APPLIED OPTIMIZATION

Siriphong Lawphongpanich
Donald W. Hearn
Michael J. Smith (Eds.)

MATHEMATICAL AND COMPUTATIONAL MODELS FOR CONGESTION CHARGING

Springer

MATHEMATICAL AND COMPUTATIONAL MODELS FOR CONGESTION CHARGING

Applied Optimization

VOLUME 101

MATHEMATICAL AND COMPUTATIONAL MODELS FOR CONGESTION CHARGING

Edited by

SIRIPHONG LAWPHONGPANICH
University of Florida, Gainesville, Florida, U.S.A.

DONALD W. HEARN
University of Florida, Gainesville, Florida, U.S.A.

MICHAEL J. SMITH
The University of York, U.K.

Springer

Library of Congress Control Number: 2005934225

ISBN-10: 0-387-29644-1 e-ISBN: 0-387-29645-X

ISBN-13: 978-0387-29644-9

AMS codes: 90C90, 90C35, 65K10, 90C31, 90C33

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springeronline.com

Contents

Preface

Although transportation economists have been advocating tolling of urban streets as a mechanism for controlling congestion and managing travel demands for over 50 years, it is only recently that this idea has become practical. When compared to the alternative of building more roads, congestion pricing, in particular via electronic tolling, is now more attractive and has been adopted in countries around the world. Singapore implemented its Area Licensing Scheme to restrict vehicular traffic into the city's central area in 1975. Later (1988) it was renamed 'Electronic Road Pricing,' in part to reflect the use of new technology. In Norway, the first toll ring was operational in Bergen in 1986 and, subsequently, two additional toll rings were established in Oslo and Trondheim. More recently, the city of London introduced in February 2003 a five pound daily fee on cars entering the city center. In spite of public resistance to the concept of tolling, some cities in the United States have also employed congestion pricing in recent years. This is due in part to the Congestion Pricing Pilot Program established by Congress in 1991 that authorized the FHWA to enter into cooperative agreements with up to 15 state or local governments to establish, maintain, and monitor congestion pricing projects. Later, this program was given a broader scope and named the Value Pricing Pilot Program.

Papers in this volume focus on the development and the analysis of mathematical and computational models for determining tolls or setting prices in an effort to control congestion or, more generally, demands. Interestingly, the first paper by **Abrams** and **Hagstrom** discusses improving traffic flow without charging tolls. Instead, they introduce the possibility of blocking entries into certain roads. **Bai**, **Hearn**, and **Lawphongpanich** consider congestion tolls based on a system optimal traffic pattern and provide methodologies for resolving computational issues associated with using a traffic pattern that is only approximately system optimal. Then, the paper by **Bertsimas** and **Perakis** addresses the problem of setting prices when the demand as a function of price is not known, but is learned over time. In general, travelers or users of transportation networks are heterogeneous, e.g., they may value time differently. **Engelson** and **Lindberg** show in their paper that different values of time (e.g., in time or monetary units) can lead to models with different properties. Computationally, some of these models are more advantageous than others. Similarly, **Florian** assumes in his paper that there are multiple classes of users, some of whom are not willing to pay tolls, and describes approaches that have been used in many countries to predict the usage of tolled facilities among different user classes. Sensitivity analysis is a useful technique for predicting changes in, e.g., the level of congestion, due to changes in, e.g., toll prices, infrastructure, and user behavior. **Josefsson** and **Patriksson** describe a method for generating sensitivity information that is more general and efficient than those in the literature. The paper by **De Palma** and **Nesterov** proposes 'stable dynamics' models for commuters who must make parking decisions as part of their commute. **Smith** considers the 'bilevel' problem of estimating road prices and signal green-times which approximately minimizes a smooth measure of disbenefit subject to equilibrium and other constraints. He proposes a method that finds an approximate equilibrium that is stationary with respect to the measure of disbenefit. **Stewart** and **Maher**'s paper considers the problem of finding toll prices that yield the least revenue in order to minimize the financial impact on the traveling public. Their procedure is based on the stochastic user equilibrium and (deterministic) system optimal traffic assignment models. Concluding the volume, the paper by **Sumalee**, **Connors** and **Watling** considers an optimal toll design problem based on stochastic user equilibrium with Probit route-choice. Their algorithm for solving the problem uses the sensitivity information discussed in the paper by Josefsson and Patriksson.

We would like to take the opportunity to thank the authors of the papers, anonymous referees, and the publisher for helping us to produce this volume. We also want to thank Altannar Chinchuluun for his help in preparing the volume and putting all of the papers in their final form.

S. Lawphongpanich, D. W. Hearn, and M. J. Smith
Gainesville, Florida and York, England
September 2005

List of Contributors

Robert A. Abrams
Information and Decision Sciences,
Univ. of Illinois at Chicago,
Chicago, IL 60607, U.S.A.
rabrams@uic.edu

Lihui Bai
College of Business Administration,
Valparaiso University,
Valparaiso, IN 46383, U.S.A.
Lihui.Bai@valpo.edu

Dimitris Bertsimas
Sloan School of Management,
MIT,
Cambridge, MA 02139, U.S.A.
dbertsim@mit.edu

Richard Connors
Institute for Transport Studies,
University of Leeds,
Leeds, LS2 9JT, United Kingdom
rconnors@its.leeds.ac.uk

Leonid Engelson
Department of Infrastructure,
Royal Institute of Technology,
SE-100 44 Stockholm, Sweden
lee@infra.kth.se

Michael Florian
Center for Research on
Transportation,
University of Montreal,
Montreal H3C 3J7, Canada,
mike@crt.umontreal.ca

Jane N. Hagstrom
Information and Decision Sciences,
Univ. of Illinois at Chicago,
Chicago, IL 60607, U.S.A.
hagstrom@uic.edu

Donald W. Hearn
Dept. of Industrial and Systems
Engineering,
University of Florida,
Gainesville, FL 32611, U.S.A.
hearn@ise.ufl.edu

Magnus Josefsson
Department of Mathematics,
Chalmers University of Technology,
SE-412 96 Gothenburg, Sweden
f98majf@dd.chalmers.se

Siriphong Lawphongpanich
Dept. of Industrial and Systems
Engineering,
University of Florida,
Gainesville, FL 32611, U.S.A.
lawphong@ise.ufl.edu

Per Olov Lindberg
Department of Mathematics,
Linköping University,
SE-581 83 Linköping, Sweden
polin@mai.liu.se

Mike Maher
Transport Research Institute and
School of the Built Environment,
Napier University,
Edinburgh, EH10 5DT, Scotland
m.maher@napier.ac.uk

Yurii Nesterov
Center for Operations Research and
Econometrics,
Université Catholique de Louvain,
B-1348 Louvain-la-Neuve, Belgium
nesterov@core.ucl.ac.be

André de Palma
Institut Universitaire de
France & THEMA,
Université de Cergy-Pontoise,
F-95011 Cergy-Pontoise, France
andre.depalma@eco.u-cergy.fr

Michael Patriksson
Department of Mathematics,
Chalmers University of Technology,
SE-412 96 Gothenburg, Sweden
mipat@math.chalmers.se

Georgia Perakis
Sloan School of Management,
MIT,
Cambridge, MA 02139, U.S.A.
georgiap@mit.edu

Michael J Smith
Department of Mathematics,
University of York, Heslington,
York, Y010 5DD, United Kingdom
mjs7@york.ac.uk

Kathryn Stewart
Transport Research Institute and
School of the Built Environment,
Napier University,
Edinburgh, EH10 5DT, Scotland
k.stewart@napier.ac.uk

Agachai Sumalee
Institute for Transport Studies,
University of Leeds,
Leeds, LS2 9JT, United Kingdom
asumalee@its.leeds.ac.uk

David Watling
Institute for Transport Studies,
University of Leeds,
Leeds, LS2 9JT, United Kingdom
dwatling@its.leeds.ac.uk

Improving Traffic Flows at No Cost

Robert A. Abrams[1] and Jane N. Hagstrom[2]

[1] Information and Decision Sciences, University of Illinois at Chicago, 601 S. Morgan, Chicago, IL 60607-7124, U.S.A., rabrams@uic.edu

[2] Information and Decision Sciences, University of Illinois at Chicago, 601 S. Morgan, Chicago, IL 60607-7124, U.S.A., hagstrom@uic.edu

Summary. The standard model of traffic flow used in the analysis of urban traffic is the Wardrop equilibrium. The existence of traffic flows that reduce costs for some travelers without increasing the costs for any other travelers when compared to the equilibrium defines a Generalized Braess Paradox. We provide a practical methodology for detecting such flows and report the existence of such a flow in the Sioux Falls study network when links with equilibrium flows in the free-flow range are regarded as constant cost.

Key words: Multicommodity Traffic, Noncooperative Equilibrium, Nonlinear Programming, Braess Paradox.

1 Introduction

Traffic congestion is becoming a more and more pressing issue for society and a major concern for urban planners. In 1968, Braess [Bra68] identified the possibility that more roads can make traffic worse. In this paper, we take an "inverse" view, that is, that fewer roads, or more-restricted roads, can make traffic better. Specifically, we look for situations in which the total cost of congestion is reduced at no additional cost to any traveler. We provide a methodology for identifying such situations and demonstrate that the Sioux Falls study network is an example in which restricting traffic on certain links leads to 33% lower travel times for some travelers while costs for other travelers increase no more than 0.25%.

In urban road networks, individual travelers decide on their own travel routes on the basis of factors such as time, cost, and convenience. Since they are not acting cooperatively, it is not surprising that these individually chosen routes are not best from society's point of view. In this paper we show how to detect cases where redirecting traffic flows reduces the travel time for some travelers while not increasing travel time for any travelers. Since this redirection can be enforced by restricting access to certain links in the network, or

Mathematical and Computational Models for Congestion Charging, pp. 1-22

by imposing tolls, it is possible to improve society's traffic costs while costs to individual travelers are reduced or remain the same.

The standard model of traffic flow assumes that travelers distribute themselves according to Wardrop's user-equilibrium principle. (See [Pat94, Chapter 2], and [War52].) This principle states that all used paths between an origin-destination pair will have the same cost, which is no more than the cost on any unused path. Cost is measured as time or some combination of tolls, time, and other factors. Braess [Bra68] used this model to construct a seemingly paradoxical example in which adding a link to a simple network results in a user-equilibrium distribution of flows that is worse for all travelers than the network without the added link. One can also view Braess's network example with the added link as an example in which a nonequilibrium flow (with no flow on the added link) reduces costs for all travelers.

In a previous paper [HA02], the authors defined a Generalized Braess Paradox to occur whenever there is an alternative distribution of flows which makes some travelers better off and none worse off than in the Wardrop equilibrium distribution. In game-theoretic terms, a Generalized Braess Paradox occurs whenever the user equilibrium is not strongly Pareto optimal. In this paper we show how to detect a Generalized Braess Paradox and report the detection of a Generalized Braess Paradox in the widely known Sioux Falls study network, when certain links with equilibrium flows in the free-flow range are assigned constant cost. We thus demonstrate the feasibility of detecting opportunities in which society can improve its total costs without increasing the cost to any individual travelers. The procedure that we develop will also detect occurrences of the "classic" Braess Paradox, in which removing a link results in improved travel cost.

This work is related to, but distinct from, work on finding system-optimal flows in a network. A system-optimal flow in a traffic network minimizes the sum of the costs of all travelers. A system-optimal flow is desirable from society's point of view because it minimizes consumption of resources and production of pollution. The system-optimal flow is usually distinct from the user-equilibrium flow, but typically will require some travelers to incur higher travel costs than in the equilibrium flow. Braess's example is one in which the system-optimal distribution demonstrates the existence of a Generalized Braess Paradox. This, however, is unusual. In more usual cases (See [HA02].), the system-optimal distribution makes some travelers worse off than in the equilibrium distribution, even when a Generalized Braess Paradox exists. Finding a distribution that demonstrates the existence of a Generalized Braess Paradox is significantly more difficult than finding a system-optimal distribution.

In [HA02], we showed that a Generalized Braess Paradox can be characterized in terms of a mathematical program. In this paper, we use that characterization to develop a method for detecting the occurrence of a Generalized Braess Paradox. We make the mathematical program of [HA02] tractable by relaxing its constraints to obtain a convex mathematical programming problem that will detect occurrences of the Generalized Braess Paradox. However, due to the particular structure of the relaxed problem, first-order optimality conditions may not hold for the optimal solution, thus rendering inapplicable any algorithm based on standard first-order conditions. Therefore we adopt a special method to solve the problem. The method first uses a sequence of linear programs to identify which nonlinear constraints always hold as equalities, and then whether a Generalized Braess Paradox exists. The number of linear programs is no more than the number of links in the network and usually much less. We apply the method to two small examples and to the well-known Sioux Falls study network with 24 nodes, 76 links, and 528 origin-destination pairs. A Generalized Braess Paradox is found to occur in the Sioux Falls study network when links with equilibrium flows in the free-flow range are regarded as constant cost. The second of the small examples illustrates that the first-order optimality conditions (Karush-Kuhn-Tucker conditions) cannot be expected to hold for the optimal solution to the relaxed mathematical program, even though it is a convex nonlinear programming problem.

2 Notation and Definition of the Equilibrium Problem

We consider a transportation network with multiple origin-destination (*o-d*) pairs. Depending on circumstances, demand (usually given as a trip table, specifying for each origin-destination pair the volume per unit time of travelers desiring to move between that origin and destination) may be either elastic or fixed. For the purposes of this paper we assume fixed demand.

We make the following two assumptions. Neither is restrictive in that if either fails to hold, existing methods in the literature (See, e.g., [AM81].) can be used to reduce these cases to situations satisfying the assumptions.

1. Travel costs are *additive*, that is, the travel cost of a route is the sum of the traversal costs of the links on the route.
2. The cost of traversing a link is the same for all travelers, and the cost depends on the vector of total link flows, where the *total flow on a single link is the sum of the individual flows on the link between each of the origin-destination pairs.*

An *equilibrium distribution* of flows is a distribution of flows that meets demands and satisfies Wardrop's User-Equilibrium Principle, i.e., every used path between an o-d pair must have the same cost, and all unused paths between the same o-d pair must have cost greater than or equal to that of the used paths. A Wardrop equilibrium corresponds, in a game-theoretic framework, to a (noncooperative) Nash equilibrium. (See [Pat94, pages 32, 54].)

For a Wardrop equilibrium to be reached, one must assume that travelers have perfect information about travel costs and act to minimize their individual travel costs. Although this may seem to be a strong assumption, most models used in traffic network analysis and planning assume that traffic will be distributed according to a Wardrop equilibrium.

2.1 Notation

As is common in traffic flow theory, our model is built on a network structure with travel costs on each link and known supplies and demands for each node. Our notation accounts for the network structure, properties of links, and properties of the travelers using the network.

Table 1. Notation

k	a link
$t(k)$	the node that link k is directed out of
$h(k)$	the node that link k is directed into
i	a node
d	a destination node
$\mathcal{N}$	the set of nodes in the network
$\mathcal{A}$	the set of links in the network
$\mathbf{A}$	the $\lvert\mathcal{N}\rvert \times \lvert\mathcal{A}\rvert$ node-link incidence matrix of the network
$\mathcal{D}$	the set of destination nodes
b_i^d	for $i \neq d$, the demand for travel from node i to destination d
b_d^d	the negative of the sum of all demands for travel to destination d
$\mathcal{O}^d$	the set of nodes with positive demand for travel to destination d
x_k^d	the amount of flow on link k destined for d
$\mathbf{x}^d$	the vector of link flows destined for d
x_k	the total flow on link k
$\mathbf{x}$	the vector of total link flows
u_i^d	a price (or potential) for node i associated with destination d
$\mathbf{u}^d$	the vector of node prices associated with destination d
z_k^d	a surplus quantity associated with link k and destination d
$\mathbf{z}^d$	the vector of link surpluses associated with destination d
$F_k(\mathbf{x})$	the traversal cost for link k of one unit of flow
$\mathbf{F}(\mathbf{x})$	the vector of link costs
$\bar{x}_k^d$	the equilibrium solution flow on link k destined for d
$\bar{x}_k$	the total flow on link k in the equilibrium solution
$\bar{u}_i^d$	the equilibrium cost of traveling from node i to destination d
$\bar{z}_k^d$	the reduced cost $F_k(\bar{\mathbf{x}}) - \bar{u}_{t(k)}^d + \bar{u}_{h(k)}^d$

Table 1 summarizes the notation we use. In addition, the elements $a_{i,k}$ of the node-link incidence matrix $\mathbf{A}$ are defined by

$$a_{i,k} = \begin{cases} 1 \text{ if link } k \text{ is directed out of node } i \\ -1 \text{ if link } k \text{ is directed into node } i \\ 0 \text{ otherwise.} \end{cases}$$

In describing the flow on the network, we partition travelers according to their destination. In our previous work, we partitioned travelers according to both their origin and destination. The latter approach is conceptually easier; however, from a computational point of view, the smaller number of classes of travelers is desirable, and does not lose any generality. It is well known (LeBlanc, [LeB73]) that it is not necessary to discriminate between travelers starting from different origins if they are bound for the same destination, or equivalently, that it is not necessary to discriminate between travelers bound for different destinations if they have all started at the same origin.

2.2 The Equilibrium Problem

The Wardrop equilibrium solution can be defined in several equivalent ways (See [Pat94, BMW55, Roc80].), e.g., as a variational principle, as the solution of an optimization problem, etc. The particular formulation chosen turns out to be critical in developing a tractable characterization of the Generalized Braess Paradox. For this purpose, we use a Lagrange multiplier definition. For given $\mathbf{A}$, $\mathbf{F}$, and demand vectors $\mathbf{b}^d$, the equilibrium problem can be expressed as seeking a solution to

(EQ)

$$\begin{aligned}
\mathbf{F}(\mathbf{x}) - \mathbf{A}^T\mathbf{u}^d - \mathbf{z}^d &= \mathbf{0} \; \forall\, d \in \mathcal{D} && (1)\\
\mathbf{A}\mathbf{x}^d &= \mathbf{b}^d \qquad \forall\, d \in \mathcal{D} && (2)\\
\mathbf{x} &= \textstyle\sum_{d\in\mathcal{D}} \mathbf{x}^d && (3)\\
\textstyle\sum_{d\in\mathcal{D}} \mathbf{z}^d \cdot \mathbf{x}^d &= 0 && (4)\\
\mathbf{x}^d &\geq \mathbf{0} \qquad \forall\, d \in \mathcal{D} && (5)\\
\mathbf{z}^d &\geq \mathbf{0} \qquad \forall\, d \in \mathcal{D} && (6)\\
u_d^d &= 0 \qquad \forall\, d \in \mathcal{D} && (7)
\end{aligned}$$

Equation Sets (1) and (4) state that on a link with positive flow destined for destination d, the cost of travel on the link k, $F_k(\mathbf{x})$, is equal to the price difference, u_i^d-u_j^d, corresponding to destination d, between the two end nodes, i and j, of the link. If there is no flow directed towards d on link k, then (4) allows z_k^d to be positive and the cost of travel on link k may be greater than or equal to the difference in prices. Equation Set (2) requires that flows directed toward d satisfy demand at origin and destination nodes and conserve flow at other nodes. Equation (3) defines the total flow on a link to be the sum of the flows on that link headed to the different destinations. There is always one node price u_i^d for each d that is arbitrary. Equation Set (7) removes this ambiguity by defining the price at the destination nodes to be zero. An equilibrium solution, denoted $\{(\bar{\mathbf{x}}^d, \bar{\mathbf{u}}^d, \bar{\mathbf{z}}^d)\}_{d\in\mathcal{D}}$, is a solution to (EQ). In an equilibrium solution, the node price $\bar{u}_i^d$ is the cost of traveling from node i to destination d along links with $\bar{x}_k^d > 0$.

3 The Existence of Improved Flows

Given a Wardrop equilibrium set of flows, we wish to determine whether there is another distribution of flows that makes some travelers better off and no travelers worse off than in this equilibrium. To that end, we define a nonlinear program which minimizes system cost subject to the constraint that no traveler has cost greater than in the given equilibrium. The constraints are similar to those of the equilibrium problem, (EQ), except that instead of requiring that the traversal cost on a used link equal the price difference of its nodes, we allow the traversal cost of the link to be less than or equal to the price difference of its nodes. In this way, the formulation allows nonequilibrium flows, and as discussed following the formulation, the potential at each node becomes an upper bound on the travel cost from that node to the relevant destination.

For given $\mathbf{A}$, $\mathbf{F}$, demand vectors $\mathbf{b}^d$, and equilibrium travel costs $\bar{u}_s^d$, define the following optimization problem, originally introduced in [HA02], which we henceforth call the ***Equilibrium Improvement Problem***, (EIP). (In [HA02], we referred to this as the Braess Optimization Problem.)

(EIP)

$$\begin{aligned}
\min \quad & \mathbf{x} \cdot \mathbf{F}(\mathbf{x}) & & \\
\text{s.t. } & \mathbf{F}(\mathbf{x}) - \mathbf{A}^T\mathbf{u}^d - \mathbf{z}^d \leq \mathbf{0} & \forall\, d \in \mathcal{D} & \qquad (8)\\
& \mathbf{A}\mathbf{x}^d = \mathbf{b}^d & \forall\, d \in \mathcal{D} & \qquad (9)\\
& \mathbf{x} = \textstyle\sum_{d\in\mathcal{D}} \mathbf{x}^d & & \qquad (10)\\
& \textstyle\sum_{d\in\mathcal{D}} \mathbf{z}^d \cdot \mathbf{x}^d = 0 & & \qquad (11)\\
& \mathbf{x}^d \geq \mathbf{0} & \forall\, d \in \mathcal{D} & \qquad (12)\\
& \mathbf{z}^d \geq \mathbf{0} & \forall\, d \in \mathcal{D} & \qquad (13)\\
& u_d^d = 0 & \forall\, d \in \mathcal{D} & \qquad (14)\\
& u_s^d \leq \bar{u}_s^d & \forall\, s \in \mathcal{O}^d & \qquad (15)
\end{aligned}$$

The constraints of (EIP) are very similar to the equilibrium problem (EQ). The differences are:

1. As noted above, Constraint Set (8) is a set of inequalities instead of equations. The inequality requires that, on a link with positive flow for a particular destination, the travel cost is less than or equal to the price difference of the nodes connected by the link. Thus for any feasible, possibly nonequilibrium, flow, and for any used route from node i to destination d, the inequality constraint (8) implies that u_i^d is an upper bound on the travel cost from node i to destination d along that route.
2. There is an additional constraint set, (15), which forces the travel cost from any origin to destination to be less than or equal to that of the equilibrium flow.

When $\mathbf{x}$ is nonnegative, the components of $\mathbf{F}(\mathbf{x})$ are convex, and $\mathbf{F}(\mathbf{x})$ is monotone ($(\mathbf{y} - \mathbf{x}) \cdot (\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})) \geq 0$ for all feasible $\mathbf{x}$, $\mathbf{y}$ [HP90]), the objective function of (EIP) is convex, as is shown in Appendix B. Thus without Equation (11), (EIP) would be a convex optimization problem.

Any feasible solution to (EIP) with objective function value less than that of the equilibrium flow reduces the travel cost for some travelers and, due to the last set of constraints, does not increase the travel cost for any travelers. Thus if an equilibrium solution is not optimal for (EIP), a Generalized Braess Paradox exists. As shown in [HA02], the converse also holds when each component of $F(\mathbf{x})$ is nonnegative and nondecreasing. It follows that under these mild conditions on F_k, determining the existence of flows that improve on a Wardrop equilibrium is equivalent to testing (EIP) to see if a Wardrop equilibrium is optimal. In the following sections we will develop methods to test optimality of the Wardrop equilibrium. We first establish that if there is a feasible solution to (EIP) for which some constraint corresponding to a used link in Set (8) holds strictly, then the equilibrium solution is not optimal for (EIP) and a Generalized Braess Paradox exists.

Proposition 1. *If there exists a feasible solution to* (EIP) *with the property that for some link k and destination d,*

$$x_k^d > 0 \textit{ and } F_k(\mathbf{x}) - u_{t(k)}^d + u_{h(k)}^d < 0,$$

then a Generalized Braess Paradox exists.

Proof. Suppose that the triples $(\mathbf{x}^d, \mathbf{u}^d, \mathbf{z}^d)$ define a feasible solution to (EIP) and there exists a link k^* and destination d^* such that

$$x_{k^*}^{d^*} > 0 \text{ and } F_{k^*}(\mathbf{x}) - u_{t(k^*)}^{d^*} + u_{h(k^*)}^{d^*} < 0.$$

From constraint set (15), we know that no traveler is worse off than in equilibrium. Since $x_{k^*}^{d^*} > 0$, there exists an origin s^* which contributes flow to link k^* that is destined for d^*; more specifically, there is a path P of links k joining s^* to d^* such that $k^* \in P$ and $x_k^{d^*} > 0$ for all links $k \in P$. Since $x_k^{d^*} > 0$ for these links, $z_k^{d^*} = 0$ on these links. Then for each of these links, Constraint Set (8) gives

$$F_k(\mathbf{x}) \leq u_{t(k)}^{d^*} - u_{h(k)}^{d^*}.$$

Our assumption of strict inequality gives

$$F_{k^*}(\mathbf{x}) < u_{t(k^*)}^{d^*} - u_{h(k^*)}^{d^*}.$$

Summing over $k \in P$, and using Constraint Sets (14) and (15) we have

$$\sum_{k \in P} F_k(\mathbf{x}) < u_{s^*}^{d^*} \leq \bar{u}_{s^*}^{d^*},$$

Thus we have travelers using path P to go from s^* to t^* with a lower cost than in equilibrium. ■

4 A Computational Approach for Local Improvements

(EIP) provides a direct method of checking for the existence of a Generalized Braess Paradox by solving an optimization problem. However, for even moderately large networks (EIP) is difficult to solve because the complementarity constraint (11), which essentially defines for each destination d the subnetwork of arcs that may be used by flows destined for d, is not convex. Solving (EIP) implies the need to (implicitly or explicitly) enumerate all feasible subnetworks of the network. Since for each destination, flows may use a different subnetwork, solving (EIP) may require an extremely large enumeration. We therefore treat a more tractable version of the problem for which we can detect many instances of the Generalized Braess Paradox using a finite sequence of linear programs.

In order to develop the more tractable test, we replace the troublesome Constraint (11) with a more restrictive, but more tractable, condition. This new problem will identify a local Generalized Braess Paradox, in the sense that our search for an improved flow is restricted to using essentially the same set of links used in the Wardrop equilibrium solution. The existence of a solution of the more restrictive problem that has a lower objective function value than the equilibrium solution will guarantee the existence of a Generalized Braess Paradox. However, because the problem is more restrictive, an improved solution using a different subnetwork may remain undetected. Therefore even when the equilibrium solution is optimal for the modified problem, a Generalized Braess Paradox may exist as shown by Example 1 of [HA02]. This limitation is shared by all tests for the Braess Paradox of which we are aware ([DN84a], [SZ83]), in that none will detect a Braess Paradox that uses flows on a subnetwork distinct from that of the equilibrium solution.

The ***Restricted Equilibrium Improvement Problem***, (R-EIP), is

(R-EIP)

$$\begin{array}{lll}
\min & \mathbf{x}\cdot\mathbf{F}(\mathbf{x}) & \\
\text{s.t. } \mathbf{F}(\mathbf{x}) - \mathbf{A}^T\mathbf{u}^d - y\bar{\mathbf{z}}^d \le \mathbf{0} & \forall\, d \in \mathcal{D} & (16)\\
\mathbf{A}\mathbf{x}^d = \mathbf{b}^d & \forall\, d \in \mathcal{D} & (17)\\
\mathbf{x} = \sum_{d\in\mathcal{D}} \mathbf{x}^d & & (18)\\
\sum_{d\in\mathcal{D}} \bar{\mathbf{z}}^d \cdot \mathbf{x}^d = 0 & & (19)\\
\mathbf{x}^d \ge \mathbf{0} & \forall\, d \in \mathcal{D} & (20)\\
u_d^d = 0 & \forall\, d \in \mathcal{D} & (21)\\
u_s^d \le \bar{u}_s^d & \forall\, s \in \mathcal{O}^d & (22)
\end{array}$$

This formulation entails two changes from (EIP). Constraint Set (8) has been replaced with Constraint Set (16), which contains a new variable y. Since the variable y can be set to a very large number, when the equilibrium value $\bar{z}_k^d > 0$, the corresponding constraint is redundant just as is the case for Constraint Set (8) when $z_k^d > 0$. Constraint (11) has been replaced with Constraint (19), which requires that $x_k^d = 0$ whenever $\bar{z}_k^d > 0$. Thus any feasible solution to (R-EIP) has flow going to destination d only on links k with $\bar{z}_k^d = 0$.

As previously noted, if $\mathbf{F}$ is convex and monotone, the objective function of (R-EIP) is convex. Therefore (R-EIP) is a convex optimization problem. All constraints except those involving $\mathbf{F}$ are linear. Our aim is to determine if the Wardrop equilibrium solution $\{(\bar{\mathbf{x}}^d, \bar{\mathbf{u}}^d, \bar{\mathbf{z}}^d)\}_{d \in \mathcal{D}}$ is optimal for (R-EIP). If a constraint qualification held for the problem, one might use the first-order necessary (KKT) conditions. However, as shown by Example 2 in Section 5, even when the equilibrium solution is optimal for (R-EIP), the KKT conditions do not necessarily hold. Therefore no constraint qualification can be assumed to hold for the problem, and methods other than those based on the standard first-order conditions must be used.

Convex programming problems for which no constraint qualification holds have been studied extensively by Ben-Israel, Ben-Tal and Zlobec [BBZ81]. Using their approach and an algorithm proposed by Kerzner [AK78], we first determine if one or more nonlinear constraints hold as strict inequalities for some feasible solution. If even one such (nonvacuous) constraint exists, Proposition 1 states that there is a Generalized Braess Paradox. If it is determined that no such constraint exists, we formulate a single linear program that searches for a feasible direction of improvement. The existence of such a direction will establish the existence of a Generalized Braess Paradox. If no such direction exists, then the Wardrop equilibrium solution is optimal for (R-EIP).

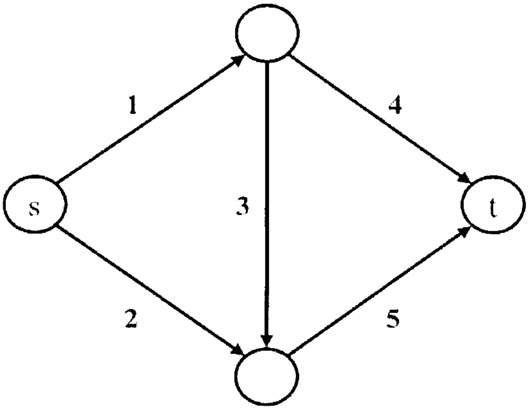


Fig. 1. Bridge Network

In our case of faithfully convex differentiable functions (See Appendix A for a definition.), Kerzner's algorithm for finding the constraints that can be satisfied strictly solves a sequence of linear programs. The number of linear programs that must be solved is no more than the number of constraints and usually far less. In a network model of form (R-EIP), we point out in Appendix A that the number of linear programs that must be solved is bounded by the number of links. The model (R-EIP) for the Sioux Falls study network has 1655 nonlinear constraints after removing those corresponding to positive $\bar{z}_k^d$, but requires the solution of only three linear programs to determine which constraints can be satisfied strictly. The algorithm as adapted for (R-EIP) is described in Appendix A.

5 Computational Results

We consider three examples in detail. The first two use the five-link bridge network studied by Braess [Bra68] to illustrate his paradox, and the third is the well-known Sioux Falls study network [Bar02]. The first of the small examples is a straightforward application of the method as described in Appendix A. The second example on the same network is a case in which there is no Generalized Braess Paradox, that is, the equilibrium solution is optimal for (EIP), but the Karush-Kuhn-Tucker conditions do not hold. The data, models and numerical results for all three examples are given in [Hag04].

Table 2. Data for Example 1, with a Generalized Braess Paradox but no Classic Braess Paradox

		Equilibrium	System Optimal	(R-EIP) Optimal	Equilibrium without Link 3
Link	Cost Function	Link Flows			
1	$1.4x_1$	4.000	3.630	3.464	2.914
2	$5.4+\sqrt{4x_2^2+9}$	2.000	2.370	2.536	3.086
3	$2.4x_3$	2.000	1.010	1.011	0.000
4	$7.8+\sqrt{4x_4^2+9}$	2.000	2.620	2.453	2.914
5	$2x_5$	4.000	3.380	3.547	3.086
Route		Route Costs			
1, 4		18.400	18.920	18.400	18.434
1, 3, 5		18.400	14.266	14.372	
2, 5		18.400	17.770	18.387	18.434
Most Costly Used Route		18.400	18.920	18.400	18.434
System Cost		110.400	106.093	106.293	110.607

Example 1. The network for Example 1 is shown in Figure 1. The demand for travel between the origin s and the destination t is 6 units of flow. The

cost functions for the five links, the equilibrium solution, the system optimal solution, and an improved solution illustrating a Generalized Braess Paradox are shown in Table 2. Note that the cost functions are strictly convex in their arguments, and monotone, and as a result (See Appendix B.) the objective function of (EIP) is convex. There is no classic Braess Paradox for this problem because, as is also shown in Table 2, eliminating Link 3 does not result in an improved equilibrium travel cost from node s to node t. Because the problem is small, it is a simple matter to solve (EIP) or (R-EIP) directly to find a Generalized Braess Paradox if one exists. The (R-EIP) optimal solution shown in Table 2 reduces the travel cost for travelers using the route consisting of Links 1, 3, and 5 by 22 percent, and does not increase the cost for any other travelers, thus establishing the existence of a Generalized Braess Paradox. In this particular example, any convex optimizer can be counted on to give a correct solution to (R-EIP) because the nonlinear constraints can be satisfied strictly for some feasible solution. The details of our general approach as applied to this example are given in Appendix A.

Example 2. The network for Example 2 is the same simple network as for Example 1, with the same demand for travel. The cost structure has been changed to eliminate the occurrence of a Generalized Braess Paradox. The equilibrium solution is optimal for (R-EIP). However, the first order optimality conditions do not hold at the equilibrium solution. The link costs, the equilibrium solution and the system-optimal solution are shown in Table 3.

Table 3. Data for Example 2, with No Generalized Braess Paradox

Link	Cost Function	Equilibrium	System Optimal
1	$1.6x_1$	4.00	3.52
2	$5.4 + \sqrt{4x_2^2 + 9}$	2.00	2.48
3	$2x_3$	2.00	1.04
4	$5.4 + \sqrt{4x_4^2 + 9}$	2.00	2.48
5	$1.6x_5$	4.00	3.52
Highest Used Route Cost		16.80	17.99
System Cost		100.8	97.3

Applying the algorithm described in Appendix A to (R-EIP), we find at the first iteration that both of the nonlinear constraints (those in (16) corresponding to Links 2 and 4) must hold with equality for all feasible solutions of (R-EIP). We then conclude (See Appendix A.) that the flows on Links 2 and 4 are constant for all feasible solutions of (R-EIP). Due to the simple structure of this example, it is immediately apparent that if there are no feasible changes to the flows on Links 2 and 4, there are no feasible changes to the flows on the other three links. Thus we know that the equilibrium solution is optimal for (R-EIP) and there is no local Generalized Braess Paradox.

If the example were not so simple, we would replace the arguments of the strictly convex cost functions for Links 2 and 4 by their constant values to obtain linear constraints. (R-EIP) then becomes a nonlinear program with convex objective function and linear constraints. Thus we have converted a formulation of the problem for which the Slater condition does not hold to one for which it does. Solving a linear program that searches for a feasible direction of decrease leads to the conclusion that the equilibrium solution is optimal for (R-EIP) and there is no local Generalized Braess Paradox. This example illustrates a case in which the equilibrium solution is optimal for (R-EIP), yet the Karush-Kuhn-Tucker conditions do not hold for the original formulation.

Example 3. The Sioux Falls study network is often used as a test network for transportation models. It consists of 24 nodes, 76 one-way links and 528 origin-destination pairs. Thus, most of the nodes are both origins and destinations. The network is shown in Figure 2.

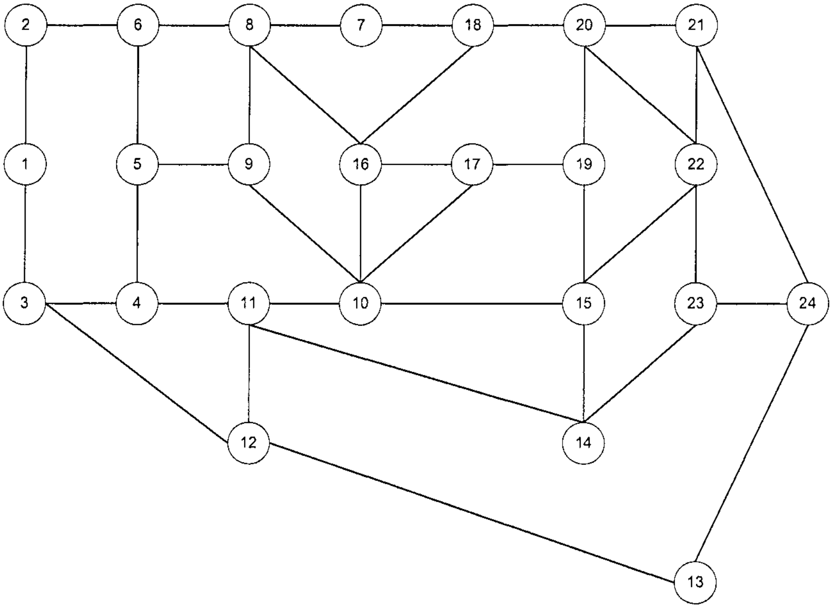


Fig. 2. Sioux Falls Network

The network structure, the trip table specifying required flows, and an accurate equilibrium solution can be found at [Bar02]. The linear programs described in Appendix A were formulated using the LINGO modeling language. (Details are available at [Hag04].) For Bar-Gera's equilibrium solution [Bar02], the linear programs have approximately 2500 constraints and 4000 variables. With one important modification described below, the cost function used is the fourth-power polynomial common in traffic analysis; it is also available at [Bar02].

The standard description of traffic flow time (cost) on a link is that it is constant in the low volume free-flow range and increasing for larger flows, with a sharp rise for flows in excess of a specific "capacity" level. This behavior is usually modeled with a fourth degree polynomial. Although the fourth degree polynomial does represent the desired behavior fairly well, it is strictly increasing in any range no matter how small the flow on the link. While this may be of little importance in most situations, it can be important in the solution of (EIP). Typically many constraints in Set (15) will hold with equality for all feasible solutions. In such cases a flow that is otherwise feasible for (EIP) might be eliminated because it mathematically violates the constraint even though the "violation" may be so small as to have no practical significance. To minimize the possibility of this occurring, we replace the fourth degree polynomial by a constant function on links for which the equilibrium flows are under half of the given capacity level, that is, for those well in the free-flow range. After finding the optimal changes for (R-EIP) using the constant cost functions for the free-flow ranges, we recalculated link costs using the original quartic functions. We found that while reducing some travelers costs by 33%, no link costs increased by more than 0.25%. These increases occur on the arcs that were set to a constant cost, and the flows on these links were still under 50% of capacity, again in the free-flow range. With the original quartic functions in (R-EIP), an improved solution was not found. This is due to the strictly increasing cost function and the dense structure of the network. A similar result to that obtained with the modified cost function can be found by retaining the quartic functions and relaxing Constraint Set (15) to allow the cost of travel to slightly exceed the equilibrium value.

After solving the first LP as described in Appendix A (See also [Hag04].), we found that 64 of the 68 (one-way) links with nonlinear cost functions must have constant cost for all feasible flows. As pointed out in Appendix A, this implies that the total flow on these links must be constant for all feasible solutions, and the nonlinear constraints may be replaced by linear constraints. Solving a second LP added two links to the set with constant cost. After solving a third LP, we found that nonlinear constraints corresponding to the last two links [5-6] and [6-5] can be satisfied strictly. All three of the LP's were solved in seconds on a desktop Windows machine using LINGO. The dual prices from the third LP give the changes in destination-based flows that will make links [5-6] and [6-5] have costs that are lower than the price differentials. This interior direction involves changing flows around two circuits, each involving a separate destination. The magnitudes of all the changes are equal. The links that have changes in flow, and the directions of these changes, are shown in Figure 3. By Proposition 1, these changes demonstrate the occurrence of a Generalized Braess Paradox for the Sioux Falls study network.

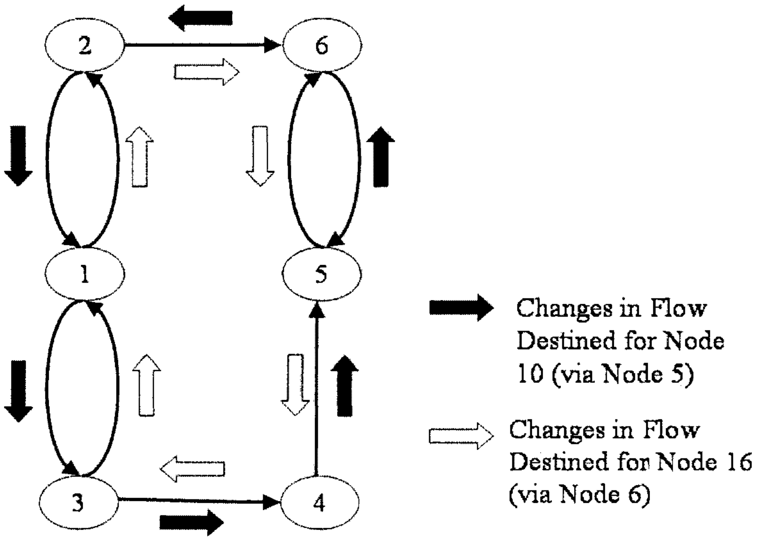

2
6
1
5
3
4
Changes in Flow Destined for Node 10 (via Node 5)
Changes in Flow Destined for Node 16 (via Node 6)

Fig. 3. Improving Changes in Flow for Sioux Falls Network

After determining that a Generalized Braess Paradox exists and that all costs were constant except those corresponding to links [5-6] and [6-5], we found the optimal solution to (R-EIP). In the optimal solution the total flow changes were consistent with the direction found at the third iteration of the algorithm. However, different destination flow changes occurred. Details may be found at [Hag04]. As might be expected in multiple origin-destination models, these changes are not unique, and may be different each time the problem is solved.

6 Conclusions

The method presented in this paper identifies situations in which, when compared with the Wardrop equilibrium, alternate flows exist that reduce cost for some travelers without increasing cost for any other travelers. The network presented by Braess [Bra68] is an example of such a situation. That the phenomenon can occur in much more complex (nonlinear cost structure, multiple origins and destinations, etc.) situations is shown by the small examples in [HA02] and, in this paper, by the Sioux Falls study network. Models of urban areas can easily involve thousands of links and origin-destination pairs. Because the method presented involves only the solution of linear programs, we expect that it can be directly applied to large urban networks. When improved flows are found, congestion and societal costs can be reduced, but individual travelers face negligible increases in costs. This is in contrast to system-optimal flows, where typically some travelers face significant increases in cost.

Acknowledgments. The authors thank Professor David Boyce of the University of Illinois at Chicago and Professor Hillel Bar-Gera of Ben-Gurion University for their interest and for providing the Sioux Falls data. We thank Professor Linus Schrage of the University of Chicago for providing the version of LINGO used for the Sioux Falls example and for helping with the implementation.

References

[AM81] Aashtiani, H.Z., Magnanti, T.L.: Equilibria on a Congested Transportation Network. SIAM Journal on Algebraic and Discrete Methods, **3**, 213–226 (1981)

[AK78] Abrams, R., Kerzner, L.: A Simplified Test for Optimality. Journal of Optimization Theory and Applications, **25**, 161–170 (1978)

[Bar02] Bar-Gera, H.: Transportation Network Test Problems. http://www.bgu.ac.il/~bargera/tntp/, (2002)

[BSS93] Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Second ed., Wiley, New York (1993)

[BMW55] Beckmann, M., McGuire, C.B., Winsten, C.B.: Studies in the Economics of Transportation. Yale University Press, New Haven, CT (1955)

[BBZ81] Ben-Israel,A., Ben-Tal, A., Zlobec, S.: Optimality in Nonlinear Programming: A Feasible Directions Approach. Wiley, New York (1981)

[Bra68] Braess, D.: Über ein Paradoxon aus der Verkehrsplanung. Unternehmenforschung, **12**, 258–268 (1968)

[DN84a] Dafermos, S., Nagurney, A.: On Some Traffic Equilibrium Theory Paradoxes. Transportation Research B, **18B**, 101–110 (1984)

[Hag04] Hagstrom, J.N.: Improving Traffic Flows at No Cost - Data, Models and Analysis. http://www.uic.edu/~hagstrom/Research/TrfImprv.htm, (2004)

[HP90] Harker, P.T., Pang, J.-S.: Finite-dimensional Variational Inequality and Nonlinear Complementarity Problems: A Survey of Theory, Algorithms and Applications. Mathematical Programming, **48**, 161–220 (1990)

[HA02] Hagstrom, J.N., Abrams, R.A.: Characterizing Braess's Paradox for Traffic Networks. Proceedings of IEEE 2001 Conference on Intelligent Transportation Systems, 837–842 (2002)

[HLN84] Hearn, D.W., Lawphongpanich, S., Nguyen, S.: Convex Programming Formulations of the Asymmetric Traffic Assignment Problem. Transportation Research B, **18B**, 357–365 (1984)

[LeB73] LeBlanc, L.J.: Mathematical programming algorithms for large scale network equilibrium and network design problems. Ph.D. Thesis, Northwestern University (1973)

[Pat94] Patriksson, M.: The Traffic Assignment Problem - Models and Methods. VSP, Utrecht (1994)

[Roc70] Rockafellar, R.T.: Some Convex Programs Whose Duals are Linearly Constrained. In: Rosen, J.B., Magasarian, O.L., Ritter, K. (eds) Nonlinear Programming. Academic Press, 293-322 (1970)

[Roc80] Rockafellar, R.T.: Lagrange Multipliers and Variational Inequalities. In: Cottle, R.W., Giannessi, F., Lions, J.L. (eds) Variational Inequalities and Complementarity Problems: Theory and Applications. Wiley, New York, 303–322 (1980)

[SZ83] Steinberg, R., Zangwill, W.I.: On the Prevalence of Braess's Paradox. Transportation Science, **17**, 301–318 (1983)

[War52] Wardrop, J.G.: Some Theoretical Aspects of Road Traffic Research. Proceedings of the Institute of Civil Engineers, Part II, 325–378 (1952)

Appendix

A Finding the Set of Always Binding Constraints

In this appendix we apply Kerzner's algorithm [AK78] to (R-EIP). Because (R-EIP) has many sets of constraints and variables, the details become complicated. We first describe a version of the algorithm for a general convex program (CP) with nonlinear convex inequality constraints and linear equality constraints. The treatment of linear inequalities is discussed at the end of the description of the algorithm.

We wish to determine whether a vector $\mathbf{x}^*$ is optimal for the following convex programming problem.

$$\text{(CP)} \quad \begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t. } g_i(\mathbf{x}) \leq 0, \ i = 1, 2, ...m \\ &\qquad \mathbf{Ax} = \mathbf{b} \end{aligned}$$

where f and the g_i are differentiable and faithfully convex. [A function f is faithfully convex (See [Roc70] and [BBZ81, page 18].) if it can be written as

$$f(\mathbf{x}) = h(\mathbf{Mx} + \mathbf{b}) + \mathbf{a} \cdot \mathbf{x} + c$$

where h is a strictly convex function, $\mathbf{M}$ is a not necessarily square matrix, $\mathbf{a}$ and $\mathbf{b}$ are vectors of appropriate dimensions, and c is a scalar constant.]

We assume, without loss of generality, that at $\mathbf{x}^*$ all of the nonlinear constraints, those involving g_i, are binding. Any constraints that are not binding at $\mathbf{x}^*$ may be ignored for our current purposes.

Let $P^=$ be the set of always binding nonlinear constraints, that is, $P^=$ is the set of nonlinear constraints which are binding for all feasible solutions of (CP). By definition all constraints not in the set $P^=$ can be satisfied strictly for some feasible solution, and by taking a convex combination of such feasible solutions, we may obtain a single feasible solution for which all constraints not in $P^=$ hold strictly.

As will be shown below, if the set $P^=$ is known, the nonlinear constraints in $P^=$ may be replaced by linear constraints, and the remaining nonlinear constraints (those not in $P^=$) will hold strictly for some feasible solution. Thus the Slater constraint qualification [BSS93, page 190] will hold, and it will be a simple matter to check the optimality of $\mathbf{x}^*$ by solving a single linear program.

Kerzner's algorithm for finding the set $P^=$ of (R-EIP) incrementally builds up the set of constraints known to be in $P^=$. It starts by assuming $P^=$ is empty, and at each iteration finds at least one more constraint (often many more) that is a member of $P^=$, or determines that $P^=$ is already completely specified, in which case the algorithm terminates.

To start the algorithm, we check for the existence of a feasible direction $\mathbf{d}$ so that all of the nonlinear constraints evaluated at $\mathbf{x}^* + t\mathbf{d}$ for sufficiently small $t > 0$ will hold strictly, that is, we look for a $\mathbf{d}$ that satisfies

$$\begin{aligned} \nabla g_i(\mathbf{x}^*)^T \mathbf{d} &\leq -1 \; i = 1, 2, \ldots m \\ \mathbf{Ad} &= \mathbf{0} \end{aligned}$$

If such a $\mathbf{d}$ exists, then $P^=$ is empty and all of the nonlinear constraints can be satisfied strictly and the algorithm terminates. If no such $\mathbf{d}$ exists, then the linear program

$$\text{(FD-P)} \quad \begin{aligned} &\max \quad \mathbf{0} \cdot \mathbf{d} \\ &\text{s.t.} \quad \nabla g_i(\mathbf{x}^*)^T \mathbf{d} \leq -1, \; i = 1, 2, \ldots m \\ &\mathbf{Ad} = \mathbf{0} \end{aligned}$$

has no feasible solution and by the duality theorem of linear programming its dual

$$\text{(FD-D)} \quad \begin{aligned} &\min -\sum \alpha_i \\ &\text{s.t.} \sum \nabla g_i(\mathbf{x}^*) \alpha_i + \mathbf{A}^T \beta = \mathbf{0} \\ &\alpha \geq \mathbf{0} \end{aligned} \tag{23}$$

must be either infeasible or unbounded. However, $\mathbf{0}$ is a feasible solution of (FD-D), and thus if (FD-P) has no feasible solution, (FD-D) must be unbounded. In that case some feasible solution to (FD-D) has at least one positive component in α.

Let α be a feasible solution of (FD-D) with at least one component, for example the first component, α_1, positive. We will show that the first nonlinear constraint must hold as an equality for all feasible solutions. Suppose, to the contrary, that for some feasible solution of (CP), the first nonlinear constraint holds strictly. Then there must exist a feasible direction $\mathbf{d}$ of (CP) satisfying

$$\begin{aligned} \mathbf{d}^T \nabla g_1 &< 0 \\ \mathbf{d}^T \nabla g_i &\leq 0 \; i \neq 1 \\ \mathbf{d}^T \mathbf{A}^T &= 0 \end{aligned}$$

Now "left" multiply the constraint (23) of (FD-D) by this feasible direction vector $\mathbf{d}$. The first summand of the left hand side consists of the positive α_1 times the (negative) inner product $\mathbf{d}^T \nabla g_1$. Thus the first summand is negative. Similarly all others are nonpositive. Noting that $\mathbf{d}^T \mathbf{A}^T \beta = 0$, we see that the left hand side is negative and the right hand side is zero – a contradiction. Therefore we conclude that whenever an α_i is positive the corresponding

$\mathbf{d}^T \nabla g_i$ cannot be negative. Hence the corresponding constraints of (CP) can never hold strictly and are members of $P^=$. Thus if we solve (FD-D), and find one or more positive α_i, we know that the corresponding constraints are members of $P^=$.

The next step is to replace nonlinear constraints known to be members of $P^=$ with linear constraints. The assumption of faithful convexity means that $g_i(\mathbf{x})$ in $P^=$ can be broken into linear and strictly convex parts. Since by definition a constraint belonging to $P^=$ is constant on the feasible set, the strictly convex part of the constraint and the linear part must each be constant on the feasible set. It then follows that the argument of the strictly convex part of the function must be constant on the feasible set. Therefore we can replace the nonlinear constraint by linear constraints which require i) that the argument of the strictly convex part of the constraint equal its unique value on the feasible set, and ii) that the linear part of the constraint equal its unique value on the feasible set. The result is that we can write (CP) as a convex program with at least one fewer nonlinear constraint.

We repeatedly apply the above method until all constraints are determined to be in $P^=$ and have been replaced by linear constraints, or the only solutions of (FD-D) have $\alpha = \mathbf{0}$. When the only solutions of (FD-D) have $\alpha = \mathbf{0}$, all of the remaining nonlinear constraints hold strictly for some feasible solution of (CP), which by Proposition 1 means that a Generalized Braess Paradox exists. If all nonlinear constraints are in $P^=$, we will have reduced the (CP) to a problem with a convex objective function and linear constraints. The optimality of $\mathbf{x}^*$ may then be determined by solving a single linear program, e.g., by seeking a feasible direction of descent of the objective function.

If the problem contains linear inequality constraints, these should be put in a separate grouping and handled in a manner similar to that of the linear equality constraints. The only differences would be that the feasible direction problem (FD-P) needs an additional set of constraints, the corresponding dual variables are nonnegative in (FD-D), and Equation (23) has one additional term. These modifications will be made in the example that follows.

As an illustration we apply the above method to Example 1 of Section 5. The network is shown in Figure 1, and the link cost functions are given in Table 2. The equilibrium solution for a total flow of 6 units is shown in Table 2, as are the system optimal flows and improved flows to be determined by the above method.

First formulate the convex program (R-EIP) using the data from Figure 2. Because the equilibrium solution has flow on all links, $\bar{\mathbf{z}}$ is equal to zero and may be omitted from the formulation. As there is only one destination, we omit the superscripts from the formulation. Also denote the top node in Figure 1 as node 2 and the bottom node as node 3.

