Traffic equilibrium

Patrice Marcotte* and Michael Patriksson†

June 26, 2006

1 Background

The subject of traffic equilibrium is the description, through analytical tools, of the stationary distribution of vehicles in a transportation network. Assuming that travelers seek to minimize their individual travel cost, an equilibrium is reached when no traveler has an incentive to modify its travel decision. Historically, the term *traffic assignment* was used to describe the same phenomenon, reflecting the fact that the practice was not so much of estimating the traffic distribution through analytical models —including design or pricing aspects— than performing an assignment of travelers onto the network, typically in order to assess the performance of traffic control policies.

Traffic equilibrium is the cornerstone, or the 'inner loop', of any modern network analysis. Its definition dates back to the 1950s and hundreds of papers have been devoted to its study. In the last two decades, it has served as the quintessential benchmark by which the performance of equilibrium algorithms could be measured, and became a field of research of its own. In fact, much of the early development of algorithms for finite-dimensional variational inequality problems defined over convex sets took place within this field, and at one point or another several of the foremost researchers in mathematical programming and operations research have contributed to the theory of traffic equilibrium. Presenting a comprehensive account of the traffic equilibrium problem (or TEP, for short) would however require an entire book and, since such books already exist, is beyond the scope of the current chapter.

In the 21st century, we understand rather well how to solve the basic TEP; consequently, the challenges and interests of the transportation community have shifted towards enhanced models that generalize the basic equilibrium model. In this chapter, the authors briefly present the main theoretical and algorithmical results pertaining to the TEP, along the way improving theoretical results that were established some 20 years ago, and then focus their attention on topics that have been overlooked, such as the relationship between the Nash and Wardrop concepts, as well as on new paradigms built around the basic TEP. In that respect, the presentation style and choice of topics are highly personal, and strongly reflect the authors' inclinations. Each section is completed by a 'Bibliographical notes' section, where we outline the relevant literature and briefly mention topics not covered in the main text. Finally, the appendices provide a list of notation, as well as a primer on variational inequalities, which constitute the adequate framework for modeling network equilibrium problems, as was recognized independently by Stella Dafermos and Michael J. Smith more than 25 years ago.

Bibliographical notes

Most books on quantitative transportation analysis discuss the issue of traffic assignment. A book entirely devoted to the topic has been written by Patriksson [Pat94b]. Among several surveys in handbooks, let us mention that of Florian and Hearn [FlH95]. These works all assume that costs increase with volumes, although this does not hold in reality, where flow rates may *decrease* as the network becomes congested. Although such effects are difficult to assess in a static model, let us mention the proposal by Nesterov [Nest00] for addressing this issue in situations where congestion levels are not critical.

Much of the analysis in this chapter can be directly transposed from traffic networks to computer communication networks, where origin–destination pairs are connected computer communication centers,

*Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada H3C 3J7. E-mail: marcotte@iro.umontreal.ca

†Department of Mathematics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. E-mail: mipat@math.chalmers.se

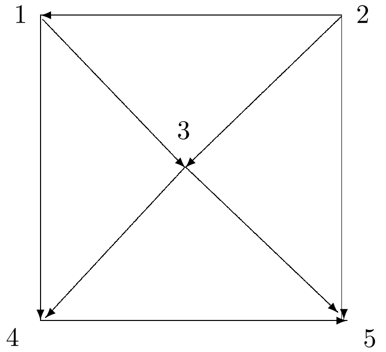


Figure 1: An example traffic network.

and where vehicles correspond to message packets being routed between them. In the fixed demand case —the most natural framework in computer communication networks—the goal is to minimize total (that is, average) delay, which usually is modeled as that of minimizing the function

$$\phi(\boldsymbol{v}) := \sum_{l \in \mathcal{L}} \frac{v_l}{c_l - v_l} + p_l v_l,$$

where v_l denotes the arrival rate of packets at link l, c_l is the transmission capacity of the link, and p_l is the processing and propagation delay on the link. Provided that the capacities c_l are sufficiently large, i.e., capacity constraints are not tight at equilibrium, most of the results presented in this chapter go through with few alterations.

2 The basic theme

This section is concerned with the basic traffic equilibrium model, including the elastic demand case, together with its many mathematical statements. Indeed, Wardrop's conditions, which express a variational principle, can be formulated in terms of a complementarity, variational inequality or optimization program, either in terms of route or link flows. Throughout the section, key notations and definitions are introduced, while a small numerical example will help in absorbing them. A comprehensive list of the notation is provided in Appendix B.

2.1 The Wardrop conditions

The basic elements of the standard traffic equilibrium model are: (i) a transportation network, (ii) travel requirements, and (iii) cost functions,[1] from which traffic volumes (or flows), expressed as vehicular rates, must be deduced. More precisely, let us consider a directed graph $\mathcal{G} := (\mathcal{N}, \mathcal{L})$, consisting of a set of nodes $i \in \mathcal{N}$ and directed links $l \in \mathcal{L}$, sometimes also denoted $l = (i, j)$ by the tail and head nodes i and j, respectively. Consider also a set $\mathcal{C} \subset \mathcal{N} \times \mathcal{N}$ of origin–destination (OD) pairs (p, q), defining starting and ending nodes of network trips in $\mathcal{G}$. Throughout the section we consider an example based on the graph illustrated in Figure 1, that involves five nodes, eight links and two OD pairs: the first has node 1 as starting node and node 5 as ending node, and the second OD pair has starting and ending nodes 2 and 4, respectively.

Any well-founded traffic model recognizes the individual network user's right to decide when, where and how to travel. A traffic equilibrium 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 sound route-choice behavioural principle.

The *equilibrium condition* refers to the concept initially introduced by the statistician J. G. Wardrop of the British Road Research Laboratory. Ever since his seminal paper, user equilibrium conditions have been known as *Wardrop's first principle*. The precise definition is as follows:

[1]Throughout the paper, the term 'cost' refers to the disutility experienced by a user of the network. In the basic models, it is a synonym for 'travel time' or 'travel delay'.

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 also introduced a second principle, whereby users choose their routes such that the *average* (or *total*) travel cost is minimal minimal. This concept will be used in Section 2.9.

To translate the concept of user equilibrium into mathematical terms, we denote by $\mathcal{R}_{pq}$ the (finite) set of simple (cycle-free) routes for OD pair (p,q), by h_r the volume of traffic on route $r \in \mathcal{R}_{pq}$, and by c_r the travel cost on the route as experienced by an individual user, given the current volume of traffic. Fixing the travel costs to these values, the user equilibrium conditions can be written as follows:

$$h_r > 0 \implies c_r = \pi_{pq}, \qquad r \in \mathcal{R}_{pq}, \qquad (p,q) \in \mathcal{C}, \tag{1a}$$

$$h_r = 0 \implies c_r \geq \pi_{pq}, \qquad r \in \mathcal{R}_{pq}, \qquad (p,q) \in \mathcal{C}, \tag{1b}$$

where π_{pq} denotes the minimal (that is, equilibrium) route cost for OD pair (p,q).

These conditions express the optimality conditions of a shortest route problem for each OD pair, where route costs are given by c_r. What makes the problem more complex than a shortest route problem is that the supply and demand characteristics are not fixed: normally, the route costs c_r depend on the volume of traffic on the routes, and the total traffic volume in an OD pair (that is, the OD demand) may depend on the least cost π_{pq} of travel, perhaps even upon other OD pairs' costs as well. In order for the above system to describe an equilibrium state, we must further incorporate the cost perception of the users given the volume of traffic, and the mechanism by which flow demand is generated.

Let therefore $c_r : \Re^{|\mathcal{R}|}_+ \to \Re$ be a real-valued function, describing the cost of utilizing route $r \in \mathcal{R}$, and whose argument is the vector $\boldsymbol{h} \in \Re^{|\mathcal{R}|}$ of route flows h_r. The vector of these route costs forms the aggregated vector-valued function $\boldsymbol{c} : \Re^{|\mathcal{R}|}_+ \to \Re^{|\mathcal{R}|}$. Further, we assume that the demand on OD pair $(p,q) \in \mathcal{C}$ is given by the real-valued, non-negative function $g_{pq} : \Re^{|\mathcal{C}|} \to \Re_+$, whose argument is the vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ of OD pair least travel costs, π_{pq}, introduced in (1). The vector-valued demand function of the least travel costs is denoted by $\boldsymbol{g} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}_+$. We also introduce the route–OD pair incidence matrix $\boldsymbol{\Gamma} \in \Re^{|\mathcal{R}|\times|\mathcal{C}|}$ whose element γ_{rk} is set to 1 if route r joins OD pair $k = (p,q) \in \mathcal{C}$, and 0 otherwise.

Based on the above notation, one may express the conditions (1) and the demand constraints as the system:

$$\mathbf{0}^{|\mathcal{R}|} \leq \boldsymbol{h} \perp (\boldsymbol{c}(\boldsymbol{h}) - \boldsymbol{\Gamma}\boldsymbol{\pi}) \geq \mathbf{0}^{|\mathcal{R}|}, \tag{2a}$$

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{h} = \boldsymbol{d}, \tag{2b}$$

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{\pi}), \tag{2c}$$

where, for two arbitrary vectors $\boldsymbol{a}, \boldsymbol{b} \in \Re^n$, the notation $\boldsymbol{a} \perp \boldsymbol{b}$ means that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = 0$. An interesting special case of the above conditions occurs when demand is *fixed*, that is, when $\boldsymbol{g} \equiv \bar{\boldsymbol{d}} \in \Re^{|\mathcal{C}|}_+$. In the sequel, we denote by H_d the set of couples $(\boldsymbol{h}, \boldsymbol{d})$ that satisfies equalities (2b)–(2c), and by H the corresponding set when demand is fixed.

Formulation (2) is a special instance of a *mixed complementarity* problem; it is the basic traffic equilibrium model from which we will derive several equivalent ones, in the forms of variational inequality, nonlinear complementarity, and optimization problems. This framework involves the pair $(\boldsymbol{h}, \boldsymbol{\pi})$ and the demand vector $\boldsymbol{d} \in \Re^{|\mathcal{C}|}$, although one could substitute its value $\boldsymbol{g}(\boldsymbol{\pi})$ to the latter. If demand is fixed, condition (2c) vanishes.

The notion of equilibrium, as described in (2), should be thought of as a steady-state evolving after a transient (disequilibrium) phase in which travelers successively adjust their route choices, seeking to minimize travel costs under prevailing traffic conditions, until a situation with stable route travel costs and route flows has been reached.

2.2 Link flow representations

As the number of routes in a practical application can be enormous, we are interested in developing representations of the Wardrop conditions that are defined in terms of link flows only.

Suppose that for every route $r \in \mathcal{R}$, the route cost $c_r(\boldsymbol{h})$ is *additive*, i.e., $c_r(\boldsymbol{h})$ is the sum of the costs of using all links defining r. Let $\boldsymbol{\Lambda} \in \{0,1\}^{|\mathcal{L}|\times|\mathcal{R}|}$ be the link–route incidence matrix, whose element λ_{lr} equals one if route $r \in \mathcal{R}$ utilizes link $l \in \mathcal{L}$, and zero otherwise. We can then write $c_r(\boldsymbol{h}) = \sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\boldsymbol{v})$, or, in compact form,

$$\boldsymbol{c}(\boldsymbol{h}) = \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{v}), \tag{3}$$

where $\boldsymbol{v} \in \Re^{|\mathcal{L}|}$ is the vector of total volumes of traffic on the links. The cost $t_l(\boldsymbol{v})$ of traversing link $l \in \mathcal{L}$ at the flow $\boldsymbol{v}$ is given by a function $t_l : \Re^{|\mathcal{L}|}_+ \to \Re$; the vector of these link costs define the function $\boldsymbol{t} : \Re^{|\mathcal{L}|}_+ \to \Re^{|\mathcal{L}|}$.

Implicit in the relationship (3) is the assumption that the pair $(\boldsymbol{h}, \boldsymbol{v})$ is consistent, in the sense that $\boldsymbol{v}$ equals the sum of the route flows:

$$v_l = \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \lambda_{lr} h_r, \qquad l \in \mathcal{L},$$

or, in vector-matrix form,

$$\boldsymbol{v} = \boldsymbol{\Lambda}\boldsymbol{h}. \tag{4}$$

From the above we derive the formula $\boldsymbol{c}(\boldsymbol{h}) = \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{\Lambda}\boldsymbol{h})$ for evaluating a set of route costs consistent with given route flows. Adding the two equations (3) and (4) to (2) yields a formulation where route flow variables are relegated to the constraints.

In our example, the link flow vector $\boldsymbol{v}$ is $\boldsymbol{v} = (v_{13}, v_{14}, v_{21}, v_{23}, v_{25}, v_{34}, v_{35}, v_{45})^{\mathrm{T}}$. The graph in Figure 1 then corresponds to the following link–route incidence matrix: in the first OD pair we identify three loop-free routes, namely the node ordering $1 \to 3 \to 5$, $1 \to 4 \to 5$, and $1 \to 3 \to 4 \to 5$, while the second OD pair has the three routes $2 \to 1 \to 4$, $2 \to 3 \to 4$, and $2 \to 1 \to 3 \to 4$. The corresponding link–route matrix $\boldsymbol{\Lambda} \in \Re^{8\times 6}$ and vector $\boldsymbol{t}$ of link cost functions are defined as

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{t}(\boldsymbol{v}) = \begin{pmatrix} t_{13}(v_{13}) \\ t_{14}(v_{14}) \\ t_{21}(v_{21}) \\ t_{23}(v_{23}) \\ t_{25}(v_{25}) \\ t_{34}(v_{34}) \\ t_{35}(v_{35}) \\ t_{45}(v_{45}) \end{pmatrix} := \begin{pmatrix} 1 + \frac{1}{2}v_{13}, \\ 1 + v_{14}, \\ 1 + 5v_{21}, \\ 2 + 3v_{23}, \\ 4 + v_{25}, \\ 3 + v_{34}, \\ 2 + 2v_{35}, \\ 1 + 4v_{45} \end{pmatrix}.$$

In this network we also define the (separable) OD demand functions $g_{15}(\pi_{15}) := 11 - \pi_{15}$ and $g_{24}(\pi_{24}) := 11 - \pi_{24}$, with inverses $\pi_{15} = g^{-1}(d_{15}) =: \xi_{15}(d_{15}) = 11 - d_{15}$ and $\pi_{24} = g^{-1}(d_{24}) =: \xi_{24}(d_{24}) = 11 - d_{24}$, respectively.

The reader is invited to check that $\boldsymbol{\pi}^* = (8, 9)^{\mathrm{T}}$, with $\boldsymbol{g}(\boldsymbol{\pi}^*) = (3, 2)^{\mathrm{T}}$. Further, a vector of equilibrium route flows is $\boldsymbol{h}^* = (h^*_{11}, h^*_{12}, h^*_{13}, h^*_{21}, h^*_{22}, h^*_{23})^{\mathrm{T}} = (2, 1, 0, 1, 1, 0)^{\mathrm{T}}$, and a vector of equilibrium link flows is $\boldsymbol{v}^* = (2, 2, 1, 1, 0, 1, 2, 1)^{\mathrm{T}}$. The corresponding link costs are $\boldsymbol{t}(\boldsymbol{v}^*) = (2, 3, 6, 5, 4, 4, 6, 5)^{\mathrm{T}}$. Regarding the equilibrium conditions, we observe, for the first OD pair, that the route costs are 8, 8, and 11, respectively. According to the equilibrium principle, the more costly third root will not be used. Similarly, in the second OD pair, the route costs are 9, 9, and 12, respectively. The reader may check that the equations $\boldsymbol{v}^* = \boldsymbol{\Lambda}\boldsymbol{h}^*$ and $\boldsymbol{c}(\boldsymbol{h}^*) = \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{v}^*)$ are satisfied.

Another formulation involves the origin-destination flows, also referred to as *commodity flows*,[2] and dispenses with route flow variables. Towards this aim, we introduce the *link–node incidence matrix* $\boldsymbol{E} \in \{-1, 0, 1\}^{|\mathcal{N}|\times|\mathcal{L}|}$, whose element e_{il} equals -1 if node i is the origin node of link l, 1 if node i is the destination node of link l, and 0 otherwise. The link–node version of Wardrop's equilibrium conditions states that at an equilibrium link flow $\boldsymbol{v}$, equal to the aggregate of commodity (OD pair) volumes $\boldsymbol{w}_k \in \Re^{|\mathcal{L}|}$, $k := (p, q) \in \mathcal{C}$, there exist vectors $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$, $k \in \mathcal{C}$, of node prices (laternatively node potentials or dual variables) such that for a given link $(i, j) \in \mathcal{L}$,

$$0 \le w_{ijk} \perp (t_{ij}(\boldsymbol{v}) - [\pi_{jk} - \pi_{ik}]) \ge 0, \qquad k \in \mathcal{C}, \tag{5}$$

or,

$$\boldsymbol{0}^{|\mathcal{L}|} \le \boldsymbol{w}_k \perp \left(\boldsymbol{t}(\boldsymbol{v}) - \boldsymbol{E}^{\mathrm{T}}\boldsymbol{\pi}_k\right) \ge \boldsymbol{0}^{|\mathcal{L}|}, \qquad k \in C. \tag{6a}$$

[2] The term *commodity* stems from the economic literature, where an OD flow is associated with a specific good.

To confirm the agreement with (2a), we select any OD pair $k = (p, q) \in \mathcal{C}$ and a route $r \in \mathcal{R}_{pq}$, and consider a consistent set of volumes $\boldsymbol{w}$, $\boldsymbol{v}$, and $\boldsymbol{h}$. Summing the above conditions (5) over route r, we obtain that

$$\sum_{l=(i,j)\in\mathcal{L}} \lambda_{lr} \left(t_l(\boldsymbol{v}) - [\pi_{jk} - \pi_{ik}]\right) = c_r(\boldsymbol{h}) - [\pi_{qk} - \pi_{pk}],$$

and since $\sum_{l=(i,j)\in\mathcal{L}} \lambda_{lr} w_{kl} = h_r$ can be made to hold by the Flow Decomposition Theorem 2 (see below), we conclude that (2a) and (6a) are equivalent, provided we identify π_k with $\pi_{qk} - \pi_{pk}$.

Indeed, let us introduce the indicator vector $\boldsymbol{i}_k \in \{-1, 0, 1\}^{|\mathcal{N}|}$, which is zero in all positions but two where, by the sign convention introduced earlier for the incidence matrix $\boldsymbol{E}$, the element with value 1 (-1) corresponds to the sink (source) node. The demand-feasibility relation (2b) is then expressed as

$$\boldsymbol{E}\boldsymbol{w}_k = \boldsymbol{i}_k g_k([\boldsymbol{i}_\kappa^{\mathrm{T}} \boldsymbol{\pi}_\kappa]_{(\kappa\in\mathcal{C})}), \qquad k \in \mathcal{C}. \tag{6b}$$

In the case of fixed demands, the latter right-hand sides are simply replaced by $\boldsymbol{i}_k \bar{d}_k$, where $\bar{d}_k \in \Re_+$ is the fixed demand for the OD pair k. We will later establish under which additional conditions the representations (2) and (6) of Wardrop's user equilibrium conditions are equivalent.

Some comments are in order. First, the vector $\boldsymbol{i}_k$ relates the OD node price vectors $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$ appearing in the link–node flow representation (6) and the OD least cost vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ appearing in the link–route version (2), through the relation

$$\boldsymbol{\pi} = (\pi_{pq})_{(p,q)\in\mathcal{C}} = (\boldsymbol{i}_k^{\mathrm{T}} \boldsymbol{\pi}_k)_{(k\in\mathcal{C})};$$

this is the same as stating that, for each $k \in \mathcal{C}$, $\pi_k = \pi_{qk} - \pi_{pk}$, as discussed above.

This relationship allows for a simplification of the right-hand-side of (6b). Further, the vectors $\boldsymbol{i}_k$ clarify the relationships between the respective demand vectors. In link-flow based transportation problems, the (fixed) demand in some commodity k is denoted by an $|\mathcal{N}|$-vector $\boldsymbol{d}_k$, wherein a positive (negative) element represents a sink (source) node and a null element a transshipment node. In the traffic equilibrium problem, commodities are normally associated with OD pairs, whence the vector $\boldsymbol{d}_k$ just mentioned equals the $|\mathcal{N}|$-vector $\boldsymbol{i}_k \bar{d}_k$, which has precisely two non-zero entries.

Returning to our numerical example, the node–link incidence matrix $\boldsymbol{E} \in \Re^{5\times 8}$ is

$$\boldsymbol{E} = \begin{pmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

A commodity link flow in equilibrium is given by $\boldsymbol{w}_1^* = (0, 1, 1, 1, 0, 1, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{w}_2^* = (2, 1, 0, 0, 0, 0, 2, 1)^{\mathrm{T}}$; notice that $\boldsymbol{w}_1^* + \boldsymbol{w}_2^* = \boldsymbol{v}^*$. Further, we can relate the equilibrium OD travel costs $\boldsymbol{\pi}^* = (9, 8)^{\mathrm{T}}$ to the node prices associated with the above network at equilibrium. Under the natural ordering of the nodes, the node price vector for the respective OD pair is

$$\boldsymbol{\pi}_1^* = \begin{pmatrix} 0 \\ +\infty \\ 2 \\ 3 \\ 8 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\pi}_2^* = \begin{pmatrix} 6 \\ 0 \\ 5 \\ 9 \\ 5 \end{pmatrix},$$

as a simple calculation, (in principle) using Dijkstra's algorithm, would show. With the vectors $\boldsymbol{i}_1 = (-1, 0, 0, 0, 1)^{\mathrm{T}}$ and $\boldsymbol{i}_2 = (0, -1, 0, 1, 0)^{\mathrm{T}}$, we see that the difference in node prices between the ending and starting nodes of the two OD pairs yield precisely the elements of $\boldsymbol{\pi}^* = (9, 8)^{\mathrm{T}}$, which confirms the above connection. This also shows the familiar behaviour of node prices in network flow models: their values are only given up to an arbitrary common additive constant; we have resolved this problem by setting the node price for the starting node in each OD pair to zero.

In Section 2.7 we show that suitable properties of the demand and travel cost functions imply that the entities $\boldsymbol{\pi}^*$, $\boldsymbol{d}^* = \boldsymbol{g}(\boldsymbol{\pi}^*)$ and $\boldsymbol{v}^*$ are unique at equilibrium, while $\boldsymbol{h}^*$ and $\boldsymbol{w}^*$ are not guaranteed to be unique. The reader is asked to verify that in our case they are unique regardless.

2.3 Variational inequality representations

Variational inequalities provide a convenient framework for the modeling of equilibrium problems, and are briefly reviewed in Appendix A. From now on we often use the short-hand 'VIP' to refer to a variational inequality problem. The notation VIP$(\boldsymbol{f}, X)$ refers to a variational inequality problem of the following form: find a vector $\boldsymbol{x}^* \in X$ such that

$$\boldsymbol{f}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X,$$

where $X \subseteq \Re^n$ is nonempty and closed and $\boldsymbol{f} : \Re^n \to \Re^n$ is continuous. An equivalent statement of the variational inequality problem that will also often be used is the following:

$$-\boldsymbol{f}(\boldsymbol{x}^*) \in N_X(\boldsymbol{x}^*),$$

where $N_X(\boldsymbol{x})$ is the *normal cone* to the set X at $\boldsymbol{x} \in X$ [cf. (94)].

Link-based variational inequality models are readily developed from model (2) by incorporating the definitional constraints (4), and from the model (6) by adding the compatibility constraint

$$\boldsymbol{v} = \sum_{k \in \mathcal{C}} \boldsymbol{w}_k. \tag{7}$$

Starting from the link-route representation, the fixed demand model is VIP$(\boldsymbol{t}, \widehat{F})$, where

$$\widehat{F} := \left\{ \boldsymbol{v} \in \Re^{|\mathcal{L}|} \;\middle|\; \exists \boldsymbol{h} \in H \text{ with } \boldsymbol{v} = \boldsymbol{\Lambda h} \right\}.$$

In the case of invertible demand functions, we can write the elastic model as

$$[-\boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\xi}(\boldsymbol{d})] \in N_{\widehat{F}_d}(\boldsymbol{v}, \boldsymbol{d}), \tag{8}$$

where

$$\widehat{F}_d := \left\{ (\boldsymbol{v}, \boldsymbol{d}) \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \;\middle|\; \exists (\boldsymbol{h}, \boldsymbol{d}) \in H_d \text{ with } \boldsymbol{v} = \boldsymbol{\Lambda h} \right\}.$$

The corresponding link-node representations are as follows. The fixed demand model is VIP$(\boldsymbol{t}, F)$, with

$$F := \left\{ \boldsymbol{v} \in \Re^{|\mathcal{L}|} \;\middle|\; \exists \boldsymbol{w}_k \in \Re^{|\mathcal{L}|}_+ \text{ with } \boldsymbol{E w}_k = \boldsymbol{i}_k \bar{d}_k,\ k \in \mathcal{C}, \text{ and } v_l = \sum_{k \in \mathcal{C}} w_{lk},\ l \in \mathcal{L} \right\}.$$

and the elastic demand model is

$$[-\boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\xi}(\boldsymbol{d})] \in N_{F_d}(\boldsymbol{v}, \boldsymbol{d}), \tag{9}$$

where

$$F_d := \left\{ (\boldsymbol{v}, \boldsymbol{d}) \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \;\middle|\; \exists \boldsymbol{w}_k \in \Re^{|\mathcal{L}|}_+ \text{ with } \boldsymbol{E w}_k = \boldsymbol{i}_k d_k,\ k \in \mathcal{C}, \text{ and } v_l = \sum_{k \in \mathcal{C}} w_{lk},\ l \in \mathcal{L} \right\}.$$

Note that both sets $\widehat{F}$ and $\widehat{F}_d$ are implicitly defined, and therefore not available in closed form (since the routes in $\mathcal{R}$ normally are not enumerated and therefore unavailable), whereas the sets F and F_d are explicit.

2.4 A fixed demand reformulation of the elastic demand model

Suppose that the demand function $\boldsymbol{g}$ is upper bounded by a positive vector $\bar{\boldsymbol{g}} \in \Re^{|\mathcal{C}|}$, that is, $\boldsymbol{g}(\boldsymbol{\pi}) < \bar{\boldsymbol{g}}$ holds for every $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$. Then, it is possible to construct a *fixed* demand model which is equivalent to the elastic demand Wardrop conditions (2). The model is constructed as follows. For each OD pair $(p, q) \in \mathcal{C}$, we introduce an additional link from node p to node q which carries a non-negative flow e_{pq}, and which has as its link cost $\xi_{pq}(\bar{\boldsymbol{g}} - \boldsymbol{e})$, $\boldsymbol{e} \in \Re^{|\mathcal{C}|}$. This link carries a flow which is the difference between the upper bound $\bar{g}_{pq}$ and the flow $\sum_{r \in \mathcal{R}_{pq}} h_r$ on that OD pair in the *original* network; hence, $\boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{h} + \boldsymbol{e} = \bar{\boldsymbol{g}}$ holds. We call this new link the *excess demand* link for OD pair (p, q).

The form of this fixed demand model is such that, at equilibrium, the excess demand link carries some flow. Therefore, the costs on all used routes on that OD pair must match the value of the inverse demand function for that OD pair, and the flow in the original network equals that specified by the demand function on a least-cost route. We conclude that the fixed demand equilibrium in the modified network is equivalent to an elastic demand in the original network. The complete fixed demand model consists in finding flows $(\boldsymbol{h}, \boldsymbol{e})$ such that for some vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$,

$$\mathbf{0}^{|\mathcal{R}|} \leq \boldsymbol{h} \perp (\boldsymbol{c}(\boldsymbol{h}) - \boldsymbol{\Gamma}\boldsymbol{\pi}) \geq \mathbf{0}^{|\mathcal{R}|}, \tag{10a}$$

$$\mathbf{0}^{|\mathcal{C}|} \leq \boldsymbol{e} \perp (\boldsymbol{\xi}(\bar{\boldsymbol{g}} - \boldsymbol{e}) - \boldsymbol{\pi}) \geq \mathbf{0}^{|\mathcal{C}|}, \tag{10b}$$

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{h} + \boldsymbol{e} = \bar{\boldsymbol{g}}. \tag{10c}$$

We remark that this model is equivalent to a fixed demand VIP of the form

$$-\begin{pmatrix} \boldsymbol{c}(\boldsymbol{h}) \\ \boldsymbol{\xi}(\bar{\boldsymbol{g}} - \boldsymbol{e}) \end{pmatrix} \in N_{H_{\bar{g}}}(\boldsymbol{h}, \boldsymbol{e}), \tag{11a}$$

where

$$H_{\bar{g}} := \left\{ (\boldsymbol{h}, \boldsymbol{e}) \in \Re^{|\mathcal{R}|}_{+} \times \Re^{|\mathcal{C}|}_{+} \;\middle|\; \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{h} + \boldsymbol{e} = \bar{\boldsymbol{g}} \right\}. \tag{11b}$$

We further note that the excess demand cost

$$\boldsymbol{\kappa}(\boldsymbol{e}) := \boldsymbol{\xi}(\bar{\boldsymbol{g}} - \boldsymbol{e}), \tag{12}$$

whose domain is the open box $\{\, \boldsymbol{e} \in \Re^{|\mathcal{C}|} \mid \mathbf{0}^{|\mathcal{C}|} < \boldsymbol{e} < \bar{\boldsymbol{g}} \,\}$, is strictly monotone if and only if $-\boldsymbol{g}$ is. For this model, we have the following result:

PROPOSITION 1 (equivalent fixed demand model) *Suppose that the function $\boldsymbol{g} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$ is invertible, with inverse $\boldsymbol{\xi} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$. Suppose further that $\boldsymbol{g}$ is upper bounded by a positive demand vector $\bar{\boldsymbol{g}} \in \Re^{|\mathcal{C}|}$. Then, the fixed demand model* (10) *is equivalent to the elastic demand model* (2). □

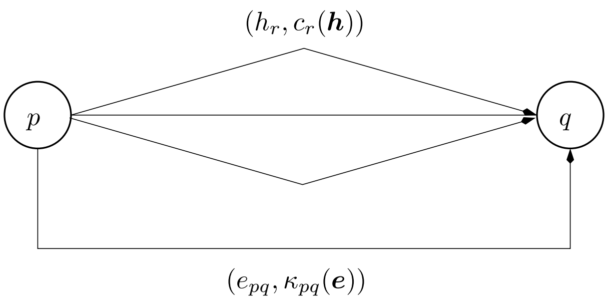


Figure 2: The excess demand link for OD pair (p, q).

2.5 The equivalence between equilibria in different representations

The models discussed in Section 2.3 are distinct. The difference between the models VIP$(\boldsymbol{t}, \widehat{F})$ [respectively, VIP$([\boldsymbol{t}, -\boldsymbol{\xi}], \widehat{F}_d)$] and VIP$(\boldsymbol{t}, F)$ [respectively, VIP$([\boldsymbol{t}, -\boldsymbol{\xi}], F_d)$] lies in possible occurrence of *cycles* in the latter's admissible sets. (Note that the definition of the sets $\mathcal{R}_{pq}$ specifies that routes are cycle-free.) We therefore have that $\widehat{F} \subseteq F$.[3] The equivalence between the link-route and link-node representations in terms of equilibria thus hinges on whether an equilibrium flow in the latter can contain a cycle; we provide such a result below. The interest in this topic is not only theoretical, as a poor modeling clearly can lead to spurious results; normally, in traffic networks, no one would travel in a cycle, but the improper modeling of a traffic equilibrium problem may lead to such a result in theory, when using the F (or, F_d) representation of flows. We note that no cycles are present in the numerical example.

[3]The $\widehat{\cdot}$ notation reflects that the set $\widehat{F}$ is bounded, while F may be unbounded.

Since the demand part of the model plays no role in the analysis, we can focus on the fixed demand case. The result below carries over trivially to elastic demand models involving bounded demand functions. The proof rests on the following flow decomposition result.

THEOREM 2 (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), which utilizes at most $|\mathcal{L}| + |\mathcal{N}|$ routes and cycles, and of which at most $|\mathcal{N}|$ are cycles.* □

We say that a link flow $\boldsymbol{v} \in F$ is *weakly acyclic* if there exists a decomposition into commodity link flows $\boldsymbol{w}_k$ that does not use all links of a directed cycle. The flow is said to be *strongly acyclic* if there exists no decomposition into commodity link flows $\boldsymbol{w}_k$ that use all the links of a directed cycle. The following relationships are immediate from the Flow Decomposition Theorem:

$$\{\, \boldsymbol{v} \in F \mid \boldsymbol{v} \text{ is strongly acyclic} \,\} \subseteq \{\, \boldsymbol{v} \in F \mid \boldsymbol{v} \text{ is weakly acyclic} \,\} \subseteq \widehat{F} \subseteq F.$$

The decomposition of a link flow solution can of course be performed with respect to other entities than routes and cycles, as was done in The Flow Decomposition Theorem 2. For example, for a fixed link flow vector $\boldsymbol{v} \in F$, consider the linear program to

$$\begin{aligned}
&\underset{(\hat{\boldsymbol{v}},\boldsymbol{\delta})}{\text{minimize}} \quad \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{v}}, && \text{(13a)}\\
&\text{subject to} \quad \hat{\boldsymbol{v}} \in F, && \text{(13b)}\\
&\qquad \hat{\boldsymbol{v}} + \boldsymbol{\delta} = \boldsymbol{v}, && \text{(13c)}\\
&\qquad \boldsymbol{\delta} \geq \mathbf{0}^{|\mathcal{L}|}. && \text{(13d)}
\end{aligned}$$

The result of this linear program is a decomposition of the link flow $\boldsymbol{v}$ into a sum $\hat{\boldsymbol{v}} + \boldsymbol{\delta}$ of a *strongly acyclic* link flow $\hat{\boldsymbol{v}}$, and a *circulation* link flow $\boldsymbol{\delta}$, that is, a flow which satisfies $\boldsymbol{E}\boldsymbol{\delta} = \mathbf{0}^{|\mathcal{N}|}$. Moreover, $\boldsymbol{v}$ is strongly acyclic if and only if $\boldsymbol{v} = \hat{\boldsymbol{v}}$. This linear program thus also provides a decomposition of a link flow in terms of extreme points and directions of the individual flow sets F_{pq}, although not disaggregated into routes and cycles in the individual commodity spaces. It is therefore different from the representation described in the Flow Decomposition Theorem, as there may be a circulation in $\hat{\boldsymbol{v}}$, although such a flow must necessarily include more than one OD pair.

It follows from the above analysis that one has to careful when implementing algorithms based on link (vs. route) flows. Indeed, convex combinations of cycle-free flows may fail to be cycle-free, unless restrictive assumptions hold.

Finally, we say that the link cost vector $\boldsymbol{t}$ is *cycle-wise non-negative* (respectively, *cycle-wise positive*) on F if the sum of the link costs of every cycle is non-negative (respectively, positive) on F.

PROPOSITION 3 (equilibria for two representations)

(a) *If $\boldsymbol{t}$ is cycle-wise non-negative on F, then every solution to* $\mathrm{VIP}(\boldsymbol{t}, \widehat{F})$ *is also a solution to* $\mathrm{VIP}(\boldsymbol{t}, F)$.

(b) *Every weakly acyclic solution to* $\mathrm{VIP}(\boldsymbol{t}, F)$ *solves* $\mathrm{VIP}(\boldsymbol{t}, \widehat{F})$.

