science, 16 (1970), pp. 652–691.

[419] ——, *Generalized Benders decomposition*, Journal of Optimization Theory and Applications, 10 (1972), pp. 237–260.

[420] M. Gerla, *The design of store-and-forward (S/F) networks for computer communications*, PhD thesis, University of California, Los Angeles, CA, 1973.

[421] ——, *Routing and flow control*, in Protocols and Techniques for Data Communication Networks, F. F. Kuo, ed., Prentice-Hall, Englewood Cliffs, NJ, 1981, pp. 122–174.

[422] M. O. Ghali and M. J. Smith, *Traffic assignment, traffic control and road pricing*, in Transportation and Traffic Theory, Proceedings of the 12th International Symposium on the Theory of Traffic Flow and Transportation, Berkeley, CA, July 21–23, 1993, C. F. Daganzo, ed., Elsevier, Amsterdam, 1993, pp. 147–169.
[423] A. Gibert, *A method for the traffic assignment problem*, Report LBS-TNT-95, Transportation Network Theory Unit, London Business School, London, 1968.
[424] ——, *A method for the traffic assignment problem when demand is elastic*, Report LBS-TNT-85, Transportation Network Theory Unit, London Business School, London, 1968.
[425] A. Glazer, *Congestion tolls and consumer welfare*, Public Finance, 36 (1981), pp. 77–83.
[426] F. Glover, D. Klingman, and N. V. Phillips, *Network Models in Optimization and Their Applications in Practice*, John Wiley & Sons, New York, NY, 1992.
[427] R. Glowinski, J. L. Lions, and R. Trémolières, *Analyses Numérique des Inéquations Variationnelles*, Dunod, Paris, 1976.
[428] J. L. Goffin, *Affine methods in nondifferentiable optimization*, CORE Discussion Paper 8744, Center for Operations Research & Econometrics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1987.
[429] ——, *The ellipsoid method and its predecessors*, in Recent Advances in System Modelling and Optimization, Proceedings of the IFIP Working Conference Held in Santiago, Chile, August 27–31, 1984, L. Contesse, R. Correa, and A. Weintraub, eds., vol. 87 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1987, pp. 127–141.
[430] B. L. Golden, *A minimum-cost multicommodity network flow problem concerning imports and exports*, Networks, 5 (1975), pp. 331–356.
[431] A. A. Goldstein, *Cauchy's method of minimization*, Numerische Mathematik, 4 (1962), pp. 146–150.
[432] ——, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society, 70 (1964), pp. 709–710.
[433] T. F. Golob and M. J. Beckmann, *A utility model for travel forecasting*, Transportation Science, 5 (1971), pp. 79–89.
[434] E. G. Gol'shtein, *Method of modification for monotone mappings*, Economics and Mathematical Methods, 11 (1975), pp. 1144–1159.
[435] J. C. Goodman, *A note on existence and uniqueness of equilibrium points for concave n-person games*, Econometrica, 48 (1980), p. 251.
[436] F. J. Gould and J. W. Tolle, *A necessary and sufficient qualification for constrained optimization*, SIAM Journal on Applied Mathematics, 20 (1971), pp. 164–172.
[437] Greater London Council, *Movement in London*, County Hall, London, 1969.
[438] A. Greenbaum, *Synchronization costs on multiprocessors*, Parallel Computing, 10 (1989), pp. 3–14.
[439] J. Guélat, *Algorithmes pour le problème d'affectation du trafic d'équilibre avec demandes fixes: comparaisons*, PhD thesis, Centre de recherche sur les transports, Université de Montréal, Montréal, 1983.
[440] J. Guélat and P. Marcotte, *Some comments on Wolfe's 'away step'*, Mathematical Programming, 35 (1986), pp. 110–119.
[441] E. J. Gumbel, *Statistics of Extremes*, Columbia University Press, New York, NY, 1958.
[442] S. O. Gunnarsson, *An algorithm for multipath traffic assignment*, in Proceedings of the PTRC Urban Traffic Model Research Seminar, London, May 8–12, 1972, 1972.
[443] Y. J. Gur, M. Turnquist, M. Schneider, L. LeBlanc, and D. Kurth, *Estimation of an origin-destination trip table based on observed link volumes and turning movements, volume 1: technical report*, Final Report DOT-FH-11-9292, U. S. Federal Highway Administration, Washington, D.C., 1979.
[444] H. Brézis and M. Sibony, *Méthodes d'approximation et d'itération pour les opérateurs monotones*, Archive for Rational Mechanics and Analysis, 27 (1969), pp. 59–82.
[445] P. C. Haarhoff and J. D. Buys, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, The Computer Journal, 13 (1970), pp. 178–184.
[446] J. K. Hadden, *The use of public transportation in Milwaukee, Wisconsin*, Traffic Quarterly, 18 (1964), p. 230.
[447] W. W. Hager and D. W. Hearn, *Application of the dual active set algorithm to quadratic network optimization*, Computational Optimization and Applications, 1 (1993), pp. 349–373.
[448] A. E. Haghani and M. S. Daskin, *Network design application of an extraction algorithm for network aggregation*, Transportation Research Record, 944 (1983), pp. 37–46.
[449] A. K. Halder, *The method of competing links*, Transportation Science, 4 (1970), pp. 36–51.
[450] M. A. Hall, *Hydraulic network analysis using (generalized) geometric programming*, Networks, 6 (1976), pp. 105–130.
[451] ——, *Properties of the equilibrium state in transportation networks*, Transportation Science, 12 (1978), pp. 208–216.

[452] M. A. Hall and E. L. Peterson, *Traffic equilibria analyzed via geometric programming*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 53–105.

[453] M. D. Hall, D. Van Vliet, and L. G. Willumsen, *SATURN: a simulation-assignment model for the evaluation of traffic management schemes*, Traffic Engineering & Control, 21 (1980), pp. 168–176.

[454] Å. Hallefjord and K. Jørnsten, *Multicommodity network flows with conversions*, Opsearch, 20 (1983), pp. 89–98.

[455] Å. Hallefjord, K. Jørnsten, and S. Storøy, *Traffic equilibrium paradoxes when travel demand is elastic*, Report 79, Department of Informatics, University of Bergen, Bergen, Norway, 1993. To appear in Asia-Pacific Journal of Operational Research.

[456] J. R. Hamburg, *Land use projection for predicting future traffic*, Highway Research Board Bulletin, 224 (1959), pp. 72–84.

[457] J. H. Hammond, *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*, PhD thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[458] J. H. Hammond and T. L. Magnanti, *A contracting ellipsoid method for variational inequality problems*, Working Paper OR 160-87, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[459] S. Hansen, *Entropy and utility in traffic modelling*, in Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory, University of New South Wales, Sydney, Australia, 26–28 August, 1974, D. J. Buckley, ed., Elsevier, New York, NY, 1974, pp. 435–452.

[460] ——, *A survey of traffic models with emphasis on behavioral assumptions*, in Optimising Bus Systems in Urban Areas, Proceedings of a Seminar on Theoretical Models for Fixed Scheduled and Demand Actuated Bus Routing in Urban Areas, Linköping University, May 28–30, 1973, S. Erlander, ed., Linköping University, Linköping, Sweden, 1974, pp. 179–220.

[461] W. G. Hansen, *Land use forecasting for transportation planning*, Highway Research Board Bulletin, 253 (1960), pp. 145–151.

[462] P. T. Harker, *A note on the existence of traffic equilibria*, Applied Mathematics and Computation, 18 (1986), pp. 277–283.

[463] ——, *Accelerating the convergence of the diagonalization and projection algorithms for finite-dimensional variational inequalities*, Mathematical Programming, 41 (1988), pp. 29–59.

[464] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications*, Mathematical Programming, 48 (1990), pp. 161–220.

[465] D. Hartgren and G. Tanner, *Behavioral model of mode choice*, preliminary report, New York State Department of Transportation, Albany, NY, 1970.

[466] P. Hartman and G. Stampacchia, *On some non-linear elliptic differential-functional equations*, Acta Mathematica, 115 (1966), pp. 271–310.

[467] J. M. Harwood and V. Miller, *Urban Traffic Planning*, Printerhall, London, 1964.

[468] A. Haurie and P. Marcotte, *On the relationship between Nash–Cournot and Wardrop equilibria*, Networks, 15 (1985), pp. 295–308.

[469] ——, *A game-theoretic approach to network equilibrium*, Mathematical Programming Study, 26 (1986), pp. 252–255.

[470] M. Hazelton, *Specification, aggregation and bias estimation for transport system performance models: a preliminary report*, Working Paper 772, Transport Studies Unit, University of Oxford, Oxford, 1993.

[471] K. E. Heanue and C. E. Pyers, *A comparative evaluation of trip distribution procedures*, Highway Research Record, 114 (1966), pp. 20–50. Also in Public Roads 34 (1966), pp. 43–51.

[472] D. W. Hearn, *Network aggregation in transportation planning, volume I: summary and survey*, Mathtec Final Report DOT-TSC-RSPD-78-8.I, Mathtec, Inc., Princeton, NJ, 1978.

[473] ——, *Bounding flows in traffic assignment models*, Research Report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 1980.

[474] ——, *The gap function of a convex program*, Operations Research Letters, 1 (1982), pp. 67–71.

[475] ——, *Practical and theoretical aspects of aggregation problems in transportation planning models*, in Transportation Planning Models, Proceedings of the Course Given at The International Center for Transportation Studies (ICTS), Amalfi, Italy, October 11–16, 1982, M. Florian, ed., North-Holland, Amsterdam, 1984, pp. 257–287.

[476] D. W. Hearn and S. Lawphongpanich, *A dual ascent algorithm for traffic assignment problems*, in Dynamic Control and Flow Equilibrium, Proceedings of the Italy-U. S. A. Joint Seminar on Urban Traffic Networks, Naples and Capri, Italy, June 20–23, 1989, pp. 35–53.

[477] D. W. Hearn and S. Lawphongpanich, *A dual ascent algorithm for traffic assignment problems*, Transportation Research, 24B (1990), pp. 423–430.

[478] D. W. Hearn, S. Lawphongpanich, and S. Nguyen, *Convex programming formulations of the asymmetric traffic assignment problem*, Transportation Research, 18B (1984), pp. 357–365.

[479] D. W. Hearn, S. Lawphongpanich, and J. A. Ventura, *Finiteness in restricted simplicial decomposition*, Operations Research Letters, 4 (1985), pp. 125–130.

[480] ——, *Optimization algorithms for congested network models*, in Flow Control of Congested Networks, Proceedings of the NATO Advanced Research Workshop on Flow Control of Congested Networks, Capri, Italy, October 12–18, 1986, A. R. Odoni, L. Bianco, and G. Szegö, eds., vol. F38 of NATO ASI Series, Springer-Verlag, Berlin, 1987, pp. 11–24.

[481] ——, *Restricted simplicial decomposition: computation and extensions*, Mathematical Programming Study, 31 (1987), pp. 99–118.

[482] D. W. Hearn, S. Lawphongpanich, J. A. Ventura, and K. C. Yang, *RSDNET: restricted simplicial decomposition network code*, European Journal of Operational Research, 38 (1989), pp. 121–122.

[483] D. W. Hearn and S. Nguyen, *Dual and saddle functions related to the gap function*, Research Report 82-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 1982.

[484] D. W. Hearn and J. Ribera, *Bounded flow equilibrium problems by penalty methods*, in Proceedings of the 1980 IEEE International Conference on Circuits and Computers, 1980, pp. 162–166.

[485] ——, *Convergence of the Frank–Wolfe method for certain bounded variable traffic assignment problems*, Transportation Research, 15B (1981), pp. 437–442.

[486] R. V. Helgason and J. L. Kennington, *En efficient specialization of the convex simplex method for nonlinear network flow problems*, Technical Report 77017, Department of Industrial Engineering and Operations Research, Southern Methodist University, Dallas, TX, 1978.