(R-EIP)

$$\min \ 1.4x_1^2 + 5.4x_2 + x_2\sqrt{4x_2^2+9} + 2.4x_3^2 + 7.8x_4 + x_4\sqrt{4x_4^2+9} + 2x_5^2$$

$$\begin{aligned}
\text{s.t.} \qquad 1.4x_1 + u_2 - u_s &\le 0 \\
5.4 + \sqrt{4x_2^2 + 9} + u_3 - u_s &\le 0 \\
2.4x_3 + u_3 - u_2 &\le 0 \\
7.8 + \sqrt{4x_4^2 + 9} + u_t - u_2 &\le 0 \\
2x_5 + u_t - u_3 &\le 0 \\
x_1 + x_2 &= 6 \\
-x_1 + x_3 + x_4 &= 0 \\
-x_2 - x_3 + x_5 &= 0 \\
-x_4 - x_5 &= -6 \\
x_j &\ge 0 \; j = 1, 2, ..5 \\
u_t &= 0 \\
u_s &\le 18.4
\end{aligned}$$

Next we formulate the primal feasible direction problem (FD-P) which searches for a direction, **d**, that will make the nonlinear constraints hold strictly.

$$\begin{aligned}
\text{(FD-P)} \qquad 1.4d_{x_1} + d_{u_2} - d_{u_s} &\le 0 \\
1.6d_{x_2} + d_{u_3} - d_{u_s} &\le -1 \\
2.4d_{x_3} + d_{u_3} - d_{u_2} &\le 0 \\
1.6d_{x_4} + d_{u_t} - d_{u_s} &\le -1 \\
2d_{x_5} + d_{u_t} - d_{u_3} &\le 0 \\
d_{x_1} + d_{x_2} &= 0 \\
-d_{x_1} + d_{x_3} + d_{x_4} &= 0 \\
-d_{x_2} - d_{x_3} + d_{x_5} &= 0 \\
-d_{x_4} - d_{x_5} &= 0 \\
d_{u_t} &= 0 \\
d_{u_s} &\le 0
\end{aligned}$$

The existence of a solution to this set of inequalities is equivalent to the Slater Condition, which requires that the nonlinear constraints be satisfied strictly at some feasible point. Although the Slater Condition will turn out to hold for this problem, in general it cannot be expected to hold, and, in fact, it does not hold in Examples 2 and 3. We first convert the above set of inequalities into a linear program by forming an objective function with cost coefficients all equal to zero. Then form the dual of the linear program which we denote (FD-D). We denote the dual variables corresponding to the nonlinear constraints of (R-EIP) (second and fourth constraints) by α_j for $j = 2, 4$ and those corresponding to the linear constraints by β_j for all other values of j.

$$
\begin{aligned}
\text{(FD-D)} \qquad & \min -\alpha_2 - \alpha_4 \\
& \text{s.t. } 1.4\beta_1 + \beta_6 - \beta_7 = 0 \\
& 1.6\alpha_2 + \beta_6 - \beta_8 = 0 \\
& 2.4\beta_3 + \beta_7 - \beta_8 = 0 \\
& 1.6\alpha_4 + \beta_7 - \beta_9 = 0 \\
& 2\beta_5 + \beta_8 - \beta_9 = 0 \\
& \beta_1 + \alpha_2 + \beta_{11} = 0 \\
& -\beta_1 + \beta_3 + \alpha_4 = 0 \\
& -\alpha_2 - \beta_3 + \beta_5 = 0 \\
& -\alpha_4 - \beta_5 + \beta_{10} = 0 \\
& \alpha_2, \alpha_4, \beta_1, \beta_3, \beta_5, \beta_{11} \geq 0
\end{aligned}
$$

As pointed out above, (FD-D) is either unbounded or it has an optimal solution with optimal objective function value equal to zero. We find that the optimal solution of (FD-D) has $\alpha_2 = 0$ and $\alpha_4 = 0$. Therefore, we know that (FD-P) has an optimal solution which is a feasible direction leading to a point that satisfies both nonlinear constraints of (R-EIP) strictly. From Proposition 1 it follows that a Generalized Braess Paradox exists.

The optimal solution to (FD-P), that is, the direction vector leading to an interior point, may be obtained from the dual prices of (FD-D). An improved flow for the original network, demonstrating the Generalized Braess Paradox, may then be obtained by moving in this direction. The optimal solution to (R-EIP) is shown in Table 2. It is very close to the solution obtained by moving in the direction given by the solution to (FD-P). The optimal solution to (R-EIP) reduces travel cost by 22% for travelers moving along the path defined by Links 1, 3, and 5, and does not increase travel cost for any travelers when compared with the equilibrium solution.

When the algorithm is applied to a network with multiple origin-destination pairs, a significant benefit becomes apparent. For each link k, the constraints corresponding to the various destinations d all have the same cost function F_k. Suppose that for a link k with strictly convex cost function it is known that for one destination d, the constraint indexed by (k, d) belongs to $P^=$. Then the unique feasible value of F_k will be known, not only for the constraint indexed by (k, d), but for every constraint involving that link k and other destinations. Thus the nonlinear constraints of all destinations involving that link may be replaced by linear constraints. As a result of this special structure of (R-EIP), at each iteration of Kerzner's algorithm we will eliminate all of the nonlinear constraints corresponding to at least one link. If α_k^d, corresponding to a cost constraint indexed by (k, d), is positive for more than one value of k, we will eliminate all of the nonlinear constraints corresponding to those values of k also. It follows that the number of linear programs that need to be solved is bounded above by the number of links in the network. General formulations of (FD-P) and (FD-D) for Kerzner's algorithm as applied to (R-EIP) may be found in an earlier version of this paper at [Hag04] .

B Convexity of the Objective Function of (EIP)

In this appendix, we provide an elementary proof of the convexity of the system cost when the link cost function is monotone and component-wise convex. We thank the editors for pointing out that this result can be viewed as a special case of Equation (9) in [HLN84]. The approach in [HLN84] uses derivatives of $\mathbf{F}$, which are not required in the proof below.

Proposition 2. *If for* $\mathbf{x} \geq 0$*, the components of* $\mathbf{F}(\mathbf{x})$ *are convex, and* $\mathbf{F}(\mathbf{x})$ *is monotone, then the system cost function* $\mathbf{x} \cdot \mathbf{F}(\mathbf{x})$ *is convex.*

Proof. By definition $g(\mathbf{x}) = \mathbf{x} \cdot \mathbf{F}(\mathbf{x})$ is convex if

$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y}), \quad \lambda \in [0,1],$$

or in terms of the vectors $\mathbf{x}$ and $\mathbf{F}(\mathbf{x})$,

$$[\lambda\mathbf{x} + (1-\lambda)\mathbf{y}] \cdot \mathbf{F}(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda\mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + (1-\lambda)\mathbf{y} \cdot \mathbf{F}(\mathbf{y}) \tag{24}$$

Thus we must show that (24) holds. The convexity of the components of $\mathbf{F}$ and the nonnegativity of the components of $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ give

$$\begin{aligned}
&[\lambda\mathbf{x} + (1-\lambda)\mathbf{y}] \cdot \mathbf{F}(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \\
&\leq [\lambda\mathbf{x} + (1-\lambda)\mathbf{y}] \cdot [\lambda\mathbf{F}(\mathbf{x}) + (1-\lambda)\mathbf{F}(\mathbf{y})] \\
&= \lambda^2\mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + (1-\lambda)^2\mathbf{y} \cdot \mathbf{F}(\mathbf{y}) + \lambda(1-\lambda)[\mathbf{x} \cdot \mathbf{F}(\mathbf{y}) + \mathbf{y} \cdot \mathbf{F}(\mathbf{x})]
\end{aligned} \tag{25}$$

The vector function $\mathbf{F}(\mathbf{x})$ is monotone for $\mathbf{x} \geq \mathbf{0}$ if

$$[\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})] \cdot [\mathbf{x} - \mathbf{y}] \geq 0 \ \forall\ \mathbf{x}, \mathbf{y} \geq 0$$

Expanding this definition of the monotonicity of $\mathbf{F}$ gives:

$$\mathbf{x} \cdot \mathbf{F}(\mathbf{y}) + \mathbf{y} \cdot \mathbf{F}(\mathbf{x}) \leq \mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + \mathbf{y} \cdot \mathbf{F}(\mathbf{y}) \tag{26}$$

Using (26) to substitute into the last term of the right-hand side of (25) yields

$$\begin{aligned}
&[\lambda\mathbf{x} + (1-\lambda)\mathbf{y}] \cdot \mathbf{F}(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \\
&\leq \lambda^2\mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + (1-\lambda)^2\mathbf{y} \cdot \mathbf{F}(\mathbf{y}) + \lambda(1-\lambda)[\mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + \mathbf{y} \cdot \mathbf{F}(\mathbf{y})] \\
&= \left[\lambda^2 + \lambda(1-\lambda)\right] \mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + \left[(1-\lambda)^2 + \lambda(1-\lambda)]\right] \mathbf{y} \cdot \mathbf{F}(\mathbf{y}) \\
&= \lambda\mathbf{x} \cdot \mathbf{F}(\mathbf{x}) + (1-\lambda)\mathbf{y} \cdot \mathbf{F}(\mathbf{y})
\end{aligned}$$

which establishes (24), as was to be shown. ∎

Relaxed Toll Sets for Congestion Pricing Problems

Lihui Bai[1], Donald W. Hearn[2], and Siriphong Lawphongpanich[3]

[1] College of Business Administration, Valparaiso University, Valparaiso, IN 46383, U.S.A., Lihui.Bai@valpo.edu
[2] Industrial and Systems Engineering Department, University of Florida, Gainesville, FL 32611, U.S.A., hearn@ise.ufl.edu
[3] Industrial and Systems Engineering Department, University of Florida, Gainesville, FL 32611, U.S.A., lawphong@ise.ufl.edu

Summary. Congestion or toll pricing problems in [HeR98] require a solution to the system problem (the traffic assignment problem that minimizes the total travel delay) to define the set of all valid tolls or the toll set. For practical problems, it may not be possible to obtain an exact solution to the system problem and the inaccuracy in an approximate system solution may render the toll set empty. When this occurs, this paper offers alternative toll sets based on relaxed optimality conditions. With carefully chosen parameters, tolls from the relaxed toll sets are shown theoretically and empirically (using four transportation networks in the literature) to induce route choices that are nearly system optimal.

Key words: Congestion Pricing, Traffic Equilibrium, Perturbation Analysis

1 Introduction

To encourage each traveller to choose a route in a transportation network that would collectively benefit all travellers, Hearn and Ramana [HeR98] proposed in 1998 a framework for determining the prices and locations at which to toll the network. This framework requires solving a congestion or toll pricing problem, an optimization problem with linear constraints that describe the set of all valid tolls or the toll set. Coefficients for the constraints depend on an optimal solution to the system problem, i.e., the traffic assignment problem (see, e.g., Florian and Hearn, [FlH95]) that minimizes the total travel delay among all travellers.

For small transportation networks, it is possible to compute an exact optimal solution to the system problem. However, obtaining such a solution for larger networks may be either impossible or impractical. When implemented on computers, algorithms for the system problem must perform all numerical computations using finite precision. This naturally induces small numerical

Mathematical and Computational Models for Congestion Charging, pp. 23-44

inaccuracies because to perform some mathematical operations precisely requires infinite precision. Furthermore, the system problem is generally a nonlinear program for which most algorithms require in theory an infinite number of iterations to reach an exact optimal solution. In practice, it is common to terminate these algorithms when they find a solution with a small optimality gap, e.g., 10E-4.

On the other hand, using an approximate solution for the system problem (or an approximate system solution) to determine the coefficients for the constraints defining the toll set may cause the toll pricing problem to become infeasible, numerically (e.g., because of finite precision) or otherwise. To overcome this infeasibility, Hearn and Ramana [HeR98] employ a penalty function approach and Hearn et al. [HYR01] relax one of the constraints defining the toll set. For the latter, the relaxation is based on a parameter defined by an optimal solution to the penalty problem in [HeR98].

This paper studies the viability of using an approximate system solution in defining the toll set. Specifically, when an approximate system solution makes the toll set empty, this paper alleviates this inconsistency by relaxing one or more constraints, some of which are similar to those used in [HYR01]. However, our approach to relaxation does not require solving a penalty problem. Moreover, this paper also addresses three issues relating to the use of an approximate system solution. The first issue is whether an approximate system solution yields a consistent set of constraints defining the toll set. When it does not, the second issue is to find practical methods for relaxing the constraints in order to generate tolls that causes travellers to use the transportation network in nearly the most efficient manner. Finally, the last issue is to ascertain how well these methods work theoretically and empirically.

The remainder of the paper assumes that the travel demands are fixed. Results for the elastic demand case are similar and given in the Appendix. Section 2 defines two types of toll sets, system and non-system, and discusses their properties. Section 3 derives a relaxed toll set using an approximate system solution and shows that the tolls from this set have the desirable property. Section 4 gives an alternate representation of the relaxed toll set. Section 5 reports encouraging results for four transportation networks from the literature and Section 6 concludes the paper.

2 System and Non-System Toll Sets

To define toll sets, consider the two traffic assignment problems in transportation science literature, the system and user problems. (See, e.g., [FlH95].) Let $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ be a network with $\mathcal{N}$ and $\mathcal{A}$ denoting the set of nodes and arcs, respectively. Associated with $\mathcal{G}$, there are also a node-arc incidence matrix, A, and a set of commodities or origin-destination pairs, $\mathcal{K}$. For each commodity $k \in \mathcal{K}$, $b_k \in R^{|\mathcal{N}|}$ and $x^k \in R_+^{|\mathcal{A}|}$ denote the corresponding (fixed) demand and arc flow vectors, respectively. Hence, $v = \sum_k x^k$ is the vector of the total (or aggregate) flow on every arc and the set of feasible aggregate flow vectors can be described as follows:

$$V = \{v | v = \sum_k x^k, Ax^k = b_k, x^k \geq 0\},$$

where $x^k \geq 0$ means $x_a^k \geq 0, \forall a$. (More generally, $x \geq y$ means $x_a \geq y_a, \forall a$.) Without loss of generality, we can assume throughout the paper that V is bounded and, therefore, compact. (See, e.g., [FlH95].)

Let $s(v)$ be a travel cost vector in which each element, $s_a(v)$, is the cost to traverse arc a. This cost does not include any toll and can be measured in monetary or time units. For simplicity, we assume that $s_a(v)$ is differentiable for all a, i.e., $\bigtriangledown s(\overline{v})$, the Jacobian of $s(v)$, exists for all v. Then, the system optimal (SOPT) problem (or, more simply, system problem) is to find a feasible aggregate flow vector that minimizes the total travel cost or delay among all travellers. Mathematically, the system problem can be stated as follows:

$$\overline{v} = \text{argmin}\{s(v)^T v : v \in V\}.$$

Instead of minimizing the total travel delay, an alternate traffic assignment problem, i.e., the user equilibrium problem (or, more simply, the user problem), assumes that each traveller tries to minimize his or her own travel time. The objective of the user problem is to find a solution for which no traveller can improve his or her travel time by unilaterally changing routes. In particular, v^* solves the user problem (or UOPT) if it satisfies the following variational inequality:

$$s(v^*)^T(v - v^*) \geq 0, \quad \forall v \in V.$$

Alternately, we say that v^* solves VI$[s(v), V]$.

The travel delay at the user solution, $s(v^*)^T v^*$, is generally larger than the one at the system solution, $s(\overline{v})^T\overline{v}$. In this sense, the user solution does not utilize the network in the most efficient manner. Mathematically, we can impose tolls on arcs in order to make travellers use the network more efficiently. For a given toll vector, β, let $v^*(\beta)$ solve VI$[s(v) + \beta, V]$, i.e., $v^*(\beta)$ satisfies the following tolled user equilibrium condition:

$$(s(v^*(\beta)) + \beta)^T(v - v^*(\beta)) \geq 0, \quad \forall v \in V.$$

We refer to $v^*(\beta)$ as the solution to the tolled user equilibrium problem and it is the user equilibrium flow resulting from imposing the toll β on the network.

Similar to [HeR98], we assume herein that $\overline{v}$ is a unique solution to SOPT and $v^*(\beta)$ is a unique solution to VI$[s(v) + \beta, V]$ for all $\beta \in R^{|\mathcal{A}|}$. Below, we refer to these two assumptions as [A] and [B], respectively. For example, both [A] and [B] hold when we use the Bureau of Public Road (BPR) function for travel costs, i.e., $s_a(v) = \tau_a(1.0 + \theta_a(v_a/\gamma_a)^4)$ and τ_a, θ_a, and γ_a are positive. More generally, both assumptions hold when $s_a(v)$ is a continuous convex function for each a and the cost vector $s(v)$ is strictly monotone on $\{v : v \geq 0\}$. These two assumptions allow us to define β as a valid or feasible toll vector if $v^*(\beta) = \overline{v}$, i.e., if the tolled user equilibrium solution associated with β equals the system solution. (See [HeR98] for a more general definition of a valid toll.) Then, the toll set is the set of all valid toll vectors, i.e., $\mathcal{T}(\overline{v}) = \{\beta | v^*(\beta) = \overline{v}\}$. The following result from [HeR98] describes $\mathcal{T}(\overline{v})$ algebraically.

Theorem 1. *The toll set, $\mathcal{T}(\overline{v})$, is the set consisting of the β component of every pair (β, ρ) that satisfies the following linear system*

$$s(\overline{v}) + \beta \geq A^T \rho^k, \quad \forall\, k \in \mathcal{K}, \tag{1}$$

$$(s(\overline{v}) + \beta)^T \overline{v} = \sum_k {b_k}^T \rho^k. \tag{2}$$

Observe that the above toll set is based on $\overline{v}$, the optimal solution to SOPT. To distinguish this toll set from others (to be defined later), we refer to $\mathcal{T}(\overline{v})$ as the "(unrestricted) system toll set." As defined above, β in the system toll set is unrestricted. (In practice, positive tolls represent payment for road usage and negative tolls represent subsidies for the same purpose.) Moreover, the system toll set is nonempty. In fact, $\beta = -s(\overline{v})$ belongs to the system toll set because $\beta = -s(\overline{v})$ and $\rho^k = 0$ for all k trivially satisfy (1) and (2). In addition, the optimality condition for the system problem also implies that $\beta_{\text{mscp}} = \bigtriangledown s(\overline{v})^T \overline{v} \in \mathcal{T}(\overline{v})$. (See, e.g., [HeR98].) Transportation economists (see, e.g., Arnott and Small [ArS94]), generally refer to β_{mscp} as the *marginal social cost vector.* Using $-s(\overline{v})$ and $\bigtriangledown s(\overline{v})^T \overline{v}$, Hearn and Ramana show in [HeR98] that $\mathcal{T}(\overline{v})$ is unbounded. Because an arbitrarily large toll is impractical, we assume that all toll vectors in $\mathcal{T}(\overline{v})$ are bounded and, when not explicitly stated, the constraint $||\beta|| \leq B$, where B is a sufficiently large number, is included in all toll sets described herein.

When β is required to be nonnegative, we refer to the set $\mathcal{T}^+(\overline{v}) = \{\beta \geq 0 | v^*(\beta) = \overline{v}\}$ as the "nonnegative system toll set." Algebraically, $\mathcal{T}^+(\overline{v})$ is the toll set described in Theorem 1 with an additional nonnegativity constraint on β. In practice, $\mathcal{T}^+(\overline{v})$ is nonempty. Practical traffic assignment models (see, e.g., [FGS87], [FlH95], [HLV87], and [LMP75]) typically use travel cost functions whose Jacobians, $\bigtriangledown s(\overline{v})$, are both nonnegative and diagonal. This makes β_{mscp} nonnegative and $\mathcal{T}^+(\overline{v})$ nonempty. Later, we provide a condition under which the latter holds without requiring the Jacobian to be nonnegative.

Consider the case when it is not practical to compute $\overline{v}$ exactly. Let $\widetilde{v}$ denote an approximate solution to SOPT. Without specifying the quality of the approximation, all that can be claimed is that $\widetilde{v}$ is a feasible aggregate flow vector and the toll set based on $\widetilde{v}$, or the non-system toll set, is $\mathcal{T}(\widetilde{v}) = \{\beta | v^*(\beta) = \widetilde{v}\}$. In words, this is the set of toll vectors whose tolled user equilibrium solution equals the aggregate flow vector $\widetilde{v}$. As shown below, the algebraic characterization of $\mathcal{T}(\widetilde{v})$ is essentially the same as that of $\mathcal{T}(\overline{v})$.

Theorem 2. *The nonsystem toll set, $\mathcal{T}(\widetilde{v})$, is the set consisting of the β component of every pair (β, ρ) that satisfies the following linear system:*

$$s(\widetilde{v}) + \beta \geq A^T\rho^k, \quad \forall\ k \in \mathcal{K}, \tag{3}$$

$$(s(\widetilde{v}) + \beta)^T\widetilde{v} = \sum_k {b_k}^T\rho^k. \tag{4}$$

Proof. Because of assumption [B], $\widetilde{v}$ must solve VI$[s(v) + \beta, V]$ uniquely for every $\beta \in \mathcal{T}(\widetilde{v})$. From Proposition 1.2.1 in Facchinei and Pang [FaP03], $\widetilde{v}$ solves VI$[s(v) + \beta, V]$ if and only if there exist ρ^k and σ^k that satisfy the following KKT conditions:

$$\begin{aligned} s(\widetilde{v}) + \beta - A^T\rho^k - \sigma^k &= 0, \quad \forall\ k \in \mathcal{K}, \\ (\widetilde{x}^k)^T\sigma^k &= 0, \quad \forall\ k \in \mathcal{K}, \\ \sigma^k &\geq 0, \quad \forall\ k \in \mathcal{K}. \end{aligned}$$

The pair (β, ρ), where ρ is determined by the above KKT conditions, satisfies (3) and (4). The first and third KKT conditions imply that (3) holds. Multiplying the first KKT conditions by $\widetilde{x}^k$ and summing the resulting equations together yield

$$(s(\widetilde{v}) + \beta)^T \sum_k \widetilde{x}^k = \sum_k (A\widetilde{x}^k)^T\rho^k + \sum_k (\widetilde{x}^k)^T\sigma^k.$$

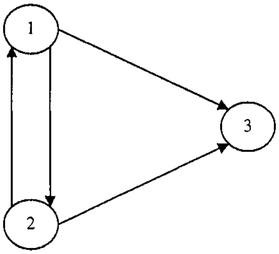
Because $\sum_k \widetilde{x}^k = \widetilde{v}$, $A\widetilde{x}^k = b_k$ and $(\widetilde{x}^k)^T\sigma^k = 0$, the above equation reduces to (4).

Conversely, let (β, ρ) satisfy (3) and (4) and set σ^k equal to $s(\widetilde{v})+\beta-A^T\rho^k$ for all k. Then, it follows from (3) and (4) that $\sigma^k \geq 0$ and $\sum_k(\widetilde{x}^k)^T\sigma^k = 0$, respectively. The latter also implies that $(\widetilde{x}^k)^T\sigma^k$ must individually equal to zero because $\widetilde{x}^k \geq 0$ and $\sigma^k \geq 0$. Thus, the above KKT conditions are satisfied and, using the above result from [FaP03], $\widetilde{v}$ solves VI$[s(v) + \beta, V]$, i.e., $\beta \in \mathcal{T}(\widetilde{v})$. ■

In the above proof, assumption [B] is essential. Without uniqueness, there may be alternate tolled user equilibrium solutions not equal to $\widetilde{v}$. In addition, the non-system toll set described above is always nonempty because, as in the system toll set, $-s(\widetilde{v}) \in \mathcal{T}(\widetilde{v})$.

Consider the nonnegative and non-system toll set, i.e., $\mathcal{T}^+(\widetilde{v}) = \{\beta \geq 0 | v^*(\beta) = \widetilde{v}\}$. The algebraic representation of $\mathcal{T}^+(\widetilde{v})$ is the same as described in the above theorem with the addition of the nonnegativity constraint on β. However, the example below illustrates that $\mathcal{T}^+(\widetilde{v})$ can be empty.

The network in Figure 1 represents a transportation system with three nodes and four arcs where the travel cost function for every arc is constant and equals 1. There are two OD pairs, (1,3) and (2,3), each with a travel demand of 2 units. Table 1 displays a set of feasible flow vectors for the transportation system. For OD pair (1,3), the flow vector $\widetilde{x}^{(1,3)}$ corresponds to sending one unit of flow along each of the two possible paths, 1 → 2 → 3



OD pairs	Demands
(1, 3)	2
(2, 3)	2

Travel Demands

Travel Cost function: $s_{ij}(v_{ij}) = 1$, $\forall$ arc (i, j)

Fig. 1. A Counterexample

and 1 → 3. Similarly, $\tilde{x}^{(2,3)}$ corresponds to send one unit of flow along paths 2 → 1 → 3 and 2 → 3. Clearly, the aggregate flow vector $\tilde{v} = \tilde{x}^{(1,3)} + \tilde{x}^{(2,3)}$ is not system optimal because sending two units of flow along arcs (1,3) and (2,3) satisfies both travel demands and is less costly.

Table 1. Feasible Flow Vectors for the Network in Figure 1

Arc	$\tilde{x}^{(1,3)}$	$\tilde{x}^{(2,3)}$	$\tilde{v}$
(1,2)	1	0	1
(1,3)	1	1	2
(2,1)	0	1	1
(2,3)	1	1	2

Because the two OD pairs can be treated as one commodity, the nonnegative and nonsystem toll set, $\mathcal{T}^{+}(\tilde{v})$, reduces to the following linear system:

$$\begin{aligned} 1+\beta_{12} &\geq \rho_1 - \rho_2 \\ 1+\beta_{13} &\geq \rho_1 - \rho_3 \\ 1+\beta_{21} &\geq \rho_2 - \rho_1 \\ 1+\beta_{23} &\geq \rho_2 - \rho_3 \\ (1+\beta_{12}) + 2(1+\beta_{13}) + (1+\beta_{21}) + 2(1+\beta_{23}) &= 2\rho_1 + 2\rho_2 - 4\rho_3 \\ \beta_{ij} &\geq 0, \qquad\qquad \forall(i,j) \end{aligned}$$

The equality constraint in the above system can be equivalently written as

$$\begin{aligned} &(1+\beta_{12} - [\rho_1 - \rho_2]) + 2(1+\beta_{13} - [\rho_1 - \rho_3]) \\ &\qquad + (1+\beta_{21} - [\rho_2 - \rho_1]) + 2(1+\beta_{23} - [\rho_2 - \rho_3]) = 0. \end{aligned}$$

This equation implies that the four inequalities in the system must hold at equality, i.e.,

$$\begin{aligned}1+\beta_{12} &= \rho_1 - \rho_2\\ 1+\beta_{13} &= \rho_1 - \rho_3\\ 1+\beta_{21} &= \rho_2 - \rho_1\\ 1+\beta_{23} &= \rho_2 - \rho_3\end{aligned}$$

Adding the first and third equations together yields

$$2+\beta_{12}+\beta_{21}=0.$$

However, this is impossible because $\beta_{ij} \geq 0, \forall(i,j)$. Thus, $\mathcal{T}^+(\widetilde{v}) = \emptyset$.

The following theorem provides a necessary and sufficient condition under which $\mathcal{T}^+(\widetilde{v})$ is nonempty. Independently, Fleischer et al. [FJM04] provide a different, but equivalent, condition for the nonemptiness of the nonnegative and non-system toll set. The condition in the theorem below is related to an earlier work on bounded traffic assignment problem in [Hea80] that was later continued in [Ber95] and [BHR97] under the setting of congestion pricing.

Theorem 3. *For any $\widetilde{v} \in V$, the set $\mathcal{T}^+(\widetilde{v})$ is nonempty if and only if $\widetilde{v}$ solves VI[$s(v)$,$\mathcal{V}$], where $\mathcal{V} = \{v | v = \sum_k x^k, Ax^k = b_k, x^k \geq 0, v \leq \widetilde{v}\}$.*

Proof. Using Proposition 1.2.1 in [FaP03], $\widetilde{v}$ solve VI[$s(v)$, $\mathcal{V}$] if and only if there exist ρ^k, σ^k, and β that satisfies the following KKT conditions:

$$\begin{aligned}s(\widetilde{v}) - A^T\rho^k - \sigma^k + \beta &= 0, \forall k,\\ (\widetilde{x}^k)^T\sigma^k &= 0, \forall k,\\ \sigma^k &\geq 0, \forall k,\\ \beta &\geq 0.\end{aligned}$$

In the above system, β is the multiplier vector corresponding to the upper bounds $v \leq \widetilde{v}$ in $\mathcal{V}$ and the complementarity condition $\beta^T(v - \widetilde{v}) = 0$ is not required because $v = \widetilde{v}$ satisfies every upper bound in $\mathcal{V}$ exactly. By an argument similar to the one in Theorem 2, the above conditions are equivalent to those that describe $\mathcal{T}^+(\widetilde{v})$. Thus, the theorem holds. ■

The corollary below provides a similar condition for the nonnegative system toll set and follows immediately from the above theorem.

Corollary 1. *The nonnegative system toll set, $\mathcal{T}^+(\overline{v})$, is nonempty if and only if $\overline{v}$ solves VI[$s(v)$,$\overline{\mathcal{V}}$], where $\overline{\mathcal{V}} = \{v | v = \sum_k x^k, Ax^k = b_k, x^k \geq 0, v \leq \overline{v}\}$.*

3 Relaxed Toll Set

Consider the situation in which an algorithm terminates and produces $\widetilde{v}$ as an approximate solution to SOPT with some desired optimality gap. Using Theorem 3 from the previous section, it is possible to determine whether $\mathcal{T}^+(\widetilde{v})$ is nonempty. However, $\mathcal{T}^+(\widetilde{v})$ is often empty in practice. This section resolves this difficulty by finding nonnegative tolls that satisfy the conditions in Theorem 2 approximately. Moreover, the focus is on defining a nonnegative relaxed toll set based on $\widetilde{v}$ when $\bigtriangledown s(\widetilde{v})$ is nonnegative. (When β is unrestricted, the system and non-system toll sets are nonempty. As such, they require no relaxation. When $\bigtriangledown s(\widetilde{v})$ is nonnegative, the same holds for the nonnegative system toll set.)

The first condition in Theorem 2 is

$$s(\widetilde{v}) + \beta \geq A^T \rho^k, \quad \forall k \in \mathcal{K}.$$

When multiplied by $\widetilde{x}^k$ and summed together, the above implies that $(s(\widetilde{v}) + \beta)^T \widetilde{v} \geq \sum_k b_k^T \rho^k$ because $A\widetilde{x}^k = b_k$ and $\sum_k \widetilde{x}^k = \widetilde{v}$. Therefore, the equality in (4) can be replaced by an inequality '$\leq$.' This replacement motivates the definition of a relaxed toll set $\mathcal{T}^+(\widetilde{v}, \epsilon)$, for some $\epsilon > 0$, as the set of all β for which there exists a corresponding ρ satisfying the following conditions:

$$\begin{aligned} s(\widetilde{v}) + \beta &\geq A^T \rho^k, & \forall k \in \mathcal{K}, \\ (s(\widetilde{v}) + \beta)^T \widetilde{v} &\leq \sum_{k \in \mathcal{K}} b_k^T \rho^k + \epsilon, & \\ \beta &\geq 0. & \end{aligned}$$

Let $-\epsilon_{\text{mscp}} = \min\{(s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v})^T (u - \widetilde{v}) : u \in V\}$. In Hearn [Hea82], ϵ_{mscp} is the optimality gap for SOPT at $\widetilde{v}$ and the following theorem shows that $\mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$ is nonempty.

Theorem 4. *If $\bigtriangledown s(\widetilde{v})$ is nonnegative, then $\mathcal{T}^+(\widetilde{v}, \epsilon_{mscp}) \neq \emptyset$, where $\epsilon_{mscp} > 0$ is as defined above.*

Proof. Note that

$$\begin{aligned} \epsilon_{\text{mscp}} &= (s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v})^T \widetilde{v} - \min\{(s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v})^T u : u \in V\} \\ &= (s(\widetilde{v}) + \beta_{\text{mscp}})^T \widetilde{v} - \min\{(s(\widetilde{v}) + \beta_{\text{mscp}})^T u : u \in V\}. \end{aligned}$$

From linear programming duality, the following holds

$$\begin{aligned} \epsilon_{\text{mscp}} &= (s(\widetilde{v}) + \beta_{\text{mscp}})^T \widetilde{v} - \max_{\rho}\{\textstyle\sum_k b_k^T \rho^k : s(\widetilde{v}) + \beta_{\text{mscp}} \geq A^T \rho^k, \forall k\} \\ &= \min_{\rho}\{(s(\widetilde{v}) + \beta_{\text{mscp}})^T \widetilde{v} - \textstyle\sum_k b_k^T \rho^k : s(\widetilde{v}) + \beta_{\text{mscp}} \geq A^T \rho^k, \forall k\} \end{aligned}$$

Let $\widetilde{\rho}$ denote an optimal solution to the linear program in the last equation. Then, the pair $(\beta_{\text{mscp}}, \widetilde{\rho})$ satisfies the relaxed toll condition with $\epsilon = \epsilon_{\text{mscp}}$, i.e.,

$$\begin{aligned} s(\widetilde{v}) + \beta_{\text{mscp}} &\geq A^T \widetilde{\rho}^k, & \forall k \in \mathcal{K}, \\ (s(\widetilde{v}) + \beta_{\text{mscp}})^T \widetilde{v} &= \sum_{k \in \mathcal{K}} b_k^T \widetilde{\rho}^k + \epsilon_{\text{mscp}}. & \end{aligned}$$

Because $\bigtriangledown s(\widetilde{v})$ is nonnegative, $\beta_{\text{mscp}} \geq 0$. So, $\beta_{\text{mscp}} \in \mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$ and $\mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}}) \neq \emptyset$. ■

As shown above, ϵ_{mscp} can be computed with little effort because many algorithms (see, e.g., [FGS87], [LMP75], and [HLV87]) for SOPT compute ϵ_{mscp} and terminate when they find a $\widetilde{v} \in V$ such that the corresponding

$\epsilon_{\text{mscp}} \leq \epsilon$, for some small $\epsilon > 0$. Instead of ϵ_{mscp}, it is also possible to choose ϵ by solving the following linear program:

$$\begin{aligned} \epsilon^* = \min_{(\beta,\rho)} \; & (s(\widetilde{v}) + \beta)^T \widetilde{v} - \sum_k b_k^T \rho^k \\ \text{s.t.} \; & s(\widetilde{v}) + \beta \geq A^T \rho^k, \qquad \forall k, \\ & \beta \geq 0. \end{aligned}$$

Because $\mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}}) \neq \emptyset$, the above optimization is feasible. In addition, $\epsilon^* \leq \epsilon_{\text{mscp}}$.

One important property of the system toll sets (unrestricted or otherwise) is that, for any β in $\mathcal{T}(\overline{v})$ (or $\mathcal{T}^+(\overline{v})$), $\overline{v}$ solves VI$[s(v) + \beta, V]$, i.e., the system solution also solves the user equilibrium problem with the toll vector β. However, this property only holds approximately for the relaxed toll set. Assume that $\widetilde{v}$ solves SOPT approximately, i.e.,

$$\min\{(s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v})^T (u - \widetilde{v}) : u \in V\} = -\epsilon_{\text{mscp}} \geq -\epsilon.$$

Then, for any $\beta \in \mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$, the following must hold

$$\begin{aligned} s(\widetilde{v}) + \beta &\geq A^T \rho^k, \qquad \forall k \in \mathcal{K}, \\ (s(\widetilde{v}) + \beta)^T \widetilde{v} &\leq \sum_{k \in \mathcal{K}} b_k^T \rho^k + \epsilon_{\text{mscp}}. \end{aligned}$$

It follows from the above that

$$\begin{aligned} -\epsilon \leq -\epsilon_{\text{mscp}} &\leq \sum_{k \in \mathcal{K}} b_k^T \rho^k - (s(\widetilde{v}) + \beta)^T \widetilde{v}, \forall \rho \in \{\rho : A^T \rho^k \leq s(\widetilde{v}) + \beta, \forall k\} \\ &\leq \max_{\rho}\{\sum_{k \in \mathcal{K}} b_k^T \rho^k : A^T \rho^k \leq s(\widetilde{v}) + \beta, \forall k\} - (s(\widetilde{v}) + \beta)^T \widetilde{v}, \\ &= \min_u\{(s(\widetilde{v}) + \beta)^T u : u \in V\} - (s(\widetilde{v}) + \beta)^T \widetilde{v}, \\ &= \min_u\{(s(\widetilde{v}) + \beta)^T (u - \widetilde{v}) : u \in V\} \\ &\leq (s(\widetilde{v}) + \beta)^T (u - \widetilde{v}), \forall u \in V, \end{aligned}$$

where the first equality holds because of the strong duality theorem in linear programming. Observe that the last inequality implies that $\widetilde{v}$ solves VI$[s(v) + \beta, V]$ approximately. Thus, $\widetilde{v}$ approximately solves both SOPT and the tolled user equilibrium problem.

For a slightly stronger statement, Theorem 5 below demonstrates that, for any $\eta > 0$, there exists a $\delta > 0$ such that $||v^*(\beta') - \overline{v}|| \leq \eta$ when $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$ and $||\widetilde{v} - \overline{v}|| \leq \delta$. (Here, $||\cdot||$ represents the Euclidean norm.) In words, the theorem states that a toll vector from the relaxed toll set yields a tolled user equilibrium solution that is approximately system optimal. To establish this theorem, the following lemmas are necessary.

Lemma 1. *Let P be a compact set, $c_1(\cdot)$ be continuous and strongly monotone with modulus α (i.e., $(c_1(v_1) - c_1(v_2))^T (v_1 - v_2) \geq \alpha ||v_1 - v_2||^2$), and $c_2(\cdot)$ be continuous. If p_1 and p_2 solve VI$[c_1(\cdot), P]$ and VI$[c_2(\cdot), P]$, respectively, then $||p_2 - p_1|| \leq \frac{1}{\alpha} ||c_2(p_2) - c_1(p_2)||$.*

Proof. See Dafermos and Nagurney [DaN84]. ■

Lemma 2. *For $i = 1$ and 2, let $F_i = \{x | U_i x \leq r_i,\ W_i x = t_i\}$, where U_i and W_i are $(l \times n)$ and $(m \times n)$ matrices, respectively, and r_i and t_i are vectors in R^l and R^m, respectively. If $x_2 \in F_2$, then there exists a $x_1 \in F_1$ such that $\|x_1 - x_2\| \leq \sigma(U_1, W_1) \left\| \begin{matrix} [(U_1 - U_2)x_2 - (r_1 - r_2)]^+ \\ [(W_1 - W_2)x_2 - (t_1 - t_2)] \end{matrix} \right\|_2$, where $\sigma(U_1, W_1)$ is a finite real number associated with U_1 and W_1.*

Proof. See Robinson [Rob73]. ■

Lemma 3. *Let $s(v)$ be continuously differentiable and $\|\widetilde{v} - \overline{v}\| \leq \delta$ for some $\delta > 0$. For any $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon)$, there must exist a $\beta \in \mathcal{T}^+(\overline{v})$ and constants K_1 and K_2 such that $\|\beta' - \beta\| \leq K_1\delta + K_2\epsilon$.*

Proof. The conditions defining $\mathcal{T}^+(\overline{v})$ and $\mathcal{T}^+(\widetilde{v}, \epsilon)$ can be written more compactly as follows:

$$\begin{cases} -\beta \leq 0, \\ \boldsymbol{A}^T\rho - \boldsymbol{I}\beta \leq \boldsymbol{I}s(\overline{v}), \\ -\boldsymbol{b}^T\rho + \overline{v}^T\beta \leq -s(\overline{v})^T\overline{v}, \end{cases} \tag{5}$$

and

$$\begin{cases} -\beta \leq 0, \\ \boldsymbol{A}^T\rho' - \boldsymbol{I}\beta' \leq \boldsymbol{I}s(\widetilde{v},) \\ -\boldsymbol{b}^T\rho' + \widetilde{v}^T\beta' \leq -s(\widetilde{v})^T\widetilde{v} + \epsilon. \end{cases} \tag{6}$$

where $\boldsymbol{A} = \mathrm{diag}(A, A, \cdots, A)$, and $\boldsymbol{b}^T = ({b_1}^T, {b_2}^T, \cdots, {b_{|\mathcal{K}|}}^T)$. To further simplify our notation, let (U_1, r_1) and (U_2, r_2) denote the pairs of matrix and right-hand-side vector for (5) and (6), respectively. Because $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon)$, there must exists a ρ' such that (β', ρ') solves (6). From Lemma 2, there must exist a pair (β, ρ) satisfying (5) for which the following hold

$$\begin{aligned}
&\left\| \begin{pmatrix} \rho' \\ \beta' \end{pmatrix} - \begin{pmatrix} \rho \\ \beta \end{pmatrix} \right\|_2 \\
&\leq \sigma(U_1) \left\| [(U_1 - U_2) \begin{pmatrix} \rho' \\ \beta' \end{pmatrix} - (r_1 - r_2)]^+ \right\|_2 \\
&\leq \sigma(U_1) \left\| (U_1 - U_2) \begin{pmatrix} \rho' \\ \beta' \end{pmatrix} - (r_1 - r_2) \right\|_2 \\
&\leq \sigma(U_1)(\left\| \begin{pmatrix} 0 \\ 0 \\ (\overline{v} - \widetilde{v})^T\beta' \end{pmatrix} \right\|_2 + \left\| \begin{pmatrix} 0 \\ \boldsymbol{I}(s(\overline{v}) - s(\widetilde{v})) \\ -s(\overline{v})^T\overline{v} + s(\widetilde{v})^T\widetilde{v} - \epsilon \end{pmatrix} \right\|_2) \\
&\leq \sigma(U_1)(\|\beta'\|_2\|\overline{v} - \widetilde{v}\|_2 + \|s(\overline{v}) - s(\widetilde{v})\|_2 + \|s(\overline{v})^T\overline{v} - s(\widetilde{v})^T\widetilde{v}\| + \epsilon) \\
&\leq \sigma(U_1)(\|\beta'\|_2\|\overline{v} - \widetilde{v}\|_2 + \| \nabla s(u_1)\|_2\|\overline{v} - \widetilde{v}\|_2 \\
&\qquad + \|s(u_2) + \nabla s(u_2)^T s(u_2)\|_2\|\overline{v} - \widetilde{v}\|_2 + \epsilon) \\
&\leq \sigma(U_1)(B\|\overline{v} - \widetilde{v}\|_2 + L_1\|\overline{v} - \widetilde{v}\|_2 + L_2\|\overline{v} - \widetilde{v}\|_2 + \epsilon),
\end{aligned}$$

where the first inequality follows from Lemma 2, the second from the fact that $||[x]^+|| \leq ||x||$, the third from the definition of U_i and r_i and the triangle inequality and the fourth from Cauchy-Schwarz inequality. In the fifth inequality, u_1 and u_2 are some points between $\overline{v}$ and $\widetilde{v}$ and the gradient of $s(v)^T v$ is $s(v) + \bigtriangledown s(v)^T v$. Furthermore, the inequality holds because of Cauchy-Schwarz inequality, the differentiability of $s(v)$, and the mean value theorem. Finally, the last inequality is true because we assume earlier that $||\beta|| \leq B$ and the continuous functions $\bigtriangledown s(v)$ and $s(v)+\bigtriangledown s(v)^T v$ are bounded on the compact set V by some constants L_1 and L_2, respectively. By letting $K_1 = (B + L_1 + L_2)\sigma(U_1)$ and $K_2 = \sigma(U_1)$, the above reduces to $||\beta' - \beta|| \leq \left\| \begin{pmatrix} \rho' - \rho \\ \beta' - \beta \end{pmatrix} \right\|_2 \leq K_1\delta + K_2\epsilon.$ ■

Theorem 5. *Let $s(\cdot)$ be strongly monotone with modulus α. For any $\eta > 0$, there exists a $\delta > 0$ such that $||v^*(\beta') - \overline{v}|| \leq \eta$ whenever $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon_{mscp})$ and $||\widetilde{v} - \overline{v}|| \leq \delta$.*

Proof. For any $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$, Lemma 3 implies that there exists a $\beta \in \mathcal{T}^+(\overline{v})$ such that $||\beta' - \beta|| \leq K_1\delta + K_2\epsilon_{\text{mscp}}$. As defined earlier, ϵ_{mscp} depends on $\widetilde{v}$. In particular, $\epsilon_{\text{mscp}} \to 0$ as $\delta \to 0$. When combining the latter with the fact that α, K_1, and K_2 are constant and independent of η it must be possible to choose δ so that $(1/\alpha)(K_1\delta + K_2\epsilon_{\text{mscp}}) \leq \eta$.

Let $\beta' \in \mathcal{T}^+(\widetilde{v}, \epsilon_{\text{mscp}})$ and $\beta \in \mathcal{T}^+(\overline{v})$. Then Lemma 1 implies that the solutions $v^*(\beta')$ and $\overline{v}$ to VI$[s(v) + \beta', V]$ and VI$[s(v) + \beta, V]$, respectively, must satisfy

$$\begin{aligned} ||\overline{v} - v^*(\beta')|| &\leq (1/\alpha)||s(\overline{v}) + \beta - s(\overline{v}) - \beta'|| \\ &\leq (1/\alpha)||\beta - \beta'|| \\ &\leq (1/\alpha)(K_1\delta + K_2\epsilon_{\text{mscp}}). \end{aligned}$$

Therefore, the above choice of δ implies the theorem holds. ■

Because $\epsilon^* \leq \epsilon_{\text{mscp}}$, the above theorem also holds when ϵ^* replaces ϵ_{mscp}.

4 Disaggregate Representation of Relaxed Toll Sets

The second condition (2) in Theorem 1 is an aggregation of a number of complementarity conditions as shown in the proof of Theorem 2. When (2) is relaxed, the resulting relaxed toll set $\mathcal{T}^+(\widetilde{v}, \epsilon)$ may be larger than necessary. To define smaller relaxed toll sets, (2) can be disaggregated into its original form.

Using the argument in, e.g., Theorem 2, it is possible to show that $\mathcal{T}^+(\overline{v})$ is equivalent to the set consisting of the β component of every vector (β, ρ, σ) that satisfies the following linear system

$$\begin{aligned} s(\overline{v}) + \beta - A^T\rho^k &= \sigma^k, \forall k \in \mathcal{K}, \\ (\overline{x})^T \sigma^k &= 0, \quad \forall k \in \mathcal{K}, \\ \sigma^k &\geq 0, \quad \forall k \in \mathcal{K}, \\ \beta &\geq 0. \end{aligned}$$

The second equation is an aggregation of the complementarity condition for each arc $(i,j) \in \mathcal{A}$, i.e., $x^k_{ij}\sigma^k_{ij} = 0$. Thus, the above system is equivalent to the following:

$$\begin{aligned} s_{ij}(\overline{v}) + \beta_{ij} &\geq \rho^k_i - \rho^k_j, \forall k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ s_{ij}(\overline{v}) + \beta_{ij} &\leq \rho^k_i - \rho^k_j, \forall k \in \mathcal{K}, (i,j) \in \mathcal{A} : \overline{x}^k_{ij} > 0, \\ \beta_{ij} &\geq 0 \qquad \forall (i,j) \in \mathcal{A}. \end{aligned}$$

As before, let $\widetilde{v} = \sum_k \widetilde{x}^k$ denote an approximate SOPT solution. Then, a relaxed toll set in the disaggregate form, $\Pi^+(\widetilde{v}, \xi)$, is the set consisting of the β component of every vector (β, ρ) that satisfies the following linear system

$$\begin{aligned} s_{ij}(\widetilde{v}) + \beta_{ij} &\geq \rho^k_i - \rho^k_j, \qquad \forall k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ s_{ij}(\widetilde{v}) + \beta_{ij} &\leq \rho^k_i - \rho^k_j + \xi^k_{ij}, \forall k \in \mathcal{K}, (i,j) \in \mathcal{A} : \widetilde{x}^k_{ij} > 0, \\ \beta_{ij} &\geq 0 \qquad \forall (i,j) \in \mathcal{A}. \end{aligned}$$

Unlike ϵ (a constant) in $\mathcal{T}^+(\widetilde{v}, \epsilon)$, ξ is a nonnegative vector in the relaxed toll set $\Pi^+(\widetilde{v}, \xi)$. Below are two properties of this (disaggregate) toll set.

Theorem 6. *For any $\widetilde{v} \in V$, let $\beta_{mscp} = \nabla s(\widetilde{v})^T \widetilde{v}$ and, for all $k \in \mathcal{K}$ and $(i,j) \in \mathcal{A}$ such that $\widetilde{x}^k_{ij} > 0$, let $\widetilde{\xi}^k_{ij} = s_{ij}(\widetilde{v}) + [\beta_{mscp}]_{ij} - \widetilde{\rho}^k_i + \widetilde{\rho}^k_j$, where $\widetilde{\rho}$ is an optimal solution to the linear program in Theorem 4. If $\nabla s(\widetilde{v})$ is nonnegative, then $\Pi^+(\widetilde{v}, \widetilde{\xi}) \neq \emptyset$.*

Proof. Recall from Theorem 4 that

$$\epsilon_{\text{mscp}} = \min_{\rho}\{(s(\widetilde{v}) + \beta_{\text{mscp}})^T\widetilde{v} - \sum\nolimits_k b^T_k \rho^k : s(\widetilde{v}) + \beta_{\text{mscp}} \geq A^T\rho^k, \forall k\}.$$

For all $k \in \mathcal{K}$ and $(i,j) \in \mathcal{A}$ such that $\widetilde{x}^k_{ij} > 0$, let $\widetilde{\xi}^k_{ij} = s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} - \widetilde{\rho}^k_j + \widetilde{\rho}^k_i$, where $\widetilde{\rho}$ is an optimal solution to the above linear program. Moreover, its constraints also ensure that $\widetilde{\xi}^k_{ij} \geq 0$ and, when combined with the definition of $\widetilde{\xi}^k_{ij}$, the following must hold

$$\begin{aligned} s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} &\geq \widetilde{\rho}^k_i - \widetilde{\rho}^k_j, \qquad \forall k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} &= \widetilde{\rho}^k_i - \widetilde{\rho}^k_j + \widetilde{\xi}^k_{ij}, \forall k \in \mathcal{K}, (i,j) \in \mathcal{A} : \widetilde{x}^k_{ij} > 0. \end{aligned}$$

Then, the nonnegativity of $\nabla s(\widetilde{v})^T \widetilde{v}$ implies that $[\beta_{\text{mscp}}]_{ij} \geq 0$ for all $(i,j) \in \mathcal{A}$. Therefore, $\beta_{\text{mscp}} \in \Pi^+(\widetilde{v}, \widetilde{\xi})$ and $\Pi^+(\widetilde{v}, \widetilde{\xi}) \neq \emptyset$. ■

Theorem 7. *If $\Pi^+(\widetilde{v}, \xi) \neq \emptyset$, then $\Pi^+(\widetilde{v}, \xi) \subseteq \mathcal{T}^+(\widetilde{v}, \epsilon)$, where $\epsilon = \sum_k \sum_{(i,j)\in\mathcal{A}} \widetilde{x}^k_{ij}\xi^k_{ij}$.*

Proof. Multiplying the second equation in the definition of $\Pi^+(\tilde{v},\xi)$ by $\tilde{x}^k_{ij}$ yields

$$(s_{ij}(\tilde{v})+\beta_{ij})\tilde{x}^k_{ij} \le (\rho^k_i-\rho^k_j)\tilde{x}^k_{ij}+\xi^k_{ij}\tilde{x}^k_{ij},\ \forall(i,j,k):\tilde{x}^k_{ij}>0.$$

Then, summing the above equations together and recognizing that $A\tilde{x}^k=b_k$ yield that $(s(\tilde{v})+\beta)^T\tilde{v} \le \sum_{k\in\mathcal{K}} b^T_k\rho^k+\epsilon$, where $\epsilon=\sum_k\sum_{(i,j)\in\mathcal{A}}\tilde{x}^k_{ij}\xi^k_{ij}$. Thus, $\beta\in\Pi^+(\tilde{v},\xi)$ implies that $\beta\in\mathcal{T}^+(\tilde{v},\epsilon)$, i.e., $\Pi^+(\tilde{v},\xi)\subseteq\mathcal{T}^+(\tilde{v},\epsilon)$. ■

Instead of choosing ξ as in Theorem 6, it is also possible to choose ξ that solves one of the following two problems:

$$\begin{array}{lll}
\min\limits_{(\beta,\rho,\xi)} & \sum_k\sum_{(i,j):\tilde{x}^k_{ij}>0}\xi^k_{ij} & \\
s.t. & s_{ij}(\tilde{v})+\beta_{ij}=\rho^k_i-\rho^k_j+\xi^k_{ij}, & \forall k\in\mathcal{K},(i,j)\in\mathcal{A},\\
 & \xi^k_{ij}\ge 0, & \forall k\in\mathcal{K},(i,j)\in\mathcal{A},\\
 & \beta_{ij}\ge 0 & \forall(i,j)\in\mathcal{A},
\end{array}$$

or

$$\begin{array}{lll}
\min\limits_{(\beta,\rho,\xi,z)} & z & \\
s.t. & s_{ij}(\tilde{v})+\beta_{ij}=\rho^k_i-\rho^k_j+\xi^k_{ij}, & \forall k\in\mathcal{K},(i,j)\in\mathcal{A},\\
 & \xi^k_{ij}\le z, & \forall k\in\mathcal{K},(i,j)\in\mathcal{A}:\tilde{x}^k_{ij}>0,\\
 & \xi^k_{ij}\ge 0 & \forall k\in\mathcal{K},(i,j)\in\mathcal{A},\\
 & \beta_{ij}\ge 0 & \forall(i,j)\in\mathcal{A}.
\end{array}$$

Both problems yield a ξ that makes $\Pi^+(\tilde{v},\xi)$ nonempty.

5 Numerical Results

To illustrate the effectiveness of the relaxed toll sets, $\Pi^+(\tilde{v},\xi)$ and $\mathcal{T}^+(\tilde{v},\epsilon)$, we solved the MINSYS problem originally introduced in [Ber95] and [BHR97] and later referred to as the minimum toll revenue problem in [Dia99]. Using the (aggregate) relaxed toll set $\mathcal{T}^+(\tilde{v},\epsilon)$, the (aggregate) minimum toll revenue (AMR) problem can be stated as follows:

$$\min\{\tilde{v}^T\beta:\beta\in\mathcal{T}^+(\tilde{v},\epsilon)\}.$$

The objective function in AMR is simply the sum of the product of the flow and the toll amount on each arc, i.e., the toll revenue. Using $\Pi^+(\tilde{v},\xi)$ instead of $\mathcal{T}^+(\tilde{v},\epsilon)$, the disaggregate minimum toll revenue (DMR) problem can be defined as follows:

$$\min\{\tilde{v}^T\beta:\beta\in\Pi^+(\tilde{v},\xi)\}.$$

Data for our experiments are from four transportation networks whose attributes are listed in Table 2. For each network, we used the restricted simplicial decomposition or RSD (see, Hearn et al. [HLV87]) to obtain a solution

Table 2. Network Attributes

Network	Links	Nodes	Commodities
Sioux Falls[LMP75]	76	24	528
Hull [FGS87]	798	501	138
Stockholm [HeR98]	962	416	1,623
Winnipeg [FGS87]	2836	1052	4,344

to SOPT with a relative optimality gap of 10^{-4}. Because they are readily available from RSD, we set $\epsilon = \epsilon_{\text{mscp}}$ and $\xi^k_{ij} = s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} - \rho^k_i + \rho^k_j$ in AMR and DMR, respectively. Both problems, AMR and DMR, were implemented in GAMS [GAM80] and solved using CPLEX 8.1 [CPL96].

Table 3 reports the results for the four networks. For Sioux Falls, the SOPT solution from RSD provides a consistent toll set and ϵ can be set to zero. The same does not hold for the remaining three networks. The values of their ϵ_{mscp} are listed in the table along with the ratio $\epsilon_{\text{mscp}}/s(\widetilde{v})^T\widetilde{v}$ to provide the magnitude of ϵ_{mscp} relative to the total travel delay at the approximate SOPT solution. The last two sets of columns compare the tolled user equilibrium solutions, $v^*(\beta)$, using toll vectors from AMR and DMR against the approximate SOPT solution, $\widetilde{v}$, from RSD. The two columns under the heading "Total Delay Error" reports $(s(v^*(\beta))^T v^*(\beta) - s(\widetilde{v})^T\widetilde{v})/(s(\widetilde{v})^T\widetilde{v})$, i.e., the error in travel delay relative to the delay at the approximate system solution, $\widetilde{v}$. The remaining two columns (under the heading "Link Flow Error") reports the percentage of arcs with relatively large link flow errors. In calculating this percentage, we only consider arcs with a moderately large amount of flow, i.e., we consider arcs in the set $\mathcal{A}' = \{a | v^*_a(\beta) \geq 0.25C_a \text{ or } \widetilde{v}_a \geq 0.25C_a\}$, where C_a is the capacity of arc a. Then, the link flow error is the percentage of arcs in $\mathcal{A}'$ such that $\frac{|v^*_a(\beta)-\widetilde{v}_a|}{\widetilde{v}_a} > 0.10$. Observe that results in the last two columns indicate that the relaxation based on the marginal social costs produces good toll vectors for they yield tolled user equilibrium solutions that are approxi-

Table 3. Numerical Results

			Total Delay Error		Link Flow Error	
Networks	ϵ_{mscp}	$\frac{\epsilon_{\text{mscp}}}{s(\widetilde{v})^T\widetilde{v}}$	(AMR)	(DMR)	(AMR)	(DMR)
Sioux Falls	0	0	0%	0%	0%	0%
Hull	4.85	9.59E-5	0.07%	0.07%	2.6%	1.3%
Stockholm	1,134.12	9.74E-5	0.06%	0.01%	0%	0%
Winnipeg	107.82	9.22E-5	0.05%	0.04%	0.1%	0.3%

mately optimal to SOPT. However, DMR on average yields tolls with slightly less error.