(c) *If $\boldsymbol{t}$ is cycle-wise positive on F, then every solution to* $\mathrm{VIP}(\boldsymbol{t}, F)$ *is strongly acyclic.* □

Clearly, then, if link costs are everywhere positive the issue of possible cyclic equilibrium flows is avoided. One must however note that when considering general link toll models, zero and even negative link costs may occur, as they may appeal to certain travelers.

2.6 Reduction to an optimization problem

In the case where $\boldsymbol{t}$ is a gradient mapping,[4] the model $\mathrm{VIP}(\boldsymbol{t}, \widehat{F})$ defines the first-order optimality conditions for an optimization problem of the form

$$\underset{\boldsymbol{v} \in \widehat{F}}{\text{minimize}} \ \phi(\boldsymbol{v}) := \oint_{\mathbf{0}^{|\mathcal{L}|}}^{\boldsymbol{v}} \boldsymbol{t}(s)\, ds, \qquad (14)$$

[4]If $\boldsymbol{t}$ is continuously differentiable relative to an open convex set containing S, then $\boldsymbol{t}$ is a gradient mapping over S if and only if its Jacobian matrix $\nabla \boldsymbol{t}(\boldsymbol{v})$ is symmetric over S. The gradient property is a more general property than this symmetry property, since $\boldsymbol{t}$ need not always be differentiable; nevertheless, the term 'symmetry' is frequently used in traffic planning when it should really refer to the gradient property, also called the 'integrability' property.

where $\oint$ denotes a line integral. The equivalence between equilibria, that is, solutions to the model VIP$(\boldsymbol{t}, \widehat{F})$, and stationary points to the problem (14) follows immediately from the fact that $\nabla\phi \equiv \boldsymbol{t}$. In particular, every stationary point of problem (14) is an equilibrium link flow.

Further, if $\boldsymbol{t}$ is separable, so that t_l is a function only of v_l, $l \in \mathcal{L}$, the optimization problem assumes the more familiar form

$$\underset{\boldsymbol{v}\in\widehat{F}}{\text{minimize}} \ \phi(\boldsymbol{v}) := \sum_{l\in\mathcal{L}} \int_0^{v_l} t_l(s)\, ds. \tag{15}$$

When the demand function is separable and invertible, that is, $g_{pq}^{-1} = \xi_{pq}$, we similarly have the equivalent optimization problem

$$\underset{(\boldsymbol{v},\boldsymbol{d})\in\widehat{F}_d}{\text{minimize}} \ \phi(\boldsymbol{v}, \boldsymbol{d}) := \sum_{l\in\mathcal{L}} \int_0^{v_l} t_l(s)\, ds - \sum_{(p,q)\in\mathcal{C}} \int_0^{d_{pq}} \xi_{pq}(s)\, ds. \tag{16}$$

It follows as above from the identity $\nabla\phi \equiv (\boldsymbol{t}, -\boldsymbol{\xi})$ that elastic demand equilibria are stationary points for (16).

Since the cost and demand mappings are identical irrespective of the flow representation, the traffic equilibrium models for the link-node flow representation correspond to optimization models that have the same form as above, except that $\widehat{F}$ and $\widehat{F}_d$ are replaced, respectively, by F and F_d. The reader may provide the corresponding optimization formulation to the above numerical example, and show that the equilibrium solution satisfies the first-order optimality conditions associatedi with the mathematical program 16.

2.7 Properties of equilibria

In this section we derive the basic properties of traffic equilibria, i.e., the existence and uniqueness of the following entities: flows, demands, and equilibrium travel costs.[5]

We state the properties of the models that hold throughout this section: (i) *The network $\mathcal{G}$ is strongly connected with respect to the OD pairs;*[6]*; (ii) the function $\boldsymbol{t} : \Re_+^{|\mathcal{L}|} \to \Re^{|\mathcal{L}|}$ is single-valued and continuous on $\Re_+^{|\mathcal{L}|}$; (iii) the function $\boldsymbol{g} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$ is single-valued, continuous, non-negative and upper bounded on $\Re^{|\mathcal{C}|}$.*

2.7.1 Existence of equilibria

THEOREM 4 (existence of an equilibrium)

(a) [link–route representation] *There exists a compact set of vectors $(\boldsymbol{h}, \boldsymbol{d}, \boldsymbol{\pi})$ of flows, demands, and least travel costs that solve the traffic equilibrium model* (2).

(b) [link–node representation] *Suppose further that the function $\boldsymbol{c}$ is additive, and that the function $\boldsymbol{t}$ is cycle-wise non-negative. Then, there exists a compact set of vectors $((\boldsymbol{w}_k)_{k\in\mathcal{C}}, \boldsymbol{v}, \boldsymbol{d}, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}})$ of flows, demands, and node prices, solving the traffic equilibrium model* (6)*. Moreover, the link flows $((\boldsymbol{w}_k)_{k\in\mathcal{C}}, \boldsymbol{v})$ can be taken to be strongly acyclic.*

PROOF. (a) The proof hinges on a general existence result for variational inequalities, which we will utilize by converting the traffic equilibrium model into an equivalent VIP involving a continuous operator defined over a non-empty, convex and *compact* set. Since $\boldsymbol{g}$ is non-negative and upper bounded, the variable $\boldsymbol{h}$ resides in a bounded subset of the non-negative orthant; since $\boldsymbol{c}$ is continuous, the variable $\boldsymbol{\pi}$ resides in a compact subset of $\Re^{|\mathcal{C}|}$, although not necessarily in the non-negative orthant. Hence, the variables of the problem, $(\boldsymbol{\pi}, \boldsymbol{h})$, lie in a compact set. We can therefore, with no loss of generality, introduce additional, redundant, bounds for each variable vector, such that

$$\underline{\boldsymbol{\pi}} \le \boldsymbol{\pi} \le \overline{\boldsymbol{\pi}},$$
$$\boldsymbol{h} \le \overline{\boldsymbol{h}},$$

[5]For ease of presentation, we work with additive link costs throughout although, with regards to the existence result in Theorem 4(a), this condition is not necessary.

[6]This condition means that there exists at least one route between each OD pair.

where $\underline{\boldsymbol{\pi}} \leq \overline{\boldsymbol{\pi}} \in \Re^{|\mathcal{C}|}$ are small and large enough, respectively, and $\overline{\boldsymbol{h}} \in \Re^{|\mathcal{R}|}$ is large enough. Incorporating these constraints yields a VIP model for (2) that involves a continuous operator defined over a nonempty, convex and compact set. Hence, its solution set is nonempty and compact.[7] Furthermore, the bounds can clearly be selected such that they all are fulfilled strictly at any equilibrium solution.

(b) Suppose that the vectors $(\boldsymbol{h}, \boldsymbol{\pi})$ and $\boldsymbol{d} := \boldsymbol{g}(\boldsymbol{\pi})$ solve (2), and set $\boldsymbol{v} := \boldsymbol{\Lambda h}$; the existence of a solution follows from the result in (a) above. Then, the vector $\boldsymbol{v}$ solves the fixed demand traffic equilibrium model VIP$(\boldsymbol{t}, \widehat{F})$. We now make use of Proposition 3(a), to conclude that $\boldsymbol{v}$ also solves VIP$(\boldsymbol{t}, F)$. We next need to establish that this flow is consistent with the elastic demand model (6). To this end, we define the $|\mathcal{N}|$-vectors $\boldsymbol{d}_k$ and $\boldsymbol{\pi}_k$, $k \in \mathcal{C}$, by the use of the incidence vectors $\boldsymbol{i}_k$, $k \in \mathcal{C}$, discussed in Section 2.2, and note that the vectors $\boldsymbol{\pi}_k$, $k \in \mathcal{C}$ are consistent with the vector $\boldsymbol{\pi}$. The first result then follows.

Last, we establish that the link flow $\boldsymbol{v}$ can be taken to be strongly acyclic. Consider the perturbed link cost operator $\boldsymbol{t} + \varepsilon_\tau \cdot \mathbf{1}^{|\mathcal{L}|}$, $\varepsilon_\tau > 0$. This mapping is cycle-wise positive, so Proposition 3(c) states that every equilibrium link flow is strongly acyclic. Reasoning as above, and letting $\{\varepsilon_\tau\}$ tend to zero as τ tends to infinity, we obtain the desired results, since the set of strongly acyclic flows is compact. ∎

We note that since a fixed demand vector is a special case of the function $\boldsymbol{g}$ stated previously, the above result provides an existence result in both the elastic and fixed demand cases.

2.7.2 Uniqueness of equilibria

To establish the uniqueness of the equilibrium solution is advantageous from several perspectives, but most of all from that of practice: using an equilibrium model to make predictions about entities that are not unique at equilibrium is a hazardous at best. Secondly, uniqueness is required for the convergence of some algorithms that seek an equilibrium. Thirdly, when traffic equilibrium is but a sub-model in a more general model, uniqueness is crucial in order for the overall model to be solvable, or even consistent. For example, a sensitivity analysis is impossible to perform in entities that are not properly defined.

While the existence of an equilibrium was ensured under conditions that are almost identical for all of the models that we have formulated, this is not the case with the uniqueness issue. For example, there is a clear distinction between uniqueness results for variational inequality models and optimization models, when the cost mapping is not integrable. Further, uniqueness is a property that depends on the level of aggregation. For example, route flows (or, commodity link flows) are almost never unique at equilibrium, at least when travel costs are additive. The reason is of course that a link flow $\boldsymbol{v}$ almost always can be distributed across the routes (or, disaggregated over the commodities on the links) in infinitely many ways into a route flow $\boldsymbol{h}$ while the relation $\boldsymbol{v} = \boldsymbol{\Lambda h}$ is satisfied (or, similarly, into an infinite variety of OD link flows $\boldsymbol{w}_k$ satisfying $\sum_{k \in \mathcal{C}} \boldsymbol{w}_k = \boldsymbol{v}$). In the results to follow, we will see that the uniqueness of the entities $\boldsymbol{c}$ (and $\boldsymbol{t}$ if present), $\boldsymbol{\pi}$, and $\boldsymbol{d}$ are ensured under the mildest conditions. In contrast, stronger conditions are required to ensure the uniqueness of *even* the link flow vector $\boldsymbol{v}$.

The results below are provided for the link–route representation only. Indeed, provided that the conditions of Theorem 4(b) hold, the assumptions that imply uniqueness are the same for both representations. The theorem is stated under fairly weak conditions that involve technical concepts that are detailed in Appendix A.

THEOREM 5 (uniqueness of the equilibrium: the general case)

(a) [convexity] *Suppose that $(\boldsymbol{c}, -\boldsymbol{g})$ is monotone. Then, the sets of equilibrium route and link flows $(\boldsymbol{h}, \boldsymbol{v})$, least travel costs $\boldsymbol{\pi}$, and travel demands $\boldsymbol{d}$, are convex. Suppose further that the travel demand is fixed. Then, the same conclusion holds if the monotonicity requirement on $\boldsymbol{c}$ is replaced by the weaker pseudo-monotonicity condition.*

(b) [travel costs and demands] *Suppose that $(\boldsymbol{c}, -\boldsymbol{g})$ is monotone$^+$. Then both the functions $\boldsymbol{c}$ and $\boldsymbol{g}$ are constant over the set of equilibria. This implies that both the route costs $\boldsymbol{c}(\boldsymbol{h})$, least travel costs $\boldsymbol{\pi}$, and travel demands $\boldsymbol{d}$, are unique. If $\boldsymbol{c}$ is additive, and $\boldsymbol{t}$ is monotone$^+$, then $\boldsymbol{t}(\boldsymbol{v})$ is unique on the set of equilibria as well.*

If further, the travel demand function is constant, the same conclusion holds for the entities $\boldsymbol{c}(\boldsymbol{h})$ and $\boldsymbol{\pi}$ under the weaker requirement that $\boldsymbol{c}$ is pseudo-monotone$^+_*$. If $\boldsymbol{c}$ is additive and $\boldsymbol{t}$ is pseudo-monotone$^+_*$, then $\boldsymbol{t}(\boldsymbol{v})$ is unique on the set of equilibria as well.

[7] The idea is to construct, by means of the projection operator, a fixed point problem equivalent to the original VIP, and then to apply Brouwer's fixed point theorem.

(c) [link flows] *Suppose that* $(\boldsymbol{c}, -\boldsymbol{g})$ *is monotone, that* $\boldsymbol{c}$ *is additive, and that* $\boldsymbol{t}$ *is strictly monotone. Then, the equilibrium link flow* $\boldsymbol{v}$ *is unique, as well as* $\boldsymbol{c}$, $\boldsymbol{\pi}$, *and* $\boldsymbol{d}$, *and the set of equilibrium route flows is a polytope.*

PROOF. [Sketch] (a) That the solution set is non-empty and compact follows from the existence Theorem 4. The convexity of the solution set and its entities is a consequence of pseudo-monotonicity and the properties of projections onto convex sets. The reason why pseudo-monotonicity is not enough for convexity to hold in the elastic demand case is that pseudo-monotonicity is not preserved under addition.

(b) This result follows from a general one for the variational inequality problem VIP$(\boldsymbol{f}, X)$: if $\boldsymbol{f}$ is pseudo-monotone$^+_*$ on X, then the value $\boldsymbol{f}(\boldsymbol{x})$ is constant on SOL$(\boldsymbol{f}, X)$. Again, the property pseudo-monotone$^+_*$ must be replaced by monotone$^+$ for the elastic demand case, because the former property is not preserved under additivon.

(c) Follows easily by arguing through contradiction. $\square$

The result for the separable case follows as a corollary, because the property of monotonicity$_+$ of a mapping is identical to that of monotonicity when it is integrable, and pseudo-monotonicity$^+_*$ reduces to pseudo-monotonicity for functions defined over $\Re$.

COROLLARY 6 (uniqueness of the equilibrium: the separable case)

(a) [travel costs and demands] *Suppose that each function* g_{pq}, $(p, q) \in \mathcal{C}$, *is decreasing, and that each function* t_l, $l \in \mathcal{L}$, *is increasing. Then, the route and link costs* $\boldsymbol{c}$ *and* $\boldsymbol{t}$, *least travel costs* $\boldsymbol{\pi}$, *and travel demands* $\boldsymbol{d}$, *are unique. In particular, if each function* g_{pq}, $(p, q) \in \mathcal{C}$, *is invertible, then it is strictly decreasing, and on the set of optimal solutions for the convex optimization formulation* (16), *the equilibrium link travel costs* $\boldsymbol{t}(\boldsymbol{v})$, *least costs* $\boldsymbol{\pi}$, *and demands* $\boldsymbol{d}$ *are unique.*

Suppose further that the travel demand is fixed. Then, the same conclusion holds for the entities $\boldsymbol{c}$, $\boldsymbol{t}$, *and* $\boldsymbol{\pi}$ *on the set of optimal solutions to the convex optimization formulation* 15) *if the monotonicity requirement on* t_l *is replaced by pseudo-monotonicity.*

(b) [link flows] *Suppose that each each function* t_l, $l \in \mathcal{L}$, *is strictly increasing. Then, in the solutions to the optimization formulations* (15) *and* (16), *the equilibrium link flow* $\boldsymbol{v}$ *is unique.* $\square$

To summarize in a more convenient and compact form, we have for the elastic demand case:

$$\begin{aligned}
\boldsymbol{t} \text{ monotone}, -\boldsymbol{g} \text{ strictly monotone} &\Longrightarrow \boldsymbol{\pi}, \boldsymbol{d} \text{ unique}, \\
(\boldsymbol{t}, -\boldsymbol{g}) \text{ monotone}^+ &\Longrightarrow \boldsymbol{c}(\boldsymbol{h}), \boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\pi}, \boldsymbol{d} \text{ unique}, \\
\boldsymbol{t} \text{ strictly monotone}, -\boldsymbol{g} \text{ monotone} &\Longrightarrow \boldsymbol{v}, \boldsymbol{c}(\boldsymbol{h}), \boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\pi}, \boldsymbol{d} \text{ unique}.
\end{aligned}$$

The first, and slightly weaker, of these results is not established above, but follows easily by arguing through contradiction. For fixed demands, the corresponding results are:

$$\begin{aligned}
\boldsymbol{t} \text{ pseudo-monotone}^+_* &\Longrightarrow \boldsymbol{c}(\boldsymbol{h}), \boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\pi} \text{ unique}, \\
\boldsymbol{t} \text{ strictly monotone} &\Longrightarrow \boldsymbol{v}, \boldsymbol{c}(\boldsymbol{h}), \boldsymbol{t}(\boldsymbol{v}), \boldsymbol{\pi} \text{ unique}.
\end{aligned}$$

2.8 Alternative notions of traffic equilibria and their relations

In this section, we present a few equilibrium conditions that have emerged since the original one defined by Wardrop, and relate them to his. Some of the motivation for this development is that when modeling the effect of congestion pricing instruments, the cost term associated with the pricing scheme might take the form of a (discontinuous) step function; in order to be able to study the existence and computations of equilibria for such discontinuous travel costs, we require that the traditional equilibrium conditions be generalized to allow for certain types of discontinuities.

Alternative definitions We consider (without any loss of generality) a fixed demand traffic network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. In this network, we assume that the traffic volume $\boldsymbol{h} \in H$ can be observed. We further focus on an arbitrary OD pair $(p, q) \in \mathcal{C}$. More precisely, we consider a route $r \in \mathcal{R}_{pq}$ for which $h_r > 0$. We also consider a cost function $\boldsymbol{c} : \Re^{|\mathcal{R}|}_+ \to \Re^{|\mathcal{R}|}$, each component of which $c_r : \Re^{|\mathcal{R}|}_+ \to \Re$ is assumed to be single-valued. Last, we consider a vector $\boldsymbol{p}$ in $\Re^{|\mathcal{R}|}$ describing the move of one unit of traffic volume from route $r \in \mathcal{R}_{pq}$ to route s across the same OD pair, that is,

$$p_i = \begin{cases} -1, & \text{if } i = r, \\ 1, & \text{if } i = s, \\ 0, & \text{otherwise.} \end{cases}$$

The following definitions summarize the equilibrium conditions considered in this section. Notice that we here distinguish the terms 'Wardrop equilibrium' and 'user equilibrium', although they are frequently used interchangeably in this text. These notions coincide whenever the travel cost is continuous, as established below, whence there is no confusion in most situations.

Definition 7 (equilibrium conditions)

(a) [Wardrop equilibrium] *The traffic volume* $\boldsymbol{h}$ *is a* Wardrop equilibrium *if*

$$c_r(\boldsymbol{h}) \leq c_s(\boldsymbol{h}). \tag{17}$$

The traffic volume is a Wardrop equilibrium if for each driver, the present travel cost on any alternative route is at least as high as the present one on its present route.

(b) [user optimized] *The traffic volume* $\boldsymbol{h}$ *is* user optimized *if for some* $\alpha > 0$,

$$c_r(\boldsymbol{h}) \leq c_s(\boldsymbol{h} + \varepsilon \boldsymbol{p}), \qquad \varepsilon \in [0, \text{minimum}\,\{\alpha, h_r\}]. \tag{18}$$

The traffic volume is user optimized if any driver who switches routes experiences a travel cost that is at least as high as the present one on its present route.

(c) [equilibrated] *The traffic volume* $\boldsymbol{h}$ *is* equilibrated *if*

$$c_r(\boldsymbol{h} + \varepsilon \boldsymbol{p}) \leq c_s(\boldsymbol{h} + \varepsilon \boldsymbol{p}), \qquad \varepsilon \in [0, h_r]. \tag{19}$$

The traffic volume is equilibrated if any driver who switches routes experiences a travel cost that is at least as high as the new one on its present route.

(d) [user equilibrium] *The traffic volume* $\boldsymbol{h}$ *is a* user equilibrium *if*

$$c_r(\boldsymbol{h}) \leq \liminf_{\varepsilon \downarrow 0} c_s(\boldsymbol{h} + \varepsilon \boldsymbol{p}). \tag{20}$$

The traffic volume is a user equilibrium if no arbitrarily small packet of flow can reroute itself and lower its travel cost.

(e) [normal equilibrium] *The traffic volume* $\boldsymbol{h}$ *is a* normal equilibrium *if it satisfies*

$$\boldsymbol{c}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{h}) \geq 0, \qquad \boldsymbol{y} \in H. \tag{21}$$

The definition states that $\boldsymbol{h} \in H$ *solves the* Minty variational inequality *associated with the traffic system.* □

Relationships In the following, we show how the above definitions are related to each other. We assume here that the travel costs are additive, that is, $c_r(\boldsymbol{h}) = \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\boldsymbol{v})$ for every consistent pair $(\boldsymbol{h}, \boldsymbol{v})$ of route and link volumes. Note that the l.s.c. and u.s.c. properties[8] are additive, so if link travel costs are l.s.c. (respectively, u.s.c.) then the route costs are l.s.c. (respectively, u.s.c.) as well.

Proposition 8 (relationships between equilibrium definitions) *Let the route travel cost* $\boldsymbol{c}$ *be additive.*

(a) [Wardrop equilibria vs. user optimized flows] *Suppose that the route cost function* $\boldsymbol{c}$ *is continuous. Then, user optimized flows are Wardrop equilibria. If the route cost function* $\boldsymbol{c}$ *is in* C^1 *on* H *and, for every* $\boldsymbol{h} \in H$ *and routes* $r, s \in \mathcal{R}_{pq}$, $(p,q) \in \mathcal{C}$, *the inequality*

$$\partial c_s(\boldsymbol{h})/\partial h_s \geq \partial c_s(\boldsymbol{h})/\partial h_r \tag{22}$$

[8] A function $\phi : \Re^n \to \Re \cup \{-\infty, +\infty\}$ is *lower semicontinuous (l.s.c.)* at $\boldsymbol{x} \in \Re^n$ if $\phi(\boldsymbol{x}) = \liminf_{\boldsymbol{y} \to \boldsymbol{x}} \phi(\boldsymbol{y})$, $\boldsymbol{x} \in \Re^n$, and *upper semicontinuous (u.s.c.)* at $\boldsymbol{x}$ if $\phi(\boldsymbol{x}) = \limsup_{\boldsymbol{y} \to \boldsymbol{x}} \phi(\boldsymbol{y})$, $\boldsymbol{x} \in \Re^n$. A function ϕ is l.s.c. (respectively, u.s.c.) on the set $S \subseteq \Re^n$ if it is l.s.c. (respectively, u.s.c.) at every point $\boldsymbol{x} \in S$. The function ϕ is *continuous* at $\boldsymbol{x}$ if it is both l.s.c. and u.s.c. at $\boldsymbol{x}$.

holds, Wardrop equilibria are user optimized flows. In particular, if the link cost function $\boldsymbol{t}$ is separable and increasing, Wardrop equilibria are user optimized.

(b) [Wardrop equilibria vs. user equilibrated flows] *Suppose that the route cost function $\boldsymbol{c}$ is continuous. Then, equilibrated flows are Wardrop equilibria. If the route cost function $\boldsymbol{c}$ is pseudo-monotone on H, Wardrop equilibria are equilibrated flows.*

(c) [Wardrop equilibria vs. user equilibria] *Suppose that for each link $l \in \mathcal{L}$, t_l is l.s.c. Then, Wardrop equilibria are user equilibria. If, for each link $l \in \mathcal{L}$, t_l is u.s.c., then user equilibria are Wardrop equilibria. Whenever the link cost is continuous, the concepts of Wardrop and user equilibria coincide.*

(d) [Wardrop equilibria vs. normal equilibria] *Suppose that the link cost is l.s.c. Then, normal equilibria are Wardrop equilibria. If the link cost operator $\boldsymbol{t}$ is pseudo-monotone, Wardrop equilibria are normal equilibria.*

(e) [user equilibria vs. normal equilibria] *Suppose that the link cost is l.s.c. Then, normal equilibria are user equilibria.*

Proof. In the case of (b), that Wardrop equilibria are equilibrated flows, suppose that $\boldsymbol{c}$ is pseudo-monotone on H, and let $\boldsymbol{h} \in H$ be a Wardrop equilibrium. Then, by definition and for all $\varepsilon > 0$ with $\boldsymbol{h} + \varepsilon\boldsymbol{p} \in H$, there holds

$$0 \leq \boldsymbol{c}(\boldsymbol{h} + \varepsilon\boldsymbol{p})^{\mathrm{T}}([\boldsymbol{h} + \varepsilon\boldsymbol{p}] - \boldsymbol{h}) = \varepsilon\boldsymbol{c}(\boldsymbol{h} + \varepsilon\boldsymbol{p})^{\mathrm{T}}\boldsymbol{p} = \varepsilon[c_s(\boldsymbol{h} + \varepsilon\boldsymbol{p}) - c_r(\boldsymbol{h} + \varepsilon\boldsymbol{p})],$$

where the inequality follows since $\boldsymbol{h} \in H$ is a Wardrop equilibrium and $\boldsymbol{c}$ is pseudo-monotone on H. As $\varepsilon > 0$, we are done.

In the case of (d), the result is established by referring to the equivalence (under the assumption of pseudo-monotonicity of $\boldsymbol{t}$) between the variational inequality problem $\mathrm{VIP}(\boldsymbol{c}, H)$ and the Minty formulation (21); see also (104). In the proof of the first part (that solutions to the Minty variational inequality are solutions to the standard variational inequality problem), continuity can be replaced by lower semi-continuity. The reverse implication utilizes the notion of pseudo-monotonicity. For the other results, we refer to the Notes section. □

Remark 9 (on the relationships between the equilibrium conditions) The matrix condition (22) appearing in (a) describes a restriction on the dependence between costs on distinct routes. In broad terms, if routes that are alternatives to each other do not interact too strongly, then Wardrop equilibria are user optimized. Differentiable, separable and increasing link travel cost functions satisfy the matrix condition (22).

Under weak conditions, viz. continuity and monotonicity, four of the concepts collapse to Wardrop's definition of user equilibrium. The outlier is the user optimized concept, akin to a non-atomic Nash equilibrium solution.[9] The following counter-example shows that a Wardrop equilibrium may not be user optimized when the travel cost is asymmetric.

Consider a two-node network with parallel routes (links) and with respective costs

$$c_1(h_1, h_2) = 2h_1 + 3h_2$$
$$c_2(h_1, h_2) = 2h_2 + 25.$$

Note that this mapping is monotone. If demand is equal to 15, the flow vector $\boldsymbol{h} = (10, 5)$ is an equilibrium, with common path cost 34. Now, let us shift one flow unit from the second to the first path ($\delta = 1$). The cost vector of the new flow pattern $\boldsymbol{h}' = (11, 4)$ is $c(h_1', h_2') = (34, 33)$. Since the cost to users that have switched path has decreased[10], the initial flow $\boldsymbol{h}$ is not a Nash equilibrium. Actually, there exist neither user optimized (nor Nash) equilibria for this example. □

The relationships between the various concepts are illustrated in Figure 3.

2.9 User versus system optimality

In his seminal paper, Wardrop introduced the concepts of user and system equilibrium, the second depicting flows that minimize the total system cost $S(\boldsymbol{v}) := \boldsymbol{t}(\boldsymbol{v})^{\mathrm{T}}\boldsymbol{v}$.

[9] A Nash equilibrium between finitely many players is achieved when no player, acting on its own, can improve its payoff. A non-atomic game involves a continuum of players.

[10] In fact, all user costs have decreased.

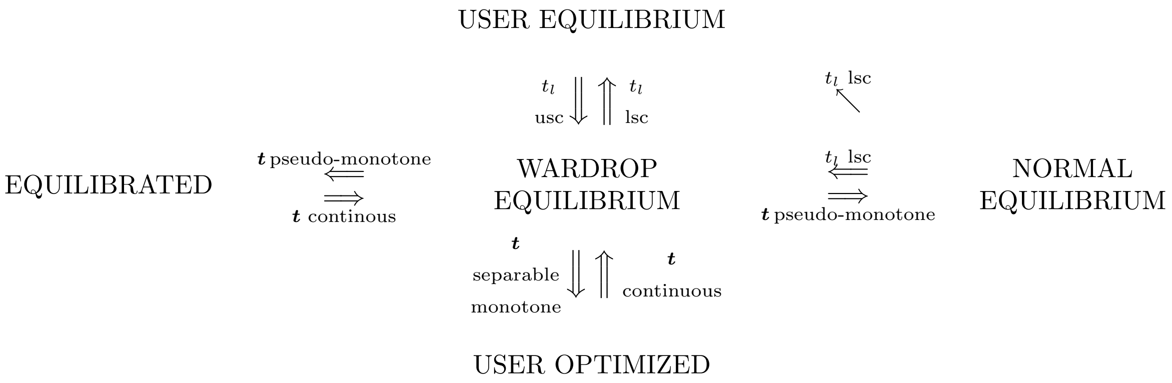
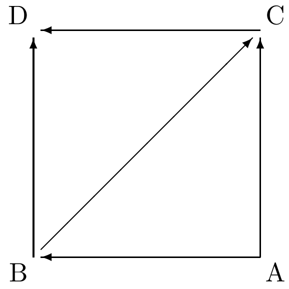


Figure 3: Relations among definitions of equilibria



$$
\begin{aligned}
t_{AB}(v_{AB}) &= 10v_{AB}\\
t_{AC}(v_{AC}) &= 50 + v_{AC}\\
t_{BD}(v_{BD}) &= 50 + v_{BD}\\
t_{CD}(v_{CD}) &= 10v_{CD}\\
t_{BC}(v_{BC}) &= 10 + v_{BC}\\
d_{AD} &= 6
\end{aligned}
$$

Figure 4: The Braess paradox

A question that naturally arises is: how different are equilibrium and system-optimal flows? How much would be gained, in terms of efficiency, from a centralized control of all traffic flow? Looking at the question from a worst-case perspective, we want to determine

$$
r(\Delta) = \underset{\Delta}{\text{maximum}} \; \frac{\boldsymbol{t}(\boldsymbol{v}_{\text{UO}})^{\text{T}}\boldsymbol{v}_{\text{UO}}}{\boldsymbol{t}(\boldsymbol{v}_{\text{SO}})^{\text{T}}\boldsymbol{v}_{\text{SO}}}, \tag{23}
$$

where $\boldsymbol{v}_{\text{SO}}$ (respectively $\boldsymbol{v}_{\text{UO}}$) denotes a system-optimal (respectively user-optimal) flow pattern, and Δ regroups the data of the equilibrium problem, that is, $\Delta = \mathcal{G} \cup \boldsymbol{g} \cup \mathcal{C} \cup \boldsymbol{t}$ in the fixed demand case. This ratio $r(\Delta)$ is the *price of anarchy* and, in order that it be well defined, we need that the system cost associated with user equilibria be unique. This condition will be satisfied under the monotonicity requirements outlined in Theorem 5.

In certain circumstances, for instance if the cost function is link-separable and assumes the monomial form

$$
t_l(v_l) = \kappa_l v_l^{\mu}, \qquad \kappa_l, \mu \geq 0, \quad l \in \mathcal{L},
$$

user and system-optimal flows coincide. This is not the case in general, as can be observed on the example of Figure 2.9, where ${\boldsymbol{v}_{\text{UO}}}^{\text{T}} = (v_{AB}, v_{AC}, v_{BC}, v_{BD}, v_{CD})^{\text{T}} = (4, 2, 2, 2, 4)$[11] and ${\boldsymbol{v}_{\text{SO}}}^{\text{T}} = (3, 3, 0, 3, 3)$. In this situation, every user is worse off at (user) equilibrium, and the price of anarchy is equal to 92/83, as can readily be computed. The following theorem provides an exact value of $r(\Delta)$ for the important class of polynomial cost functions.

Theorem 10 (the price of anarchy) *If the mapping $\boldsymbol{t}$ is link-separable, increasing and polynomial of degree p, the price of anarchy is equal to*

$$
\frac{(p+1)^{1+\frac{1}{p}}}{(p+1)^{1+\frac{1}{p}} - p} \in \Theta\left(\frac{p}{\ln p}\right).
$$

If, further, $\boldsymbol{t}$ is affine and monotone ($p = 1$), then the price of anarchy is equal to 4/3. □

[11] This corresponds to assigning flow in equal proportions on each of the three routes of the network.

Let us now define the *steepness* of a cost mapping $\boldsymbol{t}$ as

$$\sigma(\boldsymbol{t}) = \sup_{l \in \mathcal{L}, \boldsymbol{v} \in F} \frac{\boldsymbol{t}(\boldsymbol{v})^{\mathrm{T}} v}{\boldsymbol{t}(\boldsymbol{v})^{\mathrm{T}} \boldsymbol{v} + \gamma_l(\boldsymbol{v}) t_l(\gamma_l(\boldsymbol{v})) - \gamma_l(\boldsymbol{v}) t_l(\boldsymbol{v})}, \tag{24}$$

where β is a function such that $\nabla S(\beta(\boldsymbol{v})) = \boldsymbol{t}(\boldsymbol{v})$.

THEOREM 11 (the price of anarchy revisited) *Let $\mathcal{S}$ denote the class of cost mappings that are monotone gradient mappings and for which the system cost S is convex. Then*

$$r(\Delta) \leq \sup_{0 \neq \boldsymbol{t} \in \mathcal{S}} \sigma(\boldsymbol{t})$$

holds. □

Note that the above result is mainly of theoretical interest, as the computation of the steepness appears intractable for general asymmetric mappings. Moreover, these bounds are unlikely to be tight in practical situations.

Bibliographical notes

While Wardrop was first in providing a complete mathematical description of the user equilibrium conditions, the first to use the term *equilibrium* to describe the traffic pattern is perhaps the economist F. H. Knight [Kni24]. The early historical development of the equilibrium concept within the transportation field has been traced in [Boy04, Pat06].

The presentation in Section 2.1–2.8 follows that of the forthcoming book by Patriksson [Pat06]. The modeling development differs from the traditional one (e.g., [AaM81, FeP97, FaP03a]) in which the elastic demand Wardrop conditions are transformed into a nonlinear complementarity problem (NCP). The reason why we adopted the *mixed* complementarity approach is that it is the most natural, and that it possesses stronger properties: in the case of the NCP model, the equivalence with the Wardrop conditions requires that travel cost function be non-negative everywhere—this is associated with having to impose the requirement that $\boldsymbol{\pi} \geq \mathbf{0}^{|\mathcal{C}|}$ holds. In the MCP model, such a condition is not needed. For a general introduction to NCP and MCP problems, see [FeP97, FaP03a].

The fixed demand reformulation and Proposition 1, presented in Section 2.4, stem from [Pat06], and extend the work of Gartner [Gar80b] on problems with a separable costs structure. Sheffi [She85, Section 6.3] remarks that the selection of a good upper bound $\bar{\boldsymbol{g}}$ on the demand function is crucial to the performance of the fixed-demand algorithm, and proposes means to iteratively update it during the computational process, when a linearization (Frank–Wolfe) strategy is adopted.

The Flow Decomposition Theorem 2 can be found, for example, in [BuS65, Theorems 3.7 and 7.2], [Roc84, Sections 4.B and 4.J], or [AMO93, Theorem 3.5]. The linear program (13) which can be used to decompose a link flow into the sum of a strongly acyclic flow and a circulation flow is due to Gallager [Gal77]. Proposition 3 in Section 2.5 stems from [HaT98]; its extension to the case of elastic demands is immediate and is found in [Pat06].

The existence results of Section 2.7 (taken from [Pat06]) are new. This is surprising, given that variational inequality models of traffic equilibrium problems have been around since the late 1970s (e.g., [Smi79, Daf80]) and the optimization models in Section 2.6, corresponding to problems with separable costs and demands, have been around since the 1950s ([BMW56]). The improvements have been made possible for two reasons; for the case of the link–route representation, we have already remarked that the cost function need not be non-negative. For the link–node representation, we have been able to utilize the recent results on the acyclic nature of equilibria, stated in Proposition 3. Cost additivity was assumed in our presentation for simplicity only.