[487] R. V. Helgason, J. L. Kennington, and B. D. Stewart, *Dijkstra's two-tree shortest path algorithm*, Technical Report 88-OR-13, Department of Operations Research and Engineering Management, Southern Methodist University, Dallas, TX, 1988.

[488] J. V. Henderson, *Road congestion: a reconsideration of pricing theory*, Journal of Urban Economics, 1 (1974), pp. 346–365.

[489] R. J. Hensen and W. L. Grecco, *Evaluation of the effectiveness of transportation planning in the smaller urban areas*, Traffic Quarterly, 24 (1970), pp. 393–406.

[490] A. M. Hershdorfer, *Predicting the equilibrium of supply and demand: location theory and transportation network flow models*, Transportation Research Forum Papers, (1966), pp. 131–143.

[491] M. R. Hestenes, *Multiplier and gradient methods*, Journal of Optimization Theory and Applications, 4 (1969), pp. 303–320.

[492] ——, *Optimization Theory: The Finite Dimensional Case*, John Wiley & Sons, New York, NY, 1975.

[493] B. G. Heydecker, *Some consequences of detailed junction modeling in road traffic assignment*, Transportation Science, 17 (1983), pp. 263–281.

[494] ——, *On the definition of traffic equilibrium*, Transportation Research, 20B (1986), pp. 435–440.

[495] D. M. Hill and H. G. von Cube, *Development of a model for forecasting travel mode choice in urban areas*, Highway Research Record, 38 (1963), pp. 78–96.

[496] T. J. Hillegass, *Urban transportation planning: a question of emphasis*, Traffic Engineering, 39 (1969), pp. 46–48.

[497] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms, II: Advanced Theory and Bundle Methods*, Springer-Verlag, Berlin, 1993.

[498] H. H. Hoc, *Implementation and use of nonlinear cost multicommodity flow subroutines*, in Computers and Mathematical Programming, vol. 502 of National Bureau of Standards Special Publication, Department of Commerce, Washington, D.C., 1978, pp. 51–58.

[499] R. W. Hockney and C. R. Jesshope, *Parallel Computers 2: Architecture, Programming and Algorithms*, Adam Hilger, Bristol, 1988.

[500] W. W. Hogan, *Convergence results for some extensions of the Frank–Wolfe method*, Working Paper 169, Western Management Science Institute, University of California, Los Angeles, CA, 1971.

[501] ——, *Point-to-set maps in mathematical programming*, SIAM Review, 15 (1973), pp. 591–603.

[502] ——, *Project independence evaluation system: structure and algorithms*, Proceedings of Symposia in Applied Mathematics of the American Mathematical Society, 21 (1977), pp. 121–137.

[503] C. A. Holloway, *A generalized approach to Dantzig–Wolfe decomposition for concave programs*, Operations Research, 21 (1973), pp. 210–220.

[504] ——, *An extension of the Frank and Wolfe method of feasible directions*, Mathematical Programming, 6 (1974), pp. 14–27.

[505] H. S. Houthakker, *The capacity method of quadratic programming*, Econometrica, 28 (1960), pp. 62–87.

[506] R. T. Howe, *A theoretical prediction of work-trip patterns*, Highway Research Board Bulletin, 253 (1960), pp. 155–165.
[507] ——, *A theoretical prediction of work trips in the Minneapolis-St. Paul area*, Highway Research Board Bulletin, 347 (1962), pp. 156–181.
[508] T. C. Hu, *Minimum-cost flows in convex-cost networks*, Naval Research Logistics Quarterly, 13 (1966), pp. 1–9.
[509] C.-I. Hua and F. Porell, *A critical review of the development of the gravity model*, International Regional Science Review, 4 (1979), pp. 97–126.
[510] R. D. Huchingson, R. W. McNees, and C. L. Dudek, *Survey of motorist route-selection criteria*, Highway Research Record, 643 (1977), pp. 45–48.
[511] T. F. Humphrey, *A report on the accuracy of traffic assignment when using capacity restraint*, Highway Research Record, 191 (1967), pp. 53–75.
[512] B. G. Hutchinson, *Principles of Urban Transport Systems Planning*, McGraw-Hill, New York, NY, 1974.
[513] K. Hwang and F. A. Briggs, *Computer Architecture and Parallel Processing*, McGraw-Hill, Singapore, 1985.
[514] S. Ibaraki, M. Fukushima, and T. Ibaraki, *Dual-based Newton methods for nonlinear minimum cost network flow problems*, Journal of the Operations Research Society of Japan, 34 (1991), pp. 263–286.
[515] H. Inouye, *A traffic assignment problem with nonlinear travel-time function*, in Proceedings of the Annual Meeting of the Kansai Branch of the Institute of Civil Engineers of Japan, 1971.
[516] ——, *Traffic equilibria and its solution in congested road networks*, in Proceedings of the IFAC Conference on Control in Transportation Systems, Vienna, 1986, R. Genser, ed., Pergamon Press, Oxford, 1987, pp. 267–272.
[517] C. L. Irwin, *Convergence properties of a PIES-type algorithm for non-integrable functions*, Technical Report SOL 77-33, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1977.
[518] N. A. Irwin, N. Dodd, and H. G. von Cube, *Capacity restraint in assignment programs*, Highway Research Board Bulletin, 297 (1961), pp. 109–127.
[519] N. A. Irwin and H. G. von Cube, *Capacity restraint in multi-travel mode assignment programs*, Highway Research Board Bulletin, 347 (1962), pp. 258–289.
[520] T. Itoh, M. Fukushima, and T. Ibaraki, *An iterative method for variational inequalities with application to traffic equilibrium problems*, Journal of the Operations Research Society of Japan, 31 (1988), pp. 82–103.
[521] G. R. Jansen, *A pilot study in trip assignment*, master's thesis, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA, 1966.
[522] G. R. M. Jansen and P. H. L. Bovy, *The effect of zone size and network detail on all-or-nothing and equilibrium assignment outcomes*, Traffic Engineering & Control, (1982), pp. 311–317, 328.
[523] B. N. Janson, *Most likely origin-destination link uses from equilibrium assignment*, Transportation Research, 27B (1993), pp. 333–350.
[524] B. N. Janson and C. Zozaya-Gorostiza, *The problem of cyclic flows in traffic assignment*, Transportation Research, 21B (1987), pp. 299–310.
[525] S. R. Jara-Díaz, P. P. Donoso, and J. Araneda, *Best partial flow aggregation in transportation cost functions*, Transportation Research, 25B (1991), pp. 329–339.
[526] S. R. Jara-Díaz, P. P. Donoso, and J. A. Araneda, *Estimation of marginal transport costs: the flow aggregation function approach*, Journal of Transport Economics and Policy, 26 (1992), pp. 35–48.
[527] I. Jeevanantham, *A new look at the traffic assignment problem*, in Traffic Flow and Transportation, Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation, G. F. Newell, ed., American Elsevier, New York, NY, 1972, pp. 131–153.
[528] L. P. Jennergren, *Decentralization on the basis of price schedules in linear decomposable resource-allocation problems*, Journal of Financial and Quantitative Analysis, 7 (1972), pp. 1407–1417.
[529] P. Jennergren, *A price schedules decomposition algorithm for linear programming problems*, Econometrica, 41 (1973), pp. 965–980.
[530] W. S. Jewell, *Models for traffic assignment*, Transportation Research, 1 (1967), pp. 31–46.
[531] K. L. Jones, I. J. Lustig, J. M. Farvolden, and W. B. Powell, *Multicommodity network flows: the impact of formulation on decomposition*, Mathematical Programming, 62 (1993), pp. 95–117.
[532] P. C. Jones and E. S. Theise, *On the equivalence of competitive transportation markets and congestion in spatial price equilibrium models*, Transportation Science, 23 (1989), pp. 112–117.
[533] N. O. Jorgensen, *Some aspects of the urban traffic assignment problem*, master's thesis, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA, 1963.

[534] K. O. Jörnsten and P. Värbrand, *An augmented Lagrangian approach to multi-commodity flow problems with conversions*, Opsearch, 23 (1986), pp. 96–106.
[535] N. H. Josephy, *Hogan's PIES example and Lemke's algorithm*, Technical Summary Report 1972, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1979.
[536] ——, *A Newton method for the PIES energy model*, Technical Summary Report 1971, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1979.
[537] ——, *Newton's method for generalized equations*, Technical Summary Report 1965, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1979.
[538] ——, *Quasi-Newton methods for generalized equations*, Technical Summary Report 1966, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1979.
[539] S. Kakutani, *A generalization of Brouwer's fixed point theorem*, Duke Mathematical Journal, 8 (1941), pp. 457–459.
[540] P. V. Kamesam and R. R. Meyer, *Multipoint methods for separable nonlinear networks*, Mathematical Programming Study, 22 (1984), pp. 185–205.
[541] S. Karamardian, *The nonlinear complementarity problem with applications, part 1*, Journal of Optimization Theory and Applications, 4 (1969), pp. 87–98.
[542] ——, *The nonlinear complementarity problem with applications, part 2*, Journal of Optimization Theory and Applications, 4 (1969), pp. 167–181.
[543] ——, *Generalized complementarity problem*, Journal of Optimization Theory and Applications, 8 (1971), pp. 161–168.
[544] ——, *The complementarity problem*, Mathematical Programming, 2 (1972), pp. 107–129.
[545] R. Katsura, M. Fukushima, and T. Ibaraki, *Interior methods for nonlinear minimum cost network flow problems*, Journal of the Operations Research Society of Japan, 32 (1989), pp. 174–199.
[546] R. B. Kellogg, T. Y. Li, and J. Yorke, *A constructive proof of the Brouwer fixed-point theorem and computational results*, SIAM Journal on Numerical Analysis, 13 (1976), pp. 473–483.
[547] J. Kennington and M. Shalaby, *An effective subgradient procedure for minimal cost multicommodity flow problems*, Management Science, 23 (1977), pp. 994–1004.
[548] J. L. Kennington and R. V. Helgason, *Algorithms for Network Programming*, John Wiley & Sons, New York, NY, 1980.
[549] E. N. Khobotov, *Modification of the extra-gradient method for solving variational inequalities and certain optimization problems*, USSR Computational Mathematics and Mathematical Physics, 27 (1987), pp. 120–127.
[550] K. Kim and J. L. Nazareth, *The decomposition principle and algorithms for linear programming*, Linear Algebra and Its Applications, 152 (1991), pp. 119–133.
[551] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, NY, 1980.
[552] G. Kirchoff, Pogendorff Annalen, 72 (1847), pp. 497–508.
[553] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, vol. 1133 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1985.
[554] ——, *A decomposition method for linear programs with dual block angular structure*, Control and Cybernetics, 15 (1986), pp. 395–400.
[555] M. Klein, *A primal method for minimal cost flows with applications to the assignment and transportation problems*, Management Science, 14 (1967), pp. 205–220.
[556] L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, NY, 1964.
[557] R. W. Klessig, *An algorithm for nonlinear multicommodity flow problems*, Networks, 4 (1974), pp. 343–355.
[558] J. G. Klincewicz, *Algorithms for network flow problems with convex separable costs*, PhD thesis, School of Organization and Management, Yale University, New Haven, CT, 1979.
[559] ——, *A Newton method for convex separable network flow problems*, Networks, 13 (1983), pp. 427–442.
[560] ——, *Implementing an "exact" Newton method for separable convex transportation problems*, Networks, 19 (1989), pp. 95–105.
[561] R. Kluge, *Zur approximativen Lösung nichtlinearer Variationsungleichungen*, Deutsche Akademie der Wissenschaften zu Berlin. Monatsberichte, 12 (1970), pp. 120–134.
[562] F. H. Knight, *Some fallacies in the interpretation of social cost*, Quarterly Journal of Economics, 38 (1924), pp. 582–606.
[563] W. Knödel, *Graphentheoretische Methoden und ihre Anwendungen*, Springer-Verlag, Berlin, 1969.
[564] J. E. Kohl, *Der Verkehr und die Ansiedelungen der Menschen in ihrer Abhängigkeit von der Gestaltung der Erdoberfläche*, Dresden, Leipzig, 1841.
[565] M. Kojima, *A unification of the existence theorems of the nonlinear complementarity problem*, Mathematical Programming, 9 (1975), pp. 257–277.