6 Conclusions

Congestion or toll pricing problems in [HeR98] require a solution to the system problem (the traffic assignment problem that minimizes the total travel delay) to define a toll set, i.e., a set of all valid tolls. Instead of an exact solution, it is more practical to obtain an approximate solution to the system problem for large networks. In this paper, we provide necessary and sufficient conditions to determine whether the toll set constructed from an approximate solution is empty. When it is so, we derive alternative toll sets based on relaxed optimality conditions. With carefully chosen parameters, tolls from the relaxed toll sets possess the desirable property, i.e., they induce travellers to choose routes that are nearly system optimal. Numerical solutions from four transportation networks in the literature also verify empirically the previous statement.

Acknowledgements. This research was partially supported by NSF grants DMI-9978642 and DMI-0300316.

References

[ArS94] Arnott, R., Small, K.: The Economics of Traffic Congestion. American Scientist, **82**, 446–455 (1994).

[BHL03] Bai, L., Hearn, D.W., Lawphongpanich, S.: A Heuristic Method for the Minimum Toll Booth Problem. Technical Report, Department of Industrial and Systems Engineering, Univeristy of Florida, Gainesville, FLorida (2003)

[BHL04] Bai, L., Hearn, D.W., Lawphongpanich, S.: Decomposition Techniques on the Minimum Toll Revenue Problem. Networks, forthcoming in 2004.

[BSS93] Bazaraa, M., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithm. John Wiley & Sons (1993)

[Ber95] Bergendorff, P.: The Bounded Flow Approach to Congestion Pricing. Masters Thesis, Royal Institute of Technology, Stockholm (1995)

[BHR97] Bergendorff, P., Hearn, D.W., Ramana, M.V.: Congestion Toll Pricing of Traffic Networks. In: Pardalos, P.M., Hearn, D.W., Hagers, W.H. (eds) Network Optimization. Lecture Notes in Economics and Mathematical Systems, Vol **450**, Springer-Verlag, 51–71 (1997)

[CPL96] CPLEX, CPLEX Optimization, Inc., Incline Village, NV (1996)

[DaN84] Dafermos, S., Nagurney, A.: Sensitivity Analysis for the Asymmetric Network Equilibrium Problem. Mathematical Programming, **28**, 174–184 (1984)

[Dia99] Dial, R.: Minimal Revenue Congestion Pricing Part I: A Fast Algorithm for the Single-Origin Case. Transportation Research-B, **33**, 189–202 (1999)

[FaP03] Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol **1**, Springer (2003)

[FJM04] Fleischer, L., Jain, K., Mahdian, M.: Tolls for heterogeneous users in Multicommodity Networks and Generalized Congestion Games. In Proceedings 45th Annual Symposium on Foundations of Computer Science, IEEE, 277–285 (2004)

[FGS87] Florian, M., Guélat, J., Spiess, H.: An Efficient Implementation of the PARTAN Variant of the Linear Approximation Method for the Network Equilibrium Problem. Networks, **17**, 319–339 (1987)

[Hea80] Hearn, D.W.: Bounding Flows in Traffic Assignment Models. Research Report No. 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL (1980)

[FlH95] Florian, M., Hearn, D.W.: Network Equilibrium Model and Algorithms. Network Routing, In: Ball, M.O. et al. (eds.) Handbook in OR and MS. Vol. **8**, Elsevier Science (1995)

[GAM80] GAMS, General Algebraic Modeling System, GAMS Development Corporation (1995)

[Gar80] Gartner, N.H.: Optimal traffic assignment with elastic demands: A Review, Part I. Analysis Framework. Transportation Science, **14**(2), 174–191 (1980)

[Hea82] Hearn, D.W.: The Gap Function of a Convex Program. Operations Research Letters, **1**, 67–71 (1982)

[HLV87] Hearn, D.W., Lawphongpanich, S., Ventura, J.: Restricted Simplicial Decomposition: Computation and Extensions. Mathematical Programming Study, **31**, 99–118 (1987)

[HeR98] Hearn, D.W., Ramana, M.: Solving Congestion Toll Pricing Models. In: Marcotte, P., Nguyen, S. (eds.) Equilibrium and Advanced Transportation Modeling. Kluwer Academic Publishers, 109–124 (1998)

[HYR01] Hearn, D.W., Yildirim, M.B., Ramana, M., Bai, L.: Computational Methods for Congestion Toll Pricing Models. In IEEE Intelligent Transportation System Conference Proceedings, Oakland, 257–262 (2001)

[HeY01] Hearn, D.W., Yildirim, M.B.: A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand. In: Gendreau, M., Marcotte, P. (eds.) Current Trends in Transportation and Network Analysis: Miscellanea in Honor of Michael Florian, Kluwer Academic Publishers (2001)

[LMP75] LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P.: An Efficient Approach to Solving the Road Network Equilibrium Traffic Assignment Problem. Transportation Research, **9**, 309–318 (1975)

[Rob73] Robinson, S.: Bounds for Error in the Solution Set of a Perturbed Linear Program. Linear Algebra and its Applications, **6**, 69–81 (1973)

[YaB97] Yang, H., Bell, M.G.H.: Traffic Restraint, Road Pricing and Network Equilibrium. Transportation Research B: Methodological, **33(4)**, 303–314 (1997)

[ZhG97] Zhang, H.M., Ge, Y.E.: Modeling variable demand equilibrium under second-best road pricing. Working Paper, Institute of Transportation Studies, University of California at Davis (2002)

Appendix

Relaxed Toll Sets for the Elastic Demand Case

This appendix describes the results concerning the toll sets when demands are elastic. Many results for the fixed demand case naturally extend to the case with elastic demands. The presentation below follows the same outline as in the main part of the paper.

A Elastic Demand System and User Problems

To state the traffic assignment problems with elastic demands, let t_k and $w_k(t_k)$ denote the travel demand and the inverse demand function for commodity k, respectively. For each k, E_k is a vector in $R^{|\mathcal{N}|}$ with exactly two nonzero elements, one equals 1 at the origin node and the other equals -1 at the destination node. Then, the set of feasible flow-demand vectors is

$$V_{\text{ED}} = \{(v,t)|v = \sum_k x^k, Ax^k = E_k t_k, x^k \geq 0, t_k \geq 0\}.$$

Without loss of generality, we assume V_{ED} is bounded, thus, compact. (See, e.g., [FlH95].)

Among several alternatives (see, e.g., [Gar80], [YaB97] and [ZhG97]), one system problem with elastic demands maximizes the net user benefit, i.e., the difference between the user benefit as measured by $\sum_{k\in\mathcal{K}} \int_0^{t_k} w_k(z)dz$ and the total delay (or cost) $s(v)^T v$. In its minimization form, this system problem can be written as

$$(\overline{v},\overline{t}) = \text{argmin}\{s(v)^T v - \sum_{k\in\mathcal{K}} \int_0^{t_k} w_k(z)dz : (v,t) \in V_{\text{ED}}.\}$$

As in the fixed demand case, the corresponding user problem with elastic demand is a variational inequality. In particular, (v^*,t^*) is a solution to the user equilibrium problem if the pair satisfies the following:

$$s(v^*)^T(v - v^*) - w(t^*)^T(t - t^*) \geq 0, \quad \forall\ (v,t) \in V_{\text{ED}}.$$

For a given toll vector β, $(v^*(\beta), t^*(\beta))$ is a solution to the tolled user equilibrium problem if the pair satisfies the following:

$$(s(v^*(\beta)) + \beta)^T(v - v^*(\beta)) - w(t^*(\beta))^T(t - t^*(\beta)) \geq 0, \quad \forall\ (v,t) \in V_{\text{ED}}.$$

As in the fixed demand case, we assume throughout this appendix that the system, user, and tolled user equilibrium problems have unique solutions.

B System and Non-system Toll Sets

Analogous to the fixed demand case, the system toll set when demands are elastic is $\mathcal{T}(\overline{v},\overline{t}) = \{\beta | v^*(\beta) = \overline{v}, t^*(\beta) = \overline{t}\}$. Under the uniqueness assumptions stated earlier, Hearn and Yildirim [HeY01] prove that $\mathcal{T}(\overline{v},\overline{t})$ consists of the β component of every pair (β, ρ) that satisfies the following system:

$$\begin{aligned} s(\overline{v}) + \beta &\geq A^T\rho^k, \quad \forall\ k \in \mathcal{K}, \\ w_k(\overline{t}_k) &\leq E_k^T\rho^k, \quad \forall\ k \in \mathcal{K}, \\ (s(\overline{v}) + \beta)^T\overline{v} &= w(\overline{t})^T\overline{t}. \end{aligned}$$

In [HeY01], Hearn and Yildirim show that both $\mathcal{T}(\overline{v},\overline{t})$ and $\mathcal{T}^+(\overline{v},\overline{t})$ are nonempty. The latter assumes that $\bigtriangledown s(\overline{v})$ is nonnegative.

Let $(\widetilde{v},\widetilde{t})$ denote a flow-demand vector feasible to V_{ED}. Then, the non-system toll set is $\mathcal{T}(\widetilde{v},\widetilde{t}) = \{\beta | v^*(\beta) = \widetilde{v}, t^*(\beta) = \widetilde{t}\}$ and, using an argument similar to the one in Theorem 2, the following holds.

Theorem 8. *The toll set $\mathcal{T}(\widetilde{v},\widetilde{t})$, where $(\widetilde{v},\widetilde{t}) \in V_{ED}$, is the set consisting of the β component of every pair (β, ρ) that satisfies the following linear system:*

$$\begin{aligned} s(\widetilde{v}) + \beta &\geq A^T\rho^k, \quad \forall\ k \in \mathcal{K}, \\ w_k(\widetilde{t}_k) &\leq E_k^T\rho^k, \quad \forall\ k \in \mathcal{K}, \\ (s(\widetilde{v}) + \beta)^T\widetilde{v} &= w(\widetilde{t})^T\widetilde{t}. \end{aligned}$$

The theorem below shows that the non-system toll set is nonempty for any non-trivial $(\widetilde{v},\widetilde{t}) \in V_{\text{ED}}$.

Theorem 9. *For any $(\widetilde{v},\widetilde{t}) \in V_{ED}$ such that $\widetilde{v} \neq 0$, $\widetilde{\beta} = \nabla s(\widetilde{v})^T\widetilde{v} - \alpha\widetilde{v} \in \mathcal{T}(\widetilde{v},\widetilde{t})$ when $\alpha = \left[\left(s(\widetilde{v}) + \nabla s(\widetilde{v})^T\widetilde{v}\right)^T \widetilde{v} - w(\widetilde{t})^T\widetilde{t}\right]/\widetilde{v}^T\widetilde{v}$.*

Proof. Consider the following direction finding problem associated with $\text{VI}(s(v)+\widetilde{\beta}, V_{\text{ED}})$ at $(\widetilde{v},\widetilde{t})$:

$$\begin{aligned} \text{DIR-ED}(\widetilde{\beta}) : \min\ & \left(s(\widetilde{v}) + \widetilde{\beta}\right)^T \sum_{k\in K} x^k - w(\widetilde{t})^T d \\ s.t\ & Ax^k - E_k d_k = 0, && \forall k, \\ & x^k \geq 0, && \forall k, \\ & d_k \geq 0, && \forall k. \end{aligned}$$

The dual of $\text{DIR-ED}(\widetilde{\beta})$ is

$$\begin{aligned} & \max 0 \\ s.t.\ & A^T\rho^k \leq s(\widetilde{v}) + \widetilde{\beta}, \forall k, \\ & E_k^T\rho^k \geq w_k(\widetilde{t}_k), \quad \forall k, \\ & \rho^k \text{ unrestricted}, \quad \forall k. \end{aligned}$$

The relationships between the primal and dual problems in linear programming imply that the objective value of the direction finding problem is bounded below by zero. Thus, $(u, d) = (0, 0)$, where $u = \sum_k x^k$, is an optimal solution because its objective value equals the lower bound. Furthermore, the dual of DIR-ED has a feasible solution, say $\widetilde{\rho}$. Then the pair $(\widetilde{\beta}, \widetilde{\rho})$ satisfies the linear system in Theorem 1. The first two conditions of the linear system in Theorem 8 follow from the first two constraints of the dual problem and our choice of $\widetilde{\beta}$ ensures that the following holds

$$\left(s(\widetilde{v}) + \widetilde{\beta}\right)^T \widetilde{v} = \left(s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v}\right)^T \widetilde{v} - \alpha \widetilde{v}^T \widetilde{v} = w(\widetilde{t})^T \widetilde{t}.$$

Thus, the last condition in the linear system is also satisfied and $\widetilde{\beta} \in \mathcal{T}(\widetilde{v}, \widetilde{t})$. ■

As defined above, α is zero and $\widetilde{\beta} = \nabla s(\widetilde{v})^T \widetilde{v}$, when $(\widetilde{v}, \widetilde{t})$ solves the system problem. Moreover, other choices for α and $\widetilde{\beta}$ exist. For example, $\widetilde{\beta} = \nabla s(\widetilde{v})^T \widetilde{v} - \alpha s(\widetilde{v})$, where $\alpha = [\left(s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v}\right)^T \widetilde{v} - w(\widetilde{t})^T \widetilde{t}] / s(\widetilde{v})^T \widetilde{v}$, is also valid when $s(\widetilde{v})^T \widetilde{v} \neq 0$.

The following theorem provides a necessary and sufficient condition under which the nonnegative and non-system toll set is nonempty. The proof is omitted because it is similar to that of Theorem 3 in the main part of the paper.

Theorem 10. *For any* $(\widetilde{v}, \widetilde{t}) \in V_{ED}$, $\mathcal{T}^+(\widetilde{v}, \widetilde{t})$ *is nonempty if and only if* $(\widetilde{v}, \widetilde{t})$ *solves* VI*[*$(s(v), -w(t)), \mathcal{V}_{ED}$*], where* $\mathcal{V}_{ED} = \{(v, t) | v = \sum_k x^k, Ax^k = E_k t_k, x^k \geq 0, t_k \geq 0, v \leq \widetilde{v}\}$.

C Relaxed Toll Set

In this and the following sections, we focus on the relaxations of the (unrestricted) non-system toll set. However, similar results also hold for the non-system toll set requiring tolls to be nonnegative.

For a given $\epsilon > 0$, the relaxed toll set $\mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$ is the set consisting of the β component of the pair (β, ρ) that satisfies the following:

$$\begin{aligned} s(\widetilde{v}) + \beta &\geq A^T \rho^k, && \forall k \in \mathcal{K}, \\ w_k(\widetilde{t}_k) &\leq E_k^T \rho^k, && \forall k \in \mathcal{K}, \\ (s(\widetilde{v}) + \beta)^T \widetilde{v} &\leq w(\widetilde{t})^T \widetilde{t} + \epsilon. \end{aligned}$$

Then, the following results are analogous to those in Section 3.

Theorem 11. *For any* $(\widetilde{v}, \widetilde{t}) \in V_{ED}$, *let* $\epsilon_{mscp} = (s(\widetilde{v}) + \nabla s(\widetilde{v})^T \widetilde{v})^T \widetilde{v} - w(\widetilde{t})^T \widetilde{t}$. *Then,* $\mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon_{mscp}) \neq \emptyset$.

Proof. From the discussion in Section B, the optimal objective value of DIR-ED (β_{mscp}) is zero, where $\beta_{\text{mscp}} = \nabla s(\widetilde{v})^T \widetilde{v}$ as before. Thus ϵ_{mscp} can be equivalently expressed as follows:

$$\begin{aligned}\epsilon_{\text{mscp}} = (s(\widetilde{v}) + \beta_{\text{mscp}})^T \widetilde{v} - w(\widetilde{t})^T \widetilde{t} \\ - \min\{(s(\widetilde{v}) + \beta_{\text{mscp}})^T u - w(\widetilde{t})^T t : (u, t) \in V_{\text{ED}}\}.\end{aligned}$$

As in Theorem 4 in the main part of the paper, the vector β_{mscp} and an optimal solution, $\widetilde{\rho}$, to the dual of DIR-ED(β_{mscp}) form a pair of (β, ρ) that belongs to $\mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon_{\text{mscp}})$. ■

Theorem 12. *Let $s(v)$ and $w(t)$ be strongly monotone with modula α and γ, respectively. For any $\eta > 0$, there exists a $\delta > 0$ such that $\|(v^*(\beta) - \overline{v}, t^*(\beta) - \overline{t})\|_2 \leq \eta$ whenever $\beta \in \mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon_{mscp})$ and $\|(\widetilde{v} - \overline{v}, \widetilde{t} - \overline{t})\| \leq \delta$.*

Proof. Because both $s(v)$ and $w(t)$ are strong monotone, $(s(v), -w(t))$ is also strongly monotone with modulus $\min\{\alpha, \gamma\}$. The rest of the proof requires lemmas and uses an argument similar to the one in Theorem 5. ■

The following linear program also provides an ϵ for which $\mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$ is nonempty.

$$\begin{aligned}\epsilon^* = \min_{(\beta,\rho)} \ & (s(\widetilde{v}) + \beta)^T \widetilde{v} - w(\widetilde{t})^T \widetilde{t} \\ s.t. \ & s(\widetilde{v}) + \beta \geq A^T \rho^k, \quad \forall k \in \mathcal{K}, \\ & w_k(\widetilde{t}_k) \leq E_k^T \rho^k, \quad \forall k \in \mathcal{K}.\end{aligned}$$

D Disaggregate Representation of Relaxed Toll Sets

Let $(\widetilde{v}, \widetilde{t})$ be an approximate system solution. For a given pair of (ξ, μ) such that $\xi, \mu \geq 0$, the following are three possible disaggregate representations of a relaxed toll set, all of which are analogous to the one presented in Section 4.

1. $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$ = the set of the β component of every pair (β, ρ) that satisfies the following:

$$\begin{aligned}& s_{ij}(\widetilde{v}) + \beta_{ij} \geq \rho_i^k - \rho_j^k, && \forall\, k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ & s_{ij}(\widetilde{v}) + \beta_{ij} \leq \rho_i^k - \rho_j^k + \xi_{ij}^k, && \forall\, k \in \mathcal{K}, (i,j) \in \mathcal{A} : \widetilde{x}_{ij}^k > 0, \\ & w_k(\widetilde{t}_k) \leq E_k^T \rho^k, && \forall\, k \in \mathcal{K}, \\ & w_k(\widetilde{t}_k) \geq E_k^T \rho^k - \mu_k, && \forall\, k \in \mathcal{K} : \widetilde{t}_k > 0.\end{aligned}$$

2. $\Pi^2(\widetilde{v}, \widetilde{t}, \xi)$ = the set of the β component of every pair (β, ρ) that satisfies the following:

$$\begin{aligned}& s_{ij}(\widetilde{v}) + \beta_{ij} \geq \rho_i^k - \rho_j^k, && \forall\, k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ & s_{ij}(\widetilde{v}) + \beta_{ij} \leq \rho_i^k - \rho_j^k + \xi_{ij}^k, && \forall\, k \in \mathcal{K}, (i,j) \in \mathcal{A} : \widetilde{x}_{ij}^k > 0, \\ & w_k(\widetilde{t}_k) \leq E_k^T \rho^k, && \forall\, k \in \mathcal{K}.\end{aligned}$$

3. $\Pi^3(\widetilde{v}, \widetilde{t}, \mu)$ = the set of the β component of every pair (β, ρ) that satisfies the following:

$$\begin{array}{ll} s_{ij}(\widetilde{v}) + \beta_{ij} \geq \rho_i^k - \rho_j^k, & \forall\ k \in \mathcal{K}, (i,j) \in \mathcal{A}, \\ w_k(\widetilde{t}_k) \leq E_k^T \rho^k, & \forall\ k \in \mathcal{K}, \\ w_k(\widetilde{t}_k) \geq E_k^T \rho^k - \mu_k, & \forall\ k \in \mathcal{K} : \widetilde{t}_k > 0. \end{array}$$

Because the last two sets contain subsets of the constraints appeared in the first, $\Pi^2(\widetilde{v}, \widetilde{t}, \xi)$ and $\Pi^3(\widetilde{v}, \widetilde{t}, \mu)$ are relaxations of $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$. Thus, $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$ must be a subset of both $\Pi^2(\widetilde{v}, \widetilde{t}, \xi)$ and $\Pi^3(\widetilde{v}, \widetilde{t}, \mu)$.

The following theorem shows that $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$ is nonempty. This in turn implies that both $\Pi^2(\widetilde{v}, \widetilde{t}, \xi)$ and $\Pi^3(\widetilde{v}, \widetilde{t}, \mu)$ are also nonempty.

Theorem 13. *For any $(\widetilde{v}, \widetilde{t}) \in V_{ED}$, let*

1. $\widetilde{\xi}_{ij}^k = s_{ij}(\widetilde{v}) + [\beta_{mscp}]_{ij} - \widetilde{\rho}_i^k + \widetilde{\rho}_j^k$, for all k and arc (i,j) such that $\widetilde{x}_{ij}^k > 0$, and

2. $\widetilde{\mu}_k = E_k^T \widetilde{\rho}^k - w_k(\widetilde{t}_k)$, for all k such that $\widetilde{t}_k > 0$,

where $\widetilde{\rho}$ is an optimal solution to the dual of GAP-ED(β_{mscp}). Then, $\Pi^1(\widetilde{v}, \widetilde{t}, \widetilde{\xi}, \widetilde{\mu}) \neq \emptyset$.

Proof. Because $\widetilde{\rho}$ solves the dual of GAP-ED(β_{mscp}), it satisfies

$$\begin{array}{ll} s(\widetilde{v}) + \beta_{\text{mscp}} \geq A^T \widetilde{\rho}^k, & \forall k \in \mathcal{K}, \\ w_k(\widetilde{t}_k) \leq E_k^T \widetilde{\rho}^k, & \forall k \in \mathcal{K}. \end{array}$$

The above implies that both $\widetilde{\xi}_{ij}^k$ and $\widetilde{\mu}_k$ defined above are nonnegative and satisfy the following:

$$\begin{array}{lll} s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} & \geq \widetilde{\rho}_i^k - \widetilde{\rho}_j^k, & \forall k, (i,j) \in \mathcal{A}, \\ s_{ij}(\widetilde{v}) + [\beta_{\text{mscp}}]_{ij} & = \widetilde{\rho}_i^k - \widetilde{\rho}_j^k + \widetilde{\xi}_{ij}^k, & \forall k, (i,j) : \widetilde{x}_{ij}^k > 0, \\ w_k(\widetilde{t}_k) & \leq E_k^T \widetilde{\rho}^k, & \forall k, \\ w_k(\widetilde{t}_k) & = E_k^T \widetilde{\rho}^k - \widetilde{\mu}_k, & \forall k : \widetilde{t}_k > 0. \end{array}$$

Thus, $(\beta_{\text{mscp}}, \widetilde{\xi}_{ij}^k, \widetilde{\mu}^k)$ satisfies the conditions defining $\Pi^1(\widetilde{x}, \widetilde{t}, \widetilde{\xi}, \widetilde{\mu})$, i.e., $\Pi^1(\widetilde{v}, \widetilde{t}, \widetilde{\xi}, \widetilde{\mu}) \neq \emptyset$. ■

Corollary 2. *For any $(\widetilde{v}, \widetilde{t}) \in V_{ED}$, let $\widetilde{\xi}_{ij}^k = s_{ij}(\widetilde{v}) + [\beta_{mscp}]_{ij} - \widetilde{\rho}_i^k + \widetilde{\rho}_j^k$, where $\widetilde{\rho}$ is an optimal solution to the dual of GAP-ED(β_{mscp}). Then, $\Pi^2(\widetilde{v}, \widetilde{t}, \widetilde{\xi}) \neq \emptyset$.*

Corollary 3. *For any $(\widetilde{v}, \widetilde{t}) \in V_{ED}$, let $\widetilde{\mu}_k = E_k^T \widetilde{\rho}^k - w_k(\widetilde{t}_k)$, where $\widetilde{\rho}$ is an optimal solution to the dual of GAP-ED(β_{mscp}). Then, $\Pi^3(\widetilde{v}, \widetilde{t}, \widetilde{\mu}) \neq \emptyset$.*

Furthermore, if $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$ is nonempty, then multiplying the second and fourth conditions in the relaxed toll set by $\widetilde{x}_{ij}^k$ and $\widetilde{t}_k$ yields the following:

$$\begin{aligned}(s_{ij}(\widetilde{v}) + \beta_{ij})\widetilde{x}_{ij}^k &\leq (\rho_i^k - \rho_j^k)\widetilde{x}_{ij}^k + \xi_{ij}^k \widetilde{x}_{ij}^k, \ \forall\ k \in \mathcal{K}, (i,j) : \widetilde{x}_{ij}^k > 0, \\ (E_k^T \rho^k)\widetilde{t}_k &\leq w_k(\widetilde{t}_k)\widetilde{t}_k + \mu_k \widetilde{t}_k, \qquad \forall\ k : \widetilde{t}^k > 0.\end{aligned}$$

Because $\widetilde{v} = \sum_k \widetilde{x}^k$ and $A\widetilde{x}^k = E_k \widetilde{t}_k$, the above equations imply that $(s(\widetilde{v}) + \beta)^T \widetilde{v} \leq w(\widetilde{t})^T \widetilde{t} + \epsilon$, where $\epsilon = \sum_k \sum_{(i,j)\in\mathcal{A}} \widetilde{x}_{ij}^k \xi_{ij}^k + \sum_k \widetilde{t}_k \mu_k$. Thus, if $\beta \in \Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu)$, β must be in $\mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$ as well, i.e., $\Pi^1(\widetilde{v}, \widetilde{t}, \xi, \mu) \subseteq \mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$. Similarly, $\Pi^2(\widetilde{v}, \widetilde{t}, \xi) \subseteq \mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$ and $\Pi^3(\widetilde{v}, \widetilde{t}, \mu) \subseteq \mathcal{T}(\widetilde{v}, \widetilde{t}, \epsilon)$ when ϵ is chosen in a similar manner.

Dynamic Pricing: A Learning Approach

Dimitris Bertsimas[1] and Georgia Perakis[2]

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A., dbertsim@mit.edu

[2] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A., georgiap@mit.edu

Summary. We present an optimization approach for jointly learning the demand as a function of price, and dynamically setting prices of products in order to maximize expected revenue. The models we consider do not assume that the demand as a function of price is known in advance, but rather assume parametric families of demand functions that are learned over time. In the first part of the paper, we consider the noncompetitive case and present dynamic programming algorithms of increasing computational intensity with incomplete state information for jointly estimating the demand and setting prices as time evolves. Our computational results suggest that dynamic programming based methods outperform myopic policies often significantly. In the second part of the paper, we consider a competitive oligopolistic environment. We introduce a more sophisticated model of demand learning, in which the price elasticities are slowly varying functions of time, and allows for increased flexibility in the modeling of the demand. We propose methods based on optimization for jointly estimating the Firm's own demand, its competitors' demands, and setting prices. In preliminary computational work, we found that optimization based pricing methods offer increased expected revenue for a firm independently of the policy the competitor firm is following.

Key words: Dynamic Pricing, Learning, Dynamic Programming, MPEC

1 Introduction

In this paper we study pricing mechanisms for firms competing for the same products in a dynamic environment. Pricing theory has been extensively studied by researchers from a variety of fields over the years. These fields include, among others, economics (see for example, [Wo93]), marketing (see for example, [LKM92]), revenue management (see for example, [MV99]) and telecommunications (see for example, [Ke94], [KMT98], [PT98], [VD99], [Va99]). In recent years, the rapid development of information technology, the Internet and E-commerce has had very strong influence on the development of pricing and revenue management.

Mathematical and Computational Models for Congestion Charging, pp. 45-80

The overall goal of this paper is to address the problem of setting prices for a firm in both noncompetitive and competitive environments, in which the demand as a function of price is not known, but is learned over time. A firm produces a number of products which require (and compete for in the competitive case) scarce resources. The products must be priced dynamically over a finite time horizon, and sold to the appropriate demand. Our research (contrasted with traditional revenue management) considers pricing decisions, and takes capacity as given.

Problem Characteristics

The pricing problem we will focus on in this paper has a number of characteristics:

a) The demand as a function of price is *unknown* a priori and is learned over time. As a result, part of the model we develop in this paper deals with learning the demand as the firm acquires more information over time. That is, we exploit the fact that over time firms are able to acquire knowledge regarding demand behavior that can be utilized to improve profitability. Much of the current research does not consider this aspect but rather considers demand to be an exogenous stochastic process following a certain distribution. See [BM97], [CSS00], [FH97], [F95], [GvR94], [GvR97], [Gi00], [PT98]. Assuming that the demand follows an exogenous distribution that is known in advance, is often too strong an assumption. It is often unrealistic to consider that demand can be known in advance accurately. Rather, in this paper we use a learning approach. The advantage of this approach is that it is data driven. That is, it uses the data acquired so far, in order to estimate what the true demand is. As a firm acquires more data it keeps re-evaluating the true demand and hence gets a better estimate of it. As a result, this approach is practical and does not rely on assumptions that are often unrealistic.

b) Products are priced *dynamically* over a finite time horizon. This is an important aspect since the demand and the data of the problem evolve dynamically. There exists a great deal of research that does not consider the dynamic and the competitive aspects of the pricing problem jointly. An exception to this involves some work that applies differential game theory (see [Bag84], [Bas86], [DJ88]).

c) We explicitly allow competition in an *oligopolistic* market, that is, a market characterized by a few firms on the supply side, and a large number of buyers on the demand side. A key feature of such a market (in contrast to a monopoly) is that the profit one firm receives depends not just on the prices it sets, but also on the prices set by the competing firms. That is, there is no perfect competition in an oligopolistic market since decisions made by all the firms in the market impact the profits received by each firm. One can consider a cooperative oligopoly (where firms collude) or a noncooperative oligopoly. In this paper we focus on the latter. The theory of oligopoly dates back to the work of [Fr77], [Fr82], [Fr83].

d) We consider products that are *perishable*, that is, there is a finite horizon to sell the products, after which any unused capacity is lost. Moreover, *the marginal cost* of an extra unit of demand is relatively *small*. For this reason, our models in this paper ignore the cost component in the decision-making process and refer to revenue maximization rather than profit maximization.

Literature Review

Several models have been proposed for monopolistic versions of this problem. McGill and Van Ryzin [MV99], Weatherford and Bodily [WB92] as well as Williamson [Wi92] and the references therein provide a thorough review of revenue management models. More recently, Bitran and Caldentey [BC02] provide an overview of pricing models for the monopolistic version of the revenue management problem in which a perishable and non-renewable set of resources satisfy stochastic price-sensitive demand processes over a finite period of time. They survey results for deterministic as well as non-deterministic, single as well as multi-product, and static as well as dynamic pricing cases. Elmaghraby and Keskinocak [EK] review the literature and current practices in dynamic pricing in industries where capacity or inventory is fixed in the short run and perishable. They classify monopolistic models on the basis of whether inventory can be replenished or not, whether demand is dependent over time or not, and whether customers are myopic or strategic optimizers. Yano and Gilbert [YS04] review models for joint pricing and production under a monopolistic setup.

On the competitive side, Vives [Vi99] discusses the development of oligopoly pricing models. A survey by Chan et al. [CSSS01] summarize research on joint pricing, inventory control and production decisions in a supply chain. Furthermore, they survey literature on price and quantity competition in supply chain settings. Cachon and Netessine [CN] also survey the problem of competition from a supply chain perspective where the problem is characteristically a periodic production-review model. They discuss both non-cooperative and cooperative games in static and dynamic settings.

Pricing models in traditional revenue management research can be classified into two broad categories: static and dynamic. Static pricing models are based on aggregated demand distributions and can be seen as a special case of the multi-product newsvendor problem with fixed production costs and perishable product with no salvage. The extension of the newsvendor problem with price as a decision variable was studied by Zabel [Za70], Young [Yo78], Dada and Petruzzi [DP01], etc. Other relevant research includes [Za72], [Th74], [DP99] and [FH97] who study the single-product, multi-period combined pricing and inventory control problem that is typically solved by dynamic programming.

Dynamic pricing models represent demand as a controllable stochastic point process with price dependent intensity. Gallego and van Ryzin [GvR94] and Zhao and Zheng [ZZ00] consider the problem of optimally pricing a given inventory of a single product over a finite planning period before it perishes or is sold at salvage value. There is no reordering. Gallego and van Ryzin [GvR97] and Paschalidis and Tsitsiklis [PT98] extend this type of model to the dynamic pricing of multiple products whose production draws from a shared supply of resources. Kleywegt [Kl01] gives an optimal control formulation of the multiperiod dynamic pricing problem. Kachani and Perakis [KP02] propose a deterministic fluid model for dynamic pricing and inventory management for non-perishable products in capacitated and competitive make-to-stock manufacturing systems.

Some recent work explicitly considers the presence of competition within the pricing framework. Dockner and Jorgensen [DJ88] provide a treatment of the optimal pricing strategies for oligopolistic markets from a marketing perspective but not a computational perspective. Federgruen and Heching [BF99] develops a stochastic general equilibrium inventory model for supply chains in an oligopoly environment where the policies involve prices, service level targets and inventory control with linear models of demand.

There are two different classes of dynamic pricing models in the literature. The first one, is a periodic production-review model suitable for supply chain problems. In this model each firm starts with a given level of inventory/capacity at the beginning of the time horizon. At each period the firm sets his/her price level and realizes a certain demand that is a function of all price levels. A decision regarding additional production is also made at every period after reviewing inventory/capacity levels. Production costs, inventory holding costs, and cost of back orders are part of such models. On the other hand, in the second class of models (such as the model we introduce in this paper), the firm does not have the option to produce additional inventory/capacity between periods but rather the total capacity of the product the firm has available for sale is a given in the beginning of the time horizon. As a result, the product for sale is perishable. For example, such a model is suited for airlines that are selling seats on a particular flight, or hotels selling advance room reservations for a particular day or weekend. It is difficult, if not impossible, to increase the capacity of an aircraft or a hotel at short notice and requires considerable expense, advance notice and planning. For the purposes of the pricing process, the capacity (or in general the inventory of the product) can be assumed to be fixed. Note that for these problems, there are no holding or backorder costs. There are no holding costs since there is no tangible product that the seller has to hold on to from period to period if unsold. There are no backorder costs since the seller can sell only if she has the product in inventory and loses the sale otherwise. Note that this case is not a trivial extension of the periodic production review model.

1.1 Application Areas

There are many markets where the framework we consider in this paper applies. Examples include airline ticket pricing, as well as toll pricing in a transportation network. In this market the products that the consumers are demanding, are represented by the origin-destination (O-D) pairs during a particular time window. In the airline application, the resources are the flight legs (more appropriately seats on a particular flight leg) which have limited capacity. There is a finite horizon to sell the products, after which any unused capacity is lost (perishable products). The airlines compete with one another for the product demand which is of stochastic nature. Another application is toll pricing in a transportation network (see for example, [SGK01], [SKSK] and [Su02]). As in the case of airlines, the firm (e.g. the transportation authorities) are seeking to set prices (represented in this case by tolls) on the network's roads (for example, highways). The capacity for each road in this case is represented by the maximum number of cars that can be on the road without degrading the travel time to an undesirable level. The tolls can be updated dynamically as traffic conditions (represented for example through the demand) change. It is important for the authorities to get a good estimate of the demand in order to set accurately the corresponding tolls. This update can be done for example, every 30 minutes, using as information the past toll prices as well as the number of cars (demand) that entered the road (for example, the highway) during the previous 30 minute periods. The transportation authorities seek to maximize toll revenue since some highways are privately financed. This is the case for example in the SR 91 Express Lane in Orange County California (see [Su02]). Other industries sharing the same features include the service industry (for example, hotels, car rentals, and cruise-lines), the retail industry (for example, department stores) and finally, pricing in an e-commerce environment. All these industries attempt to intelligently match capacity with demand via *revenue management.*

1.2 Contributions

a) In the first part of the paper, we develop pricing mechanisms when there is incomplete demand information, by jointly setting prices and learning the firm's demand without assuming any knowledge of it in advance.

b) In the second part of the paper, we introduce a model of demand learning, in which there is competition but also price elasticities are slowly varying functions of time. This model allows for increased flexibility in the modeling of the demand in the presence of both uncertainty and competition. We propose methods based on optimization for jointly estimating the Firm's own demand, its competitors' demands, and setting prices.

1.3 Structure

The remainder of this paper is organized as follows. In Section 2, we focus on the dynamic pricing problem in a non-competitive environment. We consider jointly the problem of demand estimation and pricing using ideas from dynamic programming with incomplete state information. We present an exact algorithm as well as several heuristic algorithms that are easy to implement and discuss the various resulting pricing policies. In Section 3, we extend our previous model to also incorporate the aspect of competition. We propose an optimization approach to perform the firm's own demand estimation, its competitors' prices prediction and finally, its own price setting. Finally, in Section 4, we conclude with conclusions and open questions.

2 A Learning Approach for Dynamic Pricing, Part I: Without Competition

In this section, we consider a dynamic pricing problem for a perishable product in a non-competitive environment. We focus on a market with a single product and a single firm with overall capacity c available for sale over the time horizon T. In the beginning of each period t, the firm knows the previous price and demand realizations, that is, $d_1, \ldots, d_{t-1}$ and $p_1, \ldots, p_{t-1}$ (and as a result, the leftover capacity $c_t = c - \sum_{s=1}^{t-1} d_s$). This is the data available to the firm. In this section, we assume that the firm's true demand is an unknown linear function of the form

$$d_t = \beta^0 + \beta^1 p_t + \epsilon_t,$$

that is, it depends on the current period prices p_t, unknown parameters β^0, β^1 and a random noise $\epsilon_t \sim N(0, \sigma^2)$. Notice that in this model we assume that the parameters of the demand as a function of the price are not time dependent. In part II of this paper, due to the presence of competition we consider a more general demand model where the true parameters also vary with time. The firm's objectives are to estimate its demand dynamically and set prices in order to maximize its total expected revenue. Let $\mathcal{P} = [p_{\min}, p_{\max}]$ be the set of feasible prices. We assume that $\mathcal{P}$ is selected in such a way that the demand $d_i > 0$, $i = 1, \ldots, T$ with probability 1.

This part of the paper is organized as follows. In Section 2.1, we present a demand estimation model. In Section 2.2, we consider the joint demand estimation and pricing problem through a dynamic programming formulation. Using ideas from dynamic programming with incomplete state information, we are able to reduce this dynamic programming formulation to an eight-dimensional one. Nevertheless, this formulation is still difficult to solve, and we propose an approximation that allows us to further reduce the problem to a five dimensional dynamic program. In Section 2.3, we separate the demand estimation from the pricing problem and consider several heuristic algorithms.

In particular, we consider a one-dimensional dynamic programming heuristic as well as a myopic policy heuristic. To gain intuition, we find closed form solutions in the deterministic case. Finally, in Section 2.4, we consider some examples and offer insights.

2.1 Demand Estimation

As we mentioned at time t the firm has observed the previous price and demand realizations, that is, $d_1, \dots, d_{t-1}$ and $p_1, \dots, p_{t-1}$ and assumes a linear demand model $d_t = \beta^0 + \beta^1 p_t + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2)$. The parameters β^0, β^1 and σ are unknown and are estimated as follows.
We denote by $\mathbf{x}_s = [1,\ p_s]'$ and by $\widehat{\beta}_s$ the vector of the parameter estimates at time s, $(\widehat{\beta}^0_s, \widehat{\beta}^1_s)$. We estimate this vector of the demand parameters through the solution of the least square problem,

$$\widehat{\beta}_t = \arg\min_{\mathbf{r}\in\Re^2} \sum_{s=1}^{t-1} (d_s - \mathbf{x}'_s \mathbf{r})^2, \qquad t = 3, \dots, T. \tag{1}$$

Given $d_1, \dots, d_{t-1}$ and $p_1, \dots, p_{t-1}$, the least squares estimates are

$$\widehat{\beta}^1_t = \frac{(t-1)\sum_{s=1}^{t-1} p_s d_s - \sum_{s=1}^{t-1} p_s \sum_{s=1}^{t-1} d_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2}, \qquad \widehat{\beta}^0_t = \frac{\sum_{s=1}^{t-1} d_s}{t-1} - \widehat{\beta}^1_t \frac{\sum_{s=1}^{t-1} p_s}{t-1}.$$

The next proposition gives a recursive formula for these estimates that will be useful in the remainder of the paper.

Proposition 1. *The least squares estimates (1) can be generated by the following iterative process*

$$\widehat{\beta}_t = \widehat{\beta}_{t-1} + \mathbf{H}^{-1}_{t-1} \mathbf{x}_{t-1} \left(d_{t-1} - \mathbf{x}'_{t-1} \widehat{\beta}_{t-1} \right), \qquad t = 3, \dots, T$$

where $\widehat{\beta}_2$ is an arbitrary vector, and the matrices H_{t-1} are generated by

$$\mathbf{H}_{t-1} = \mathbf{H}_{t-2} + \mathbf{x}_{t-1}\mathbf{x}'_{t-1}, \qquad t = 3, \dots, T,$$

with $\mathbf{H}_1 = \begin{bmatrix} 1 & p_1 \\ p_1 & p_1^2 \end{bmatrix}$. *Therefore,* $\mathbf{H}_{t-1} = \begin{bmatrix} t-1 & \sum_{s=1}^{t-1} p_s \\ \sum_{s=1}^{t-1} p_s & \sum_{s=1}^{t-1} p_s^2 \end{bmatrix}$.

Although a proof can be easily derived using standard results, see for example [Ri95], for the sake of completeness we provide a proof in the appendix.

Notice that the matrix $\mathbf{H}_{t-1}$ is singular, and hence not invertible, when

$$(t-1)\sum_{s=1}^{t-1} p_s^2 = \left(\sum_{s=1}^{t-1} p_s\right)^2 . \qquad (2)$$

Notice that the only solution to the above equality is $p_1 = p_2 = \cdots = p_{t-1}$. If the matrix $\mathbf{H}_{t-1}$ is nonsingular, then the inverse is

$$\mathbf{H}_{t-1}^{-1} = \begin{bmatrix} \dfrac{\sum_{s=1}^{t-1} p_s^2}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} & \dfrac{-\sum_{s=1}^{t-1} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \\ \dfrac{-\sum_{s=1}^{t-1} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} & \dfrac{t-1}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \end{bmatrix}.$$

Therefore,

$$\mathbf{H}_{t-1}^{-1}\mathbf{x}_{t-1} = \begin{bmatrix} \dfrac{\sum_{s=1}^{t-1} p_s^2}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} & \dfrac{-\sum_{s=1}^{t-1} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \\ \dfrac{-\sum_{s=1}^{t-1} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} & \dfrac{t-1}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \end{bmatrix} \begin{bmatrix} 1 \\ p_{t-1} \end{bmatrix} =$$

$$\begin{bmatrix} \dfrac{\sum_{s=1}^{t-2} p_s^2 - p_{t-1}\sum_{s=1}^{t-2} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \\ \dfrac{(t-2)p_{t-1} - \sum_{s=1}^{t-2} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \end{bmatrix}.$$

As a result, we can express the estimates of the demand parameters in period t in terms of earlier estimates as

$$\begin{bmatrix} \widehat{\beta}_t^0 \\ \widehat{\beta}_t^1 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_{t-1}^0 \\ \widehat{\beta}_{t-1}^1 \end{bmatrix} + (d_{t-1} - \widehat{\beta}_{t-1}^0 - \widehat{\beta}_{t-1}^1 p_{t-1}) \begin{bmatrix} \dfrac{\sum_{s=1}^{t-2} p_s^2 - p_{t-1}\sum_{s=1}^{t-2} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \\ \dfrac{(t-2)p_{t-1} - \sum_{s=1}^{t-2} p_s}{(t-1)\sum_{s=1}^{t-1} p_s^2 - \left(\sum_{s=1}^{t-1} p_s\right)^2} \end{bmatrix}.$$

The estimate for the variance σ^2 at time t is given by

$$\widehat{\sigma}_t^2 = \sum_{\tau=1}^{t-1} \frac{\left(d_\tau - \widehat{\beta}_t^0 - \widehat{\beta}_t^1 p_\tau\right)^2}{t-3}.$$

Notice that the variance estimate is based on $t-1$ pieces of data, with two parameters already estimated from the data, hence there are $t-3$ degrees of freedom. Such an estimate is unbiased (see [Ri95]).

2.2 An Eight-Dimensional DP for Determining Pricing Policies

The difficulty in coming up with a general framework for dynamically determining prices is that the parameters β^0 and β^1 of the true demand are not directly observable. What is observable though are the realizations of demand and price in the previous periods, that is, $d_1, \ldots, d_{t-1}$ and $p_1, \ldots, p_{t-1}$. This seems to suggest that ideas from dynamic programming with incomplete state information may be useful (see [Be95]). As a first step in this direction, during the current period t, we consider a dynamic program with state space $(d_1, \ldots, d_{t-1}, p_1, \ldots, p_{t-1}, c_t)$, control variable the current price p_t and randomness coming from the noise ϵ_t. We observe though that as time t increases, the dimension of the state space becomes huge and therefore, solving this dynamic programming formulation is not possible. In what follows we will illustrate that we can considerably reduce the high dimensionality of the state space.

First we introduce the notation, $\widehat{\beta}_{s,t} = \left(\widehat{\beta}_{s,t}^0, \widehat{\beta}_{s,t}^1\right)$, $s = t, \ldots, T$, which is the current time t estimate of the parameters for future times $s = t, \ldots, T$. Notice that $\widehat{\beta}_{t,t} = \widehat{\beta}_t$. Similarly to Proposition 1, we can update our least squares estimates through $\widehat{\beta}_{t+1,t} = \widehat{\beta}_{t,t} + \mathbf{H}_t^{-1}\mathbf{x}_t \left(\widehat{D}_t - \mathbf{x}_t' \widehat{\beta}_{t,t}\right)$. Notice that since in the beginning of period t demand d_t is not known, we replaced it with $\widehat{D}_t = \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_t + \varepsilon_t$. As a result, vector $\widehat{\beta}_{t+1,t}$ is a random variable. A useful observation we need to make is that in order to calculate matrix $\mathbf{H}_t$ we need to keep track of the quantities $\sum_{\tau=1}^{t-1} p_\tau^2$ and $\sum_{\tau=1}^{t-1} p_\tau$. These will be as a result part of the state space in the new dynamic programming formulation.

It is natural to assume that the variance estimates change with time and do not remain constant in future periods. This is the case since the estimate of the variance will be affected by the prices. That is,

$$\varepsilon_{s,t} \sim N\left(0, \widehat{\sigma}_{s,t}^2\right)$$

$$\widehat{\sigma}_{s,t}^2 = \frac{\sum_{\tau=1}^{s-1} \left(d_\tau - \widehat{\beta}_{s,t}^0 - \widehat{\beta}_{s,t}^1 p_\tau\right)^2}{s-3}, \quad s = t, \ldots, T.$$

This observation implies that we need to find a way to estimate the variance for the future periods from the current one. We denote by $\widehat{\sigma}_{t+1,t}^2$ the estimate (in the current period, t) of next period's variance.

Proposition 2. *The estimate of next period's variance in the current period t is given by,*

$$
\begin{aligned}
&\widehat{\sigma}^2_{s+1,\,t} = \\
&\frac{\widehat{\sigma}^2_{s,t}(s-3) + 2\widehat{\beta}^0_{s,t}\sum_{i=1}^{s-1} d_i + 2\widehat{\beta}^1_{s,t}\sum_{i=1}^{s-1} d_i p_i - (s-1)\left(\widehat{\beta}^0_{s,t}\right)^2 - 2\widehat{\beta}^0_{s,t}\widehat{\beta}^1_{s,t}\sum_{i=1}^{s-1} p_i}{s-2} \\
&+ \frac{-\left(\widehat{\beta}^1_{s,t}\right)^2\sum_{i=1}^{s-1} p_i^2 + \left(\widehat{\beta}^0_{s,t}\right)^2 + \left(\widehat{\beta}^1_{s,t}p_s\right)^2 + \widehat{\sigma}^2_{s,t} + 2\widehat{\beta}^0_{s,t}\widehat{\beta}^1_{s,t}p_t - 2\widehat{\beta}^0_{s+1,t}\sum_{i=1}^{s-1} d_i}{s-2} \\
&+ \frac{-2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^0_{s,t} - 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s,t}p_s - 2\widehat{\beta}^1_{s+1,t}\sum_{i=1}^{s-1} p_i d_i}{s-2} \\
&+ \frac{-2\widehat{\beta}^1_{s+1,t}\widehat{\beta}^0_{s,t}p_s - 2\widehat{\beta}^1_{s+1,t}\widehat{\beta}^1_{s,t}p_s^2 + s\left(\widehat{\beta}^0_{s+1,t}\right)^2}{s-2} \\
&+ \frac{2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t}\sum_{i=1}^{s-1} p_i + 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t}p_s + \left(\widehat{\beta}^1_{s+1,t}\right)^2\sum_{i=1}^{s-1} p_i^2 + \left(\widehat{\beta}^1_{s+1,t}\right)^2 p_s^2}{s-2} \qquad (3)
\end{aligned}
$$

Although a proof of this proposition easily follows using standard results (see for example, [Ri95]), for the sake of completeness we also provide a proof in the Appendix.

This proposition suggests that in order to estimate in period s, the next period $s+1$ variance from the variance in period s, we need to keep track of the following quantities

$$\widehat{\beta}^0_{s,t},\ \widehat{\beta}^1_{s,t},\ \sum_{\tau=1}^{s-1} p_\tau^2,\ \sum_{\tau=1}^{s-1} p_\tau,\ \sum_{\tau=1}^{s-1} p_\tau d_\tau,\ \sum_{\tau=1}^{s-1} d_\tau,\ \widehat{\sigma}^2_{s,t}.$$

This observation allows us to provide an eight-dimensional dynamic programming formulation with state space given by,

$$\left(c_s,\ \widehat{\beta}^0_{s,t},\ \widehat{\beta}^1_{s,t},\ \sum_{\tau=1}^{s-1} p_\tau^2,\ \sum_{\tau=1}^{s-1} p_\tau,\ \sum_{\tau=1}^{s-1} p_\tau d_\tau,\ \sum_{\tau=1}^{s-1} d_\tau,\ \widehat{\sigma}^2_{s,t}\right), \qquad s = t, \ldots, T.$$

We are now able to formulate the following dynamic program where the control is the price and the randomness is the noise. The idea behind this dynamic programming formulation is that we set the prices by optimizing the revenue over the time horizon and at the same time learning the parameters of the demand by appropriately updating them from the previous period estimates.

An Eight-Dimensional DP Pricing Policy

$$J_T(c_T,\ \widehat{\beta}^0_{T,t},\ \widehat{\beta}^1_{T,t},\ \widehat{\sigma}^2_{T,t}) = \max_{p_T} E_{\varepsilon_{T,t}}\left[p_T \min\left\{\left(\widehat{\beta}^0_{T,t} + \widehat{\beta}^1_{T,t}p_T + \varepsilon_{T,t}\right),\ c_T\right\}\right]$$

$for\ s = \max(4,t),\ldots,T-1:$

$$J_s(c_s,\ \widehat{\beta}^0_{s,t},\ \widehat{\beta}^1_{s,t},\ \sum_{\tau=1}^{s-1} p_\tau^2,\ \sum_{\tau=1}^{s-1} p_\tau,\ \sum_{\tau=1}^{s-1} p_\tau d_\tau,\ \sum_{\tau=1}^{s-1} d_\tau,\ \widehat{\sigma}^2_{s,t})$$

$$= \max_{p_s} E_{\varepsilon_{s,t}}\left[p_s \min\left\{\left(\widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t}p_s + \varepsilon_{s,t}\right),\ c_s\right\}\right.$$

$$\left. + J_{s+1}\left(\begin{array}{c} c_s - \min\left\{\left(\widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t}p_s + \varepsilon_{s,t}\right),\ c_s\right\}, \\ \widehat{\beta}^0_{s+1,t},\ \widehat{\beta}^1_{s+1,t}, \\ \sum_{\tau=1}^{s-1} p_\tau^2 + p_s^2,\ \sum_{\tau=1}^{s-1} p_\tau + p_s, \\ \sum_{\tau=1}^{s-1} p_\tau d_\tau + p_s\left(\widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t}p_s + \varepsilon_{s,t}\right), \\ \sum_{\tau=1}^{s-1} d_\tau + \left(\widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t}p_s + \varepsilon_{s,t}\right), \\ \widehat{\sigma}^2_{s+1,t} \end{array}\right)\right], \tag{4}$$

where

$$\begin{bmatrix} \widehat{\beta}^0_{s+1,t} \\ \widehat{\beta}^1_{s+1,t} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}^0_{s,t} \\ \widehat{\beta}^1_{s,t} \end{bmatrix} + \varepsilon_{s,t} \begin{bmatrix} \dfrac{\sum_{\tau=1}^{s-1} p_\tau^2 - p_s \sum_{\tau=1}^{s-1} p_\tau}{s\sum_{\tau=1}^{s-1} p_\tau^2 + sp_s^2 - \left(\sum_{\tau=1}^{s-1} p_\tau + p_s\right)^2} \\ \dfrac{(s-1)p_s - \sum_{\tau=1}^{s-1} p_\tau}{s\sum_{\tau=1}^{s-1} p_\tau^2 + sp_s^2 - \left(\sum_{\tau=1}^{s-1} p_\tau + p_s\right)^2} \end{bmatrix},$$

with noise $\varepsilon_{s,t} \sim N(0, \widehat{\sigma}^2_{s,t})$ and variance $\widehat{\sigma}^2_{s,t}$ given from the recursive formula in (3). Eq. (4) represents the Bellman equation for the eight dimensional DP. Specifically, the term $p_s \min\left\{\left(\widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t}p_s + \varepsilon_{s,t}\right),\ c_s\right\}$ represents the revenue at period s, and the arguments of $J_{s+1}(\cdot)$ the evolution of the eight dimensional state.

Notice that in the DP recursion $s = \max(4,t),\ldots,T$, because we need at least three data points in order to estimate three parameters.

2.3 A Five-Dimensional DP for Determining Pricing Policies

Although the previous DP formulation is the correct framework for determining pricing policies, it has an eight-dimensional state space which makes the problem computationally intractable. For this reason we consider in this section an approximation that gives rise to a lower dimensional dynamic program that is computationally tractable. In particular, we relax the assumption that the noise at time t changes in time and is affected by future pricing decisions. In particular, we consider

$$\varepsilon_s \sim N\left(0,\ \widehat{\sigma}_t^2\right), \qquad s = t, \ldots, T$$
$$\widehat{\sigma}_t^2 = \frac{\sum\limits_{\tau=1}^{t-1}\left(d_\tau - \widehat{\beta}_t^0 - \widehat{\beta}_t^1 p_\tau\right)^2}{t-3}.$$

Specifically, we assume (as an approximation) that the estimate of the variance of the noise only depends on the current time t, and does not change with future times s.