Some of the uniqueness results (taken from [Pat06]) are likewise new, in particular regarding costs and demands. The classical result is that $\boldsymbol{t}$ monotone, $-\boldsymbol{g}$ strictly monotone $\Longrightarrow \boldsymbol{\pi}, \boldsymbol{d}$ unique (e.g., [Flo79, AaM81]). That the cost mapping of a variational inequality is constant over the solution set whenever the mapping is monotone$^+$ (see [CMZ00]) is used to good effect, as well as the result that "$\boldsymbol{f}$ monotone$^+$" coincides with "$\boldsymbol{f}$ monotone" whenever $\boldsymbol{f}$ is a gradient.

The content of Section 2.8 is taken from [Pat06]. It is clear that under continuity and some form of monotonicity of the travel cost function, all definitions collapse to the traditional one. Proposition 8 is collected partly from the following sources: (a) For the first result, see [Smi84, Theorem 1]. See also

[Daf71, Theorem 2.6] and [Daf82] (the latter for the elastic demand case). For the second result, see [Hey86, Theorem 7]. For the third result, see [Daf71, Theorem 2.6] and [Hey86]. (b) That equilibrated flows are Wardrop equilibria is established in [Hey86, Theorem 5]. Our result improves upon it by reducing the requirements to pseudo-monotonicity. (c) See [BeS94, Theorem 2.1]. (e) See [dPN98, Theorem 4(i)]; however, their result is weaker in that monotonicity is assumed.

In Section 2.9, the worst-case behavior of Wardrop equilibria, with respect to a system-optimal flow pattern, has been analyzed in [RoT02, ChS03]. We remark that the result that user and system optimal solutions coincide for the exponent $\mu = 0$ was established first by Jorgensen [Jor63], while the result for $\mu \geq 1$ is due to Dafermos and Sparrow [DaS69, eq. (1.34)]; see also Bennett [Ben93].

3 Variations

3.1 Multi-mode and Multi-attribute models

A model of traffic equilibrium where user behavior is captured by a single function, either route-based or link-based, is a mathematical idealization of real life. It fails to account for travel time (mis)perceptions and for the fact that delay is but one attribute of travel, together with cost, driving stress, landscape, number of stops, vehicle type, etc. The first issue, mis-perception, is best dealt with by using stochastic models (logit, mixed logit, probit) while the second issue yields multi-mode or multi-attribute[12] models. In multi-mode models, distinct cost functions $\boldsymbol{t}_m$ are assigned to distinct modes $m \in \mathcal{M}$. One possible functional form is

$$\boldsymbol{t}_m(\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{|\mathcal{M}|}) = \boldsymbol{t}_m\left(\sum_{i=1}^{|\mathcal{M}|} \alpha_i \boldsymbol{v}^i\right), \tag{25}$$

where $\boldsymbol{v}^m$ denotes the flow of mode m and α_i may be interpreted as the car-equivalent associated with mode m. An equilibrium is then characterized as a solution to the variational inequality

$$-\boldsymbol{t}_m(\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{|\mathcal{M}|}) \in N_{F_m}(\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{|\mathcal{M}|}), \qquad m \in \mathcal{M}, \tag{26}$$

where F_m denotes the set of demand-feasible flows for mode m. The theoretical drawbacks of such a functional form have been pointed out by some researchers, who suggested more complex but better behaved functional forms. Multi-mode models can easily be converted to the standard model by assigning to each mode its own copy of the transportation network, with a cost function that becomes non-separable, and most likely asymmetric.

Closely related to multi-mode models are multi-attribute (or multi-class) models, where the disutility (generalized cost) of each user is a function of route attributes, and the perception of disutility associated with each attribute varies across the population. To fix ideas, let us consider a model with two attributes, namely travel time $\boldsymbol{t}$ and travel cost $\boldsymbol{f}$, the latter including both out-of-pocket cost (tolls, for example) and variable costs such as petrol or maintenance. Let us adopt monetary cost as numéraire, define α_g as the value of one time unit (VOT) associated with population group $g \in \mathcal{G}$, and make the hypothesis that the disutility of each class g assumes the linear form

$$u_g(\boldsymbol{v}) = \alpha_g \boldsymbol{t}(\boldsymbol{v}) + \boldsymbol{f}(\boldsymbol{v}). \tag{27}$$

where $\boldsymbol{v} = (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{|\mathcal{G}|})$. Let us denote by d_k^g the set of demand-feasible flow for population group g and origin-destination pair k, and by F_g the corresponding set of demand-feasible link flows. Due to the separability of the feasible sets, the equilibrium is then characterized by the variational inequalities

$$-u_g(\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{|\mathcal{G}|}) \in N_{F_g}(\boldsymbol{v}^g), \quad g \in \mathcal{G}. \tag{28}$$

Although the above model is subsumed, from the mathematical point of view, by the variational inequality (26), its analysis yields insight into models involving distinct classes of vehicles or customers, and is well worth investigating for its own sake. As an introductory example, let us consider a two-group, one-link example, with cost mapping

$$\boldsymbol{u}(v^1, v^2) = \begin{pmatrix} \alpha_1 t(v^1 + v^2) + f(v^1 + v^2) \\ \alpha_2 t(v^1 + v^2) + f(v^1 + v^2) \end{pmatrix} \tag{29}$$

[12] We refrain from using the term 'multi-criterion' that usually refers to vector optimization or vector equilibrium problems.

and with Jacobian matrix

$$\nabla \boldsymbol{u}(v^1, v^2) = \begin{pmatrix} \alpha_1 t'(v^1+v^2) + f'(v^1+v^2) & \alpha_1 t'(v^1+v^2) + f'(v^1+v^2) \\ \alpha_2 t'(v^1+v^2) + f'(v^1+v^2) & \alpha_2 t'(v^1+v^2) + f'(v^1+v^2) \end{pmatrix}. \tag{30}$$

It is instructive to check the monotonicity of the mapping $\boldsymbol{u}$. To this aim, we compute the determinant of the symmetrized Jacobian matrix:

$$\begin{aligned} \det(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^{\mathrm{T}}) &= \begin{vmatrix} 2\alpha_1 t'(v^1+v^2) + 2f'(v^1+v^2) & (\alpha_1+\alpha_2)t'(v^1+v^2) + 2f'(v^1+v^2) \\ (\alpha_1+\alpha_2)t'(v^1+v^2) + 2f'(v^1+v^2) & 2\alpha_2 t'(v^1+v^2) + 2f'(v^1+v^2) \end{vmatrix} \\ &= [4\alpha_1\alpha_2 - (\alpha_1+\alpha_2)^2](t'(v^1+v^2))^2 \quad = \quad -(\alpha_1-\alpha_2)^2(t'(v^1+v^2))^2. \end{aligned}$$

This calculation shows that $\boldsymbol{u}$ can only be monotone if either $\alpha_1 = \alpha_2$ or the cost function t is constant. In the first case, the problem reduces to the single-group case while, if t is constant, congestion is absent. Now, since the feasible domain is separable by group, one may scale each group cost mapping by an arbitrary positive number, to derive an equivalent formulation of the equilibrium conditions. For instance, if the parameters α_1 and α_2 are not zero, we can divide the cost functions by their respective VOT parameter, yielding the equivalent cost mapping, which we still denote by $\boldsymbol{u}$:

$$\boldsymbol{u}(v^1, v^2) = \begin{pmatrix} t(v^1+v^2) + \alpha_1 f(v^1+v^2) \\ t(v^1+v^2) + \alpha_2 f(v^1+v^2) \end{pmatrix}, \tag{31}$$

and where α_g represents the inverse VOT of group g. Although both formulations are equivalent from the equilibrium viewpoint, their mathematical properties are quite different. Indeed, repeating the previous analysis, we obtain that $\boldsymbol{u}$ is monotone if f (rather than t previously) is constant. This will be satisfied if tolls and variable travel costs are flow-independent, a reasonable first-approximation assumption. For this reason we will, from now on, switch to *travel time* as numéraire. Note that the monotonicity results extend to the general multi-criteria case, i.e., a sufficient condition that the cost mapping $\boldsymbol{u}$ be monotone is that all cost functions, with the possible exception of travel time, be flow independent. We shall see, in a more general setting, that this condition is necessary as well.

A natural question that arises concerning the variational inequality involving the mapping $\boldsymbol{u}$ is the existence of a 'natural' optimization problem whose set of stationary (Karush–Kuhn–Tucker) points coincides with the set of multi-attribute equilibria. In other words: under what conditions is the mapping $\boldsymbol{u}$ a gradient mapping? Let $\mathcal{C}$ denote the index set of attributes and $\overline{\boldsymbol{v}} = \sum_{i\in\mathcal{C}} \boldsymbol{v}^i$ the total link flow vector. The answer to our question is that $\boldsymbol{f}$ be constant and $\boldsymbol{t}$ be a gradient mapping, and the relevant optimization program is simply to

$$\underset{\boldsymbol{v}}{\text{minimize}} \oint_0^{\overline{\boldsymbol{v}}} \boldsymbol{t}(s)\, ds + \sum_{i\in\mathcal{C}} \alpha_i \boldsymbol{f}^{\mathrm{T}} \boldsymbol{v}^i, \tag{32}$$

where $\mathcal{C}$ denotes the index set of criteria. Indeed, it is easy to check that the optimality conditions of (32) fulfill the multi-attribute equilibrium conditions. Furthermore, if the mapping $\boldsymbol{t}$ is monotone, the program (32) is convex. If $\boldsymbol{t}$ is strictly monotone, the total flow equilibrium is unique, although group and origin-destination flows will not in general be unique.

We now focus our attention on a two-attribute model where the VOT (or inverse VOT) parameter is continuously distributed across the population, and described by a continuous probability density function h defined over the nonnegative axis. For ease of notation, we assume that the density h is identical for all origin-destination pairs. For each origin-destination pair k, we then have that the demand density for the α-group[13] is equal to $h(\alpha)d_k$, and $F_\alpha = h(\alpha)F$ is the feasible set for the α-group. The variables of the equilibrium problem are regrouped into a link flow density vector $\boldsymbol{v}(\alpha) = \{v_l(\alpha)\}_{l\in\mathcal{L}}$ and, similar to the finite-dimensional case, we denote by $\overline{\boldsymbol{v}}$ the total link flow vector which, in this case, is defined by the integral

$$\overline{\boldsymbol{v}} = \int_0^\infty \boldsymbol{v}(\alpha)\, d\alpha. \tag{33}$$

[13] Although the terminology α-group is convenient, the notion actually relates to an infinitesimal user.

Replacing the index g by α yields the equilibrium conditions:

$$-\left(\boldsymbol{t}(\overline{\boldsymbol{v}})+\alpha \boldsymbol{f}(\overline{\boldsymbol{v}})\right) \in N_{F_\alpha}(\boldsymbol{v}(\alpha)), \quad \alpha \in [0,\infty). \tag{34}$$

The above formulation involves an infinite collection of variational inequalities, one for each α-group. Upon introduction of the vectorial disutility function $\boldsymbol{u}(\overline{\boldsymbol{v}})$ defined as

$$[\boldsymbol{u}(\overline{\boldsymbol{v}})](\alpha)=\boldsymbol{t}(\overline{\boldsymbol{v}})+\alpha \boldsymbol{f}(\overline{\boldsymbol{v}}), \tag{35}$$

the vector function $\boldsymbol{v}=\boldsymbol{v}(\alpha)_{\alpha\geq 0}$ and the set $F=\prod_\alpha F_\alpha$, the equilibrium conditions (34) can be aggregated into the infinite-dimensional variational inequality

$$-\boldsymbol{u}(\overline{\boldsymbol{v}}) \in N_F(\boldsymbol{v}). \tag{36}$$

While this may be considered elegant from the mathematical point of view, the approach has drawbacks. First, the proper setting for such formulation is the set of square-integrable functions. Since two functions that differ over a set of measure zero are equivalent, standard existence results cannot be invoked to prove the existence of a solution to the system (35) for *every* value of the parameter α. Second, this framework hides the network structure of the problem, which is essentially discrete.

Both these drawbacks can be remedied by considering finite-dimensional formulations. The remainder of the section is devoted to this topic, while algorithms will be discussed in Section 5.1. To ease the presentation, we assume from now on that F denotes the unit simplex $F=\{\,\overline{\boldsymbol{v}}_i \geq 0 \mid \sum_{i=1}^n \overline{\boldsymbol{v}}_i=1\,\}$ and that $h(\alpha)=0$ for all α larger than some finite threshold $\overline{\alpha}$. Since the original set F is a bounded polyhedron, the simplicial form of F can be achieved by expressing each of the points of F as a convex combination of its vertices. Alternatively, one may consider a traffic equilibrium problem involving one origin, one destination and a unit demand. For this reason, we will use, without any loss of generality, the term 'path' to denote an extreme point of the unit simplex.

Our finite-dimensional approach is based on the fixed point formulation of a variational inequality. In our setting, this takes the form:

$$v(\alpha) \quad \in \quad h(\alpha)F \,\cap\, \arg \underset{\boldsymbol{w}(\alpha)\in F_\alpha}{\operatorname{minimum}} \left(\boldsymbol{t}(\overline{\boldsymbol{v}})+\alpha \boldsymbol{f}(\overline{\boldsymbol{v}})\right)^{\mathrm{T}} \boldsymbol{w}(\alpha), \quad \alpha \in [0,\infty). \tag{37}$$

A key observation is that the right-hand-side of (37) is equivalent to the set of linear programs

$$\underset{\boldsymbol{w}\in F}{\operatorname{minimize}} \left(\boldsymbol{t}(\overline{\boldsymbol{v}})+\alpha \boldsymbol{f}(\overline{\boldsymbol{v}})\right)^{\mathrm{T}} \boldsymbol{w}(\alpha), \quad \alpha \in [0,\infty). \tag{38}$$

This is nothing else than a parametric linear program, where the parameter α scans the nonnegative real axis. Because the number of extreme points of F, that is, combinations of shortest paths, is finite, this infinite-dimensional program can actually be solved in *finite* time by the parametric simplex algorithm. To this aim, let us sort the paths in decreasing order of their slopes. As the valuation of time decreases with α, the cost-conscious users will use high-index paths, while the time-conscious users will travel on low-index paths. The solution to the parametric linear program is characterized by a vector $\boldsymbol{\alpha}=(\alpha_0=0,\alpha_1,\ldots,\alpha_n=\overline{\alpha})$ of critical values, where users having an inverse VOT α belonging to the interval (α_{i-1},α_i) are assigned to path i (see Figure 3.1). Several comments are in order:

(i) The flow on path i is given by the integral

$$\overline{\boldsymbol{w}}(\overline{\boldsymbol{v}})=\int_{\alpha_{i-1}}^{\alpha_i} h(\alpha)\,d\alpha. \tag{39}$$

If the total flow vector $T(\overline{\boldsymbol{v}})=(T_i(\overline{\boldsymbol{v}}))_{i=1}^n$ agrees with $\overline{\boldsymbol{v}}$, then the solution to the parametric LP is an equilibrium for the bi-attribute problem. Since this solution only depends on $\overline{\boldsymbol{v}}$, it can be recovered from the solution to the finite-dimensional fixed-point problem $\overline{\boldsymbol{w}}(\overline{\boldsymbol{v}}) \in \overline{\boldsymbol{v}}$. Alternatively, a finite-dimensional fixed point formulation can be built around the vector of critical points $\boldsymbol{\alpha}$. This amounts to initializing with $\boldsymbol{\alpha}$ the sequence of evaluations

$$\text{total flow } \overline{\boldsymbol{v}} \rightarrow \text{ critical vector } \boldsymbol{\alpha} \rightarrow \text{ (through integration) } \overline{\boldsymbol{w}}.$$

(ii) The assignment of users whose inverse VOT is equal to one of the critical points $(\alpha_1,\ldots,\alpha_{n-1})$ is ambiguous. Generically, the number of such points is finite, i.e., of zero Lebesgue measure.

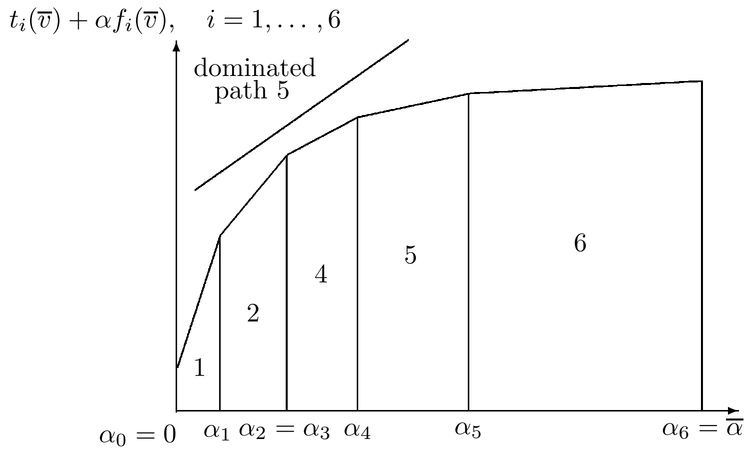


Figure 5: Parametric path costs. Path 3 is dominated and therefore carries no flow.

(iii) In the degenerate situation where two or more paths have identical $\boldsymbol{t}$ and $\boldsymbol{f}$ values over an interval (α_{i-1}, α_i), users may be assigned to any of these paths, a situation similar to that occurring in the standard, single-attribute model.

(iv) If all slopes are distinct, the path assignment is unambiguous, with the exception of at most $n-1$ values of the parameter α, and the solution of the parametric LP is unique, almost everywhere. This regularity condition, which is easily achieved by perturbing the f-value of each path (or link), allows for a finite-dimensional variational inequality formulation and has beneficial algorithmic implications.

We now consider four important issues related to any fixed point or variational formulation, namely: existence and uniqueness of solutions, monotonicity and the gradient property. In the infinite-dimensional setting, existence of a solution can be proved using measure-theoretic arguments. In either finite-dimensional fixed point formulation, one can actually show that the fixed point mapping $\overline{\boldsymbol{w}}$ is upper semi-continuous whenever the ordering of path slopes remain constant, and existence follows from Kakutani's theorem. As in the discrete case, monotonicity holds if and only if $\boldsymbol{t}$ is monotone and $\boldsymbol{f}$ is constant. Similarly, the gradient property is satisfied if and only if $\boldsymbol{t}$ is a gradient mapping (not necessarily monotone) and, again, $\boldsymbol{f}$ is constant. The equivalent convex optimization problem is to

$$\underset{\overline{\boldsymbol{v}} \in F}{\text{minimize}} \int_{\boldsymbol{0}}^{\overline{\boldsymbol{v}}} \boldsymbol{t}(s)\, ds + \sum_{i=1}^{n} \Big[(f_i - f_{i+1}) \nu \Big(\sum_{j=1}^{i} \overline{\boldsymbol{v}}_j \Big) \Big], \tag{40}$$

where $\nu = \int \phi^{-1}$ and ϕ denotes the cumulative distribution function ϕ of the density h.

The case for uniqueness deserves more attention. If $\boldsymbol{t}$ is constant and $\boldsymbol{f}$ is strictly monotone, the (total flow) solution $\overline{\boldsymbol{v}}$ to is unique. Moreover, if path slopes are distinct, the solution to the parametric LP (37) is unique, except for the critical points α_i. It follows that path flows are unique, a surprising result. Indeed, the Jacobian matrix $\nabla \boldsymbol{u}(\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots, \boldsymbol{v}^{|\mathcal{G}|})$ of the disutility function $\boldsymbol{u}$ has rank one and is therefore 'less and less' definite as the number of user groups increases, making futile the standard argument for establishing uniqueness. Although the situation might be expected to worsen in the infinite-dimensional case, exactly the opposite occurs. This is reminiscent of stochastic assignment models, for which path uniqueness can be established under weak regularity assumptions.

3.2 Side constrained traffic equilibrium models

3.2.1 Introduction and equilibrium characterizations

Traffic flows are frequently subject to restrictions, physical or otherwise, that have not been taken account in the basic model. For example, in the presence of a stationary traffic control system, flows might have to obey constraints of the form

$$s_k(\boldsymbol{v}) \leq 0, \qquad k \in \mathcal{K}. \tag{41}$$

where, for mathematical convenience, we assume that the functions $s_k : \Re^{|\mathcal{L}|}_+ \to \Re$, $k \in \mathcal{K}$, are convex and differentiable. Simple flow capacities on some links, describing the stationary effect of the traffic control, correspond to letting $\mathcal{K} := \bar{\mathcal{L}} \subseteq \mathcal{L}$, and

$$s_l(\boldsymbol{v}) := v_l - c_l, \qquad l \in \bar{\mathcal{L}},$$

where $c_l > 0$ is the (stationary) link volume capacity. Side constraints can also be used to force equilibrium flows to comply with traffic management goals. For instance, system optimal link flow patterns can be enforced by setting c_l to the elements of a system optimal link flow vector.

Unfortuntaly, under the presence of joint constraints, the Cartesian product structure of the feasible sets of, for example, (15) and (16), are not satisfied any more. This has rather far reaching consequences on the equilibrium characterization of optimal flows. An equilibrium, in the sense of Wardrop, may no longer exist in terms of the original cost structure. If, however, one incorporates into the cost the Lagrange multipliers associated with the side constraints (41), then on can describe an equilibrium in terms of a generalized cost which, under some circumstances, can be given a natural interpretation. We introduce multipliers $\beta_k \geq 0$ for each constraint in (41) append to (2) the additional condition

$$\mathbf{0}^{|\mathcal{K}|} \leq \boldsymbol{\beta} \perp -\boldsymbol{s}(\boldsymbol{v}) \geq \mathbf{0}^{|\mathcal{K}|}. \tag{42}$$

Letting $S := \{\, \boldsymbol{v} \in \Re^{|\mathcal{L}|} \mid s_k(\boldsymbol{v}) \leq 0, \quad k \in \mathcal{K} \,\}$ the side constrained inelastic demand model becomes

$$-\boldsymbol{t}(\boldsymbol{v}) \in N_{F\cap S}(\boldsymbol{v}) = N_F(\boldsymbol{v}) + N_S(\boldsymbol{v}) = N_F(\boldsymbol{v}) + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \in N_S(\boldsymbol{v});$$

if S satisfies a constraint qualification, for instance if each function s_k is affine, then for some vector $\boldsymbol{\beta}$ satisfying (42), we have that, at equilibrium,

$$-\left[\boldsymbol{t}(\boldsymbol{v}) + \sum_{k\in\mathcal{K}} \beta_k \nabla s_k(\boldsymbol{v})\right] \in N_F(\boldsymbol{v})$$

holds. Hence, for any optimal vector of Lagrange multipliers, $\boldsymbol{v}$ is a Wardrop equilibrium volume in terms of the generalized link costs $\bar{\boldsymbol{t}}(\boldsymbol{v}) := \boldsymbol{t}(\boldsymbol{v}) + \sum_{k\in\mathcal{K}} \beta_k \nabla s_k(\boldsymbol{v})$. In the case of link capacities, this reduces to $\bar{\boldsymbol{t}}(\boldsymbol{v}) := \boldsymbol{t}(\boldsymbol{v}) + \boldsymbol{\beta}$, where the elements of $\boldsymbol{\beta}$ for $l \notin \bar{\mathcal{L}}$ are fixed to 0. This expression may be interpreted as a generalized delay, where $\boldsymbol{\beta}$ is a vector of *queueing* delays at intersections. While this interpretation may be valid, it is important to note that the vector $\boldsymbol{\beta}$ of Lagrange multipliers is seldom unique, even in this simple case. This implies that link queueing delays are not unique, even if link flows $\boldsymbol{v}$ are. The following example illustrates how this situation can be utilized to improve traffic conditions.

We consider the directed graph illustrated in Figure 6, that involves two OD pairs, $(1,5)$ and $(2,5)$, with a demand of 2 and 3, respectively. Travel cost functions and link capacities are given in Table 1. The solution to the traffic equilibrium problem (that is, in the absence of upper bounds) is $\boldsymbol{v} = (2.00, 2.09, 0.91, 1.27, 2.82, 2.18)^{\mathrm{T}}$, while the solution to the capacitated model is $\boldsymbol{v}^* = (2,2,1,2,2,3)^{\mathrm{T}}$.

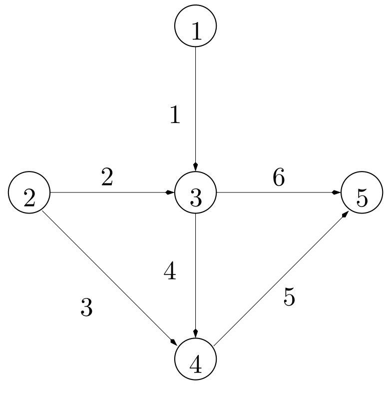


Figure 6: A capacitated traffic network.

link	$t_l(v_l)$	c_l
1: $(1,3)$	$t_1(v_1) = v_1$	2
2: $(2,3)$	$t_2(v_2) = 4v_2$	2
3: $(2,4)$	$t_3(v_3) = 12v_3$	1
4: $(3,4)$	$t_4(v_4) = 2v_4$	2
5: $(3,5)$	$t_5(v_5) = 4v_5$	2
6: $(4,5)$	$t_6(v_6) = 4v_6$	4

Table 1: Network data.

Applying an augmented Lagrangean algorithm (see further Section 5.2) to this problem produces the multiplier vector $\boldsymbol{\beta} := (0, 0.219, 0.219, 0, 8.0, 0)^{\mathrm{T}}$. This is, however, far from the only multiplier vector in this problem. We first note that the set of multiplier vectors $\boldsymbol{\beta}$ for this problem satisfies

$$\begin{array}{ccccccccrr}
\beta_1 & & & +\beta_4 & & +\beta_6 & -\pi_{15} & & \geq & -18, \\
\beta_1 & & & & +\beta_5 & & -\pi_{15} & & \geq & -10, \\
 & \beta_2 & & +\beta_4 & & +\beta_6 & & -\pi_{25} & \geq & -24, \\
 & \beta_2 & & & +\beta_5 & & & -\pi_{25} & \geq & -16, \\
 & & \beta_3 & & & +\beta_6 & & -\pi_{25} & \geq & -24, \\
2\beta_1 & +2\beta_2 & +\beta_3 & +2\beta_4 & +2\beta_5 & +3\beta_6 & -2\pi_{15} & -3\pi_{25} & = & -92, \\
\beta_1, & \beta_2, & \beta_3, & \beta_4, & \beta_5, & \beta_6, & & & \geq & 0, \\
 & & & & & \beta_6 & & & = & 0.
\end{array}$$

Clearly, this set is not a singleton, and is even unbounded. We see that link 6 is uncapacitated at $\boldsymbol{v}^*$, whence its multiplier must be zero by complementarity, while the remaining link multipliers are non-negative, by definition. If the model is valid, it is then interesting to note that we may interpret the multipliers β_l as link tolls. These ensure that the uncapacitated traffic equilibrium problem with link cost $\boldsymbol{t}(\cdot) + \boldsymbol{\beta}$ has the same solution $\boldsymbol{v}^*$ as the capacitated model with the original link travel costs. Actually, the set of multipliers shown above is only a subset of interesting choices. It may even turn out to be beneficial to impose *negative* tolls, or to set tolls on links that are *not* saturated, in order to achieve some preset management target. This possibility brings forward an alternative perspective on how to induce a favored equilibrium solution.

Suppose that we wish to impose upon the users link tolls (be they positive or negative) such that a link flow $\boldsymbol{v}^* \in \hat{F}$ is in equilibrium. Then, if one interprets the side constraints as an enforcement of the equality $\boldsymbol{v} = \boldsymbol{v}^*$ (that is, $s_l(\boldsymbol{v}) = v_l - v_l^*$, $l \in \mathcal{L}$, and $s_{|\mathcal{L}|+l}(\boldsymbol{v}) = v_l^* - v_l$, $l \in \mathcal{L}$), one observes that the set of link tolls corresponds to the polyhedron defined by the solutions in $\boldsymbol{\beta}$ of the linear system

$$\boldsymbol{\Lambda}^{\mathrm{T}}(\boldsymbol{t}(\boldsymbol{v}^*) + \boldsymbol{\beta}) \geq \boldsymbol{\Gamma}\boldsymbol{\pi}, \tag{43a}$$

$$(\boldsymbol{t}(\boldsymbol{v}^*) + \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{v}^* = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{\pi}. \tag{43b}$$

In the above example, this set corresponds to removal of the last two conditions, i.e., non-negativity and complementarity.

Using the previous polyhedron as the feasible set of a mathematical program, we may, for example, devise a minimum-revenue toll by minimizing the revenue function $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{v}^*$, whose minimal value 16 is achieved for the toll vector $\boldsymbol{\beta}^* = (0,0,0,0,8,0)^{\mathrm{T}}$ (compare with the value 16.657 obtained from the augmented Lagrangean solution mentioned above) and $\boldsymbol{\pi} = (18, 24)^{\mathrm{T}}$. Obviously, there does not exist a maximum-revenue toll schedule, since the revenue can be made arbitrarily large along the line $\boldsymbol{\beta} = (0, \gamma, \gamma, 0, 8, 0)^{\mathrm{T}}$, $\gamma \geq 0$. (In the case of the set (43), we can let $\gamma \to -\infty$ to prove that no minimum-revenue toll exists.) Such toll problem will be investigated in more detail in Section 6.2.

In the elastic demand case, the toll polyhedron that induces given link flow and demand patterns $(\boldsymbol{d}^*, \boldsymbol{v}^*)$ is

$$\boldsymbol{\Lambda}^{\mathrm{T}}(\boldsymbol{t}(\boldsymbol{v}^*) + \boldsymbol{\beta}) \geq \boldsymbol{\Gamma}\boldsymbol{\pi}, \tag{44a}$$

$$(\boldsymbol{t}(\boldsymbol{v}^*) + \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{v}^* = \boldsymbol{\xi}(\boldsymbol{d}^*)^{\mathrm{T}}\boldsymbol{d}^*, \tag{44b}$$

that is, the immediate extension to the system (43). We note that the value of $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{v}^*$ is *constant* over this set, i.e., the toll revenue is constant! This paradox can be explained by the fact that network users might *not* travel when the travel cost becomes too large.

3.2.2 Wardrop-like principles

According to the above, side constrained equilibria satisfy the Wardrop conditions in terms of *generalized* travel costs, but they will, in general, not satisfy any similar conditions in terms of *actual* travel costs. One can therefore, in general, not relate the actual travel costs of the unused routes to those of the used ones; for example, the least costly route in an OD pair may be unused because its generalized cost is too high. This deficiency is due to the fact that the side constrained problem does, in contrast to the standard model, lacks a Cartesian product structure.[14]

Wardrop-like principles in terms of actual travel costs may however be established if the following assumption is fulfilled at the solution to the side constrained model:

ASSUMPTION 12 (nondecreasing side constraint functions) *At the flow $\boldsymbol{v} \in \hat{F}$,*

$$\frac{\partial s_k(\boldsymbol{v})}{\partial v_l} \geq 0, \qquad l \in \mathcal{L}, \quad k \in \mathcal{K},$$

holds. □

If this assumption holds for any flow $\boldsymbol{v} \in \hat{F}$, then a flow increase along one or more links of the network can never result in any side constraint becoming 'looser'. Conversely, this assumption holds whenever the side constraints correspond to general capacity restrictions, that is, when they represent upper bounds on traffic volumes on certain links or routes, or within an area of the traffic system. In the sequel, we shall use the notions of links and routes that are *unsaturated* with respect to the side constraints.

DEFINITION 13 (unsaturated link and route) *A link $l \in \mathcal{L}$ is* unsaturated *at the flow $\boldsymbol{v} \in \hat{F}$ if for all $k \in \mathcal{K}$,*

$$\frac{\partial s_k(\boldsymbol{v})}{\partial v_l} > 0 \Longrightarrow s_k(\boldsymbol{v}) < 0.$$

A route $r \in \mathcal{R}$ is unsaturated *at the flow $\boldsymbol{v} \in \hat{F}$ if all links $l \in \mathcal{L}$ on route r are unsaturated.* □

A route is clearly *saturated* at the flow $\boldsymbol{v} \in \hat{F}$ if $\frac{\partial s_k(\boldsymbol{v})}{\partial v_l} > 0$ holds for some link l on the route and some $k \in \mathcal{K}$ such that $s_k(\boldsymbol{v}) = 0$.

THEOREM 14 (Wardrop-type principles) *Suppose that $(\boldsymbol{h}^*, \boldsymbol{v}^*)$ solves the side constrained model and that Assumption 12 holds at $\boldsymbol{v}^*$. Then, the following conclusions hold for any OD pair $(p, q) \in \mathcal{C}$.*

(a) *The routes utilized in the OD pair have equal and minimal generalized route costs.*

(b) *Assume, without loss of generality, that the first ℓ routes are utilized in the OD pair and that m of these are unsaturated. Then, the routes may be ordered so that*

$$c_1 = c_2 = \ldots = c_m \geq c_{m+1} \geq c_{m+2} \geq \ldots \geq c_\ell.$$

(c) *For any pair of routes $r, s \in \mathcal{R}_{pq}$,*

$$\left.\begin{array}{c}\text{route } r \text{ is unsaturated} \\ c_s > c_r\end{array}\right\} \quad \Longrightarrow \quad h_s^* = 0.$$

(d) *For any pair of routes $r, s \in \mathcal{R}_{pq}$,*

$$\left.\begin{array}{c}\text{route } r \text{ is utilized} \\ c_s < c_r\end{array}\right\} \quad \Longrightarrow \quad \text{route } s \text{ is saturated.}$$

The (simple) proof of the above results is based upont the equilibrium characterization of a side constrained equilibrium.

If the implication in either of the results (c) and (d) were not fulfilled for some pair of routes, then some traveller might shift to a less costly and unsaturated alternative route; hence, these results are quite natural. As touched upon above, the OD routes that are unused in a solution to the side constrained model are not necessarily more costly (in actual travel cost) than those used in the OD pair; this is implied by the result (d) since a route may be saturated at zero flow.

[14]For capacitated networks, an alternative approach is analyzed in Section 3.3.

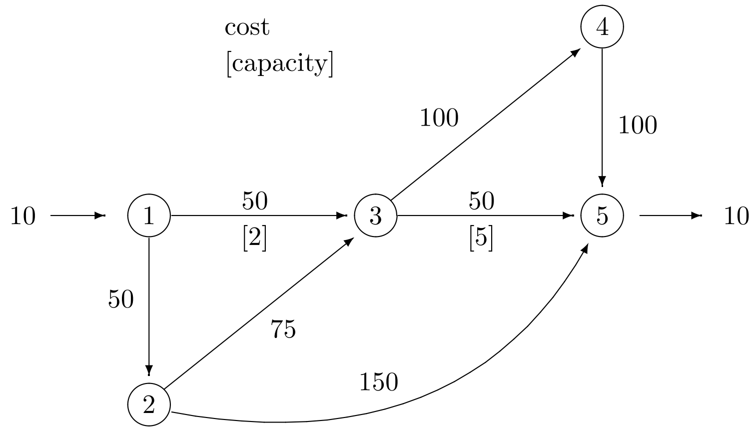


Figure 7: A capacitated network

3.3 Strategic models

The standard traffic equilibrium models assume that path selection, performed at the origin nodes, does not change en route. This assumption is reasonable as a first approximation, but is not entirely compatible with user behavior. In this section, we outline an approach that allows users to take on-line decisions, based on current traffic conditions. While such an approach is common place in transit systems, where users may need to transfer one or more times to reach their destination, its application to congested traffic assignment is recent.