[566] T. C. Koopmans, ed., *Activity Analysis of Production and Allocation*, John Wiley & Sons, New York, NY, 1951.

[567] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 13 (1977), pp. 35–49.

[568] D. T. Kresge and P. O. Roberts, *Systems Analysis and Simulation Models*, vol. 2 of Techniques of Transport Planning, The Brookings Institution, Washington, D.C., 1971.

[569] J. Kruithof, *Telefoonverkeersrekening*, De Ingenieur, 52 (1937), pp. E15–E25. English translation, *Calculation of telephone traffic*, by U. K. Post Office Research Department Library, No. 2663, Dollis Hill, London.

[570] H. W. Kuhn, *Pathfix: an algorithm for computing traffic equilibria*, in Proceedings of the 10th IEEE Conference on Decision and Control, New Orleans, TX, 1977, pp. 831–834.

[571] ——, *Network aggregation in transportation planning, volume II: a fixed point method for treating traffic equilibria*, Mathtec Final Report DOT-TSC-RSPD-78-8.II, Mathtec, Inc., Princeton, NJ, 1978.

[572] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, in Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, CA, 1951, pp. 481–492.

[573] D. Kulash, *A transportation equilibrium model*, Urban Institute Paper 708-45, Urban Institute, Washington, D.C., 1971.

[574] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, New York, NY, 1959.

[575] H. T. Kung, *Synchronized and asynchronous parallel algorithms for multiprocessors*, in Algorithms and Complexity: New Directions and Recent Results, J. F. Traub, ed., Academic Press, New York, NY, 1976, pp. 153–200.

[576] J. Kyparisis, *Sensitivity analysis framework for variational inequalities*, Mathematical Programming, 38 (1987), pp. 203–213.

[577] W. H. K. Lam and H.-J. Huang, *A combined trip distribution and assignment model for multiple user classes*, Transportation Research, 26B (1992), pp. 275–287.

[578] B. Lamond and N. F. Stewart, *Bregman's balancing method*, Transportation Research, 15B (1981), pp. 239–248.

[579] Lancaster Area Transportation Study, *Volume II: Analyses and Forecasts*, Pennsylvania Department of Transportation, Harrisburg, PA, 1970.

[580] M. G. Langdon, *Multiple choice models in transport assessment*, TRRL Laboratory Report 1048, Department of Transport, Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1982.

[581] T. Larsson and Z. Liu, *A Lagrangean relaxation scheme for structured linear programs with application to multicommodity networks flows*, Report LiTH-MAT-R-1989-24, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1989.

[582] T. Larsson, Z. Liu, and M. Patriksson, *A dual scheme for traffic assignment problems*, Report LiTH-MAT-R-1992-21, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1992.

[583] T. Larsson and A. Migdalas, *An algorithm for nonlinear programs over Cartesian product sets*, Optimization, 21 (1990), pp. 535–542.

[584] T. Larsson, A. Migdalas, and M. Patriksson, *A partial linearization method for the traffic assignment problem*, Optimization, 28 (1993), pp. 47–61.

[585] ——, *A generic column generation scheme*, Report LiTH-MAT-R-94-18, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1994.

[586] T. Larsson and M. Patriksson, *Simplicial decomposition with disaggregated representation for the traffic assignment problem*, Transportation Science, 26 (1992), pp. 4–17.

[587] ——, *An augmented Lagrangean scheme for capacitated traffic assignment problems*, in Proceedings of the 2nd Meeting of the EURO Working Group on Urban Traffic and Transportation, Paris, France, September 15–17, 1993, F. Boillot, N. Bhouri, and F. Laurent, eds., vol. 38 of Actes INRETS, Institut National de Recherche sur les Transport et leur Sécurité (INRETS), Arcueil, France, 1993, pp. 163–199. Also as Report LiTH-MAT-R-93-22, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[588] ——, *A class of gap functions for variational inequalities*, Mathematical Programming, 64 (1994), pp. 53–79.

[589] ——, *On the relationship between side constrained and asymmetric models of traffic equilibria*, Report LiTH-MAT-R-94-06, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1994.

[590] L. S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, NY, 1970.

[591] L. S. Lasdon and A. D. Waren, *Survey of nonlinear programming applications*, Operations Research, 28 (1980), pp. 1029–1073.

[592] F. Laurent, *Cost versus time equilibrium over a network*, European Journal of Operational Research, 71 (1993), pp. 205–221.

[593] S. Lawphongpanich, *Decomposition techniques for the traffic assignment problem*, PhD thesis, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 1983.

[594] S. Lawphongpanich and D. W. Hearn, *Simplicial decomposition of the asymmetric traffic assignment problem*, Transportation Research, 18B (1984), pp. 123–133.

[595] ——, *Restricted simplicial decomposition with application to the traffic assignment problem*, Ricerca Operativa, 38 (1986), pp. 97–120.

[596] ——, *Benders decomposition for variational inequalities*, Mathematical Programming, 48 (1990), pp. 231–247.

[597] M. C. Lawson and J. A. Dearinger, *A comparison of four work trip distribution models*, Proceedings of the American Society of Civil Engineers, 93 (1967), pp. 1–25.

[598] R. Layard, *The distributional effects of congestion taxes*, Economica, 44 (1977), pp. 297–304.

[599] L. J. LeBlanc, *Mathematical programming algorithms for large scale network equilibrium and network design problems*, PhD thesis, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 1973.

[600] ——, *An algorithm for the discrete network design problem*, Transportation Science, 9 (1975), pp. 183–199.

[601] ——, *The use of large scale mathematical programming models in transportation systems*, Transportation Research, 10 (1976), pp. 419–421.

[602] L. J. LeBlanc and M. Abdulaal, *A comparison of user-optimum versus system-optimum traffic assignment in transportation network design*, Transportation Research, 18B (1984), pp. 115–121.

[603] L. J. LeBlanc and K. Farhangian, *Efficient algorithms for solving elastic demand traffic assignment problems and mode-split assignment problems*, Transportation Science, 15 (1981), pp. 306–317.

[604] ——, *Selection of a trip table which reproduces observed link flows*, Transportation Research, 16B (1982), pp. 83–88.

[605] L. J. LeBlanc, R. V. Helgason, and D. E. Boyce, *Improved efficiency of the Frank–Wolfe algorithm for convex network programs*, Transportation Science, 19 (1985), pp. 445–462.

[606] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, *An accurate and efficient approach to equilibrium traffic assignment on congested networks*, Transportation Research Record, 491 (1974), pp. 12–23.

[607] ——, *An efficient approach to solving the road network equilibrium traffic assignment problem*, Transportation Research, 9 (1975), pp. 309–318.

[608] S. Lee, *A simplicial decomposition algorithm to solve the traffic assignment problem.* Paper presented at the Universities Transport Study Group Conference, Newcastle University, Newcastle, January 1992.

[609] C. E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Management Science, 11 (1965), pp. 681–689.

[610] A. Lent, *A convergent algorithm for maximum entropy image restoration with a medical X-ray application*, in Image Analysis and Evaluation, Proceedings of the SPSE Conference, Toronto, Canada, 1976, R. Shaw, ed., Society of Photographic Scientists and Engineers, Washington, D.C., 1977, pp. 249–257.

[611] T. Leventhal, G. Nemhauser, and L. Trotter, Jr., *A column generation algorithm for optimal traffic assignment*, Transportation Science, 7 (1973), pp. 168–176.

[612] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1–50.

[613] E. Lill, *Das Reisegesetz und Seine Anwendung auf den Eisenbahnverkehr*, Vienna, 1891.

[614] Y. Y. Lin and J.-S. Pang, *Iterative methods for large convex quadratic programs: a survey*, SIAM Journal on Control and Optimization, 25 (1987), pp. 383–411.

[615] J. L. Lions and G. Stampacchia, *Variational inequalities*, Communications on Pure and Applied Mathematics, 20 (1967), pp. 493–519.

[616] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[617] Z. Liu, *A Lagrangean dual scheme for structured linear programs with applications and extensions*, PhD thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1992.

[618] G. D. Long and V. G. Stover, *The effect of network detail on traffic assignment results*, Report 60-11, Texas Transportation Institute, Texas A & M University, College Station, TX, 1967.

[619] M. Los, *A discrete-convex programming approach to the simultaneous optimization of land use and transportation*, Transportation Research, 13B (1979), pp. 33–48.

[620] M. Los and S. Nguyen, *Spatial allocation on a network with congestion*, Transportation Research, 15B (1981), pp. 113–126.

[621] P. S. Loubal, *A network evaluation procedure*, Highway Research Record, 205 (1967), pp. 96–109.

[622] R. Luce, *Individual Choice: a Theoretical Analysis*, John Wiley & Sons, New York, NY, 1959.

[623] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, second ed., 1984.

[624] J. T. Lundgren, *Optimization approaches to travel demand modelling*, PhD thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1989.

[625] S. E. Lunn, *Route choice by drivers*, TRRL Report SR374, Department of Transport, Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1978.

[626] Z.-Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, Journal of Optimization Theory and Applications, 72 (1992), pp. 7–35.

[627] ——, *On the convergence rate of dual ascent methods for linearly constrained convex minimization*, Mathematics of Operations Research, 18 (1993), pp. 846–867.

[628] M. Lupi, *Convergence of the Frank–Wolfe algorithm in transportation networks*, Civil Engineering Systems, 3 (1986), pp. 7–15.

[629] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM Journal on Control and Optimization, 22 (1984), pp. 277–293.

[630] T. L. Magnanti, *Models and algorithms for predicting urban traffic equilibria*, in Transportation Planning Models, Proceedings of the Course Given at The International Center for Transportation Studies (ICTS), Amalfi, Italy, October 11–16, 1982, M. Florian, ed., North-Holland, Amsterdam, 1984, pp. 153–185.

[631] T. L. Magnanti and B. L. Golden, *Transportation planning: network models and their implementation*, in Studies in Operations Management, A. C. Hax, ed., North-Holland, Amsterdam, 1978, pp. 465–518.

[632] T. L. Magnanti and R. T. Wong, *Network design and transportation planning: models and algorithms*, Transportation Research, 18 (1984), pp. 1–55.

[633] H. S. Mahmassani and K. C. Mouskos, *Vectorization of transportation network equilibrium assignment codes*. Paper presented at the ORSA Conference on the Impact of Recent Computer Advances on Operations Research, Williamsburg, VI, January, 1989.

[634] ——, *Some numerical results on the diagonalization algorithm for network assignment with asymmetric interactions between cars and trucks*, Transportation Research, 22B (1988), pp. 275–290.

[635] C. Mandl, *Applied Network Optimization*, Academic Press, London, 1979.

[636] O. L. Mangasarian, *A simple characterization of solution sets of convex programs*, Operations Research Letters, 7 (1988), pp. 21–26.

[637] M. L. Manheim, *Principles of transport systems analysis*, Highway Research Record, 180 (1966), pp. 11–20.

[638] ——, *Fundamentals of Transportation Systems Analysis, Volume 1: Basic Concepts*, vol. 4 of MIT Press Series in Transportation Studies, MIT Press, Cambridge, MA, 1979.

[639] M. L. Manheim and E. R. Ruiter, *DODOTRANS I: a decision-oriented computer language for analysis of multimode transportation systems*, Highway Research Record, 314 (1970), pp. 135–163.

[640] A. S. Manne, R. G. Richels, and J. P. Weyant, *Energy policy modeling: a survey*, Operations Research, 27 (1979), pp. 1–36.

[641] P. Marcotte, *A new algorithm for solving variational inequalities with application to the traffic assignment problem*, Mathematical Programming, 33 (1985), pp. 339–351.