Moreover, as in the previous section

$$\widehat{\beta}_{t+1,t} = \widehat{\beta}_{t,t} + \mathbf{H}_t^{-1}\mathbf{x}_t\left(\widehat{d}_t - \mathbf{x}_t'\widehat{\beta}_{t,t}\right).$$

To calculate the matrix $\mathbf{H}_t$ we need to keep track of the quantities $\sum\limits_{\tau=1}^{t-1} p_\tau^2$ and $\sum\limits_{\tau=1}^{t-1} p_\tau$.

This gives rise to a dynamic programming formulation with state variables,

$$\left(c_s,\ \widehat{\beta}_{s,t}^0,\ \widehat{\beta}_{s,t}^1,\ \sum_{\tau=1}^{s-1} p_\tau^2,\ \sum_{\tau=1}^{s-1} p_\tau\right) \qquad s = t, \ldots, T. \tag{5}$$

A Five-Dimensional DP Pricing Policy

$$J_T\left(c_T,\ \widehat{\beta}_{T,t}^0,\ \widehat{\beta}_{T,t}^1\right) = \max_{p_T \in \mathcal{P}} \mathrm{E}_{\varepsilon_T}\left[p_T \min\left\{\left(\widehat{\beta}_{T,t}^0 + \widehat{\beta}_{T,t}^1 p_T + \varepsilon_T\right),\ c_T\right\}\right]$$

for $s = \max(4, t), \ldots, T-1$:

$$J_s\left(c_s,\ \widehat{\beta}_{s,t}^0,\ \widehat{\beta}_{s,t}^1,\ \sum_{\tau=1}^{s-1} p_\tau^2,\ \sum_{\tau=1}^{s-1} p_\tau\right) = \max_{p_s \in \mathcal{P}} \mathrm{E}_{\varepsilon_s}[p_s \min\left\{\left(\widehat{\beta}_{s,t}^0 + \widehat{\beta}_{s,t}^1 p_s + \varepsilon_s\right),\right.$$
$$\left. c_s\right\} + J_{s+1}\left(\begin{array}{c} c_s - \min\left\{\left(\widehat{\beta}_{s,t}^0 + \widehat{\beta}_{s,t}^1 p_s + \varepsilon_s\right),\ c_s\right\}, \\ \widehat{\beta}_{s+1,t}^0,\ \widehat{\beta}_{s+1,t}^1, \\ \sum\limits_{\tau=1}^{s-1} p_\tau^2 + p_s^2,\ \sum\limits_{\tau=1}^{s-1} p_\tau + p_s \end{array}\right)\Bigg],$$

with

$$\begin{bmatrix} \widehat{\beta}^0_{s+1,t} \\ \widehat{\beta}^1_{s+1,t} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}^0_{s,t} \\ \widehat{\beta}^1_{s,t} \end{bmatrix} + \varepsilon_s \begin{bmatrix} \dfrac{\sum_{\tau=1}^{s-1} p_\tau^2 - p_s \sum_{\tau=1}^{s-1} p_\tau}{s \sum_{\tau=1}^{s-1} p_\tau^2 + s p_s^2 - \left(\sum_{\tau=1}^{s-1} p_\tau + p_s \right)^2} \\ \dfrac{(s-1) p_s - \sum_{\tau=1}^{s-1} p_\tau}{s \sum_{\tau=1}^{s-1} p_\tau^2 + s p_s^2 - \left(\sum_{\tau=1}^{s-1} p_\tau + p_s \right)^2} \end{bmatrix}.$$

Although this latter approach is more tractable it is still fairly complex to solve. To make the computations tractable we discretize the values of the parameters. See Subsection 2.5 for some preliminary numerical examples.

2.4 Pricing Heuristics

In the previous two subsections, we considered two dynamic programming formulations for determining pricing policies. The first was an exact formulation with an eight-dimensional state space that was computationally intractable, while the second was an approximation with a five-dimensional state space that is more tractable. Nevertheless, although this latter approach is tractable it is still fairly complex to solve. Both of these formulations were based on the idea of performing jointly the demand estimation with the pricing problem. In this section, we consider two heuristics that are approximations but yet are computationally very easy to perform. They are based on the idea of separating the demand estimation from the pricing problem.

One-Dimensional DP Pricing Policy

In the beginning of period t, the firm computes the estimates $\widehat{\beta}^0_t$ and $\widehat{\beta}^1_t$ and solves a one-dimensional dynamic program assuming that these parameter estimates are valid over all future periods. That is, this heuristic approach ignores the fact that these estimates will in fact be affected by the current pricing decisions. In particular,

$$\begin{aligned} \widehat{d}_s &= \widehat{\beta}^0_t + \widehat{\beta}^1_t p_s + \varepsilon_s, \quad & s = t, \ldots, T \\ \varepsilon_s &\sim N\left(0,\ \widehat{\sigma}^2_t\right), \quad & s = t, \ldots, T, \end{aligned}$$

with

$$\widehat{\sigma}^2_t = \sum_{s=1}^{t-1} \frac{\left(d_s - \widehat{\beta}^0_t - \widehat{\beta}^1_t p_s\right)^2}{t-3}.$$

Subsequently, the firm solves the following dynamic program in the beginning of period t:

$$J_T(c_T) = \max_{p_T \in \mathcal{P}} E_{\varepsilon_T} \left[p_T \min \left\{ \left(\widehat{\beta}^0_t + \widehat{\beta}^1_t p_T + \varepsilon_T \right),\ c_T \right\} \right]$$

for $s = \max(4, t), \ldots, T-1$:

$$J_s(c_s) = \max_{p_s \in \mathcal{P}} E_{\varepsilon_s} \begin{bmatrix} p_s \min \left\{ \left(\widehat{\beta}^0_t + \widehat{\beta}^1_t p_s + \varepsilon_s \right),\ c_s \right\} + \\ J_{s+1} \left(c_s - \min \left\{ \left(\widehat{\beta}^0_t + \widehat{\beta}^1_t p_s + \varepsilon_s \right),\ c_s \right\} \right) \end{bmatrix}.$$

In this dynamic programming formulation the remaining capacity represents the state space, the prices are the controls and the randomness comes from the noise.

Deterministic One-Dimensional DP Policy

To gain some intuition, in what follows we examine the deterministic case (that is, when the noise $\varepsilon_s = 0$). As a first step we formulate the dynamic optimization problem as a strictly concave optimization problem. In period t and for the remaining of the time horizon after having computed the estimates $\widehat{\beta}_t^0$ and $\widehat{\beta}_t^1$, the firm sets prices by solving the following problem.

$$\begin{aligned} \max_{\mathbf{p},\mathbf{d}} \ & \sum_{s=t}^{T} d_s p_s \\ \text{s.t.} \ & d_s \leq \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s, \quad s = t, \ldots, T \\ & \sum_{s=t}^{T} d_s \leq c_t \\ & p_s \in \mathcal{P}, \quad s = t, \ldots, T. \end{aligned} \tag{6}$$

Since the parameter $\widehat{\beta}_t^1 < 0$, the demand is strictly decreasing with respect to the price. As a result, at the optimal solution $d_s = \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s$. Since if this was not the case and at the optimal solution $d_s < \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s$, by increasing the price to p_s^{new} so that $d_s = \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s^{new}$ and keeping d_s as before, we would get a higher revenue. This is a contradiction, (see also [PS03] for more details).

This observation allows us to reformulate the problem as a strictly concave optimization problem,

$$\begin{aligned} \max_{\mathbf{p}} \ & \sum_{s=t}^{T} (\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s) p_s \\ \text{s.t.} \ & \sum_{s=t}^{T} (\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s) \leq c_t \\ & p_s \in \mathcal{P}, \quad s = t, \ldots, T. \end{aligned} \tag{7}$$

After having computed the estimates $\widehat{\beta}_t^0$ and $\widehat{\beta}_t^1$, the firm solves the following DP in the beginning of period t ($t = 1, \ldots, T$),

$$\begin{aligned} J_T(c_T) &= \max_{p_T \in \mathcal{P}} p_T \min \left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_T \right), \ c_T \right\} \\ &\text{for } s = t, \ldots, T-1 : \\ J_s(c_s) &= \max_{p_s \in \mathcal{P}} p_s \min \left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s \right), \ c_s \right\} \\ & \quad + J_{s+1} \left(c_s - \min \left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s \right), \ c_s \right\} \right). \end{aligned}$$

This deterministic one-dimensional DP policy has a closed form solution. We establish its solution in two parts. Since the dynamic program is deterministic, an optimal solution is given by an open-loop policy (that is, we can solve for an optimal price path versus an optimal pricing policy, i.e. there is no dependence on the state). For the proofs that follow, we need to introduce the following definition.

Definition 1. *A price vector* $\mathbf{p} = (p_t, \ldots, p_T)'$ *leads to* **premature stock-out** *if*

$$\sum_{s=t}^{T} \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s \right) > c_t.$$

Lemma 1. *The optimal solution given by the one-dimensional DP is unique and satisfies* $p_t = \cdots = p_T$.

Proof. First we will show that any optimal solution must satisfy $p_t = \cdots = p_T$, then we will prove uniqueness. Suppose there exists an optimal solution $\mathbf{p}^*$ for which the above does not hold. Then at least two of the prices are different and at least one price is less than $p_{\max}$. Without loss of generality, assume that $p_t \neq p_{t+1}$ (the argument holds for any two prices). We will show that such a solution cannot be optimal. Next we will show that the optimal solution must satisfy,

$$\sum_{s=t}^{T} \widehat{d}_s = \sum_{s=t}^{T} \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s^* \right) \leq c_t.$$

This is true since otherwise we could increase at least one of the prices by a small amount (since at least one is strictly less than $p_{\max}$), and achieve greater revenue by selling the same number of units c_t at a slightly higher average price (contradicting the optimality of the solution). Therefore, the firm does not expect a premature stock-out and the optimal objective value is given by, $z^* = \sum_{s=t}^{T} p_s^* \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s^* \right)$. Notice that the revenue generated in periods t and $t+1$ is given by,

$$\begin{aligned} & p_t^* \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_t^* \right) + p_{t+1}^* \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_{t+1}^* \right) \\ & = \widehat{\beta}_t^0 p_t^* + \widehat{\beta}_t^0 p_{t+1}^* + \widehat{\beta}_t^1 \left((p_t^*)^2 + (p_{t+1}^*)^2 \right). \end{aligned} \tag{8}$$

In what follows, consider setting price $\frac{p_t^* + p_{t+1}^*}{2}$ in periods t and $t+1$. Therefore, the revenue generated in periods t and $t+1$ is given by,

$$\widehat{\beta}_t^0 p_t^* + \widehat{\beta}_t^0 p_{t+1}^* + \frac{\widehat{\beta}_t^1}{2} \left(p_t^* + p_{t+1}^* \right)^2. \tag{9}$$

Comparing (9) with (8) we notice that the total revenue has been increased. This is a contradiction. Hence, any optimal solution must satisfy $p_t = \cdots = p_T$.

Since problem (7) is strictly concave, the solution is unique. ■

We use this result to prove the following theorem.

Theorem 1. *Let*

$$p^* = \max\left\{ -\frac{\widehat{\beta}_t^0}{2\widehat{\beta}_t^1}, \frac{c_t - (T-t+1)\,\widehat{\beta}_t^0}{(T-t+1)\,\widehat{\beta}_t^1} \right\}.$$

Then, in the deterministic case, the one-dimensional DP has the following closed form solution for $s = t, \ldots, T$:

$$p_s^* = \begin{cases} p_{\max} & \textit{if } p^* \geq p_{\max} \\ p^* & \textit{if } p_{\min} < p^* p_{\max} \\ p_{\min} & \textit{if } p^* \leq p_{\min}. \end{cases}$$

Proof. From Lemma 1, the optimal solution satisfies $p_t = \ldots = p_T$. Thus, the capacity constraint in (7) simplifies to $(T-t+1)(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p) \leq c_t$. Thus, problem (7) can be reformulated as a single variable concave quadratic optimization problem, the result follows. ■

We note that in the deterministic case the policies given by the one and five-dimensional DPs are equivalent. This follows since in the deterministic case $\varepsilon_s = 0$ and as a result, the future demand parameter estimates are not affected by the current pricing decision. Hence, $(\widehat{\beta}_{s+1,t}^0, \widehat{\beta}_{s+1,t}^1)' = (\widehat{\beta}_{s,t}^0, \widehat{\beta}_{s,t}^1)'$. Therefore, the five-dimensional DP can be reduced to the following three dimensional DP,

$$\begin{aligned} J_T\left(c_T, \widehat{\beta}_T^0, \widehat{\beta}_T^1\right) &= \max_{p_T \in \mathcal{P}} p_T \min\left\{ \left(\widehat{\beta}_T^0 + \widehat{\beta}_T^1 p_T\right), c_T \right\} \\ &\text{for } s = t, \ldots, T-1: \\ J_s\left(c_s, \widehat{\beta}_s^0, \widehat{\beta}_s^1\right) &= \max_{p_s \in \mathcal{P}} p_s \min\left\{ \left(\widehat{\beta}_s^0 + \widehat{\beta}_s^1 p_s\right), c_s \right\} \\ &\quad + J_{s+1}\left(c_s - \min\left\{ \left(\widehat{\beta}_s^0 + \widehat{\beta}_s^1 p_s\right), c_s \right\}, \widehat{\beta}_s^0, \widehat{\beta}_s^1\right). \end{aligned}$$

Moreover, notice that the one-dimensional DP policy in the deterministic case is given by,

$$\begin{aligned} J_T(c_T) &= \max_{p_T \in \mathcal{P}} p_T \min\left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_T\right), c_T \right\} \\ &\text{for } s = t, \ldots, T-1: \\ J_s(c_s) &= \max_{p_s \in \mathcal{P}} p_s \min\left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s\right), c_s \right\} \\ &\quad + J_{s+1}\left(c_s - \min\left\{ \left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_s\right), c_s \right\}\right). \end{aligned}$$

When the firm uses the five-dimensional DP policy, since in the beginning of period t, $\left(\widehat{\beta}_{s,t}^0, \widehat{\beta}_{s,t}^1\right) = \left(\widehat{\beta}_t^0, \widehat{\beta}_t^1\right)$, for all $s = t, \ldots, T$, it follows, just like in the case of the one-dimensional DP policy, that the current parameter estimates are valid over all future periods. The DPs solved for both policies are in that case equivalent. The only difference is that the five-dimensional DP *explicitly* treats $\widehat{\beta}_t^0$ and $\widehat{\beta}_t^1$ as (constant) states while the one-dimensional DP *implicitly* treats $\widehat{\beta}_t^0$ and $\widehat{\beta}_t^1$ as (constant) states. This observation leads us to conclude that the two policies are equivalent.

The Myopic Pricing Policy

Finally, we introduce the last heuristic pricing policy, the myopic pricing policy. This policy maximizes the expected current period revenue over each period, without considering future implications of the pricing decisions. In period t

$$p_t \in \arg\max_{p\in\mathcal{P}} p\mathrm{E}_{\varepsilon_t}\left[\min\left\{\left(\widehat{\beta}_t^0 + \widehat{\beta}_t^1 p + \widehat{\varepsilon}_t\right),\ c_t\right\}\right].$$

Quantity c_t denotes the remaining capacity in the beginning of period t. Clearly the myopic policy is suboptimal since it does not take into account the number of periods left in the planning horizon. However, when capacity is sufficiently large the expected revenue obtained through the myopic and the one-dimensional DP policy become the same. This follows from the observation that when capacity is sufficiently large, both methods maximize current expected revenue. This myopic approach is optimal since the firm does not run the risk of stocking out before the end of the planning horizon that is, there are no future implications of the current pricing decision.

2.5 Computational Results

In the previous subsections, we introduced dynamic pricing policies for revenue maximization with incomplete demand information based on DP (one, five and eight dimensional) as well as a myopic policy which we consider as a benchmark. We have implemented all methods except the eight-dimensional DP, which is outside today's computational capabilities.

We consider an example where true demand is given by $d_t = 60 - p_t + \varepsilon_t$, with $\varepsilon_t = 0$ initially and $\varepsilon_t \sim N(0, \sigma^2)$, where $\sigma = 4$. The prices belong in the set $\mathcal{P} = \{20,\ 21, \ldots,\ 40\}$, the total capacity is $c = 400$ and the time horizon is $T = 20$. As we discussed in the previous subsections we consider a linear model for estimating the demand, that is, $\widehat{d_t} = \widehat{\beta}_t^0 + \widehat{\beta}_t^1 p_t$.

We first assume a model of demand assuming that $\varepsilon_t = 0$, and we apply both the myopic and the one-dimensional DP policies, which are optimal in this case. In order to show the effect of demand learning we plot in Figures 1 and 2 the least squares estimates of the intercept $\widehat{\beta}_t^0$ and the slope $\widehat{\beta}_t^1$. In particular, we plot the average estimate of the parameters within one standard deviation. We notice that the estimates of the demand parameters indeed tend to the true demand parameters over time.

In Table 1, we compare the total revenue and average price from the myopic and the one-dimensional DP policies, over 1,000 simulation runs. In general, as we mentioned earlier, for very large capacities both policies lead to

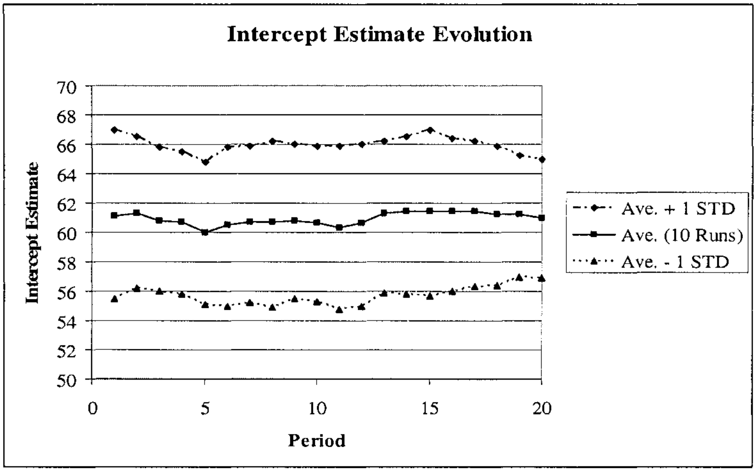


Fig. 1. The estimate $\widehat{\beta}_t^0$

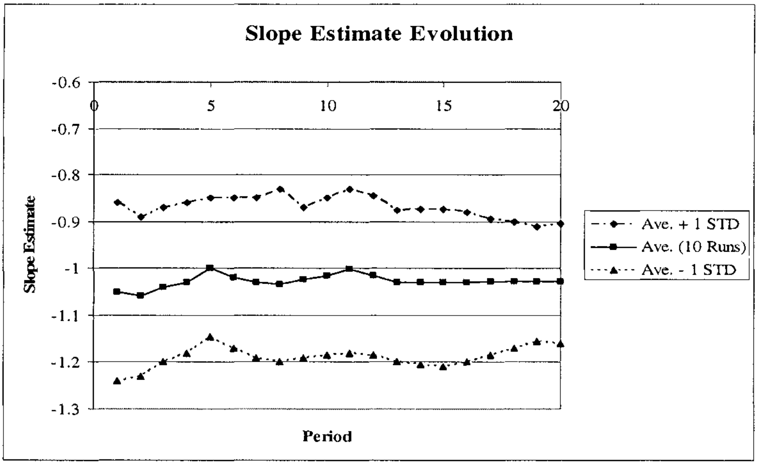


Fig. 2. The estimate $\widehat{\beta}_t^1$

the same revenue. The results of Table 1 suggest that the one-dimensional DP outperforms the myopic policy significantly (by 28.65%). Moreover, the standard deviation of the revenue given by the one-dimensional DP policy is 3.5 times lower than the revenue given by the myopic policy. In addition, the one-dimensional DP leads to a higher and more stable price compared to the price given by the myopic policy.

We next consider the case that $\varepsilon_t \sim N(0, 16)$. In Table 2, we report the total revenue and average price from the myopic, one-dimensional DP and five-dimensional DP policies, over 1,000 simulation runs. We consider $T = 8$ periods. For all policies, we compute average revenue over the periods $t =$

Table 1. Comparison of total revenue and average price for the myopic and the one-dimensional DP policies for $\varepsilon_t = 0$, over 1000 simulation runs with $T = 20$ and $c = 400$

$T = 20$, $c = 400$	**Myopic**	**1-dim. DP**
Ave (Total Revenue)	12,194	15,688
Std (Total Revenue)	1,162.9	303.6
Ave(Price)	30.93	39.36
Std (Price)	2.81	0.65

$4, \ldots, 8$, as in the first 3 periods we do not have enough observations to start the five dimensional DP. For the first three periods, we use $p = 30$.

Table 2. Comparison of total revenue and average price for the myopic, the one-dimensional and five-dimensional DP policies for $\varepsilon_t \sim N(0, 16)$, over 1000 simulation runs with $T = 8$ and $c = 125$

$T = 8$, $c = 125$	**Myopic**	**1-dim DP**	**5-Dim DP**
Ave.(Total Revenue)	3,884.6	4,250.1	4,339.3
Std (Total Revenue)	302.6	282.0	394.2
Ave.(Price)	32.5	35.7	36.7
Std (Price)	2.5	1.8	1.89

The results of Table 2 agree with intuition that the more computationally intensive methods lead to higher revenues. In particular, the one-dimensional DP policy outperforms the myopic policy (by 9.4%), and the five-dimensional DP policy outperforms the one-dimensional DP policy (by 2.09%). In this experiment, the variability of the revenue and the price was comparable among the three policies.

Overall, we feel that this example (as well as several others of similar nature) offers the following insights:

1. All the methods we considered succeed in estimating accurately the demand parameters over time.
2. The class of DP policies outperforms the myopic policy. In addition, revenue increases with higher complexity of the DP method, that is the five-dimensional DP policy outperforms the one-dimensional DP policy.

3 A Learning Approach for Dynamic Pricing, Part II: With Competition

In this section, we consider a dynamic pricing model in a competitive setting. In particular, we focus on an oligopolistic market where several firms compete for a single perishable product in a dynamic environment. As time progresses the firms competing in the market are learning their demand and setting prices over the leftover time horizon. As a result, the firm apart from trying to estimate its own demand, it also needs to predict its competitors' demands and pricing policies. Given the increased uncertainty due to competition, we use a more flexible model of demand, in which the firm considers that its own true demand as well as its competitors' demands have parameters that are time varying. Models of the type we consider in this section, were introduced in a more general context in [BGT99], and have nice asymptotic properties that we review shortly. Specifically, the J competing firms have total capacity c_1, c_2, ..., c_J respectively, over a finite time horizon T. At time t, firm k has leftover capacity of the product for sale $c_{k,t}$, $k = 1, \ldots, J$, for the remainder of the time horizon. In the beginning of each period t, Firm 1 knows the past realizations of its own demand $d_{1,s}$, its own prices $p_{1,s}$ as well as its competitors' prices $p_{k,s}$, where $k \in \{-1\} = \{2, \ldots, J\}$ and $s = 1, \ldots, t-1$. Notice that it is not realistic to assume that the firm directly observes its competitors' demands.

We assume that each firm's true demand is an unknown linear function, where the true demand parameters are time varying, that is, for firm $k = 1, \ldots, J$ demand is of the form

$$d_{k,t} = \beta^0_{k,t} + \sum_{l \in \{-k\}} \beta^l_{k,t} p_{l,t} + \beta^k_{k,t} p_{k,t} + \epsilon_{k,t}, \quad k = 1, 2, \ldots, J,$$

the coefficients $\beta^0_{k,t}, \beta^l_{k,t} \geq 0$, $l \in \{-k\} = \{1, \ldots, k-1, k+1, \ldots, J\}$, $\beta^k_{k,t} \leq 0$. The coefficients vary slowly with time, i.e.,

$$|\beta^i_{k,t} - \beta^i_{k,t+1}| \leq \delta_k(i), \qquad k = 1, \ldots, J; \ i = 0, 1, \ldots, J; \ t = 1, \ldots, T-1.$$

This model assumes that demand for each firm $k = 1, \ldots, J$ depends on its own as well as its competitors current period prices $p_{1,t}, p_{2,t}, \ldots, p_{J,t}$, unknown parameters $\beta^0_{k,t}$, $\beta^1_{k,t}, \ldots, \beta^J_{k,t}$, and a random noise $\epsilon_{k,t} \sim N(0, \sigma^2_{k,t})$, $k = 1, \ldots, J$. The parameters $\delta_k(i)$, $i = 0, 1, \ldots, J$ are pre-specified constants, called *volatility parameters*, and impose the condition that the coefficients $\beta^0_{k,t}, \beta^1_{k,t}, \ldots, \beta^J_{k,t}$ are Lipschitz continuous. For example setting $\delta_k(i) = 0$, for some i, implies that the i^{th} parameter of the demand is constant in time (this is the usual regression condition).
Firm 1's objectives are to estimate its own demand, its competitors' reactions and finally, set its own prices dynamically in order to maximize its total expected revenue.

The results in [BGT99] suggest that if the true demand is Lipschitz continuous, then the linear model of demand with time varying parameters we consider will indeed converge to the true demand. Moreover, the rate of convergence is faster than other alternative models. While we could use this model in the noncompetitive case of the previous section, it would lead to very high dimensional DPs that we could not solve exactly.

The remainder of this section is organized as follows. In Section 3.1, we present the firm's demand estimation model. In Section 3.2, we present a model that will allow the firm to predict its competitors' prices but also a model that the firm performs to set its own prices. Finally, in Section 3.3, we present some computational results.

3.1 Demand Estimation

Each firm at time t estimates its own demand to be

$$\widehat{D}_{k,t} = \widehat{d}_{k,t} + \varepsilon_{k,t}, \qquad k = 1, \ldots, J$$

where $\widehat{d}_{k,t}$ is a point estimate of the current period demand and $\varepsilon_{k,t}$ is a random noise for firm $k = 1, \ldots, J$. The point estimate of the demand in current period t is given by $\widehat{d}_{k,t} = \widehat{\beta}^0_{k,t} + \widehat{\beta}^k_{k,t} p_{k,t} + \sum_{l \neq k} \widehat{\beta}^l_{k,t} p_{l,t}$, $k = 1, \ldots, J$. The parameter estimates are based on the price and demand realizations in the previous periods.

We assume that the parameter estimates $\widehat{\beta}^k_{k,t}$, $k = 1, \ldots, J$ that describe how each firm's own price affects its own demand, are negative. This is a reasonable assumption since it states that the demand is decreasing in the firm's own price. Moreover, the parameter estimates $\widehat{\beta}^l_{k,t}$, $k \neq l$, $k, l \in \{1, \ldots, J\}$ are nonnegative, indicating that if the competitors set for example, high prices they will increase the firm's own demand.

The firm makes the following distributional assumption on the random noise for each firm's demand,

$$\varepsilon_{k,t} \sim N(0, \widehat{\sigma}^2_{k,t}), \quad \text{where} \quad k = 1, \ldots, J$$

and the demand variance estimated for each firm is,

$$\widehat{\sigma}^2_{k,t} = \frac{\sum_{\tau=1}^{t-1} \left(d_{k,\tau} - \widehat{\beta}^0_{k,t} - \widehat{\beta}^k_{k,t} p_{k,\tau} - \sum_{l \neq k} \widehat{\beta}^l_{k,t} p_{l,\tau} \right)^2}{t - J - 2}, \qquad k = 1, \ldots, J.$$

Notice that for the same reason as in the noncompetitive case, the variance estimates $\widehat{\sigma}^2_{k,t}$, for $k = 1, \ldots, J$, have $t - J - 2$ degrees of freedom. Notice that when the market is a duopoly (i.e., there are two firms competing in the market), then $J = 2$ and the degrees of freedom are $t - 4$ and hence the denominator in the variance estimates is $t - 4$.

For each firm $k = 1, \ldots, J$ we denote by $\widehat{\beta}_k = (\widehat{\beta}_{k,1}, \ldots, \widehat{\beta}_{k,t-1})$, the vector of the estimate of its demand parameters, where $\widehat{\beta}_{k,t} = (\widehat{\beta}^0_{k,t}, \widehat{\beta}^1_{k,t}, \ldots, \widehat{\beta}^J_{k,t})$.

In order to estimate its own demand, Firm 1 solves a regression-type problem. It minimizes the absolute value of the error, that is, the sum over the data acquired so far (i.e., from the past time periods) of the absolute value of the difference between the observed demand $d_{1,\tau}$, and the estimate of the demand $\widehat{d}_{1,\tau} = \widehat{\beta}^0_{k,\tau} + \widehat{\beta}^k_{k,\tau} p_{k,\tau} + \sum_{l \neq k} \widehat{\beta}^l_{k,\tau} p_{l,\tau}$, $\tau = 1, \ldots, t-1$. Alternatively, we could replace the absolute value with a square and consider a more traditional regression-type model. Nevertheless, the absolute value will allow us to convert the problem into a linear optimization problem which is computationally more tractable. That is, we solve the following optimization problem.

$$\begin{aligned}
\min \sum_{\tau=1}^{t-1} & |d_{1,\tau} - (\widehat{\beta}^0_{1,\tau} + \widehat{\beta}^1_{1,\tau} p_{1,\tau} + \sum_{k \neq 1} \widehat{\beta}^k_{1,\tau} p_{k,\tau})| \\
\text{s.t. } & |\widehat{\beta}^i_{1,\tau} - \widehat{\beta}^i_{1,\tau+1}| \leq \delta_1(i), \quad i = 0, 1, \ldots, J, \ \tau = 1, 2, \ldots, t-2 \\
& \widehat{\beta}^1_{1,\tau} \leq 0, \widehat{\beta}^2_{1,\tau} \geq 0, \ldots, \widehat{\beta}^J_{1,\tau} \geq 0.
\end{aligned}$$

Note that we impose the constraint that the parameters are varying slowly with time. This is reflected in the numbers $\delta_1(i)$. As we mentioned above, this problem can be transformed to a linear optimization model, which makes it attractive computationally.

Let $(\widehat{\beta}^i_{1,\tau})^*$, $i = 0, 1, \ldots, J$, $\tau = 1, \ldots, t-1$ be an optimal solution of this problem. Firm 1 would like now to use this information in order to estimate for the parameters in the future, for example in period t parameters $(\widehat{\beta}^0_{1,t}, \widehat{\beta}^1_{1,t}, \ldots, \widehat{\beta}^J_{1,t})$. We propose as an estimate the average:

$$\widehat{\beta}^i_{1,t} = \frac{1}{N} \sum_{l=t-1-N}^{t-1} (\widehat{\beta}^i_{1,l})^*, \qquad i = 0, 1, \ldots, J. \tag{10}$$

that is, the new estimate is an average of the estimates of the N previous periods. In particular, if we choose $N = 1$, we take the new estimate to be equal to the estimate for the previous period.

3.2 Competitor's price prediction and own price setting

In order for Firm 1 to set its own prices in current period t, apart from estimating its own demand, it also needs to predict how its competitors' (Firm $2, \ldots, J$) will react and set their prices in period t. Unfortunately, it is not realistic to assume that Firm 1 observes the past demand realizations of its competitors. Nevertheless, it is more realistic to assume that the information available to Firm 1 at each time period, includes, apart from the realizations of its own demand, also the prices each firm has set in all the previous periods (for example, one can easily find out the fares airlines are charging over the internet). We will assume that Firm 1 believes that its competitors are also setting prices optimally. As a result, Firm 1 will estimate the demand parameters of its competitors using as data the past realizations of prices. That is, Firm 1 tries to guess the parameters of its competitors' demands (by assuming the demand of each competitor also belongs to a parametric family with unknown parameters) through an optimization problem that would exploit the actual observed competitors' prices. This suggests that Firm 1 needs to solve an *inverse* optimization problem. For ease of reading, we first describe the formulation in the setting of two competing firms. We denote by $\widehat{\beta}_2^{t,T}$ the vector of demand parameters for the remaining periods $t, \ldots, T$. Vector $\widehat{\mathbf{p}}_2(\widehat{\beta}_2^{t,T})$ denotes the estimate Firm 1 makes for the price of Firm 2, as a function of the estimates of the parameters of the demand function of Firm 2. It is a vector of the prices over the remaining time periods $t, \ldots, T$. Notice that Firm 2 is an optimizer, therefore, at time t it sets its prices by optimizing its total expected revenue (price times expected demand) over the leftover time horizon $[t, T]$, under the constraint that the demand for the remaining time should not exceed the leftover capacity of the product $c_{2,t}$.

$$\begin{aligned}
\widehat{\mathbf{p}}_2(\widehat{\beta}_2^{t,T}) = \operatorname{argmax} & \sum_{\tau=t}^{T} p_{2,\tau}.(\widehat{\beta}_{2,\tau}^0 + \widehat{\beta}_{2,\tau}^1 p_{1,\tau}^1 + \widehat{\beta}_{2,\tau}^2 p_{2,\tau}) \qquad (11)\\
\text{s.t.} \qquad & \sum_{\tau=t}^{T}(\widehat{\beta}_{2,\tau}^0 + \widehat{\beta}_{2,\tau}^1 p_{1,\tau}^1 + \widehat{\beta}_{2,\tau}^2 p_{2,\tau}) \leq c_{2,t}\\
& p_{\min,\tau} \leq p_{2,\tau} \leq p_{\max,\tau}, \ \forall \tau \in \{t, \ldots, T\}\\
& \widehat{\beta}_{2,\tau}^0 + \widehat{\beta}_{2,\tau}^1 p_{1,\tau}^1 + \widehat{\beta}_{2,\tau}^2 p_{2,\tau} \geq d_{2,min}, \ \forall \tau \in \{t, \ldots, T\}.
\end{aligned}$$

We denote with $d_{2,min} \geq 0$ the minimum allowable allocation Firm 2 is willing to make at each time period (note that this can be equal to zero). $c_{2,t}$ denotes the leftover capacity at time t for Firm 2 and as a result, is equal to $c_2 - \sum_{\tau=0}^{t-1}(\widehat{\beta}_{2,\tau}^0 + \widehat{\beta}_{2,\tau}^2 p_{2,\tau} + \widehat{\beta}_{2,\tau}^1 p_{1,\tau}^1)$. Notice that part of optimization problem (11) involves the estimate of the price of Firm 1 as perceived by Firm 2. As a result, we use notation $p_{1,\tau}^1$ to denote what Firm 1's estimate is of what Firm 2 believes for Firm 1's pricing. The solution of this optimization problem (i.e. the price for Firm 2) is a function of the parameters of the competitor's (Firm 2) demand from period t to T. Note that we set these parameters (see also a discussion in the previous subsection) as an average of the past parameter estimates (see for example, (10) for Firm 1). In conclusion, the previous discussion leads us to conclude that problem (11) gives an estimate $\widehat{p}_{2,\tau}(\widehat{\beta}_{2,\tau})$, $\tau = 1, \ldots, t-1$.

The reason for the previous analysis came from the fact that Firm 1 was trying to estimate the demand parameters for Firm 2 without being able to directly observe the past demand realizations but rather deduce this information through the past price realizations. The last step in this process is for Firm 1 to estimate the demand parameters for Firm 2 by solving a regression-type of model. That is, minimizing the sum over the time periods so far of the absolute value of the difference between the so far observed prices of Firm 2 and the parametric solution of the price of Firm 2 in terms of its demand parameters from (11). Notice that as in the previous section an alternative is to minimize the squared difference. Nevertheless, absolute values allow us to reformulate the problem as a linear optimization problem which is tractable.

In summary, Firm 1 solves the following optimization problem in order to estimate the demand parameters of Firm 2,

$$\begin{aligned}\min \sum_{\tau=1}^{t-1} & \left| p_{2,\tau} - \widehat{p}_{2,\tau}(\widehat{\beta}_{2,\tau}) \right| \\ \text{s.t. } & |\widehat{\beta}^i_{2,\tau} - \widehat{\beta}^i_{2,\tau+1}| \le \delta_2(i), \qquad i = 0,1,2,\ \tau = 1,2,\dots,t-2, \\ & \widehat{\beta}^1_{2,\tau} \ge 0,\ \widehat{\beta}^2_{2,\tau} \le 0.\end{aligned}$$

Since parameter $\widehat{\beta}^2_{2,\tau} \le 0$, it follows that problem (11) is a concave quadratic optimization problem.

Notice that this formulation extends to J competing firms. Similarly to before, Firm 1 estimates the demand parameters for Firm $k \in \{-1\} = \{2,\dots,J\}$ by solving the following optimization problem

$$\begin{aligned}\min \sum_{\tau=1}^{t-1} & \left| p_{k,\tau} - \widehat{p}_{k,\tau}(\widehat{\beta}_{k,\tau}) \right| \\ \text{s.t. } & |\widehat{\beta}^i_{k,\tau} - \widehat{\beta}^i_{k,\tau+1}| \le \delta_k(i), \qquad i = 0,1,\dots,J,\ \tau = 1,2,\dots,t-2, \\ & \widehat{\beta}^1_{k,\tau} \ge 0,\dots,\widehat{\beta}^{k-1}_{k,\tau} \ge 0, \widehat{\beta}^k_{k,\tau} \le 0, \widehat{\beta}^{k+1}_{k,\tau} \ge 0,\dots,\widehat{\beta}^J_{k,\tau} \ge 0,\end{aligned}$$

where $\widehat{p}_k(\widehat{\beta}_k)$, for $k \in \{-1\} = \{2,\dots,J\}$ is the vector solving

$$\begin{aligned}\max \sum_{\tau=t}^{T} & p_{k,\tau}(\widehat{\beta}^0_{k,\tau} + \sum_{l \neq k} \widehat{\beta}^l_{k,\tau} p^1_{l,\tau} + \widehat{\beta}^k_{k,\tau} p_{k,\tau}) \qquad (12) \\ \text{s.t. } & \sum_{\tau=t}^{T} (\widehat{\beta}^0_{k,\tau} + \sum_{l \neq k} \widehat{\beta}^l_{k,\tau} p^1_{k,\tau} + \widehat{\beta}^k_{k,\tau} p_{k,\tau}) \le c_{k,t} \\ & p_{\min,\tau} \le p_{k,\tau} \le p_{\max,\tau},\ \forall \tau \in \{t,\dots,T\} \\ & \widehat{\beta}^0_{k,\tau} + \sum_{l \neq k} \widehat{\beta}^l_{k,\tau} p^1_{l,\tau} + \widehat{\beta}^k_{k,\tau} p_{k,\tau} \ge d_{k,min},\ \forall \tau \in \{t,\dots,T\}, \\ & \text{with } \widehat{\beta}^i_{k,t} = \frac{1}{N} \sum_{l=t-1-N}^{t-1} (\widehat{\beta}^i_{k,l}),\ i = 0,\dots,J, k \neq 1.\end{aligned}$$

Price $p^1_{k,\tau}$ denotes what Firm 1's estimate is of what Firm k believes for its corresponding competitors' pricing. Furthermore, $d_{k,min}$ denotes the minimum allowable allocation Firm k is willing to make at each time period (note that this can be equal to zero). We denote the feasible region of problem (12) as $K(p^1)$, where is a vector of prices representing what Firm 1's estimates are of what the competitors' believe for their corresponding competitors' pricing. Since problem (12) is a concave quadratic optimization problem, we can reformulate it as a variational inequality problem.

Lemma 2. *Problem (12) that Firm k solves in order to determine its pricing policy as a function of its demand parameters, is equivalent to the following variational inequality problem*

$$\text{Find } \widehat{\mathbf{p}}_{\mathbf{k}}(\widehat{\beta}_{\mathbf{k}}) \in K(p^1) : -\sum_{\tau=t}^{T}(\widehat{\beta}^0_{k,\tau} + \sum_{l\neq k}\widehat{\beta}^l_{k,\tau}p^1_{l,\tau} + 2\widehat{\beta}^k_{k,\tau}p_{k,\tau})(p_{k,\tau} - \widehat{p}_{k,\tau}(\widehat{\beta}_{k,\tau})) \geq 0, \ \forall \mathbf{p}_{\mathbf{k}} \in K(p^1). \tag{13}$$

The proof follows easily since the feasible region is a compact, convex set and the optimization objective is a concave function.

Furthermore, since Firm 1's competitors $k \in \{-1\} = \{2, \ldots, J\}$ simultaneously solve this problem, we can combine variational inequality problems (13) into the following single quasi variational inequality.

$$\text{Find } \widehat{\mathbf{p}}(\widehat{\beta}) \in K(p^1) : -\sum_{k=2}^{J}\sum_{\tau=t}^{T}(\widehat{\beta}^0_{k,\tau} + \sum_{l\neq k}\widehat{\beta}^l_{k,\tau}p^1_{l,\tau} + 2\widehat{\beta}^k_{k,\tau}p_{k,\tau})(p_{k,\tau} - \widehat{p}_{k,\tau}(\widehat{\beta}_{k,\tau})) \geq 0, \ \forall \mathbf{p} \in K(p^1). \tag{14}$$

In this case the feasible region $K(p^1) = \{p = (p_2, \ldots, p_J) : \ p_k \in K(p^1_{-k}), \ k = 2, \ldots, J\}$ is the joint feasible region that combines the feasible regions $K(p^1_{-k})$.

This formulation gives rise to the following MPEC formulation describing the problem Firm 1 is solving in order to guess its competitors' parameters.

$$\begin{aligned} \min \ & \sum_{\tau=1}^{t-1}\left|p_{k,\tau} - \widehat{p}_{k,\tau}(\widehat{\beta}_{k,\tau})\right| \\ \text{s.t. } & |\widehat{\beta}^i_{k,\tau} - \widehat{\beta}^i_{k,\tau+1}| \leq \delta_k(i), \qquad i = 0, 1, \ldots, J, \ \tau = 1, 2, \ldots, t-2, \\ & \widehat{\beta}^1_{k,\tau} \geq 0, \ldots, \widehat{\beta}^{k-1}_{k,\tau} \geq 0, \widehat{\beta}^k_{k,\tau} \leq 0, \widehat{\beta}^{k+1}_{k,\tau} \geq 0, \ldots, \widehat{\beta}^J_{k,\tau} \geq 0, \end{aligned}$$

where $\widehat{p}_k(\widehat{\beta}_k)$, $k \in \{-1\} = \{2, \ldots, J\}$ satisfies quasi variational inequality (14).

Own Price Setting Policy

The last step involves Firm 1's own price setting problem. Firm 1 sets its prices by maximizing expected revenues over the remaining time horizon. That is,

$$\begin{aligned}\max \sum_{\tau=t}^{T} & p_{1,\tau}.(\widehat{\beta}^0_{1,\tau} + \widehat{\beta}^1_{1,\tau} p_{1,\tau} + \sum_{k\in\{-1\}} \widehat{\beta}^k_{1,\tau}\widehat{p}_{k,\tau}) \\ \text{s.t.} \sum_{\tau=t}^{T} & (\widehat{\beta}^0_{1,\tau} + \sum_{k\in\{-1\}} \widehat{\beta}^k_{1,\tau}\widehat{p}_{k,\tau} + \widehat{\beta}^1_{1,\tau} p_{1,\tau}) \leq c_{1,t}, \\ & p_{\min,\tau} \leq p_{1,\tau} \leq p_{\max,\tau}, \ \forall \tau \in \{t,\dots,T\} \\ & \widehat{\beta}^0_{1,\tau} + \sum_{k\in\{-1\}} \widehat{\beta}^k_{1,\tau}\widehat{p}_{k,\tau} + \widehat{\beta}^1_{1,\tau} p_{1,\tau} \geq d_{1,min}, \ \forall \tau \in \{t,\dots,T\}.\end{aligned}$$

As before, $d_{1,min}$ denotes the minimum allowable allocation Firm 1 is willing to make at each period (note that this can be zero). This optimization model uses the estimates of the parameters $\widehat{\beta}^i_{1,\tau} = \frac{1}{N}\sum_{l=t-1-N}^{t-1}(\widehat{\beta}^i_{1,l})^*$, $i = 0,1,\dots,J$, for $\tau = t,\dots,T$ (i.e., an average of Firm 1's own demand estimation problem from the past periods), as well as the prediction of its competitors' price $\widehat{p}_{k,\tau} = \frac{1}{N}\sum_{l=t-1-N}^{t-1}(\widehat{p}_{k,l})^*$, $\tau = t,\dots,T$, $k = 2,\dots,J$.

To make the analysis more transparent in the remainder of the paper we will present in detail the case of two competing firms. Nevertheless, the analysis easily extends to the case of several competing firms. We will distinguish between the uncapacitated and the capacitated versions of the problem.

Uncapacitated Case

First we would like to point out that in the uncapacitated case problem (11) separates by time period. Furthermore, as we mentioned above, we assume that Firm 1 believes that Firm 2 is also a revenue maximizer. As a result, Firm 2 solves the optimization problem,

$$\max_{p_{2,\tau}} \ p_{2,\tau}.(\widehat{\beta}^0_{2,\tau} + \widehat{\beta}^1_{2,\tau} p^1_{1,\tau} + \widehat{\beta}^2_{2,\tau} p_{2,\tau}), \qquad \tau = 1,\dots,t.$$

This problem has a closed form solution of the form

$$\widehat{p}_{2,\tau} = \frac{\widehat{\beta}^0_{2,\tau} + \widehat{\beta}^1_{2,\tau} p^1_{1,\tau}}{-2\widehat{\beta}^2_{2,\tau}}, \qquad \tau = 1,\dots,t.$$

Price $p^1_{1,\tau}$ denotes what Firm 1's estimate is of what Firm 2 believes for Firm 1's pricing. Examples of such estimates include: $p^1_{1,\tau} = p_{1,\tau}$, $p^1_{1,\tau} = p_{1,\tau-1}$, or an average of price realizations from several periods prior to period τ.

Firm 1 will then estimate the demand parameters for Firm 2 by solving the following optimization problem

$$\begin{aligned}\min \sum_{\tau=1}^{t-1} & \left| p_{2,\tau} - \frac{\widehat{\beta}^0_{2,\tau} + \widehat{\beta}^1_{2,\tau} p^1_{1,\tau}}{-2\widehat{\beta}^2_{2,\tau}} \right| \\ \text{s.t.} \ & |\widehat{\beta}^i_{2,\tau} - \widehat{\beta}^i_{2,\tau+1}| \leq \delta_2(i), \qquad i = 0,1,2, \ \tau = 1,2,\dots,t-2, \\ & \widehat{\beta}^1_{2,\tau} \geq 0, \ \widehat{\beta}^2_{2,\tau} \leq 0.\end{aligned}$$

As in the model for estimating the current period demand for Firm 1, $\delta_2(i)$, $i = 0, 1, 2$, are *volatility parameters* that we assume to be prespecified constants. The solutions $(\widehat{\beta}^i_{2,\tau})^*$, $i = 0, 1, 2$, of this optimization model allow Firm 1 to estimate its competitor's current period demand by setting:

$$\widehat{\beta}_{2,t} = \frac{1}{N} \sum_{l=t-1-N}^{t-1} (\widehat{\beta}_{2,l})^*.$$

Own Price Setting Policy

After the previous analysis, Firm 1's own price setting problem follows easily. As before, since the problem is uncapacitated, this optimization problem also separates by time period. As a result, Firm 1 sets its prices by maximizing its current period t revenues. That is,

$$\max_{p_{1,t}} \; p_{1,t}.(\widehat{\beta}^0_{1,t} + \widehat{\beta}^1_{1,t} p_{1,t} + \widehat{\beta}^2_{1,t} \widehat{p}_{2,t}).$$

This optimization model uses the estimates of the parameters $\widehat{\beta}^i_{1,t}$, $i = 0, 1, 2$, that we described in Firm 1's own demand estimation problem, as well as the prediction of the competitor's price $\widehat{p}_{2,t} = \frac{\widehat{\beta}^0_{2,t} + \widehat{\beta}^1_{2,t} p^1_{1,t}}{-2\widehat{\beta}^2_{2,t}}$. Notice that this latter part also involves the estimates of the demand parameters $\widehat{\beta}^i_{2,t}$, $i = 0, 1, 2$ arising through the inverse optimization problem in the competitor's price prediction problem.

Capacitated Case

We assume that both firms face a total capacity c_1 and c_2 respectively that they need to allocate in the total time horizon. Quantities $c_{1,t}$ and $c_{2,t}$ denote the leftover capacities of Firm 1 and 2 respectively in the beginning of period t. As before, Firm 1 makes the behavioral assumption that Firm 2 is also a revenue maximizer. As a result, in general Firm 2 will solve problem (11). In order to perform some computations and derive some insights, in what follows we will assume that the firms solve their price setting problems myopically. As a result, the price prediction problem that Firm 1 solves for predicting its competitor's prices becomes

$$\hat{p}_{2,t} = \arg\max_{p \in \mathcal{P}_2} \; p \min \Big\{ \left(\widehat{\beta}^0_{2,t} + \widehat{\beta}^2_{2,t} p_{2,t} + \widehat{\beta}^1_{2,t} p^1_{1,t} \right), \\ c_2 - \sum_{\tau=0}^{t-1} (\widehat{\beta}^0_{2,\tau} + \widehat{\beta}^2_{2,\tau} p_{2,\tau} + \widehat{\beta}^1_{2,\tau} p^1_{1,\tau}) \Big\}.$$

As in the uncapacitated case, $p^1_{1,\tau}$ denotes Firm 1's estimate of what Firm 2 assumes for Firm 1's own pricing. Examples include: $p^1_{1,\tau} = p_{1,\tau}$, or $p_{1,\tau-1}$, or considering an average of the prices Firm 1 sets in several previous periods. We can now estimate Firm 2's demand parameters through the following optimization model

$$\begin{aligned}
&\min \sum_{\tau=1}^{t-1} |p_{2,\tau} - \hat{p}_{2,\tau}| \\
&\text{s.t. } |\widehat{\beta}^i_{2,\tau} - \widehat{\beta}^i_{2,\tau+1}| \leq \delta_2(i), \qquad i = 0,1,2,\ \tau = 1,2,\dots,t-2 \\
&\quad \widehat{\beta}^1_{2,\tau} \geq 0,\ \widehat{\beta}^2_{2,\tau} \leq 0,
\end{aligned}$$

where $\hat{p}_{2,t} \in \arg\max_{p \in \mathcal{P}_2}\ p \min\left\{ \left(\widehat{\beta}^0_{2,t} + \widehat{\beta}^2_{2,t} p + \widehat{\beta}^1_{2,t} p^1_{1,t}\right),\ c_{2,t} \right\}$.

Let $(\widehat{\beta}^i_{2,\tau})^*$, $i = 0,1,2$, $\tau = 1,\dots,t-1$ be optimal solutions to this optimization problem. As before, Firm 1 estimates its competitor's current period demand parameters as

$$\widehat{\beta}^i_{2,t} = \frac{1}{N} \sum_{l=t-1-N}^{t-1} (\widehat{\beta}^i_{2,l})^*, \qquad i = 0,1,2.$$

Myopic Own Price Setting Policy

After computing its own and its competitor's demand parameter estimates and establishing a prediction on the price of its competitor for the current period, Firm 1 is ready to set its own current period price. As in the uncapacitated case, Firm 1 solves the current period revenue maximization problem, that is,

$$p_{1,t} \in \arg\max_{p \in \mathcal{P}} \left[p \min\left\{ \left(\widehat{\beta}^0_{1,t} + \widehat{\beta}^1_{1,t} p + \widehat{\beta}^2_{1,t} \hat{p}_{2,t}\right), c_{1,t} \right\} \right],$$

where $c_{1,t} = c_1 - \sum_{\tau=1}^{t-1} d_{1,\tau}$ is Firm 1's remaining capacity in period t. Moreover, the demand parameters $\widehat{\beta}^i_{1,t} = \frac{1}{N}\sum_{k=t-1-N}^{t-1} (\widehat{\beta}^i_{1,k})^*$, $\widehat{\beta}^i_{2,t} = \frac{1}{N}\sum_{l=t-1-N}^{t-1} (\widehat{\beta}^i_{2,l})^*$, $i = 0,1,2$, and finally, the estimates of the competitor's prices are

$$\hat{p}_{2,t} \in \arg\max_{p \in \mathcal{P}_2}\ p \min\left\{ \left(\widehat{\beta}^0_{2,t} + \widehat{\beta}^2_{2,t} p + \widehat{\beta}^1_{2,t} p^1_{1,t}\right),\ c_{2,t} \right\}.$$

3.3 Computational Results

We consider two firms competing for one product. The true models of demand for the two firms respectively are as follows:

$$d_{1,t} = 50 - .05 p_{1,t} + .03 p_{2,t} + \varepsilon_{1,t}$$

$$d_{2,t} = 50 + .03 p_{1,t} - .05 p_{2,t} + \varepsilon_{2,t}$$

where the $\varepsilon_{1,t}$, $\varepsilon_{2,t}$ $\sim N(0,16)$. Moreover, the prices for both firms range in the sets $\mathcal{P}_1 = \mathcal{P}_2 = [100,\ 900]$, the time horizon is $T = 150$ and finally we assume that $p_{1,1} = p_{2,1} = 500$. Finally, we assume an uncapacitated setting.

We compare three pricing policies: (a) random pricing, (b) price matching, and (c) optimization based pricing using the methods we outlined in this section. A firm employing the random pricing policy chooses a price at random from the feasible price set. In particular, we consider a discrete uniform distribution over the set of integers [100, 900]. A firm employing the price matching policy sets, in the current period, the price its competitor set in the previous period. Finally, a firm employing optimization based pricing first solves the demand estimation problem in order to estimate its current period parameter estimates using linear programming, supposes its competitor will repeat its previous period pricing decision, and then uses myopic pricing in order to set its prices. In Table 3, we report the revenue from the three strategies, over 1000 simulation runs.

Table 3. A comparison of revenues under random, matching, optimization based pricing policies

Firm 1	Firm 2	1 Avg(Rev)	2 Avg(Rev)	1 Std(Rev)	2 Std(Rev)
Opt	Rand	3,126,000	2,909,200	70,076	109,790
Rand	Rand	2,638,800	2,616,900	63,112	61,961
Match	Rand	2,602,700	2,603,200	117,470	123,070
Opt	Match	3,791,100	3,779,400	177,540	197,370
Rand	Match	2,603,200	2,602,700	123,070	117,470
Opt	Opt	3,757,700	3,804,700	70,577	129,530
Rand	Opt	2,909,200	3,126,000	109,790	70,076
Match	Opt	3,779,400	3,791,100	197,370	177,540

In order to obtain intuition from Table 3, we fix the strategy the competitor is using, and then see the effect on revenue of the policy followed by Firm 1. If Firm 2 is using the random pricing policy, it is clear that Firm 1 has a significant increase in revenue by using an optimization based policy. Similarly, if Firm 2 is using a matching policy, again the optimization based policy leads to significant improvements in revenue. Finally, if Firm 2 is using an optimization based policy, then the matching policy is slightly better than the optimization based policy. However, given that the margin is small and given the variability in the estimation process, it might still be possible for the optimization based policy to be stronger. It is thus fair to say, that at least in this example, no matter what policy Firm 2 is using, Firm 1 seems to be better off by using an optimization based policy.

4 Conclusions

We introduced models for dynamic pricing in an oligopolistic market. In the first part of the paper, we studied models in a noncompetitive environment in order to understand the effects of demand learning. By considering the framework of dynamic programming with incomplete state information for jointly estimating the demand and setting prices for a firm, we proposed increasingly more computationally intensive algorithms that outperform myopic policies. Our overall conclusion is that dynamic programming models based on incomplete information are effective in jointly estimating the demand and setting prices for a firm.

In the second part of the paper, we studied pricing in a competitive environment. We introduced a more sophisticated model of demand learning in which the price elasticity is a slowly varying function of time. This allows for increased flexibility in the modeling of the demand. We outlined methods based on optimization for jointly estimating the Firm's own demand, its competitors' demands, and setting prices. In preliminary computational work, we found that optimization based pricing methods offer increased revenue for a firm independently of the policy the competitor firm is following.