Let us first consider a transit system where, at each stop served by several transit lines, waiting customers must make strategic choices of the form: Should one board the incoming vehicle, or wait for another, more attractive (quicker) transit line? Such decision must balance the travel time of the incoming vehicle against travel times *plus* waiting times of vehicles yet to reach the boarding station. Let us assume that the transit stop is served by two lines with respective frequencies $\phi_1 < \phi_2$, and travel times to destination $t_1 < t_2$. Assuming that the arrival process is random and memoryless (Poisson process), and that users are rational and risk-neutral, they will naturally board a vehicle of line 1 if it shows up first, and a vehicle of line 2 if the expected waiting time of line 1, i.e., $1/\phi_1$, exceeds the difference $t_2 - t_1$, i.e.,

$$t_2 \leq 1/\phi_1 + t_1.$$

This reasoning can be generalized to n common bus lines indexed in decreasing order of their travel times to destination. The optimal strategy is characterized by a threshold index $\bar{l}$, below which all transit lines are attractive, that satisfies the relation

$$\bar{l} \in \arg \operatorname*{minimum}_{1 \leq l \leq n} \frac{1}{\sum_{i=1}^{l} \phi_i} \left(1 + \sum_{i=1}^{l} \phi_i t_i \right). \tag{45}$$

The key feature of this approach is that, while strategies are *deterministic* objects, the route traveled from day to day by users is *stochastic*. The applicability of the method relies on the existence of efficient algorithms for computing optimal strategies. Such algorithms, akin to shortest path methods, are able to address large-scale problems.

In private transportation, the situation is different, since randomness is associated with demand (users) rather than supply (vehicles). Defining equilibrium meaningfully in this context represents a nontrivial task. To gain some insight into the situation, let us consider the network illustrated in Figure 7, where each link is endowed with a cost (shown next to the corresponding link) and, possibly, a capacity (bracketed number). Paths from origin node 1 to destination node 5 are listed, together with their features, in Table 2.

Clearly, no Wardrop equilibrium is compatible with this data, since the shortest path 1-3-5 cannot accommodate all demand of 10 units. Quite naturally, one could settle for an equilibrium satisfying the following Wardrop-like principle:

path index	path	cost	capacity
1	$1-3-5$	100	2
2	$1-2-3-5$	175	5
3	$1-2-5$	200	∞
4	$1-3-4-5$	250	2
5	$1-2-3-4-5$	325	∞

Table 2: Network paths for a small example

node	1	2	3	4	5
s_1	$[3,2]$	$[3]$	$[5,4]$	$[5]$	$[\ ]$
s_2	$[3,2]$	$[5]$	$[5,4]$	$[5]$	$[\ ]$
s_3	$[2]$	$[3]$	$[5,4]$	$[5]$	$[\ ]$
s_4	$[2]$	$[5]$	$[\ ]$	$[\ ]$	$[\ ]$
s_5	$[3,2]$	$[5,3]$	$[5,4]$	$[5]$	$[\ ]$
s_6	$[3]$	$[\ ]$	$[5,4]$	$[5]$	$[\ ]$
s_7	$[3,2]$	$[3]$	$[4,5]$	$[5]$	$[\ ]$

Table 3: A set of strategies for the small example

> At equilibrium, the cost of a path with a positive residual capacity is larger or equal to the cost of any path carrying positive flow.

In our small example, any path flow vector of the form

$$\boldsymbol{x} = (5-\delta, \delta, 5, 0, 0)^{\mathrm{T}}$$

with $\delta \in [3, 5]$ satisfies the above principle. However, whenever δ is strictly greater than 3 and the users of the second path are endowed with the gift of prescience, it is tempting for them to switch to the less costly first path on which some capacity has been set free by *themselves*. In this sense, the equilibrium associated with the value $\delta = 3$ is more natural. Indeed this solution, associated with a queueing delay of 75 on link (1,3) and 25 on link (3,5), corresponds to the equilibrium notion of Section 3.2. Notwithstanding the queueing delay, it is a system-optimal solution as well.

Let us now consider the strategic approach where a subset of strategies, represented as vectors whose elements consist of an ordered list of outgoing nodes, is provided in Table 3. For instance, when leaving its origin node 1, a user adopting strategy s_1 will prefer going to node 3 (first node in the ordered list) over the alternative choice, that is, going to node 2, provided that link (1,3) is not saturated yet. A similar situation occurs at node 3, where the outgoing link (3,5) is preferred over (3,4). At nodes 2 and 4, however, the strategic choice shrinks to a singleton, due to the fact that the respective preferred outgoing links are uncapacitated. It follows that a user adopting strategy s_1 could end up traveling on path 1, 2, 4 or 5, depending on the availability of the capacitated links (1,3) and (3,5). The strategies that, such as s_4, avoid capacitated links, correspond to ordinary paths in the network.

Whenever the number v_a of users that wish to access link l exceeds its capacity u_l, we assume that the probability p of accessing the link is set to $p = u_l/v_l$. This is equivalent to assuming that users are independently and uniformly distributed at the tail node of link l, in a dimensionless queue. These *access probabilities* (or *access proportions*) allow us to compute the expected cost of strategies. For instance, if all 10 users adopt strategy s_1, the probability of accessing link (1,3) is equal to 2/10. The users, whether they access link (1,3) or not, clash again at node 3, where the access probability of link (3,5) is 5/10. These numbers, which are required to compute the access probabilities associated with the paths of the network, are shown in Table 4.

The expected value of each user's *expected* delay is then equal to the sum of the path costs, weighted by the respective path access probabilities, i.e.,

$$\left(\frac{1}{10} \times 100\right) + \left(\frac{4}{10} \times 175\right) + \left(\frac{0}{10} \times 200\right) + \left(\frac{1}{10} \times 250\right) + \left(\frac{4}{10} \times 325\right) = 235.$$

path	access probability	cost
$1-3-5$	$\frac{2}{10}\times\frac{5}{10}=\frac{1}{10}$	100
$1-2-3-5$	$\frac{8}{10}\times\frac{5}{10}=\frac{4}{10}$	175
$1-2-5$	0	200
$1-3-4-5$	$\frac{2}{10}\times\frac{5}{10}=\frac{1}{10}$	250
$1-2-3-4-5$	$\frac{8}{10}\times\frac{5}{10}=\frac{4}{10}$	325

Table 4: Path access probabilities for the small example

path	flow from s_1	flow from s_2	cost
$1-3-5$	$\frac{5}{6}$	$\frac{5}{6}$	100
$1-2-3-5$	$\frac{20}{6}$	0	175
$1-2-5$	0	4	200
$1-3-4-5$	$\frac{1}{6}$	$\frac{1}{6}$	250
$1-2-3-4-5$	$\frac{4}{6}$	0	325

Table 5: Equilibrium strategic path flows

A *strategic equilibrium* is reached when all users are assigned to strategies of minimal expected delays. This condition is clearly violated if all commuters are assigned to strategy s_1, since strategy s_4, which corresponds to the uncapacitated path $1-2-5$, is available at cost $200 < 235$. It is not too difficult to verify that the unique equilibrium corresponds to the assignment of 5 users to strategy s_1 and of the remaining 5 to strategy s_2; the resulting strategic flows and path flows are shown in Table 5. One checks that the expected delay of each strategy is equal to 185, which is less than the expected delay of unused strategies, thus fulfilling the equilibrium conditions.

In general, a vector $\boldsymbol{x} = \{\boldsymbol{x}_k\}_{k\in\mathcal{C}}$ of *strategic flows* (one for each commodity or OD pair $k \in \mathcal{C}$) is a strategic equilibrium if and only if it is demand-feasible ($\boldsymbol{x}_k$ belongs to the set of feasible strategic flows X_k for every commodity $k \in \mathcal{C}$) and satisfies the variational inequality

$$-\boldsymbol{c}(\boldsymbol{x}) \in N_X(\boldsymbol{x}), \tag{46}$$

where X denotes the Cartesian product of the sets X_k. For a given strategy s, the component $\boldsymbol{c}^s(\boldsymbol{x})$ of $\boldsymbol{c}(\boldsymbol{x})$ represents the expected delay associated with strategy s and the strategic vector $\boldsymbol{x}$. Not that the evaluation of $\boldsymbol{c}$ requires the knowledge of total link flows. Unfortunately, in contrast with the the standard model where this information is readily available from path flows (via the link-route incidence matrix), the situation is different in the strategic model, as the dependence of the link-hyperpath matrix on strategies complicates matters significantly. In particular, the task of determining strategic flows that are compatible with a given link flow vector is an algorithmic challenge by itself, and most likely 'intractable'.

In our introductory example, the link access probabilities were easy to compute. However, the situation becomes complex when the forward star of a node involves more than two links, and when strategies active at that node have different priority orders. In this situation, should one assume that users strictly obey a FIFO (First-In-First-Out) rule, or does one allow a user to jump the queue if its preferred choice is available, as in banks with tellers dedicated to specific services? The distinction is important, since it impacts the assignment of users to outgoing links. In the first case, users are assigned to their preferred node according to their position in the queue. In the second case, users are assigned *simultaneously* to their preferred node, until some residual capacity becomes zero. Throughout the process, users that are denied their current preferred link keep a priority compatible with their arrival instant. Both situations are illustrated in Figure 8 where each square, for ease of understanding, represents an atomic user whose preferred outgoing node is either A or B. The queue is 'virtual' (sometimes denoted 'vertical') in the sense that it occupies no physical space.

At node j, the loading of the flow onto the outgoing links is an iterative process. In the single queue case, it is initiated by constructing the set $\overline{K}$ of first choices. Next, one computes[15]

$$\eta = \operatorname*{minimum}_{k\in\overline{K}} \{\overline{u}_{jk}/d_k, 1\},$$

[15]Here, the index k refers to a node, not an OD pair.

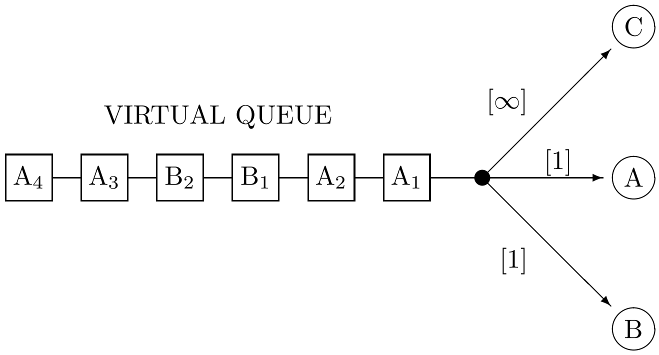


Figure 8: Flow dispersion at a node. In the FIFO case, A_1 is assigned to A, A_2 is assigned to B and all other users are assigned to C. In the second case, A_1 is assigned to A while, in parallel, B_1 is assigned to B; the four remaining users are assigned to C.

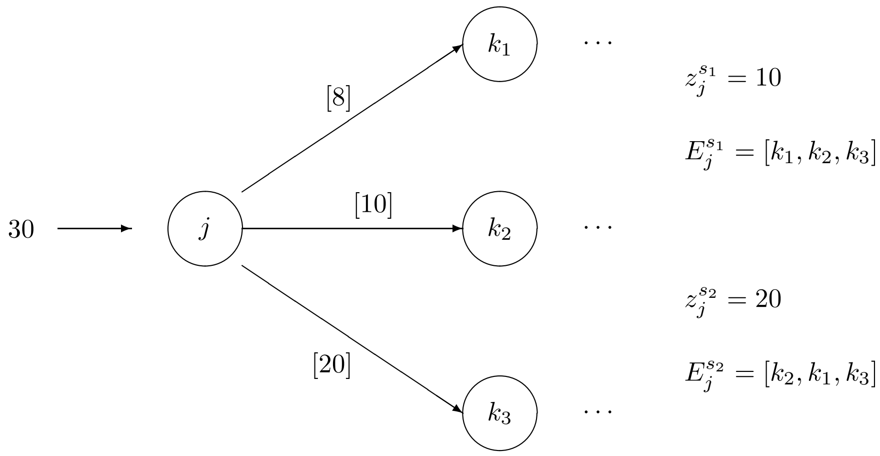


Figure 9: Flow assignment at a node

where d_k is the total demand for outgoing node k and $\overline{u}_{jk}$ is the residual capacity of link (j, k). If $\eta = 1$, all users access their preferred node k and the loading terminates trivially. Otherwise, let

$$\overline{k} \in \arg\underset{k \in \overline{K}}{\text{minimum}} \{\overline{u}_{jk}/d_k\}$$

denote the head node of the link (j, k) that gets saturated first. One then loads a fraction η of each demand to node k, updates demands, removes node k from the preference lists and repeats the process until demand is exhausted.

Let us consider a nontrivial example involving three outgoings links and two user groups of size 10 and 20 having respective preference lists $[k_1, k_2, k_3]$ and $[k_2, k_1, k_3]$. We set the capacities of the three links to 8, 10 and 20. At the first iteration, link (j, k_2) gets saturated first; 5 flow units from strategy s_1 are assigned to k_1, et 10 units from s_2 to k_2. Next, one deletes node k_2 from the preference orders $E_j^{s_1}$ and $E_j^{s_2}$. At the second iteration, 15 units (5 from s_1 and 10 from s_2) select link (j, k_1), whose residual capacity is 3. A third iteration is required to terminate the process, as the 10 residual units are assigned to link (j, k_3). Note that a by-product of the loading process is the link access probabilities π_{jk} that allow to compute the strategic cost function $\boldsymbol{c}$. The process is summarized in Table 6 while, based on the same data, the outcome of the 'parallel' assignment is summarized in Table 7.

Although there are considerable differences in the assignments resulting from the two rules, these agree

	link:	(j,k_1)	(j,k_2)	(j,k_3)
iteration 1	residual capacity:	8	10	20
	flow:	$5(s_1)$	$10(s_2)$	0
iteration 2	residual capacity:	3	0	20
	flow:	$5(s_1)$+$1(s_1)$+$2(s_2)$	$10(s_2)$	0
iteration 3	residual capacity:	0	0	20
	flow:	$6(s_1)$+$2(s_2)$	$10(s_2)$	$4(s_1)$+$8(s_2)$

probabilities: $\pi^{s_1}_{jk_1} = \frac{6}{10} \quad \pi^{s_1}_{jk_3} = \frac{4}{10} \quad \pi^{s_2}_{jk_1} = \frac{2}{20} \quad \pi^{s_2}_{jk_2} = \frac{10}{20} \quad \pi^{s_2}_{jk_3} = \frac{8}{20}$

Table 6: Outcome of single file loading at node j

	link:	(j,k_1)	(j,k_2)	(j,k_3)
iteration 1	residual capacity:	8	10	20
	flow:	$8(s_1)$	$8(s_2)$	0
iteration 2	residual capacity:	0	2	20
	flow:	$8(s_1)$	$8(s_2)+\frac{2}{7}(s_1)+\frac{12}{7}(s_2)$	0
iteration 3	residual capacity:	0	0	20
	flow:	$8(s_1)$	$\frac{68}{7}(s_2)+\frac{2}{7}(s_1)$	$\frac{12}{7}(s_1)+\frac{72}{7}(s_2)$

probabilities: $\pi^{s_1}_{jk_1} = \frac{8}{10} \quad \pi^{s_1}_{jk_2} = \frac{2}{7\cdot 10} \quad \pi^{s_1}_{jk_3} = \frac{12}{7\cdot 10} \quad \pi^{s_2}_{jk_2} = \frac{68}{7\cdot 20} \quad \pi^{s_2}_{jk_3} = \frac{72}{7\cdot 20}$

Table 7: Outcome of the parallel loading at node j

in important situations, namely when at most two choices are available at each decision node. Other theoretical points worth mentioning are:

(i) The cost function $\boldsymbol{c}$ is continuous which, together with the compactness of the feasible set, implies that the set of equilibrium solutions is nonempty.

(ii) The loading procedure only makes sense if the network is acyclic, and nodes are processed in the relevant topological order. If this condition is not fulfilled, it is yet possible to perform the loading operation by iterating, à la Gauss–Seidel, with respect to the OD pairs. At a given iteration, the access probabilities for all OD pairs, with the exception of the current one, are frozen to their previous values.

(iii) The cost mapping $\boldsymbol{c}$ may fail to be monotone under both queue disciplines.

(iv) In the single queue case, the issue of convexity of the equilibrium set is open.

From the practical side, static strategic models are of limited scope. However, they can be adapted to time-dependent networks and help in developing dynamic models that take into account the rational reaction of users to online information. In a dynamic context, models must take into account the natural temporal priorities; this leads to a more complex loading procedure involving a cost mapping $\boldsymbol{c}$ that may fail to be continuous. On the positive side, the mapping's upper semi-continuity ensures that the set of equilibria is nonempty, and the underlying time-space network is trivially acyclic.

3.4 Non-additive route costs

In our previous discussions, we have assumed that route costs were additive, that is, $c_r(\boldsymbol{h}) = \sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\boldsymbol{v})$ holds for all $r \in \mathcal{R}$ and consistent flows $(\boldsymbol{h}, \boldsymbol{v})$. This is not true, however, in several situations involving route-specific tolls or route-specific vehicle emissions. Cost additivity also fails in loss networks arising, among others, in telecommunication systems.

Let us reconsider the model with two attributes, time and money, introduced in Section 3.1. Empirical studies show that, in contrast with our previous assumption, the VOT parameter can be nonlinear. Consider, for example, that

$$c_r(\boldsymbol{h}) = f_r + \beta_r \left(\sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\boldsymbol{v}) \right), \tag{47}$$

where f_r is the monetary outlay and β_r is now a *function* of travel time, for route $r \in \mathcal{R}$. In this model, time is converted into money through the function β. It is also possible to work with time as the numéraire, thus considering instead a cost model of the form

$$c_r(\boldsymbol{h}) = \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\boldsymbol{v}) + \alpha_r(f_r), \tag{48}$$

where the scalar α_r converts money into time along route r. Interestingly, even if we set α_r to β_r^{-1}, the two models are *not* equivalent. From the analytic point of the view, the latter model has the advantage that there exists an equivalent optimization formulation in the case of separable link costs: provided that f_r is flow-independent, and that the value of ν_r does not vary within OD pair r, the partial derivative of the objective function

$$\sum_{l \in \mathcal{L}} \int_0^{v_l} t_l(s) + \sum_{k \in \mathcal{C}} \sum_{r \in \mathcal{R}_k} h_r \alpha_k(f_r)$$

with respect to h_r is exactly $c_r(\boldsymbol{h})$ given in (48), and Wardrop's conditions reduce to the optimality conditions of this optimization problem. We also see that the objective function is convex under the same conditions as for the additive model. Of course, these results hold because monetary costs are flow independent; observe the similarity between this result and those of Section 3.3.

It is clear that an algorithm in link-flow space based on first-order approximation cannot be immediately extended to this model; indeed the 'best' route cannot be determined through a standard shortest route calculation.

Bibliographical notes

References to the basic multi-mode model of Section 3.1 can be found in [Flo77, Flo79, ToW96, FWH02, MaW04]. The introduction of the bi-criterion model in the transportation literature is due to Dial [Dia79], who also proposed a solution algorithm ([Dia96a, Dia96b]). Early on, Dafermos [Daf81] gave a necessary condition in order that the variational formulation of the problem be monotone. Similar results appear in [NaD02]. Marcotte [Mar98] proved that these conditions were necessary for both the monotonicity and integrability properties to hold. Equivalent finite-dimensional formulations of the problem have been proposed in [Leu93, Mar98].

Section 3.2 introduces a modeling paradigm from the viewpoint of a control policy. This way of interpreting the model leads to interesting interpretations of the Lagrange multipliers, and the use of the 'side constraint' that $\boldsymbol{v} = \boldsymbol{v}^*$ must hold for some a priori set feasible flow $\boldsymbol{v}^*$ is the direct route to the use of link tolls for achieving a system optimum or any other desired flow, as presented for example in Section 6.2. Side constraints can emerge from other sources than controls, however, and Larsson and Patriksson [LaP94b, LaP97, LaP99] describe several such circumstances, and show how the Lagrange multiplier terms can be associated with link queues. That the Lagrange multipliers are not unique was established in [LaP98], whence interpretations must be used with great care, for example in the context of equilibrium queues. Previously, equilibrium characterizations of link capacitated models have been provided in [Jor63, Hea80], and the interpretation of the multipliers β_l as stationary link queues can be found in [MPT75, PaT75]. See also Ferrari [Fer95]. The numerical example stems from [LaP98], as does the description of the toll polyhedron development. Related, and in many ways parallel, work exists for the special case where one wishes to achieve a *system optimal* solution. In this field, the master's thesis by Bergendorff [Ber95] was followed by a series of work by Don Hearn et al. (e.g., [BHR97, HeR98]).

The strategic approach outlined in Section 3.3 was motivated by the transit equilibrium model described in [NPM01], and is based on [MNS04]. Additional information, including technical details, is available in [MaN98, HMS04a, HMS04b]. Mathematical programming formulations of transit equilibrium problems, based on the notion of strategy (or hyperpath), were independently proposed in [NgP88] and [SpF89].

The non-additive models described in Section 3.4 are due to Gabriel and Bernstein [GaB97] (see also [ChB03]), for the case (47) of converting money into time, and Larsson et al. [Lar+02], for the case (48) of converting time into money. That the two models are not equivalent was observed in [Lar+02], and later established in more generality theoretically in [BeW00]. The logit-based stochastic user equilibrium model is due to Fisk [Fis80]. For further reading on stochastic traffic equilibrium models, see [She85, Aka96, Wat99, Cas01]. The effects of modeling emission effects are discussed in [Lar+02]. Recent models of loss networks in telecommunication can be found in [AEA02]. As is stated in the section, the model considered converts money into time, and models based on this transformation have better properties.

4 Solution algorithms: The basic problem

4.1 General guidelines

Finding a flow pattern satisfying Wardrop's user equilibrium conditions (2) amounts to solving a large-scale, network structured variational inequality problem. The computational challenge is threefold: How to exploit the underlying network structure of the problem efficiently? How to obtain fast local convergence? How to ensure global convergence?

Since the constraint set F, in the link–node representation, is the feasible set of an uncapacitated multi-commodity flow problem, minimizing a linear function over F is a task that can be efficiently performed by finding the shortest route trees rooted at the origin nodes. At this level, the distinction between link, route or commodity flows becomes irrelevant, as the shortest route procedure provides information at all three levels of (dis)aggregation.

As far as convergence speed is concerned, algorithms combining the above-mentioned linear subproblems with simple line searches only achieve sub-linear convergence. To achieve linear or super-linear convergence in such a framework, one may have to solve a *quadratic* problem over the multi-commodity network; unfortunately, such task cannot be performed on large scale instances. Suppose, however, that we consider the link–route representation, and consider the set of routes carrying positive flow to be fixed. Then, solving the corresponding restricted problem by fast algorithms is indeed possible. Since, in practical applications, a relatively small set of routes is required to describe an equilibrium flow, this fact suggests that route (or, column) generation strategies are prime suspects for solving traffic equilibrium problems. It must be understood that the maximum size of the working set of routes will influence the behavior of the algorithm: The larger the maximum cardinality of this set, the better the convergence. However, a trade-off must be achieved between a good convergence rate on the one hand and, on the other hand, the size and the computational difficulty associated with the restricted problems. As a limiting case, the slow converging Frank–Wolfe method is obtained by limiting the size of the working set to two feasible points, one corresponding to the current iterate, and the second to the solution to the linear subproblem.

If $\boldsymbol{t}$ is not a gradient mapping, then one must be careful that an aggressive column dropping scheme will not result in a sequence of iterates that cycles between identical working sets. The third issue is more or less irrelevant if $\boldsymbol{t}$ is a gradient mapping, since the potential $\phi = \int \boldsymbol{t}$ then may be used to monitor descent towards a stationary (Karush–Kuhn–Tucker) point, that is, a user equilibrium. If $\boldsymbol{t}$ is *not* a gradient, then some monotonicity property (such as pseudo-monotonicity) is required in order to be able to implement a provably convergent method. In the following, we elaborate more on these issues for the basic traffic equilibrium model.

4.2 Model structures and the verification of equilibria

Some favorable problem structures are observed when looking at the models (15) and (16):

OD pair separability The feasibility of the demand and traffic volume in one OD pair does not affect that of another. The problems indeed are multi-commodity flow problems, where there are no side constraints acting on more than one commodity, such as link flow capacity constraints. This property holds for more general cost and demand models as well.

Cost separability The objective function, as well as the link cost and demand functions, is separable in the link volume and demand variables. This property does not extend to more general cost and demand models.

Primal–dual relations In both the link–route and link–node representations, the vector π is a Lagrange multiplier vector, which further measures the least cost. This property holds for more general cost and demand models as well.

The link–route representation has a simpler constraint structure than the link–node one; the constraints in the former describe a simplex, while those of the latter describe flow conservation. The former, on the other hand, has an exponential number of variables, which must be enumerated iteratively, while the latter has a polynomial number of variables, and a richer constraint structure.

The very definition of a variational inequality provides a convenient criterion for verifying whether a vector is in equilibrium. Indeed, the inequality

$$\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in X.$$

is satisfied if and only if $\boldsymbol{x}$ is optimal for the linear program

$$\underset{\boldsymbol{y} \in X}{\operatorname{minimize}}\, \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{y}.$$

If one keeps demand fixed, this amounts to checking whether the *gap function* (see also Appendix A)

$$\operatorname{gap}(\bar{\boldsymbol{h}}) := \underset{(\boldsymbol{h},\bar{\boldsymbol{d}}) \in \widehat{H}_d}{\operatorname{maximum}}\, \boldsymbol{c}(\bar{\boldsymbol{h}})^{\mathrm{T}}(\bar{\boldsymbol{h}}-\boldsymbol{h}) = \sum_{k \in \mathcal{C}} \underset{(\boldsymbol{h}_k,\bar{d}_k) \in \widehat{H}_{k,d}}{\operatorname{maximum}}\, c_k(\bar{\boldsymbol{h}})^{\mathrm{T}}(\bar{\boldsymbol{h}}_k-\boldsymbol{h}_k). \tag{49}$$

is zero[16]. This LP problem defines a set of shortest route problems; an equilibrium is reached if and only if the total cost of transportation along the shortest routes in each OD pair equals the total travel cost of the current OD flow. The value of $\operatorname{gap}(\boldsymbol{h})$ is always non-negative, and a flow vector $\boldsymbol{h}$ is an equilibrium if and only if $\operatorname{gap}(\boldsymbol{h}) = 0$. If $\operatorname{gap}(\boldsymbol{h})$ is positive, one has identified a set of attractive routes that are not currently used (at least not enough).

In the elastic demand case, it seems natural to consider the iteration

$$\boldsymbol{d} \mapsto (\boldsymbol{g} \circ \boldsymbol{\pi} \circ \boldsymbol{t} \circ \boldsymbol{v})(\boldsymbol{d}) \tag{50}$$

that takes the current demand and calculates a corresponding inelastic demand solution, from which the least route cost is calculated. The value of the demand function then yields the new demand estimate. A sufficient condition that this fixed-point iteration converge is that it be contractive, that is, for an equilibrium demand $\boldsymbol{d}^*$ and any feasible demand vector $\boldsymbol{d}$, there exists a scalar $\theta \in (0,1)$ such that $\|\boldsymbol{g}(\boldsymbol{\pi}(\boldsymbol{t}(\boldsymbol{v}(\boldsymbol{d})))) - \boldsymbol{d}^*\| \leq \|\boldsymbol{d}-\boldsymbol{d}^*\|$ holds. Given the complicated form of this mapping, whether the contraction condition is satisfied is difficult to analyze from the original data, and updating rules for flows and demands are instead based on approximations of one or more of these mappings, examples of which will be provided below.

4.3 Decomposition–coordination

The separability of the OD pairs (or, user classes, etc.) in the constraints suggests the use of a decomposition strategy, wherein flow of a given type is updated upon while keeping the others fixed; the remaining problem is then a single-commodity network flow problem. Among the advantages of such a scheme we mention first that the link cost function has stronger monotonicity properties when viewed on this smaller subspace than on the entire space of flows; for example, if $v_l \mapsto t_l(v_l)$ is strictly monotonically increasing, then it is (usually) non-strictly monotone in the vector of OD link flows $\boldsymbol{w}_k$, $k \in \mathcal{C}$, but it is strictly monotonically increasing when keeping flows in all but one OD pair fixed. Second, several single-commodity network flow algorithms are available, among which are (potentially) super-linearly convergent ones. Third, algorithms of this type can be viewed as block versions of Gauss–Seidel (or Jacobi) algorithms, which have a sound convergence theory. They can however be slow when the number of blocks is large, since too much of the cost interaction among the commodities is lost. One can then take advantage of the problem structure by creating a small number of large blocks of variables such that, for example, all flow from one given origin is updated at any one time. Contrary to the use of decomposition over OD pairs or origins, a decomposition over links is less convenient, as costs are not always defined by separable functions.

A natural means to decompose the problem over OD pairs is to temporarily fix link costs, thereby removing the cost dependence between OD pairs. In the fixed demand problem, what is left is a number of shortest route problems, on which demand is assigned. In the elastic deman case, the demand to be assigned onto the shortest routes is given by the demand function evaluated at the shortest route costs. In this way, we obtain the algorihms of Frank–Wolfe and Evans', respectively.

[16]The index k in the second equality relates to feasibility sets and sub-vectors for individual OD pairs.

4.3.1 Decomposition by Lagrangean relaxation

In order to take advantage of the link cost separability in the models (15) and (16), we must eliminate their dependence in the constraints by introducing some form of relaxation. One possibility is to introduce explicit multipliers, $\boldsymbol{\nu} \in \Re^{|\mathcal{L}|}$, for the constraints '$\boldsymbol{v} = \sum_{k\in\mathcal{C}} \boldsymbol{w}_k$.' With the (redundant) constraints $\boldsymbol{v} \in \Re_+^{|\mathcal{L}|}$ appended to the Lagrangean relaxed problem, we obtain a problem, equivalent to VIP($\boldsymbol{t}, F_d$), which amounts to finding $(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{d}, \boldsymbol{\pi}, \boldsymbol{\nu})$ such that

$$\mathbf{0}^{|\mathcal{L}|} \leq \boldsymbol{v} \perp (\boldsymbol{t}(\boldsymbol{v}) - \boldsymbol{\nu}) \geq \mathbf{0}^{|\mathcal{L}|}, \tag{51a}$$

$$\boldsymbol{E}\boldsymbol{w}_k = \boldsymbol{i}_k d_k, \qquad k \in \mathcal{C}, \tag{51b}$$

$$\mathbf{0}^{|\mathcal{L}|} \leq \boldsymbol{w}_k \perp \left(\boldsymbol{\nu} - \boldsymbol{E}^{\mathrm{T}}\boldsymbol{\pi}_k\right) \geq \mathbf{0}^{|\mathcal{L}|}, \qquad k \in \mathcal{C}, \tag{51c}$$

$$\boldsymbol{g}(\boldsymbol{d}) = \boldsymbol{\pi}, \tag{51d}$$

$$\boldsymbol{v} - \sum_{k\in\mathcal{C}} \boldsymbol{w}_k = \mathbf{0}^{|\mathcal{L}|} \tag{51e}$$

holds. Fixing $\bar{\boldsymbol{\nu}} \geq \boldsymbol{t}(\mathbf{0})$, and noting that (51e) is the dual optimality condition, the problem separates into one over $\boldsymbol{v}$ and $\boldsymbol{w}_k$, $k \in \mathcal{C}$. In the separable case, (51a) further decomposes into a set of single-variable problems with the objective $\int_0^{v_l} (t_l(s) - \bar{\nu}_l)\, ds$, to be minimized over $\Re_+$, and with a solution given by the inverse of t_l evaluated at $\bar{\nu}_l$, whenever the latter is non-negative (otherwise it is set to zero). In the general case, this is a nonlinear complementarity problem (NCP) in $\Re^{|\mathcal{L}|}$. The second problem, (51b)–(51d), amounts to finding, for each $k \in \mathcal{C}$, the shortest route based on the vector $\bar{\boldsymbol{\nu}}$ of link costs. The shortest route costs are used to calculate the demand through the demand function Demand is then distributed onto the shortest routes to form the vectors $\boldsymbol{w}_k$. A step is then taken along the dirction $\boldsymbol{\nu}$, in order to produce a result that better fulfills the last constraint. Although being a primal–dual algorithm, the aggregate of the $\boldsymbol{w}_k$ is a feasible link flow; this is a consequence of the fact that the dualization is made with respect to the definitional constraint (51e).

Another way to decompose the problem is to relax the linear equality constraints defining the network, that is, (51b). Introducing explicit multipliers $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$ for each commodity $k \in \mathcal{C}$, we obtain a problem, equivalent to VIP($\boldsymbol{t}, \widehat{F}_d$), which amounts to finding $(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{d}, \boldsymbol{\pi})$ such that

$$\mathbf{0}^{|\mathcal{L}|} \leq \boldsymbol{w}_k \perp \left(\boldsymbol{t}(\boldsymbol{v}) - \boldsymbol{E}^{\mathrm{T}}\boldsymbol{\pi}_k\right) \geq \mathbf{0}^{|\mathcal{L}|}, \qquad k \in \mathcal{C}, \tag{52a}$$

$$\boldsymbol{E}\boldsymbol{w}_k = \boldsymbol{i}_k d_k, \qquad k \in \mathcal{C}, \tag{52b}$$

$$\boldsymbol{g}(\boldsymbol{d}) = \boldsymbol{\pi}, \tag{52c}$$

$$\boldsymbol{v} - \sum_{k\in\mathcal{C}} \boldsymbol{w}_k = \mathbf{0}^{|\mathcal{L}|} \tag{52d}$$

holds. Fixing $\bar{\boldsymbol{\pi}}_k$ (and letting $\bar{\pi}_{pq} := \boldsymbol{i}_k^{\mathrm{T}}\bar{\boldsymbol{\pi}}_k$), and noting that (52b) is the dual optimality condition, the remaining problem is solved as follows. The demand $\boldsymbol{d}$ is given, through (52c), by the demand function at $\bar{\pi}$. The complementarity system (52a) is solved, in the separable case, by first defining, for each link $l = (i,j) \in \mathcal{L}$, $\bar{p}_{ij}(\boldsymbol{\pi}) := \text{maximum}_{k\in\mathcal{C}} \{\pi_{jk} - \pi_{ik}\}$ to be the largest node potential difference among the commodities. Let $\mathcal{C}_{ij}(\boldsymbol{\pi}) \subseteq \mathcal{C}$ denote the set of commodities for which $\pi_{jk} - \pi_{ik} = \bar{p}_{ij}(\boldsymbol{\pi})$ holds. Then, $v_{ij}(\boldsymbol{\pi}) := \text{maximum}\,\{0, t_{ij}^{-1}(\bar{p}_{ij}(\boldsymbol{\pi}))\}$ is the total link flow in link (i,j). The commodity link flows $w_{ij,k}$, $k \in \mathcal{C}$, can be taken as any distribution of the total flow $v_{ij}(\boldsymbol{\pi})$ onto the commodities in $\mathcal{C}_{ij}(\boldsymbol{\pi})$. It remains to define an updating rule for the vectors $\bar{\boldsymbol{\pi}}_k$ such that we approach also a solution to (52b). If this Lagrangean relaxation is utilized within a cyclic decomposition algorithm over OD pairs, where the subproblems are single-commodity network flow problems, then more efficient dual algorithms are available, since the duals then are differentiable.