[642] ——, *Gap-decreasing algorithms for monotone variational inequalities*. Paper presented at the 22nd ORSA/TIMS Joint National Meeting, Miami Beach, FL, 1986.

[643] ——, *Application of Khobotov's algorithm to variational inequalities and network equilibrium problems*, INFOR, 29 (1991), pp. 258–270.

[644] P. Marcotte and J.-P. Dussault, *A modified Newton method for solving variational inequalities*, in Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 1433–1436.

[645] ——, *A note on a globally convergent Newton method for solving monotone variational inequalities*, Operations Research Letters, 6 (1987), pp. 35–42.

[646] ——, *A sequential linear programming algorithm for solving monotone variational inequalities*, SIAM Journal on Control and Optimization, 27 (1989), pp. 1260–1278.

[647] P. Marcotte and J. Guélat, *Adaptation of a modified Newton method for solving the asymmetric traffic equilibrium problem*, Transportation Science, 22 (1988), pp. 112–124.

[648] J. Marklund, *A study of Lagrangian heuristics for convex network flow problems*, master's thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, Linköping, Sweden, 1993.

[649] B. V. Martin and M. L. Manheim, *A research program for comparison of traffic assignment techniques*, Highway Research Record, 88 (1965), pp. 69–84.

[650] B. V. Martin, F. W. Memmot, III, and A. J. Bone, *Principles and Techniques of Predicting Future Demand for Urban Area Transportation*, MIT Press, Cambridge, MA, 1961.

[651] B. Martinet, *Regularisation d'inéquations variationnelles par approximations successives*, Revue Française d'Informatique et de Recherche Opérationnelle, R-3 (1970), pp. 154–158.
[652] ——, *Détermination approchée d'un point fixe d'une application pseudo-contractante*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 274 (1972), pp. 163–165.
[653] B. Martos, *Nonlinear Programming Theory and Methods*, North-Holland, Amsterdam, 1975.
[654] L. G. Mason, *Equilibrium flows, routing patterns and algorithms for store-and-forward networks*, Large Scale Systems, 8 (1985), pp. 187–209.
[655] E. C. Matsoukis, *Road traffic assignment—a review. Part I: non-equilibrium methods*, Transportation Planning and Technology, 11 (1986), pp. 69–79.
[656] E. C. Matsoukis and P. C. Michalopoulos, *Road traffic assignment—a review. Part II: equilibrium methods*, Transportation Planning and Technology, 11 (1986), pp. 117–135.
[657] A. Maugeri, *Applications des inéquations variationnelles au problème de l'éuilibre du trafic*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série I, 295 (1982), pp. 649–652.
[658] ——, *Convex programming, variational inequalities, and applications to the traffic equilibrium problem*, Applied Mathematics and Optimization, 16 (1987), pp. 169–185.
[659] J. P. Mayberry, *Structural requirements for abstract-mode models of passenger transportation*, in The Demand for Travel: Theory and Measurement, R. Quandt, ed., Heath, Lexington, MA, 1970.
[660] D. McFadden, *Conditional logit analysis of qualitative choice behaviour*, in Frontiers in Econometrics, P. Zarempka, ed., Academic Press, New York, NY, 1974, pp. 105–142.
[661] D. McFadden, A. P. Talvitie, et al., *Demand model estimation and validation*, Special Report UCB-ITS-SR-77-9, Institute of Transportation Studies, University of California, Berkeley, CA, 1977.
[662] P. T. McIntosh and B. V. Martin, *Use of the computer in transportation planning*, The Journal of the Institution of Highway Engineers, (1968), pp. 25–31.
[663] W. A. McLaughlin, *Multi-path system traffic assignment algorithm*, Report RB108, Ontario Joint Highway Research Programme, Department of Highways, Ontario, 1966.
[664] V. V. Menon, *The minimal cost flow problem with convex costs*, Naval Research Logistics Quarterly, 12 (1965), pp. 163–172.
[665] D. K. Merchant and G. L. Nemhauser, *A model and an algorithm for the dynamic traffic assignment problems*, Transportation Science, 12 (1978), pp. 183–199.
[666] ——, *Optimality conditions for a dynamic traffic assignment model*, Transportation Science, 12 (1978), pp. 200–207.
[667] W. L. Mertz, *Review and evaluation of electronic computer traffic assignment programs*, Highway Research Board Bulletin, 297 (1961), pp. 94–105.
[668] W. L. Mertz and L. B. Hamner, *A study of factors related to urban travel*, Public Roads, 29 (1957), pp. 170–174.
[669] G. G. L. Meyer, *Accelerated Frank–Wolfe algorithms*, SIAM Journal on Control, 12 (1974), pp. 655–663.
[670] J. R. Meyer, J. F. Kain, and M. Wohl, *The Urban Transportation Problem*, Harvard University Press, Cambridge, MA, 1965.
[671] J. R. Meyer and M. R. Straszheim, *Pricing and Project Evaluation*, vol. 1 of Techniques of Transport Planning, The Brookings Institution, Washington, D.C., 1971.
[672] R. M. Michaels, *Attitudes of drivers determine choice between alternate highways*, Public Roads, 33 (1965), pp. 225–236.
[673] ——, *Attitudes of drivers toward alternative highways and their relation to route choice*, Highway Research Record, 122 (1966), pp. 50–74.
[674] J.-C. Miellou, *Méthodes de Jacobi, Gauss–Seidel, sur-(sous) relaxation par blocs: appliquées à une classe de problèmes non linéaires*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A, 273 (1971), pp. 1257–1260.
[675] A. Migdalas, *Mathematical programming techniques for analysis and design of communication and transportation networks*, PhD thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1988.
[676] ——, *Cyclic linearization vs. Frank–Wolfe decomposition for nonlinear problems over Cartesian product sets*, unpublished note, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1990.
[677] ——, *A regularization of the Frank–Wolfe method*, Report LiTH-MAT-R-1990-10, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1990. To appear in Mathematical Programming.
[678] S. D. Miller, H. J. Payne, and W. A. Thompson, *An algorithm for traffic assignment on capacity constrained transportation networks with queues.* Paper presented at the Johns Hopkins Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, MD, April 2–4, 1975.

[679] S. Mimis, *Equilibrium traffic assignment models for urban networks*, master's thesis, Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[680] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Mathematical Journal, 29 (1962), pp. 341–346.

[681] R. B. Mitchell and C. Rapkin, *Urban Traffic: a Function of Land Use*, Columbia University Press, 1954.

[682] H. Mohring, *Relation between optimum congestion tolls and present highway user charges*, Highway Research Record, 47 (1964), pp. 1–14.

[683] ——, *Transportation Economics*, Ballinger Publishing Company, Cambridge, MA, 1976.

[684] L. Montero, *A simplicial decomposition approach for solving the variational inequality formulation of the general traffic assignment problem for large scale networks*, PhD thesis, Departament d'Estadística i Investigació Operativa, Facultat d'Informàtica, Universitat Politècnica de Catalunya, Barcelona, Spain, 1991.

[685] E. F. Moore, *The shortest path through a maze*, in Proceedings of the International Symposium on the Theory of Switching, Part II, Harvard University, Cambridge, MA, 1957. Also in The Annals of the Computation Laboratory of Harvard University, 30 (1957), pp. 285–292.

[686] J. J. Moré, *Coercivity conditions in nonlinear complementarity problems*, SIAM Review, 16 (1974), pp. 1–16.

[687] J.-J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299.

[688] U. Mosco, *Dual variational inequalities*, Journal of Mathematical Analysis and Applications, 40 (1972), pp. 202–206.

[689] ——, *Implicit Variational Problems and Quasi-Variational Inequalities*, vol. 543 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1976.

[690] W. W. Mosher, Jr., *A capacity-restraint algorithm for assigning flow to a transport network*, Highway Research Record, 6 (1963), pp. 41–70.

[691] K. Moskowitz, *California method of assigning diverted traffic to proposed freeways*, Highway Research Board Bulletin, 130 (1956), pp. 1–26.

[692] R. A. Moyer, *Comprehensive urban transportation study methods*, Journal of the Highway Division, 91 (1965), p. 62.

[693] D. Munby, ed., *Transport: Selected Readings*, Penguin Books, Harmondsworth, Middlesex, England, 1968.

[694] J. D. Murchland, *Some remarks on the gravity model of traffic distribution, and an equivalent maximization formulation*, Report LSE-TNT-38, Transport Network Theory Unit, London School of Economics, London, 1966.

[695] ——, *Braess's paradox of traffic flow*, Transportation Research, 4 (1970), pp. 391–394.

[696] ——, *A fixed method for all shortest distances in a directed graph and for the inverse problem*, PhD thesis, University of Karlsruhe, Karlsruhe, Germany, 1970.

[697] ——, *Road network traffic distribution in equilibrium*, in Mathematical Models in the Social Sciences, II Oberwolfach-Tagung über Operations Research, Mathematisches Forschungsinstitut, Oberwolfach, 20–25 October 1969, R. Henn, H. P. Künzi, and H. Schubert, eds., vol. 8, Anton Hain Verlag, Meisenheim am Glan, 1970, pp. 145–183. In German, translation by H. A. Paul.

[698] B. A. Murtagh and M. A. Saunders, *Large-scale linearly constrained optimization*, Mathematical Programming, 14 (1978), pp. 41–72.

[699] A. Nagurney, *Comparative tests of multimodal traffic equilibrium methods*, Transportation Research, 18B (1984), pp. 469–485.

[700] ——, *Computational comparisons of algorithms for general asymmetric traffic equilibrium problems with fixed and elastic demands*, Transportation Research, 20B (1986), pp. 78–84.

[701] ——, *An equilibration scheme for the traffic assignment problem with elastic demands*, Transportation Research, 22B (1988), pp. 73–79.

[702] ——, *Network Economics: A Variational Inequality Approach*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[703] I. Nakahori, K. Nakazaki, and Y. Nishikawa, *A graph-theoretic consideration and algorithms for a multicommodity flow assignment problem*, Electronics and Communications in Japan, 60-A (1977).

[704] J. Nash, *Equilibrium points in n-person games*, Proceedings of the National Academy of Sciences of the United States of America, 36 (1950), pp. 48–49.

[705] ——, *Non-cooperative games*, Annals of Mathematics, 54 (1951), pp. 286–295.

[706] W. W. Nash and J. R. Voss, *Analyzing the socio-economic impacts of urban highways*, Highway Research Board Bulletin, 268 (1960), pp. 80–94.

[707] J. R. Nelson, ed., *Marginal Cost Pricing in Practice*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[708] E. R. Nestle, *Cost versus accuracy tradeoffs for the transition between conventional and equilibrium traffic assignment methods*, master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1974.

[709] M. Netter, *Affectations de trafic et tarification au coût marginal social: critique de quelques idées admises*, Transportation Research, 6 (1972), pp. 411–429.

[710] ——, *Equilibrium and marginal cost pricing on a road network with several traffic flow types*, in Traffic Flow and Transportation, Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation, G. F. Newell, ed., American Elsevier, New York, NY, 1972, pp. 155–163.

[711] G. F. Newell, *Traffic Flow on Transportation Networks*, vol. 5 of MIT Press Series in Transportation Studies, MIT Press, Cambridge, MA, 1980.

[712] I. S. Newton, *Philosophiae Naturalis Principia Mathematica*, London, 1687.

[713] S. Nguyen, *A mathematical programming approach to equilibrium methods of traffic assignment with fixed demands*, Publication 138, Département d'Informatique et de Recherche Opérationelle, Université de Montréal, Montréal, 1973.

[714] ——, *An algorithm for the traffic assignment problem*, Transportation Science, 8 (1974), pp. 203–216.

[715] ——, *Une approche unifiée des méthodes d'équilibre pour l'affectation du trafic*, Publication 171, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, 1974.

[716] ——, *Procedures for equilibrium traffic assignment with elastic demand*, Publication 39, Centre de recherche sur les transports, Université de Montréal, Montréal, 1976.