Acknowledgments. We would like to thank the referees for their detailed comments that have helped improve the presentation and content of this paper. We would also like to thank Marc Coumeri for performing some of the computations in this paper. Research supported by the Singapore-MIT Alliance, the PECASE Award DMI-9984339 from the National Science Foundation, the Charles Reed Faculty Initiative Fund and the New England University Transportation Research Grant.

References

[Bag84] Bagchi, A.: Stackleberg Differential Games in Economic Models. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York (1984)

[Bas86] Basar, T.: Dynamic Games and Applications in Economics. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York (1986)

[BF99] Bernstein, F., Federgruen, A.: Pricing and Replenishment Strategies in a Distribution System with Competing Retailers, Working Paper (1999)

[Be95] Bertsekas, D.: Dynamic Programming and Optimal Control I. Athena Scientific, MA (1995)

[BeT96] Bertsekas, D., Tsitsiklis, J.: Neuro-Dynamic Programming, Athena Scientific, MA (1996)

[BGT99] Bertsimas, D., Gamarnik, D., Tsitsiklis, J.: Estimation of Time-Varying Parameters in Statistical Models: An Optimization Approach. Machine Learning, **35**, 225–245 (1999)

[BiT97] Bertsimas, D., Tsitsiklis, J.: Introduction to Linear Optimization. Athena Scientific, MA (1997)

[BC02] Bitran, G., Caldentey, R.: An Overview of Pricing Models for Revenue Management. Submitted to MSOM (2002)

[BM97] Bitran, G., Mondschein, S.: Periodic Pricing of Seasonal Products in Retailing. Management Science, **43(1)**, 64–79 (1997)

[CN] Cachon, G., Netessine, S.: Game theory in supply chain analysis. In: Simchi-Levi, D., Wu, S.D., Shen, M. (eds) Supply Chain Analysis in E-business era. Eds, Kluwer, Forthcoming

[CSSS01] Chan, L.M.A., Shen, Z.J.M., Simchi-Levi, D., Swann, J.: Coordinating Pricing, Inventory, and Production: A Taxonomy and Review, In: Handbook on Supply Chain Analysis in the eBusiness Era, Kluwer Academic Publishers, to appear(2001)

[CSS00] Chan, LMA., Simchi-Levi, D., Swann, J.: Flexible Pricing Strategies to Improve Supply Chain Performance. Working Paper (2000)

[DP99] Dada, J.D., Petruzzi, N.C.: Pricing and the Newsvendor Problem: A Review with Extensions. Operations Research, **47**, 183–194 (1999)

[DP01] Dada, J.D., Petruzzi, N.C.: Note: The Newsvendor Model with Endogenous Demand. Management Science, **47**, 1488–1497 (2001)

[DJ88] Dockner, E., Jorgensen, S.: Optimal Pricing Strategies for New Products in Dynamic Oligopolies. Marketing Science, **7(4)**, 315-334 (1988)

[EK] Elmaghraby, W., Keskinocak, P.: Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices and Future Directions. to appear in Management Science.

[FH97] Federgruen, A., Heching, A.: Combined Pricing and Inventory Control Under Uncertainty, Operations Research, **47(3)**, 454–475 (1997)

[F95] Feng, Y., Gallego, G.: Optimal Starting Times for End-of-Season Sales and Optimal Stopping Times for Promotional Fares. Management Science, **41(8)**, 1371–1391 (1995)

[Fr77] Friedman, J.W.: Oligopoly and the Theory of Games. North Holland, Amsterdam (1977)

[Fr82] Friedman, J.W.: Oligopoly Theory. In: Handbook of Mathematical Economics II chapter 11. North Holland, Amsterdam (1982)

[Fr83] Friedman, J.W.: Oligopoly Theory. Cambridge University Press, Cambridge (1983)

[FT86] Fudenberg, D., Tirole, J.: Dynamic Models of Oligopoly. Harwood Academic, London (1986)

[GvR94] Gallego, G., van Ryzin, G.: Optimal Dynamic Pricing of Inventories with Stochastic Demand Over Finite Horizons. Management Science, **40(8)**, 999–1020 (1994)

[GvR97] Gallego, G., van Ryzin, G.: A Multiproduct Dynamic Pricing Problem and its Applications to Network Yield Management, Operations Research, **45(1)**, 24–41 (1997)

[GK98] Gibbens, R.J., Kelly, F.P.: Resource Pricing and the Evolution of Congestion Control. Working Paper (1998)

[Gi00] Gilbert, S. 2000. Coordination of Pricing and Multiple-Period Production Across Multiple Constant Priced Goods. *Management Science*, **46**(12), 1602-1616.

[KP02] Kachani, S., Perakis, G.: A Fluid Dynamics Model of Dynamic Pricing and Inventory Control for Make-to-Stock Manufacturing Systems. Working Paper. Operations Research Center. MIT (2002)

[Ka96] Kalyanam, K.: Pricing Decisions Under Demand Uncertainty: A Bayesian Mixture Model Approach. Marketing Science, **15(3)**, 207–221 (1996)

[Ke94] Kelly, F.P.: On Tariffs, Policing and Admission Control for Multiservice Networks. Operations Research Letters, **15**, 1–9 (1994)

[KMT98] Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stabilit. Journal of the Operational Research Society, **49**, 237–252 (1998)

[Kl01] Kleywegt, A.J.: An Optimal Control Problem of Dynamic Pricing. GaTech ISyE Working Paper, Atlanta (2001)

[KRA96] Kopalle, P., Rao, A., Assuncao, J.: Asymmetric Reference Price Effects and Dynamic Pricing Policies. Marketing Science, **15(1)**, 60–85 (1996)

[Ku97] Kuhn, H.: *Classics in Game Theory*, Princeton University Press, NJ (1997)

[LKM92] Lilien, G., Kotler, P., Moorthy, K.: Marketing Models. Prentice Hall, NJ (1992)

[MWG95] Mas-Colell, A., Whinston, M., Green, J.: Microeconomic Theory. Oxford University Press, New York (1995)

[MV99] McGill, J., Van Ryzin, G.: Focused Issue on Yield Management in Transportation. Transportation Science, **33(2)**, (1999).

[Na93] Nagurney, A.: Network Economics A Variational Inequality Approach. Kluwer Academic Publishers, Boston (1993)

[PT98] Paschalidis, I., Tsitsiklis, J.: Congestion-Dependent Pricing of Network Services. Technical Report (1998)

[PS03] Perakis, G., Sood, A.: Competitive Multi-period Pricing for Perishable Products. Working Paper, Operations Research Center, MIT (2003)

[Ri95] Rice, J.: Mathematical Statistics and Data Analysis. Second Edition, Duxbury Press, California (1995)

[SGK01] Supernak, J., Golub, J.T.F., Kaschade, C., Kazimi, C., Schreffler, E., Steffey, D.: Phase II Year Three Overall Report: I-15 Congestion Pricing Project Monitoring and Evaluation Services, San Diego State University, San Diego, CA (2001)

[SKSK] Supernak, J., Kaschade, C., Steffey, D., Kubiak, G.: I-15 Congestion Pricing Project Monitoring and Evaluation Services: Year Three Traffic Study. San Diego State University, San Diego, CA (2001)

[Su02] Sullivan, E.C.: Updated Observations for the State Route 91 Value Priced Express Lane. 81st Annual Transportation Research Board Meeting, Preprint Paper No. 02-2554, Washington, DC (2002)

[Th74] Thomas, L.G.: Price and Production Decision with Random Demand. Operations Research, **22**, 513–518 (1974)

[TM85] Tirole, J., Maskin, E.: A Theory of Dynamic Oligopoly II: Price Competition. MIT Working Papers (1985)

[VD99] Van Mieghen, J., Dada, M.: Price vs Production Postponement. Management Science, **45(12)**, 1631–1649 (1999)

[Va99] Van Mieghen, J.: Differentiated Quality of Service: Price and Service Discrimination in Queueing Systems. Working Paper (1999)

[Vi99] Vives, X.: Oligopoly Pricing - Old Ideas and New Tools. MIT Press (1999)

[WB92] Weatherford, L., Bodily, S.: A Taxonomy and Research Overview of Perishable Asset Revenue Management: Yield Management, Overbooking and Pricing. Operations Research, **40(5)**, 831–844 (1992)

[Wi92] Williamson, E.: Airline Network Seat Inventory Control: Methodologies and Revenue Impacts. Ph.D. Thesis Flight Transportation Lab, MIT (1992)

[Wo93] Wilson, R.: Nonlinear Pricing. Oxford University Press (1993)

[YS04] Yano, C., Gilbert, S.: Coordinated Pricing and Production/ Procurement decisions: A Review. In: Amiya Chakravarty, Jehoshua Eliashberg (eds) Managing Business Interfaces: Marketing, Engineering and Manufacturing Perspectives. Kluwer Academic Publishers (2004)

[Yo78] Young, L.: Price, Inventory, and the Structure of Uncertain Demand. New Zealand Operations Research, **6**, 157–177 (1978)

[Za70] Zabel, E.: Monopoly and Uncertainty. Rev. Econom. Studies, **37**, 205–219 (1970)

[Za72] Zabel, E.: Multi-period Monopoly and Uncertainty. Journal of Economic Theory, **5**, 524–536 (1972)

[ZZ00] Zhao, W., Zheng, Y.S.: Optimal Dynamic Pricing for Pershable Assets with Nonhomogenous Demand. Management Science, **46**, 375–388 (2000)

Appendix

Proof of Proposition 1

The first order conditions of the least squares problem for $\widehat{\beta}_t$ and $\widehat{\beta}_{t-1}$ respectively, imply that

$$\sum_{s=1}^{t-1}\left(d_s - \mathbf{x}_s'\widehat{\beta}_t\right)\mathbf{x}_s = \mathbf{0} \tag{15}$$

$$\sum_{s=1}^{t-2}\left(d_s - \mathbf{x}_s'\widehat{\beta}_{t-1}\right)\mathbf{x}_s = \mathbf{0}. \tag{16}$$

If we write, $\widehat{\beta}_t = \widehat{\beta}_{t-1} + \mathbf{a}$, where $\mathbf{a}$ is some vector, it follows from (15) that

$$\sum_{s=1}^{t-1}\left(d_s - \mathbf{x}_s'\widehat{\beta}_{t-1} - \mathbf{x}_s'\mathbf{a}\right)\mathbf{x}_s = \mathbf{0}.$$

This in turn implies that,

$$\sum_{s=1}^{t-2}\left(d_s - \mathbf{x}_s'\widehat{\beta}_{t-1} - \mathbf{x}_s'\mathbf{a}\right)\mathbf{x}_s + \left(d_{t-1} - \mathbf{x}_{t-1}'\widehat{\beta}_{t-1} - \mathbf{x}_{t-1}'\mathbf{a}\right)\mathbf{x}_{t-1} = \mathbf{0}. \tag{17}$$

Subtracting (16) from (17) we obtain that

$$\sum_{s=1}^{t-1}\left(\mathbf{x}_s'\mathbf{a}\right)\mathbf{x}_s = \left(d_{t-1} - \mathbf{x}_{t-1}'\widehat{\beta}_{t-1}\right)\mathbf{x}_{t-1}.$$

Therefore, $\mathbf{a} = \mathbf{H}_{t-1}^{-1}\mathbf{x}_{t-1}\left(d_{t-1} - \mathbf{x}_{t-1}'\widehat{\beta}_{t-1}\right)$, with $\mathbf{H}_{t-1} = \sum_{s=1}^{t-1}\left(\mathbf{x}_s\mathbf{x}_s'\right) = \begin{bmatrix} t-1 & \sum\limits_{s=1}^{t-1} p_s \\ \sum\limits_{s=1}^{t-1} p_s & \sum\limits_{s=1}^{t-1} p_s^2 \end{bmatrix}.$

Proof of Proposition 2

Let t be the current time and $s \geq t$. We first relate the variance in period s,

$$\widehat{\sigma}_{s,t}^2 = \frac{\sum\limits_{i=1}^{s-1}\left(d_i - \widehat{\beta}_{s,t}^0 - \widehat{\beta}_{s,t}^1 p_i\right)^2}{s-3} \tag{18}$$

with the variance in the next period $s+1$,

$$E[\widehat{\sigma}_{s+1,t}^2|d_s] = \frac{\sum\limits_{i=1}^{s}\left(d_i - \widehat{\beta}_{s+1,t}^0 - \widehat{\beta}_{s+1,t}^1 p_i\right)^2}{s-2}.$$

By expanding this last equation and separating the period s terms from the previous period $s-1$ we obtain

$$E[\widehat{\sigma}^2_{s+1,t}|d_s] = \frac{\sum\limits_{i=1}^{s-1} d_i^2 + d_s^2 - 2\widehat{\beta}^0_{s+1,t}\sum\limits_{i=1}^{s-1} d_i - 2\widehat{\beta}^0_{s+1,t} d_s - 2\widehat{\beta}^1_{s+1,t}\sum\limits_{i=1}^{s-1} p_i d_i - 2\widehat{\beta}^1_{s+1,t} p_s d_s}{s-2} + \tag{19}$$

$$\frac{s\left(\widehat{\beta}^0_{s+1,t}\right)^2 + 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t}\sum\limits_{i=1}^{s-1} p_i + 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t} p_s + \left(\widehat{\beta}^1_{s+1,t}\right)^2 \sum\limits_{i=1}^{s-1} p_i^2 + \left(\widehat{\beta}^1_{s+1,t}\right)^2 p_s^2}{s-2}.$$

Substituting Eq. (18) we obtain

$$\sum_{i=1}^{s-1} d_i^2 = \widehat{\sigma}^2_{s,t}(s-3) + 2\widehat{\beta}^0_{s,t}\sum_{i=1}^{s-1} d_i + 2\widehat{\beta}^1_{s,t}\sum_{i=1}^{s-1} d_i p_i - (s-1)\left(\widehat{\beta}^0_{s,t}\right)^2 \tag{20}$$
$$-2\widehat{\beta}^0_{s,t}\widehat{\beta}^1_{s,t}\sum_{i=1}^{s-1} p_i - \left(\widehat{\beta}^1_{s,t}\right)^2 \sum_{i=1}^{s-1} p_i^2.$$

We substitute (20) into (19) to obtain that $E[\widehat{\sigma}^2_{s+1,t}|d_s]$ is equal to:

$$\frac{\widehat{\sigma}^2_{s,t}(s-3) + 2\widehat{\beta}^0_{s,t}\sum\limits_{i=1}^{s-1} d_i + 2\widehat{\beta}^1_{s,t}\sum\limits_{i=1}^{s-1} d_i p_i - (s-1)\left(\widehat{\beta}^0_{s,t}\right)^2 - 2\widehat{\beta}^0_{s,t}\widehat{\beta}^1_{s,t}\sum\limits_{i=1}^{s-1} p_i}{s-2} +$$

$$\frac{-\left(\widehat{\beta}^1_{s,t}\right)^2 \sum\limits_{i=1}^{s-1} p_i^2 + d_s^2 - 2\widehat{\beta}^0_{s+1,t}\sum\limits_{i=1}^{s-1} d_i - 2\widehat{\beta}^0_{s+1,t} d_s - 2\widehat{\beta}^1_{s+1,t}\sum\limits_{i=1}^{s-1} p_i d_i - 2\widehat{\beta}^1_{s+1,t} p_s d_s}{s-2} +$$

$$\frac{s\left(\widehat{\beta}^0_{s+1,t}\right)^2 + 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t}\sum\limits_{i=1}^{s-1} p_i + 2\widehat{\beta}^0_{s+1,t}\widehat{\beta}^1_{s+1,t} p_s + \left(\widehat{\beta}^1_{s+1,t}\right)^2 \sum\limits_{i=1}^{s-1} p_i^2 + \left(\widehat{\beta}^1_{s+1,t}\right)^2 p_s^2}{s-2}.$$

Since in the beginning of period s, $d_s = \widehat{\beta}^0_{s,t} + \widehat{\beta}^1_{s,t} p_s + \varepsilon_{s,t}$, and taking expectations over $\varepsilon_{s,t}$ with $E[\varepsilon_{s,t}] = 0$ and $E[\varepsilon^2_{s,t}] = \widehat{\sigma}^2_{s,t}$, we obtain Eq. (3).

Congestion Pricing of Road Networks with Users Having Different Time Values

Leonid Engelson[1] and Per Olov Lindberg[2]

[1] Department of Infrastructure, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden lee@infra.kth.se

[2] Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden polin@mai.liu.se

Summary. We study congestion pricing of road networks with users differing only in their time values. In particular, we analyze the marginal social cost (MSC) pricing, a tolling scheme that charges each user a penalty corresponding to the value of the delays inflicted on other users, as well as its implementation through fixed tolls. We show that the variational inequalities characterizing the corresponding equilibria can be stated in symmetric or nonsymmetric forms. The symmetric forms correspond to optimization problems, convex in the fixed-toll case and nonconvex in the MSC case, which hence may have multiple equilibria. The objective of the latter problem is the total value of travel time, which thus is minimized at the global optima of that problem. Implementing close-to-optimal MSC tolls as fixed tolls leads to equilibria with possibly non-unique class specific flows, but with identical close-to-optimal values of the total value of travel time. Finally we give an adaptation, to the MSC setting, of the Frank-Wolfe algorithm, which is further applied to some test cases, including Stockholm.

Key words: Multi-Class Traffic Assignment, Congestion Pricing, Marginal Social Cost

1 Introduction

Traffic in large cities has become a major problem for society. It is inefficient, causes accidents and pollutes the environment. It has become a common viewpoint among transportation economists that charging some kind of fee from the users of the road network is necessary. The European Commission [EC01, p. 77] plans to propose a framework directive, setting out the principles of an infrastructure-charging system, including a common methodology for setting charging levels which incorporate external costs. In 1998, the Swedish Government [SG98] recommended that transport taxes and fees should correspond as close as possible to the marginal costs caused by the transport. Road pricing

Mathematical and Computational Models for Congestion Charging, pp. 81-104

has further been implemented in Singapore, London, and several Norwegian cities. In the Stockholm region, various studies have considered different toll patterns and performed social cost and benefit analyses for various time horizons. Moreover, the Stockholm city government has decided to carry out a full-scale trial of road pricing. Events such as these make questions related to the choice of pricing system and fee levels highly timely.

By request of the Swedish Institute for Transport and Communications Analysis (SIKA), the consulting firm Inregia [Ing01] attempted to calculate marginal cost road charges for Stockholm County for three user classes (work and school trips, business trips, and other trips) with different time values (0.98, 3.30 and 0.19 SEK/min. respectively) estimated from travel surveys. In this implementation, the marginal cost tolls were updated by the method of successive average, resulting in slow convergence and large link volume oscillations. This led to the initiation of a research project whose results are presented here.

In transportation science, the classical marginal social cost pricing theory (e.g., [BMW56]) suggests that for the most efficient usage of a congested road network with homogeneous users, each user should be charged a toll equal to the total value of time loss inflicted on other users of the network. In the case of fixed travel demand, this will induce an equilibrium that is system optimal in the sense that the total cost of network usage is minimal, *assuming that all users have fixed and identical time values.* To calculate this toll pattern, one modifies the link cost functions by adding the external cost term and solves for a user-optimal solution, using e.g. the Frank-Wolfe algorithm. The solution is unique in the terms of link flows and tolls, provided that the modified link travel cost functions are positive and strictly increasing (see, e.g., [Pat94, Ch 2]). Once the tolls are fixed and implemented, the user-optimal flow pattern will be system-optimal.

However, it is well known that travelers may have widely varying time values. In Stockholm, for instance, estimated time values for different trip purposes vary by a factor of more than seventeen, as indicated above. Hence, since tolls cannot be charged in time units, but have to be levied in monetary equivalents, different user groups will react differently to a given toll scheme. Therefore, methods to compute tolled equilibria need to account for these different reactions, leading to multi-class user equilibrium problems.

Dafermos [Daf73] has shown that in the case with multiple user classes, a modification of the link cost functions similar to the one above yields a user-optimizing flow that is also system-optimizing (assuming, however, convexity of the system objective).

Netter [Net71], on the other hand, argues that the assumption of convexity of the total travel cost is unrealistic in the context of marginal cost pricing in multi-class transportation networks. When link travel times depend on class specific volumes on the links and are different for different user classes, the user equilibrium is not generally unique even in toll-free networks or in networks with fixed tolls. So, even if the planner knows the tolls corresponding to the system optimum, the achievement of this optimum will not necessarily follow from the implementation of these tolls. In Section 5 of this article, we provide an example that supports Netter's statement in [Net71].

Notwithstanding the practical difficulties in its implementation, tolls based on marginal social costs are useful for evaluating other tolling policies when used in conjunction with the relative welfare index introduced by Verhoef et al. [VNR95].

Hearn and co-authors (e.g., [HY02], [HR98]) argue that instead of marginal social cost tolls, it might be worthwhile contemplating alternative tolls achieving the systems optimum; optimizing some other objective, such as the number of toll booths.

Using an entertaining parable in two companion papers ([Dia99a] and [Dia99b]), Dial studies the problem of determining "optimal" congestion tolls under continuous distribution of time values over the users. He addresses how such tolls can be determined by solving a variational inequality and provides a solution method. However, Lindberg [Lin05] indicates that [Dia99a] contains several flaws.

Yang and Huang [YH04] consider the social optimum in terms of cost, as well as the system optimum in terms of time, in the context of users with different time values. For the cost optimum they demonstrate that the optimum flows are equilibrium flows for a fixed-toll problem with marginal social cost tolls. However, they claim that the total social cost is a strictly convex function (Section 3.1). We provide a counterexample to this in Section 5 below. Concerning the time optimum, they show by an interesting argument that there exists a monetary toll pattern that minimizes the total travel time in the network. The corresponding tolls can be calculated by consecutively solving two optimization problems with linear constraints – one with a convex objective and the other with a linear objective.

While minimizing the total travel time might be an interesting task from a pure transportation planning view, the overall economic efficiency, in the case of fixed travel demand for all user classes, requires minimization of the *total value of travel time*

In many case studies, problems related to user heterogeneity have been circumvented by application of an average time value to all users. However, as shown by Eliasson [Eli00], such models can lead to erroneous conclusions about the efficiency of the resulting toll system.

The subject of the current paper is the study of tolled equilibria and marginal cost pricing in networks with several user classes that differ only by their time values. Possible applications include modeling of individual travelers that have different trip purposes (e.g. work, business, leisure etc.) and therefore perceive the relation between travel time and monetary cost in dissimilar ways. Forerunners of the current paper are Engelson, Lindberg and Daneva [ELD03] and Engelson and Lindberg [EL02].

For the remainder, Section 2 of the paper is mainly devoted to basic definitions, including that of a multi-class equilibrium, and to statements of the variational inequalities characterizing equilibria. For cases with symmetric cost functions we show that all stable equilibria correspond to local minima of the corresponding objective.

Realizing that marginal social cost tolls need to be implemented as fixed tolls, we consider multi-class equilibria under fixed tolls in Section 3. We show that the variational inequality defining the fixed-toll equilibrium can be stated in nonsymmetric or symmetric forms, and thus it has a corresponding "equivalent" optimization formulation based on the symmetric form. We demonstrate that the optimization formulation is convex, but show that the class specific link flows are not necessarily unique. In spite of this, the total value of travel time is unique.

Section 4 is devoted to the case with flow-dependent tolls based on marginal social cost (MSC tolls). Again the variational inequality can be stated in symmetric or nonsymmetric forms. The optimization problem corresponding to the symmetric version has a non-convex objective function, which turns out to be the total value of travel time. Finding the MSC tolls thus corresponds to a form of welfare optimization. However, due to non-convexity, there may be multiple local optima which implies multiple equilibria. Implementing equilibrium tolls as fixed tolls does not necessarily achieve the corresponding equilibrium, but still gives flows with the same total value of travel time. Thus, using fixed tolls we can achieve the same levels of welfare (in the form of total value of travel time) as when optimizing over all feasible flows.

In Section 5, we first consider a simple example illustrating the nonconvexity of the total value of travel time function; an example which is then expanded to demonstrate that this function is non-convex in general. Section 6 outlines an algorithm of Frank-Wolfe type for the marginal cost toll case, an algorithm that is applied in Section 7 to the classical Sioux Falls network and to that of Stockholm.

2 User equilibria in networks with several user classes

This section defines multi-class equilibria and characterizes them as variational inequalities (VI). In addition, the stability of such equilibria and the conditions under which these VI's can be addressed as optimization problems are also considered.

As noted in the introduction, when studying tolled equilibria one needs to consider multi-class equilibria, i.e. with user classes having different perceptions of travel costs. Dafermos [Daf73] studies such equilibria, with equilibrium definitions, however, that are nonstandard today. Multi-class equilibrium definitions have also been given by Netter [Net71] and Van Vliet [VBS86], but in publications not easily accessible. Due to these circumstances, we will state equilibrium definitions of Wardrop type. We also state corresponding variational inequalities characterizing equilibria, as well as symmetry conditions guaranteeing the existence of corresponding optimization problems.

Consider a road network consisting of *nodes* $n \in N$ and directed *links* $a \in A$. Let $W \subset N \times N$ be the set of *OD pairs.* Assume that OD *demands* q_w^k between the OD pairs $w \in W$ for each *user class* $k \in K$ are given. For each link a, there are associated continuously differentiable *cost functions* $c_a^k : \Re_+^{A \times K} \to \Re_+$ that represent the cost of traversing link a for a user in class k and depends on the class specific volumes in all links. For the time being, c_a^k is a general function, but it will be endowed with special structure in the next section. We also let c^k denote the (column) vector of link costs (or functions) for class k, c_a the (row) vector of class costs for link a, and c the (matrix of) class specific link costs. We will use the same convention for other entities indexed by a and k.

Let R_w be the set of *routes* (or paths) connecting OD pair w and $R = \bigcup_{w \in W} R_w$, the set of all routes. In analogy with c, let $h = (h_r^k)$ be the matrix of class k flows on routes r, with columns h^k (of class k *route flows*) and rows h_r (of class flows on route r). Let the set of feasible *route flows* be $H = \left\{h \in \Re_+^{R \times K} : \sum_{r \in R_w} h_r^k = q_w^k, \forall w \in W,\ k \in K\right\}$. Further, denote by F the set of feasible *link flows* (or volumes), i.e.
$F = \left\{f \in \Re_+^{A \times K} : \exists h \in H, \forall a \in A,\ k \in K \quad f_a^k = \sum_{r \in R} \delta_{ar} h_r^k\right\}$, where δ_{ar} is 1 if route r traverses link a, and 0 otherwise. Introducing the link-route *incidence matrix* $\Delta = (\delta_{ar})$, and using the indexing convention, we see that $f^k = \Delta h^k$, $f = \Delta h$ and $F = \Delta H$.

Let $C = (C_r^k)$ be the matrix of total travel costs for users of class k on route r, with columns C^k. We assume that C_r^k is additive over the links, i.e. that $C_r^k(h) = \sum_a \delta_{ar} c_a^k (\Delta h)$, or with our notation conventions, $C^k = \Delta^T c^k$ and $C = \Delta^T c$.

Definition 1. *(Multi-class Wardrop equilibrium) The route flow matrix $\hat{h} \in H$ is a (route flow) multi-class equilibrium if, for any OD pair and class, each route that is used by the class (i.e. has positive flow) has cost not greater than the cost of any other route for that OD pair and class.*

In similarity to the single class case, the equilibrium definition can alternatively be stated as a *variational inequality (VI)* in the set of feasible route flows or in the set of link flows. We will use $\langle *, * \rangle$ to denote the inner product between vectors (or matrices) of appropriate dimensions.

Lemma 1. *A route flow $\hat{h} \in H$ is an equilibrium if and only if $\hat{h}$ fulfills the variational inequality*

$$\left\langle C\left(\hat{h}\right), h - \hat{h}\right\rangle \geq 0 \quad \forall h \in H. \tag{1}$$

Using the relationship $\hat{f} = \Delta \hat{h}$, (1) is equivalent to

$$\left\langle c\left(\hat{f}\right), f - \hat{f}\right\rangle \geq 0 \quad \forall f \in F. \tag{2}$$

Proof. The lemma is a simple extension of the single class result (pp. 299 - 300) in Smith [Smi79]. ■

In view of the lemma, we introduce the following notion.

Definition 2. *The link flow matrix $\hat{f} \in F$ is a (link flow multi-class) equilibrium if it satisfies the VI in (2).*

Variational inequalities are usually solved by reduction to optimization problems (or series of such). When the VI is *symmetric*, in the sense that the link specific cost functions are symmetric, i.e. when

$$\frac{\partial c_a^k(f)}{\partial f_b^l} = \frac{\partial c_b^l(f)}{\partial f_a^k} \quad \forall f \in \Re_+^{A \times K},\ a, b \in A,\ k, l \in K, \tag{3}$$

then $c(f) = \nabla I(f)$, the gradient of some differentiable *primitive* function $I : \Re_+^{A \times K} \to \Re$. (This follows e.g. from Green's/Stoke's theorem, or the Symmetry Principle, see, e.g., [OR70, p. 95]) In this case, the VI (2) says that there are no feasible descent directions for I at $\hat{f}$, a necessary condition for a local minimum of I over F (see, e.g., [Zan69, Lemma 2.11]). Note however that the VI (2) can be fulfilled also at other points, such as saddle points. Dafermos [Daf73] claims that the symmetry condition (3) is usually satisfied in real transportation networks. Netter [Net71], on the other hand, argues that, for general link travel cost functions c_a^k, condition (3) is not fulfilled in general. In sections 3 and 4 of this paper, we shall show that validity of the symmetry conditions may depend on the units in which the costs are specified.

If, in addition to c_a^k being symmetric, I is convex, then the VI (2) is equivalent to the optimization problem

$$\min I(f) \text{ s.t. } f \in F, \tag{4}$$

since in this case (2) is a necessary and sufficient condition for a global minimum of I over F. Summing up:

Proposition 1. *When c is symmetric, i.e. fulfilling (3), it has a primitive function I, such that $c(f) = \nabla I(f)$. In this case the variational inequality (2) is equivalent to the condition that there is no feasible descent direction to I at $\hat{f}$. Moreover, if I is convex, then the multi-class equilibria $f \in F$ correspond exactly to the global optima of problem (4). Uniqueness of the solution to (4) and hence uniqueness of the equilibrium is guaranteed if I is strictly convex.*

Sandholm [San02] studies single class traffic equilibria, and introduces a type of continuous time, dynamic adjustment process whereby route flow (on the average) shifts from costlier routes to cheaper routes (in the sense that $\langle C, dh/dt \rangle < 0$ unless h is an equilibrium). Such a shift is quite rational from the point of view of the users. Therefore we will call such a process a *rational adjustment process.* (In [San02], Sandholm uses the more neutral term valid.) For single class equilibrium problems with a primitive function, Sandholm

[San02] shows that such a process will converge to an equilibrium in the route flow space, and hence also in the link flow space. Such adjustment processes can in an obvious way be introduced also in the multi-class case, converging to equilibria also in this case. It is now natural to give the following definition.

Definition 3. *A multi class equilibrium f is locally stable, if any rational adjustment process started in a neighborhood of a route flow matrix $h \in H$corresponding to f, will converge to an $\bar{h}$ corresponding to f, and unstable otherwise.*

Stable equilibria are of interest because if an equilibrium is not locally stable, it can typically not be upheld, since if the route flow pattern is exposed to a small change (e.g. due to a temporary change of the traffic conditions) then users will deviate from the equilibrium (by the dynamic adjustment process). We are now in a position to state and prove

Theorem 1. *Assume that the matrix cost function c has a primitive function I. Then all locally stable multi-class equilibria are local optima to problem (4).*

Proof. If $\bar{f} = \Delta\bar{h}$ is an equilibrium which is not a local optimum to $I(f)$, then there is $f = \Delta h \in F$, in the neighborhood of $\bar{f}$ with lower objective values than $\bar{f}$. An adjustment process started in such an h cannot converge to an h^* such that $\bar{f} = \Delta h^*$, since the objective values have to decrease during the process (due to that $\frac{d}{dt}I(f) = \left\langle \nabla I, {}^{df}/_{dt} \right\rangle = \langle c, \frac{d}{dt}\Delta h\rangle = \left\langle \Delta^T c, {}^{dh}/_{dt} \right\rangle = \langle C, dh/dt\rangle < 0$). Hence, all locally stable equilibria correspond to local optima. ∎

3 Fixed-Toll Multi-Class Equilibria with Class Specific Time Values

As noted above, marginal social cost tolls typically need to be implemented as fixed tolls. Further, travelers with different time values react differently to such tolls. Therefore, in this section we specialize general multi-class equilibria to the case where the classes only differ in their time values, and where the tolls are fixed. In particular we show that the VI's characterizing equilibria can be stated in symmetric or nonsymmetric forms, hence allowing corresponding optimization formulations. We further show that, although this optimization problem is convex, the equilibrium class flows need not be unique. In spite of this the total value of travel time turns out to be unique.

Assumption 1 *Below, it is assumed that the class specific travel cost of link a for users of class k depends linearly on two components: the link toll p_a and the travel time $t_a(f_a^{tot})$, which is a positive, nondecreasing, nonconstant, and twice differentiable function of the total volume $f_a^{tot} = \sum_k f_a^k$ on the link. In particular, this linear relation is mediated through class specific time values $v_k > 0$, assumed distinct.*

Remark 1. Please note that the tolls p_a in this section could as well be any flow independent monetary cost (e.g. the gasoline cost if one assumes the gasoline consumption to be just proportional to the trip distance). Hence the derived results are still valid in this "more general" case.

Under Assumption 1, the travel disutilities can be expressed either in time or in monetary units. Thus we define the *generalized cost* $\bar{c}_a^k$ and the *generalized time* $\bar{t}_a^k$ of link a for class k respectively as

$$\bar{c}_a^k(f) = v_k t_a\left(f_a^{tot}\right) + p_a, \tag{5}$$

and

$$\bar{t}_a^k(f) = t_a\left(f_a^{tot}\right) + p_a/v_k. \tag{6}$$

Note that $\bar{c}$ and $\bar{t}$ are *separable* over the links, in the sense that $\bar{c}_a^k(f)$ and $\bar{t}_a^k(f)$ only depend on f_a. For this reason we will write $\bar{c}_a^k(f_a)$ rather than $\bar{c}_a^k(f)$ and correspondingly for $\bar{t}_a^k(f)$.

Switching between $\bar{c}_a^k$ and $\bar{t}_a^k$, one just scales all link and route costs for a given user class with the same scalar (v_k or $1/v_k$). This does not change the equilibria, since Def. 1 is scale invariant in the cost.

Definition 4. *A fixed-toll multi-class user equilibrium (with class specific time values), is a multi-class Wardrop equilibrium with link costs c_a^k equal to $\bar{c}_a^k$ or, equivalently, to $\bar{t}_a^k$.*

By Lemma 1, these equilibria are the solutions to the VI's (1) or (2), with these same link costs.

Checking symmetry of $\bar{c}$ and $\bar{t}$, we only have to check the "intra-link" version of (3), by separability. We then see, using $\partial f_a^{tot}/\partial f_a^k = 1$, that $\frac{\partial \bar{c}_a^k(f_a)}{\partial f_a^l} = v_k t_a'(f_a^{tot})$ which in general differs from $v_l t_a'(f_a^{tot}) = \frac{\partial \bar{c}_a^l(f_a)}{\partial f_a^k}$ for $k \neq l$, since $t_a'(f_a^{tot}) > 0$ for some f_a^{tot}. Thus the $\bar{c}_a^k$ do not fulfill (3). On the other hand,

$$\frac{\partial \bar{t}_a^k(f_a)}{\partial f_a^l} = t_a'(f_a^{tot}) = \frac{\partial \bar{t}_a^l(f_a)}{\partial f_a^k}$$

whence the $\bar{t}_a^k$ do fulfill (3), implying that $\bar{t}(f) = \nabla \bar{I}(f)$ for an appropriate primitive function $\bar{I}$, which can be seen to be (up to an additive constant)

$$\bar{I}(f) = \sum_{a \in A}\left[\int_0^{f_a^{tot}} t_a(u)du + \sum_{k \in K} f_a^k p_a/v_k\right]. \tag{7}$$

Since the link times t_a are assumed nondecreasing, $\bar{I}(f)$ is convex. Hence equilibrium link flows $\hat{f}_a^k$ can be obtained as solutions to the optimization problem (4) with $I = \bar{I}$. In summary we have showed

Theorem 2. *The equivalent cost functions $\bar{c}_a^k$ and $\bar{t}_a^k$ are nonsymmetric and symmetric respectively. The fixed-toll multi-class equilibria can be determined as the optima to problem (4) with convex objective $I = \bar{I}$, according to (7).*

Van Vliet et al. [VBS86] also recognize similar symmetric and nonsymmetric properties in a traffic equilibrium problem with multiple user classes. On the other hand, the transportation science literature does not seem to recognize that these properties indirectly lead to the existence of an optimization problem equivalent to a nonsymmetric VI. In particular, standard references, such as Nagurney [Nag93] and Patriksson [Pat94], often claim that a VI, $\langle F(x^*), x - x^* \rangle \geq 0$, $\forall x \in S$ (a feasible region), has an equivalent optimization problem only when F is a gradient of a function or (equivalently) when F has a symmetric Jacobian. However, the above discussion demonstrates that it is possible, in some cases, to obtain an optimization problem equivalent to a VI via a reformulation even when the Jacobian of F is nonsymmetric or when F is not a gradient of a function.

When implementing a computed set of tolls $(p_a)_{a \in A}$, uniqueness of the fixed-toll equilibrium is important, so that one does indeed achieve the solution computed. The following proposition is an extension of the well known uniqueness theorem for the single class user equilibrium (e.g. [Pat94, Thm. 2.5].)

Proposition 2. *Assume that the link times t_a are strictly increasing. Let the link flows $f, g \in F$ be two fixed-toll multi-class equilibria corresponding to the same set of tolls $(p_a)_{a \in A}$. Then*

(a) the total volume and the travel time on each link are the same in both equilibria;
(b) for each user class and each link, the generalized link time and the generalized link cost are the same in both equilibria;
(c) for each user class and each route, the generalized route cost and generalized route time are the same in both equilibria.

Proof. Since the link travel times t_a are increasing, the objective (7) of problem (4) is strictly convex with respect to the total link volumes. Hence the total link volumes are unique, whence the link travel times are unique too. This proves (a). Assertions (b) and (c) follow because generalized link times and costs as well as generalized route times and costs only depend on the link travel times, the (fixed) tolls and the class specific time values. ■

Note, however, that the solution need not be unique in the terms of the class specific link volumes (f_a^k), e.g. if there are two routes (between the same nodes) with the same sum of tolls. Indeed, if both routes are used by two different classes in an equilibrium (whence the route times must be equal too), then part of the users of the first class can be moved from one route to the other and exchanged for users of the other class. The new flow pattern obviously is an equilibrium too, since the total link flows and hence the link times are not changed. This non-uniqueness implies that it is possible that implementation of a computed equilibrium may lead to other solutions than the computed one. The next result shows that the situation is still well behaved, though, in that the *total value of travel time* ,

$$V(f) = \sum_a \sum_k v_k t_a(f_a^{tot}) f_a^k \quad ,$$

is unique. Let us further introduce the *total generalized cost* in the network, $G(f) = \sum_a \sum_k \bar{c}_a^k f_a^k = \sum_a \sum_k v_k t_a f_a^k + \sum_a \sum_k p_a f_a^k$, and $P(f) = \sum_a \sum_k p_a f_a^k$, the *total toll revenue.* Then, obviously,

$$G(f) = V(f) + P(f). \tag{8}$$

Proposition 3. *Assume that the link travel times are strictly increasing. Consider a fixed-toll multi-class equilibrium for a given set of tolls. Then the total value of travel time is unique, i.e. $V(f) = V(g)$ for any two distinct equilibria f and g.*

Proof. By Prop. 2(a), the total toll revenue, $P(f)$, is unique. Hence, by (8), $V(f)$ is unique if $G(f)$ is. On the other hand, the total generalized cost for class k can be expressed in route flows instead of link flows. Using Def. 1 we thus get $G(f) = \sum_{a,k} \bar{c}_a^k f_a^k = \sum_{w,k} \sum_{r \in R_w} \bar{C}_r^k h_r^k = \sum_{w,k} \sum_{r \in R_w} \pi_w^k h_r^k = \sum_{w,k} \pi_w^k q_w^k$ where π_w^k is the minimal generalized class k cost for routes connecting OD pair w. It follows from Prop. 2(c) that the π_w^k are the same in both equilibria. Thus the total generalized cost $G(f)$ is unique and the proposition is proved. ■

The possibility of nonuniqueness of multi-class equilibria, when the link costs depend only on the total flows, was noted by Toint and Wynter [TW96], in observing that the Jacobian $(\partial c_a^k / \partial f_a^l)$ in a is singular. Toint and Wynter considered this to be a problematic property which should be avoided for link cost functions. In view of Proposition 3, we consider this nonuniqueness no more problematic in our case than the standard nonuniqueness of route flows in single class equilibrium problems. Further discussions of nonuniqueness of multi-class equilibria may be found in Konishi [Kon04].

We will finally discuss the continuity of the total link flows (of fixed-toll equilibria) with respect to the tolls. This is an interesting property in its own right, but it will also be instrumental in proving that one can through fixed tolls (at least approximately) achieve the same levels of total travel time as through flow dependent tolls (Thm 5, below).

First note that since the total equilibrium link flows $f^{tot} = (f_a^{tot})$ are unique for given tolls p, f^{tot} is a *function* $\hat{f}^{tot}(p)$ of the tolls p.

Proposition 4. *Assume that the link travel times are strictly increasing. Then the total equilibrium link flows $\hat{f}^{tot}(p)$ is a continuous function of the tolls p.*

Proof. The proof will be by contradiction. Thus assume that there is a set of tolls $\bar{p}$ and a sequence $(p^{(n)})$ converging to $\bar{p}$, such that $f^{tot,(n)} = \hat{f}^{tot}(p^{(n)})$ does not converge to $\bar{f}^{tot} = \hat{f}^{tot}(\bar{p})$. By compactness of F we may assume that the class specific link flows $f^{(n)} = (f_a^{k,(n)})$, corresponding to $f^{tot,(n)}$, converge to some $\tilde{f} = (\tilde{f}_a^k)$ with $\tilde{f}^{tot} \neq \bar{f}^{tot}$. Let $\bar{c}^{(n)} = (v_k t_a(f_a^{tot,(n)}) + p_a^{(n)})$ be the user specific link costs corresponding to $f^{(n)}$. Since $f^{(n)}$ is an equilibrium for the tolls $p^{(n)}$, it fulfills the VI $\langle \bar{c}^{(n)}, f - f^{(n)} \rangle \geq 0, \forall f \in F$.

But $f^{(n)}$converges to $\tilde{f}$ and $\bar{c}^{(n)}$ converges to $\bar{\bar{c}} = (v_k t_a(\tilde{f}^{tot}) + \bar{p}_a)$. Thus by continuity, $\tilde{f}$ fulfills the VI $\left\langle \bar{\bar{c}}, f - \tilde{f} \right\rangle \geq 0, \forall f \in F$, whence $\tilde{f}$ is an equilibrium for $\bar{p}$. Hence $\tilde{f}^{tot}$and $\bar{f}^{tot}$ are different equilibrium link flows for $\bar{p}$, contradicting uniqueness. Thus the introductory assumption is false, and $\hat{f}^{tot}(p)$ is a continuous function of p. ■

Remark 2. Please note that we only used the uniqueness of $\hat{f}^{tot}(p)$ plus the standard properties of compactness of F, continuous dependence of $\bar{c}$ on its parts and the continuity of the inner product. Hence, the proof will go through in other similar cases.

As a corollary we have

Theorem 3. *For the fixed-toll multi-class equilibrium problem the total equilibrium link flows $\hat{f}^{tot}(p)$ as well as the total (equilibrium) value of travel time, $V(\hat{f}^{tot}(p))$, the total (equilibrium) generalized cost, $G(\hat{f}^{tot}(p))$, and the total (equilibrium) toll revenue, $P(\hat{f}^{tot}(p))$, depend continuously on the tolls p.*

Proof. The continuity of $\hat{f}^{tot}(p)$was proved already in Prop. 3. From this follows that the equilibrium link times are continuous functions of p. Thus also the generalized costs $\bar{c}$ are continuous, whence the same is true for the minimal generalized route costs, π_w^k, and hence also for the total generalized cost $G(f) = \sum_{w \in W} \sum_{r \in R_w} \pi_w^k q_w^k$.

The continuity of the total toll revenue, P, follows directly from that of $\hat{f}^{tot}(p)$. Finally, since $V = G - P$, the continuity of V follows. ■

4 Tolls based on marginal social costs

In this section we first look at flow dependent marginal social cost tolls. Again the VI characterizing equilibria can be stated in symmetric or non-symmetric forms. The symmetric one corresponds to an optimization problem, where the objective is the total value of travel time, later shown to be nonconvex in general (Section 5). Then we look at the implementation of these tolls as fixed tolls. We show that the flow dependent equilibria are indeed equilibria to the corresponding fixed-toll problem, which however may also have other equilibria. All these equilibria, however have the same total value of travel time.

Thus, we now first consider a situation where the tolls are based on marginal congestion costs, or *marginal social costs* (MSC). Internalizing the external congestion costs, the authorities make the users pay for the delays they inflict on other users and are interested in the traffic volumes that are established in the network and the corresponding toll values. A marginal user inflicts a delay $t'_a\left(f_a^{tot}\right)$ on all other users on link a. However, the monetary value of this delay is different for the different users and equal to $v_m t'_a\left(f_a^{tot}\right)$ for users belonging to class m. The flow-dependent *MSC toll* then is a sum of all delay values for the users of the link caused by a marginal user, i.e.

$$p_a = p_a(f) = t'_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m. \tag{9}$$

Substituting (9) into (5) or (6) gives the *MSC link costs*

$$\tilde{c}_a^k\left(f_a\right) = v_k t_a\left(f_a^{tot}\right) + t'_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m, \tag{10}$$

or the *MSC link times*

$$\tilde{t}_a^k\left(f_a\right) = t_a\left(f_a^{tot}\right) + t'_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m / v_k. \tag{11}$$

Definition 5. *A multi-class MSC equilibrium (with class specific time values), is a multi-class Wardrop equilibrium with link costs c_a^k equal to $\tilde{c}_a^k$ or, equivalently, to $\tilde{t}_a^k$.*

By Lemma 1, these equilibria are the solutions to the VI's (1) or (2), with these same link costs.

For the fixed-toll multi-class equilibrium problem, generalized time $\bar{t}_a^k$ was symmetric. Differentiating the MSC link times (11) with respect to flow variables yields $\frac{\partial \tilde{t}_a^k(f_a)}{\partial f_a^l} = t'_a\left(f_a^{tot}\right) + \frac{1}{v_k} t''_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m + \frac{1}{v_k} t'_a\left(f_a^{tot}\right) v_l$, which is different from $\frac{\partial \tilde{t}_a^l(f_a)}{\partial f_a^k}$ in general.

(To see this in more detail, note that equality holds if and only if

$$\begin{aligned} 0 = {} & \frac{1}{v_k}\left[t''_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m + t'_a\left(f_a^{tot}\right) v_l\right] \\ & - \frac{1}{v_l}\left[t''_a\left(f_a^{tot}\right) \sum_{m \in K} v_m f_a^m + t'_a\left(f_a^{tot}\right) v_k\right], \end{aligned}$$

or

$$
\begin{aligned}
0 &= v_l t''_a\left(f_a^{tot}\right) \sum_{m\in K} v_m f_a^m + t'_a\left(f_a^{tot}\right) v_l^2 \\
&\quad - \left[v_k t''_a\left(f_a^{tot}\right) \sum_{m\in K} v_m f_a^m + t'_a\left(f_a^{tot}\right) v_k^2 \right] \\
&= (v_l - v_k) t''_a\left(f_a^{tot}\right) \sum_{m\in K} v_m f_a^m + (v_l^2 - v_k^2) t'_a\left(f_a^{tot}\right),
\end{aligned}
$$

or

$$
0 = t''_a\left(f_a^{tot}\right) \sum_{m\in K} v_m f_a^m + (v_l + v_k) t'_a\left(f_a^{tot}\right),
$$

which cannot hold for all f_a^m summing up to f_a^{tot}.)

Hence the symmetry condition (3) does not hold, which precludes direct application of optimization methods to equilibrium search based on $\tilde{t}_a^k$. However, differentiating the MSC costs gives

$$
\frac{\partial \tilde{c}_a^k\left(f_a\right)}{\partial f_a^l} = v_k t'_a\left(f_a^{tot}\right) + t''\left(f_a^{tot}\right) \sum_{n\in K} v_n f_a^n + t'_a\left(f_a^{tot}\right) v_l = \frac{\partial \tilde{c}_a^l\left(f_a\right)}{\partial f_a^k}
$$

Hence, $\tilde{c}_a^k$ is symmetric, and there is a primitive function $\tilde{I}(f)$, such that $\nabla \tilde{I}\left(f\right) = \tilde{c}\left(f\right)$. It is easily checked that (up to a constant)

$$
\tilde{I}\left(f\right) = \sum_{a\in A} t_a\left(f_a^{tot}\right) \sum_{k\in K} v_k f_a^k \tag{12}
$$

Note that $\tilde{I}(f) = V(f)$, the total value of travel time in the network. As explained in the introduction, minimization of $V(f)$ corresponds to the most efficient usage of the road network.

Summing up, using Prop. 1, we have the following result.

Proposition 5. *The equivalent cost functions $\tilde{c}_a^k$ and $\tilde{t}_a^k$ are symmetric and non-symmetric respectively. The MSC multi-class equilibria are flow matrices $f \in F$ where the total value of travel time $V(f)$ has no feasible descent directions.*

$V(f)$ is in general non-convex and the VI (2) can have multiple solutions (see section 5). As noted before, all local minima of $V(f)$ (and maybe also some other points) in F are MSC multi-class equilibria.

Theoretically, one can distinguish between the three kinds of equilibria: *global minima* of $V(f)$ on F, *local minima* that are not global minima, and *other equilibria.* From the application point of view, the most interesting equilibria are the ones that minimize $V(f)$ globally on F. However, there are no efficient methods for finding global minima of general non-convex functions. Various iterative descent methods can be used for finding local minima. The quality of achieved minima depends however, on the starting points of the algorithm. Equilibria of the third kind are of little interest. Since they are not local minima, there are points with better objective values $V(f)$ arbitrary close to them, and starting an iterative descent in such a neighbor, will give a local optimum of lower objective value. By Thm. 1 such equilibria are further unstable.

MSC tolls typically need to be implemented as fixed tolls. The theorem below, shows that the multi-class MSC equilibrium flows are fixed-toll equilibria with the computed MSC tolls as fixed tolls. But, as mentioned before, the equilibrium with fixed tolls is not necessarily unique. Therefore, implementing MSC tolls as fixed tolls, the resulting fixed-toll equilibrium need not coincide with the MSC equilibrium. Prop. 3, however, saves the day.

Theorem 4. *Assume that the travel time functions t_a are strictly increasing. Let $\hat{f}$ be a multi-class MSC equilibrium, and let $\hat{p} = (\hat{p}_a) = (p_a(\hat{f}))$, defined by (9), be the corresponding vector of link tolls. Then $\hat{f}$ is also a fixed-toll multi-class equilibrium for fixed tolls $p = \hat{p}$.*

If $\tilde{f}$is another fixed-toll equilibrium for tolls $p = \hat{p}$, then $V(\hat{f}) = V(\tilde{f})$, i.e. the total value of travel time is unique.

Proof. Being an MSC equilibrium, $\hat{f}$ fulfills (by Lemma 1) the VI (2) with $c_a^k = \tilde{c}_a^k$, i.e. $\forall g \in F$,

$$\sum_{a\in A}\sum_{k\in K}\left[t_a\left(\hat{f}_a^{tot}\right) + p_a(\hat{f})\Big/v_k\right]\left(g_a^k - \hat{f}_a^k\right) \geq 0.$$

But since $\hat{p}_a = p_a(\hat{f})$, $\hat{f}$ also fulfills the VI $\forall g \in F$,

$$\sum_{a\in A}\sum_{k\in K}\left[t_a\left(\hat{f}_a^{tot}\right) + \hat{p}_a/v_k\right]\left(g_a^k - \hat{f}_a^k\right) \geq 0,$$

implying that $\hat{f}$is a fixed-toll multi-class equilibrium for fixed tolls $p = \hat{p}$.

If $\tilde{f}$ is another fixed toll equilibrium for $p = \hat{p}$, it follows from Prop. 3 that $V(\tilde{f}) = V(\hat{f})$. ■

To clarify the above discussion, it might be illuminating to consider the following multi-class problems studied in sections 3 and 4, explicitly or implicitly.

(P_1): determination of fixed-toll equilibria,

(P_2): determination of MSC equilibria, i.e. equilibria under flow dependent MSC tolls, and

(P_3): finding $f \in F$ minimizing the total cost of travel time, $V(f)$.

Further as a combination of (P_1) and (P_3) we may consider

(P_4): determining fixed tolls p, minimizing the total equilibrium cost of travel time $V(\hat{f}(p))$ over all fixed-toll equilibria $\hat{f}(p)$.

Of these four problems (P_4) is the most important from an application viewpoint.

In section 3 we showed how to solve (P_1). Theorems 4 and 1 show that problems (P_2)-(P_4) are in fact equivalent if we restrict our interest to locally stable equilibria and are content with local minima. They are all solved by minimizing $V(f)$ over F. In particular the minimal value of $V(\hat{f}(p))$ over fixed-toll equilibria is the same as the minimal value of $V(f)$ over all of F.

The nonuniqeness displayed above also shows that it is not altogether trivial, that implementing MSC tolls, as fixed tolls, will give flows that minimize total value of travel time. Yang and Huang [YH04] show that the minimum of V is also a fixed-toll equilibrium. This is however only *necessary* for being able to implement the equilibrium through fixed tolls, since there are also other equilibria to the fixed-toll problem. Theorem 4 proves the *sufficiency*, namely that all such equilibria have the same (minimal) value for V.

There are further problems with implementing MSC tolls. Computing the MSC tolls, e.g. by applying the Frank-Wolfe method to problem (4) with objective $I(f) = V(f)$, one will never arrive at the equilibrium. Thus, one will have to implement *close-to-equilibrium* tolls as fixed tolls. We know that equilibrium tolls, implemented as fixed tolls, will give fixed-toll equilibria with the same total value of travel time as the MSC equilibrium (Thm. 4). When implementing close-to-equilibrium tolls the situation is not a priori obvious, though.

Theorem 5. *Let the functions t_a be strictly increasing and $f^{(n)} = (f_a^{k,(n)})$ be a sequence of multi-class flow matrices, converging to an MSC-equilibrium $\bar{f}$. Let $f^{tot,(n)}$ be the corresponding total link flows, and $p^{(n)} = \left(p_a^{(n)}\right) = \left(t_a'(f_a^{tot,(n)}) \sum_k v_k f_a^{k,(n)}\right)$ the corresponding MSC tolls. Further let $\hat{f}^{tot}(p^{(n)})$ be the (unique) fixed-toll equilibrium link flows corresponding to $p^{(n)}$, and $V^{(n)}$ the corresponding unique total values of travel time. Then $V^{(n)}$ converges to $V(\bar{f})$.*

Proof. Since V is continuous, $V(f^{(n)}) \to V(\bar{f})$. Further, $p^{(n)}$ converges to $\bar{p} = \left(p_a(\bar{f})\right) = \left(t'(\bar{f}_a^{tot}) \sum_k v_k \bar{f}_a^k\right)$ by continuity of $p_a(f)$, see (9). Since the achieved V in the fixed-toll case depends continuously on p (by Thm 3), $V^{(n)}$ converges to $\tilde{V}$, the total value of travel time for fixed tolls $\bar{p}$. But, by Thm 4, the values of V agree in the MSC and the fixed-toll problem, i.e. $\tilde{V} = V(\bar{f})$. Thus $V^{(n)} \to V(\bar{f})$. ■

The theorem says that one is justified in implementing close-to-equilibrium MSC tolls as fixed tolls, but it does not tell how close one needs to be. For that, a more elaborate analysis probably is needed.