4.3.2 Decomposition by linearization

A decomposition (relaxation) can also be achieved by manipulating the cost structure. Suppose, for example, that the link costs are fixed at their current values, and consider the resulting relaxation. Taking VIP($[\boldsymbol{t}, \boldsymbol{g}], \widehat{F}_d$) as the model for discussion, and the iterate $(\boldsymbol{v}^\tau, \boldsymbol{d}^\tau, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}}^\tau)$ at iteration τ, fixing the link costs to $\boldsymbol{t}(\boldsymbol{v}^\tau)$ yields the following procedure for solving this relaxation of VIP($[\boldsymbol{t}, \boldsymbol{g}], \widehat{F}_d$): first, for each $k = (p,q) \in \mathcal{C}$, calculate a least-cost route, cf. (6a), whence the vectors $\bar{\boldsymbol{\pi}}_k^\tau$, $k \in \mathcal{C}$, $(\bar{\boldsymbol{\pi}}^\tau)_{(p,q)\in\mathcal{C}} = (\boldsymbol{i}_k^{\mathrm{T}}\bar{\boldsymbol{\pi}}_k^\tau)_{k\in\mathcal{C}}$

are given; second, calculate the demand through $\bar{\boldsymbol{d}}^\tau := \boldsymbol{g}(\bar{\boldsymbol{\pi}}^\tau)$; third, assign this demand to each OD pair's shortest route, giving rise to the link flows $\bar{\boldsymbol{w}}_k^\tau$, $k \in \mathcal{C}$, and $\bar{\boldsymbol{v}}^\tau = \sum_{k\in\mathcal{C}} \bar{\boldsymbol{w}}_k^\tau$. If link costs and demands are gradient mappings, we can perform a line search in the direction of $(\bar{\boldsymbol{v}}^\tau, \bar{\boldsymbol{d}}^\tau, (\bar{\boldsymbol{\pi}}_k)_{k\in\mathcal{C}}^\tau) - (\boldsymbol{v}^\tau, \boldsymbol{d}^\tau, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}}^\tau)$, whence we have defined a partial linearization algorithm known as Evans' algorithm; whenever demand is inelastic, this is precisely the Frank–Wolfe algorithm.

The route information generated during the process of calculating shortest routes is badly utilized in line search algorithms, considering the effort in generating it. If, instead, we were to store the routes that carry flow or that are new and promising, and solve the equilibrium problem over those, we would in fact be addressing a restriction to the link–route formulation of the original model (a restricted master problem). As we have remarked earlier, this model form has an advantage in that the constraint structure can be effectively utilized; if we keep down the size of the route set, we also have a chance of avoiding the drawback of this formulation. Consider then the model VIP$(\boldsymbol{t}, \widehat{F})$, where we have available a restriction $\widehat{\mathcal{R}} \subset \mathcal{R}$ of the route set. We recognize that the simplex structure can be put to good use, for instance by devising a decomposition over OD pairs or origins. This is achieved in a cyclic Newton-like algorithm, wherein we keep the flow from all but one group of users fixed, and then utilize a diagonalized second-order approximation of the remaining objective function. At a given iteration where the flow is $\boldsymbol{h}^\tau$, we would, for one OD pair k, say, solve the following problem:

$$\underset{\boldsymbol{h}_k}{\text{minimize}} \sum_{r\in\widehat{\mathcal{R}}_k} \left\{ c_r(\boldsymbol{h}^\tau) h_r + \frac{b_r^\tau}{2\gamma_\tau} h_r^2 \right\}, \tag{53a}$$

$$\text{subject to} \sum_{r\in\widehat{\mathcal{R}}_k} h_r \bar{d}_k, \tag{53b}$$

$$\boldsymbol{h}_k \geq \boldsymbol{0}^{|\widehat{\mathcal{R}}_k|}, \tag{53c}$$

where $0 < b_r^\tau \approx \partial c_r(\boldsymbol{h}^\tau)/\partial h_r$, and the value of $\gamma_\tau > 0$ is chosen such that the improvement made is the best possible; the resulting solution is taken to be $\boldsymbol{h}_k^{\tau+1}$. The above problem has a low complexity: solving for the optimal Lagrange multiplier for the equation can be done in $O(|\widehat{\mathcal{R}}_k|)$ time, that is, linear in the number of routes. Achieving low complexity is essential in the subproblems of an iterative algorithm, as the size of traffic problems can be very large.

A final remark on algorithms for the basic models is that it seems difficult to devise truly super-linearly convergent algorithms that utilize problem structure efficiently. One main reason is the presence of multiple commodities; for separable cost nonlinear single-commodity flow problems, super-linearly convergent algorithms are available, which also utilize the network property in an efficient way. However, the Gauss–Seidel algorithm—which is needed to coordinate the search in the multi-commodity setting—is linearly convergent, at best. There is therefore still plenty of room for algorithmic improvements, in particular for new and clever ways of decomposing and coordinating the problem.

4.4 Merit functions and convergence for asymmetric models

When solving models involving cost and demand functions that are gradient mappings[17], we have access to natural merit functions [cf. (15) or (16)], with which we can measure the progress of an algorithm and perform line searches; within this framework, the whole area of nonlinear optimization is open to exploration. When turning to non-integrable models, these merit functions are no longer well-defined, and progress toward an equilibrium must be measured by some artificial construct. One such function is the primal gap function given in (49) or (106), which can be calculated entirely in link flow space. As with most other merit functions used for variational inequality problems, its its evaluation requires the solution of an approximation to the original problem. The computation of equilibria for non-separable models are in fact much more complex than for separable ones:

Decomposition In the integrable case, some of the most natural decompositions lead to low-complexity subproblems. In the non-integrable case, however, they either become unavailable altogether, or they become more complex. For example, a decomposition over links will no longer be natural or even possible. Further, if the form of the cost and demand is complex, it also means that computing their values and derivatives will be more complex, and this will also affect the computational complexity of an algorithm.

[17]This is referred to as the 'integrable case.'

Search direction → merit function In the integrable case, virtually every algorithm proposed is a feasible direction-based descent method which utilizes the natural objective in (15) or (16). Establishing convergence of such algorithms is relatively straightforward, as it is easy to satisfy the descent condition that the gradient of the merit function makes an obtuse angle with the search direction, and it does not rely on the monotonicity properties of the cost and demand functions. In the non-separable case, however, the angle condition becomes much more difficult to establish for a suitable merit function. The merit function is intimately associated with the search direction, and in some cases it is even defined by the solution to the search direction-finding problem. That also means that, in contrast to the separable case, the merit function and its gradient depends on the search direction chosen. In order to derive convergence results, more stringent requirements on both the search direction and the step length must be enforced.

As an example of the complications that arise in the general, we mention that an algorithm that would mimic the Frank–Wolfe strategy by treating the cost function as if it were a gradient mapping, need not converge for asymmetric cost models[18]. However, it is straightforward to show that, whenever the 'Frank–Wolfe' direction is unique and the cost mapping is monotone, it constitutes a feasible descent direction for the primal gap function. If the set of Frank–Wolfe directions is not a singleton, it is yet possible to prove that there exists a member of that set that is gap-decreasing. Based upon this result, an implementable and convergent algorithm that operates in link flow space can be designed.

There exist several variational inequality algorithms that so far have not caught the attention of the traffic research community, and that converge under rather mild assumptions on the travel cost and demand mappings. For one, the projection algorithm, which takes an iterate $\boldsymbol{v}^\tau$ and maps it onto the next one through the formula

$$\boldsymbol{v}^{\tau+1} := \mathrm{Proj}_{\hat{F}}\,(\boldsymbol{v}^\tau - \gamma_\tau \boldsymbol{t}(\boldsymbol{v}^\tau)), \qquad \gamma_\tau > 0,$$

where $\mathrm{Proj}_{\hat{F}}$ denotes the Euclidean projection onto $\hat{F}$, can be supplied with special line searches that make it convergent for monotone cost mappings. (In its original statement it requires strong monotonicity or co-coercivity, and relies on the estimate of the cost mapping's strong monotonicity modulus and Lipschitz constant.) The class of extra-gradient algorithms involves two projections of the above type, and is convergent under even milder monotonicity requirements: pseudo-monotonicity.

The most natural environment for these algorithms is the singly constrained restricted master problem in a route (column) generation algorithm, that is, a restriction of the feasible sets defined in (15) or (16) to a subset $\widehat{\mathcal{R}}$ of the routes. The reason why we favour this environment is that projections are easy to perform onto such sets, cf. (53).

An especially interesting avenue for constructing algorithms for non-strictly monotone problems is to combine an algorithm of one's choice (and which perhaps requires strong monotonicity, such as the Jacobi algorithm) and the proximal point algorithm. It is defined thus: for the given cost function $\boldsymbol{t}$ in the variational inequality $\mathrm{VIP}(\boldsymbol{t}, \hat{F})$, one would add a multiple ($\gamma_\tau > 0$) times an affine cost of the form $\boldsymbol{v} \mapsto \boldsymbol{v} - \boldsymbol{v}^\tau$, if we work in link space. The resulting mapping $\boldsymbol{v} \mapsto \boldsymbol{t}(\boldsymbol{v}) + \gamma_\tau(\boldsymbol{v} - \boldsymbol{v}^\tau)$ is strongly monotone with a modulus of at least γ_τ, provided that the original cost function $\boldsymbol{t}$ is monotone. Therefore, if we have an algorithm which requires strong monotonicity, we apply it instead to this perturbed problem. At the (approximate) solution to the perturbed problem, we again perturb the original function at the new iterate, $\boldsymbol{v}^{\tau+1}$, and proceed, perhaps with a different value of γ.

Diagonalization is a favorite among heuristics for non-separable problems. It amounts to, at an iteration, temporarily removing the dependency of a link's cost on the flow in other links. This is related to the Jacobi algorithm, and may suffer from quite poor convergence characteristics: if the correlation in reality is quite large, one can expect the algorithm to diverge, or at least converge slowly; on the other hand, if the cost dependence between links is mild, or local to a few links only, then convergence can be expected to be fast.

We close this section by mentioning that minimizing the dual gap function

$$\underset{\boldsymbol{x} \in X}{\text{maximum}} \qquad \boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y})$$

can be achieved by solving the semi-infinite linear program to

$$\begin{aligned} &\underset{\boldsymbol{x} \in X, z \in \Re}{\text{minimize}} && z \\ &\text{subject to} && z \geq \boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}), \qquad \boldsymbol{y} \in X. \end{aligned} \tag{54}$$

[18] In such an algorithm, line searches are performed based on first-order information, not the objective function.

This structure lends itself naturally to decomposition algorithms similar to the linear programming method of Dantzig and Wolfe, and to more recent interior point cutting plane methods.

4.5 Implementations and computational experience

Commercial implementations of equilibrium solvers are frequently based on the the Frank–Wolfe method, or variants thereof. Although it has been known for quite some time that the Frank–Wolfe algorithm has the worst (sub-linear) convergence rate among equilibrium algorithms, its high initial efficiency, small memory requirements, efficient handling of the underlying network structure and the natural interpretation behind its workings—it is reminiscent of heuristics that were devised already in the mid 1960's—have made it popular. Its counterpart for elastic demand models is Evans' algorithm, which has similar convergence characteristics: demand quickly converges, and once demand stabilizes, the algorithm's behavior is similar to Frank–Wolfe's.

Among the favored extensions of the Frank–Wolfe algorithm is the class of simplicial decomposition/column generation algorithms (see the discussion on the DSD algorithm above) which also are based on shortest route subproblems. While the simplicity of the subproblem is preserved, the results of the computations are better utilized, as the shortest routes are kept in memory and demand is optimally assigned onto them. This restricted master problem, which has the same form as that of the original problem except that the set $\mathcal{R}$ is replaced by the known subset $\widehat{\mathcal{R}}$, is then solved using a specialized, perhaps second-order, method. In later implementations of the DSD algorithm a diagonalized Newton algorithm is applied, which also efficiently detects routes with zero flow; these routes may be removed from the set $\widehat{\mathcal{R}}$ in order to save space. Other implementations use different versions of reduced gradient and gradient projection methods. As projection methods are available also for non-separable models, it is anticipated that the best versions can be transferred to solve also such more general models.

For a numerical illustration of two of the above discussions, let us consider the 'famous' Sioux Falls network, composed of 24 nodes, 76 links, and 528 OD pairs. For a MATLAB[19] implementation of the DSD and FW algorithms, Figure 10 illustrates how fast (in terms of CPU time) convergence occurs towards an equilibrium solution. The implementation of the DSD algorithm comes with several methods for solving the restricted master problems: two versions of the Goldstein–Levitin–Polyak gradient projection method (marked 'Gradient P.' in the figure), and a diagonalized Newton method. For the small-scale network representing Sioux Falls, the gradient projection method is faster, but for larger networks, Newton's method becomes relatively more favorable. Clearly, the Frank–Wolfe method is very slow for this problem, and its relative efficiency is even further reduced for larger problems.

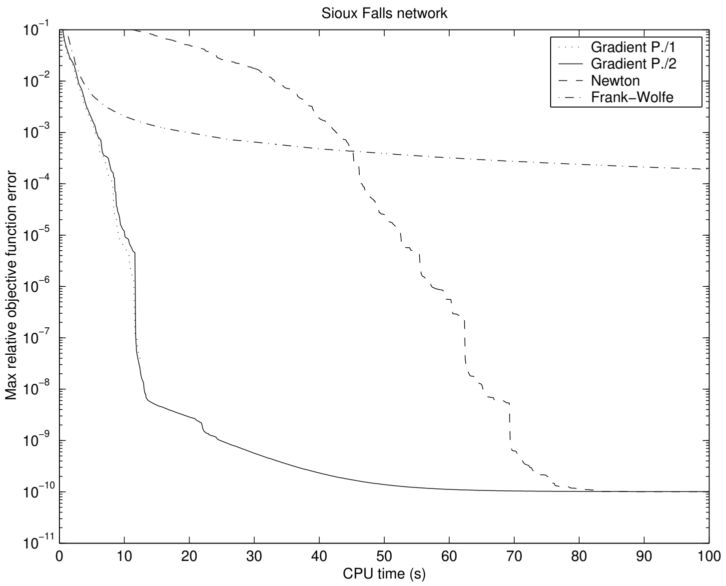


Figure 10: The performance of DSD vs. FW on the Sioux Falls network.

[19]The absolute CPU times should not be taken at face value; a corresponding C implementation can be two to three orders of magnitude faster.

A recent contribution to the field is the quasi-Newton algorithm of Bar-Gera. It utilizes the fact that the equilibrium flow from a given origin is acyclic [cf. Theorem 4(b)], and tries in a combinatorial fashion to determine the correct acyclic networks. For each origin and acyclic network, a quasi-Newton algorithm is implemented for finding the best single-commodity flow, based not only on link flows, but also on the flow through the nodes. The numerical experience so far is impressive, and it is believed that it will, if not replace, then at least become an alternative to the algorithms currently used in practice. The main drawback of this algorithm is that it has not yet been shown how to make it work for non-separable models.

Somewhat surprisingly, the field of computational testing, in particular comparative ones, remains sparse. Comparisons are still quite often performed against the Frank–Wolfe algorithm; such results are not very enlightening, given the bad convergence behavior of the latter (see Figure 10).

Several interesting alternatives to the classical algorithms have never been tested, and the protocols of those that are tested on a regular basis would not pass the scrutiny of a numerical analysis specialist. The latter is not so surprising, considering that most of the algorithmic development relates to specific applications. The former is not so surprising either, given that most researchers in the field come from an area different from mathematical programming.

The low-complexity subproblems, like the shortest route problem and the quadratic knapsack problem (53), are also mathematical objects of interest. Although there exist efficient algorithms for these problems, the fact that they have to be solved repeatedly leaves room for algorithms that are less efficient according to a worst-case criterion, but lend themselves better to re-optimization. As is the case for sorting, the worst-case behavior does not tell the whole story.

Bibliographical notes

The gap function defined by (49) is a measure of the distance to an equilibrium. Its negative is the directional derivative of the traffic equilibrium objective function ϕ [see (15)] along the direction towards a shortest route solution—which explains its role in the Frank–Wolfe algorithm. The gap function was originally considered by Zuhovickiĭ et al. [ZPP69], and introduced into the field of transportation science by Murchland [Mur70]; see also [Aus76].

Fixed-point type results concerning the algorithmic map (50) are based on contraction arguments (which are difficult to verify in practice), and can be found in [Dun73, Bai75]. One reason we do not cover the fixed-point literature is that such formulations, through the use of problem mappings similar to (50), hide the salient structures of the underlying model. We prefer, departing somewhat from orthodoxy, to view such fixed-point problems not as traffic equilibrium models in their own right, but rather as algorithmic constructs.

The uncapacitated and link capacitated single-commodity flow problems discussed in Section 4.3, and algorithms for their solution, are reviewed in [KeH80, AMO93, Pat94b, Ber98, Pat06].

The procedure following the statement (51) is taken from [LLP97], and the procedure following the statement (52) stems from [Pat94b, Section 4.3.2]; both are covered in [Pat06]. The Evans and Frank–Wolfe methods were first implemented for traffic equilibrium models by Evans [Eva76], and Nguyen [Ngu74] and LeBlanc et al. [LMP75], respectively.

The first successful attempt at building an efficient route-based column generation algorithm was made by Larsson and Patriksson [LaP92], who devised the *disaggregate simplicial decomposition* (DSD) framework. Although similar approaches had been proposed earlier, for example in [LNT73, BeG82], the restricted master problems were then deemed too difficult to solve, and route-based methods were therefore for many years considered inefficient. A few years later, Hearn et al. ([HLV87]) designed a similar algorithm where routes however are aggregated into all-or-nothing solutions, thereby restricting the number of variables in the restricted problems and at the same time reducing the potential speed of convergence for the respective route flows. The quadratic subproblem shown in (53) is the one preferred for the restricted problems, and is also used in the DSD algorithm. The convergence of this scaled gradient projection method has been known for many years (e.g., [Ber76]). What makes it attractive is that the quadratic term can incorporate derivative information on route costs in order to enhance convergence, and also that the solution to (53) can be computed in linear time ([Bru84]). Although route-based formulations yield variational formulations that are not *strongly* monotone, convergence can yet be proved under the assumption that the *link* cost functions are strongly monotone (see [BeG82, ZhM96]). (In the separable case, the problem's convexity is always enough to ensure convergence provided that a line search is performed in each iteration.)

For separable cost single-commodity problems, super-linearly convergent Newton-like methods that exploit network structure have been proposed [BeG75, Kli83, Esc86, GaB84]; they are also overviewed in [Pat06]. In the multi-commodity, non-separable, case, a modified Newton method has been implemented by Marcotte and Guélat [MaG88] within the framework of simplicial decomposition. The linear convergence of the Gauss–Seidel algorithm is established in [BeT89]; the convergence rate depends of course on the extent to which the different variable components (that is, flow variables) interact in the objective function. Choosing the right decomposition—by OD pair, by origin, or even by grouping together some origins or OD pairs—may therefore be crucial for the practical convergence rate of the algorithm.

In the non-separable case, the Frank–Wolfe strategy may lead to cycling (see, e.g., Hammond [Ham84]). It is interesting to note that column generation/simplicial decomposition algorithms are guaranteed to converge, as long as all routes are kept in memory, or if column dropping is done with care. Contributions to the theory and practice of such methods are found in [LaH84, Pat98]. The existence of gap-decreasing Frank–Wolfe directions, and the design of a convergent framework built around such directions, is due to Marcotte [Mar86a].

The remainder of Section 4.4 discusses several ways in which convergence can be achieved by utilizing recent contributions to the field of variational inequality algorithms. General references for the development of line search methods for variational inequality problems until the late 1990s can be found in the monographs of Patriksson [Pat98] and Facchinei and Pang [FaP03b]. Special line searches incorporated into the projection method are found in [Bru77, Pat98]. The class of extra-gradient algorithms has developed quite far since the original paper by Korpelevich [Kor77]; recent contributions can be found in [Kon01, WXW01, Sol03]. The proximal point algorithm was developed largely by Rockafellar [Roc76], following previous developments in the 1960s for more general problems. In that paper, an inexact solution of the regularized problems is already shown to be valid for monotone problems; more recent contributions, where the proximal point method is combined with projection methods, and where previous convergence rate analyses are improved, are found in [SoS99a, SoS99b, Sol03]. 'Diagonalization' is the favorite term in the transportation research literature (e.g., [She85]) for the method known as 'Jacobi' to researchers in numerical analysis and mathematical programming. Cutting-plane methods for solving the Minty formulation or, equivalently, the semi-infinite linear program (54), have been proposed by Zuhovickiĭ et al. [ZPP69] in a game-theoretical framework. It has been modified and adapted to the traffic assignment problem, using a column generation framework, by Nguyen and Dupuis [NgD84]. The analytic center cutting plane method described in [GMZ97] may provide a viable alternative.

The contributions by Bar-Gera to the practice of solving separable, fixed demand problems are found in [Bar02a, Bar02b]. Algorithms based on the DSD framework that have emerged since 1992 are found in, for example, [JTPR94, CLJ02]. The web site

`http://www.bgu.ac.il/~bargera/tntp/`

provides test problems and computational results. The reader may also consult

`http://www-rocq.inria.fr/metalau/ciudadsim/` .

Finally, the reoptimization issue has been addressed in [NPS02, PaS03].

5 Solution algorithms: variations

5.1 Multi-mode and Multi-attribute models

In a multi-modal context, decomposition is the strategy of choice for computing an equilibrium. If all modal flows are fixed, save one, the model reduces to a standard traffic equilibrium model which frequently involves a gradient mapping. At sub-iteration j of a major cycle τ, one solves the single-mode problem of finding the solution $\boldsymbol{w}^{j\,\tau}$ to the variational inequality

$$\boldsymbol{t}_m(\boldsymbol{v}^{1\,\tau}, \boldsymbol{v}^{j-1\,\tau}, \dots, \boldsymbol{w}^{j\,\tau}, \boldsymbol{v}^{j+1\,\tau-1}, \dots, v^{|\mathcal{M}|\,\tau-1}) \in N_{F_j}(\boldsymbol{w}^{j\,\tau}), \tag{55}$$

in the Gauss–Seidel variant and to the variational inequality

$$\boldsymbol{t}_m(\boldsymbol{v}^{1\,\tau-1}, \boldsymbol{v}^{j-1\,\tau-1}, \dots, \boldsymbol{w}^{j\,\tau}, \boldsymbol{v}^{j+1\,\tau-1}, \dots, v^{|\mathcal{M}|\,\tau-1}) \in N_{F_j}(\boldsymbol{w}^{j\,\tau}). \tag{56}$$

in the Jacobi (diagonalization) implementation. After the completion of an entire cycle, one updates the multi-attribute flow vector to

$$\boldsymbol{v}^{\tau+1} = \lambda \boldsymbol{v}^{\tau+1} + (1-\lambda)\boldsymbol{w}^{\tau}, \tag{57}$$

where $\lambda \in (0,1)$ is a relaxation parameter. The convergence analysis of the above two schemes is available in classical numerical analysis textbooks and is reproduced in the transportation literature. It relies on the strong monotonicity and Lipschitz moduli of the cost mapping, two constants for which reliable estimates may however be difficult to obtain; a trade-off has to be achieved between large values for λ that may yield to cycling, and small values that induce slow convergence.

The two-attribute structure is more interesting and lends itself to three lines of attacks. The first consists in discretizing the density function h and solving the resulting finite-dimensional problem by the decomposition approaches outlined above. A second strategy consists in solving a fixed point problem, either with respect to the total flow vector $\overline{\boldsymbol{v}}$ or the critical vector $\boldsymbol{\alpha}$, with the drawback that the choice of a suitable relaxation parameter may be difficult to estimate. We favor a third approach, viz. addressing the infinite-dimensional formulation (34) by a cost approximation method akin to the Frank–Wolfe linearization scheme. At iteration $\tau = 1$, one sets

$$\boldsymbol{v}^{\tau+1} = \boldsymbol{v}^{\tau} + \lambda_{\tau}(\boldsymbol{w}^{\tau} - \boldsymbol{v}^{\tau}) \tag{58}$$

where $\boldsymbol{w}^{\tau}$ is the solution to the infinite-dimensional parametric shortest paths problem. Since the information provided by the total flow vector $\overline{\boldsymbol{v}}$ is sufficient to recover the flow densities $\boldsymbol{v}(\alpha)$, the algorithm can also be defined in the finite-dimensional total flow space. If the conditions ensuring the validity of the optimization formulation (40) are fulfilled, λ can be determined by performing a line search along the direction $\boldsymbol{d}^{\tau} = \boldsymbol{w}^{\tau} - \boldsymbol{v}^{\tau}$. Otherwise, one may minimize, along $\boldsymbol{d}^{\tau}$, the primal gap function

$$\begin{aligned}
\text{gap}(\boldsymbol{v}) &= \underset{\boldsymbol{w}\in F}{\text{maximum}}\ (\boldsymbol{t}(\overline{\boldsymbol{v}}) + \alpha \boldsymbol{f}(\overline{\boldsymbol{v}}))^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{w}), \\
&= \underset{\boldsymbol{w}\in F}{\text{maximum}} \int_0^{\overline{\alpha}} \langle \boldsymbol{t}(\overline{\boldsymbol{v}}) + \alpha \boldsymbol{f}(\overline{\boldsymbol{v}}), \boldsymbol{v}(\alpha) - \boldsymbol{w}(\alpha)\rangle \, d\alpha, \\
&= \underset{\boldsymbol{w}\in F}{\text{maximum}}\ \langle \boldsymbol{t}(\overline{\boldsymbol{v}}), \overline{\boldsymbol{v}} - \overline{\boldsymbol{w}}\rangle + \int_0^{\overline{\alpha}} \alpha \langle \boldsymbol{f}(\overline{\boldsymbol{v}}), \boldsymbol{v}(\alpha)\rangle \, d\alpha - \int_0^{\overline{\alpha}} \alpha \langle \boldsymbol{f}(\overline{\boldsymbol{v}}), \boldsymbol{w}(\alpha)\rangle \, d\alpha.
\end{aligned}$$

The first term of this expression is the term found in the standard model, while the last two are the average costs associated with flow densities $\boldsymbol{v}$ and $\boldsymbol{w}$, respectively. If $\boldsymbol{f}$ is constant then the update of the latter quantity can be performed efficiently.

Without further assumptions on the problem's structure, the above described algorithm is a heuristic. However, the algorithm can be made globally convergent by imposing two simple conditions on the cost functions:

Invariance condition: The mapping $\boldsymbol{f}$ is constant over its domain F.

Slope condition: For any two distinct routes r_1 and r_2 there holds

$$\sum_{l\in r_1} f_l(\overline{\boldsymbol{v}}) \neq \sum_{l\in r_2} f_l(\overline{\boldsymbol{v}}), \qquad \overline{\boldsymbol{v}} \in F. \tag{59}$$

The slope condition derives its name from the graph of the parametric LP, illustrated in Figure 3.1, whose solution yields the search direction $\boldsymbol{d}^{\tau} = \boldsymbol{w}^{\tau} - \boldsymbol{v}^{\tau}$. It states that the monetary cost associated with distinct routes are distinct, and can be enforced through a perturbation of the link costs, whenever the invariance condition holds.

Theorem 15 *Let the invariance and slope conditions hold. If $\boldsymbol{t}$ is monotone over F, then the link-flow solution $\boldsymbol{v}^*$ of the multi-attribute problem is unique, almost everywhere. Moreover, the path-flow solution is unique as well.* □

The above results are surprising, as both conclusions may fail to hold for the standard model. Indeed, the first conclusion usually requires a strict monotonicity condition while the second fails if there exist more than one path flow vector compatible with a given link-flow solution. The situation is altogether different in the multi-attribute model, where the slope condition suffices to insure the uniqueness of the solution to a parametric LP. At equilibrium, every α-group is assigned to a unique path. While such a solution is extremal in the infinite-dimensional setting, the corresponding total flow vector will in all likelihood not coincide with an extreme point of the polyhedron F. Another consequence of the slope condition is that the search direction $\boldsymbol{d}^\tau$ induces a feasible but non-extremal total flow direction, relative to the polyhedron F. It is superior to an ordinary Frank–Wolfe direction in the sense that it is a descent direction for the gap function and induces convergence at a linear rate!

THEOREM 16 *Let the invariance and slope conditions hold. Let $\boldsymbol{t}$ be strongly monotone (with modulus b) and Lipschitz continuous (with modulus L) over F. If the step size if fixed and satisfies $0 < \lambda_\tau \equiv \lambda <$ minimum $\{1, 2b/L\}$, then the cost approximation algorithm converges geometrically to the equilibrium solution. More precisely:*

$$\begin{aligned} g(\boldsymbol{v}^{\tau+1}) &\leq (1-\lambda) g(\boldsymbol{v}^\tau), \\ \|\overline{\boldsymbol{v}}^\tau - \overline{\boldsymbol{v}}^*\| &= \mathcal{O}((1-\lambda)^{\tau/2}), \\ \|\overline{\boldsymbol{w}}^\tau - \overline{\boldsymbol{w}}^*\| &= \mathcal{O}((1-\lambda)^{\tau/2}) \end{aligned}$$

hold. □

The last equality in the theorem states that the total flow vector $\overline{\boldsymbol{w}}^\tau$ corresponding to the solution $\boldsymbol{w}^\tau$ of the parametric shortest paths problem converges to the optimal total-flow equilibrium. Although the asymptotic rate of convergence is similar to that of the sequence $\overline{\boldsymbol{v}}^\tau$, the actual convergence is much slower, as observed empirically on large scale problems.

One drawback of the fixed step size scheme is that its validity rests on parameters b and L that might be difficult to estimate. This can be fixed by adopting a step size rule that obeys the Armijo condition with respect to the primal gap function. Alternatively, if this solution is deemed to expensive computationally, the step size can be made to obey one of the following two rules:

$$\begin{aligned} &\text{RULE I:} \qquad 0 < \lambda_\tau < 1, \quad \lim_{\tau\to\infty} \lambda_\tau = 0, \quad \sum_{\tau=0}^{\infty} \lambda_\tau = +\infty; \\ &\text{RULE II:} \qquad \lambda_0 = 1, \quad \lambda_{\tau+1} = \lambda_\tau - (1/2)\lambda_\tau^2. \end{aligned}$$

In either case, global convergence is preserved, but asymptotic convergence may be sub-linear.

Implementation. The main challenge in implementing the cost approximation algorithm lies in the numerical resolution of the parametric shortest paths problem. One may of course solve this parametric LP for its optimal solution by a parametric network simplex procedure, but this would prove costly on large networks involving thousands of paths, many of them irrelevant, and many others only active within a critical interval (α_{i-1}, α_i) of insignificant width. An alternative is to replace the density function h by a suitable discrete approximation. In practice, a 20-bar histogram yields results that are virtually indistinguishable from the exact solution. Might we then argue that the running time of a multi-attribute algorithm is roughly 20 times that of a standard model of similar size? No. We claim that the actual ratio is much smaller, as the convergence rate partially offsets the number of discrete problems (20) to be solved. Finally, one might ask what is the advantage of this approach over discretizing h *before* linearizing the functions $\boldsymbol{t}$ and $\boldsymbol{f}$, a dilemma reminiscent of the Jacobi–Newton vs. Newton-Jacobi paradigm. Our answer is twofold. First, it is easier to monitor the convergence process through the control of a well-defined combinatorial subproblem. Second, let us not be shy about it: the analysis has some mathematical elegance.

5.2 Side constraints

The derivation of the side constrained traffic equilibrium model and its equilibrium characterization suggests the following algorithmic construct: guess a value of the multiplier vector $\boldsymbol{\beta}$, solve a standard traffic equilibrium model in terms of generalized link costs, and update the value of $\boldsymbol{\beta}$ based on a convergent algorithm for solving the Lagrangean dual problem. If the number of side constraints is relatively small, then this approach can be extremely efficient, since the subproblems are standard traffic equilibrium problems and the dual space has a small dimension.

Provided that the link travel cost functions are strictly increasing, the Lagrangean is strictly convex Numerical experiments show however that, whenever an *augmented* Lagrangean algorithm is used for a link capacitated problem, convergence to a near-optimal, near-feasible solution is reached even more quickly. Starting from a near-optimal and near-feasible solution, it is then relatively straightforward to generate a *primal feasible* and near-optimal solution by means of specialized graph search techniques working in a residual graph.

5.3 Strategic model

In this section, we develop a strategy for solving the basic model. These ideas can be extended to the dynamic and priority models. The challenge in finding an equilibrium solution is threefold:

(i) The cost mapping $\boldsymbol{c}$ is not available in closed form.

(ii) $\boldsymbol{c}$ is neither differentiable nor monotone.

(iii) The entire information pertaining to the strategic flow x must be preserved at every iteration.

The second issue implies that only heuristic methods based on first order information are implementable. One such proposal is based on the solution of a linear program; at iteration $\tau + 1$, one sets

$$\boldsymbol{x}^{\tau+1} = \boldsymbol{x}^{\tau} + \lambda_\tau(\boldsymbol{y}^{\tau} - \boldsymbol{x}^{\tau}),$$

where $\lambda_\tau \in [0, 1]$ and

$$\boldsymbol{y}^{\tau} \in \arg \underset{\boldsymbol{y}\in X}{\text{minimum}}\ \boldsymbol{c}(\boldsymbol{x}^{\tau})^{\mathrm{T}}\boldsymbol{y},$$

the vector of best strategic responses to $\boldsymbol{x}^{\tau}$, is obtained by solving Bellman's dynamic programming equations, for every destination node q:

$$\omega_j^{s^*} = \begin{cases} \infty & \text{if } j > q, \\ 0 & \text{if } j = q, \\ \sum_{k\in E_j^{s^*}} \pi_{jk}^{s^*}(d_{jk} + \omega_k^{s^*}) & \text{if } j < q, \end{cases} \tag{60}$$

and assigning all strategic flow $\boldsymbol{y}^{\tau}$ to strategy s^*. In the above expression, $E_j^{s^*}$ denotes the preference order of strategy s^*, $\pi_{jk}^{s^*}$ is the probability of accessing node k from node j, using s^*, t_{jk} is the traversal time of link (j,k) and $\omega_k^{s^*}$ is the cost-to-go function. Since the vector $\boldsymbol{\pi}$ is a by-product of the loading process[20], the recursion will be well-defined once the preference orders are set. At a given node, one may of course consider all possible such orders. However, under the FIFO queue discipline, one can prove the optimality of the following greedy approach:

$$E_j^{s^*} \in \arg \underset{E_j^s}{\text{minimum}} \sum_{k\in E_j^s} \pi_{jk}^{s}(t_{jk} + \omega_k^{s^*}). \tag{61}$$

The time required to compute $\boldsymbol{c}(\boldsymbol{x})$ favors the use of a preselected sequence of step sizes, such as the harmonic sequence $\lambda_\tau = 1/\tau$. An alternative which proves efficient in practice is to weight the harmonic step size $1/\tau$ by the nonnegative vector $\boldsymbol{c}(\boldsymbol{x}^{\tau}) - \boldsymbol{c}(\boldsymbol{y}^{\tau})$, which is a measure of departure from equilibrium or, in layperson terms, a measure of user dissatisfaction. The third point is best addressed by resorting to restriction (simplicial decomposition) techniques.

5.4 Non-additive route costs

Algorithms for non-additive traffic equilibrium models need explicit route information; a simplicial decomposition/route (column) generation algorithm à la DSD therefore seems natural. In the case of the above cost model, the only place where some complications arise is in the generation of new, profitable routes. The Wardrop conditions suggest the following extension of the shortest route problem, for an OD pair $k \in \mathcal{C}$, at some fixed link volume $\bar{\boldsymbol{v}}$:

$$\underset{r\in\mathcal{R}_k}{\text{minimize}} \sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\bar{\boldsymbol{v}}) + \alpha_r(f_r).$$

[20] This result also holds for strategies that have not been considered yet, i.e., strategies that carry no flow.