[717] ——, *A unified approach to equilibrium methods for traffic assignment*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 148–182.

[718] S. Nguyen and C. Dupuis, *Une méthode efficace de calcul d'un trafic d'équilibre dans le cas des coûts non-symétriques*, Publication 205, Centre de recherche sur les transports, Université de Montréal, Montréal, 1981.

[719] ——, *An efficient method for computing traffic equilibria in networks with asymmetric transportation costs*, Transportation Science, 18 (1984), pp. 185–202.

[720] S. Nguyen and L. James, *TRAFFIC: an equilibrium traffic assignment program*, Publication 17, Centre de recherche sur les transports, Université de Montréal, Montréal, 1975.

[721] T. A. J. Nicholson, *Finding the shortest route between two points in a network*, Computer Journal, 9 (1966), pp. 275–280.

[722] H. Nikaido, *Convex Structures and Economic Theory*, Academic Press, New York, NY, 1968.

[723] Y. Nishikawa and I. Nakahori, *New algorithms for an equilibrium solution of the traffic assignment problem*, Electronics and Communications in Japan, 59-A (1976), pp. 12–21.

[724] ——, *A network theoretic formulation and algorithms for the traffic assignment problem*, in Proceedings of the 7th International Symposium on Transportation and Traffic Theory, Kyoto, August 14–17, 1977, T.Sasaki and T. Yamaoka, eds., The Institute of Systems Science Research, Kyoto, Japan, 1977, pp. 531–544.

[725] J. Oh, *Estimation of trip matrices from traffic counts: an equilibrium approach*, in Mathematics in Transport Planning and Control, Based on the Proceedings of a Conference on Mathematics in Transport Planning and Control Organized by The Institute of Mathematics and its Applications and Held at the University of Wales College of Cardiff in September 1989, J. D. Griffiths, ed., Clarendon Press, Oxford, 1992, pp. 35–44.

[726] W. Y. Oi and P. W. Shuldiner, *An Analysis of Urban Travel Demands*, Northwestern University Press, Evanston, IL, 1962.

[727] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[728] J. Ortúzar and L. G. Willumsen, *Modelling Transport*, John Wiley & Sons, Chichester, U. K., 1990.

[729] S. Osofsky, *A multiple regression approach to forecasting urban area traffic volumes*, in Proceedings of the American Association of State Officials, Washington, D.C., 1958.

[730] V. E. Outram and E. Thompson, *Driver route choices — behavioural and motivational studies*, in Proceedings of the PTRC Seminar on Transportation Models, 1977, pp. 114–121.

[731] ——, *Drivers perceived cost in route choice*, in Proceedings of the PTRC Seminar on Transportation Models, University of Warwich, Warwich, England, July 10–13, 1978, pp. 226–257.

[732] K. R. Overgaard, *Traffic estimation in urban transportation planning*, vol. 37 of Acta Polytechnica Scandinavica, Civil Engineering and Building Construction Series, 1966.

[733] ——, *Testing a traffic assignment algorithm*, in Vehicular Traffic Science, Proceedings of the 3rd International Symposium on the Theory of Traffic Flow, New York, June 1965, L. C. Edie, R. Herman, and R. Rothery, eds., American Elsevier, New York, NY, 1967, pp. 215–221.

[734] J.-S. Pang, *An equivalence between two algorithms for quadratic programming*, Mathematical Programming, 20 (1981), pp. 152–165.

[735] J.-S. Pang, *On the convergence of dual ascent methods for large-scale linearly constrained optimization problems*, technical report, School of Management, University of Texas at Dallas, Richardson, TX, 1984.

[736] ——, *Asymmetric variational inequality problems over product sets: applications and iterative methods*, Mathematical Programming, 31 (1985), pp. 206–219.

[737] J.-S. Pang and D. Chan, *Gauss–Seidel methods for variational inequality problems over Cartesian product sets*, report, School of Management, University of Texas at Dallas, Dallas, TX, 1982.

[738] ——, *Iterative methods for variational and complementarity problems*, Mathematical Programming, 24 (1982), pp. 284–313.

[739] J.-S. Pang and C.-S. Yu, *Linearized simplicial decomposition methods for computing traffic equilibria on networks*, Networks, 14 (1984), pp. 427–438.

[740] M. Papageorgiou, ed., *Concise Encyclopedia of Traffic & Transportation Systems*, Pergamon Press, 1991.

[741] U. Pape, *Implementation and efficiency of Moore-algorithms for the shortest route problem*, Mathematical Programming, 7 (1974), pp. 212–222.

[742] C. Pasche, *Optimisation convexe dans les réseaux*, Report O.R.W.P 86/5, Département de Mathématiques, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 1986.

[743] M. Patriksson, *Algorithms for urban traffic network equilibria*, licentiate thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1991.

[744] ——, *A descent algorithm for a class of generalized variational inequalities*, Report LiTH-MAT-R-93-35, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[745] ——, *Partial linearization methods in nonlinear programming*, Journal of Optimization Theory and Applications, 78 (1993), pp. 227–246.

[746] ——, *A unified description of iterative algorithms for traffic equilibria*, European Journal of Operational Research, 71 (1993), pp. 154–176.

[747] ——, *A unified framework of descent algorithms for nonlinear programs and variational inequalities*, PhD thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[748] ——, *On the convergence of descent methods for monotone variational inequalities*, Report LiTH-MAT-R-94-01, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1994.

[749] ——, *A taxonomy of classes of descent algorithms for nonlinear programs and variational inequalities*, Report LiTH-MAT-R-94-06, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1994.

[750] H. J. Payne and W. A. Thompson, *Traffic assignment on transportation networks with capacity constraints and queueing.* Paper presented at the 47th National ORSA Meeting/TIMS 1975 North-American Meeting, Chicago, IL, April 30–May 2, 1975.

[751] E. R. Petersen, *A primal-dual traffic assignment algorithm*, Management Science, 22 (1975), pp. 87–95.

[752] D. W. Peterson, *A review of constraint qualifications in finite-dimensional spaces*, SIAM Review, 15 (1973), pp. 639–654.

[753] A. T. Philbrick, *A short history of the development of the gravity model*, Australian Road Research, 5 (1973), pp. 40–54.

[754] G. W. Pick and J. Gill, *New developments in category analysis*, in PRTC Symposium, London, 1970.

[755] G. Pierra, *Decomposition through formalization in a product space*, Mathematical Programming, 28 (1984), pp. 96–115.

[756] A. C. Pigou, *The Economics of Welfare*, Macmillan & Co, London, 1920.

[757] M. Ç. Pinar and S. A. Zenios, *Solving nonlinear programs with embedded network structures*, in Network Optimization Problems: Algorithms, Applications and Complexity, D.-Z. Du and P. M. Pardalos, eds., World Scientific, Singapore, 1993, pp. 177–202.

[758] J. Polak, *Some methodological aspects of equilibrium assignment algorithms.* Paper presented at the Annual Conference of the Universities' Transport Study Group, 1983.

[759] B. T. Polyak, *Minimization of unsmooth functionals*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.

[760] ——, *Introduction to Optimization*, Optimization Software, New York, NY, 1987.

[761] B. T. Polyak and N. V. Tret'yakov, *An iterative method for linear programming and its economic interpretation*, Matekon, 10 (1974), pp. 81–100.

[762] R. B. Potts and R. M. Oliver, *Flows in Transportation Networks*, vol. 90 of Mathematics in Science and Engineering, Academic Press, New York, NY, 1972.

[763] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, Symposium of the Institute of Mathematics and Its Applications, University of Keele, England, 1968, R. Fletcher, ed., Academic Press, New York, NY, 1969, pp. 283–298.

[764] W. B. Powell and Y. Sheffi, *The convergence of equilibrium algorithms with predetermined step sizes*, Transportation Science, 16 (1982), pp. 45–55.
[765] W. Prager, *Problems of traffic and transportation*, in Proceedings of the Symposium on Operations Research in Business and Industry, Midwest Research Institute, Kansas City, KS, 1954, pp. 105–113.
[766] M. E. Primak, *A computational process of search for equilibrium points*, Cybernetics, 9 (1975), pp. 106–113.
[767] J. A. Proudlove, *Some comments on West Midlands Transport Study*, Traffic Engineering & Control, 10 (1968), pp. 351–353, 360.
[768] B. N. Pshenichny and Yu. M. Danilin, *Numerical Methods in Extremal Problems*, MIR Publishers, Moscow, 1978.
[769] Y. Qiu and T. L. Magnanti, *Sensitivity analysis for variational inequalities defined on polyhedral sets*, Mathematics of Operations Research, 14 (1989), pp. 410–432.
[770] ——, *Sensitivity analysis for variational inequalities*, Mathematics of Operations Research, 17 (1992), pp. 61–76.
[771] D. A. Quarmby, *Choice of travel mode for the journey to work: some findings*, Journal of Transport Economics and Policy, 1 (1967), pp. 273–314.
[772] J. Randle, *A convergent probabilistic road assignment model*, Traffic Engineering & Control, 20 (1979), pp. 519–521.
[773] E. P. Ratcliffe, *A comparison of drivers' route choice criteria and those used in current assignment processes*, Traffic Engineering & Control, 13 (1972), pp. 526–529.
[774] E. J. Roberts, *Transportation networks*, in Proceedings of the 10th IEEE Conference on Decision and Control, New Orleans, TX, 1977, pp. 639–644.
[775] P. Robillard and N. F. Stewart, *Iterative numerical methods for trip distribution problems*, Transportation Research, 8 (1974), pp. 575–582.
[776] S. M. Robinson, *An implicit-function theorem for generalized variational inequalities*, Technical Report 1672, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1976.
[777] ——, *Strongly regular generalized equations*, Mathematics of Operations Research, 1 (1980), pp. 43–62.
[778] R. T. Rockafellar, *Convex functions, monotone operators and variational inequalities*, in Theory and Applications of Monotone Operators, Proceedings of the NATO Advanced Study Institute, Venice, Italy, A. Ghizzetti, ed., Edizioni Oderisi, Gubbio, Italy, 1969, pp. 35–65.
[779] ——, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[780] ——, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Mathematical Programming, 5 (1973), pp. 354–373.
[781] ——, *The multiplier method of Hestenes and Powell applied to convex programming*, Journal of Optimization Theory and Applications, 12 (1973), pp. 555–562.
[782] ——, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of Operations Research, 1 (1976), pp. 97–116.
[783] ——, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.
[784] ——, *Lagrange multipliers and variational inequalities*, in Variational Inequalities and Complementarity Problems: Theory and Applications, R. W. Cottle, F. Giannessi, and J.-L. Lions, eds., John Wiley & Sons, Chichester, U. K., 1980, ch. 20, pp. 303–322.
[785] ——, *Network Flows and Monotropic Optimization*, John Wiley & Sons, New York, NY, 1984.
[786] V. V. Rodionov, *The parametric problem of shortest distances*, U.S.S.R. Computational Mathematics and Mathematical Physics, 8 (1968), pp. 336–343.
[787] K. G. Rogers, G. Townsend, and A. E. Metcalf, *Planning for the work journey: a generalized explanation of modal choice*, Report C.67, Local Government Operational Research Unit, 1971. Revised edition.
[788] G. Rose, M. S. Daskin, and F. S. Koppelman, *An examination of convergence error in equilibrium traffic assignment models*, Transportation Research, 22B (1988), pp. 261–274.
[789] J. B. Rosen, *The gradient projection method for nonlinear programming, part I: linear constraints*, Journal of the Society of Industrial and Applied Mathematics, 8 (1960), pp. 181–217.
[790] ——, *Existence and uniqueness of equilibrium points for concave n-person games*, Econometrica, 33 (1965), pp. 520–534.
[791] R. W. Rosenthal, *The network equilibrium problem in integers*, Networks, 3 (1973), pp. 53–59.
[792] T. F. Rossi, S. McNeil, and C. Hendrickson, *Entropy model for consistent impact-fee assessment*, Journal of Urban Planning and Development/ASCE, 115 (1989), pp. 51–63.
[793] G. J. Roth, *An economic approach to traffic congestion*, in Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow, London, June 25–27, 1963, J. Almond, ed., The Organisation for Economic Co-operation and Development, Paris, 1965, pp. 304–316.
[794] W. Rudin, *Principles of Mathematical Analysis*, Mc-Graw Hill, Auckland, New Zealand, third ed., 1976.