5 Nonconvexity of V

In this section, we will show that the MSC objective $V(f)$ in general is nonconvex. We will however start with a small illuminating example that will be instrumental in showing non-convexity.

Consider a network consisting of a single OD pair w connected by two links a and b with identical travel time functions $t_a(u) = t_b(u) = u$ (Figure 1). Assume there are two user classes with time values $v_1 = 1$ and $v_2 = 5$, respectively, and with travel demands $q_w^1 = q_w^2 = 100$.

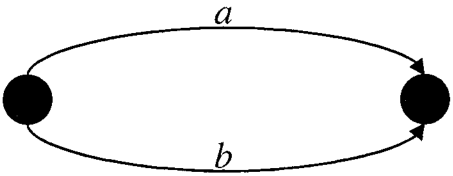


Fig. 1. The example network

The feasible set is seen to be

$$F = \left\{f = \left(f_a^1,\, f_a^2,\, f_b^1,\, f_b^2\right) \in \Re_+^4 : f_a^1 + f_b^1 = f_a^2 + f_b^2 = 100\right\}.$$

Introducing the *independent variables* $f_a = \left(f_a^1, f_a^2\right)$, F can be more compactly described as $F = \left\{f \in \Re^4 : f_a^k \in [0, 100], f_b^k = 100 - f_a^k, k = 1, 2\right\}$.

Without tolls, there is a continuous set of user equilibria

$$\hat{F} = \left\{f \in F :\ f_a^2 = 100 - f_a^1\right\}.$$

Note that the total volume and hence the travel time on each link is constant across $\hat{F}$. Considering these equilibria as fixed-toll equilibria (with toll 0) this is in line with Prop. 2.

Introduction of MSC pricing leads to the MSC objective

$$V(f) = \left(f_a^1 + f_a^2\right)\left(f_a^1 + 5f_a^2\right) + \left(f_b^1 + f_b^2\right)\left(f_b^1 + 5f_b^2\right),$$

or, in terms of the independent variables

$$V(f) = \left(f_a^1 + f_a^2\right)\left(f_a^1 + 5f_a^2\right) + \left(200 - f_a^1 - f_a^2\right)\left(600 - f_a^1 - 5f_a^2\right).$$

In Fig. 2 we display the level curves and negative gradient directions of V as functions of the independent variables. We see that there are three equilibria: first two equilibria corresponding to local (and global) minima, $f_a^{(1)} = (0,\ 80)$ and $f_a^{(2)} = (100,\ 20)$, both with objective value 56000, and with corresponding MSC tolls $p^{(1)} = (400,\ 200)$ and $p^{(2)} = (200,\ 400)$, respectively; finally one corresponding to a saddle point, $f^{(3)} = (50,\ 50)$ with toll $p^{(3)} = (300,\ 300)$ and objective 60000.

When the tolls $p^{(1)}$ or $p^{(2)}$ are enforced as fixed tolls, the only existing user equilibria are $f_a^{(1)}$ and $f_a^{(2)}$ respectively. Implementation of the tolls $p^{(3)}$, however, does not affect the route choice, whence there is the same set of equilibria $\hat{F}$ as in the situation without tolls. Thus an equilibrium flow pattern with fixed equilibrium tolls $p^{(3)}$ need not coincide with $f^{(3)}$. However, in line with Proposition 3 and Theorem 4, these flow patterns are equivalent both from the individual and the social points of view, since total flow and travel time along each link, and the total value of travel time and toll revenues at any point in $\hat{F}$ is the same as at $f^{(3)}$.

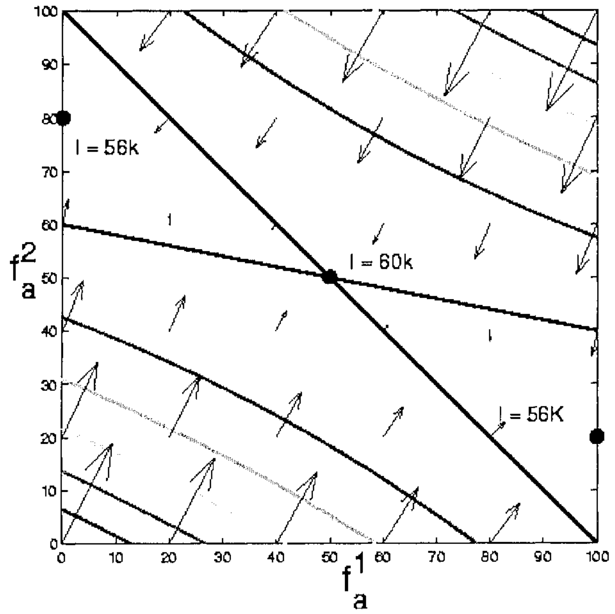


Fig. 2. Equilibria (•), level curves and negative gradients of the function V

This example shows that $V(f)$ is not convex in general. The example may seem very specialized, but the two links could represent two routes between two nodes in a network with at least two user classes. In this form it can be used to show that $V(f)$ is in general non-convex.

In the general setting, we have several (≥ 2) user classes, differing only in their time values. Road networks (and demands) moreover typically have the following property: There exist two nodes n_1 and n_2 connected by two link-disjoint paths $p_j, j = 1, 2$, such that each p_j is a subpath of two routes $r_j^k \in R, k = 1, 2$, and there are two classes, $k = 1, 2$, say, such that for a given k, $r_j^k, j = 1, 2$, connect the same OD-pair w_k with class k demand $q_{w_k}^k > 0$. Let us call such networks *multi-route*, multi-class networks. In particular, if there is an OD-pair w with at least two routes in R_w, and with positive demand for at least two classes, we have a multi-route multi-class network.

Theorem 6. *Consider a multi-route multi-class network with strictly increasing travel time functions t_a. Then the objective $V(f)$ in the tolled MSC equilibrium problem is nonconvex (on the feasible set).*

Proof. Let us use the notations of the definition of a multi-route multi-class network. Consider a feasible matrix of route flows, where $\bar{h}_j^k > 0$ is the class k route flow along r_j^k $(j = 1, 2; k = 1, 2)$, existing according to the definition. Denote $\bar{h}^k = \bar{h}_1^k + \bar{h}_2^k$, and let $(h_j^k)_{j=1,2}^{k=1,2}$ be new route flows along r_j^k varying over positive values so that $h_1^k + h_2^k = \bar{h}^k$, but keep all other route flows fixed. In this way we get feasible route (and hence link) flows.

We will show that $V(\Delta h)$ is nonconvex as a function of $(h_j^k)_{j=1,2}^{k=1,2}$. As in the example, we can view $p_j, j = 1, 2$ as two links. Let $h_1 = (h_1^k)_{k=1,2}$, be the independent variables, and $h_2 = (h_2^k)_{k=1,2}$, the "dependent" variables, thus being a linear function of h_1.

When varying h_1 (and hence h_2) the contribution to $V(\Delta h)$ from links not in $\{p_j\}_{j=1,2}$ is constant, thus giving no contribution to the hessian.

Let $h_j^{tot} = h_j^1 + h_j^2$, and for an a in p_j let $\bar{f}_a^{tot}$ be the sum of the route flows in a other than h_j^k. Thus $f_a^{tot} = h_j^{tot} + \bar{f}_a^{tot}$.

Let $\bar{V}(h_1)$ be the nonconstant part of $V(\Delta h)$ as a function of h_1 (i.e. excluding the constant terms mentioned above). Then,

$\bar{V}(h_1) = \sum_{j=1,2} \sum_{a \in p_j} t_a(h_j^{tot} + \bar{f}_a^{tot})(v_1 h_j^1 + v_2 h_j^2 + \bar{v}_a \bar{f}_a^{tot})$, where $\bar{v}_a$ is the mean time value of the route flows in $\bar{f}_a^{tot}$. Thus

$\frac{\partial \bar{V}}{\partial h_1^k} = \sum_{a \in p_1} [t'_a(h_1^{tot} + \bar{f}_a^{tot})(v_1 h_1^1 + v_2 h_1^2 + \bar{v}_a \bar{f}_a^{tot}) + t_a(h_1^{tot} + \bar{f}_a^{tot}) v_k] -$
$\sum_{a \in p_2} [t'_a(h_2^{tot} + \bar{f}_a^{tot})(v_1 h_2^1 + v_2 h_2^2 + \bar{v}_a \bar{f}_a^{tot}) + t_a(h_2^{tot} + \bar{f}_a^{tot}) v_k]$,

$\frac{\partial^2 \bar{V}}{\partial (h_1^k)^2} = \sum_{j=1,2} \sum_{a \in p_j} [t''_a(h_j^{tot} + \bar{f}_a^{tot})(v_1 h_j^1 + v_2 h_j^2 + \bar{v}_a \bar{f}_a^{tot}) +$
$2t'_a(h_j^{tot} + \bar{f}_a^{tot}) v_k]$, and

$\frac{\partial^2 \bar{V}}{\partial h_1^1 \partial h_1^2} = \sum_{j=1,2} \sum_{a \in p_j} [t''_a(h_j^{tot} + \bar{f}_a^{tot})(v_1 h_j^1 + v_2 h_j^2 + \bar{v}_a \bar{f}_a^{tot}) +$
$t'_a(h_j^{tot} + \bar{f}_a^{tot})(v_1 + v_2)]$.

Using $A = \sum_{j=1,2} \sum_{a \in p_j} t''_a(h_j^{tot} + \bar{f}_a^{tot})(v_1 h_j^1 + v_2 h_j^2 + \bar{v}_a \bar{f}_a^{tot})$
and $B = \sum_{j=1,2} \sum_{a \in p_j} t'_a(h_j^{tot} + \bar{f}_a^{tot})$, positive by assumption, we get

$$\nabla^2 \bar{V} = \begin{pmatrix} A + 2Bv_1 & A + B(v_1 + v_2) \\ A + B(v_1 + v_2) & A + 2Bv_2 \end{pmatrix}.$$

An easy check gives that $\det(\nabla \bar{V}) = -B^2(v_1 + v_2)^2 < 0$. Thus $\bar{V}$ and hence V are nonconvex. ■

This theorem shows that the MSC toll problem is in general nonconvex except for very special networks. This resolves the question, raised in Dial [Dia99a], whether $V(f)$ is convex or not in general, and refutes the statement in Yang and Huang [YH04] that $V(f)$ is convex.

6 A Frank-Wolfe algorithm for the multi-class MSC equilibria

As noted in section 4, the equilibria in the case with flow-dependent MSC tolls, can be determined by solving the optimization problem (4), with the objective V of (12). To this end one can use (an adaptation of) the Frank-Wolfe method.

To streamline the algorithm, let $\omega_a(f_a) = \sum_k v_k f_a^k$ be the *total "flow value"* in link a. Note that the MSC link cost can be written

$$\tilde{c}_a^k = v_k t_a(f_a^{tot}) + t_a'(f_a^{tot})\omega_a(f_a), \tag{13}$$

and that the objective can be written $V(f) = \sum_a t_a(f_a^{tot})\omega_a(f_a)$.

Also note that both $f^{tot} = (f_a^{tot})$ and $\omega = (\omega_a)$ vary linearly with $f = (f_a^k)$. Thus, instead of storing the whole vector $f_a = (f_a^k)_{k \in K}$ for each link a it is enough to store $x_a = (f_a^{tot}, \omega_a(f_a))$ to be able to compute the link costs for the different classes. This may be important when there are many user classes.

In analogy to the standard single class case, linearizing the objective V, problem (4) decomposes into independent shortest path problems, one for each OD pair and class, and the extreme point solution to the linearized problem is composed of the all-or-nothing solutions corresponding to these shortest paths. The detailed implementation of this is straightforward.

In the same way that the classical Frank-Wolfe method can be shown to converge to a global optimum for a convex problem, this version can be shown to converge to a solution to the VI (2) (see, e.g., [Zan69, p 158-162]).

7 Some experimental results

We have applied the methods and results of the current paper to 3 test problems: the two link network (presented in section 5), the classical Sioux Falls network and the large Stockholm network.

7.1 The two link network

The algorithm has first been applied to the two link network with two user classes described in the example of Section 5, although with a quadratic volume delay function $t_a(f) = t_b(f) = 1 + f^2$, and time values $v_1 = 1$ and $v_2 = 2$. Qualitatively, the location of equilibria and their properties are the same as in the example. Due to the symmetric network structure with two identical links, the algorithm, when started under free flow conditions, quickly reaches the saddle point equilibrium and gets stuck there. This behavior, though improbable for real networks, suggests that it may be worthwhile, after obtaining an equilibrium, to make a short step in a random direction and make additional iterations to see if the process converges to the same equilibrium.

When the algorithm was started from another feasible link volume matrix, it converged to one of the local optima, although experiencing a lot of zigzagging.

7.2 Sioux Falls

In this network, we used three user classes, with time values from the Stockholm case, $v = (.98, 3.30, .19)$. We also set the class fractions of demand for each OD-pair equal to the Stockholm fractions (.754, .036, .210). See [ELD03] for more details of the experiments.

Note, that since the problem is nonconvex, we do not get an underestimate of the optimal value, when we solve the linearized problem. Instead, we have to stop the iterations when the improvement gets to too small.

Starting in the free flow solution, and performing the large number of iterations (N=10000) that would give a relative error of 10^{-6} in the classical single class case, we get an objective value of $V(f) = 71.09$, compared to the untolled value $V(f) = 74.80$, i.e. a decrease of 5%. The iteration history of the "relative error" $(V(f^{(i)}) - V(f^{(N)}))/V(f^{(N)})$ versus i becomes approximately linear in a log-log diagram, similar to the single class case, showing that convergence is comparable to that case (see [ELD03]).

To test for the existence of multiple local optima, we started at 10 random extreme point solutions. For iteration counts that would give relative errors of 10^{-3} in the free flow run these runs all gave relative errors of the same magnitude (assuming that the previous long run gave the optimum) This indicates that there is only one local optimum (see [ELD03]) conforming with the observations in Dial [Dia99b].

7.3 Stockholm

To apply the algorithm to the Stockholm case (1250 centroids, 4635 regular nodes and 18044 links), it has been implemented as a macro in EMME/2. In the initial iterations of the algorithm, we minimize the convex hull, $convV$, of V, rather than V itself (see [LE04]). This approach on the one hand provides lower bounds for V, which we do not get from the linearizations in the Frank–Wolfe algorithm, due to the nonconvexity of V; on the other hand it speeds up the initial convergence (Figure 3).

As can be seen in Figure 4, satisfactory link flow differences between consecutive iterations (i.e. lower than 100 veh./h) are obtained after approximately 50 iterations of the Frank-Wolfe algorithm. This is a substantial progress compared to the method of successive averages used in Inregia's study (see Section 1).

To check the uniqueness of the MSC equilibrium, ten initial flow patterns have been generated as random convex combinations (with exponentially distributed weights) of fifteen different extreme solutions to the multi-class assignment problem. Starting from each initial pattern, 80 iterations of the

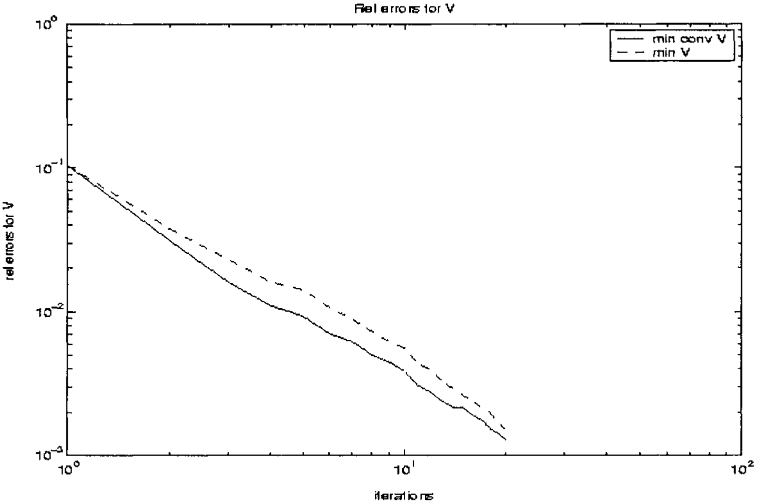


Fig. 3. Rel. errors for $V(f)$ (Stockholm), when minimizing *convV* (solid) or V (dashed)

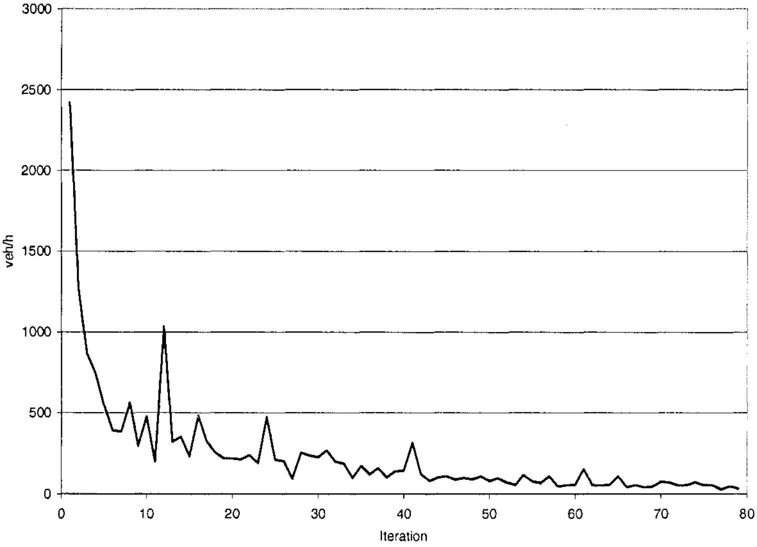


Fig. 4. Convergence of link volumes (Stockholm). Horizontal axis: iteration number. Vertical axis: maximal absolute difference of total link flows between consecutive iterations

Frank-Wolfe method have been performed. As might be seen from Figure 5, the interseries standard deviation diminishes approximately as the iteration number to the power −0.75. Thus minimizations starting from different initial flows result, after a large number of steps, in essentially the same link flow pattern. This suggests that there is only one local minimum and conforms to observations above and in Dial [Dia99b].

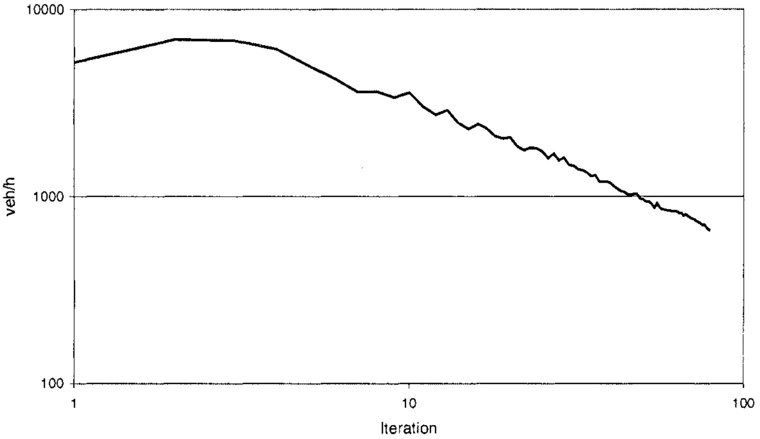


Fig. 5. Uniqueness check. Horizontal axis (log scale): iteration number (i)

Vertical axis (log scale): Interseries standard deviation $\sigma(i) = \sqrt{\frac{1}{10}\sum_{s=1}^{10}\sum_{a\in A}\sum_{k\in K}\left[f_a^k(s,i) - \overline{f}_a^k(i)\right]^2}$ where $\overline{f}_a^k(i) = \frac{1}{10}\sum_{s=1}^{10} f_a^k(s,i)$ and $f_a^k(s,i)$ is the flow of class k on link a at iteration i of series s.

8 Concluding remarks

In this paper we have studied tolled multi-class traffic equilibria. In particular we have pointed at some problematic points (concerning symmetry) in stating the equilibrium problems, in the non-uniqueness of their solutions, and in the implementation of computed MSC equilibria through fixed tolls, as well as suggested some solutions. In our opinion, the main contributions of this paper are the following:

It elucidates that some asymmetric variational inequalities may be restated in a symmetric form, and hence have a corresponding optimization formulation, contrary to their first appearance. This is in particular true for fixed-toll multi-class equilibria and for MSC-toll equilibria.

It clarifies the relation between the (flow dependent) MSC-tolls and their implementation as fixed tolls in a multi-class setting. In particular, it shows that the matrix of class specific flows at a stable MSC-equilibrium, which also is a local minimum to the total value of travel time, is an equilibrium for the corresponding fixed-toll problem. Further, in spite of that this latter equilibrium is not unique, the total value of travel time is. It moreover demonstrates that implementing close-to-optimal MSC-tolls as fixed-toll equilibria, will lead to close-to-optimal fixed-toll equilibria.

The paper further shows that the total value of travel time of heterogeneous users in general is nonconvex, settling a question raised by Dial [Dia99a], and disproving a claim made by Yang and Huang [YH04].

Acknowledgments. This research was partially supported by the Swedish Agency for Innovation Systems (VINNOVA), grant 2001-03833.

The authors thank M. Daneva, Department of Mathematics, Linköping University, for the computations for the Sioux Falls case.

References

[BMW56] Beckmann, M., McGuire, C.B., Winsten, C.B.: Studies in the Economics of Transportation. Yale University Press, New Haven (1956)

[Daf73] Dafermos, S.C.: Toll Patterns for Multiclass-User Transportation Networks. Transportation Science, **7**, 211–223 (1973)

[Dia99a] Dial, R.B.: Network-optimized road pricing: Part I: A parable and a model. Operations Research, **47**, 54–64 (1999)

[Dia99b] Dial, R.B.: Network-optimized road pricing: Part II: Algorithms and examples. Operations Research, **47**, 327–336 (1999b)

[Eli00] Eliasson, J.: The use of average values of time in road pricing. A note on a common misconception. In: Eliasson, J.: Transport and location analysis. Dissertation, Dept. of Infrastructure and Planning, Royal Institute of Technology, Stockholm (2000)

[ELD03] Engelson, L., Lindberg, P.O., Daneva, M.: Multi-Class User Equilibria under Social Marginal Cost Pricing. Operations Research Proceedings, 174–179, Springer (2003)

[EL02] Engelson, L., Lindberg, P.O.: Congestion Pricing of Road Networks with Users having Different Time Values. Technical Report, LiTH-MAT-R-2002-10, revised 2004-06-30 (2002)

[EC01] European Commission: White Paper. European Transport Policy for 2010: Time to Decide. European Communities, Brussels (2001)

[HR98] Hearn, D.W., Ramana, M.V.: Solving Congestion Toll Pricing Models. In: Marcotte, P., Nguyen, S. (eds) Equilibrium and Advanced Transportation Modeling. Kluwer Academic Publishers, 109–124 (1998)

[HY02] Hearn, D.W., Yildirim, M.B.: A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand. In: Gendreau, M., Marcotte, P. (eds) Transportation and Network Analysis: Current Trends. Kluwer Academic Publishers, 135-145 (2002)

[Ing01] Inregia: Case Study: Österleden. A basis for planning of transport systems in cities. Stockholm, (In Swedish) (2001)

[Kon04] Konishi, H.: Uniqueness of User Equilibrium in Transportation Networks with Heterogeneous Commuters. to appear in Transp. Sc. (2004)

[Lin05] Lindberg, P.O. A note on two papers by Dial, forthcoming (2005)

[LE04] Lindberg, P.O., Engelson, L.: Convexification of the Traffic Equilibrium Problem with Social Marginal Cost Tolls. Operations Research Proceedings 2003, Springer, Berlin, 141–148 (2004)

[Nag93] Nagurney, A.: Network Economics: A Variational Inequality Approach. Kluwer, Boston (1993)

[Net71] Netter, M.: Equilibrium and Marginal Cost Pricing on a Road Network with Several Traffic Flow Types. In: Newell, G.F. (ed.) Proceedings of the 5^{th} International Symposium on the Theory of Traffic Flow and Transportation. Elsevier, 155–163 (1971)

[OR70] Ortega, J.M., Rheinboldt, W.C.: Iterative solution of nonlinear equations in several variables. Academic Press (1970)

[Pat94] Patriksson, M.: The Traffic Assignment Problem: Models and Methods. VSP, Utrecht (1994)

[San02] Sandholm, W.H.: Evolutionary Implementation and Congestion Pricing. Review of Economic Studies, **69**, 667–689 (2002)

[Smi79] Smith, M.J.: Existence, uniqueness, and stability of traffic equilibria, Transp. Res., **13B**, 259–304 (1979)

[SG98] Swedish Government: Government proposition 1997/98:56. Transport politics for sustainable development. (In Swedish) (1998)

[TW96] Toint, Ph., Winter, L.: Asymmetric Multiclass Traffic Assignment: A Coherent Formulation. In: Lesort, J.-B. (ed.) Transportation and Traffic Theory: Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France, 24-26 July (1996)

[VBS86] Van Vliet, D., Bergman, T., Scheltes, W.H.: Equilibrium Traffic Assignment with Multiple User Classes. Proceedings PTRC 14^{th} Summer Annual Meeting, 111–122 (1986)

[VNR95] Verhoef, E.T., Nijkamp, P., Rietveld, P.: Second-best regulation of road transport externalities. Journal of Transport Economics and Policy, **29**, 147–167 (1995)

[YH04] Yang, H., Huang, H-J.: The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. Transp. Res., **38B**, 1–15 (2004)

[Zan69] Zangwill, W.: Nonlinear Programming: A Unified Approach. Prentice Hall, Englewood Cliffs, N.J. (1969)

Network Equilibrium Models for Analyzing Toll Highways

Michael Florian[1]

Center for Research on Transportation, University of Montreal, Montreal, H3C 3J7, Canada, mike@crt.umontreal.ca

Summary. The construction of toll highways by concessions awarded to private companies leads to the need of forecasting their usage in order to estimate the future stream of revenues. Two main modeling approaches for this problem that result in variants of multiclass network equilibrium models, are presented and commented upon.

Key words: Traffic equilibrium, congestion pricing, transportation.

1 Introduction

The construction of new highways, in both developed and developing countries, is often assigned to private companies which operate these new facilities as concessions. The users are charged tolls according to the extent that they travel on the new facilities. The derived revenues finance the construction and operation of the highway for a certain period of time, after which the highway becomes property of the state government that awarded the concession.

Economic theory is not respected by such toll highway enterprises. If one were to follow the dictates of the economic literature on tolling congested facilities, then a toll would have to be imposed on some or all of the links of the congested network. In 1952, William Vickrey, a Nobel Prize winner in Economics and the father of Congestion Pricing, suggested that fares for New York City subways should be increased in peak times and in high-traffic sections and be lowered in others. Later, he made a similar proposal for road pricing. Vickrey considered time-of-day pricing as a classic application of market forces to balance supply and demand. Those who are able can shift their schedules to cheaper hours, reducing congestion, air pollution and energy use – and increasing use of roads or other utilities. According to Vickrey, "you're not reducing traffic flow, you're increasing it, because traffic is spread more evenly over time." He also claimed that "even some proponents of congestion pricing don't understand that."

Mathematical and Computational Models for Congestion Charging, pp. 105-116

Despite the sound economic theory that supports it, the public in general opposes tolling. Because of this, elected government officials are reluctant to impose tolls on roads and highways, a resource often thought of as "free" good. When the committee chaired by Professor Reuben Smead of University College in the U.K. supported the proposition that charging road tolls would increase economic welfare, Sir Alec D. Home, the prime minister at the time, lamented that "if we are re-elected we will never again set up a study like this one." Notwithstanding the public's opposition, the mayor of London (Mr. Ken Livingstone) recently implemented a flat congestion toll of £5 for access to the city center. In doing so, London joins Singapore and Oslo as one of a few cities around the world to impose systematic congestion tolls.

The study of tolls within the context of network equilibrium models was advanced recently by the contributions of Hearn and Ramana [HR98] and Hearn and Yildirim [HY02]. This line of research, initiated by Don Hearn, led to the understanding of a variety of toll schemes that all render a "user optimal" route choice to a "system optimal" route choice. The latter minimizes travel time for all the travelers on the network.

However, the models described in this paper correspond to the actual analyses carried out in many countries for the construction of toll highways as stand alone enterprises. There is no value judgment implied by the statement of these models; rather, they are a testimony to the flexibility and adaptability of network equilibrium models to a variety of different situations and circumstances. The purpose of this paper is to identify and analyse the various approaches that have been used to predict the usage of tolled facilities among different classes of users. Essentially, the new facilities (e.g., new highways) provide shorter travel times and, given the value of time of different classes of users, one must determine the trade-off between increased travel cost and reduced travel time in order to predict their usage.

The paper is organized as follows. The next section introduces notation and definitions. Section 3 deals with deterministic models and Section 4 described a demand function based approach to this problem. In Section 5 a small numerical example is given. In Section 6, some large scale applications of these models are described and Section 7 offers some conclusions.

2 Notation and Definitions

A road network $R = (N, A)$ consists of nodes n, $n \in N$ and directed arcs a, $a \in A$ which may carry vehicular traffic. The demand for travel is subdivided into classes c, $c \in C$ which may correspond to different vehicle types or different socio-economic characteristics. The demand for travel of class c for origin-destination (O-D) pairs i, $i \in I \subset N \times N$ is denoted g_i^c. These demands use paths k, $k \in K_i^c$ where K_i^c is the set of paths used by class c for travel between O-D pair i. In its simplest form, the travel cost function for class c on arc a is the sum of the travel time function denoted as $s_a(\cdot)$ and a toll, t_a^c, that is converted into time units by the factor θ^c:

$$s_a^c(v_a) = s_a(v_a) + \theta^c t_c^a, \quad \forall a \in A. \tag{1}$$

In the above equation, v_a^c denotes the number of class c vehicles on arc a and $v_a = \sum\limits_{c \in C} v_a^c$.

To determine the choices that travelers make between "toll" and "non-toll" alternatives, stated preference analyses are carried out. Usually, the result of a stated preference analysis is a set of logit functions of the form

$$p(\text{using toll facility}) = \frac{1}{1 + \exp(\alpha^c \Delta\text{cost} + \beta^c \Delta\text{time})}, \quad \forall c \in C, \tag{2}$$

where α^c and β^c are nonnegative parameters, Δcost is the difference in the cost of the trip (usually positive if a toll facility is used) and Δtime is the difference in the trip time (usually negative if a toll facility is used).

The perceived value of time for each class c of travellers is determined as the ratio $\theta^c = \beta^c / \alpha^c$. The cost of a path is denoted $s_k^c(v)$ and is simply

$$s_k^c(v) = \sum_{a \in A} \delta_{ak} s_a^c(v_a) = \sum_{a \in A} \delta_{ak} \left(s_a(v_a) + \theta^c t_a^c\right), \quad \forall k \in K_i^c, i \in I, c \in C, \tag{3}$$

where $\delta_k^a = 1$ if arc a belongs to path k and zero otherwise. Later, it is useful to write the cost of a path as

$$s_k^c(v) = \sum_{a \in A} \delta_{ak} s_a(v_a) + t_k^c, \quad \forall k \in K_i^c, i \in I, c \in C, \tag{4}$$

where $t_k^c = \sum\limits_{a \in A} \delta_{ak} \theta^c t_a^c$ may be viewed as the toll cost of path k. The link fixed costs t_a^cmay be used to model toll plazas or tolls which vary with the distance traveled on the toll facility. It suffices to define t_a^c proportional to the length of the arc.

3 Models Based on Generalized Cost Path Choice

In such models, the demand for each class, g_i^c, $c \in C$, $i \in I$ is fixed and known and users are assumed to make their choice of a toll based only on the generalized cost differences between paths that include tolled facilities and those that do not. The usage of the tolled facilities may then be deduced from the flows on links a, $a \in A$ with positive tolls, i.e., $t_a^c > 0$. The resulting model is the classical multiclass (or multi-user) network equilibrium models which satisfies the user equilibrium condition of [Wa52]

$$\left.\begin{array}{l} s_k^c(v) = u_i^c \text{ if } h_k > 0 \\ s_k^c(v) \geq u_i^c \text{ if } h_k = 0 \end{array}\right\} k \in K_i^c,\ i \in I,\ c \in C, \tag{5}$$

where u_i^c are the shortest travel times for O-D pairs i, $i \in I$ and classes c, $c \in C$ and subject to conservation of flow and nonnegativity constraints. It is well-known (see [Da73], [Va76], [Sp95] that this network equilibrium problem is equivalent to solving the convex cost minimisation problem

$$\min \sum_{a \in A} \int_0^{v_a} s_a(x)dx + \sum_{c \in C} \sum_{a \in A} v_a^c \theta^c t_a^c \tag{6}$$

$$s.t. \sum_{k \in K_i^c} h_k = g_i^c, \; i \in I, \; c \in C \tag{7}$$

$$h_k \geq 0, \; k \in K_i^c, \; i \in I \tag{8}$$

$$(v_a^c = \sum_{k \in K_i^c} \delta_{ak} h_k, \; a \in A, \; c \in C). \tag{9}$$

The numerical solution of this model by the linear approximation method is well-known and will not be repeated again here. It is perhaps worthwhile to point out that the flows by class, v_a^c, are not unique, nor are the path flows h_k, but the arc flows v_a are indeed unique.

This model has been used extensively in many toll facility studies since most popular transportation planning software packages offer, as a standard model, a generalized cost multi-class network equilibrium model. The only published references known to the author are [Me95] and [Me95]. These articles describe the models used for the analysis of Highway 407, a toll facility which bypasses the city of Toronto, Canada.

In this formulation the link cost functions, $s_a(v_a)$, $a \in A$, are relatively simple, since they do not model asymmetric costs due to different vehicle types. If more complex functions were used, the resulting multiclass model would be considerably more complex and would require the solution of a variational inequality model (see [FH95]).

4 Models Based on Explicit Choice of Tolled Facilities

Such models are based on logit functions obtained from stated preference analyses to determine the probability (or proportion) that a user in each class will use paths that include tolled facilities. Let g_i^{ct} denote the number of users in class c who are willing to pay tolls and g_i^{cn} denote the number of those who are not. That is, $g_i^c = g_i^{ct} + g_i^{cn}$, $i \in I$, $c \in C$, and, as in the path based approach, the total demand for each class g_i^c, $i \in I$, $c \in C$ is assumed to be fixed and known. Also, let K_i^{ct} and K_i^{cn} denote the sets of paths that contain tolled facilities and those that do not, respectively. The resulting multi-class network equilibrium model with explicit choice functions may be stated as follows:

$$\left. \begin{array}{l} s_k^{ct}(v) = u_i^{ct}, \text{ if } h_k > 0 \\ s_k^{ct}(v) \geq u_i^{ct}, \text{ if } h_k = 0 \end{array} \right\} k \in K_i^{ct}, \; i \in I, \; c \in C \tag{10}$$

$$\left.\begin{array}{l} s_k^{cn}(v) = u_i^{cn}, \text{ if } h_k > 0 \\ s_k^{cn}(v) \geq u_i^{cn}, \text{ if } h_k = 0 \end{array}\right\} k \in K_i^{cn},\ i \in I,\ c \in C \tag{11}$$

$$\sum_{k \in K_i^{ct}} h_k - g_i^{ct} = 0,\ \ i \in I,\ c \in C \tag{12}$$

$$\sum_{k \in K_i^{cn}} h_k - g_i^{cn} = 0,\ \ i \in I,\ c \in C \tag{13}$$

$$g_i^{ct} = g_i^c / \left\{1 + \exp\left(\alpha^c t_i^c + \beta^c \left(u_i^{ct} - u_i^{cn}\right)\right)\right\},\ i \in I,\ c \in C;\ \left(g_i^{cn} = g_i^c - g_i^{ct}\right) \tag{14}$$

$$h_k \geq k \in K_i^{ct},\ \ k \in K_i^{cn},\ c \in C,\ i \in I \tag{15}$$

$$g_i^{cn},\ g_i^{ct} \geq 0,\ \ i \in I,\ \ c \in C, \tag{16}$$

where t_i^{ct} is the average toll paid for all traffic of O-D pair i that uses toll roads.

Clearly there is a difficulty with this formulation, since the paths used are not known before computing the equilibrium flows. In addition, the number of possible paths is exceedingly large. On the other hand, there is an equivalent formulation in terms of p_k, the *proportion* of demand that uses path k. In particular, the path flows, h_k, may be written as

$$\left.\begin{array}{l} h_k = p_k g_i^{ct},\ k \in K_i^{ct} \\ h_k = p_k g_i^{cn},\ k \in K_i^{cn} \end{array}\right\} i \in I,\ c \in C, \tag{17}$$

and the arc flows may be expressed as

$$v_a^{ct} = \sum_{k \in K_i^{ct}} \delta_{ak} p_k g_i^{ct},\ a \in A,\ c \in C \tag{18}$$

$$v_a^{cn} = \sum_{k \in K_i^{cn}} \delta_{ak} p_k g_i^{cn},\ a \in A^n \tag{19}$$

$$A^n = A - \{\text{toll links}\} \tag{20}$$

$$v_a = \sum_{c \in C} \left(v_a^{ct} + v_a^{cn}\right),\ a \in A. \tag{21}$$

Then, the costs of paths containing and not containing tolled facilities are, respectively,

$$s_k^{ct}(v) = \sum_{a \in A} \delta_{ak} s_a(v_a), \quad k \in K_i^{ct},\ i \in I \tag{22}$$

$$s_k^{cn}(v) = \sum_{a \in A^n} \delta_{ak} s_a(v_a), \quad k \in K_i^{cn},\ i \in I \tag{23}$$

The formulation in the space of path flow proportions , p_k, consists of

1) The user equilibrium inequalities (9) – (10).

2) The conservation of flow equations

$$\sum_{k\in K_i^{ct}} p_k g_i^{ct} - g_i^{ct} = 0 \quad\Rightarrow\quad g_i^{ct}\left(\sum_{k\in K_i^{ct}} p_k - 1\right) = 0,\ i\in I,\ c\in C \tag{24}$$

$$\sum_{k\in K_i^{cn}} p_k g_i^{cn} - g_i^{cn} = 0 \quad\Rightarrow\quad g_i^{cn}\left(\sum_{k\in K_i^{cn}} p_k - 1\right) = 0,\ i\in I,\ c\in C \tag{25}$$

3)

$$g_i^{ct} = g_i^c / \left\{1 + \exp\left(\alpha^c\left[\sum_{k\in K_i^{ct}} p_k t_k^c\right] + \beta^c\left(u_i^{ct} - u_i^{cn}\right)\right)\right\},$$
$$i\in I,\ c\in C; \left(g_i^{cn} = g_i^c - g_i^{ct}\right) \tag{26}$$

4) Nonnegativity constraints (14) – (15).

This formulation highlights the importance of the path proportions p_k and the large dimension of the problem. For example, with 9 user classes, one would have 18 flow vectors for each link. For a network of 1000 × 1000 O-D pairs, one would consider explicitly a number of paths of the order of 10^6 and one would need to keep at least 18 matrices, each of size 10^6. While it is possible to restate this model in the form of a variational inequality and search for rigorous solution algorithms, the actual solution methods used in most applications rely on heuristic algorithms that have performed well but that are not supported by convergence proofs.

In order to simplify the model, it is sometimes assumed that the vehicles of the different classes are homogeneous and that the O-D travel costs (impedances) may be simplified to

$$u_i^{ct} = u_i^t \text{ and } u_i^{cn} = u_i^n,\ i\in I,\ c\in C \tag{27}$$

that is, all the toll payers may be aggregated into one class and all the non toll payers may be aggregated into one class. This is partly justified by the implicit assumption that the toll is perceived at the demand function level, prior to the trip, and once the decision to pay or not to pay the toll is made, the path choice is no longer governed by generalized cost, but only by time. However, this assumption is *not* made in the following "heuristic" solution algorithm:

Explicit Choice Tolled Assignment Heuristic

Step 0 (Initialization) : $l = 0$, choose $g_i^{ct(0)}$, $g_i^{cn(0)}$, $i\in I$, $c\in C$;

Step 1 (Compute path costs and times)
Solve a two-class network equilibrium problem by the linear approximation method:

$$\min \sum_{a\ \in A} \int_0^{v_a} s_a(x)dx + \sum_{c\ \in C} \sum_{a\ \in A} v_a^c \theta^c t_a^c \tag{28}$$

$$s.t. \sum_{k \in K_i^t} h_k = g_i^{ct(l)}, \ \sum_{k \in K_i^n} h_k = g_i^{cn(l)}, \ i \in I \tag{29}$$

$$h_k \geq 0, \ \ k \in K_i^t, \ \ k \in K_i^c, \ \ i \in I, \ \ c \in C \tag{30}$$

to find $u_i^{t(l)}$ and $u_i^{n(l)}$ and, while doing so compute

$$t_i^{t(l)} = \sum_{k \in K_i^t} p_k^{(l)} t_k^{ct(l)} \tag{31}$$

which are the tolls for each class and O-D pair. The path proportions $p_k^{(l)}$, $k \in K_i^{ct(l)}$, $k \in K_i^{cn(l)}$ are computed from the step sizes of the linear approximation method at each iteration.

Step 2 (Modify demand): $l = l + 1$;
$\tilde{g}_i^{ct}$, $\tilde{g}_i^{cn}$are recomputed by using the logit functions for each class c:

$$\tilde{g}_i^{ct} = g_i^c / \left\{ 1 + \exp\left(\alpha^c t_i^{ct(l)} + \beta^c \left(u_i^{t(l)} - u_i^{n(l)} \right) \right) \right\}, \ c = 1, 2, ..., c \tag{32}$$

and

$$\left. \begin{array}{l} g_i^{ct(l)} = \left(1 - \lambda^{(l)}\right) g_i^{ct(l-1)} + \lambda^{(l)} \tilde{g}_i^{ct} \\ g_i^{cn(l)} = g_i^c - g_i^{ct(l)} \end{array} \right\} \ i \in I, \ \ c \in C \tag{33}$$

$$0 \leq \lambda^{(l)} \leq 1 \tag{34}$$

Step 3 (Convergence test)
If $\max_{i,c} \left\| g_i^{ct(l)} - g_i^{ct(l-1)} \right\| \leq \varepsilon$, STOP ;
otherwise, return to Step 1.

The step sizes $\lambda^{(l)}$ may be chosen to implement the method of successive averages (MSA) or any other reasonable sequence of step sizes. The algorithm still requires at least $2\,|C|$ O-D matrices, and there are $2\,|C|$ link flow vectors, v_a^{ct} and v_a^{cn}, $a \in A$.

No convergence proof is given in this paper, however in numerous applications in practice, the algorithm has demonstrated good empirical convergence. It is evident that, if the algorithm terminates, the resulting demands and flows satisfy approximately the model formulation.

A block diagram representation of this heuristic algorithm is in Figure 1.

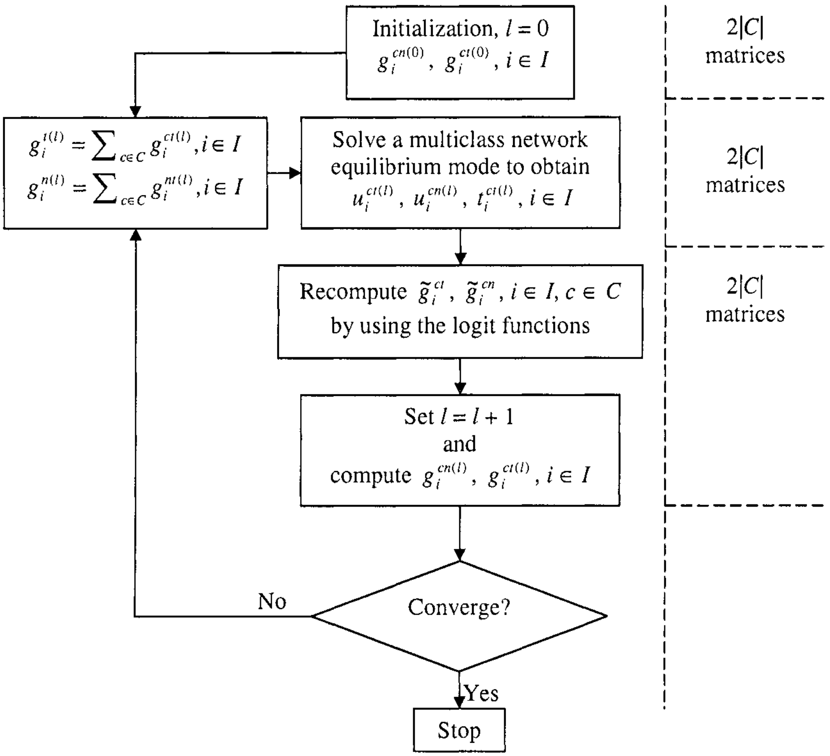

Initialization, $l = 0$
$g_i^{cn(0)}, g_i^{ct(0)}, i \in I$
$g_i^{t(l)} = \sum_{c \in C} g_i^{ct(l)}, i \in I$
$g_i^{n(l)} = \sum_{c \in C} g_i^{nt(l)}, i \in I$
Solve a multiclass network equilibrium mode to obtain
$u_i^{ct(l)}, u_i^{cn(l)}, t_i^{ct(l)}, i \in I$
Recompute $\tilde{g}_i^{ct}, \tilde{g}_i^{cn}, i \in I, c \in C$ by using the logit functions
Set $l = l + 1$ and compute $g_i^{cn(l)}, g_i^{ct(l)}, i \in I$
Converge?
No
Yes
Stop
2|C| matrices
2|C| matrices
2|C| matrices

Fig. 1. Block diagram of heuristic algorithm

5 A Small Numerical Example

A small network of three links and one origin-destination pair is used to illustrate the difference between the two approaches for predicting the usage of toll highways. The network is given in Figure 2.

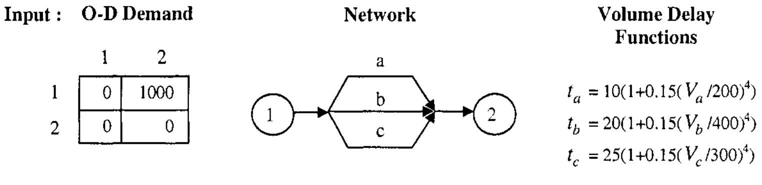

Input : O-D Demand
1
2
1
0
1000
2
0
0
Network
a
b
c
1
2
Volume Delay Functions
$t_a = 10(1+0.15(V_a/200)^4)$
$t_b = 20(1+0.15(V_b/400)^4)$
$t_c = 25(1+0.15(V_c/300)^4)$

Fig. 2. The network and the demand

Figures 3,4,5 show successively the equilibrium flows without tolls, with a toll of 2 units on the middle link and a value of time of 1.2 and with the application of a demand function where the probability of using a toll facility

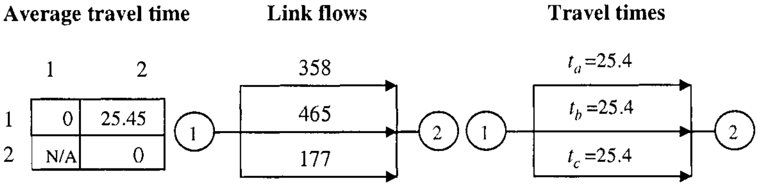


Fig. 3. The equilibrium flows without tolls

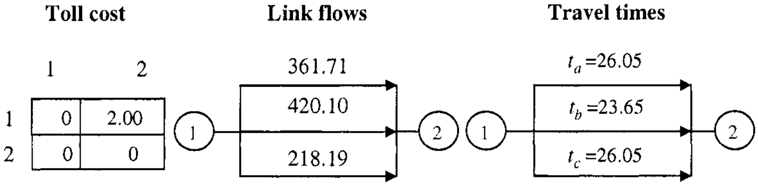


Fig. 4. The equilibrium flows with a toll of 2 units and value of time of 1.2

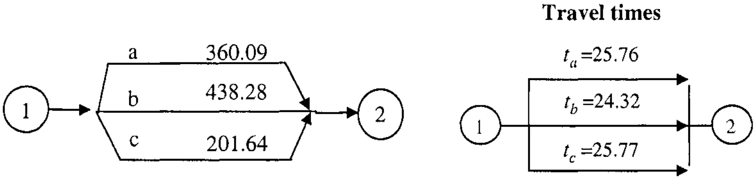


Fig. 5. The equilibrium flows obtained by using the demand function

is given by the function

$$P_r(\text{toll}) = 1/(1 + \exp(.2556(\text{time difference}) + .3067 * \text{toll})[1].$$

For this model convergence was reached after 4 iterations. The toll facility, which is the middle link, carries 420.10 trips in the simple model compared to 438.28 trips when the demand function is used. For this solution the proportion of toll trips given by the logit function is .438972

6 Some Large-Scale Applications

The algorithm described in Section 4 has been used in numerous applications in Europe, North America and Asia. Most of these applications are confidential and the results may not be reported in an academic paper. However, a pilot application of very large scale, carried out on the network used for transportation planning in Southern California may be reported in this paper. The network consists of 2,450 zones, 46,000 arcs. The demand for travel is subdivided into High Occupancy Vehicles (HOV) and Low Occupancy Vehicles

[1] The constants in this function were chosen so that their ratio is exactly 1.2.

(LOV). Tolls were envisioned on some of the regional highways. The logit function

$$P_r(\text{using toll}) = 1/(1 + \exp(0.5647(u_i^t - u_i^n) + 0.4199(t_i^t))$$

was used to determine the probability of using the toll facility. The model described in the previous section was adapted to handle the HOV and LOV demand. A two-class (HOV, LOV) network equilibrium model was used to find the initial travel times and toll costs. The logit function was used to obtain four matrices corresponding to the demand for HOV_{toll}, HOV_{notoll}, LOV_{toll}, and LOV_{notoll}, and a four-class network equilibrium assignment was carried out in Step 1 of the heuristic algorithm. The convergence criterion for an $\varepsilon = 1$ (1 trip) was satisfied after four iterations of the algorithm. The computations were carried out with the EMME/2 (INRO, 1996) software package.

Both these models were applied in Mexico City for the evaluation of a 26 km section of an urban autoroute (Chamapa Highway). They produced different results, which is not surprising. The explicit choice model was used in the final analysis. The generalized cost path based approach was used in several applications in North America, Europe, Asia and Australia.

7 CONCLUSIONS

The intuitive heuristic solution algorithm for the explicit choice function approach was used successfully in practice in numerous applications. It is an example of the compromises that one must make in order to solve large-scale non-standard network equilibrium models. The results obtained are quite sensitive to the coding of the network and the quality of the stated preference model calibrations. The costs of building toll highways are so large that they justify careful use of travel demand and network models to predict the potential ridership.

References

[Da73] Dafermos, S.: The Traffic Assignment Problem for Multi-class User Transportation Networks. Transportation Science, **6**, 73–87 (1973)

[FH95] Florian, M., Hearn, D.: Network Equilibrium Models and Algorithms. In: Ball, M.O. et al (eds) Handbooks in OR & MS. **8**, 485–550 (1995)

[Go97] Goodwin, P.B.: Solving Congestion. Inaugural Lecture for the Professorship of Transport Policy University College London. 23^{rd} October (1997)

[HR98] Hearn, D.W., Ramana, M.V.: Solving Congestion Toll Pricing Models. In: Marcotte, P., Nguyen, S. (eds) Equilibrium and Advanced Transportation Modeling. Kluwer Academic Publishers, 109–124 (1998)

[HY02] Hearn, D.W., Yildirim, M.B.: A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand. In: Gendreau, M., Marcotte, P. (eds) Transportation and Network Analysis: Current Trends. Kluwer Academic Publishers, 135–145 (2002)

[INRO96] INRO Consultants Inc.: EMME/2 User's Manual. Release 8, 815 (1996)

[Me95] Mekky, A.: Toll Revenue and Traffic Study of Highway 407 in Toronto. Transportation Research Record, **1498**, Transportation Research Board, Network Research Council, Washington, C.D., U.S.A., 5–15 (1995)

[Me97] Mekky, A.: Comparing Tolling Strategies for Highway 407 in the Greater Toronto Area. Paper presented at the 76^{th} Annual Meeting, Transportation Research Board, Washington, D.C., U.S.A., 8 (1997)

[Sp95] Spiess, H.: A Note on Multi-class Network Equilibrium Models. Private Communication, 2 (1995)

[Va76] Van Vliet, D.: Road Assignment-I. Principles and Parameters of Model Formulation. Transportation Research, **10**, 137–143 (1976)

[Wa52] Wardrop, J.G.: Some Theoretical Aspects of Road Traffic Research. Proc. Inst. Civil Engineers, Part II, 325–378 (1952)

On the Applicability of Sensitivity Analysis Formulas for Traffic Equilibrium Models

Magnus Josefsson[1] and Michael Patriksson[2]

[1] Department of Mathematics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden, f98majf@dd.chalmers.se
[2] Department of Mathematics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden, mipat@math.chalmers.se

Summary. The paper by Tobin and Friesz [ToF88] brought the classic nonlinear programming subject of sensitivity analysis to transportation science. It is still the most widely used device by which "gradients" of traffic equilibrium solutions (that is, flows and/or demands) are calculated, for use in bilevel transportation planning applications such as network design, origin–destination (OD) matrix estimation and problems where link tolls are imposed on the users in order to reach a traffic management objective. However, it is not widely understood that the regularity conditions proposed by them are stronger than necessary. Also, users of their method sometimes misunderstand its limitations and are not aware of the computational advantages offered by more recent methods. In fact, a more often applicable formula was proposed already in 1989 by Qiu and Magnanti [QiM89], and Bell and Iida [BeI97] describe one of the cases in practice in which the formula by Tobin and Friesz [ToF88] would not be able to generate sensitivity information, because one of their regularity conditions fails to hold.

This paper provides a short overview of a sensitivity formula that provides directional derivatives of traffic equilibrium flows, route and link costs, and demands, exactly when they exist, and which are found in [PaR03] and [Pat04]. For the simplicity of the presentation, we provide the analysis for the simplest cases, where the link travel cost and demand functions are separable, so that we can work with optimization formulations; this specialization was first given in [JoP04]. The connection between directional derivatives and the gradient is that exactly when the directional derivative mapping of the traffic equilibrium solution is linear in the parameter, the solution is differentiable.

The paper then provides an overview of the formula of Tobin and Friesz [ToF88], and illustrates by means of examples that there are several cases where it is not applicable: First, the requirement that the equilibrium solution be strictly complementary is too strong—differentiable points may not be strictly complementary. Second, the special matrix invertibility condition implies a strong requirement on the topology of the traffic network being analyzed and which may not hold in practice, as noted by Bell and Iida [BeI97](page 97); moreover, the matrix condition may fail to hold at differentiable points.

Mathematical and Computational Models for Congestion Charging, pp. 117-142

The findings of this paper are hoped to motivate replacing the previous approach with the more often applicable one, not only because of this fact but equally importantly because it is intuitive and also can be much more efficiently utilized: the sensitivity problem that provides the directional derivative is a linearized traffic equilibrium problem, and the sensitivity information can be generated efficiently by only slightly modifying a state-of-the-art traffic equilibrium solver. This is essential for bringing the use of sensitivity analysis in transportation planning beyond the solution of only small problems.

Key words: Traffic equilibrium, sensitivity analysis, bilevel programming, MPEC.