If the current solution is not in equilibrium, it can be improved through the reallocation of flow onto shortest routes, exactly as in Franc-Wolfe's linearization method. Unfortunately, the above shortest route problem cannot be solved using Dijkstra's algorithm, since each link possesses two attributes: travel time and money outlay; in particular, the dynamic programming principle—that sub-routes are optimal if the entire routes are—does not apply. Instead, more complex multi-label shortest route methods must be used.

Suppose that $f_r = \sum_{l \in \mathcal{L}} \lambda_{lr} f_{lr}$, so that the monetary outlay on the route is the sum of link tolls. Then we can devise an algorithm in which node labels incorporate both accumulated travel times and monetary outlays, from a specific node on to the destination nodes. Domination tests (based on the principle of Pareto optimality) are used to keep down the number of labels stored, and those remaining at the terminal node are compared with respect to the function ν_k to determine the best route. The corresponding extension of the DSD algorithm works well for this problem.

Bibliographical notes

Algorithms for the bi-attribute model have been analyzed in [MaZ97]. An implementation based on the parametric network simplex method is described in [MNT96] in the congestion-free case.

The side constrained model discussed in Section 5.2 had previously been considered algorithmically mainly for the link capacitated case, typically through penalty methods. The algorithm in [LaP95] (whose literature section traces the history of such methods within the context of transportation science) combines an augmented Lagrangean method with the DSD algorithm used for the subproblem. One of the few algorithms for the general side constrained model is found in [LPR04], which combines Lagrangean relaxation and column generation.

The implementation of the strategic model is discussed in [MNS04], while numerical results on the non-additive model are found in [Lar+02].

6 Optimization in a user equilibrium context

Equilibrium models are by construction descriptive, passive mathematical objects that can be used to assess the impact of managerial policies. Ideally, a design model should be able to integrate equilibrium equations within the overall control process. In this section, we address the design problem both from the local and global perspectives. In this respects, the first subsection provides sensitivity results for the traffic equilibrium problem, while the next two subsections are concerned with the problem of imposing tolls either for controlling flow patterns through the network, or for raising revenues.

6.1 Stability and sensitivity of traffic equilibria

Although Braess' paradox, which was discovered in the late 1960s, prompted a study into the qualitative behaviour of a traffic equilibrium under varying conditions, a thorough quantitative study, in the form of a variational and differential analysis, is much more recent. This section traces the main variational characteristics of a traffic equilibrium, and shows when the equilibrium is (directionally) differentiable as a function of input data; the results cited improve quite substantially on previous analyses that relied heavily on the Implicit Function Theorem and are therefore only applicable in very restrictive settings.

6.1.1 Introduction

The basis of our sensitivity analysis is a result which is stated for a general variational inequality problem over a polyhedron and with a differentiable mapping, $\boldsymbol{f} : \Re^d \times \Re^n \to \Re^n$ in the parameters $\boldsymbol{\rho} \in \Re^d$ and variables $\boldsymbol{x} \in \Re^n$: find $\boldsymbol{x}^* \in X$ such that

$$\boldsymbol{f}(\boldsymbol{\rho}, \boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X, \tag{62}$$

where $X \subseteq \Re^n$ is a polyhedral set.[21] We let $\boldsymbol{S} : \Re^d \rightrightarrows \Re^d$ denote the mapping that assigns to each vector $\boldsymbol{\rho} \in \Re^d$ the set $\boldsymbol{S}(\boldsymbol{\rho})$ of solutions to this problem. Letting $\boldsymbol{\rho} = \boldsymbol{\rho}^*$ be the current value of the parameter vector, we are interested in the direction of change of the solution $\boldsymbol{x}^*$ as $\boldsymbol{\rho}^*$ is perturbed along a direction

[21] That the set X is polyhedral is crucial. It is however possible to take into account nonlinear constraints by reformulating the problem such that Lagrange multipliers are explicitly used.

$\boldsymbol{\rho}'$. This directional derivative of $\boldsymbol{S}$ is the solution to an auxiliary variational inequality, which has the following form: find $\boldsymbol{x}' \in K$ such that

$$\boldsymbol{r}(\boldsymbol{\rho}', \boldsymbol{x}')^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}') \geq 0, \qquad \boldsymbol{x} \in K, \tag{63a}$$

where

$$K := T_X(\boldsymbol{x}^*) \cap \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*)^{\perp}, \tag{63b}$$

and

$$\boldsymbol{r}(\boldsymbol{\rho}', \boldsymbol{x}') := \nabla_{\boldsymbol{\rho}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*)\boldsymbol{\rho}' + \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*)\boldsymbol{x}'. \tag{63c}$$

We let $\boldsymbol{DS}(\boldsymbol{\rho}^*|\boldsymbol{x}^*) : \Re^d \rightrightarrows \Re^d$ denote the mapping that assigns to each perturbation $\boldsymbol{\rho} \in \Re^d$ the set $\boldsymbol{DS}(\boldsymbol{\rho}^*|\boldsymbol{x}^*)(\boldsymbol{\rho}')$ of solutions to this problem. The set K above is referred to as the *critical cone*, and denotes the set of variations around $\boldsymbol{x}^*$ that, roughly speaking, retains feasibility and optimality to the first order. T_X denotes the tangent cone to X, which means that if X is defined by linear constraints, we have that

$$X = \{\, \boldsymbol{x} \in \Re^n |\, \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b};\ \boldsymbol{B}\boldsymbol{x} = \boldsymbol{d} \} \quad \Longrightarrow \quad T_X(\boldsymbol{x}^*) = \left\{\, \boldsymbol{z} \in \Re^n |\, \bar{\boldsymbol{A}}\boldsymbol{z} \geq \boldsymbol{0};\ \boldsymbol{B}\boldsymbol{z} = \boldsymbol{0} \right\},$$

where $\bar{\boldsymbol{A}}$ consists of the rows $\boldsymbol{A}_i$ of $\boldsymbol{A}$ corresponding to the binding inequality constraints at $\boldsymbol{x}^*$, that is, the indices i with $\boldsymbol{A}_i\boldsymbol{x}^* = b_i$. Further, for any vector $\boldsymbol{z} \in \Re^n$, $\boldsymbol{z}^{\perp} := \{\, \boldsymbol{y} \in \Re^n \mid \boldsymbol{z}^{\mathrm{T}}\boldsymbol{y} = 0 \,\}$ is the orthogonal subspace associated with the vector $\boldsymbol{z}$. The mapping $\boldsymbol{r}$ is a linearization of $\boldsymbol{f}$ around $(\boldsymbol{\rho}^*, \boldsymbol{x}^*)$; it is an affine mapping in $\boldsymbol{x}'$.

Suppose now that $\boldsymbol{f}(\boldsymbol{\rho}, \cdot)$ is monotone on X around $\boldsymbol{\rho} = \boldsymbol{\rho}^*$, and that the parameterization is such that the rank of the matrix $\nabla_{\boldsymbol{\rho}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*)$ is equal to n. (The latter result can always be fulfilled by including dummy parameters.) We say that the mapping $\boldsymbol{S}$ is *strongly regular* at $\boldsymbol{\rho}^*$ if $\boldsymbol{S}$ is single-valued and Lipschitz continuous on some neighborhood of $\boldsymbol{\rho}^*$. Then,

$$\boldsymbol{S} \text{ is strongly regular at } \boldsymbol{\rho}^* \tag{64a}$$

$$\Longleftrightarrow$$

$$\boldsymbol{DS}(\boldsymbol{\rho}^*|\boldsymbol{x}^*) \text{ is single-valued.} \tag{64b}$$

Moreover, the unique solution $\boldsymbol{x}' \in \Re^n$ to (63) is the *directional derivative* of the solution $\boldsymbol{x}^*$ to (62) at $\boldsymbol{\rho}^*$, in the direction of $\boldsymbol{\rho}'$. A sufficient condition for the property (64b) to hold is that

$$\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*) \text{ is positive definite on } (K - K). \tag{65}$$

We refer to this as a sufficient *second-order* condition. A result stronger than directional differentiability can also be obtained under additional assumptions. Indeed, under strong regularity,

$$\boldsymbol{S} \text{ is differentiable at } \boldsymbol{\rho}^* \tag{66a}$$

$$\Longleftrightarrow$$

$$\boldsymbol{DS}(\boldsymbol{\rho}^* \mid \boldsymbol{x}^*)(\boldsymbol{\rho}') \in -K, \qquad \boldsymbol{\rho}' \in \Re^d. \tag{66b}$$

Moreover, if the critical cone K is a subspace, that is, if $K = K \cap (-K)$, then the gradient can be represented as

$$\nabla_{\boldsymbol{\rho}} \boldsymbol{x}(\boldsymbol{\rho}^*) = -\boldsymbol{Z} \left[\boldsymbol{Z}^{\mathrm{T}} \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*) \boldsymbol{Z} \right]^{-1} \boldsymbol{Z}^{\mathrm{T}} \nabla_{\boldsymbol{\rho}} \boldsymbol{f}(\boldsymbol{\rho}^*, \boldsymbol{x}^*), \tag{67}$$

for any $n \times \ell$ matrix $\boldsymbol{Z}$ such that $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ is nonsingular and $\boldsymbol{z} \in K \cap (-K)$ if and only if $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{y}$ for some $\boldsymbol{y} \in \Re^{\ell}$, where ℓ is the dimension of $K \cap (-K)$. This differentiability result is a kind of implicit function theorem; the relationships in (64) and (66) show how the implicit function theorem naturally extends to more general cases.

The latter property has been invoked in many sensitivity analysis frameworks. That is, they hinge on being able to invoke the implicit function theorem. Unfortunately, not only does the property $\boldsymbol{DS}(\boldsymbol{\rho}^* \mid \boldsymbol{x}^*)(\boldsymbol{\rho}') \in -K$ fail to hold in many cases (cf. below), but also there may not exist a nonsingular matrix of the kind referred to above.

6.1.2 Sensitivity analysis of separable traffic equilibrium

For simplicity of presentation, we focus on the case of separable cost and demand functions. First, we cast the VIP (8) in the sensitivity framework. Let

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{h} \\ \boldsymbol{v} \\ \boldsymbol{d} \end{pmatrix}; \qquad \boldsymbol{f}(\boldsymbol{\rho}, \boldsymbol{x}) \begin{pmatrix} \boldsymbol{0}^{|\mathcal{R}|} \\ \boldsymbol{t}(\boldsymbol{\rho}, \boldsymbol{v}) \\ -\boldsymbol{\xi}(\boldsymbol{\rho}, \boldsymbol{d}) \end{pmatrix}; \qquad X = \Re_+^{|\mathcal{R}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|}.$$

Then, we can identify the sensitivity problem through the following identifications:

$$K = \left\{ \begin{pmatrix} \boldsymbol{h}' \\ \boldsymbol{v}' \\ \boldsymbol{d}' \end{pmatrix} \in \Re^{|\mathcal{R}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \;\middle|\; \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{h}' = \boldsymbol{d}'; \quad \boldsymbol{v}' = \boldsymbol{\Lambda} \boldsymbol{h}'; \quad \boldsymbol{h}' \in H' \right\},$$

where

$$H' = \left\{ \boldsymbol{h}' \in \Re^{|\mathcal{R}|} \;\middle|\; \begin{array}{l} h'_r \text{ free if } h^*_r > 0 \\ h'_r \geq 0 \text{ if } h^*_r = 0 \text{ and } c_r(\boldsymbol{\rho}^*, \boldsymbol{h}^*) = \pi^*_k \\ h'_r = 0 \text{ if } h^*_r = 0 \text{ and } c_r(\boldsymbol{\rho}^*, \boldsymbol{h}^*) > \pi^*_k \\ [r \in \mathcal{R}_k, \ k \in \mathcal{C}] \end{array} \right\},$$

and

$$\boldsymbol{r}(\boldsymbol{\rho}', \boldsymbol{x}') = \begin{pmatrix} \boldsymbol{0}^{|\mathcal{R}|} \\ \nabla_{\boldsymbol{\rho}} \boldsymbol{t}(\boldsymbol{\rho}^*, \boldsymbol{v}^*)\boldsymbol{\rho}' + \nabla_{\boldsymbol{v}} \boldsymbol{t}(\boldsymbol{\rho}^*, \boldsymbol{v}^*)\boldsymbol{v}' \\ -[\nabla_{\boldsymbol{\rho}} \boldsymbol{\xi}(\boldsymbol{\rho}^*, \boldsymbol{d}^*)\boldsymbol{\rho}' + \nabla_{\boldsymbol{d}} \boldsymbol{\xi}(\boldsymbol{\rho}^*, \boldsymbol{d}^*)\boldsymbol{d}'] \end{pmatrix}.$$

By the monotonicity and separability of $\boldsymbol{t}$ and $-\boldsymbol{\xi}$, the resulting 'sensitivity variational inequality' (63) can be equivalently written as the convex quadratic optimization problem to

$$\begin{aligned} \underset{(\boldsymbol{v}', \boldsymbol{d}')}{\text{minimize}} \quad & \phi'(\boldsymbol{v}', \boldsymbol{d}') := [\nabla_{\boldsymbol{\rho}} \boldsymbol{t}(\boldsymbol{\rho}^*, \boldsymbol{v}^*)\boldsymbol{\rho}']^{\mathrm{T}} \boldsymbol{v}' + \frac{1}{2} \sum_{l \in \mathcal{L}} \frac{\partial t_l(\boldsymbol{\rho}^*, v^*_l)}{\partial v_l} (v'_l)^2 \\ & \qquad - [\nabla_{\boldsymbol{\rho}} \boldsymbol{\xi}(\boldsymbol{\rho}^*, \boldsymbol{d}^*)\boldsymbol{\rho}']^{\mathrm{T}} \boldsymbol{d}' - \frac{1}{2} \sum_{k \in \mathcal{C}} \frac{\partial \xi_k(\boldsymbol{\rho}^*, d^*_k)}{\partial d_k} (d'_k)^2, && \text{(68a)} \\ \text{subject to} \quad & \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{h}' = \boldsymbol{d}', && \text{(68b)} \\ & \boldsymbol{v}' = \boldsymbol{\Lambda} \boldsymbol{h}', && \text{(68c)} \\ & \boldsymbol{h} \in H'. && \text{(68d)} \end{aligned}$$

The sensitivity problem is closely related to the original model, with two notable differences: the link cost and demand functions are replaced by their linearizations, and the sign restrictions on $\boldsymbol{h}$ are replaced by individual restrictions on the route flow perturbations h'_r that depend on whether the route in question was used at equilibrium or not, cf. the set H'. Although the appearance of H' depends on the choice of route flow solution $\boldsymbol{h}^*$, it is an interesting fact that the possible choices of $\boldsymbol{v}'$ in K do *not*; this is a general consequence of aggregation, and amounts to the possibility of (essentially) eliminating route flow variables from explicit consideration; this is only possible because of the special connection between route and link flow variables. Finally, we see that the resemblance to the original problem implies that the sensitivity variational inequality problem can be solved using software similar to those for the original traffic equilibrium model, provided of course that route flow information can be extracted.

Before we apply the sensitivity analysis results of the previous section to the present problem, we mention an important fact which allows us to provide slightly stronger results for optimization formulations than for general variational inequality models: for a differentiable, convex problem, the gradient of the objective function is invariant over the solution set (cf. the discussion following the proof of Theorem 5).

Assumption 17 (properties of the network model).

(a) For each $l \in \mathcal{L}$, the link travel cost function $t_l(\cdot, \cdot)$ is continuously differentiable, and strictly increasing in its second argument.

(b) For each $k \in \mathcal{C}$, the demand function $g_k(\cdot, \cdot)$ is continuously differentiable, non-negative, upper bounded, and strictly decreasing in its second argument. The function $g_k(\rho, \cdot)$ is therefore invertible, and has a single-valued inverse, $\xi_k(\rho, \cdot)$, which also is continuously differentiable and strictly decreasing. □

Theorem 18 (sensitivity of separable traffic equilibrium problems). *Let Assumption 17 hold, and consider an arbitrary vector $\boldsymbol{\rho}^* \in \Re^d$. Then, the solution $(\boldsymbol{v}^*, \boldsymbol{d}^*)$ to (16) is unique. Assume that the link travel cost function $\boldsymbol{t}(\boldsymbol{\rho}^*, \cdot)$ is such that*

$$\frac{\partial t_l(\boldsymbol{\rho}^*, v_l^*)}{\partial v_l} > 0, \qquad l \in \mathcal{L}. \tag{69}$$

Assume further that the demand function $\boldsymbol{g}(\boldsymbol{\rho}^*, \cdot)$ is such that[22]

$$\frac{\partial g_k(\boldsymbol{\rho}^*, \pi_k^*)}{\partial \pi_k} < 0, \qquad k \in \mathcal{C}. \tag{70}$$

Then, in the solution to (68), the values of the link flow and demand perturbation $\boldsymbol{v}'$ and $\boldsymbol{d}'$ are unique; therefore, the value $\boldsymbol{v}'$ (respectively, $\boldsymbol{d}'$) is the directional derivative of the equilibrium link flow (respectively, demand), at $\boldsymbol{\rho}^*$, in the direction of $\boldsymbol{\rho}'$. ∎

6.1.3 Limitations of the classical sensitivity framework

Sensitivity analyses that extend standard techniques for nonlinear optimization and variational inequality problems are widespread, and have been adopted by several researchers over the past fifteen years, particularly as a subroutine in the solution of more complex problems of a bilevel nature (see, for example, Section 6.4). These analyses have however a drawback: they often are inapplicable because of the strong assumptions underlying their validity. Part of the problem lies in the utilization of the implicit function theorem, which requires the problem to be expressed as a system of equations; this leads to a strict complementarity type of condition, which may fail to be in force at differentiable points. Additional requirements pose, for example, conditions on the topology of the network itself, and are not necessary either. It appears that some researchers are not aware of these limitations, and of the existence of more widely applicable analyses. This justifies devoting some space to examples where this 'classical' sensitivity framework fails, while the techniques of the previous section still apply.

Example 19 (a non-strictly complementary example) The classical analysis requires that the solution be strictly complementary. In our context this means that there must exist some equilibrium route flow solution $\boldsymbol{h}^*$ which has the property that for every route in $\mathcal{R}$, $h_r^* > 0$ holds if and only if route r is a shortest route given the equilibrium travel costs. The definition in the classical work is however based on the total link flows and an aggregated 'node price', common to all OD pairs, which may not exist (see, for example, the numerical example in Section 2).

To show that strict complementarity is not necessary for differentiability (as we have already seen from the above analysis), we consider the network depicted in Figure 11. It involves two OD pairs, $(1, 2)$ and $(4, 2)$, with a fixed and unperturbed demand of 2 and 1 units of flow, respectively. The link cost functions are given by

$$t_1(v_1, \rho) := 2v_1 + \rho;\ t_2(v_2) := v_2;\ t_3(v_3) := 1;\ t_4(v_4) := v_4 + 2;\ t_5(v_5) = v_5.$$

We have four routes: $\{1\}$, $\{2, 3\}$, $\{4\}$, and $\{5, 3\}$, two for each OD pair.

If $\rho^* = 0$, the unperturbed traffic equilibrium solution is $\boldsymbol{v}^* = (1, 1, 1, 1, 1)^{\mathrm{T}}$ and the route flow $\boldsymbol{h}^* = (1, 1, 0, 1)^{\mathrm{T}}$ is unique. We observe that the travel cost on route 3 is equal to 2, as is the case for route 4, so this equilibrium solution is not strictly complementary.

In order to check whether the solution $\boldsymbol{v}^*$ is nevertheless differentiable at $\rho^* = 0$, we solve the sensitivity problem for both $\rho' := 1$ and $\rho' := -1$. For $\rho' = 1$, we obtain the following unique solution to the sensitivity problem, thus being the directional derivative of $\boldsymbol{v}^*$ with respect to the direction $\rho' = 1$ at $\rho^* = 0$: $\boldsymbol{v}' = (-\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0)^{\mathrm{T}}$. The effect of perturbing link 1's cost, such that it becomes more expensive, is that of sending flow along the cycle $\{-1, 2, 3\}$, where the minus sign reflects that flow is sent backwards on link 1. When solving the sensitivity problem for $\rho' := -1$, we obtain the directional derivative $\boldsymbol{v}' = (\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 0, 0)^{\mathrm{T}}$, that is, the negative of the directional derivative of $\boldsymbol{v}^*$ in the direction of $\rho' := 1$. This proves that the directional derivative mapping is linear, and thus that the derivative of $\boldsymbol{v}^*$ with respect to ρ' at $\rho^* = 0$ equals $\mathrm{d}\boldsymbol{v}^*/\mathrm{d}\rho = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0)^{\mathrm{T}}$. ∎

[22] The two derivative conditions (69) and (70) imply that the functions $t_l(\boldsymbol{\rho}^*, \cdot)$ and $-g_k(\boldsymbol{\rho}^*, \cdot)$ are strictly increasing.

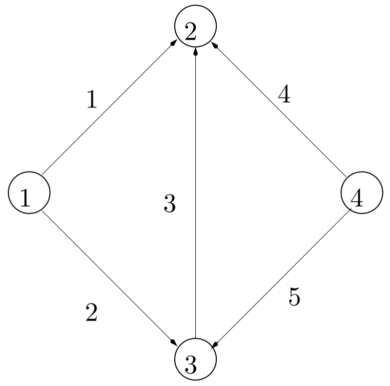


Figure 11: Network for the first example.

Example 20 (second example) A key assumption of the classical analysis is that the travel cost function $\boldsymbol{t}(\boldsymbol{\rho}, \cdot)$ be strongly monotone in a neighborhood of the equilibrium solution. This is clearly not needed, as we have shown above.

More serious, however, is the presence of a condition on the topology of the graph representing the traffic network. Suppose that we were to limit the discussion to a subgraph of $\mathcal{G}$ in which only the links carrying positive flow are included. In this subnetwork, the analysis must be performed from an equilibrium route flow solution that is an extreme point of H (the routes using the above-stricken zero-flow links also having been stricken) which has exactly as many routes with positive flow as the rank of the matrix $[\boldsymbol{\Lambda}_+^{\mathrm{T}} \mid \boldsymbol{\Gamma}_+]$ (the + sign indicates that we have eliminated the zero-flow routes, as discussed). The rank of this matrix is never larger than the number of links with positive flow at $\boldsymbol{v}^*$ plus $|\mathcal{C}|$. Although the choice of the route flow solution is immaterial, it must be extremal, as explained above. The resulting formula then takes the form

$$\begin{pmatrix} \nabla_{\boldsymbol{\rho}}\, \boldsymbol{h}_+ \\ \nabla_{\boldsymbol{\rho}}\, \boldsymbol{\pi} \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{h}}\, \boldsymbol{c}_+(\boldsymbol{\rho}^*, \boldsymbol{h}_+^*) & -\boldsymbol{\Lambda}_+^{\mathrm{T}} \\ \boldsymbol{\Lambda}_+ & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} -\nabla_{\boldsymbol{\rho}}\, \boldsymbol{c}_+(\boldsymbol{\rho}^*, \boldsymbol{h}_+^*) \\ \nabla_{\boldsymbol{\rho}}\, \boldsymbol{g}(\boldsymbol{\rho}^*) \end{pmatrix}. \tag{71}$$

Unfortunately, it is not difficult to construct examples where the solution is differentiable but where the matrix condition fails, so that the formula (71) breaks down. Consider the network shown in Figure 12, involving a single OD pair, $(1, 3)$, and a fixed demand of 2 flow units. Set the link cost functions to

$$t_1(v_1, \rho) = v_1 + \rho; \quad t_2(v_2) = v_2; \quad t_3(v_3) = v_3; \quad t_4(v_4) = v_4.$$

In this example, we have four routes: $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, and $\{2, 4\}$.

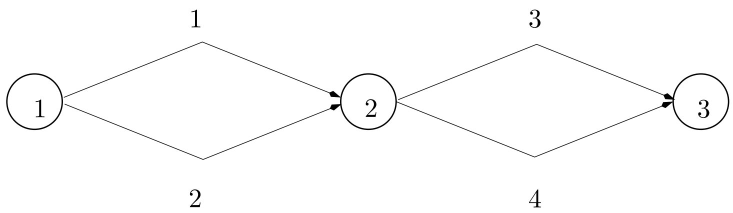


Figure 12: Network for the second example.

If $\rho^* = 0$, the unperturbed traffic equilibrium solution is $\boldsymbol{v}^* = (1, 1, 1, 1)^{\mathrm{T}}$. We can easily see that the solution is differentiable, and even strictly complementary. The derivative with respect to ρ at ρ^* is $(-\frac{1}{2}, 0, \frac{1}{2}, 0)^{\mathrm{T}}$. This result is intuitive: if the value of ρ increases, then the flow on link 1 should decrease, whence the flow on link 2 must increase by the same amount. If, on the other hand, the value of ρ decreases, the reverse should happen.

Consider then the workings of the classical formula (71). We obviously fulfill the strong monotonicity conditions on the travel cost function. Since all links carry flow at equilibrium, we need not remove any links or routes when considering the sensitivity analysis problem. We last try to comply with the linear independence condition, by choosing the right equilibrium route flow solution. Note then that

$$[\boldsymbol{\Lambda}^{\mathrm{T}} \mid \boldsymbol{\Gamma}] = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix},$$

which has rank 3. So, we should find a route flow solution, $\boldsymbol{h}^*$, in which exactly 3 routes have a positive flow. This is however impossible; the only alternatives are 2 or 4. To see why, let us suppose that the flow on the first route, $\{1,3\}$, is $\alpha \in [0,1]$. Then, the flows on routes $\{1,4\}$ and $\{2,3\}$ must both be $1-\alpha$, in order to comply with the total flow on the links. This implies that the flow on route $\{2,4\}$ is α and that, for any value of $\alpha \in [0,1]$, the number of routes carrying non-zero flow is either 2 or 4. We can therefore not comply with the matrix condition stated, and the classical formula (71) fails, even though the gradient does exist. □

Note that the technique of Section 6.1.2 carries over to the asymmetric case. This approach yields the recommended source for sensitivity information, not only because it is widely applicable, but also because the sensitivity analysis problem, being closely related to that of the original equilibrium problem, can be solved using similar computational tools. This is especially important when relying on sensitivity information for an extended traffic model, for example of the bilevel type, where many such analyses must be performed (cf. Section 6.4). The classical formula may then fail, either because of the topology being 'wrong' or for lack of strict complementarity (and the latter *will* typically be the case at a stationary point of a bilevel program); in either case, the formula breaks down, regardless of the existence, or not, of a gradient. It is common to claim that the possibility of the non-differentiability of a given point in the space of $\boldsymbol{\rho}$ can be ignored because of Rademacher's Theorem, which essentially states that a locally Lipschitz continuous function is differentiable everywhere except possibly on a set with a zero (Lebesgue) measure. This type of argument ignores that several of the most interesting points to look at *are* points of the latter category; for example, optimal solutions to bilevel programs and, more generally, MPECs (Mathematical Program with Equilibrium Constraints) problems, are indeed extremal.[23]

6.2 Inducing system-optimal flows through link tolls

Tolls, either physical or virtual, may be used not only to raise money but also to alter user behavior in order to improve the performance of a transportation system with respect to indicators such as travel time or pollution. In this regard, let $\boldsymbol{v}_{\mathrm{SO}}$ be a globally optimal solution to the mathematical program

$$\underset{\boldsymbol{v} \in F}{\operatorname{minimize}} \; \boldsymbol{t}(\boldsymbol{v})^{\mathrm{T}} \boldsymbol{v}, \tag{72}$$

and let us first consider the problem of setting tolls that induce system-optimal flows $\boldsymbol{v}_{\mathrm{SO}}$, while being compatible with the selfish behavior of users.[24] Then any toll vector $\boldsymbol{\rho}$ that satisfies

$$-[\boldsymbol{t}(\boldsymbol{v}_{\mathrm{SO}}) + \boldsymbol{\rho}] \in N_F(\boldsymbol{v}_{\mathrm{SO}}) \tag{73}$$

clearly satisfies our requirements. We say that Equation (73) describes an *inverse optimization problem* where, given an optimal solution to a mathematical program, one looks for a vector of parameters ($\boldsymbol{\rho}$ in the above) that reproduces the given solution. To pursue the analysis, we assume that the polyhedron F takes the form $F = \{\, \boldsymbol{v} \mid \boldsymbol{B}\boldsymbol{v} = \boldsymbol{b}, \boldsymbol{v} \geq \boldsymbol{0} \,\}$, so that one can rewrite the toll equation (73) in the Karush–Kuhn–Tucker format

$$\boldsymbol{t}(\boldsymbol{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\pi} \geq \boldsymbol{0}, \tag{74a}$$

$$\boldsymbol{v}_{\mathrm{SO}}{}^{\mathrm{T}} [\boldsymbol{t}(\boldsymbol{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - B^{\mathrm{T}} \boldsymbol{\pi}] = \boldsymbol{0}, \tag{74b}$$

$$\boldsymbol{v}_{\mathrm{SO}} \in F. \tag{74c}$$

[23] The reader may draw a piece-wise linear convex function in $\Re$ and see for herself that if there exists a global minimum, then there exists also a 'non-differentiable' global minimum.

[24] This is is a particular instance of (43) with $\boldsymbol{v}^* = \boldsymbol{v}_{\mathrm{SO}}$. For notational simplicity, we have adopted a link flow formulation of the equilibrium problem.

Because $\boldsymbol{v}_{\rm SO}$ is known and feasible, we obtain that the solution set of (74) associated with the system-optimal flow $\boldsymbol{v}_{\rm SO}$ is the polyhedron P defined as

$$\begin{aligned} &\boldsymbol{\rho} \geq \boldsymbol{B}^{\rm T}\boldsymbol{\pi} - t(\boldsymbol{v}_{\rm SO}), \\ &\boldsymbol{\rho}_i = [\boldsymbol{B}^{\rm T}\boldsymbol{\pi} - t(\boldsymbol{v}_{\rm SO})]_i \quad \text{if} \quad (\boldsymbol{v}_{\rm SO})_i > 0. \end{aligned}$$

Since $\boldsymbol{v}_{\rm SO}$ is solution to the variational inequality

$$-[\boldsymbol{t}(\boldsymbol{v}_{\rm SO}) + \boldsymbol{t}'(\boldsymbol{v}_{\rm SO})\boldsymbol{v}_{\rm SO}] \in N_F(\boldsymbol{v}), \tag{75}$$

the marginal cost solution $\boldsymbol{\rho} = \boldsymbol{t}'(\boldsymbol{v}_{\rm SO})\boldsymbol{v}_{\rm SO}$ is an element of P. If the system-optimal solution is not unique in link-flow space, then the set P must be enlarged to include all polyhedra associated with system-optimal flows. The resulting union of polyhedra is, in the general situation, not convex.

Under suitable assumptions, such as strict monotonicity of the cost mapping $\boldsymbol{t}$ over the set F, $\boldsymbol{v}_{\rm SO}$ is uniquely defined. However, the set P contains several elements, which suggests optimizing over a secondary criterion. For technical or political reasons, one may wish to minimize the total revenue levied, and insist that link tolls be nonnegative. This yields the mathematical program

$$\begin{aligned} \underset{\boldsymbol{\rho}\geq \boldsymbol{0},\boldsymbol{\pi}}{\text{minimize}} \quad & \boldsymbol{\rho}^{\rm T}\boldsymbol{v}_{\rm SO}, && (76\text{a}) \\ \text{subject to} \quad & \boldsymbol{t}(\boldsymbol{v}_{\rm SO}) + \boldsymbol{\rho} - \boldsymbol{B}^{\rm T}\boldsymbol{\pi} \geq \boldsymbol{0}, && (76\text{b}) \\ & \boldsymbol{v}_{\rm SO}{}^{\rm T}(\boldsymbol{t}(\boldsymbol{v}_{\rm SO}) + \boldsymbol{\rho} - \boldsymbol{B}^{\rm T}\boldsymbol{\pi}) = 0. && (76\text{c}) \end{aligned}$$

Since both terms of the complementarity equation are nonnegative, this constraint can be replaced by the constraint (76c). Writing its dual linear program as that to

$$\begin{aligned} \underset{\alpha\geq 0,\boldsymbol{v}}{\text{maximize}} \quad & \boldsymbol{t}(\boldsymbol{v}_{\rm SO})^{\rm T}(\alpha\boldsymbol{v}_{\rm SO} - \boldsymbol{v}), \\ \text{subject to} \quad & \boldsymbol{v} - \alpha\boldsymbol{v}_{\rm SO} \leq \boldsymbol{v}_{\rm SO}, \\ & \boldsymbol{B}\boldsymbol{v} = \alpha\boldsymbol{b}, \end{aligned}$$

and performing the change of variable $\boldsymbol{v} = \alpha\boldsymbol{w}$ yields the parametric multi-commodity flow problem to

$$\begin{aligned} \underset{\alpha\geq 0,\boldsymbol{w}\in F}{\text{maximize}} \quad & \boldsymbol{t}(\boldsymbol{v}_{\rm SO})^{\rm T}(\boldsymbol{v}_{\rm SO} - \boldsymbol{w}), && (77) \\ \text{subject to} \quad & \boldsymbol{w} \leq (1 + \frac{1}{\alpha})\boldsymbol{v}_{\rm SO}, \end{aligned}$$

which is equivalent to $\text{maximize}_{\alpha\geq 0}\ \phi(\alpha)$, where

$$\begin{aligned} \phi(\alpha) := \underset{\boldsymbol{w}\in F}{\text{maximum}} \quad & \boldsymbol{t}(\boldsymbol{v}_{\rm SO})^{\rm T}(\boldsymbol{v}_{\rm SO} - \boldsymbol{w}), && (78) \\ \text{subject to} \quad & \boldsymbol{w} \leq (1 + \frac{1}{\alpha})\boldsymbol{v}_{\rm SO}. \end{aligned}$$

The scalar function ϕ has interesting properties: $\phi(0) = 0$, $\phi(\alpha)$ is nonnegative ($\boldsymbol{v}_{\rm SO}$ is a feasible solution), ϕ is piecewise linear, ϕ is bounded from above (the primal problem (76) is feasible and bounded), ϕ is concave, and ϕ is increasing.

The properties of the function ϕ imply that there exists a threshold value α^* such that $\phi(\alpha) = \phi(\alpha^*)$ whenever $\alpha \geq \alpha^*$. In other words, whenever α is sufficiently large, a *single* multi-commodity flow problem needs to be solved. The optimal tolls are then retrieved from the dual vector corresponding to the flow constraints $\boldsymbol{B}\boldsymbol{w} = \boldsymbol{b}$.