[795] N. Ruggles, *Recent developments in the theory of marginal cost pricing*, Review of Economic Studies, 17 (1949/50), pp. 107–126.

[796] ——, *The welfare basis of the marginal cost pricing principle*, Review of Economic Studies, 17 (1949/50), pp. 29–46.

[797] E. R. Ruiter, *ICES TRANSET I: engineering users manual*, Research Report R68-10, Laboratory of Civil Engineering Systems, Massachusetts Institute of Technology, Cambridge, MA, 1968.

[798] ——, *The prediction of network equilibrium: the state of the art*, in Proceedings of the International Conference on Transportation Research, Bruge, Belgium, College of Europe, Bruge, Belgium, 1973, pp. 717–726.

[799] ——, *Implementation of operational network equilibrium procedures*, Transportation Research Record, 491 (1974), pp. 40–51.

[800] A. Ruszczyński, *An augmented Lagrangian decomposition method for block diagonal linear programming problems*, Operations Research Letters, 8 (1989), pp. 287–294.

[801] D. P. Rutenberg, *Generalized networks, generalized upper bounding and decomposition of the convex simplex method*, Management Science, 16 (1970), pp. 388–401.

[802] R. S. Sacher, *A decomposition algorithm for quadratic programming*, Mathematical Programming, 18 (1980), pp. 16–30.

[803] K. N. A. Safwat and T. L. Magnanti, *A combined trip generation, trip distribution, modal split, and trip assignment model*, Transportation Science, 18 (1988), pp. 14–30.

[804] K. Saishu and Y. Moriwaki, *Optimal assignment of traffic flows*, Electrical Engineering in Japan, 92 (1972), pp. 113–120.

[805] R. J. Salter, *Highway Traffic Analysis and Design*, Macmillan, London, second ed., 1976.

[806] P. A. Samuelson, *Spatial price equilibrium and linear programming*, American Economic Review, 42 (1952), pp. 283–303.

[807] ——, *Intertemporal price equilibrium: a prologe to the theory of speculation*, Weltwirtschaftliches Archiv, 79 (1957), pp. 181–219.

[808] T. Sasaki, Y. Iida, and H. Yang, *User-equilibrium traffic assignment by continuum approximation of network flow*, in Transportation and Traffic Theory, Proceedings of the 11th International Symposium on Transportation and Traffic Theory, Yokohama, July 18–20, 1990, M. Koshi, ed., Elsevier, New York, NY, 1990, pp. 233–252.

[809] T. Sasaki and H. Inouye, *Traffic assignment by analogy to electric circuits*, in Transportation and Traffic Theory, Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney, August 26–28, 1974, D. J. Buckley, ed., Elsevier, New York, NY, 1974, pp. 495–518.

[810] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM Journal on Applied Mathematics, 15 (1967), pp. 1328–1343.

[811] S. Schechter, *Iteration methods for nonlinear problems*, Transactions of the American Mathematical Society, 104 (1962), pp. 179–189.

[812] ——, *Relaxation methods for convex problems*, SIAM Journal on Numerical Analysis, 5 (1968), pp. 601–612.

[813] H. Schittenhelm, *On the integration of an effective assignment algorithm with path and path flow management in a combined trip distribution and traffic assignment algorithm*, in Proceedings of the 18th PTRC Summer Annual Meeting on European Transport and Planning, 1990.

[814] R. E. Schmidt and M. E. Campbell, *Highway Traffic Estimation*, The Eno Foundation for Highway Traffic Control, Saugatuck, CT, 1956.

[815] M. Schneider, *A direct approach to traffic assignment*, Highway Research Record, 6 (1963), pp. 71–75.

[816] ——, *Probability maximization in networks*, in Proceedings of the International Conference on Transportation Research, Bruges, Belgium, 1973, pp. 748–755.

[817] M. H. Schneider and S. A. Zenios, *A comparative study of algorithms for matrix balancing*, Operations Research, 38 (1990), pp. 439–455.

[818] G. L. Schultz and R. R. Meyer, *A three-phase algorithm for block-structured optimization*, in Proceedings of the Fourth SIAM Conference on Parallel Processing for Scientific Computing, J. Dongarra, P. Messina, D. C. Sorensen, and R. G. Voigt, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990, pp. 186–191.

[819] ——, *An interior point method for block angular optimization*, SIAM Journal on Optimization, 1 (1991), pp. 583–602.

[820] A. Schwartz, *Forecasting transit use*, Highway Research Board Bulletin, 297 (1961), pp. 18–35.

[821] M. Schwartz and C. K. Cheung, *The gradient projection algorithm for multiple routing in message-switched networks*, IEEE Transactions on Communication, COM-24 (1976), pp. 449–456.

[822] Y. Sekine, *Decentralized optimization of an interconnected system*, IEEE Transactions on Circuit Theory, CT-10 (1963), pp. 161–168.

[823] J. G. Sender and M. Netter, *Équilibre offre-demande et tarification sur un réseau de transport*, Rapport de recherche 3, Département Economie, Institut de Recherche des Transports, Arcueil, France, 1970.

[824] P. Serra and A. Weintraub, *Convergence of decomposition algorithms for the traffic assignment problem*, in Studies on Graphs and Discrete Programming, P. Hansen, ed., North-Holland, Amsterdam, 1981, pp. 313–318.

[825] B. V. Shah, R. J. Beuhler, and O. Kempthorne, *Some algorithms for minimizing a function of several variables*, SIAM Journal on Applied Mathematics, 12 (1964), pp. 74–92.

[826] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, 27 (1948), pp. 379–423, 623–656.

[827] J. F. Shapiro, *Mathematical Programming: Structures and Algorithms*, John Wiley & Sons, New York, NY, 1979.

[828] C. H. Sharp, *Congestion and welfare: an examination of the case for a congestion tax*, The Economic Journal, 76 (1966), pp. 806–817.

[829] ——, *'Congestion and welfare' reconsidered*, Journal of Transport Economics and Policy, 2 (1968), pp. 119–125.

[830] A. Sheer, N. Simhairi, and L. Bennett, *Trees, cycles and sensitivity of the traffic assignment problem*, in Mathematics in Transport Planning and Control, Based on the Proceedings of a Conference on Mathematics in Transport Planning and Control Organized by The Institute of Mathematics and its Applications and Held at the University of Wales College of Cardiff in September 1989, J. D. Griffiths, ed., Clarendon Press, Oxford, 1992, pp. 145–158.

[831] Y. Sheffi, *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[832] Y. Sheffi and W. B. Powell, *A comparison of stochastic and deterministic traffic assignment over congested networks*, Transportation Research, 15B (1981), pp. 53–64.

[833] ——, *An algorithm for the equilibrium assignment problem with random link times*, Networks, 12 (1982), pp. 191–207.

[834] Y. Sheffi and R. Trexler, *A note on the accuracy of the continuum approximation spatial aggregation algorithm of traffic assignment*, Transportation Science, 14 (1980), pp. 306–323.

[835] C. M. Shetty and M. Ben Daya, *A decomposition procedure for convex quadratic programs*, Naval Research Logistics, 35 (1988), pp. 111–118.

[836] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.

[837] P. W. Shuldiner, *Trip generation and the home*, Highway Research Board Bulletin, 347 (1962), pp. 40–59.

[838] M. Sibony, *Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone*, Calcolo, 7 (1970), pp. 65–183.

[839] R. Smeed, *Road pricing: the economic and technical possibilities*, report of a panel set up by the ministry of transport, Her Majesty's Stationery Office, London, 1964.

[840] M. J. Smith, *The existence, uniqueness and stability of traffic equilibria*, Transportation Research, 13B (1979), pp. 295–304.

[841] ——, *The marginal cost taxation of a transportation network*, Transportation Research, 13B (1979), pp. 237–242.

[842] ——, *The existence of an equilibrium solution to the traffic assignment problem when there are junction interactions*, Transportation Research, 15B (1981), pp. 443–451.

[843] ——, *Properties of a traffic control policy which ensure the existence of a traffic equilibrium consistent with the policy*, Transportation Research, 15B (1981), pp. 453–462.

[844] ——, *The stability of urban traffic control systems*, in Proceedings of the 3rd IMA Conference on Control Theory, Sheffield 1980, J. E. Marshall, ed., Academic Press, London, 1981, pp. 661–681.

[845] ——, *An algorithm for solving asymmetric equilibrium problems with a continuous cost-flow function*, Transportation Research, 17B (1983), pp. 365–371.

[846] ——, *The existence and calculation of traffic equilibria*, Transportation Research, 17B (1983), pp. 291–303.

[847] ——, *A descent algorithm for solving monotone variational inequalities and monotone complementarity problems*, Journal of Optimization Theory and Applications, 44 (1984), pp. 485–496.

[848] ——, *The stability of a dynamic model of traffic assignment: an application of a method of Lyapunov*, Transportation Science, 18 (1984), pp. 245–252.

[849] ——, *Two alternative definitions of traffic equilibrium*, Transportation Research, 18B (1984), pp. 63–65.

[850] ——, *Traffic control and traffic assignment in a signal-controlled network with queueing*, in Proceedings of the 10th International Symposium on Transportation and Traffic Theory, Boston, MA, 1987.

[851] M. J. Smith and M. O. Ghali, *Dynamic traffic assignment and dynamic traffic control*, in Transportation and Traffic Theory, Proceedings of the 11th International Symposium on Transportation and Traffic Theory, Yokohama, July 18–20, 1990, M. Koshi, ed., Elsevier, New York, NY, 1990, pp. 273–290.

[852] ——, *The dynamics of traffic assignment and traffic control: a theoretical study*, Transportation Research, 24B (1990), pp. 409–422.

[853] T. E. Smith, *A cost-efficiency principle of spatial interaction behavior*, Regional Science and Urban Economics, 8 (1978), pp. 313–337.

[854] ——, *A general efficiency principle of spatial interaction*, in Spatial Interaction Theory and Planning Models, A. Karlqvist, L. Lundqvist, F. Snickars, and J. W. Weibull, eds., North-Holland, Amsterdam, 1978, pp. 97–118.

[855] ——, *A cost-efficiency approach to the analysis of congested spatial-interaction behaviour*, Environment and Planning, 15A (1983), pp. 435–464.

[856] ——, *A solution condition for complementarity problems: with an application to spatial price equilibrium*, Applied Mathematics and Computations, 15 (1984), pp. 61–69.

[857] ——, *A cost-efficiency theory of dispersed network equilibria*, Environment and Planning, A20 (1988), pp. 231–266.

[858] R. B. Smock, *A comparative description of a capacity-restrained traffic assignment*, Highway Research Record, 6 (1963), pp. 12–40.

[859] ——, *An iterative assignment approach to capacity restraint on arterial networks*, Highway Research Board Bulletin, 347 (1963), pp. 60–66.

[860] R. R. Snell, M. L. Funk, L. T. Fan, F. A. Tillman, and J. J. Wang, *Traffic assignment with a nonlinear travel-time function*, Transportation Science, 2 (1968), pp. 146–159.

[861] F. Snickars and J. Weibull, *A minimum information principle: theory and practice*, Regional Science and Urban Economics, 7 (1977), pp. 137–168.

[862] W. Solesbury and A. Townsend, *Transportation studies and British planning practice*, Town Planning Review, (1970).

[863] T. J. Soltman, *Effects of alternate loading sequences on results from Chicago trip distribution and assignment model*, Highway Research Record, 114 (1966), pp. 122–140.