1 Introduction

Performing a sensitivity analysis of traffic equilibria means evaluating the directions of change that occur in the flows and travel costs as parameters in the cost and demand functions change. A sensitivity analysis is particularly useful in control and pricing applications since, if we can anticipate the effects of a change in, say, the traffic infrastructure, on the behaviour of the travellers, then we can utilize this knowledge to optimize these changes according to some goal fulfillment, like a reduction in flows or delays, a higher revenue from congestion tolls, etc. Such problems constitute instances of *bilevel optimization* problems, or *mathematical programs with equilibrium constraints* (MPEC), which is the scientific field within operations research and mathematical programming that is associated with hierarchical optimization problems, and which also includes the origin–destination (OD) matrix estimation problem. (The monograph [LPR96] provides a good overview of MPEC models and methods.) Several algorithms for MPEC problems rely on efficiently and generally applicable sensitivity analysis tools; it is in this framework that are findings can best be utilized.

Recently, the authors have been involved in a project having the goal to provide a precise sensitivity analysis of elastic and fixed demand traffic equilibrium problems, focusing on general models involving possibly non-separable and non-invertible link cost and demand functions; cf. [PaR02, PaR03, JoP04, Pat04, Pat05]. Our focus here is on the special case of separable link cost and demand functions, the latter also being invertible, in which case we can work directly on optimization formulations. We illustrate how to perform a sensitivity analysis efficiently in practice by using a modification of state-of-the-art traffic equilibrium software.

In 1988, Tobin and Friesz did the transportation science community the great service of bringing to it the nonlinear programming topic of sensitivity analysis, with their publication [ToF88]. Their analysis is quite accessible to practitioners; for example, they utilize the rather intuitive implicit function theorem in their analysis. It also remains the most popular tool for producing sensitivity analysis information in traffic equilibrium problems.

We illustrate through examples how their formula is however less applicable in several ways. Moreover, it relies on calculations with very large matrices, and therefore cannot be applied to large-scale networks. Our sensitivity analysis problem is however quite structured and need not involve matrix calculations at all; it amounts to solving a perturbed, affine traffic equilibrium problem, which is no more difficult to solve than the original one.

2 The Traffic Model

Let $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ be a transportation network, where $\mathcal{N}$ and $\mathcal{L}$ are the sets of nodes and directed links, respectively. For certain ordered pairs of nodes, $(p,q) \in \mathcal{C}$, where node p is an origin, node q is a destination, and $\mathcal{C}$ is a subset of $\mathcal{N} \times \mathcal{N}$, there is a transport demand, which may be given by a function of the travel cost. We assume that the network is enough connected, such that at least one route joins each origin–destination (OD) pair.

Wardrop's user equilibrium principle states that for every OD pair $(p,q) \in \mathcal{C}$, the travel costs of the routes utilized are equal and minimal for each individual user. We denote by $\mathcal{R}_{pq}$ the set of simple (loop-free) routes for OD pair (p,q), by h_r the flow on route $r \in \mathcal{R}_{pq}$, and by c_r the travel cost on the route as experienced by an individual user.

We introduce the parameter (that is, control variable) to be present in the sensitivity analysis: it is denoted ρ, and is assumed to be of dimension d. This parameter could be present in one or both of the travel cost and demand functions. We assume that the travel cost function has the form $c(\rho, \cdot) : \Re_+^{|\mathcal{R}|} \mapsto \Re^{|\mathcal{R}|}$ given a value of ρ, where $|\mathcal{R}|$ denotes the total number of routes in the network. Further, for a given value of the vector ρ, the demand function is given by $g(\rho, \cdot) : \Re^{|\mathcal{C}|} \mapsto \Re_+^{|\mathcal{C}|}$. (We introduce the notation $\Re_+ := \{ x \in \Re \mid x \geq 0 \}$ and $\Re_{++} := \{ x \in \Re \mid x > 0 \}$.)

In an application to OD estimation, d is in the order of $|\mathcal{C}|$, while $d \approx |\mathcal{L}|$ holds in equilibrium network design, pricing and control models.

We also introduce the matrix $\Gamma \in \{0,1\}^{|\mathcal{R}| \times |\mathcal{C}|}$, which is the route–OD pair incidence matrix (i.e., the element γ_{rk} is 1 if route r joins OD pair $k = (p,q) \in \mathcal{C}$, and 0 otherwise). Then, demand-feasibility is described by the conditions that $h \in \Re_+^{|\mathcal{R}|}$ and

$$\Gamma^T h = g(\rho, \pi) \tag{1}$$

holds, while the Wardrop equilibrium conditions for the route flows are that

$$h_r > 0 \Longrightarrow c_r(\rho, h) = \pi_{pq}, \qquad r \in \mathcal{R}_{pq}, \qquad (p,q) \in \mathcal{C}, \tag{2a}$$

$$h_r = 0 \Longrightarrow c_r(\rho, h) \geq \pi_{pq}, \qquad r \in \mathcal{R}_{pq}, \qquad (p,q) \in \mathcal{C}, \tag{2b}$$

holds, where the value of $\pi_{pq} := \pi_{pq}(\rho, h)$ is the minimal route cost in OD pair (p,q) at equilibrium. By the non-negativity of the route flows, the system

(1)–(2) can more compactly be written as the mixed complementarity problem (MCP)

$$0^{|\mathcal{R}|} \leq h \perp (c(\rho, h) - \Gamma\pi) \geq 0^{|\mathcal{R}|}, \tag{3a}$$

$$\Gamma^T h = g(\rho, \pi), \tag{3b}$$

where $a \perp b$, for two arbitrary vectors $a, b \in \Re^n$, means that $a^T b = 0$. (By nonnegativity, this implies that $a_j \cdot b_j = 0$ for all j.)

As we are interested in the sensitivity of link flows, we will assume that the route cost is additive. For each link $l \in \mathcal{L}$, the travel cost has the form $t_l(\rho, v_l)$, where $v \in \Re^{|\mathcal{L}|}$ is the vector of link flows. The route and link travel costs and flows are related through a link–route incidence matrix, $\Lambda \in \{0,1\}^{|\mathcal{L}| \times |\mathcal{R}|}$, whose element λ_{lr} equals one if route $r \in \mathcal{R}$ utilizes link $l \in \mathcal{L}$, and zero otherwise. Route r has an additive route cost $c_r(\rho, h)$ if it is the sum of the costs of using all the links defining it. In other words, $c_r(\rho, h) = \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\rho, v_l)$. In short, then, $c(\rho, h) = \Lambda^T t(\rho, v)$. Also, implicit in this relationship is the assumption that the pair (h, v) is consistent, in the sense that v equals the sum of the route flows: $v = \Lambda h$. We shall use the representation in terms of v, as it is an entity for which we can introduce conditions ensuring that uniqueness holds at equilibrium.

As could be noted above, the link travel cost is assumed to be separable (that is, the travel cost of link l depends only on v_l). A separability assumption is made also with respect to the demand function, which is supposed to be of the form $g_k(\rho, \pi_k)$, $k \in \mathcal{C}$, for a given value of the vector ρ.

In order to be able to work with an optimization formulation, which furthermore admits a unique solution (v^*, d^*) and is such that we can apply sensitivity analysis theory, we introduce the following assumption, which is supposed to hold throughout:

Assumption 1 (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.*

The optimization formulation that we will work with is the following standard one for elastic demand traffic assignment (e.g., [Pat94]):

$$\min_{(v,d)} \phi(v,d) := \sum_{l \in \mathcal{L}} \int_0^{v_l} t_l(\rho, s)\, ds - \sum_{k \in \mathcal{C}} \int_0^{d_k} \xi_k(\rho, s)\, ds, \tag{4a}$$

$$\text{s.t.} \quad \Gamma^T h = d, \tag{4b}$$

$$v = \Lambda h, \tag{4c}$$

$$h \geq 0^{|\mathcal{R}|}. \tag{4d}$$

For future use, let C denote the set of feasible vectors (h, v, d) in the problem (4), that is,

$$C = \left\{ \begin{pmatrix} h \\ v \\ d \end{pmatrix} \in \Re_+^{|\mathcal{R}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \;\middle|\; \Gamma^T h = d; \quad v = \Lambda h \right\}.$$

The variational inequality problem, that characterizes the solution (h^*, v^*, d^*) to this problem, is stated as that of finding $(h^*, v^*, d^*) \in C$ such that

$$t(\rho, v^*)^T (v - v^*) - \xi(\rho, d^*)^T (d - d^*) \geq 0, \qquad (h, v, d) \in C. \tag{5}$$

To see that this expression characterizes the Wardrop conditions stated earlier in (3), we notice that (5) is equivalent to (h^*, v^*, d^*) solving the following linear program:

$$\min_{(v,d)} \; t(\rho, v^*)^T v - \xi(\rho, d^*)^T d, \tag{6a}$$

$$\text{s.t.} \;\; \Gamma^T h - d = 0^{|\mathcal{C}|}, \tag{6b}$$

$$v - \Lambda h = 0^{|\mathcal{L}|}, \tag{6c}$$

$$h \geq 0^{|\mathcal{R}|}. \tag{6d}$$

Its LP dual is to

$$\max_{(\pi,\alpha)} \; 0, \tag{7a}$$

$$\text{s.t.} \;\; \Gamma \pi - \Lambda^T \alpha \leq 0^{|\mathcal{R}|}, \tag{7b}$$

$$-\pi = -\xi(\rho, d^*), \tag{7c}$$

$$\alpha = t(\rho, v^*), \tag{7d}$$

where π and α are, respectively, the LP dual variables for the constraints (6b) and (6c). The dual variable α is eliminated by using (7d). The complementarity conditions between the two LP problems can then be written as

$$0^{|\mathcal{R}|} \leq h^* \perp (\Lambda^T t(\rho, v^*) - \Gamma \pi^*) \geq 0^{|\mathcal{R}|}, \tag{8}$$

which is identical to the Wardrop condition (3a). The condition (3b) is obtained as follows: from (6b) and (7c), $\Gamma^T h^* = d^* = g(\rho, \pi^*)$. As $t(\rho, \cdot)$ and $-g(\rho, \cdot)$ both are strictly monotone, the objective function of (4) is strictly convex; therefore, the solution in (v^*, d^*) to (4), and equivalently to the variational inequality (5) and to the Wardrop conditions (3), is unique. We see that from (7c)–(7d), also the dual entities (π^*, α^*) are unique.

3 The basis for our sensitivity analysis

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

$$f(\rho, x^*)^T(x - x^*) \geq 0, \qquad x \in X, \tag{9}$$

where $X \subseteq \Re^n$ is a polyhedral set.

Equivalently, we can write this in a more natural form as follows:

$$-f(\rho, x^*) \in N_X(x^*),$$

where $N_X(x)$ denotes the normal cone to X at x:

$$N_X(x) := \begin{cases} \{\, v \in \Re^n \mid v^T(y - x) \leq 0, \qquad y \in X \,\}, & \text{if } x \in X, \\ \emptyset & \text{otherwise.} \end{cases}$$

We let $S : \Re^d \rightrightarrows \Re^n$ denote the mapping that assigns to each vector $\rho \in \Re^d$ the set $S(\rho)$ of solutions to this problem:

$$S(\rho) := \{\, x^* \in X \mid f(\rho, x^*)^T(x - x^*) \geq 0, \quad x \in X \,\}, \qquad \rho \in \Re^d. \tag{10}$$

(The notation "$\rightrightarrows$" signifies that the mapping S in general is a point–to–set mapping.) Letting $\rho = \rho^*$ be the current value of the parameter vector, we are interested in the direction of change of the solution x^* as ρ^* is perturbed along a direction ρ'. This directional derivative of S is the solution to an auxiliary variational inequality, which has the following form: find $x' \in K$ such that

$$r(\rho', x')^T(x - x') \geq 0, \qquad x \in K, \tag{11a}$$

where

$$K := T_X(x^*) \cap f(\rho^*, x^*)^{\perp}, \tag{11b}$$
$$r(\rho', x') := \nabla_\rho f(\rho^*, x^*)\rho' + \nabla_x f(\rho^*, x^*)x'. \tag{11c}$$

We let $DS(\rho^*|x^*) : \Re^d \rightrightarrows \Re^n$ denote the mapping that assigns to each perturbation $\rho \in \Re^d$ the set $DS(\rho^*|x^*)(\rho')$ of solutions to this problem:

$$DS(\rho^*|x^*)(\rho') := \{\, x' \in K \mid r(\rho', x')^T(x - x') \geq 0, \quad x \in K \,\}, \qquad \rho' \in \Re^d.$$

The set K denotes the set of variations around 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

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

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

We remark that the two polyhedral cones $T_X(x)$ and $N_X(x)$ in fact are polar to each other:

$$T_X(x) := (N_X(x))^* := \{\, w \in \Re^n \mid w^T v \leq 0, \qquad v \in N_X(x) \,\}.$$

This classic result in polyhedral theory lies behind some of the development of the results of this section.

Suppose now that $f(\rho, \cdot)$ is monotone on X around $\rho = \rho^*$, and that the parameterization is such that $\operatorname{rank} \nabla_\rho f(\rho^*, x^*) = n$. (The latter result can always be fulfilled by including enough dummy parameters.) We say that the mapping S is *strongly regular* at ρ^* ([Rob80, Rob85]) if S is single-valued and Lipschitz continuous on some neighbourhood of ρ^*. Then, according to a result by Dontchev and Rockafellar [DoR01],

$$S \text{ is strongly regular at } \rho^* \quad \Longleftrightarrow \quad DS(\rho^* | x^*) \text{ is single-valued.} \tag{12}$$

Moreover, then the unique solution x' to (11) is the directional derivative of the solution x^* to (9) at ρ^*, in the direction of ρ'. A sufficient condition for the property of single-valuedness of DS in (12) to hold is, by Kyparisis [Kyp88, Lemma 2.1], that

$$\nabla_x f(\rho^*, x^*) \text{ is positive definite on } (K - K). \tag{13}$$

We refer to this as a sufficient *second-order* condition. The set $K - K$ is the subspace consisting of all vectors z of the form $z = \alpha - \beta$ for some α and β in K. A stronger result than directional differentiability can also be obtained under additional assumptions: according to a result of Kyparisis [Kyp90], under the strong regularity assumption above,

$$S \text{ is differentiable at } \rho^* \quad \Longleftrightarrow \quad DS(\rho^* \mid x^*)(\rho') \in -K, \qquad \rho' \in \Re^d.$$

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

$$\nabla_\rho x(\rho^*) = -Z \left[Z^T \nabla_x f(\rho^*, x^*) Z \right]^{-1} Z^T \nabla_\rho f(\rho^*, x^*), \tag{14}$$

for any $n \times \ell$ matrix Z such that $Z^T Z$ is nonsingular and $z \in K \cap (-K)$ if and only if $z = Zy$ for some $y \in \Re^\ell$, where ℓ is the dimension of $K \cap (-K)$. This differentiability result is a kind of implicit function theorem; the relationship in (12) shows how the implicit function theorem naturally extends to more general cases.

We refer to this latter property not because we will establish sufficient conditions for its application in the present context (this has already been done in [Pat04, Pat05]), but to remark that the heuristic sensitivity analysis that is developed in the paper [ToF88] and its follow-up [CSF00] strives to utilize (14). Unfortunately, not only does the property $DS(\rho^* \mid x^*)(\rho') \in -K$ fail to hold in many cases (cf. [Pat04, Pat05], as well as below), but also there may not exist a nonsingular matrix of the kind that is referred to above (cf. [BeI97, page 97]).

4 Sensitivity analysis of separable traffic equilibria

We first identify the sensitivity problem in our notation. Let

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

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

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

where

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

and

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

By the monotonicity and separability of t and $-\xi$, the resulting sensitivity variational inequality can be equivalently written as the following convex quadratic optimization problem to

$$\min_{(v', d')} \; \phi'(v', d') := [\nabla_\rho t(\rho^*, v^*)\rho']^T v' + \frac{1}{2} \sum_{l \in \mathcal{L}} \frac{\partial t_l(\rho^*, v^*_l)}{\partial v_l} (v'_l)^2$$
$$- [\nabla_\rho \xi(\rho^*, d^*)\rho']^T d' - \frac{1}{2} \sum_{k \in \mathcal{C}} \frac{\partial \xi_k(\rho^*, d^*_k)}{\partial d_k} (d'_k)^2, \tag{15a}$$

$$\text{s.t.} \quad \Gamma^T h' = d', \tag{15b}$$
$$v' = \Lambda h', \tag{15c}$$
$$h \in H'. \tag{15d}$$

The derivation follows the same pattern as that in [PaR02, PaR03, Pat04, Pat05]. The sensitivity problem is closely related to the original model. Two main differences are notable: the link cost and demand functions are replaced by their linearizations, and the sign restrictions on 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 h^*, it is an interesting fact that the possible choices of v' in K does *not*; this is a general consequence of aggregation, which was also utilized in [PaR02, PaR03, JoP04, Pat04, Pat05]. In summary, it appears that the sensitivity 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.

We now apply the result (12) in the previous section. The result states that the sensitivity problem provides directional derivatives, provided that the solution is unique. So, under what circumstances will the optimal solution to (15) be unique? Similarly, which entities in the solution to the problem (4) [flow, travel cost, demand] have directional derivatives?

Clearly, we cannot apply the theory of the previous section to the problem stated in (h, v, d), since h^* is not unique. As (v^*, d^*) is unique, we could project the problem onto this space. This is simply accomplished by redefining

$$C^{\text{proj}} = \left\{ \begin{pmatrix} v \\ d \end{pmatrix} \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \,\middle|\, \exists h \text{ with } \begin{pmatrix} h \\ v \\ d \end{pmatrix} \in C \right\};$$

further, we would let

$$x = \begin{pmatrix} v \\ d \end{pmatrix}; \qquad f(\rho, x) = \begin{pmatrix} t(\rho, v) \\ -\xi(\rho, d) \end{pmatrix}.$$

We stress that this type of projection of the problem is only valid thanks to the particular relationships between the link and routes flows; normally, a projection such as the one above does not preserve the regularity properties we are utilizing. (In [QiM89, Yen95] it is established that the route flow variables can be gotten rid of by always choosing a particular value of them, namely that which minimizes, over the equilibrium set of route flows, the route flow vector's Euclidean norm; this operation preserves regularity. We can consider our projection to be based on exactly that type of choice.)

As we are also interested in the sensitivity of the travel costs, we will, for the first time, introduce yet another modification: we introduce a dummy variable, $s \in \Re^{|\mathcal{L}|}$, which will take on the (negative) value

$$s^* = -t(\rho^*, v^*)$$

of the link travel cost at equilibrium, and likewise a variable, $\pi_- \in \Re^{|\mathcal{C}|}$, to take on the (negative) value

$$\pi_-^* = -\xi(\rho^*, d^*)$$

of the equilibrium OD travel costs. [Note that $\pi_-^* = -\pi^*$, where π^* is given in (8).] In the sensitivity problem, then, s' and π_-' will equal the (negative of the) link and OD travel cost perturbation, respectively.

The problem which will be analyzed is the following: in (9), let

$$x = \begin{pmatrix} v \\ d \\ s \\ \pi_- \end{pmatrix}; \qquad f(\rho, x) = \begin{pmatrix} t(\rho, v) \\ -\xi(\rho, d) \\ s + t(\rho, v) \\ -\pi_- - \xi(\rho, d) \end{pmatrix}; \qquad X = C^{\text{proj}} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|}.$$

The variational inequality corresponding to (9) states that, at ρ^*,

$$\begin{aligned} t(\rho^*, v^*)^T(v - v^*) - \xi(\rho^*, d^*)^T(d - d^*) \geq 0, &\qquad (v, d) \in C^{\text{proj}}, \\ s^* = -t(\rho^*, v^*), & \\ \pi_-^* = -\xi(\rho^*, d^*), & \end{aligned}$$

so it is entirely equivalent to the VIP in (5). The reason for introducing the last two rows of the problem, that is, the extra variables (s, π_-), is that by doing so, we have direct access to the sensitivity of the travel costs, through the corresponding elements (s', π_-') of x'.

The sensitivity problem has the form of (11), with

$$r(\rho', x') = \begin{pmatrix} \nabla_\rho t(\rho^*, v^*)\rho' + \nabla_v t(\rho^*, v^*)v' \\ -[\nabla_\rho \xi(\rho^*, d^*)\rho' + \nabla_d \xi(\rho^*, d^*)d'] \\ s' + \nabla_\rho t(\rho^*, v^*)\rho' + \nabla_v t(\rho^*, v^*)v' \\ -\pi_-' - [\nabla_\rho \xi(\rho^*, d^*)\rho' + \nabla_d \xi(\rho^*, d^*)d'] \end{pmatrix}, \tag{16}$$

and

$$K = \left\{ \begin{pmatrix} v' \\ d' \\ s' \\ \pi_-' \end{pmatrix} \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \,\middle|\, \exists h' \in H' : \ \Gamma^T h' = d'; v' = \Lambda h' \right\}.$$

The sensitivity optimization problem in (v', d') is (15), and the value of (s', π_-') is then given by

$$\begin{aligned} s' &= -[\nabla_\rho t(\rho^*, v^*)\rho' + \nabla_v t(\rho^*, v^*)v'], \\ \pi_-' &= -[\nabla_\rho \xi(\rho^*, d^*)\rho' + \nabla_d \xi(\rho^*, d^*)d'], \end{aligned}$$

that is, the cost perturbations are given by a kind of chain rule. The result to follow establishes that this chain rule provides uniquely given values of (s', π_-') even when v' is not unique.

Before we apply the sensitivity analysis results in the previous section to the present problem, we mention an important fact which allowed us in [JoP04] to provide stronger results for optimization formulations than for the general variational inequality models in [PaR02, PaR03, Pat04, Pat05]: for a differentiable, convex problem, the gradient of the objective function is invariant over the solution set (cf. [BuF91]). The following result stems from Josefsson and Patriksson [JoP04].

Theorem 1 (sensitivity of separable traffic equilibrium problems). *Let Assumption 1 hold, and consider an arbitrary vector $\rho^* \in \Re^d$. Then, the solution (v^*, d^*) to (4) is unique, and so are the (negative) travel cost entities $(s^*, \pi_-^*) = -(t(\rho^*, v^*), \xi(\rho^*, d^*))$. Let $\rho' \in \Re^d$ be an arbitrary perturbation.*

(a) *In the solution to (15), the travel cost perturbations (s', π_-') are unique; therefore, the values*

$$-s' = \nabla_\rho t(\rho^*, v^*)\rho' + \nabla_v t(\rho^*, v^*)v',$$
$$-\pi_-' = \nabla_\rho \xi(\rho^*, d^*)\rho' + \nabla_d \xi(\rho^*, d^*)d',$$

are the directional derivatives of, respectively, the equilibrium link and OD travel costs, at ρ^*, in the direction of ρ'.

(b) *Assume that the link travel cost function $t(\rho^*, \cdot)$ is such that*

$$\frac{\partial t_l(\rho^*, v_l^*)}{\partial v_l} > 0, \qquad l \in \mathcal{L}. \tag{17}$$

Assume further that the demand function $g(\rho^*, \cdot)$ is such that

$$\frac{\partial g_k(\rho^*, \pi_k^*)}{\partial \pi_k} < 0, \qquad k \in \mathcal{C}. \tag{18}$$

Then, in the solution to (15), the values of the link flow and demand perturbation v' and d' are unique; therefore, the value v' (respectively, d') is the directional derivative of the equilibrium link flow (respectively, demand), at ρ^*, in the direction of ρ'.

Note that the second-order condition (18) is equivalent to the condition that

$$\frac{\partial \xi_k(\rho^*, d_k^*)}{\partial d_k} < 0, \qquad k \in \mathcal{C}.$$

Obviously, this condition on the demand function derivative (or its inverse) is not needed in the case when we consider fixed demands.

An interesting aspect of this result is that the cost perturbations (s', π_-'), which are related to the perturbations (v', d') through a kind of chain rule, are *not* dependent on the perturbations (v', d') to be unique. This is in contrast to the type of analysis offered by Tobin and Friesz [ToF88], see also [Bel97, Section 5.4], where the sensitivity of the costs is considered an implication of that of the flows and demands. (Not to mention that it will sometimes fail, cf. [JoP04, Pat04, Pat05].)

In the Tables 1 and 2 we summarize the most important notation for the readers' convenience.

5 An illustrative example

The following small numerical example, taken from Josefsson and Patriksson [JoP04], illustrates the workings of our analysis.

Table 1. Network and traffic equilibrium notation

Notation	*Explanation*
$\mathcal{G} = (\mathcal{N}, \mathcal{L})$	directed graph of nodes and links
$\mathcal{C} \subset \mathcal{N} \times \mathcal{N}$	set of commodities/OD pairs
$\mathcal{R}_{pq}$	set of directed routes in OD pair $(p,q) \in \mathcal{C}$
$\rho \in \Re^d$	vector of control variables
$\varGamma \in \{0,1\}^{\lvert\mathcal{R}\rvert \times \lvert\mathcal{C}\rvert}$	route–OD pair incidence matrix
h_r, $r \in \mathcal{R}_{pq}$	flow on route r in the OD pair (p,q)
$c_r = c_r(\rho, h)$	travel cost of route r given control variable values and route flows
π_{pq}	least route cost in OD pair $(p,q) \in \mathcal{C}$ at equilibrium
$g_{pq} = g_{pq}(\rho, \pi)$	travel demand given control variable values and travel costs
$\xi_{pq} = \xi_{pq}(\rho, d)$	inverse travel demand given control variable values and demands
$\varLambda \in \{0,1\}^{\lvert\mathcal{L}\rvert \times \lvert\mathcal{R}\rvert}$	link–route incidence matrix
v_l, $l \in \mathcal{L}$	flow on link l
$t_l = t_l(\rho, v)$	travel cost of link l given control variable values and link flows
C	polyhedral set of feasible flows in terms of (h, v, d)
C^{proj}	polyhedral set of feasible flows in terms of (v, d)
$a \perp b$	orthogonality requirement

Table 2. Analysis notation

Notation	*Explanation*
$f : X \mapsto Y$	the mapping f maps a point in X to a point in Y
$f : X \rightrightarrows Y$	the mapping f maps a point in X to a subset of Y
x	problem variable
X	polyhedral feasible set in the variational problem
$f : X \mapsto \Re^n$	cost mapping in the variational problem
$N_X(x)$	normal cone to the set X at x
$T_X(x)$	tangent cone to the set X at x
$S(\rho)$	set of solutions to the variational problem given control variable values
ρ'	direction of change (perturbation) of control variables
x'	direction of change (perturbation) of solution vector in the variational problem
K	polyhedral feasible set in the perturbation problem
$DS(\rho^* \vert x^*)(\rho')$	set of solutions to the perturbation problem given perturbation ρ'
$z^{\perp}$	orthogonal subspace to $z \in \Re^n$
s	vector of (negative) link travel costs
π_{-}	vector of (negative) OD least travel costs

The network of Braess [Bra68] is classic in the analysis of system optimal solutions. Consider the network in Figure 1.

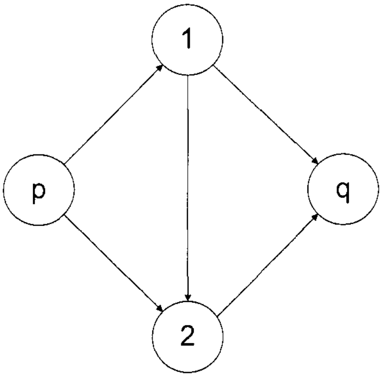


Fig. 1. Braess' traffic network

For this problem [where link $(1, 2)$ has a travel cost that is deliberately chosen so that the third route is not used but still has the same cost], we have the data given in Table 3.

Table 3. Network data

Link	$t_{ij}(\rho, v_{ij})$	*OD pair* d_{pq}
1: $(p, 1)$	$10v_{p1}$	1: (p, q) 6
2: $(p, 2)$	$50 + v_{p2}$	
3: $(1, q)$	$50 + v_{1q}$	
4: $(2, q)$	$10v_{2q}$	
5: $(1, 2)$	$23 + \rho + v_{12}$	

The data corresponds to an instance of the fixed (and unperturbed) demand traffic equilibrium problem, which can be written as that to

$$\min_{v} \phi(v) := \sum_{l \in \mathcal{L}} \int_0^{v_l} t_l(\rho, s)\, ds, \tag{19a}$$

$$\text{s.t. } \Gamma^T h = d, \tag{19b}$$

$$v = \Lambda h, \tag{19c}$$

$$h \geq 0^{|\mathcal{R}|}, \tag{19d}$$

where $d \in \Re^{|\mathcal{C}|}_{++}$ is the vector of demands.

Solving the fixed demand traffic equilibrium problem with $\rho = \rho^* = 0$, we obtain the link flow solution $v^* = (3, 3, 3, 3, 0)^T$. The cost of the three routes $\{1, 3\}$, $\{2, 4\}$, and $\{1, 5, 4\}$, are in fact the same, namely 83, but the route flows are 3 on each of the first two, and zero on the third. (This unique route flow solution is non-strictly complementary.)

Since the parameter ρ is present on link $(1,2)$, which has link flow zero at equilibrium, it looks clear that a positive value of ρ' should lead to no changes [since the flow on link $(1,2)$ cannot be negative] whereas a negative value of ρ' should imply that v'_{12} is positive. Indeed, that is the case.

In this special case where demand is fixed and unperturbed, the sensitivity problem is that to

$$\min_{v'} \phi'(v') := [\nabla_\rho t(\rho^*, v^*)\rho']^T v' + \frac{1}{2}\sum_{l\in\mathcal{L}} \frac{\partial t_l(\rho^*, v_l^*)}{\partial v_l}(v'_l)^2, \tag{20a}$$

$$\text{s.t. } \Gamma^T h' = 0^{|\mathcal{C}|}, \tag{20b}$$

$$v' = \Lambda h', \tag{20c}$$

$$h \in H'. \tag{20d}$$

For any $\rho' \in \Re$, therefore, we have that

$$\phi'(v') = \rho' v'_{12} + 5(v'_{p1})^2 + \frac{1}{2}(v'_{p2})^2 + \frac{1}{2}(v'_{1q})^2 + 5(v'_{2q})^2 + \frac{1}{2}(v'_{12})^2,$$

and the constraints specify that

$$\begin{aligned} h'_1 + h'_2 + h'_3 &= 0,\\ -v'_{p1} + h'_1 + h'_3 &= 0,\\ -v'_{p2} + h'_2 &= 0,\\ -v'_{1q} + h'_1 &= 0,\\ -v'_{2q} + h'_2 + h'_3 &= 0,\\ -v'_{12} + h'_3 &= 0,\\ h'_3 &\geq 0. \end{aligned}$$

Letting $\rho' = 1$, the optimal solution is $h' = 0^T$ and $v' = 0^T$.
Letting $\rho' = -1$, the optimal solution is $h' = \frac{1}{26}(-1,-1,2)^T$ and $v' = \frac{1}{13}(1,-1,-1,1,2)^T$.

This is therefore also a case where the traffic equilibrium solution is non-differentiable, since clearly the directional derivative mapping is not linear.

6 A sensitivity analysis tool

In [JoP04], the disaggregate simplicial decomposition (DSD) method of [LaP92] for the fixed demand problem (19) was taken as the building block of the sensitivity analysis tool. (In the case of elastic demands, one can still solve a fixed demand problem, by first utilizing the fixed demand transformation of Gartner [Gar80].) It has the advantage of utilizing route flow information, and therefore the close resemblance between the original problem (19) and the sensitivity problem (15) can be utilized fully.

The DSD algorithm was recoded in Matlab for experimentational purposes, fully aware of the fact that the CPU time will be perhaps two orders of magnitude larger than a final C or Fortran implementation. We refer to the paper [LaP92] for the basics of the DSD algorithm, but remind the reader that the most central points of the algorithm are the following: at some iteration point (h^τ, v^τ) of consistent route and link flows, the current travel costs are used to solve a shortest route problem for each origin node. The routes that then are generated are compared to sets $\hat{\mathcal{R}}_{pq} \subset \mathcal{R}_{pq}$ that have been generated and stored previously, and those sets are augmented with any routes that were not known already. (This process is the column, or route, generation one.) With those subsets of the routes at hand, the restricted master problem (RMP)—which is the original problem (19) except that $\hat{\mathcal{R}}_{pq}$ replaces $\mathcal{R}_{pq}$ for each $(p, q) \in \mathcal{C}$—is solved using one of several methodologies implemented. The highly structured RMP is either solved by using a gradient projection method or a diagonalized Newton method. (In practice, it appears that the former is the best for smaller networks, but that the Newton method wins for large enough cases.)

The similarity of the sensitivity problem to the equilibrium problem meant that much of the code from the DSD algorithm implementation could be reused. Of course, in the sensitivity context, flow and cost derivatives take the place of flows and costs themselves. For simplicity, these derivatives can be considered "virtual" costs and flows. The restricted master problem solver code then only had to be altered slightly to allow the subset of the routes that were used at equilibrium to take on negative "flow" values.

In order to set up the sensitivity problem, except for the flows and costs, the main part concerns the routes to be included. Remember that only least-cost routes are valid, some of which have a non-negativity requirement (if it is unused). In order to construct this set of routes, we first included only those routes that were used in the equilibrium solution. In order to compensate for the possibility that the equilibrium solution might not perfectly identify these routes, a "fuzz"-factor was used. In other words, to determine whether a route was used, the route flow was compared not to zero but to a very small positive number, obtained by multiplying the OD-pair demand by a tiny factor. In other words, a sign restriction may be included for a route that has a very small amount of flow at the terminal flow of the DSD algorithm. Any remaining set of routes that are potentially interesting for the sensitivity problem could then be included based on a final shortest route calculation and a graph search, together with a "fuzz"-factor similar to the above, allowing for near-shortest routes to be included as well.

This set of routes then is the one that defines the sensitivity problem; no further route generation is necessary, and so the only problem left is a convex quadratic RMP with some variables being free and some being sign restricted. Virtually the same algorithm as in the RMP for the original problem can be used; the only special case stems from the sign restrictions. Further details on the implementation of this algorithm can be found in the first author's master's thesis [Jos03].

7 A dissection of the sensitivity analysis of Tobin and Friesz

7.1 The analysis

We show, by means of both analytical and numerical tools, some examples in which the sensitivity analysis presented in [ToF88] requires too strong assumptions.

The strong monotonicity condition

The analysis is performed on a problem similar to (19) but where the fixed demand is also perturbed. In order to ensure local uniqueness, they introduce the following condition:

(Condition 1—strong monotonicity) $t(\rho, \cdot)$ is strongly monotone in a neighbourhood of ρ^*.

This condition is stronger than necessary, as we have already seen.

The strict complementarity condition

The analysis is based on first selecting a particular equilibrium route flow solution. Among the conditions stated, the route flow is supposed to be strictly complementary. The definition is however not the one commonly used, it being the following: a route flow solution h^* is *strictly complementary* if and only if that it is complementary (that is, that $0 \leq h_r^* \perp [c_r(\rho^*, h^*) - \pi_{pq}(\rho^*)] \geq 0$ holds for all $r \in \mathcal{R}_{pq}$, $(p,q) \in \mathcal{C}$), and

$$h_r^* + [c_r(\rho^*, h^*) - \pi_{pq}(\rho^*)] > 0, \qquad r \in \mathcal{R}_{pq}, \quad (p,q) \in \mathcal{C}. \tag{21}$$

In other words, our use of the term strict complementarity means that for an arbitrary route $r \in \mathcal{R}_{pq}$, it is either used ($h_r^* > 0$) or it is more expensive than the least costly route used in the OD pair $[c_r(\rho^*, h^*) > \pi_{pq}(\rho^*)]$.

Tobin and Friesz state a definition of traffic equilibrium in terms of *total* link flows v only, and which unfortunately is not consistent with the standard definition, given in (2). Their definition of a user equilibrium in terms of the vector v is that there exists a vector $\lambda^* \in \Re^{|\mathcal{N}|}$ such that for every link $l = (i,j) \in \mathcal{L}$,

$$\begin{aligned} v_l^* = 0 \qquad &\Longrightarrow \qquad t_l(v^*) \geq \lambda_j^* - \lambda_i^*, \\ v_l^* > 0 \qquad &\Longrightarrow \qquad t_l(v^*) = \lambda_j^* - \lambda_i^*. \end{aligned}$$

An inherent problem with this definition is that the *aggregated* potential differences, $\lambda_j^* - \lambda_i^*$, are not consistent with our node price vectors π_k, $k \in \mathcal{C}$, associated with the shortest route problem for OD pair k at v^*.

In order to illustrate this inconsistency, in comparing our definition with theirs, we must take into account that their definition assumes that we utilize the link–node representation of flows while we utilize the link–route representation. We can easily define vectors π_k of equilibrium costs for each commodity also in the former setting, where then these vectors are of size $|\mathcal{N}|$, since they are defined by node prices. Given an OD pair $k \in \mathcal{C}$ defined by an origin s and a destination t, both being nodes in $\mathcal{N}$, we would define these node prices by π_{ik} for each node $i \in \mathcal{N}$, and the OD travel cost would be given by the difference between the node prices at the terminal and initial node, that is, by $\pi_{tk} - \pi_{sk}$. Further, along a shortest route in which link $(i,j) \in \mathcal{L}$ is used, we would have that $t_{ij}(v_{ij}) = \pi_{jk} - \pi_{ik}$.

So, the above inconsistency would be described as follows: at a traffic equilibrium, $\pi_{jk}^* - \pi_{ik}^* \neq \pi_{j\kappa}^* - \pi_{i\kappa}^*$ may hold for two OD pairs k and κ. (For example, it can happen as soon as link (i,j) lies on a shortest route in the OD pair k but not in κ.)

Based on the equilibrium definition, however, the authors define a strict complementarity criterion for the perturbed problem:

(Condition 2—strict complementarity) For each link $l = (i,j) \in \mathcal{L}$, $v_l^* = 0 \implies t_l(\rho^*, v^*) > \lambda_j^* - \lambda_i^*$ holds.

So, whenever the total link flow vector v^* is positive, this condition is satisfied. Clearly, it is therefore not compatible with the strict complementarity condition (21).

In any case, the strict complementarity condition is not a necessary condition for the differentiability of the traffic equilibrium solution, although our strict complementarity condition is *sufficient*. An example below will illustrate this fact.

The linear independence condition

Next, we are asked to restrict the network $\mathcal{G}$ to $\mathcal{G}_+ = (\mathcal{N}, \mathcal{L}_+)$, where $l \in \mathcal{L}_+$ if and only if $v_l^* > 0$, that is, to the network corresponding to the links having a positive flow given ρ^*. Consequently, there are possibly some routes that will be removed as well. The $+$ notation to follow reflects this restriction.

Under the assumptions stated so far, the set $H_+^*(\rho^*)$ of equilibrium route flows is a bounded polyhedron. The next condition states that an equilibrium route flow vector h_+^* is selected such that it is a “non-degenerate extreme point” of $H_+^*(\rho^*)$:

(Condition 3—linear independence) An equilibrium route flow h_+^* is chosen such that it is an extreme point of $H_+^*(\rho^*)$ which has exactly as many routes with a positive flow as the rank of the matrix $[\Lambda_+^T \mid \Gamma_+]$.

The rank of this matrix is never higher than the number of links with a positive flow at v^* plus $|\mathcal{C}|$. The authors state an LP problem that can be used to generate such a point, but also remark in their Theorem 6 that the sensitivity values do not depend on this choice, as long as it is an extreme point of $H_+^*(\rho^*)$.

A final restriction is then made, such that we remove all the indices in the vector h_+^* for which the flow is zero. (We do not change the notation to reflect this restriction.) The sensitivity problem is then finally set up as follows:

$$\begin{pmatrix} \nabla_\rho h_+ \\ \nabla_\rho \pi \end{pmatrix} = \begin{pmatrix} \nabla_h c_+(\rho^*, h_+^*) & -\Lambda_+^T \\ \Lambda_+ & 0 \end{pmatrix}^{-1} \begin{pmatrix} -\nabla_\rho c_+(\rho^*, h_+^*) \\ \nabla_\rho g(\rho^*) \end{pmatrix}. \tag{22}$$

7.2 Examples

A case of differentiability where strict complementarity does not hold

To show that strict complementarity is not necessary for differentiability, we consider the network depicted in Figure 2.

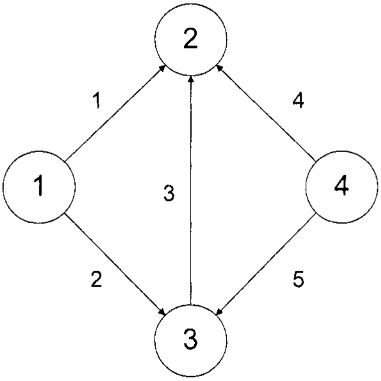


Fig. 2. A traffic network

There are 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.

With $\rho^* = 0$, the unperturbed traffic equilibrium solution is $v^* = (1,1,1,1,1)^T$. The route flow is unique: $h^* = (1,1,0,1)^T$. We see that the travel cost on route 3 is 2, as is the case for route 4, so this is a non-strictly complementary equilibrium solution. Since it is the unique route flow, we do not comply with the conditions (21) for strict complementarity.

In order to check if the solution 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 (20), thus being the directional derivative of v^* with respect to the direction $\rho' = 1$ at $\rho^* = 0$: $v' = \frac{1}{3}(-1, 1, 1, 0, 0)^T$. The effect, as we can see, of perturbing link 1's cost such that it becomes more expensive, is that of sending flow in the cycle $\{-1, 2, 3\}$, where the minus reflects that flow is sent backwards on link 1. When solving the sensitivity problem for $\rho' := -1$, we obtain the directional derivative $v' = \frac{1}{3}(1, -1, -1, 0, 0)^T$, that is, the negative of the directional derivative of v^* in the direction of $\rho' := 1$. This proves that the directional derivative mapping is linear, and thus that the derivative of v^* with respect to ρ' at $\rho^* = 0$ equals

$$\frac{\mathrm{d}\, v^*}{\mathrm{d}\, \rho} = \begin{pmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \end{pmatrix} .$$

At the same time, we have here shown that the sufficient matrix condition (13) indeed is only sufficient; it is not satisfied here because the set of feasible route flow perturbations is the entire space and the partial Jacobian of t with respect to v at the pair (v^*, ρ^*) is the non-positive definite diagonal matrix with diagonal entries $(2, 1, 0, 1, 1)$; yet the equilibrium solution is even differentiable.

This is then an example where the analysis formula (22) is not applicable, although the solution is differentiable.

A case of differentiability where the formula (22) fails

Consider the network shown in Figure 3.

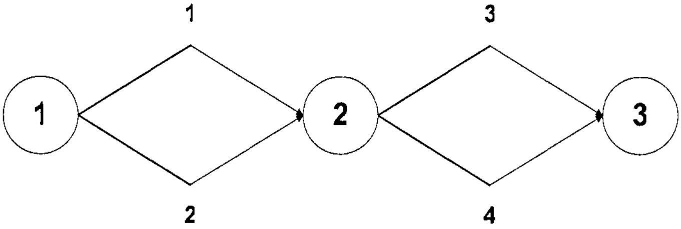


Fig. 3. Network for the first counter-example

There is a single OD pair, $(1,3)$, with a fixed demand of 2 units of flow. The link cost functions are given by

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

We have four routes: $\{1,3\}$, $\{1,4\}$, $\{2,3\}$, and $\{2,4\}$.

With $\rho^* = 0$, the unperturbed traffic equilibrium solution is $v^* = (1,1,1,1)^T$. We can easily see that the solution is differentiable; it is strictly complementary even. The derivative with respect to ρ at ρ^* moreover is

$$\begin{pmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}.$$

This is intuitive: if the value of ρ increases, then the flow on link 1 should decrease, whence link 2 must increase its flow with the same amount. If, on the other hand, the value of ρ decrease, the reverse should happen.

Consider then the workings of the formula (22) outlined above. We obviously fulfill Condition 1 on the travel cost functions. We also satisfy Condition 2, because $v^* > 0^{|\mathcal{L}|}$. Also, then, $\mathcal{G}_+ = \mathcal{G}$. We last try to comply with the linear independence Condition 3, by choosing the right equilibrium route flow solution. Note then that

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

has rank 3. So, we should find a route flow solution, h^*, in which exactly 3 routes have a positive flow. This is however impossible; the only alternatives are 2 or 4. To see why, suppose that the flow on the first route, $\{1,3\}$, is $\alpha \in [0,1]$. Then, the flows on the 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. This shows that for any value of $\alpha \in [0,1]$, the number of routes having a non-zero flow is either 2 or 4. Since we cannot comply with Condition 3, the formula (22) fails, even though the gradient exists.

The problems regarding the applicability of the formula (22) associated with the rank Condition 3 was first observed and commented on by [BeI97, p. 97]; our example however seems to be the first that has been worked out in detail.

A case of non-differentiability where the formula (22) may provide a result

Consider the network shown in Figure 4.

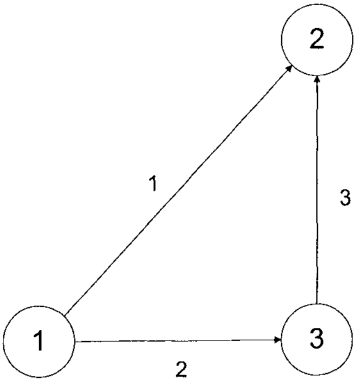


Fig. 4. Network for the second counter-example

In this example, there are three OD pairs, with the following fixed demands:

$$d_{12} := 1; \qquad d_{13} := 1; \qquad d_{32} := 1.$$

The link cost functions are given by

$$t_1(v_1, \rho) := 2v_1 + \rho; \qquad t_2(v_2) := v_2; \qquad t_3(v_3) := v_3.$$

(We thereby comply with Condition 1.) With $\rho^* = 0$, the unique equilibrium link volume is $v^* = (1, 1, 1)^T$. In this case, the route flow is unique: the flow on route $\{(1, 2)\}$ is 1; the flow on route $\{(1, 3), (3, 2)\}$ is 0; the flow on route $\{(1, 3)\}$ is 1; and the flow on route $\{(3, 2)\}$ is 1 as well.

This solution is non-strictly complementary by our definition, since the route $\{(1, 3), (3, 2)\}$ is of least cost but it cannot be used. It is however strictly complementary according to Condition 2, which we thereby satisfy.

We also see that a small negative perturbation in ρ would not affect the equilibrium solution, since the link $\{(1, 2)\}$ (that is, the first route in the first OD pair) is already utilized to send all the demand in the first OD pair. But if the perturbation is positive, we see that the flow in route $\{(1, 2)\}$ would decrease, and the flow on the route $\{(1, 3), (3, 2)\}$ would increase. This is then a case where the directional derivative (which of course exists) is not linear, so $\rho^* = 0$ is a point of non-differentiability.

What happens if we wish to apply the formula (22)? We here have that

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

which has rank 4. Since the flow on the route $\{(1, 3), (3, 2)\}$ is restricted to zero, in all fairness, the formula then breaks down. But it does not really do so for the right reason; there is every possibility of believing that the formula might still work if we either still include the route, or if we were to delete it. In both cases, the formula (22) does produce a result, which in none of the two cases can be interpreted as the value of the gradient at ρ^*.

7.3 The gradient formula of Cho, Smith, and Friesz

The sensitivity analysis of [CSF00] is somewhat related to that in [ToF88]. It replaces all the three conditions 1–3 mentioned in the previous section with weaker ones, and further provides an analysis entirely in link flows. We briefly discuss this analysis below.

(Condition 1—strict monotonicity) $t(\rho, \cdot)$ is strictly monotone in a neighbourhood of ρ^*. Further, the Jacobian matrix $\nabla_v t(\rho^*, v^*)$ is positive definite.

(Condition 2—strict complementarity) There exists a strictly complementary route flow $h^* \in H^*(\rho^*)$.

Notice that these two conditions together imply differentiability, but that they are stronger then necessary; the latter utilizes the classic definition of strict complementarity, as we do in this paper.

We are first asked to consider, as in [ToF88], the graph $\mathcal{G}_+$, which only includes links $l \in \mathcal{L}$ with $v_l^* > 0$ at ρ^*. In order to state the differences between the analysis in [CSF00] and [ToF88] more clearly, we do not introduce the $+$ notation here, and assume, from now on, that $\mathcal{G} = \mathcal{G}_+$.

Next, suppose that we, for each OD pair $(p, q) \in \mathcal{C}$, remove the routes $r \in \mathcal{R}_{pq}$ whose cost is higher than π_{pq}^*. Thus, we reach a network which we may denote by $\mathcal{G}_0$, in which the set $\mathcal{R}$ is replaced by the subset $\mathcal{R}_0$ of least-cost routes at v^*. By Condition 2, they must also be the routes with positive flow at h^*.

The sensitivity analysis proceeds with a further reduction:

(Condition 3—linear independence) Select a subset of the rows of Λ_0, such that the resulting matrix $[(\Lambda_0')^T \mid \Gamma_0]$ has full (column) rank.

Note that there is no requirement on the rank itself, and therefore this condition is milder than the Condition 3 in [ToF88].

The sensitivity formula is similar to that in (22), but provides the sensitivity in the link flow space directly, and therefore does not require the selection of a particular equilibrium route flow solution. It is however much more complicated in the sense that the translation between the spaces in h and v implies that several sub-matrices of, for example, Λ_0' must be constructed, collected, inverted and multiplied:

$$\begin{pmatrix} \nabla_\rho v \\ \nabla_\rho \pi \end{pmatrix} = \begin{pmatrix} \nabla_v t(\rho^*, v^*) & -[\Lambda_0'' N_1, -I]^T \\ [\Lambda_0'' N_1, -I] & 0 \end{pmatrix}^{-1} \begin{pmatrix} -\nabla_\rho t(\rho^*, v^*) \\ -\Lambda_0'' N_2 \nabla_\rho g(\rho^*) \end{pmatrix}, \tag{23a}$$

where Λ_0'' contains the rows of Λ_0 which are not present in Λ_0', and

$$\begin{aligned} N_1 &:= (\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T - \Lambda_0'\Gamma_0(\Gamma_0^T\Gamma_0)^{-1}\Gamma_0^T(\Lambda_0')^T]^{-1} \\ &\quad -\Gamma_0[\Gamma_0^T\Gamma_0 - \Gamma_0^T(\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T]^{-1}\Lambda_0'\Gamma_0]^{-1}\Gamma_0^T(\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T]^{-1}, \qquad (23b) \\ N_2 &:= -(\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T]^{-1}\Lambda_0'\Gamma_0[\Gamma_0^T\Gamma_0 - \Gamma_0^T(\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T]^{-1}\Lambda_0'\Gamma_0]^{-1} \\ &\quad +\Gamma_0[\Gamma_0^T\Gamma_0 - \Gamma_0^T(\Lambda_0')^T[\Lambda_0'(\Lambda_0')^T]^{-1}\Lambda_0'\Gamma_0]^{-1}. \qquad (23c) \end{aligned}$$

7.4 Conclusion

While being applicable to a wider selection of situations than the analysis in [ToF88], the analysis in [CSF00] is still not valid for problems with a non-strictly complementary equilibrium solution, since it relies on the implicit function theorem. Its main drawback is however its complexity; the formula (23) reached in [CSF00] is rather non-intuitive and computationally burdensome to use. It is also clear from the papers that have been written and referred to during the past two–three years that the analysis in [ToF88] is the one favoured, despite the fact that it is less generally applicable.

Interestingly, shortly after Tobin and Friesz published their paper, Qiu and Magnanti [QiM89] published the first paper which develops a sensitivity analysis of traffic equilibria based on Robinson's strong regularity condition (albeit under slightly stronger assumptions than necessary; cf. [Pat04]); it is based on a linearized traffic equilibrium model which is similar to (15) in the case of a separable problem. Their paper did however not get much attention from the transportation science community. (See [Pat04] for an account of the history of the sensitivity analysis of traffic equilibria, and a list of references its utilization.) One of the few who has utilized the results of Qiu and Magnanti is [Den94], who applied it in the context of OD matrix estimation. He compared numerically the Qiu/Magnanti sensitivity analysis formulas to that of Tobin/Friesz, and found that the former was significantly more robust and efficient to use. Our above findings clearly are supportive of that claim.

The paper [JoP04] provides a first application of our sensitivity formulas to network design problems, with encouraging results.

Acknowledgment. An anonymous reviewer provided remarks that improved the clarity of the presentation.

References

[BeI97] Bell, M.G.H., Iida, Y.: Transportation Network Analysis. John Wiley & Sons, Chichester, UK (1997)

[Bra68] Braess, D.: Über ein Paradox der Verkehrsplannung. Unternehmensforchung, **12**, 258–268 (1968)

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

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

[Den94] Denault, L.: Étude de deux méthods d'adjustement de matrices origina–destination à partir des flots des véhicules observés, PhD thesis, Centre de recherche sur les transports, Université de Montréal, Montréal, Canada (1994)

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

[Gar80] Gartner, N.H.: Optimal traffic assignment with elastic demands: A review. Part II: Algorithmic approaches. Transportation Science, **14**, 192–208 (1980)

[Jos03] Josefsson, M.: Sensitivity analysis of traffic equilibria. Master's thesis, Department of mathematics, Chalmers University of Technology, Gothenburg (2003)

[JoP04] Josefsson, M., Patriksson, M.: Sensitivity analysis of separable traffic equilibrium equilibria, with application to bilevel optimization in network design. report, Department of mathematics, Chalmers University of Technology, Gothenburg (2003). Revised (2004) for Transportation Research, B.

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

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

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

[LPR96] Luo, Z.Q., Pang, J.-S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge, UK (1996)

[Pat94] Patriksson, M.: The Traffic Assignment Problem—Models and Methods. Topics in Transportation, VSP BV, Utrecht, The Netherlands (1994)

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

[Pat05] Patriksson, M.: Traffic Equilibrium Problems: Analysis, Applications, and Optimization Algorithms. Springer-Verlag (under preparation).

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

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

[QiM89] Qiu, Y., Magnanti, T.L.: Sensitivity analysis for variational inequalities defined on polyhedral sets. Mathematics of Operations Research, **14**, 410–432 (1989)

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

[Rob85] Robinson, S.M.: Implicit B-differentiability in generalized equations. Technical Summary Report No. 2854, Mathematics Research Center, University of Wisconsin at Madison, Madison, WI (1985)

[ToF88] Tobin, R.L., Friesz, T.L.: Sensitivity analysis for equilibrium network flow. Transportation Science, **22**, 242–250 (1988)

[Yen95] Yen, N.D.: Lipschitz continuity of solutions of variational inequalities with a parametric polyhedral constraint. Mathematics of Operations Research, **20**, 695–708 (1995)

Park and Ride for the Day Period and Morning-Evening Commute

André de Palma[1] and Yurii Nesterov[2]

[1] Institut Universitaire de France & THEMA, Université de Cergy-Pontoise, 33 Bd du Port, F-95011, Cergy-Pontoise, France, andre.depalma@eco.u-cergy.fr
[2] Center for Operations Research and Econometrics, Université Catholique de Louvain, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium
nesterov@core.ucl.ac.be

Summary. We propose an application of a new set of models referred to as "Stable Dynamics" which provide a flexible yet rigorous way to model traffic congestion in large urban areas. Data requirements are extremely low, since supply and demand data can be given by GIS systems. This approach is based on the requirements that (1) the maximum entering flow for each link is given and that (2) Wardrop principle holds. In this paper, we supplement this basic model by parking choice. We focus on the case where the commuters use private and public transportation from the origin to the destination (and back to the origin). We propose a consistent model of both for the day commuting and the morning-evening commuting and show that such extension can be formulated as standard convex mathematical problems.

Key words: Morning and Evening Commute, Traffic Congestion, Stable Dynamics, Parking, Park and Ride.