In a multi-attribute context, marginal cost pricing may not be applicable any more, as the resulting tolls will be different for distinct classes. It is therefore surprising that 'system-optimal' tolls that achieve the same goal do actually exist. Let us first analyze a two-attribute (time and money) problem where we look for a link toll vector that induces a system-optimal solution with respect to time[25]. As in Section 3.1, we denote by α_g the VOT parameter associated with the link-flow vector $\boldsymbol{v}^g$ of group $g \in \mathcal{G}$. Let

$$\boldsymbol{v}_{\rm SO} \in \arg \underset{\boldsymbol{v}\in F}{\text{minimum}}\ \boldsymbol{t}(\bar{\boldsymbol{v}})\ ^{\rm T}\bar{\boldsymbol{v}} + \sum_{g\in\mathcal{G}} \alpha_g \boldsymbol{f}(\bar{\boldsymbol{v}})^{\rm T}\boldsymbol{v}^g$$

[25]There also exist system-optimal solutions with respect to cost. One must be aware that they need not coincide.

denote the system-optimal solution, and $\bar{\boldsymbol{v}}_{\mathrm{SO}}$ the associated total link-flow vector. Consider the linear program

$$\begin{aligned} \underset{\boldsymbol{v}\in F}{\text{minimize}} \qquad & \boldsymbol{t}(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\bar{\boldsymbol{v}} + \sum_{g\in\mathcal{G}} \alpha_g \boldsymbol{f}(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\boldsymbol{v}^g, \\ \text{subject to} \qquad & \sum_{g\in\mathcal{G}} \boldsymbol{v}^g = \bar{\boldsymbol{v}}_{\mathrm{SO}}, \end{aligned}$$

and the optimal dual vector $\boldsymbol{\alpha}^*$ associated with its sole explicit constraint. Since its primal solution is also a solution to the Lagrangean problem

$$\begin{aligned} \underset{\boldsymbol{v}\in F}{\text{minimize}} \qquad & \boldsymbol{t}(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\bar{\boldsymbol{v}} + \sum_{g\in\mathcal{G}} \alpha_g \boldsymbol{f}(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\boldsymbol{v}^g + \sum_{g\in\mathcal{G}} \alpha_g^* \boldsymbol{v}^g, \\ = \quad & \sum_{g\in\mathcal{G}} [\boldsymbol{t}(\bar{\boldsymbol{v}}_{\mathrm{SO}}) + \alpha_g \boldsymbol{f}(\bar{\boldsymbol{v}}_{\mathrm{SO}}) + \boldsymbol{\alpha}^*]^{\mathrm{T}}\boldsymbol{v}^g, \end{aligned}$$

it satisfies the Wardrop equilibrium principle for every individual class $g \in \mathcal{G}$. It follows that the dual vector $\boldsymbol{\alpha}^*$ induces system-optimal flows and is a suitable choice for a toll vector. If one insists that tolls be positive, then one can replace the compatibility constraint $\sum_{g\in\mathcal{G}} \boldsymbol{v}^g = \bar{\boldsymbol{v}}_{\mathrm{SO}}$ by the inequality $\sum_{g\in\mathcal{G}} \boldsymbol{v}^g \geq \bar{\boldsymbol{v}}_{\mathrm{SO}}$, which must be tight at the optimal solution to the linear program.

The analysis can be repeated in the infinite-dimensional case, setting tolls to the dual vector of the infinite-dimensional LP to

$$\begin{aligned} \underset{\boldsymbol{v}\in F}{\text{minimize}} \qquad & \boldsymbol{t}(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\bar{\boldsymbol{v}} + \int_0^\infty \alpha f(\bar{\boldsymbol{v}}_{\mathrm{SO}})^{\mathrm{T}}\boldsymbol{v}(\alpha)\, d\alpha, \\ \text{subject to} \qquad & \int_0^\infty \boldsymbol{v}(\alpha)\, d\alpha = \bar{\boldsymbol{v}}_{\mathrm{SO}}. \end{aligned}$$

By formulating the multi-attribute problem with respect to total flows (see Section 3.1), it is also possible to derive a finite-dimensional formulation variant, and obtain the same result via arguments invoked in the discrete case.

Note that, in the previous analysis, no monotonicity condition was required from the mappings $\boldsymbol{t}$ and $\boldsymbol{f}$, and that cycle-free flow patterns other than system-optimal ones could be induced by the above technique. Note also that, if tolls cannot be set on all links of the transportation network, it might not be possible to achieve system optimality through a toll schedule. The resulting problem, which involves equilibrium constraints, can be formulated as an MPEC (Mathematical Program with Equilibrium Constraints). This class of hard non-convex and non-differentiable problems is usually not amenable to theoretically efficient algorithms, although much attention has been paid, recently, to the design of locally convergent algorithms. Instances of such problems are presented in the next two sections.

6.3 Tolls for raising revenues

Let us reconsider the Braess paradox network illustrated in Figure 2.9. By imposing a toll ρ_{23} larger than 13 on the diagonal link (2,3), one can induce the link-flow system-optimal solution

$$\boldsymbol{v} = (v_{12}, v_{13}, v_{23}, v_{24}, v_{34}) = (3, 3, 0, 3, 3),$$

with equilibrium cost 83. Alternatively, one may wish to maximize the revenue generated by the toll link (2,3). By solving the equilibrium problem for ρ_{23} fixed, one can express the diagonal flow v_{23} as a function of ρ_{23} and obtain the revenue

$$\rho_{23} \cdot v_{23}(\rho_{23}) = \begin{cases} 4\rho_{23} & \text{if } \rho_{23} \leq -26, \\ \rho_{23}(2 - \frac{2}{13}\rho_{23}) & \text{if } \rho_{23} \in [-26, 13], \\ 0 & \text{if } \rho_{23} \geq 13. \end{cases}$$

The optimal revenue of 13/2 is achieved when $\rho_{23} = 13/2$. This toll not only maximizes revenue but, surprisingly, also reduces each user perceived travel cost from 92 to 87.5.

In the case of a general network, let $\mathcal{L}_1$ denote the set of toll links and $\mathcal{L}_2 = \mathcal{L} \setminus \mathcal{L}_1$; let $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$, $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = (\boldsymbol{\rho}_1, \mathbf{0})$, $\boldsymbol{B} = (\boldsymbol{B}_1, \boldsymbol{B}_2)$, and $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2)$ be the respective partitions of the link flow vector, toll vector, constraint matrix and cost mapping. The revenue maximizing problem is to

$$\begin{aligned} \underset{\boldsymbol{v}\in F, \boldsymbol{\rho}_1}{\text{maximize}} \quad & \boldsymbol{\rho}_1^{\mathrm{T}} \boldsymbol{v}_1, \\ \text{subject to} \quad & (\boldsymbol{t}(\boldsymbol{v}) + \boldsymbol{\rho})^{\mathrm{T}} (\boldsymbol{v} - \boldsymbol{w}) \leq 0 \quad \boldsymbol{w} \in F, \\ & \boldsymbol{\rho}_2 = \mathbf{0}. \end{aligned} \tag{79}$$

If the mapping $\boldsymbol{t}$ is monotone, one can replace the lower level variational inequality by its Karush–Kuhn–Tucker conditions. This yields the MPEC formulation

$$\begin{aligned} \underset{\boldsymbol{v}\in F, \boldsymbol{\rho}_1, \boldsymbol{\pi}}{\text{maximize}} \quad & \boldsymbol{\rho}_1^{\mathrm{T}} \boldsymbol{v}_1, \\ \text{subject to} \quad & \boldsymbol{t}_1(\boldsymbol{v}) + \boldsymbol{\rho}_1 - \boldsymbol{B}_1^{\mathrm{T}} \boldsymbol{\pi} \geq \mathbf{0}, \\ & \boldsymbol{t}_2(\boldsymbol{v}) + \boldsymbol{\rho}_2 - \boldsymbol{B}_1^{\mathrm{T}} \boldsymbol{\pi} \geq \mathbf{0}, \\ & \boldsymbol{v}_1^{\mathrm{T}} [\boldsymbol{t}_1(\boldsymbol{v}) + \boldsymbol{\rho}_1 - \boldsymbol{B}_2^{\mathrm{T}} \boldsymbol{\pi}] = 0, \\ & \boldsymbol{v}_2^{\mathrm{T}} [\boldsymbol{t}_2(\boldsymbol{v}) - \boldsymbol{B}_2^{\mathrm{T}} \boldsymbol{\pi}] = 0, \end{aligned} \tag{80}$$

or, upon the introduction of a vector $\boldsymbol{z}$ of binary variables and a suitably large constant M, the mixed discrete-continuous program

$$\begin{aligned} \underset{\boldsymbol{v}\in F, \boldsymbol{\rho}_1, \boldsymbol{\pi}, \boldsymbol{z}}{\text{maximize}} \quad & \boldsymbol{\rho}_1^{\mathrm{T}} \boldsymbol{v}_1, \\ \text{subject to} \quad & \boldsymbol{t}(\boldsymbol{v}) + \boldsymbol{\rho} - \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\pi} \geq \mathbf{0}, \\ & \boldsymbol{v} \leq M\boldsymbol{z} \geq \boldsymbol{t}(\boldsymbol{v}) + \boldsymbol{\rho} - \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\pi}, \\ & z_l \in \{0, 1\} \quad l \in \mathcal{L}. \end{aligned} \tag{81}$$

If the mapping $\boldsymbol{t}$ is increasing and separable, (81) can be approximated by a piecewise linear function, hence the importance of investigating the linear case, where $\boldsymbol{t}(\boldsymbol{v}) = \boldsymbol{t}^{\mathrm{T}} \boldsymbol{v}$. First, due the complementarity conditions, $\rho_{1l} = (\boldsymbol{B}_1^{\mathrm{T}} \boldsymbol{\pi})_l - t_{1l}$ whenever the flow v_{1l} is positive. On the other hand, ρ_{1l} may be set to any value larger than $(\boldsymbol{B}_2^{\mathrm{T}} \boldsymbol{\pi})_l - t_{1l}$ when $v_{1l} = 0$. Therefore, one may set a priori $\rho_1 = \boldsymbol{B}_1^{\mathrm{T}} \boldsymbol{\pi}$ in the optimal solution. Replacing ρ_1 by its value in the objective of (80) one obtains the simplified formulation

$$\begin{aligned} \underset{\boldsymbol{v}\in F, \boldsymbol{\pi}}{\text{maximize}} \quad & (\boldsymbol{B}_1^{\mathrm{T}} \boldsymbol{\pi} - \boldsymbol{t}_1)^{\mathrm{T}} \boldsymbol{v}_1 = \boldsymbol{\pi}^{\mathrm{T}} (\boldsymbol{b} - \boldsymbol{B}_2 \boldsymbol{v}_2) - \boldsymbol{t}_1^{\mathrm{T}} \boldsymbol{v}_1 = \boldsymbol{b}^{\mathrm{T}} \boldsymbol{\pi} - \boldsymbol{t}^{\mathrm{T}} \boldsymbol{v}, \\ \text{subject to} \quad & \boldsymbol{t}_2 - \boldsymbol{B}_2^{\mathrm{T}} \boldsymbol{\pi} \geq 0, \\ & \boldsymbol{v}_2^{\mathrm{T}} (\boldsymbol{t}_2 - \boldsymbol{B}_2^{\mathrm{T}} \boldsymbol{\pi}) = 0. \end{aligned} \tag{82}$$

If one solves the mixed integer linear program (82) by an implicit enumeration method, an upper bound on the revenue must be available at each node of the branch-and-bound tree. At such a node, let $\mathcal{L}_0 \subseteq \mathcal{L}_2$ denote the index set of toll-free links with null flow, and let us relax the complementarity constraint corresponding to flows $\boldsymbol{v}_l$, $l \in \mathcal{L}_2 \setminus \mathcal{L}_0$, yielding the upper bound

$$\begin{aligned} \underset{\boldsymbol{\pi}}{\text{maximum}} \quad & \boldsymbol{b}^{\mathrm{T}} \boldsymbol{\pi} \quad - \quad \underset{\boldsymbol{v}\in F}{\text{minimum}} \ \boldsymbol{t}^{\mathrm{T}} \boldsymbol{v}, \\ \text{subject to} \quad & B_2^{\mathrm{T}} \boldsymbol{\pi} \leq \boldsymbol{t}_2, \\ & (B_2^{\mathrm{T}} \boldsymbol{\pi})_l = t_{2l}, \quad l \in \mathcal{L}_0. \end{aligned}$$

Replacing the left-hand LP by its dual, the upper bound can be expressed as

$$\begin{aligned} \underset{\boldsymbol{v}\in F, \boldsymbol{v}_1 = \mathbf{0}}{\text{minimum}} \quad & \boldsymbol{t}^{\mathrm{T}} \boldsymbol{v} \quad - \quad \underset{\boldsymbol{v}\in F}{\text{minimum}} \ \boldsymbol{t}^{\mathrm{T}} \boldsymbol{v}, \\ \text{subject to} \quad & v_{2l} \text{ free if } l \in \mathcal{L}_0, \end{aligned} \tag{83}$$

and can be interpreted as the difference between two shortest distances, one in a network with tolls set at 0 and some link flows unrestricted in sign, the other in a network with tolls set at ∞. Once the subset of positive flows has been selected, a set of optimal tolls compatible with this choice can be retrieved from the dual solution associated with the LP that appears on the left-hand side of (83). This operation, which amounts to solving an *inverse* linear program, can be performed very efficiently, by computing shortest paths on a network from which toll links have been removed, and where backwards copies of toll-free links carrying positive flow are introduced.

6.4 A continuous network design problem

In this section, we address the problem that consists in designing a network subject to equilibrium constraints, through the specification of continuous control parameters. More specifically, let $\varphi(\boldsymbol{\rho})$ denote the cost of implementing a traffic control $\boldsymbol{\rho}$ belonging to some set Z, assumed to be convex. Based on this notation, the toll problems considered in the previous subsections constitute special instances of the general problem to

$$\begin{aligned} &\underset{\boldsymbol{\rho}\in Z,\boldsymbol{v}\in F(\boldsymbol{\rho})}{\text{minimize}} && \boldsymbol{v}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{v},\boldsymbol{\rho})+\varphi(\boldsymbol{\rho}), \\ &\text{subject to} && -\boldsymbol{t}(\boldsymbol{v},\boldsymbol{\rho})\in N_{F(\boldsymbol{\rho})}(\boldsymbol{v}), \end{aligned} \tag{84}$$

where the vector of design parameters $\boldsymbol{\rho}$ may impact both the cost mapping $\boldsymbol{t}$ and the feasible set $F(\boldsymbol{\rho})$. For instance, one may consider improving a road network through capacity enhancement, balancing long term investment costs against recurrent travel delays. In a continuous version of a network design problem, where φ_l represents the capacity of link l, $\varphi_l(z_l)$ the cost of achieving capacity z_l, we assume that the functions $\boldsymbol{t}$ and φ are link-separable. If, furthermore, t_l is a function of the flow-capacity ratio, the set F is independent of $\boldsymbol{\rho}$, and the design vector $\boldsymbol{\rho}$ is unconstrained, then a continuous variant of the network design problem (84) takes the form

$$\begin{aligned} &\underset{\boldsymbol{\rho}\in Z,\boldsymbol{v}\in F}{\text{minimize}} && \sum_{l\in\mathcal{L}} v_l t_l(v_l/\rho_l)+\varphi_l(\rho_l), \\ &\text{subject to} && \boldsymbol{v}\in\arg\underset{\boldsymbol{w}\in F}{\text{minimum}}\sum_{l\in\mathcal{L}}\int_0^{w_l} t_l(u/\rho_l)\,du. \end{aligned} \tag{85}$$

While the resulting problem is theoretically difficult, it is amenable to efficient heuristic procedures. Maybe the simplest one consists in first solving the system-optimal problem to

$$\text{H1:}\qquad \underset{\boldsymbol{\rho}\in Z,\boldsymbol{v}\in F}{\text{minimize}} \sum_{l\in\mathcal{L}} v_l t_l(v_l/\rho_l)+\varphi_l(\rho_l), \tag{86}$$

for the design vector $\boldsymbol{\rho}_{\mathrm{SO}}$ and then finding the equilibrium flows compatible with $\boldsymbol{\rho}_{\mathrm{SO}}$. This is Heuristic H1. Another natural procedure consists in iterating, à la Gauss–Seidel, between flow and capacity assignment subproblems: for fixed $\boldsymbol{\rho}$, one solves a traffic equilibrium problem while, for fixed flow pattern $\boldsymbol{v}$, each $\rho_l=\rho_l(v_l)$ satisfies the single-variable equation

$$\text{H2:}\qquad t_l(v_l/\rho_l)+(v_l/\rho_l)^2 t_l'(v_l/\rho_l)+\varphi_l'(\rho_l)=0, \tag{87}$$

and is frequently available in closed form. If the procedure converges, it does so to an equilibrium flow that is compatible with the 'greedy' optimality condition (87). This is Heuristic H2, whose solution can also be computed by solving the single variational inequality

$$-\,\boldsymbol{t}(\boldsymbol{v},\boldsymbol{\rho}(\boldsymbol{v}))\in N_F(\boldsymbol{v}). \tag{88}$$

A general class of heuristic procedures consists in solving a problem where the flow part of the solution automatically obeys Wardrop's equilibrium principle. This can be achieved by replacing the total cost by its integral in the objective and, for generality, scaling the investment cost term. This yields the mathematical program

$$\text{H3:}\qquad \underset{\boldsymbol{\rho}\in Z,\boldsymbol{v}\in F}{\text{minimize}}\ \boldsymbol{t}(\boldsymbol{v},\boldsymbol{\rho})=\sum_{l\in\mathcal{L}}\int_0^{v_l} t_l(u/\rho_l)\,du+\sum_{l\in\mathcal{L}}\xi_l\varphi_l(\rho_l). \tag{89}$$

Next, one might look for the set of parameters that yields the best solution to the original program (85). This problem, however, is of the same computational complexity as solving (85) directly, and is thus 'intractable.' In this respect, it is natural to restrict our attention to a sub-family H3, for instance that where all ξ_l are set to a common value ξ. To pursue the analysis, we suppose that φ is the monomial $\varphi_l(\rho_l)=d_l\rho_l^m$ and that the cost function $\boldsymbol{t}$ assumes the BPR form $t_l(x)=\alpha_l+\beta_l x^p$. Under these assumptions, the solution to (87) is given by

$$\rho_l(v_l)=\left(\frac{p\beta_l}{md_l}\right)^{\frac{1}{p+m}} v_l^{\frac{p+1}{p+m}}. \tag{90}$$

If one introduces the function

$$\zeta_l(v_l) = \left(\frac{1}{p+1}\right)^{\frac{1}{p}} \left(\frac{p\beta_l}{md_l}\right)^{\frac{1}{p+m}} v_l^{\frac{p+1}{p+m}},$$

one may check that the solution to the convex program

$$\text{H4:} \qquad \underset{\boldsymbol{v}\in F}{\text{minimize}} \sum_{l\in\mathcal{L}} \int_0^{v_l} [(\alpha_l + \beta_l(u/\rho_l(u))^p]\, du \tag{91}$$

coincides with the system-optimal flow of program (86).

Note that, for the above functional forms, the class H3 defined by (89) subsumes the Gauss–Seidel solution, the latter corresponding to the choice $\xi_l = 1/(p+1)$ for every $l \in \mathcal{L}$, and that all heuristics presented find a design vector z by solving a convex optimization problem, and are thus easy to implement. To assess the quality of their respective solutions, we adopt the worst-case analysis point of view, and aim to determine exactly (or estimate) the ratio

$$R_m^p(H) = \sup_{\alpha,\beta,d} \frac{\text{cost of heuristic solution}}{\text{cost of (unknown) optimal solution}} = \sup_{\alpha,\beta,\Delta} \frac{F(\boldsymbol{v}(\boldsymbol{\rho}^H), \boldsymbol{\rho}^H)}{F(\boldsymbol{v}^*, \boldsymbol{\rho}^*)} \tag{92}$$

over all possible network configurations represented by the symbol Δ. This analysis is reminiscent of the worst-case analysis presented in Section 2.9. We have the following approximation results.

Theorem 21 *For heuristics* H1 *through* H4*, the following bounds on* R_m^p *hold:*

(a) $\lim_{p\to\infty} R_1^p(\text{H1}) \geq 2$;

(b) $R_1^p(\text{H2}) = p+1$;

(c) $R_m^p(\text{H4}) = m(p+1)/(m+p) + p(m+p)(p+1)^{-\frac{m}{p}}$;

(d) $\lim_{m\downarrow 0} R_m^p(H4) = 1$ *and* $\lim_{p\to\infty} R_1^p(H4) = 2$;

(e) $1 + \frac{p}{\xi(p+1)} \leq R_1^p(H3) \leq \frac{\xi^{p/(p+1)}}{(p+1)^{1/(p+1)}} \left[1 + \frac{p}{\xi(p+1)}\right]^2$;

(f) $2 \leq \lim_{p\to\infty} R_1^p(\text{H4}) = 4$. □

Remark 22 (i) For linear investment functions ($m = 1$), the value that yields the tightest upper bound on $R_1^p(\text{H3})$ is $\xi = 2 - p/(p+1)$.

(ii) The bounds provided by the theorem are very pessimistic. Indeed, numerical experiments show that the error is of the order of one percent. This is not surprising if one observes that the objective of both 'players' are not antagonistic: the leader actually wishes to minimize the users' travel times.

(iii) If the investment function φ is linear, Equation (87) fixes a value for the ratio v_l/ρ_l. It follows that $\boldsymbol{t}(\boldsymbol{v}, \boldsymbol{\rho}(\boldsymbol{v}))$ is constant in the variational inequality (88) and that its solution is an extremal flow pattern that can be efficiently computed by shortest routes methods. □

Once a heuristic solution is obtained, it can be improved using the sensitivity results of Section 6.1 with respect to the design (toll) vector $\boldsymbol{\rho}$. At differentiable points, the procedure reduces to gradient projection. At points of non-differentiability, a descent direction can be tentatively built from the coordinate-wise directional derivatives. In the numerical example below, these were always sufficient to provide a descent direction at every iteration.

We consider the special case of problem (85) where Z is a box, declaring that $0 \leq \rho_l \leq u_l$ must hold for every l. An implementation of the DSD algorithm, the directional derivative problem, and a gradient projection algorithm with a rather simple line search,[26] form a complete descent-based framework for this network design problem. We have applied the procedure to a problem defined for the Sioux Falls network (with 10 potential links to improve), and compared it to some heuristic schemes that have been proposed for the continuous network design problem. The results are shown in Table 8.

In the table, we use the following short-hand notation: SBD stands for 'sensitivity-based descent', that is, the algorithm presented in this section; H–J stands for 'Hooke–Jeeves', a direct search method

[26]We have used the Armijo step length rule; the main motivation is that the objective value calculations are quite expensive.

	SBD	SBD	SBD	SBD	H–J	EDO
Initial value $\boldsymbol{\rho}^0$	0	2	3	6.25	2	—
Upper bound $\boldsymbol{u}$	25	25	25	25	—	25
ρ_{16}	5.3027	5.1492	5.3457	5.2773	4.8	4.59
ρ_{17}	2.0560	2.0214	1.9786	2.0533	1.2	1.52
ρ_{19}	5.3430	5.1679	5.3741	5.3002	4.8	5.45
ρ_{20}	1.9901	2.0012	1.9460	2.0369	0.8	2.33
ρ_{25}	2.5216	2.4945	2.7856	2.7670	2.0	1.27
ρ_{26}	2.5548	2.5447	2.8245	2.8222	2.6	2.33
ρ_{29}	2.9883	2.9535	2.9257	3.0124	4.8	0.41
ρ_{39}	4.8559	4.8330	4.7528	4.7348	4.4	4.59
ρ_{48}	3.0026	2.9798	2.9732	2.9746	4.8	2.71
ρ_{74}	4.8496	4.8212	4.7347	4.7511	4.4	2.71
Obj, reported	79.9969	79.9968	79.9987	80.0026	80.78	83.08
Obj, recomputed	79.9961	79.9971	79.9990	80.0043	80.67	82.34

	SA	H–J	EDO	SA	PIPA
Initial value $\boldsymbol{\rho}^0$	6.25	3	0	6.25	—
Upper bound $\boldsymbol{u}$	25	—	25	25	—
ρ_{16}	5.38	4.507	4.276	5.322	5.4680
ρ_{17}	2.26	4.509	2.288	2.596	2.0039
ρ_{19}	5.50	4.520	4.080	5.664	5.4471
ρ_{20}	2.01	4.052	1.618	1.309	1.9395
ρ_{25}	2.64	4.299	1.654	2.498	2.9448
ρ_{26}	2.47	2.949	1.130	2.732	2.8191
ρ_{29}	4.54	3.000	3.219	4.123	3.4039
ρ_{39}	4.45	3.601	3.326	4.508	4.8061
ρ_{48}	4.21	3.006	1.981	3.736	3.2364
ρ_{74}	4.67	3.200	3.190	3.903	4.7779
Obj, reported	80.87	83.316	83.703	81.983	80.8669
Obj, recomputed	80.42	81.185	81.345	80.304	80.0528

Table 8: Results for network design on the Sioux Falls network.

based only on objective values; EDO stands for 'equilibrium decomposed optimization', an algorithm which utilizes the 'classical' sensitivity analysis formula (71); SA stands for 'simulated annealing;' and PIPA stands for 'penalty interior point algorithm.' The Notes section provides the sources for the results in the table.

The values in the last row of the table corresponds to a much more accurately calculated upper-level objective value for the terminal designs; in the case of our algorithm, the difference between the nominal values are small, since accurate traffic equilibrium computations are used throughout. For the other algorithms—which typically use 50–100 Frank–Wolfe iterations on each lower-level problem—the differences are much larger. Clearly, their final solutions suffer from a poor estimation of traffic equilibrium solutions, in addition to being further from a local optimum (or, a stationary point) than in the proposed algorithm.

The reason why traffic equilibrium computations need to be accurate becomes obvious when one realizes that they serve as input to the sensitivity analysis procedure, which in turn provides the gradient values or, at least, coordinate-wise directional derivatives. Unless the equilibrium solution is accurately computed, the sensitivity analysis may provide erroneous results, leading to bad search directions, or to the algorithm halting prematurely. However, this is not a serious problem with the DSD algorithm, which can solve the traffic equilibrium problem to a sufficiently good accuracy quickly. This must be taken as an indication that the Frank–Wolfe method, or any other slowly converging method for the traffic equilibrium problem, should definitely *not* be used.

We remark finally that the necessity to provide accurate equilibrium solutions arises in several other applications involving traffic equilibrium models. We mention here only two, in addition to all models of a hierarchical nature, such as traffic management or toll optimization type models: (i) assessing the environmental impact of the transportation network (exhaust fumes, for instance) relies on accurate link usage data, which in turn rely on accurate traffic equilibrium solutions; (ii) travel forecasting, subject to changes in the network configuration, also relies on accurate flow estimates, both for the 'before' and 'after' simulations.

Bibliographical notes

The sensitivity analysis of Section 6.1 was first developed in [PaR02, PaR03] and later refined in [Pat04]. (Regarding the background material, (64) stems from [DoR01], and (66) from [Kyp90].) The material presented here is a condensed version that for the main part discusses the standard, separable model. That the sensitivity analysis does not depend on the equilibrium route flow chosen was first noticed in [ToF88], albeit in a framework that has limitations, as we remark and show; this independence was observed and utilized later in [PaR02, Pat04, PaR03, Pat06]. The examples showing the limitations of the Tobin/Friesz formula (71) are taken from [JoP03, JoP05, Pat06].

While it is a classical result from economic theory that marginal cost pricing can induce a system-optimal flow pattern, Section 6.2 focuses on toll schedules that achieve that aim at minimal user cost. Since the set of system-optimal tolls is polyhedral (see [BHR97, HeR98, LaP98]), such schedules are solutions to a linear program. Dial [Dia99, Dia00], either in the single or multi-commodity case, proposed an algorithm that exploits the underlying network structure. A similar result was obtained by Marcotte and Savard [MaS02], who made the link, in the single-commodity case, with the inverse shortest route problem (see [AhO01]). In a multi-attribute context, where marginal taxation may not be feasible, Yang and Huang [YaH04] gave a constructive proof of the existence of system-optimal tolls. The extension of their method from the discrete to the continuous case is new to this book. A more complex existence proof, of a highly non-constructive nature (even for the discrete case), was given by [CYR03].

Whenever tolls cannot be set on *every* link of a network, the nature of the problem becomes more complex, and falls within the realm of *bilevel programming*. Although bilevel programs, together with the closely related MPECs (Mathematical Programs with Equilibrium Constraints), have been topics of several studies in the past years, much work remains to be done before robust procedures may address real-life instances. The reader is referred to the books by Luo et al. [LPR96], Bard [Bar98] and Shimizu et al. [SIB97] for introductions to the subject.

Section 6.3 addresses one such instance, where one aims at maximizing toll revenues. References for theory and algorithms relevant to uncongested networks are [LMS98, LMS99, BLMS01, MaS01, MaS02, BHKL03]. A linearization algorithm applicable to the congested case can be found in [Jul98].

The continuous variant of the network design problem considered in Section 6.4 was first proposed in [AbL79]. Heuristics were analyzed from the theoretical and computational perspectives in [Mar86b, MaM92]. The numerical results displayed in Figure 10 and Table 8 demonstrate that Frank–Wolfe type methods are to recommend for use neither in the solution of the basic traffic equilibrium model nor for more complex models where accurate equilibrium solutions are needed as inputs. The numerical results are taken from [JoP03, Pat06]. In Table 8, the last two columns on the first row appeared in [SFT87], the first column on the second row in [FCMTA92], columns 2–5 in [HuB98], and column five in [Lim02]. The remarks made at the end of the section refer to application work performed in [LLPR01, BRB02].

References

[AaM81] H. Z. Aashtiani and T. L. Magnanti, "Equilibria on a congested transportation network", *SIAM Journal on Algebraic and Discrete Methods*, 2 (1981), pp. 213–226.

[AbL79] M. Abdulaal and L.J. LeBlanc, "Continuous equilibrium network design models", *Transportation Research B*, 13 (1979) 19–32.

[AMO93] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[AhO01] R. K. Ahuja and J.B. Orlin, "Inverse optimization", *Operations Research*, 49 (2001), pp. 771–783.

[Aka96] T. Akamatsu, "Cyclic flows, Markov process and stochastic traffic assignment", *Transportation Research*, 30B (1996), pp. 369–386.

[AEA02] E. Altman, R. El Azouzi, and V. Abramov, "Non-cooperative routing in loss networks", *Performance Evaluation*, 49 (2002), pp. 43–55.

[Aus76] A. Auslender, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[Bai75] J. B. Baillon, "Un théorème de type ergodique pour les contractions non linéaires dans un espace de Hilbert", *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A*, 280 (1975), pp. A1511–A1514.

[Bar98] J. F. Bard, Practical Bilevel Optimization: Algorithms and Applications, Kluwer Academic Publishers, Boston, MA, 1998.

[Bar02a] H. Bar-Gera, "Origin-based algorithm for the traffic assignment problem", *Transportation Science*, 36 (2002), pp. 398–417.

[Bar02b] H. Bar-Gera, "Origin-based network assignment", In: *Transportation Planning—State of the Art*, M. Patriksson and M. Labbé (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 1–17.

[BSS93] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY, second ed., 1993.

[BMW56] M. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*, Yale University Press, New Haven, CT, 1956.

[BeI97] M. G. H. Bell and Y. Iida, *Transportation Network Analysis*, John Wiley & Sons, Chichester, U.K., 1997.

[Ben93] 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.

[Ber95] P. Bergendorff, "The bounded flow approach to congestion pricing", Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 1995.

[BHR97] P. Bergendorff, D. W. Hearn, and M. V. Ramana, "Congestion toll pricing of traffic networks", in *Network Optimization*, Proceedings of the Network Optimization Conference, University of Florida, Gainesville, FL, February 12–14, 1996, P. M. Pardalos, D. W. Hearn, and W. W. Hager, eds., vol. 450 of Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1997, pp. 51–71.

[BeS94] D. Bernstein and T. E. Smith "Equilibria for networks with lower semicontinuous costs: With an application to congestion pricing", *Transportation Science*, 28 (1994), pp. 221–235.

[BeW00] D. Bernstein and L. Wynter, "Issues of uniqueness and convexity in non-additive bi-criteria equilibrium models", Conference paper, 8th Meeting of the EURO Working Group on Transportation, Rome Jubilee 2000 Conference, La Sapienza, Rome, Italy, September 11–14, 2000.

[Ber76] D. P. Bertsekas, "On the Goldstein–Levitin–Polyak gradient projection method", *IEEE Transactions on Automatic Control*, AC-21 (1976), pp. 174–184.

[Ber98] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*, Athena Scientific, Belmont, MA, 1998.

[BeG82] 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.

[BeG92] D. P. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, second ed., 1992.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, London, 1989.

[BeG75] M. J. Best and V. J. Griffin, "Optimisation of nonlinear performance criteria subject to flow constraints", Conference paper presented at the ORSA/TIMS National Meeting, Chicago, IL, USA, 30 April–2 May 1975.

[BHKL03] M. Bouhtou, S. van Hoesel, Anton F. van der Kraaij, and J-L Lutton, "Tariff optimization in Networks", Report RM03011, Maastricht Economic Research School on Technology and Organisation, 2003.

[Boy04] D. Boyce, "Forecasting travel on congested urban transportation networks: Review and prospects for network equilibrium models", Conference paper, presented at TRISTAN V, The Fifth Triennial Symposium on Transportation Analysis, Le Gosier, Guadeloupe, June 13–18, 2004.

[BRB02] D. E. Boyce, B. Ralevic-Dekic and H. Bar-Gera, "Convergence of traffic assignments: How much is enough? The Delaware Valley Region Case Study", paper presented at the 16th Annual International EMME/2 Users' Group Conference, Albuquerque, NM, March, 2002.

[BLMS01] L. Brotcorne, M. Labbé, P. Marcotte, and G. Savard, "A bilevel model for toll optimization on a multicommodity transportation network", *Transportation Science*, 35 (2001), 1–14.

[Bru75] R. E. Bruck, Jr, "Asymptotic convergence of nonlinear contraction semigroups in Hilbert space", *Journal of Functional Analysis*, 18 (1975), pp. 15–26.

[Bru77] R. E. Bruck, Jr, "On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space", *Journal of Mathematical Analysis and Applications*, 61 (1977), pp. 159–164.

[Bru84] P. Brucker, "An O(n) algorithm for quadratic knapsack problems", *Operations Research Letters*, 3 (1984), pp. 163–166.

[BuF91] J. V. Burke and M. C. Ferris, "Characterization of solution sets of convex programs", *Operations Research Letters*, 10 (1991), pp. 57–60.

[BuS65] R. G. Busacker and T. L. Saaty, *Finite Graphs and Networks: An Introduction with Applications*, McGraw-Hill, New York, NY, 1965.

[Cas01] E. Cascetta, *Transportation Systems Engineering: Theory and Methods*, no. 49 in Applied Optimization, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[ChS03] C.K. Chau and K.M. Sim, "The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands", *Operations Research Letters*, 31 (2003) 327–334.

[CLJ02] A. Chen, D.-H. Lee, and R. Jayakrishnan, "Computational study of state-of-the-art path-based traffic assignment algorithms", *Mathematics and Computers in Simulation*, 2060 (2002), pp. 1–10.

[ChB03] M. Chen and D. Bernstein, "Solving the toll design problem with multiple user groups", *Transportation Research B*, 37 (2003), pp. 1–19.

[CSF00] H.-J. Cho, T. E. Smith, and T. L. Friesz, "A reduction method for local sensitivity analyses of network equilibrium arc flows", *Transportation Research*, 34B (2000), pp. 31–51.

[Coh88] G. Cohen, "Auxiliary problem principle extended to variational inequalities", *Journal of Optimization Theory and Applications*, 59 (1988) pp. 325–333.

[CYR03] R. Cole, Y. Dodis, and T. Roughgarden, "Pricing network edges for heterogeneous selfish users", In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, 2003, pp. 521–530.

[CSS03] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses, "Selfish routing in capacitated networks", *Mathematics of Operations Research*, 29 (2004) pp. 961–976.

[CMS03] J.-P. Côté, P. Marcotte, and G. Savard, "A bilevel modeling approach to pricing and fare optimization in the airline industry", *Journal of Revenue and Pricing Management*, 2 (2003) pp. 23–36.