[864] H. Soroush and P. B. Mirchandani, *The stochastic multicommodity flow problem*, Networks, 20 (1990), pp. 121–155.

[865] F. Soumis, *Planification d'une flotte d'avions*, Publication 133, Centre de Recherche sur les Transports, Université de Montréal, Montréal, Canada, 1978.

[866] H. Spiess, *Conical volume-delay functions*, Transportation Science, 24 (1990), pp. 153–158.

[867] S. Spurkland, *Mathematical tools for urban studies*, Reprint 14, Norwegian Institute of Urban and Regional Research, 1966.

[868] G. Stampacchia, *Variational inequalities*, in Theory and Applications of Monotone Operators, Proceedings of the NATO Advanced Study Institute, Venice, Italy, A. Ghizzetti, ed., Edizioni Oderisi, Gubbio, Italy, 1969, pp. 101–192.

[869] F. N. Starasinic and J. J. Schuster, *Comparative analysis of capacity restraint traffic assignments*, Traffic Engineering, 42 (1972), pp. 48–52, 58–59.

[870] P. A. Steenbrink, *Optimalisering van de infrastruktur*, Verkeerstechniek, 22 (1971).

[871] ——, *Optimization of Transport Networks*, John Wiley & Sons, London, 1974.

[872] D. Stefek, *Extensions of simplicial decomposition for solving the multicommodity flow problem with bounded arc flows and convex costs*, PhD thesis, University of Pennsylvania, 1989.

[873] R. Steinberg and R. E. Stone, *The prevalence of paradoxes in transportation equilibrium problems*, Transportation Science, 22 (1988), pp. 231–241.

[874] R. Steinberg and W. I. Zangwill, *The prevalence of Braess' paradox*, Transportation Science, 17 (1983), pp. 301–318.

[875] E. Stern and D. Leiser, *Urban route choice among different driving groups*, report, Israeli National Academy of Sciences, Beer Sheva, Israel, 1986.

[876] N. F. Stewart, *Notes on the mathematical structure of equilibrium models*, Report LiTH-MAT-R-79-8, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1979.

[877] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, Berlin, 1970.

[878] P. R. Stopher and T. Lisco, *Modelling travel demand: a disaggregate behavioral approach—issues and applications*, Transportation Research Forum Proceedings, 11 (1970), pp. 195–214.

[879] P. R. Stopher and A. H. Meyburg, *Urban Transportation Modeling and Planning*, Lexington Books, Lexington, MA, 1975.

[880] S. A. Stouffer, *Intervening opportunities: a theory relating mobility and distance*, American Sociological Review, 5 (1940), pp. 845–867.

[881] S. Suh and T. J. Kim, *Solving nonlinear bilevel programming models of the equilibrium network design problem: a comparative review*, in Hierarchical Optimization, G. Anandalingam and T. L. Friesz, eds., vol. 34 of Annals of Operations Research, J. C. Baltzer AG, Basel, Switzerland, 1992, pp. 203–218.

[882] C. E. Sweet, Jr., *Guidelines for the administration of urban transportation planning*, report, Institute of Traffic Engineers, Washington, D.C., 1969.

[883] F. Tagliacozzo and F. Pirzio, *Assignment models and urban path selection criteria: results of a survey of the behaviour of road users*, Transportation Research, 7 (1973), pp. 313–329.

[884] A. Taguchi, *Braess' paradox in a two-terminal transportation network*, Journal of the Operations Research Society of Japan, 25 (1982), pp. 376–388.

[885] K. Taji, M. Fukushima, and T. Ibaraki, *A globally convergent Newton method for solving strongly monotone variational inequalities*, Mathematical Programming, 58 (1993), pp. 369–383.

[886] T. Takayama and G. G. Judge, *An intertemporal price equilibrium model*, Journal of Farm Economics, 46 (1964), pp. 477–484.

[887] ——, *Spatial and Temporal Price and Allocation Models*, North-Holland, Amsterdam, 1971.

[888] J. C. Tanner, *Pricing the use of the roads: a mathematical and numerical study*, in Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow, London, June 25–27, 1963, J. Almond, ed., The Organisation for Economic Co-operation and Development, Paris, 1965, pp. 317–345.

[889] M. A. P. Taylor, *On Davidson's flow rate-travel time relationship*, Australian Road Research, 7 (1977), pp. 3–13.

[890] ——, *A note on using Davidson's function in equilibrium assignment*, Transportation Research, 18B (1984), pp. 181–199.

[891] W. C. Taylor, *Optimization of traffic flow splits*, Highway Research Record, 230 (1968), pp. 60–77.

[892] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Mathematics of Operations Research, 17 (1992), pp. 670–690.

[893] E. N. Thomas and J. L. Schofer, *Strategies for the evaluation of alternative transportation plans*, National Cooperative Highway Research Program Report 96, Highway Research Board, Washington, D.C., 1970.

[894] R. Thomas, *Traffic Assignment Techniques*, Avebury Technical, Aldershot, Hampshire, England, 1991.

[895] W. A. Thompson, *Traffic assignment for capacitated transportation networks including queueing—with application to freeway corridor control*, PhD thesis, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 1976.

[896] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Mathematics Doklady, 4 (1963), pp. 1035–1038.

[897] A. N. Tikhonov and V. Ya. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley & Sons, New York, NY, 1977.

[898] R. L. Tobin, *An extension of Dial's algorithm utilizing a model of tripmakers' perceptions*, Transportation Research, 11 (1977), pp. 337–342.

[899] ——, *Sensitivity analysis for variational inequalities*, Journal of Optimization Theory and Applications, 48 (1986), pp. 191–204.

[900] R. L. Tobin and T. L. Friesz, *Sensitivity analysis for equilibrium network flow*, Transportation Science, 22 (1988), pp. 242–250.

[901] M. J. Todd, *The Computation of Fixed Points and Applications*, vol. 124 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976.

[902] Ph. L. Toint and D. Tuyttens, *On large scale nonlinear network optimization*, Mathematical Programming, 48 (1990), pp. 125–159.

[903] A. R. Tomazinis, *A new method of trip distribution in an urban area*, Highway Research Board Bulletin, 347 (1962), pp. 77–99.

[904] J. A. Tomlin, *Minimum-cost multicommodity network flows*, Operations Research, 14 (1966), pp. 45–51.

[905] ——, *Mathematical programming models for traffic network problems*, PhD thesis, Department of Mathematics, University of Adelaide, Adelaide, Australia, 1967.

[906] ——, *A mathematical programming model for the combined distribution-assignment of traffic*, Transportation Science, 5 (1971), pp. 122–140.

[907] J. A. Tomlin and S. G. Tomlin, *Traffic distribution and entropy*, Nature, 220 (1968), pp. 974–976.

[908] D. L. Trueblood, *Effect of travel time and distance on freeway usage*, Highway Research Board Bulletin, 61 (1952), pp. 18–37.

[909] P. Tseng, *Dual ascent methods for problems with strictly convex costs and linear constraints: a unified approach*, SIAM Journal on Control and Optimization, 28 (1990), pp. 214–242.

[910] P. Tseng, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Mathematical Programming, 48 (1990), pp. 249–263.

[911] ——, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM Journal on Control and Optimization, 29 (1991), pp. 119–138.

[912] ——, *Decomposition algorithm for convex differentiable minimization*, Journal of Optimization Theory and Applications, 70 (1991), pp. 109–135.

[913] ——, *On the rate of convergence of a partially asynchronous gradient projection algorithm*, SIAM Journal on Optimization, 1 (1992), pp. 603–619.

[914] P. Tseng and D. P. Bertsekas, *Relaxation methods for problems with strictly convex costs and linear constraints*, Mathematics of Operations Research, 16 (1991), pp. 462–481.

[915] P. Tseng, D. P. Bertsekas, and J. N. Tsitsiklis, *Partially asynchronous, parallel algorithms for network flow and other problems*, SIAM Journal on Control and Optimization, 28 (1990), pp. 678–710.

[916] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[917] J. N. Tsitsiklis and D. P. Bertsekas, *Distributed asynchronous optimal routing in data networks*, IEEE Transactions on Automatic Control, AC-31 (1986), pp. 325–332.

[918] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control, AC-31 (1986), pp. 803–812.

[919] U. S. Bureau of Public Roads, *Traffic assignment manual*, U. S. Bureau of Public Roads, U. S. Government Printing Office, Washington, D.C., 1964.

[920] ——, *Calibrating and testing a gravity model for any size urban area*, U. S. Bureau of Public Roads, U. S. Government Printing Office, Washington, D.C., 1965.

[921] ——, *Modal split: documentation of nine methods for estimating transit usage*, U. S. Bureau of Public Roads, U. S. Government Printing Office, Washington, D.C., 1966.

[922] ——, *Guidelines for trip generation analysis*, U. S. Bureau of Public Roads, U. S. Government Printing Office, Washington, D.C., 1967.

[923] U. S. Bureau of the Census, *Census of Population, Volume II: Characteristics of the Population*, U. S. Bureau of the Census, U. S. Government Printing Office, Washington, D.C., 1953.

[924] Urban Mass Transportation Administration, *UMTA Transportation Planning System Reference Manual*, U. S. Department of Transportation, Washington, D.C., 1977.

[925] M. M. Vainberg, *Variational Method and Method of Monotone Operators in the Theory of Nonlinear Equations*, John Wiley & Sons, New York, NY, 1973. Translated from the Russian by A. Libin.

[926] C. Van de Panne and A. Whinston, *The simplex and the dual method for quadratic programming*, Operational Research Quarterly, 15 (1964), pp. 355–388.

[927] ——, *Simplicial methods for quadratic programming*, Naval Research Logistics Quarterly, 11 (1964), pp. 273–302.

[928] ——, *A comparison of two methods for quadratic programming*, Operations Research, 14 (1966), pp. 422–441.

[929] ——, *A parametric simplicial formulation of Houthakker's capacity method*, Econometrica, 34 (1966), pp. 354–380.

[930] ——, *An alternative interpretation of the primal-dual method and some related parametric methods*, International Economic Review, 9 (1968), pp. 87–99.

[931] ——, *The symmetric formulation of the simplex method for quadratic programming*, Econometrica, 37 (1969), pp. 507–527.

[932] D. Van Vliet, *Road assignment: a case for incremental loading*, GLTS Note 30, Department of Planning and Transportation, Greater London Council, 1973.

[933] ——, *The choice of assignment techniques for large networks*, in Traffic Equilibrium Methods, Proceedings of the International Symposium Held at the Université de Montréal, November 21–23, 1974, M. A. Florian, ed., vol. 118 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1976, pp. 396–412.

[934] ——, *Road assignment—I: principles and parameters of model formulation*, Transportation Research, 10 (1976), pp. 137–143.

[935] ——, *Road assignment—II: the GLTS model*, Transportation Research, 10 (1976), pp. 145–149.

[936] ——, *An application of mathematical programming to network assignment*, in Urban Transportation Planning: Current Themes and Future Prospects, P. Bonsall, Q. Dalvi, and P. J. Hills, eds., Abacus Press, England, 1977, pp. 147–158.

[937] ——, *D'Esopo: a forgotten tree-building algorithm*, Traffic Engineering & Control, 18 (1977), pp. 372–373.

[938] ——, *Improved shortest path algorithms for transport networks*, Transportation Research, 12 (1978), pp. 7–20.

[939] D. Van Vliet, *SATURN: a modern assignment model*, Traffic Engineering & Control, 23 (1982), pp. 578–581.
[940] ——, *Modelling route guidance systems using SATURN*, in DRIVE Project V1054 (ASTERIX) Deliverable 4, 1990, pp. B–83–B–105.
[941] D. Van Vliet and P. D. C. Dow, *Capacity-restrained road assignment*, Traffic Engineering & Control, 20 (1979), pp. 296–305.
[942] D. Van Vliet and H. Schittenhelm, *A proportional Frank–Wolfe algorithm*, tech. report, Institute for Transport Studies, University of Leeds, Leeds, 1991.
[943] T. Van Vuren and D. Van Vliet, *Route Choice and Signal Control: The Potential for Integrated Route Guidance*, Avebury, Aldershot, Hampshire, England, 1992.
[944] G. Vanderstraeten-Tilquin, *Problèmes de circulation avec coûts convexes*, Publication 268, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, 1977.
[945] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[946] J. A. Ventura, *Computational development of a Lagrangian dual approach for quadratic networks*, Networks, 21 (1991), pp. 469–485.
[947] W. Vickrey, *Pricing in urban and suburban transport*, American Economic Review, 53 (1963), pp. 452–465.
[948] ——, *Pricing as a tool in coordination of local transportation*, in Transportation Economics, Columbia University Press, New York, NY, 1965, pp. 275–296.
[949] ——, *Optimization of traffic and facilities*, Journal of Transport Economics and Policy, 1 (1967), pp. 123–136.
[950] ——, *'Congestion charges and welfare': some answers to Sharp's doubts*, Journal of Transport Economics and Policy, 2 (1968), pp. 107–118.
[951] H. Von Falkenhausen, *Traffic assignment by a stochastic model*, in Proceedings of the 4th International Conference on Operational Science, 1966, pp. 415–421.
[952] B. von Hohenbalken, *A finite algorithm to maximize certain pseudoconcave functions on polytopes*, Mathematical Programming, 9 (1975), pp. 189–206.
[953] ——, *Simplicial decomposition in nonlinear programming algorithms*, Mathematical Programming, 13 (1977), pp. 49–68.
[954] A. M. Voorhees, *Forecasting peak hours of travel*, Highway Research Board Bulletin, 203 (1958), pp. 37–46.
[955] A. A. Walters, *The theory and measurement of private and social cost of highway congestion*, Econometrica, 29 (1961), pp. 676–699.
[956] ——, *The Economics of Road User Charges*, vol. 5, International Bank for Reconstruction and Development, Baltimore, MD, 1968.
[957] J. Wang, *Accelerated convergence in traffic assignment by social pressure approach*, master's thesis, Institute for Transport Studies, University of Leeds, Leeds, 1993.
[958] J. G. Wardrop, *Some theoretical aspects of road traffic research*, Proceedings of the Institute of Civil Engineers, Part II, (1952), pp. 325–378.
[959] S. Warshall, *A theorem on boolean matrices*, Journal of the Association of Computing Machinery, 9 (1962), pp. 11–12.
[960] H. S. Weigel and J. E. Cremeans, *The multicommodity network flow model revised to include vehicle per time period and node constraints*, Naval Research Logistics Quarterly, 19 (1972), pp. 77–89.
[961] E. Weiner, *A modal split model for Southeastern Wisconsin*, Technical Records 6, Southeastern Wisconsin Regional Planning Commission, 1966.
[962] A. Weintraub, *A primal algorithm to solve network flow problems with convex costs*, Management Science, 21 (1974), pp. 87–97.
[963] ——, *Optimal flows and games: the multicommodity flow problem in integers*, Working Paper 76/21/C, Departmento de Industrias, University of Chile, 1976.
[964] A. Weintraub and J. González, *An algorithm for the traffic assignment problem*, Networks, 10 (1980), pp. 197–209.
[965] A. Weintraub, C. Ortiz, and J. González, *Accelerating convergence of the Frank–Wolfe algorithm*, Transportation Research, 19B (1985), pp. 113–122.
[966] J. Weiss, J. Gottfried, and T. L. Friesz, *Numerical experience with diagonalization/relaxation algorithms for asymmetric demand traffic assignment.* Paper presented at the 13th TIMS/ORSA Joint National Meeting, Detroit, MI, 1982.
[967] D. J. White, *The use of the concept of entropy in system modelling*, Operational Research Quarterly, 21 (1970), pp. 279–281.
[968] P. D. Whiting and J. A. Hillier, *A method for finding the shortest route through a road network*, Operational Research Quarterly, 11 (1960), pp. 37–40.
[969] B.-W. Wie, *Dynamic models of network traffic assignment: a control theoretic approach*, PhD thesis, University of Pennsylvania, 1988.

[970] B.-W. Wie, T. L. Friesz, and R. L. Tobin, *Dynamic user optimal traffic assignment: a control theoretic formulation*, in Dynamic Control and Flow Equilibrium, Proceedings of the Italy-U. S. A. Joint Seminar on Urban Traffic Networks, Naples and Capri, Italy, June 20–23, 1989, pp. 113–149.

[971] M. R. Wigan, *Benefit assessment for network traffic models and application to road pricing*, RRL Report LR 417, Road Research Laboratory, Crowthorne, Berkshire, England, 1971.

[972] ——, *Theory and implementation*, in Techniques for Transport and Systems Analysis, Australian Road Research Board, 1977, ch. 17, pp. 135–147.

[973] M. R. Wigan and T. J. G. Bamford, *The effects of network structure on the benefits derivable from road pricing*, Laboratory Report LR 557, Department of the Environment, Road Research Laboratory, Crowthorne, Berkshire, England, 1973.

[974] M. R. Wigan and J. Y. Luk, *Equilibrium assignment for fixed travel demand: an initial appraisal of its practical utility*, ARR Report 68, Australian Road Research Board, 1976.

[975] D. F. Wilkie and R. G. Stefanek, *Precise determination of equilibrium in travel forecasting problems using numerical optimization techniques*, Highway Research Record, 369 (1971), pp. 239–252.

[976] H. C. W. L. Williams, *On the formation of travel demand models and economic evaluation measures of user benefit*, Environment and Planning, 9A (1977), pp. 285–344.

[977] A. G. Wilson, *A statistical theory of spatial distribution models*, Transportation Research, 1 (1967), pp. 253–269.

[978] ——, *The use of entropy maximizing models in the theory of trip distribution, mode split, and route split*, Journal of Transport Economics and Policy, 3 (1969).

[979] ——, *Entropy in Urban and Regional Modelling*, Pion, London, 1970.

[980] A. G. Wilson, J. D. Coelho, S. M. Macgill, and H. C. W. L. Williams, *Optimization in Locational and Transport Analysis*, John Wiley & Sons, Chichester, U. K., 1981.

[981] L. Wingo and H. Perloff, *The Washington transportation plan: technics or politics?*, Proceedings and Papers of the Regional Science Association, (1961).

[982] D. K. Witheford, *Comparison of trip distribution by opportunity model and gravity model*, Highway Research Board Bulletin, (1961).

[983] ——, *Traffic assignment analysis and evaluation*, Highway Research Record, 6 (1963), pp. 1–11.

[984] M. Wohl and B. V. Martin, *Traffic System Analysis for Engineers and Planners*, McGraw-Hill, New York, NY, 1967.

[985] P. Wolfe, *The simplex method for quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

[986] ——, *Methods of nonlinear programming*, in Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, NY, 1963, pp. 67–86.

[987] ——, *Convergence theory in nonlinear programming*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 1–36.

[988] ——, *Algorithm for a least-distance programming problem*, Mathematical Programming Study, 1 (1974), pp. 190–205.

[989] ——, *Finding the nearest point in a polytope*, Mathematical Programming, 11 (1976), pp. 128–149.

[990] R. Wollmer, *Removing arcs from a network*, Operations Research, 12 (1964), pp. 934–940.

[991] R. D. Wollmer, *Multicommodity networks with resource constraints: the generalized multicommodity flow problem*, Networks, 1 (1972), pp. 245–263.

[992] H. J. Wootton, M. P. Ness, and R. S. Burton, *Improved direction signs and the benefits for road users*, Traffic Engineering & Control, 22 (1981), pp. 264–268.

[993] H. J. Wootton and G. W. Pick, *Travel estimates from census data*, Traffic Engineering & Control, 9 (1967), pp. 142–145, 152.

[994] J. H. Wu, M. Florian, and P. Marcotte, *A general descent framework for the monotone variational inequality problem*, Mathematical Programming, 61 (1993), pp. 281–300.

[995] S. Yagar, *Dynamic traffic assignment by individual path minimization and queuing*, Transportation Research, 5 (1970), pp. 179–196.

[996] ——, *A study of the travel patterns in a corridor with reference to the assignment principles of Wardrop*, in Traffic Flow and Transportation, Proceedings of the 5th International Symposium on the Theory of Traffic Flow and Transportation, G. F. Newell, ed., American Elsevier, New York, NY, 1972, pp. 165–181.

[997] B. Yaged, Jr., *Minimum cost routing for static network models*, Networks, 1 (1971), pp. 139–172.

[998] ——, *Minimum cost routing for dynamic network models*, Networks, 3 (1973), pp. 193–224.

[999] H. Yang, Y. Iida, and T. Sasaki, *The equilibrium-based origin-destination matrix estimation problem*, Transportation Research, 28B (1994), pp. 23–33.

[1000] T.-C. Yang and R. R. Snell, *Traffic assignment by the maximum principle*, Journal of the Highway Division of the American Society of Civil Engineers, 92 (1966), pp. 1–14.

[1001] N. Zadeh, *A note on the cyclic coordinate ascent method*, Management Science, 16 (1969/1970), pp. 642–644.

[1002] W. I. Zangwill, *The convex simplex method*, Management Science, 14 (1967), pp. 221–238.

[1003] W. I. Zangwill, *Convergence conditions for nonlinear programming algorithms*, Management Science, 16 (1969), pp. 1–13.
[1004] ——, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
[1005] W. I. Zangwill and C. B. Garcia, *Equilibrium programming: the path following approach and dynamics*, Mathematical Programming, 21 (1981), pp. 262–289.
[1006] E. Zeidler, *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*, Springer-Verlag, New York, NY, 1990.
[1007] S. A. Zenios, *Data parallel computing for network-structured optimization problems*, Computational Optimization and Applications, 3 (1994), pp. 199–242.
[1008] S. A. Zenios and Y. Censor, *Massively parallel row-action algorithms for some nonlinear transportation problems*, SIAM Journal on Optimization, 1 (1991), pp. 373–400.
[1009] S. A. Zenios and J. M. Mulvey, *Relaxation techniques for strictly convex network problems*, Annals of Operations Research, 5 (1985/86), pp. 517–538.
[1010] ——, *A distributed algorithm for convex network optimization problems*, Parallel Computing, 6 (1988), pp. 45–56.
[1011] S. A. Zenios and M. Ç. Pinar, *Parallel block-partitioning of truncated Newton for nonlinear network optimization*, SIAM Journal on Scientific and Statistical Computing, 13 (1992), pp. 1173–1193.
[1012] D. L. Zhu and P. Marcotte, *Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities*, Publication CRT-926, Centre de Recherche sur les Transports, Université de Montréal, Montréal, Canada, 1993.
[1013] ——, *Modified descent methods for solving the monotone variational inequality problem*, Operations Research Letters, 14 (1993), pp. 111–120.
[1014] ——, *An extended descent framework for variational inequalities*, Journal of Optimization Theory and Applications, 80 (1994), pp. 349–366.
[1015] P. H. Zipkin, *Bounds for aggregating nodes in network problems*, Mathematical Programming, 19 (1980), pp. 155–177.
[1016] G. Zoutendijk, *Methods of Feasible Directions: a Study in Linear and Non-linear Programming*, Elsevier, Amsterdam, 1960.
[1017] S. I. Zuhovickiĭ, R. A. Polyak, and M. E. Primak, *Numerical methods of finding equilibrium points of n-person games*, in Proceedings of the First Winter School on Mathematical Programming at Drogobych, 1969, pp. 93–130.
[1018] ——, *Two methods of search for equilibrium points of n-person concave games*, Soviet Mathematics Doklady, 10 (1969), pp. 279–282.
[1019] ——, *On an n-person concave game and a production model*, Soviet Mathematics Doklady, 11 (1970), pp. 522–526.
[1020] ——, *Concave multiperson games: numerical methods*, Matekon, 9 (1973), pp. 10–30.

Index