[CMZ00] J.-P. Crouzeix, P. Marcotte, and D. Zhu, "Conditions ensuring the applicability of cutting-plane methods for solving variational inequalities", *Mathematical Programming*, 88 (2000), pp. 521–539.

[Daf71] S. C. Dafermos, "An extended traffic assignment model with applications to two-way traffic", *Transportation Science*, 5 (1971), pp. 366–389.

[Daf80] S. C. Dafermos, "Traffic equilibrium and variational inequalities", *Transportation Science*, 14 (1980), pp. 42–54.

[Daf81] S. C. Dafermos, "A multicriteria route-mode choice traffic equilibrium model", unpublished manuscript, Lefschetz Center for Dynamical Systems, June 1981.

[Daf82] S. C. Dafermos, "The general multimodal network equilibrium problem with elastic demand", *Networks*, 12 (1982), pp. 57–72.

[DaS69] 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.

[dPN98] A. de Palma and Y. Nesterov, “Optimization formulations and static equilibrium in congested transportation networks”, CORE Discussion Paper 9861, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1998.

[Dia79] R. B. Dial, “A model and algorithm for multicriteria route-mode choice”, *Transportation Research*, 13B (1979), pp. 311–316.

[Dia96a] R. B. Dial, “Bicriterion traffic assignment: basic theory and elementary algorithms”, *Transportation Science*, 30 (1996), pp. 93–111.

[Dia96b] R. B. Dial, “Bicriterion traffic assignment: efficient algorithms plus examples”, *Transportation Research B*, 31 (1996), pp. 357–359.

[Dia99] R.B. Dial, “Minimal-revenue congestion pricing part I: A fast algorithm for the single-origin case”, *Transportation Research B*, 33 (1999), pp. 189–202.

[Dia00] R.B. Dial, “Minimal-revenue congestion pricing part II: An efficient algorithm for the general case”, *Transportation Research B*, 34 (2000), pp. 645–665.

[DoR01] A. L. Dontchev and R. T. Rockafellar, “Ample parameterization of variational inclusions”, *SIAM Journal on Optimization*, 12 (2001), pp. 170–187.

[Dun73] J. C. Dunn, “On recursive averaging processes and Hilbert space extensions of the contraction mapping principle”, *Journal of the Franklin Institute*, 295 (1973), pp. 117–133.

[Esc86] 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.

[Eva76] S. P. Evans, “Derivation and analysis of some models for combining trip distribution and assignment”, *Transportation Research*, 10 (1976), pp. 37–57.

[FaP03a] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, vol. I of Springer Series in Operations Research, Springer-Verlag, Berlin, 2003.

[FaP03b] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, vol. II of Springer Series in Operations Research, Springer-Verlag, Berlin, 2003.

[Fer95] P. Ferrari, “Road pricing and network equilibrium”, *Transportation Research*, 29B (1995), pp. 357–372.

[FeP97] M. C. Ferris and J.-S. Pang, “Engineering and economic applications of complementarity problems”, *SIAM Review*, 39 (1997), pp. 669–713.

[Fis80] C. S. Fisk, “Some developments in equilibrium traffic assignment”, *Transportation Research*, 14B (1980), pp. 243–255.

[Flo77] M. Florian, “A traffic equilibrium model of travel by car and public transit modes”, *Transportation Science*, 11 (1977), pp. 166–179.

[Flo79] M. Florian, “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 Opérationnelle, Université de Montréal, Montréal, Canada, 1979.

[FlH95] M. Florian and D. W. Hearn, “Network equilibrium models and algorithms”, Chapter 6 of *Network Routing*, M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, eds., vol. 8 of Handbooks in Operations Research and Management Science, North-Holland, Amsterdam, 1995, pp. 485–550.

[FWH02] M. Florian, J. H. Wu, and S. He, “Multi-mode variable demand network equilibrium model with hierarchical logit structure”, in *Transportation and Network Analysis: Current Trends*, M. Gendreau and P. Marcotte, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 119–133.

[FGK71] L. Fratta, M. Gerla, and L. Kleinrock, “The flow-deviation method: An approach to store-and-forward computer communication network design”, *Networks*, 3 (1971), pp. 97–133.

[FCMTA92] 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.

[GaB97] S. A. Gabriel and D. Bernstein, “The traffic equilibrium problem with non-additive path costs”, *Transportation Science*, 31 (1997), pp. 337–348.

[GaB84] E. M. Gafni and D. P. Bertsekas, “Two-metric projection methods for constrained optimization”, *SIAM Journal on Control and Optimization*, 22 (1984), pp. 936–64.

[Gal77] 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.

[Gar80a] N. H. Gartner, "Optimal traffic assignment with elastic demands: A review. Part I: Analysis framework", *Transportation Science*, 14 (1980), pp. 174–191.

[Gar80b] N. H. Gartner, "Optimal traffic assignment with elastic demands: A review. Part II: Algorithmic approaches", *Transportation Science*, 14 (1980), pp. 192–208.

[GHV92] J.-L. Goffin, A. Haurie, A., and J.-P. Vial, "Decomposition and nondifferentiable optimization with the projective algorithm", *Management Science*, 38 (1992), 284–302.

[GLT81] R. Glowinski, J-L Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[GMZ97] J-L Goffin, P. Marcotte, and D. L. Zhu, "An analytic center cutting-plane method for pseudomonotone variational inequalities", *Operations Research Letters*, 20 (1997) pp. 1–6.

[HaT98] J. N. Hagstrom and P. Tseng, *Traffic equilibrium: Link flows, path flows and weakly/strongly acyclic solutions*, Technical report M/C 294, Department of Information and Decision Sciences, University of Illinois, Chicago, IL, 1998. Available at `www.math.washington.edu/~tseng/papers.html`.

[HMS04a] Y. Hamdouch, P. Marcotte, and S. Nguyen, "Capacitated transit assignment with loading priorities", *Mathematical Programming B*, 101 (2004) pp. 205–230.

[HMS04b] Y. Hamdouch, P. Marcotte, and S. Nguyen, "A strategic model for dynamic traffic assignment", *Networks and Spatial Economics*, 4 (2004) pp. 291–315.

[Ham84] 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.

[HaS66] P. Hartman and G. Stampacchia, "On some non-linear elliptic differential-functional equations", *Acta Mathematica*, 115 (1966), pp. 271–310.

[Hea80] D. W. Hearn, "Bounding flows in traffic assignment models", Research Report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 1980.

[HLV87] D. W. Hearn, S. Lawphongpanich, and J. A. Ventura, "Restricted simplicial decomposition: Computation and extensions", *Mathematical Programming Study*, 31 (1987), pp. 99–118.

[HeR98] D. W. Hearn and M. V. Ramana, "Solving congestion toll pricing models", in *Equilibrium and Advanced Transportation Modelling*, P. Marcotte and S. Nguyen, eds., Kluwer Academic Publishers, Boston, MA, 1998, pp. 109–124.

[Hey86] B. G. Heydecker, "On the definition of traffic equilibrium", *Transportation Research*, 20B (1986), pp. 435–440.

[HuB98] H.-J. Huang and M. G. H. Bell, "Continuous equilibrium network design problem with elastic demand: derivative-free solution methods", in *Transportation Networks: Recent Methodological Advances*, M. G. H. Bell (ed.), Elsevier Science, 1998.

[Ius98] A. N. Iusem, "On some properties of paramonotone operators", *Journal of convex analysis*, 5 (1998), pp. 269–278.

[JTPR94] R. Jayakrishnan, W. K. Tsai, J. N. Prashker, and S. Rajadhyaksha, "Faster path-based algorithm for traffic assignment", *Transportation Research Record*, 1443 (1994).

[Jor63] 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.

[JoP03] M. Patriksson and M. Josefsson, "Sensitivity analysis of separable traffic equilibrium solutions, with application to bilevel optimization in network design", Report, Department of Mathematics, Chalmers University of Technology, Gothenburg, Sweden, 2003. To appear in *Transportation Research, B*.

[JoP05] M. Josefsson and M. Patriksson, "On the applicability of sensitivity analysis formulas for traffic equilibrium models", in *Mathematical and Computational Methods for Congestion Charging*, Proceedings of the Theory and Practice of Congestion Charging Symposium, Imperial College, London, August 18–20, 2003, D. Hearn, S. Lawphongpanich, and M. Smith (eds.), vol. 148 in the Applied Optimization series, Springer-Verlag, Berlin, 2005, to appear.

[Jul98] H. Julsain, "Tarification dans les réseaux de télécommunications: une approche par programmation mathématique à deux niveaux", Master's thesis, École Polytechnique, Montréal, 1998.

[KeH80] J. L. Kennington and R. V. Helgason, *Algorithms for Network Programming*, John Wiley & Sons, New York, NY, 1980.

[Kli83] J. G. Klincewicz, "A Newton method for convex separable network flow problems", *Networks*, 13 (1983), pp. 427–442.

[Kni24] F. H. Knight, "Some fallacies in the interpretation of social cost", Quarterly Journal of Economics, 38 (1924), pp. 582–606.

[Kon01] I. V. Konnov, *Combined Relaxation Methods for Variational Inequalities*, Lecture Notes in Economics and Mathematical Systems, Vol. 495, Springer-Verlag, Berlin, 2001.

[Kor77] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems", *Matecon*, 13 (1977), pp. 35–49.

[Kyp88] J. Kyparisis, "Perturbed solution of variational inequality problems over polyhedral sets", *Journal of Optimization Theory and Applications*, 66 (1988), pp. 121–135.

[Kyp90] J. Kyparisis, "Solution differentiability for variational inequalities", *Mathematical Programming*, 48 (1990), pp. 285–301.

[LMS98] M. Labbé, P. Marcotte, and G. Savard, "A bilevel model of taxation and its application to optimal highway pricing", *Management Science*, 44 (1998), pp. 1595–1607.

[LMS99] M. Labbé, P. Marcotte, and G. Savard, "On a class of bilevel programs", *Nonlinear Optimization and Related Topics*, Di Pillo and Giannessi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 183–206.

[Lar+02] T. Larsson, P.-O. Lindberg, J. Lundgren, M. Patriksson, and C. Rydergren, "On traffic equilibrium models with a nonlinear time/money relation", In: *Transportation Planning—State of the Art*, M. Patriksson and M. Labbé (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 19–31.

[LLP97] T. Larsson, Z.-W. Liu, and M. Patriksson, "A dual scheme for traffic assignment problems", *Optimization*, 42 (1997), pp. 323–358.

[LLPR01] T. Larsson, J. Lundgren, M. Patriksson, and C. Rydergren, "Most likely traffic equilibrium route flows: Analysis and computation", In: *Equilibrium Problems: Nonsmooth Optimization and Variational Inequality Models*, Proceedings of an international Workshop in Memory of Marino De Luca, Hotel Villa Diodoro, Taormina, Italy, December 3–5, 1998, F. Giannessi, A. Maugeri, and P. M. Pardalos (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 129–159.

[LaP92] T. Larsson and M. Patriksson, "Simplicial decomposition with disaggregated representation for the traffic assignment problem", *Transportation Science*, 26 (1992), pp. 4–17.

[LaP94a] T. Larsson and M. Patriksson, "A class of gap functions for variational inequalities", *Mathematical Programming*, 64 (1994), pp. 53–79.

[LaP94b] T. Larsson and M. Patriksson, "Equilibrium characterizations of solutions to side constrained asymmetric traffic assignment models", *Le Matematiche*, 49 (1994), pp. 249–280.

[LaP95] T. Larsson and M. Patriksson, "An Augmented Lagrangean dual algorithm for link capacity side constrained traffic assignment problems", *Transportation Research*, 29B (1995), pp. 433–455.

[LaP97] T. Larsson and M. Patriksson, "Price-directive traffic management—an approach utilizing side constrained traffic equilibrium models", *Rendiconti del Circolo Matematico di Palermo, Serie II*, 48 (1997), pp. 147–170.

[LaP98] T. Larsson and M. Patriksson, "Side constrained traffic equilibrium models—traffic management through link tolls", in *Equilibrium and Advanced Transportation Modelling*, P. Marcotte and S. Nguyen, eds., Kluwer Academic Publishers, New York, NY, 1998, pp. 125–151.

[LaP99] T. Larsson and M. Patriksson, " Side constrained traffic equilibrium models: Analysis, computation and applications", *Transportation Research*, 33B (1999), pp. 233–264.

[LPR04] T. Larsson, M. Patriksson, and C. Rydergren, "A column generation procedure for the side constrained traffic equilibrium problem", *Transportation Research*, 38B (2004), pp. 17–38.

[LPS99] T. Larsson, M. Patriksson, and A.-B. Strömberg, "Ergodic, primal convergence in dual subgradient schemes for convex programming", *Mathematical Programming*, 86 (1999), pp. 283–312.

[LaH84] S. Lawphongpanich and D. W. Hearn, "Simplicial decomposition of the asymmetric traffic assignment problem", *Transportation Research*, 18B (1984), pp. 123–133.

[LMP75] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, "An efficient approach to solving the road network equilibrium traffic assignment problem", *Transportation Research*, 9 (1975), pp. 308–318.

[Leu93] F. Leurent, "Cost versus time equilibrium over a network", *European Journal of Operational Research*, 71 (1993), pp. 205–221.

[LNT73] T. Leventhal, G. Nemhauser, and L. Trotter, Jr., "A column generation algorithm for optimal traffic assignment", *Transportation Science*, 7 (1973), pp. 168–176.

[Lim02] A. C. Lim, "Transportation network design problems: An MPEC approach", Ph.D. thesis, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, 2002.

[LPR96] Z.-Q. Luo, J.-S Pang, and D. Ralph, Mathematical programs with equilibrium constraints, Cambridge University Press, Cambridge, U.K., 1996.

[Mar86a] P. Marcotte, "A new algorithm for solving variational inequalities, with application to the traffic assignment problem," *Mathematical Programming*, 33 (1986) pp. 339–351.

[Mar86b] P. Marcotte, "Network design problem with congestion effects: A case of bilevel programming", *Mathematical Programming*, 34 (1986), pp. 142–162.

[Mar91] P. Marcotte, "Application of Khobotov's algorithm to variational inequalities and network equilibrium problems," *INFOR*, 29 (1991), pp. 258–270.

[Mar98] P. Marcotte, "Reformulations of a bicriterion equilibrium problem", in *Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 269–292.

[MaG88] 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.

[MaM92] P. Marcotte and G. Marquis, "Efficient implementation of heuristics for the continuous network design problem", *Annals of Operations Research*, 34 (1992), pp. 163–176.

[MaN98] P. Marcotte and S. Nguyen, "Hyperpath formulations of traffic assignment problems", *Equilibrium and advanced transportation modelling*, P. Marcotte and S. Nguyen, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 175–200.

[MNS04] P. Marcotte, S. Nguyen, and A. Schoeb, "A strategic model of traffic assignment in static capacitated networks", *Operations Research*, 52 (2004) pp. 191–212.

[MNT96] P. Marcotte, S. Nguyen, and K. Tanguay, "Implementation of an efficient algorithm for the multiclass traffic assignment problem", *Transportation and Traffic Theory, Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, 24–26 July 1996, J.-B. Lesort, ed., Pergamon, 1996, pp. 217–236.

[MaS01] P. Marcotte and G. Savard, "Bilevel programming: Formulation", "Bilevel programming: Applications", and"Bilevel programming: Algorithms", in *The Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 155–159.

[MaS02] P. Marcotte and G. Savard, "A bilevel programming approach to price setting," *Decision and Control in Management Science. Essays in Honor of Alain Haurie*, Georges Zaccour, ed., Kluwer Academic Publishers, 2002, pp. 97–117.

[MaW04] P. Marcotte and L. Wynter, "A new look at the multi-class network equilibrium problem", *Transportation Science*, 38 (2004), pp. 282–292.

[MaZ97] P. Marcotte and D. L. Zhu, "Equilibria with infinitely many differentiated classes of customers", in *Complementarity and variational problems*, Proceedings of the *13th International Conference on Complementarity Problems: Engineering and Economics and Applications and Computational Methods*, J.-S. Pang and M. Ferris, eds., SIAM, Philadelphia, 1997, pp. 234–258.

[MPT75] 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.

[Mur70] J. D. Murchland, "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.

[NaD02] A. Nagurney and J. Dong, *Supernetworks: Decision-Making for the Information Age*, Edward Elgar, 2002.

[Nest00] Y. Nesterov, "Stable traffic equilibria, Properties and applications", *Optimization and Engineering*, 3 (2000), pp. 29–50.

[Ngu74] S. Nguyen, "An algorithm for the traffic assignment problem", *Transportation Science*, 8 (1974), pp. 203–216.

[NgD84] S. Nguyen and C. Dupuis, "An efficient method for computing traffic equilibria in networks with asymmetric transportation costs", *Transportation Science*, 18 (1984), pp. 185–202.

[NgP88] S. Nguyen and S. Pallottino, "Equilibrium traffic assignment for large scale transit networks", *European Journal of Operational Research*, 37 (1988), pp. 176–186.

[NPM01] S. Nguyen, S. Pallottino and F. Malucelli, "A modeling framework for the passenger assignment on a transport network with time-tables", *Transportation Science*, 35 (2001), pp. 238–249.

[NPS02] S. Nguyen, S. Pallottino and M.G. Scutellá, "A new dual algorithm for shortest path reoptimization", in *Transportation and Network Analysis: Current Trends*, M. Gendreau and P. Marcotte, eds., Kluwer Academic Publishers, New York, NY, 2002, pp. 221–235.

[PaS03] S. Pallottino and M.G. Scutellà, "A new algorithm for reoptimizing shortest paths when the arc costs change", *Operations Research Letters*, 31 (2003), pp. 149-160.

[Pat94a] M. Patriksson, "On the convergence of descent methods for monotone variational inequalities", *Operations Research Letters*, 16 (1994), pp. 265–269.

[Pat94b] M. Patriksson, *The Traffic Assignment Problem—Models and Methods*, Topics in Transportation, VSP BV, Utrecht, The Netherlands, 1994.

[Pat98] M. Patriksson, *Nonlinear Programming and Variational Inequalities—A Unified Approach*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[Pat04] M. Patriksson, "Sensitivity analysis of traffic equilibria", *Transportation Science*, 38 (2004), pp. 258–281.

[Pat06] M. Patriksson, *Traffic Equilibrium Models: Analysis, Applications, and Optimization Algorithms*, Springer-Verlag (under preparation).

[PaR02] M. Patriksson and R.T. Rockafellar, "A mathematical model and descent algorithm for bilevel traffic management", *Transportation Science*, 36 (2002), pp. 271–291.

[PaR03] M. Patriksson and R.T. Rockafellar, "Sensitivity analysis of variational inequalities over aggregated polyhedra, with application to traffic equilibria", *Transportation Science*, 37 (2003), pp. 56–68.

[PaT75] 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.

[Rob80] S. M. Robinson, "Strongly regular generalized equations", *Mathematics of Operations Research*, 5 (1980), pp. 43–62.

[Rob85] S. M. Robinson, "Implicit B-differentiability in generalized equations", Technical Summary Report 2854, Mathematics Research Center, University of Wisconsin at Madison, Madison, WI, 1985.

[Rob91] S. M. Robinson, "An implicit-function theorem for a class of nonsmooth functions", *Mathematics of Operations Research*, 16 (1991), pp. 292–309.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc76] R. T. Rockafellar, "Monotone operators and the proximal point algorithm", *SIAM Journal on Control and Optimization*, 14 (1976), pp. 877–898.

[Roc84] R. T. Rockafellar, *Network Flows and Monotropic Optimization*, John Wiley & Sons, New York, NY, 1984. Also published by Athena Scientific, Belmont, MA, 1998.

[RoW97] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, 1997.

[RoT02] T. Roughgarden and É. Tardos, "How bad is selfish routing?", *Journal of the ACM*, 49 (2002), pp. 236–259.

[She85] Y. Sheffi, *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[SIB97] K. Shimizu, Y. Ishizuka and J. F. Bard, *Nondifferentiable and Two-Level Programming*, Kluwer Academic Publishers, Boston, MA, 1997.

[Smi79] M. J. Smith, "The existence, uniqueness and stability of traffic equilibria", *Transportation Research*, 13B (1979), pp. 295–304.

[Smi84] M. J. Smith, "Two alternative definitions of traffic equilibrium", *Transportation Research*, 18B (1984), pp. 63–65.

[Sol03] M. V. Solodov, "Convergence rate analysis of iterative algorithms for solving variational inequality problems", *Mathematical Programming*, 96 (2003), pp. 513–528.

[SoS99a] M. V. Solodov and B. F. Svaiter, "A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator", *Set-Valued Analysis*, 7 (1999), pp. 323–345.

[SoS99b] M. V. Solodov and B. F. Svaiter, "A hybrid projection-proximal point algorithm", *Journal of Convex Analysis*, 6 (1999), pp. 59–70.

[SpF89] H. Spiess and M. Florian, "Optimal strategies: A new assignment model for transit networks", *Transportation Research B*, 23 (1989), pp. 83–102.

[SuK92] 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.

[SFT87] C. Suwansirikul, T.L. Friesz and R.L. Tobin, "Equilibrium decomposed optimization: A heuristic for the continuous equilibrium network design problem", *Transportation Science*, 21 (1987), pp. 254–260.

[ToF88] R. L. Tobin and T. L. Friesz, "Sensitivity analysis for equilibrium network flow", *Transportation Science*, 22 (1988), pp. 242–250.

[ToW96] Ph. Toint and L. Wynter, "Asymmetric multiclass traffic assignment: A coherent formulation", *Transportation and Traffic Theory, Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, 24–26 July, 1996, J.-B. Lesort, ed., 1996, pp. 237–260.

[WXW01] Y. J. Wang, N. H. Xiu, and C. Y. Wang, "Unified framework of extragradient-type methods for pseudomonotone variational inequalities", *Journal of Optimization Theory and Applications*, 111 (2001), pp. 641–656.

[War52] J. G. Wardrop, "Some theoretical aspects of road traffic research", *Proceedings of the Institute of Civil Engineers, Part II*, (1952), pp. 325–378.

[Wat99] D. Watling, "Stability of the stochastic equilibrium assignment problem: a dynamical systems approach", *Transportation Research B*, 33 (1999), pp. 281–312.

[YaH04] H. Yang and H.-J. Huang, "The multi-class, multi-criteria traffic network equilibrium and systems optimum problem", *Transportation Research B*, 38 (2004), pp. 1–15.

[ZhM96] D. L. Zhu and P. Marcotte, "Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities", *SIAM Journal on Optimization*, 6 (1996), pp. 714–726.

[ZPP69] S. I. Zuhovickiĭ, R. A. Polyak, and M. E. Primak, "Two methods of search for equilibrium points of n-person concave games", *Soviet Mathematics Doklady*, 10 (1969), pp. 279–282.

A A primer on variational inequality problems

A.1 Definition and formulations

Equilibrium problems naturally fit the variational inequality framework, which constitutes an extension of equation systems, in contrast with optimization problems. A *variational inequality* $\mathrm{VIP}(\boldsymbol{f}, X)$ in $\Re^n$ is characterized by a non-empty, closed and convex set $X \subseteq \Re^n$ and a mapping $\boldsymbol{f}$ from $\Re^n$ into $\Re^n$. A vector $\boldsymbol{x} \in X$ is a solution to $\mathrm{VIP}(\boldsymbol{f}, X)$ if and only if it satisfies

$$\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in X. \tag{93}$$

If one denotes by $N_X(\boldsymbol{x})$ the normal cone to the set X at the point $\boldsymbol{x} \in X$, defined as

$$N_X(\boldsymbol{x}) = \{\, \boldsymbol{p} \in R^n \mid \boldsymbol{p}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \quad \boldsymbol{y} \in X \,\}, \tag{94}$$

then $\boldsymbol{x}$ is a solution to $\mathrm{VIP}(\boldsymbol{f}, X)$ if and only if it satisfies the *generalized equation* (or, *normal cone inclusion*)

$$-\boldsymbol{f}(\boldsymbol{x}) \in N_X(\boldsymbol{x}). \tag{95}$$

If $\boldsymbol{f}$ is a gradient mapping, that is, there exists a real function ϕ such that $\boldsymbol{f} \equiv \nabla\phi$, then solving $\mathrm{VIP}(\boldsymbol{f}, X)$ reduces to finding a point that satisfies the first-order optimality condition of the mathematical program to

$$\underset{\boldsymbol{x} \in X}{\operatorname{minimize}} \ \phi(\boldsymbol{x}). \tag{96}$$

If $X = \{\, \boldsymbol{x} \in \Re^n \mid g_i(\boldsymbol{x}) \leq 0, i = 1, \dots, m \,\}$ and the functions g_i, $i = 1, \dots, m$, are continuously differentiable, then, under suitable constraint qualifications (CQs), any solution $\boldsymbol{x}$ to $\mathrm{VIP}(\boldsymbol{f}, X)$, together with a vector of multipliers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$, satisfies the Karush–Kuhn–Tucker system

$$g_i(\boldsymbol{x}) \leq \mathbf{0}^m, \tag{97a}$$

$$\boldsymbol{f}(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}) = \mathbf{0}^n, \tag{97b}$$

$$\lambda_i g_i(x) = 0, \qquad i = 1, \dots, m, \tag{97c}$$

$$\lambda_i \geq 0, \qquad i = 1, \dots, m. \tag{97d}$$

If $X = \Re^n_+$, the Karush–Kuhn–Tucker system reduces to the *nonlinear complementarity* problem

$$\boldsymbol{x} \geq \mathbf{0}^n, \qquad \boldsymbol{f}(\boldsymbol{x}) \geq \mathbf{0}^n, \qquad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{f}(\boldsymbol{x}) = 0, \tag{98}$$

or, using a more 'geometric' shorthand notation:

$$\mathbf{0}^n \leq \boldsymbol{x} \perp \boldsymbol{f}(\boldsymbol{x}) \geq \mathbf{0}^n. \tag{99}$$

If $X \neq \Re^n_+$, the KKT system (97) can still be formulated as a nonlinear complementarity problem, through a suitable redefinition of the primal and dual variables. It follows that both formulations are equivalent.

Variational inequalities can also be formulated as fixed point problems. Two such examples, involving either the projection operator Proj_X over the set X, or a linear program defined over X, are

$$\boldsymbol{x} = \mathrm{Proj}_X(\boldsymbol{x} - \alpha\boldsymbol{f}(\boldsymbol{x})), \qquad \alpha > 0; \tag{100}$$

$$\boldsymbol{x} \in \arg \underset{\boldsymbol{y} \in X}{\operatorname{minimum}} \ \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}). \tag{101}$$

A.2 Monotonicity properties

A mapping $\boldsymbol{f}$ is *monotone* over X if

$$[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})]^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) \geq 0, \qquad \boldsymbol{x}, \boldsymbol{y} \in X. \tag{102}$$

It is *strictly monotone* over X if the above inequality is strict whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and *strongly monotone* if there exists a positive number β such that

$$[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})]^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}) \geq \beta\|\boldsymbol{x} - \boldsymbol{y}\|^2, \qquad \boldsymbol{x}, \boldsymbol{y} \in X. \tag{103}$$

If $\boldsymbol{f}$ is continuously differentiable, it is monotone over X if and only if its Jacobian $\nabla\boldsymbol{f}(\boldsymbol{x})$ is positive semidefinite over X, that is, $(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}\nabla\boldsymbol{f}(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \geq 0$ holds for every pair $\boldsymbol{x}, \boldsymbol{y} \in X$.[27] Monotonicity is to variational

[27]Beware: this is not equivalent to $\nabla\boldsymbol{f}(\boldsymbol{x})$ being positive semidefinite for every $\boldsymbol{x} \in X$.

inequality problems what convexity is to optimization. Indeed, if $\boldsymbol{f}$ is the gradient of ϕ, then $\boldsymbol{f}$ is (strictly, strongly) monotone if and only if ϕ is (strictly, strongly) convex. Other related concepts of interest used in this chapter are

$$\begin{array}{rlcl} \textit{pseudo-monotonicity:} & \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) \geq 0 & \Longrightarrow & \boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) \geq 0, \quad \boldsymbol{x},\boldsymbol{y} \in X; \\ \textit{monotonicity}^{+}\textit{:} & [\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{y})]^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = 0 & \Longrightarrow & \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{y}). \end{array}$$

A mapping that is monotone (respectively pseudo-monotone) and satisfies the additional condition that

$$\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = 0 \qquad \Longrightarrow \qquad \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{y})$$

is *monotone*$_*^+$ (respectively *pseudo-monotone*$_*^+$). Pseudo-monotonicity$_*^+$ ensures that the value of $\boldsymbol{f}$ is constant over the solution set of VIP$(\boldsymbol{f}, X)$. Note that the stronger monotone$_+$ condition is satisfied by convex gradient mappings.

If $\boldsymbol{f}$ is pseudo-monotone and continuous, an equivalent formulation of the variational inequality (93) is the Minty (or dual) variational inequality

$$\boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in X. \tag{104}$$

A.3 Existence and uniqueness

Assume that the mapping $\boldsymbol{f}$ is continuous. If the set X is compact, the solution set X^* of VIP$(\boldsymbol{f}, X)$ is nonempty (and compact). If X is closed but unbounded, *coercivity* of $\boldsymbol{f}$, that is, the existence of a point $\boldsymbol{x}^0 \in X$ such that

$$\lim_{\substack{\|\boldsymbol{x}\| \to \infty \\ \boldsymbol{x} \in X}} \|\boldsymbol{f}(\boldsymbol{x})\| / \|\boldsymbol{x}-\boldsymbol{x}^0\| = \infty,$$

yields the same conclusion. Uniqueness of the solution usually requires that the mapping $\boldsymbol{f}$ be strictly monotone over X.

A.4 Algorithms

The fixed point formulation (100) suggests the iteration

$$\boldsymbol{x}^{\tau+1} = \mathrm{Proj}_X(\boldsymbol{x}^{\tau} - \alpha \boldsymbol{f}(\boldsymbol{x}^{\tau})), \qquad \tau = 1, 2, \dots, \tag{105}$$

which converges if $\boldsymbol{f}$ is strongly monotone on X and α is sufficiently small, although it is possible to modify this basic scheme and achieve convergence under weaker assumptions. A similar strategy is not directly applicable to the fixed point formulation (101), it is yet possible to construct an algorithm that finds a solution of VIP$(\boldsymbol{f}, X)$ by minimizing the *gap function*

$$\mathrm{gap}(\boldsymbol{x}) = \underset{\boldsymbol{y} \in X}{\mathrm{maximum}}\ \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}). \tag{106}$$

While the above function is in general neither convex nor differentiable, one can show that all of its stationary points are solutions to VIP$(\boldsymbol{f}, X)$, provided that $\boldsymbol{f}$ is monotone on X. The descent direction, built around the solutions of the LP problem to

$$\underset{\boldsymbol{y} \in X}{\mathrm{minimize}}\ \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{y},$$

can, for traffic equilibrium models, be obtained simply by computing shortest route trees rooted at each origin (or, destination) node of the network. Alternatively, under pseudo-monotonicity, one can minimize the dual gap function

$$\mathrm{dgap}(\boldsymbol{x}) = \underset{\boldsymbol{y} \in X}{\mathrm{maximum}}\ \boldsymbol{f}(\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}). \tag{107}$$

The resulting convex program can be expressed as a semi-infinite linear program and solved using cutting-plane methods.

A general framework for solving variational inequality problems consists in iteratively solving an approximation of the original problem (93), of a simpler form. One sets $\boldsymbol{x}^{\tau+1}$ to the solution of VIP$(\tilde{\boldsymbol{f}}(\boldsymbol{x}^{\tau}, \cdot), X)$ where $\tilde{\boldsymbol{f}}$ satisfies the condition that $\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in X. Under relevant choices of the approximation mapping $\tilde{\boldsymbol{f}}$, one recovers the Jacobi, Gauss–Seidel, Newton, quasi-Newton and projection algorithms, and many others. In particular, if ψ is a strongly convex function, the approximation

$$\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}) + \nabla\psi(\boldsymbol{y}) - \nabla\psi(\boldsymbol{x}), \qquad \boldsymbol{x}, \boldsymbol{y} \in \Re^n, \tag{108}$$

suggests the continuously differentiable merit function

$$\phi(\boldsymbol{x}) := \underset{\boldsymbol{y} \in X}{\operatorname{maximum}} \, [\boldsymbol{f}(\boldsymbol{x}) - \nabla \psi(\boldsymbol{x})]^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{y}) - \psi(\boldsymbol{y}), \qquad \boldsymbol{x} \in X. \tag{109}$$

Under suitable monotonicity conditions on $\boldsymbol{f}$, the solution $\boldsymbol{y}(\boldsymbol{x})$ to (109) yields a feasible descent direction $\boldsymbol{p} := \boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{x}$ for ϕ, around which a convergent algorithm can be designed. In particular, if one sets $\psi(\boldsymbol{x}) = \frac{\alpha}{2}\|\boldsymbol{x}\|^2$, the solution to (109) is precisely the projection $\mathrm{Proj}_X(\boldsymbol{x} - \alpha \boldsymbol{f}(\boldsymbol{x}))$.

Bibliographical notes

A classical reference on variational inequality problems, mostly in the infinite-dimensional setting, is [GLT81]. Several results pertaining to finite-dimensional variational inequalities appear in the monograph [Aus76]. A more recent and comprehensive treatment is the twin volume [FaP03a, FaP03b]. Algorithms based on cost approximations, of which the projection method is the quintessential example, have been analyzed in [Coh88, Pat94b, Pat98] and several others.

B A summary of key notations

$\boldsymbol{h} = (h_r)$:	vector of flows on routes indexed by r
$\boldsymbol{c}(\boldsymbol{h}) = c_r(\boldsymbol{h})$:	vector of route travel costs
$\boldsymbol{d} = (d_k)$:	vector of demands for commodity (OD pair) k
$\boldsymbol{\pi} = (\pi_{pq})$:	vector of travel costs for origin-destination couples (p, q)
$\boldsymbol{g}(\boldsymbol{\pi})$:	demand function
$\boldsymbol{\Gamma}$:	route-OD incidence matrix
$\boldsymbol{v} = (v_l)$:	vector of flows along links indexed by l
$\boldsymbol{t}(\boldsymbol{v}) = (t_l(\boldsymbol{v}))$:	vector of traversal costs on links indexed by l
$\boldsymbol{\Lambda}$:	link-route incidence matrix
$\boldsymbol{E}$:	link-node incidence matrix
$\boldsymbol{w}_k$:	vector of OD flows
$\boldsymbol{\xi}(\boldsymbol{d})$:	inverse demand function: $\boldsymbol{\xi} = \boldsymbol{g}^{-1}$
H_d:	set of route flows and demand vectors (variable demand case)
H:	set of route flows and demand vectors (fixed demand case)
$\widehat{H}_d$:	set of link flows and demand vectors (link-route representation)
$\widehat{F}$:	set of link flows (link-route representation, fixed demand)
F:	set of link flows (link-node representation, fixed demand)
$\widehat{F}_d$:	set of link flows (link-route representation, variable demand)
F_d:	set of link flows (link-node representation, variable demand)
$\boldsymbol{f}(\boldsymbol{v})$:	vector of out-of-pocket travel costs along links
$N_X(\boldsymbol{x})$:	normal cone to a convex set X at a point $\boldsymbol{x}$
ϕ_l:	frequency on transit line l
$\boldsymbol{\varphi}(\boldsymbol{\rho})$:	cost vector associated with a link-improvement vector $\boldsymbol{\rho}$
$\boldsymbol{c}(\boldsymbol{x})$:	vector of link costs associated with strategic flow vector $\boldsymbol{x}$
E_j^s:	preference order of strategy s at node j
π_{jk}^s:	probability of accessing node k from node j, when using strategy s