1 Introduction

Parking plays an increasingly important role in the study of traffic congestion. This is due to the fact that the number of cars has dramatically increased over the last decades (the increase in car ownership remains of course very different across countries, even in Western Europe). As a result, traffic conditions tend to deteriorate in many cities, and as a corollary, the loss of time while searching for parking has substantially increased in many congested cities. Economists and traffic engineers have advocated road pricing since the last decade. Early research in this direction is due to Vickrey and Roth (see [Vic59, Rot65]). However, road pricing remains, with a few exceptions, very difficult to implement in practice, as discussed at length for example in several European projects, such as MC-ICAM or REVENUE. Transportation experts tend to agree that such frictions are not only the result of technological constraints, but especially of legal and sociological barriers or constraints. Some

Mathematical and Computational Models for Congestion Charging, pp. 143-158

exceptions exist for example in Hong Kong, Singapore, California, Norway and more recently London, where road pricing has been successfully implemented. On the whole, public acceptance of road pricing remains low, even if it seems to have improved recently, and even if transport management has improved substantially during the last decades. Current models are focused on the standard commute, without taking park and ride into account as well as many other subtle aspects of activity patterns. In this paper, we wish to study some simple dimensions of this complex decision process, for at least two reasons.

(a) One major impact of road pricing is the modal shift, which can be encouraged by park and ride policies.
(b) Parking can be (and have been) priced with a much higher public acceptance than road pricing, and is therefore an interesting alternative to take into account in the general picture.

There are few recent developments of the economic analysis of parking pricing (see [AR99, YTT91]; for an analysis of parking pricing with time of the day dependent congestion, see also [APL91]). The perception of parking has evolved over time. About two decades ago, traffic planners believed that parking problems were a consequence of insufficient parking places. This naive view is true only if one neglects the medium and long run adjustment decisions made by the drivers (such as mode choice and relocation). Over the last few years, it has been recognized that the transport systems are more complex, in the sense that more parking places in a downtown reduces search time for parking (and cruising), which decreases the generalized cost of cars and therefore induces an undesirable modal shift in favor of automobiles [AH90]. Pricing parking can, however, be used to control congestion. Parking fees have proven to be much more acceptable, and they potentially provide easy-to-implement second-best policy instruments, to regulate automobile congestion [AG99]. Park and Ride may induce mode shift downtown. Still, adequate management of parking remains a complex task since it requires optimization along several dimensions: the optimal location of parking lots, their optimal capacity and the optimal parking charge (which, in a dynamic context, may depend on the time of the day and on the length of stay).

Many studies have been performed and models developed to study the driver's choice behavior; in particular search behavior has attracted a lot of attention (see [BMS81], [TR98], and [AP91]). However, those models are often complex and are therefore not applicable to large networks. We have preferred to develop a more comprehensive and consistent parking model at the expense of simplifying drivers' behavior when searching for parking. In particular, we focus on the choice of a parking place but do not take into account search behavior (although search costs and parking congestions are included in our setting).

The choice of a parking place is a binding decision, since it provides constraints on the subsequent trips, or on the return trip. The return trip does affect the morning commute, and it is therefore (potentially) erroneous to treat the morning and evening commute separately. Most studies have concentrated on one trip from an origin to a destination, for example for the morning commute or in the context of a shopping trip involving the choice of a parking place [TL00, VO95]. However, to the best of our knowledge, no theoretical study has considered a unified mathematical formulation for the whole day trip chaining for a population of users travelling in a large congested network. It is not clear, for example, how the standard Beckmann objective functions should be extended to take into account interdependent morning and evening commute decisions. Here, we wish to study the combined choice of private transportation, of a parking space and of the alternative modes (e.g. public transportation) used once the car is parked. In order to accomplish this task, we have developed a combined model for the morning and the evening commute. We focus our attention on car drivers and consider that each driver has access to one (or several) parking places, with endogenous access time and given parking charges (of course this latter could be zero). The complexity of this problem is due to the fact that the morning and the evening peaks cannot be treated in isolation. If a driver chooses a specific parking lot in the morning, s/he necessarily has to return to the same lot after work. One way to solve this problem is to model the whole trip explicitly as follows:

(a) Home to parking lot trip;
(b) Parking lot to work and back to the parking lot trip (using another mode, such as bus, street car or even walking);
(c) Parking lot to home trip.

If the parking lot of a user is located at home, it means that this user does not travel by car, while if it is located at destination, it means that only car is used for the morning and evening commute.

We use in our description the Stable Dynamics approach introduced in [Nes00] (see also [NP03] for more examples and developments. We have shown in [NP03] that Stable Dynamics can be used to compute the equilibrium and the optimal solutions on simple networks such as the Braess network, only using logical arguments). This formulation is based on two simple assumptions, which characterize the supply side and drivers' behavior. First (see Assumption 1), it is assumed that each driver selects the minimum travel time route. Second (see Assumption 2), it is assumed that the outflow of each road cannot exceed a given value that we refer to as the capacity: either the outflow is lower than the capacity and the travel time is the free travel time or it is equal to the capacity and the travel time is larger or equal to the minimum travel time. These set of assumptions is sufficient to characterize equilibrium in a general large network (but not the out-of-equilibrium solution as defined in game theory).

The purpose of this paper is to develop an integrated treatment of different park-and-ride problems using the Stable Dynamics approach. In Section 2, we introduce the notation, briefly summarize the concepts underlying Stable Dynamics and present the corresponding mathematical formulation. In Section 3, we consider two cases: either the drivers are travelling during the same time period from the origin to the destination and back, or the morning and the evening can be treated as two independent periods (on the supply side, but of course not on the demand side). In Section 4, we consider a charging policy for the parking lots and extend in this direction the mathematical formulation of the park and ride commute. In Section 5 we provide a simple example which illustrates our model. Concluding remarks are presented in Section 6.

2 Notations and generalities

Suppose we have a transportation network $\mathcal{R}$ composed by set of nodes $\mathcal{N}$ and set of directed arcs $\mathcal{A}$:

$$\mathcal{R} = (\mathcal{N}, \mathcal{A}), \quad n = |\mathcal{N}|, \quad m = |\mathcal{A}|.$$

For this network we define the set of origin-destination pairs (OD-pairs):

$$\mathcal{OD} = \{(i,j),\ i,j \in \mathcal{N},\ i \neq j\}.$$

Each OD-pair (i,j) generates a demand $d_{(i,j)}$. Usually, this demand is considered as an average flow of drivers, which need to travel from origin node i to destination node j; so the demand is a non-negative real number. Sometimes it is necessary to work with *cumulative demand*, which is the total number of drivers $N_{(i,j)}$ travelling from origin i to destination j during a time window Δ. In this case, we assume that the drivers generate a constant flow in the network. Thus,

$$N_{(i,j)} = d_{(i,j)} \cdot \Delta.$$

Further, for each OD-pair (i,j) let us define a (finite) set of routes connecting i and j:

$$\mathcal{R}_{(i,j)} \subset \{0,1\}^m \subset R^m.$$

The β-component of the root incidence vector $a \in \mathcal{R}_{(i,j)}$, is equal to one if the arc β is included in this route; otherwise this component is equal to zero.

For each arc $\beta \in \mathcal{A}$ we introduce the minimal free traffic travel time $\bar{t}_\beta$. Let us introduce also the second characteristic of the arc, the maximal output flow $\bar{f}_\beta$. In urban network the maximal flow depends on the number of lanes of the road, the duration of the green light at the intersection, the weather conditions, etc. As it was shown in [Nes00, NP03], even from this restricted (and easily available) information we can retrieve the equilibrium travel time. Formally, this equilibrium solution can be derived from two assumptions. The first one is behavioral. We are looking for the solutions in which all drivers travel along the shortest paths computed with respect to existing (unknown) system of arc travel time.

Assumption 1 *Given an established arc travel time pattern $t = \{t_\beta\}_{\beta\in\mathcal{A}}$ in the network $\mathcal{R}$, $t \geq \bar{t} = \{\bar{t}_\beta\}_{\beta\in\mathcal{A}}$, each driver chooses one of the cheapest (with respect to t) paths to travel from origin to destination.*

The second assumption characterizes a performance of the arc and the congestion pattern. It says that either a road is uncongested, and then the travel time is equal to the minimal free traffic travel time, or, the route is congested and then the travel time on the arc must be found from a global analysis of the congestion in the network based on Assumption 1. Formally, our assumption looks as follows.

Assumption 2 *The flow f_β, observed on arc β, never exceeds $\bar{f}_\beta$. If $f_\beta < \bar{f}_\beta$, then the travel time on this arc is equal to $\bar{t}_\beta$ (that is the free traffic travel time). If $f_\beta = \bar{f}_\beta$, then the travel time t_β can be any value greater or equal to $\bar{t}_\beta$.*

Assumption 2 can be justified as follows. Consider a congested urban road equipped with a traffic light. Let us measure the output flow of this road. Clearly, in average the flow is constant and it depends *only* on the number of lanes, on the duration and the fraction of green light, etc. On the other hand, the travel time on this road *does not depend* on the output flow. It depends only on the size of the queue, which can be arbitrarily large.

We refer to the models, which compute the equilibrium patterns satisfying Assumptions 1 and 2, the *Stable Dynamic models* [Nes00, NP03]. The main advantage of these models is that they do not use any arc travel time functions. They just rely on basic and easily available parameters of the roads. One could think that this approach oversimplifies the complexity of the real world. However, we can see that in many cases we get intuitively correct solutions. Let us provide the reader with a simple example.

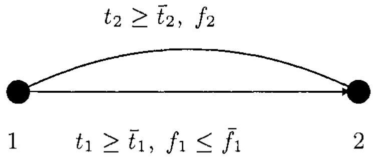

$t_2 \geq \bar{t}_2$, f_2
1
$t_1 \geq \bar{t}_1$, $f_1 \leq \bar{f}_1$
2

Fig. 1. Two routes in parallel

Example 1. Consider a network $\mathcal{R}$ consisting of two nodes 1 and 2. These nodes are connected by two parallel arcs directed to the second node. Let the performance characteristics of the arcs be related as follows:

$$\bar{t}_1 < \bar{t}_2, \quad 0 < \bar{f}_1 < \infty, \quad \bar{f}_2 = \infty.$$

Then the equilibrium solution (t^*, f^*) depends on the demand flow $d_{(1,2)}$ in the following way:

$$\begin{aligned} &1)\ d_{(1,2)} < \bar{f}_1 : t_1^* = \bar{t}_1, \quad f_1^* = d_{(1,2)}, f_2^* = 0, \\ &2)\ d_{(1,2)} > \bar{f}_1 : t_1^* = \bar{t}_2, \quad f_1^* = \bar{f}_1, \quad f_2^* = d_{(1,2)} - \bar{f}_1, \\ &3)\ d_{(1,2)} = \bar{f}_1 : t_1^* \in [\bar{t}_1, \bar{t}_2], f_1^* = d_{(1,2)}, f_2^* = 0. \end{aligned} \tag{1}$$

It is clear that for the third case in (1) we indeed cannot say more about t_1^* since it can depend on a constant queue accumulated at this arc. For more examples and discussion see Nesterov and Palma [NP03].

Note that, in general, we get user-equilibrium solutions, which are different from the social optimum. Indeed, for the second situation in (1), the optimal solution is given by

$$t_1^o = \bar{t}_1, \quad f_1^o = \bar{f}_1, \quad t_2^o = \bar{t}_2, \quad f_2^o = d_{(1,2)} - \bar{f}_1.$$

Thus, the optimal cost is $\bar{t}_1 \bar{f}_1 + \bar{t}_2 \cdot (d_{(1,2)} - \bar{f}_1)$, which is strictly less than the user-equilibrium cost $\bar{t}_2 \cdot d_{(1,2)}$.

We now show a way to find the Stable Dynamics solutions for general networks. Let us fix some arc travel time pattern t in network $\mathcal{R}$. Then, for each OD-pair (i, j) we can compute the shortest-path travel time. This value is a function of t and it has the following analytical form

$$T_{(i,j)}(t) = \min_{a \in \mathcal{R}_{(i,j)}} \langle a, t \rangle.$$

Thus, $T_{(i,j)}(t)$ is a *concave* piece-wise linear function of t. It is well defined for any $t \in R^m$. Therefore it is subdifferentiable on R^m. Its subdifferential

$$\partial T_{(i,j)}(t) = \text{Conv}\,\{a \in \mathcal{R}_{(i,j)} : \ \langle a, t \rangle = T_{(i,j)}(t)\} \tag{2}$$

has a very interesting interpretation. Any vector $g \in \partial T_{(i,j)}(t)$ describes an arc loading pattern of network $\mathcal{R}$ by a unit of flow from origin i to destination j, which satisfies Assumption 1.

Let us introduce now the cost function

$$C(t) = \sum_{(i,j) \in \mathcal{OD}} d_{(i,j)} T_{(i,j)}(t).$$

Consider the following *max-flow* arc performance model [Nes00]:

$$t_\beta \geq \bar{t}_\beta, \quad 0 \leq f_\beta \leq \bar{f}_\beta, \quad \beta \in \mathcal{A}. \tag{3}$$

Using the cost function $C(t)$ and the constraints (3), we can compose an optimization problem, the solution of which satisfies Assumptions 1, 2.

Theorem 1. *The arc travel time t^* and the arc flow vector f^* is an equilibrium solution of the model (3) if and only if t^* is a solution to the problem*

$$\max_t [C(t) - \langle \bar{f}, t - \bar{t} \rangle : t \geq \bar{t}], \tag{4}$$

and $f^* = \bar{f} - s^*$, where $s^* \geq 0$ is a vector of optimal dual multipliers for the inequality constraints in (4).

The proof of this theorem can be found in [Nes00]. Note that the optimization problem (4) has an interesting interpretation. The objective function in (4) is composed of two terms. The first one, the function $C(t)$, is the *loading* of the network; that is the total number of drivers travelling in the network at a given moment. The term $\langle \bar{f}, t - \bar{t} \rangle$ represents the total number of drivers waiting in the queues at the same moment. Thus, their difference,

$$\psi(t) = C(t) - \langle \bar{f}, t - \bar{t} \rangle$$

represents the number of cars involved in free traffic. In other words, Theorem 1 says that at user equilibrium this number is maximal.

The goal of this paper is to develop a Park-And-Ride model in the framework of Stable Dynamics.

3 Park and ride models

We are going to model the situation when a driver can leave his car at a special parking place and continue the trip to destination by public transport. The main difference with respect to the standard models is that now we have to take into account the way drivers come back. In order to describe our models, we need to introduce additional data:

1. The list of destinations $\mathcal{D} \subseteq \mathcal{N}$, $d = |\mathcal{D}|$.
2. The list of parking lots $\mathcal{P} \subseteq \mathcal{N}$, $p = |\mathcal{P}|$.
3. The cost π_k of parking lot $k \in \mathcal{P}$, $\pi = \{\pi_k\}_{k \in \mathcal{P}} \in R^p$.
4. A two-way cost $c_{(k,j)}$, $k \in \mathcal{P}$, $j \in \mathcal{D}$, for travelling between the parking lot k and destination j by public transport.

We need to use a monetary value of time per unit flow, which is denoted by α. It is assumed to be the same for all drivers.

In accordance to activity pattern of the drivers, we consider two different Park-And-Ride models.

3.1 Day period

In this situation, the drivers go to their destinations and come back in the same period of the day. So, in both directions they observe on the roads the same driving conditions. Let characterize these conditions by some arc travel time vector $t \in R^m$. Then the cost function for OD-pair (i,j) is as follows:

$$T^D_{(i,j)}(\pi, t) = \min_{k \in \mathcal{P}} \left[\alpha \cdot T_{(i,k)}(t) + \pi_k + c_{(k,j)} + \alpha \cdot T_{(k,i)}(t) \right]. \qquad (5)$$

Note that this function is qualitatively different from $T_{(i,j)}(t)$ since it *cannot* be implemented as a shortest path function of a single network. Nevertheless, $T^D_{(i,j)}(\pi, t)$ is still a concave piece-wise linear function of t, and it can be computed by a standard shortest-path technique.

Since in the current situation the strategies of drivers become more complicated, we need to modify Assumption 1. Indeed, the strategy now includes the choice of the parking place and the choice of the two-way route. Thus, we need to assume the following.

Assumption 3 *Given an established arc travel time pattern* $t = \{t_\beta\}_{\beta \in \mathcal{A}}$ *in the network* $\mathcal{R}$, $t \geq \bar{t} = \{\bar{t}_\beta\}_{\beta \in \mathcal{A}}$, *and the systems of prices* $\{\pi_k\}$ *and* $\{c_{(k,j)}\}$, *each driver chooses one of the cheapest strategies to travel from origin to destination and come back to the origin.*

In this case, similar to the Stable Dynamics approach, we can find an equilibrium solution from an appropriate convex optimization problem. Indeed, let us form the total cost function of our model:

$$C^D(\pi, t) = \sum_{(i,j) \in \mathcal{OD}} d_{(i,j)} T^D_{(i,j)}(\pi, t).$$

Then, taking into account Assumptions 2 and 3, can form the following convex optimization problem:

$$\text{Find } \phi_D(\pi) = \max_t \left[C^D(\pi, t) - \alpha \cdot \langle \bar{f}, t - \bar{t} \rangle : \; t \geq \bar{t} \right]. \qquad (6)$$

Theorem 2. *Let* t^* *be the optimal solution to the problem (6) and* $f^* = \bar{f} - s^*/\alpha$, *where* $s^* \geq 0$ *is a vector of optimal dual multipliers for the inequality constraints in (6). Then the pair* (t^*, f^*) *delivers an equilibrium solution to our problem in the following sense:*

(a) *The arc travel time vector* t^* *and the arc flow pattern* f^* *satisfy Assumption 2.*

(b) *The arc flow pattern* f^* *is composed of OD-flows* $\{g^*_{(i,j)}\}_{(i,j) \in \mathcal{OD}}$,

$$f^* = \tfrac{1}{\alpha} \cdot \sum_{(i,j) \in \mathcal{OD}} d_{(i,j)} g^*_{(i,j)}, \quad g^*_{(i,j)} \in \partial_t T^D_{(i,j)}(\pi, t^*), \qquad (7)$$

which satisfy Assumption 3 with respect to the cost functions (5).

Proof. The proof of this theorem consists of a straightforward application of Karush-Kuhn-Tucker conditions to problem (6), taking into account the flow interpretation of subgradients of functions $T^D_{(i,j)}(\pi, t^*)$. Indeed, let us fix π and write down the Lagrangean for problem (6):

$$\mathcal{L}(t,s) = C^D(\pi,t) - \alpha \cdot \langle \bar{f}, t - \bar{t} \rangle + \langle s, t - \bar{t} \rangle, \quad t \in R^m, \ s \geq 0 \in R^m.$$

By the Karush-Kuhn-Tucker conditions, for the optimal solution t^* of problem (6) there exists a vector of dual multipliers $s^* \geq 0$ and a vector $g^* \in \partial_t C^D(\pi, t^*)$ such that

$$\begin{aligned} g^* &= \alpha \cdot \bar{f} - s^*, \\ s_i^*(t_i^* - \bar{t}_i) &= 0, \ i = 1, \ldots, m. \end{aligned} \tag{8}$$

Now we need to find an interpretation of the vector g^*. Clearly,

$$\begin{gathered} g^* \in \partial_t C^D(\pi, t^*) = \sum_{(i,j) \in \mathcal{OD}} d^{(i,j)} \partial_t T^D_{(i,j)}(\pi, t^*), \\ \partial_t T^D_{(i,j)}(\pi, t^*) = \alpha \cdot \operatorname{Conv} \{ \partial T_{(i,k)}(t^*) + \partial T_{(k,j)}(t^*) : \\ \alpha T_{(i,k)}(t^*) + \pi_k + c_{(k,j)} + \alpha T_{(k,j)}(t^*) = T^D_{(i,j)}(\pi, t^*) \}, \quad (i,j) \in \mathcal{OD}. \end{gathered}$$

In view of representation (2), the set $\frac{1}{\alpha} \partial_t T^D_{(i,j)}(\pi, t^*)$ is composed by all arc loads of the network by a unit of flow, which goes from origin i to destination j and comes back, and which satisfies Assumption 3. Thus, expression (7) is justified. Item (a) of the theorem follows from the second line of conditions (8). ■

3.2 Morning-evening peak periods

In this situation, the drivers go to their destinations in the morning and come back in the evening. Hence, the driving conditions for two directions can be completely different. Therefore we introduce for the morning period an arc travel time variable t^M, and for the evening period another arc travel time variable t^E. Then the cost function for OD-pair (i,j) is as follows:

$$T^{ME}_{(i,j)}(\pi, t) = \min_{k \in \mathcal{P}} \left[\alpha \cdot T_{(i,k)}(t^M) + \pi_k + c_{(k,j)} + \alpha \cdot T_{(k,i)}(t^E) \right], \quad t = (t^M, t^E). \tag{9}$$

As before, $T^{ME}_{(i,j)}(\pi, t)$ is a concave piece-wise linear function of t.

We need to use the following modification of Assumption 3.

Assumption 4 *Given established arc travel time patterns t^M and t^E for the morning and the evening peak periods, and the systems of prices $\{\pi_k\}$ and $\{c_{(k,j)}\}$, each driver chooses one of the cheapest strategies to travel from origin to destination and to come back.*

Since we work with different periods of day, we need to pass to cumulative demands. Then, the total cost function of our model becomes

$$C^{ME}(\pi,t) = \sum_{(i,j)\in \mathcal{OD}} N_{(i,j)} T^{ME}_{(i,j)}(\pi,t).$$

Thus, taking into account Assumption 2, we come to the following convex optimization problem:

$$\text{Find } \phi_{ME}(\pi) = \max_t \left[C^{ME}(\pi,t) - \alpha \cdot \Delta \cdot \langle \bar{f}, t^M + t^E - 2\bar{t}\rangle : \; t^M \geq \bar{t}, \; t^E \geq \bar{t} \right]. \tag{10}$$

Theorem 3. *Let $t^* = (t^{M*}, t^{E*})$ be an optimal solution to the problem (10), $s^{M*} \geq 0$ be a vector of optimal dual multipliers for the inequality constraints $t^M \geq \bar{t}$, and $s^{E*} \geq 0$ be a vector of optimal dual multipliers for the inequality constraints $t^E \geq \bar{t}$. Denote $f^{M*} = \bar{f} - s^{M*}/(\alpha\Delta)$, and $f^{E*} = \bar{f} - s^{E*}/(\alpha\Delta)$. Then the pattern $(t^{M*}, t^{E*}, f^{M*}, f^{E*})$ delivers an equilibrium solution to our problem in the following sense:*

(a) *The patterns (t^{M*}, f^{M*}) and (t^{E*}, f^{E*}) satisfy Assumption 2.*
(b) *The arc flow patterns f^{M*} and f^{E*} are composed of OD-flows, which satisfy Assumption 4 with respect to the cost functions (9).*

The proof of this theorem is very similar to that of Theorem 2.

4 Pricing policy

Note that the solutions of problems (6), (10), depend on π, the prices of the parking places. Moreover, since functions $T^D_{(i,j)}(\pi,t)$ and $T^{ME}_{(i,j)}(\pi,t)$ are jointly concave in (π,t), the functions $\phi_D(\pi)$ and $\phi_{ME}(\pi)$ are concave in π. This opens a possibility to control the solutions of these problems as functions of π. Let us show how we can solve, for example, the problem of filling the parking lots. For simplicity, we do that for the day-period model (see Section 3.1).

For each parking lot $k \in \mathcal{P}$ we introduce the following characteristics:

1. n_k, the number of parking places in the lot.
2. δ_k, the upper bound for the time spent in the parking lot.

Using these characteristics, we can compute

$$\bar{\eta}_k = n_k/\delta_k,$$

the upper bound for the car flow through this parking lot. We can define also $\bar{\pi}_k$, the lower bound for the price of the parking place at lot k. As usual, we denote

$$\bar{\eta} = \{\bar{\eta}_k\}_{k\in\mathcal{P}}, \quad \bar{\pi} = \{\bar{\pi}_k\}_{k\in\mathcal{P}}.$$

The question we want to answer is as follows: *Which system of parking prices $\pi \geq \bar{\pi}$ ensures the absence of parking congestion?* In other words, we want to guarantee that the car flow through any parking lot k does not exceed $\bar{\eta}_k$.

Consider the following convex optimization problem:

$$\max_{\pi,t} \left[C^D(\pi,t) - \alpha \cdot \langle \bar{f}, t-\bar{t} \rangle - \langle \bar{\eta}, \pi - \bar{\pi} \rangle : \; t \geq \bar{t}, \quad \pi \geq \bar{\pi} \right]. \tag{11}$$

Theorem 4. *Let (π^*, t^*) be the optimal solution to the problem (11), $s^* \geq 0$ be a vector of optimal dual multipliers for the inequality constraints $t \geq \bar{t}$, and $\theta^* \geq 0$ be a vector of optimal dual multipliers for the inequality constraint $\pi \geq \bar{\pi}$. Denote $f^* = \bar{f} - s^*/\alpha$ and $\eta^* = \bar{\eta} - \theta^*$. Then the pattern $(\pi^*, t^*, f^*, \eta^*)$ delivers an equilibrium solution to problem (11) in the following sense:*

(a) *The arc travel time vector t^* and the arc flow pattern f^* satisfy Assumption 2.*
(b) *The arc flow pattern f^* is composed by OD-flows, which satisfy Assumption 3 with respect to the cost $C^D(\pi^*, t^*)$.*
(c) *The parking flow pattern η^* does not exceed the bound $\bar{\eta}$.*
(d) *The additional toll $\sigma_k = \pi_k - \bar{\pi}_k$ is strictly positive only for congested parking lots, that is $\eta_k^* = \bar{\eta}_k$.*

The proof of this theorem consists in a straightforward application of Karush-Kuhn-Tucker conditions and follows the lines of the proof of Theorem 2.

As for the pure Stable Dynamic model (4), we can provide the problem (11) with some interpretation. The objective function in this problem is composed of three terms. The first one, $C^D(\pi^*, t^*)$, represents the total flow of expenses spent in the network, which include the value of travel time, parking place and the cost of public transportation. From this amount we subtract the cost of the waiting time (social lost) and the extra charges for the parking places (pure transfer). Thus, we maximize an "efficient" payment of the drivers, which can be interpreted as a monetary value of the mobility in the system generated each unit of time.

Clearly, a similar pricing model can be developed for the situation described in Section 3.2.

5 Simple illustration of the model

Let us consider a simple example. Our network consists of three nodes. Node 1 is the origin. A Park-And-Ride facility is located at intermediate node 2. And there is a parking lot available at destination 3. All parking lots have unlimited capacities. We denote by π_i the price of parking at node i, $i = 2, 3$. The price of a two-way public transportation ticket from node 2 to 3 is $c_{2,3}$.

Our network is shown on Figure 2. All nodes are connected by two-way streets. Both directions of each street have identical characteristics (free traffic travel time and capacity). Nodes 1 and 2 are connected by a street with parameters $\bar{t}_1 > 0$ and $\bar{f}_1 = \infty$. Nodes 2 and 3 are connected by two streets. The short one has characteristics $\bar{t}_2 > 0$ and $\bar{f}_2 < \infty$, and the long one is characterized by $\bar{t}_3 > \bar{t}_2$ and $\bar{f}_3 = \infty$. Let us consider the day-period model described in Section 3.1. In view of complete symmetry in our network, the equilibrium solutions for both directions of each street will be the same. So, we will have only three travel-time variables $t = (t_1, t_2, t_3)$ and three equilibrium flow variables $f = (f_1, f_2, f_3)$. Since there is no congestion possible on the streets 2 and 3, we are sure that at equilibrium $t_1 = \bar{t}_1$ and $t_3 = \bar{t}_3$. Thus, the cost function of OD-pair $(1,3)$ depends only on the single variable t_2:

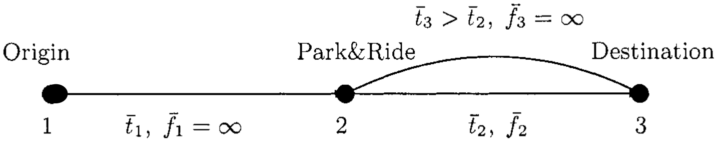

$\bar{t}_3 > \bar{t}_2,\ \bar{f}_3 = \infty$
Origin
Park&Ride
Destination
1
$\bar{t}_1,\ \bar{f}_1 = \infty$
2
$\bar{t}_2,\ \bar{f}_2$
3

Fig. 2. Simple Park-And-Ride model

$$T^D_{(1,3)}(t_2) = \min\{\, 2\alpha\bar{t}_1 + \pi_2 + c_{2,3},\ 2\alpha\,[\bar{t}_1 + \min\{t_2, \bar{t}_3\}] + \pi_3 \,\}.$$

Thus, given any demand flow $d_{(1,3)}$, the equilibrium Stable Dynamics solution can be found from the following problem:

$$\max_{t_2} \left[\, d_{(1,3)} \cdot T^D_{(1,3)}(t_2) \ - \ \alpha \cdot \bar{f}_2 t_2 \,:\, t_2 \geq \bar{t}_2 \,\right].$$

Since the cost function is bounded from above, the solution to this problem exists for any level of demand. Let us discuss its structure for different variants of parking prices π_2 and π_3, and public transportation price $c_{2,3}$. In our analysis we assume that

$$\pi_2 + c_{2,3} \geq \pi_3 + 2\alpha\bar{t}_2,$$

which means that on an empty network it is better to travel by car.

1. *Traffic ban*: $\pi_2 + c_{2,3} < 2\alpha\bar{t}_2 + \pi_3$. In this situation all drivers will park at node 2 independent of the level of demand. In this situation

$$t_2 = \bar{t}_2, \quad f_1 = d_{(1,3)}, \quad f_2 = f_3 = 0.$$

2. *Controllable congestion*: $2\alpha\bar{t}_2 + \pi_3 < \pi_2 + c_{2,3} < 2\alpha\bar{t}_3 + \pi_3$. In this situation no driver uses street 3. The usage of street 2 depends on the demand level. If $d_{(1,3)} < \bar{f}_2$, then the parking lot at node 2 is empty. Nevertheless, there is no congestion on street 2.

If $d_{(1,3)} > \bar{f}_2$, then the level of congestion on street 2 can be found from the condition:

$$2\alpha t_2 + \pi_3 = \pi_2 + c_{2,3}.$$

Again, there is no reason to use street 3. But the parking lot at node 2 is used. Note that the equilibrium travel time at street 2 is a function of the parking prices:

$$t_2 = \frac{1}{2\alpha}[\pi_2 - \pi_3 + c_{2,3}] \geq \bar{t}_2.$$

Since the number of cars on street 2 is equal to $\bar{f}_2 \cdot t_2$, it is possible to avoid an excessive usage of street 2 by regulating price differential of the parking lots.

3. *Over-pricing:* $2\alpha\bar{t}_3 + \pi_3 < \pi_2 + c_{2,3}$. In this situation the price of the parking lot at node 2 is too high. Thus, the drivers avoid it and we obtain the equilibrium solution from the model considered in Example 1.

Note that in our analysis the parking lot at node 2 can be seen as a substitute for street 3.

6 Discussion and concluding remarks

We have developed a mathematical formulation to describe Park and Ride for the morning and evening commutes. This tool is based on the Stable Dynamics theory. It allows addressing parking policies at both the local and the global or strategic levels. This approach relies on the idea that parking policies should be studied as one element of a comprehensive transportation system and cannot be studied in isolation. The proposed framework also takes into account parking management and congestion in a consistent manner and within the same standard convex formulation.

Some important steps are still missing at this stage. The proposed methodology is able to compare alternative parking policies, but is unable to select the best policies to be implemented. This would require the addition of a supplementary step involving optimization of the parking location considering a given pricing strategy.

In order to simplify the presentation, we have assumed that the morning and the evening O/D matrices were the same. This does need to be true, and the proposed approach can easily be extended, with additional notation, to the case where the morning and the evening O/D matrices differ. We have not considered more complex (and realistic) activity patterns, but it is clear that the proposed formulation can be extended to more complex set of activity patterns (including several stops).

Finally, we have assumed, for the sake of simplicity, that the O/D matrices were exogenous. This hypothesis is not restrictive since in our previous formulation (see [NP03]) we have proposed a method to describe trip generation and distribution as a part of the same mathematical formulation of Stable Dynamics. This extension adds one additional level in the optimization procedure, but does not modify the approach proposed in this paper. Recently, researchers have started to model parking within the context of the monocentric city [AP01]. In particular, they evaluate the impact of parking management on land use. These long-term issues (requiring endogenous O/D matrices) are important and should be considered for large-scale networks.

Acknowledgments. We would like to thank Richard Arnott, Moshe Ben-Akiva and Robin Lindsey for kind suggestions and fruitful comments. We would like also to thank PREDIT, French Ministry of Transportation (project URBANSIM) for partial financial support. Finally, we benefited from very helpful suggestions of Don Hearn, and an anonymous referee; their comments greatly improved the quality of this paper.

The research results presented in this paper have been supported by a grant "Action de recherche concertée ARC 04/09-315" from the "Direction de la recherche scientifique - Communauté française de Belgique". The scientific responsibility is assumed by the authors.

References

[AP01] Anderson, S., de Palma, A.: Parking in the Cities. Journal of Urban Economics, **55**, 1–20 (2001)

[APL91] Arnott, R., de Palma, A., Lindsey, R.: Temporal and Spatial Equilibrium Analysis of Commuter Parking. Journal of Publics Economics, **45**, 301–337 (1991)

[AR99] Arnott, R., Rowse, J.: Modeling Parking. Journal of Urban Economics, **45(1)**, 97–124 (1999)

[AG99] Association of Governments 1998 State of the Commute Report. Southern California of Governments Southern California Ridershare (1999)

[AH90] Axhausen, K., Hertz, R.: Simulating Activity Chains: A German Approach to Modeling Urban Travel. Journal of Transportation Engineering, American Society of Civil Engineers (1990)

[AP91] Axhausen, K., Polak, J.: Choice of Parking: stated preference approach. Transportation, **18**, 59–81 (1991)

[BMS81] Ben-Akiva, M., Manski, C.F., Sherman, C/F.: A Behavioral Approach to Modeling Household Vehicle Ownership and Applications to Aggregate Policy Analysis. Environment and Planning A, **13**, 339–441 (1981)

[Knu79] Knuth, D.E.: The Art of Computer Programming. Addsion-Wesley (1979)

[Nes00] Nesterov, Yu.: Stable traffic equilibria: properties and applications. Optimization and Engineering, **1(1)**, 29–50 (2000)

[NP03] Nesterov, Yu., de Palma, A.: Stationary dynamic solutions in congested transportation networks: summary and perspectives. Networks and Spatial Economics, **3**, 371–395 (2003)

[Rot65] Roth, G.: Paying for Parking. The Institute of Economic Affairs, London (1965)

[TL00] Tam, M., Lam, W.: Maximum Car Ownership under Constraints of Road Capacity and Parking Space. Transportation Research A, **34**, 145–170 (2000)

[TR98] Thomson, R., Richardson, A.: A Parking Search Model. Transportation Research A, **32(3)**, 159–170 (1998)

[VO95] Van der Waerden, P., Oppwal, H.: Modeling the Combined Choice of Parking lot and Shopping Destination. Paper presented at the 7th World Conference on Transport Research, Sydney, Australia July 16-21 (1995)

[Vic59] Vickrey, W.: Statement on the Pricing of Urban Street Use. In Hearing, US Congress Joint Committee on Metropolitan Washington Problem. November 11, 1959, reprinted in Journal of Urban Economics, **36**, 42–65 (1994)

[YTT91] Young, W., Thomson, R., Taylor, M.: A review of Urban Parking Models. Transport Review, **11**, 63–84 (1991)

Bilevel Optimisation of Prices and Signals in Transportation Models

Michael J Smith[1]

Department of Mathematics, University of York, Heslington, York, YO10 5DD, United Kingdom, mjs7@york.ac.uk

Summary. We suppose given a variable demand model with some control parameters to represent prices, a smooth function V which measures departure from equilibrium and a smooth function Z which measures overall disbenefit. We suppose that we wish to minimise Z subject to the constraint that the disequilibrium function V is no more than ε, where we think of ε as a small positive number. The paper suggests a *simultaneous descent direction* to solve this bilevel optimisation problem; such a direction reduces Z and V simultaneously and may often be computed by simply bisecting the angle between $-\nabla Z$ and $-\nabla V$. The paper shows that following a direction Δ which employs the simultaneous descent direction as its central element leads, under natural conditions which preclude edge effects (where a flow may be zero or a price may be maximum), to the set of those approximate equilibria (where $V \leq \varepsilon$) at which Z is stationary.

Then the method is extended on the one hand to deal with edge effects (allowing a route flow to be zero or a price to be the maximum permitted), by ensuring that the direction Δ followed anticipates nearby edges of the feasible region, using reduced gradients instead of gradients, and on the other hand to deal with signal controls.

Within the optimisation procedure proposed here, optimisation and equilibration move in parallel and the need to compute a sequence of approximate equilibria is avoided.

Key words: Bilevel Optimisation, Transportation Networks, Pricing, Control, Equilibrium

1 Introduction

1.1 Purpose of the paper

It is important to model transportation networks so that various alternative strategies designed by planners may be tested in a model prior to implementation. But it also important to devise methods of optimising network models

Mathematical and Computational Models for Congestion Charging, pp. 159-200

subject to natural assumptions concerning travellers choices, so that the computer modelling actually makes suggestions to the planners. Modelling may then further assist decision-makers with their future designs, by taking a more pro-active role in strategy design. In this paper we give and justify a possible method of optimising prices within an equilibrium transportation model. The method is based on *simultaneous descent directions*; these reduce two objective functions simultaneously.

1.2 Practical and general equilibration modelling background

The need to consider the several parts of a transportation equilibrium model as a single unity has been emphasised in [COM96], for example; this study was part of the US Travel Model Improvement Program. In the UK, this same need has also been described by the Department of the Environment, Transport and the Regions (see, [DofE98] and [SAC99]). This need leads directly to fairly general variable demand equilibrium models.

Many transport equilibrium models, together with a number of solution methods, have been proposed. See, e.g., [BMW56], [Eva76], [Gar80], [Bar02] and [BB03].

In this paper we adopt a specific variable demand equilibrium model within which some interactions, including that between the flows along certain arcs and costs felt on others and between the costs of travel between certain OD pairs and the flows generated between other OD pairs, may readily be represented; this model is similar to that suggested in [CC61] and [AM83]. The model combines the standard user equilibrium route-choice principle stated in [War52] and a demand for travel between each OD pair which may vary with the costs of travel between the various OD pairs.

1.3 Optimising prices and signals within equilibrium models

The need to optimise signal controls subject to equilibrium within an equilibrium transportation model was pointed out in [All74].

Many others have also considered this type of bilevel optimisation problem. See, e.g., [AL79], [TGA79], [Mar83],[Mar86], [Fis84], [TF88], [Dav94], [YY94], [Yan96a], [Yan96b], [Yan96c], [Chi97], [CSX99], [PR02] and [CW02]. All these essentially seek to solve the same problem as we do here. Migdalas [Mig95] provides an interesting survey of some solution techniques which have been proposed. The problem, shortly stated, is as follows.

First problem statement: *Given a smooth objective function, find optimal prices and signal green-times allowing for travellers' decisions.*

Here we weaken this problem as follows.

Second problem statement: *Given a smooth objective function, find flows, prices and signal green-times which comprise an approximately stationary point, allowing approximately for travellers' choices.*

This is a bilevel problem because we choose to allow for travellers' choices by constraining transport flows to be an equilibrium or an approximate equilibrium; already for fixed prices this is itself an optimisation problem.

Bilevel problems in a wider context have also been studied by very many people including, e.g., [GS94], [LPR96] and [OZ95].

The basic equilibration and optimisation ideas in this paper are natural developments of the Lyapunov equilibration methods described in [Smi84a] and [Smi84b] and build directly upon the half-space and cone projection methods developed (and very slightly tested) in a series of papers: [SXY97], [SXY98], [SXYG98], [CSXY01], [CS98], and [CS01]. This paper also builds directly on [Smi05a]. However the basic idea of using a direction which is the average of the steepest descent direction of the constraint function and the steepest descent direction of the given objective function may be traced back to Zoutendijk.

The advances presented here are as follows:

1. The optimisation method here may be properly applied to a wider variety of real-life problems than that in [CSXY01] as the conditions on the cost and demand functions are more general here,
2. The optimisation method here also embraces a better dynamic, Armijo-like, step length rule; the aim of this is to guarantee that a stationary point is indeed approached fairly economically from a computational viewpoint, and
3. The optimisation method here extends that in [Smi05a], by changing the search direction so as to allow for hard constraints; this permits (for example) route-flows to be zero and prices to be at their maximum at a stationary point.

These advances allow more general and more simple proofs of convergence and might also be expected to be much more efficient computationally than the methods demonstrated in previous work. Some details not given here are given in [Smi05b].

Using the algorithm here the approach to a stationary point is typically via points which are not themselves approximate equilibria, so that the optimisation and the equilibration move in parallel and the need to compute a sequence of approximate equilibria is avoided.

Previously, [FL00] and [RM04] both consider descent methods which have elements in common with the simultaneous descent method outlined here.

In [CQW02], Cohen et al. are critical of [CS01]. However they do not consider the smoothed search direction in that paper; they consider only a more simple discontinuous search direction. This means that their comments are not relevant to this current paper.

2 The Model

2.1 The main variables

We suppose that in our model network there are K OD pairs, that OD pair ij is joined by N_{ij} routes and so the total number of routes is $N = \sum_{ij} N_{ij}$. The main variables are as follows:

$X_{ijr} =$ the flow (in, e.g., vehicles per minute) along the r^{th} route joining OD pair ij,
$X =$ the route flow vector comprising all the X_{ijr},
$Y_{ij} =$ the cost of travel between OD pair ij (*in minutes per vehicle, say*), and
$Y =$ the cost vector comprising all the Y_{ij}.

Costs are in *minutes per vehicle*; so that a flow times a cost is dimensionless.

Let 0^N denote the zero N-vector, $+\infty^N$ denote the N-vector with infinite co-ordinates. Then we define $[0^N, +\infty^N)$ to be $[0, +\infty)^N$ and $[0^K, +\infty^K)$ to be $[0, +\infty)^K$.

2.2 Cost functions and demand functions

The central supposition throughout is that we are given two functions: the cost function $C(\cdot)$ and the demand function $D(\cdot)$.

Here $C_{ijr}(X)$ is the cost of traversing the r^{th} route joining OD pair ij when the flow vector is $X \geq 0$ and $D_{ij}(Y)$ is the total flow between OD pair ij when the OD cost vector is Y where $Y \geq 0$.

We suppose that the cost function $C(\cdot)$ is defined throughout $[0^N, +\infty^N)$ and that the demand function $D(\cdot)$ is defined throughout $[0^K, +\infty^K)$. Thus, including domains and co-domains, our two given functions are:

$$C : [0^N, +\infty^N) \to [0^N, +\infty^N) \text{ and } D : [0^K, +\infty^K) \to [0^K, +\infty^K).$$

2.3 The functions T and S

We define the two functions

$$T : [0^N, +\infty^N) \to [0^K, +\infty^K) \text{ and } S : [0^K, +\infty^K) \to [0^N, +\infty^N)$$

as follows. For each ij and each i, j and r:

$$T_{ij}(X) = \sum_r X_{ijr} \text{ for all } X \in [0^N, +\infty^N)$$

and

$$S_{ijr}(Y) = Y_{ij} \text{ for all } Y \in [0^K, +\infty^K).$$

$T_{ij}(X)$ gives the total flow from node i to node j and $S_{ijr}(Y)$ spreads each cost Y_{ij} over all routes joining node i and j.

2.4 Assumptions

Positivity of C and non-negativity of D

We suppose that, always, $C(X) > 0$ for all $X \in [0^N, +\infty^N)$ and $D(Y) \geq 0$ for all $Y \in [0^N, +\infty^N)$.

Boundedness

We suppose always (i) that $D(\cdot)$ is bounded and (ii) that $C(\cdot)$ is bounded on bounded sets.

Monotonicity

We also suppose that $C : [0^N, +\infty^N) \to [0^N, +\infty^N)$ is monotone. That is we suppose:

$$[C(X^1) - C(X^2)]^T(X^1 - X^2) \geq 0$$

for all $X^1 \in [0^N, +\infty^N)$ and $X^2 \in [0^N, +\infty^N)$. We also suppose that $-D : [0^K, +\infty^K) \to -[0^K, +\infty^K)$ is monotone.

It is easy to show that if C and $-D$ are both monotone then

$$\begin{pmatrix} C(X) - S(Y) \\ T(X) - D(Y) \end{pmatrix} \tag{1}$$

is also a monotone function of $\begin{pmatrix} X \\ Y \end{pmatrix} \in [0^N, +\infty^N) \times [0^K, +\infty^K)$.

Continuous differentiability

We also usually suppose that C and D are differentiable throughout their domains and that their derivatives or Jacobians C' and D' are continuous. (We assume that Jacobians or derivatives also exist at boundary points with suitable restricted definitions.)

3 Variable Demand Equilibrium

3.1 Definition of equilibrium

Suppose given a (flow-vector, cost-vector) pair $\begin{pmatrix} X \\ Y \end{pmatrix}$ in which all $X_{ijr} > 0$ and all $Y_{ij} > 0$. This vector will sometimes be written (X, Y). A vector (X, Y) will be called an equilibrium if the following hold:

1. $(X, Y) \in [0^N, +\infty^N) \times [0^K, +\infty^K)$.

2. For each ijr, the cost $C_{ijr}(X)$ of traversing the r^{th} route joining OD pair ij equals Y_{ij}.
3. For each ij the demand $D_{ij}(Y)$ generated by the cost vector Y equals the total flow $T_{ij}(X) = \sum_r X_{ijr}$ actually occurring from node i to node j.

So in this case (X, Y) is a variable demand equilibrium pair if and only if $(X, Y) \in [0^N, +\infty^N) \times [0^K, +\infty^K)$;

$$\begin{aligned} Y_{ij} - C_{ijr}(X) &= 0 \quad \text{for all } i, j, r; \text{ and} \\ D_{ij}(Y) - T_{ij}(X) &= 0 \quad \text{for all } i, j. \end{aligned} \tag{2}$$

The equilibrium equations (2) may be written, using (1), in vector form:

$$\begin{pmatrix} S(Y) - C(X) \\ D(Y) - T(X) \end{pmatrix} = \begin{pmatrix} 0^N \\ 0^K \end{pmatrix}.$$

Here we have assumed for simplicity that at equilibrium all $X_{ijr} > 0$ and all $Y_{ij} > 0$. To relax this initial assumption and so to allow the possibility that a listed route may have a high cost and zero flow at equilibrium we revise (2) to introduce a partial complementarity condition. The revised equilibrium condition is as follows. For each i, j and r:

$$\begin{aligned} &Y_{ij} - C_{ijr}(X) \leq 0 \text{ and } Y_{ij} - C_{ijr}(X) < 0 \quad \text{implies} \\ &X_{ijr} = 0; \quad \text{and} \quad D_{ij}(Y) - T_{ij}(X) = 0. \end{aligned} \tag{3}$$

[The expression $D_{ij}(Y) - T_{ij}(X) = 0$ may be similarly replaced by $D_{ij}(Y) - T_{ij}(X) \leq 0$ and $D_{ij}(Y) - T_{ij}(X) < 0$ implies $Y_{ij} = 0$. This change would make (2) exactly into a complementarity problem instead. We choose not to make this change here as it proves to be unnecessary under our positivity assumption in Section 2.4 above.]

The set of equilibria, or solutions to (3), will be denoted by E.

3.2 Proofs of existence of equilibria

We give here two proofs of existence of equilibria. The first proof is a standard proof using continuity and has some similarity to that given in [AM83]. The second proof utilises monotonicity (as well as continuity) and serves to introduce the algorithms proposed in the paper. The set F defined below is common to both proofs.

The closed bounded feasible set F and the objective function V

The natural feasible set is $[0^N, +\infty^N) \times [0^K, +\infty^K)$, but this is *unbounded.* Suppose now that our positivity and boundedness assumptions, in 2.4 above, hold. Under these conditions we are able to specify a *bounded* feasible set F.

First we define upper bounds UX_{ijr} and UY_{ij} for X_{ijr} and Y_{ij} as follows:

$$UX_{ijr} = 2sup\{D_{ij}(Y); Y \in [0^K, +\infty^K)\}$$

and

$$UY_{ij} = 2sup\{C_{ijr}(X); X \in [0^N, UX]\}$$

where UX is the vector of all the UX_{ijr}.

Now define the *bounded* feasible set F of (X, Y) pairs by putting:

$$\begin{aligned} F &= [0^N, UX] \times [0^K, UY] \\ &= \prod_{ijr}[0, UX_{ijr}] \times \prod_{ij}[0, UY_{ij}] \subset [0^N, +\infty^N) \times [0^K, +\infty^K) \end{aligned}$$

where UY is the vector of all the UY_{ij}.

Given the set F, consider the objective function $V : F \to R_+$ where, for all $(X, Y) \in F$:

$$\begin{aligned} V(X,Y) = &\sum_{ijr}\{X_{ijr}^2[C_{ijr}(X) - Y_{ij}]_+^2 + (UX_{ijr} - X_{ijr})^2[Y_{ij} - C_{ijr}(X)]_+^2\} \\ &+ \sum_{ij}\{Y_{ij}^2[T_{ij}(X) - D_{ij}(Y)]_+^2 + (UY_{ij} - Y_{ij})^2[D_{ij}(Y) - T_{ij}(X)]_+^2\}. \end{aligned}$$

Here $y_+ = \max\{y, 0\}$ and $y_+^2 = (y_+)^2$ for all real numbers y.

Proof of existence using continuity

Let $(X, Y) \in F$. It is clear that if (X, Y) is an equilibrium then $V(X, Y) = 0$. Our first task is to show the converse; that, under natural conditions, if $(X, Y) \in F$ and $V(X, Y) = 0$ then (X, Y) is an equilibrium. This result follows from Lemma 1 below.

Lemma 1. *Suppose that the functions C and D satisfy our positivity and boundedness assumptions in Section* 2.4 *above. Let $(X, Y) \in F$. Then $V(X, Y) > 0$ if (X, Y) satisfies any one of the following conditions:*

1. *For some $ijr, 0 \leq X_{ijr} \leq \frac{1}{2}UX_{ijr}$ and $Y_{ij} > \frac{1}{2}UY_{ij}$.*
2. *For some $ijr, \frac{1}{2}UX_{ijr} < X_{ijr} \leq UX_{ijr}$ and $0 < Y_{ij} \leq UY_{ij}$.*
3. *For some $ijr, 0 < X_{ijr} \leq UX_{ijr}$ and $Y_{ij} = 0$.*
4. *For some $ij, D_{ij}(Y) - T_{ij}(X) > 0$.*
5. *For some $ij, D_{ij}(Y) - T_{ij}(X) < 0$.*
6. *For some $ijr, Y_{ij} < C_{ijr}(X)$ and $X_{ijr} > 0$.*

This lemma shows that if $(X, Y) \in F$ and $V(X, Y) = 0$ then

- None of the X_{ijr} exceed $\frac{1}{2}UX_{ijr}$ (as neither 2 nor 3 can hold).
- None of the Y_{ij} exceed $\frac{1}{2}UY_{ij}$ (as neither 1 nor 2 can hold.
- The vector (X, Y) is an equilibrium (as none of 1, 2, 4, 5, nor 6 can hold).

Hence, (X, Y) satisfies the equilibrium condition (3) in Section 3.1 and also belongs to $[0^N, \frac{1}{2}UX] \times [0^K, \frac{1}{2}UY]$.

Proof. Suppose that Condition 1 holds. Then

$$\begin{aligned} V(X,Y) &\geq (UX_{ijr} - X_{ijr})^2[Y_{ij} - C_{ijr}(X)]_+^2 \\ &> (\frac{1}{2}UX_{ijr})^2[\frac{1}{2}UY_{ij} - C_{ijr}(X)]_+^2 \geq 0 \end{aligned}$$

by definition of UY_{ij}.

Suppose now that Condition 2 holds. Then

$$\begin{aligned} V(X,Y) &\geq Y_{ij}^2[T_{ij}(X) - D_{ij}(Y)]_+^2 \\ &\geq Y_{ij}^2[X_{ijr} - D_{ij}(Y)]_+^2 \\ &> Y_{ij}^2[\frac{1}{2}UX_{ijr} - D_{ij}(Y)]_+^2 \geq 0 \end{aligned}$$

by definition of UX_{ijr}.

Suppose now that Condition 3 holds. Then

$$V(X,Y) \geq X_{ijr}^2[C_{ijr}(X) - Y_{ij}]_+^2 = X_{ijr}^2[C_{ijr}(X)]_+^2 > 0$$

by positivity of C and X_{ijr}.

Suppose now that Condition 4 holds. Then there are two cases: $Y_{ij} > \frac{1}{2}UY_{ij}$ and $Y_{ij} \leq \frac{1}{2}UY_{ij}$.

$$Y_{ij} > \tfrac{1}{2}UY_{ij} \Rightarrow V(X,Y) > 0$$

since (1) or (2) must hold. Also

$$Y_{ij} \leq \tfrac{1}{2}UY_{ij} \Rightarrow V(X,Y) \geq (\tfrac{1}{2}UY_{ij})^2(D_{ij}(Y) - T_{ij}(X))^2 > 0.$$

Suppose now that Condition 5 holds. Then again there are two cases: $Y_{ij} > 0$ and $Y_{ij} = 0$. Firstly, $Y_{ij} > 0 \Rightarrow V(X,Y) \geq Y_{ij}^2[T_{ij}(X) - D_{ij}(Y)]^2 > 0$. On the other hand, condition (5) here ensures that $T_{ij}(X) > D_{ij}(Y) \geq 0$ in any case and hence there is an $X_{ijr} > 0$. But we also know that $C_{ijr}(X) > 0$ always and so

$$Y_{ij} = 0 \Rightarrow Y_{ij} - C_{ijr}(X) = -C_{ijr}(X) < 0$$

and therefore $V(X,Y) \geq X_{ijr}^2[C_{ijr}(X) - Y_{ij}]_+^2 = X_{ijr}^2 C_{ijr}(X) > 0$.

Suppose finally that Condition 6 holds. Then again

$$V(X,Y) \geq X_{ijr}^2[C_{ijr}(X) - Y_{ij}]_+^2 > 0.$$

Thus the lemma is proved. ■

A related results is as follows.

Lemma 2. *Suppose* C *and* D *satisfy the positivity, non-negativity and boundedness assumptions in Section* 2.4 *above. Suppose also that* (X,Y) *is an equilibrium in* $[0^N,+\infty^N)\times[0^K,+\infty^K)$. *Then* (X,Y) *is an equilibrium in* $\frac{1}{2}F=[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$.

Proof. We only have to rule out the possibility of an equilibrium outside $\frac{1}{2}F$. Suppose then that (X,Y) is an equilibrium, possibly outside $\frac{1}{2}F$. From the equilibrium conditions: $Y_{ij}-C_{ijr}(X)\le 0$ and $X_{ijr}\le T_{ij}(X)=D_{ij}(Y)$. Hence for all ijr:

$$Y_{ij}\le C_{ijr}(X)\le \tfrac{1}{2}UY_{ij}$$

by definition of UY_{ij} and also

$$X_{ijr}\le T_{ij}(X)=D_{ij}(Y)\le \tfrac{1}{2}UX_{ijr}$$

by definition of UX_{ijr}.

Thus any equilibrium $(X,Y)\in[0^N,+\infty^N)\times[0^K,+\infty^K)$ also belongs to the smaller set $\frac{1}{2}F=[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]\subset F$. ■

Theorem 1 (A standard existence theorem using continuity). *Suppose that* C *and* D *satisfy the positivity, non-negativity and boundedness conditions in Section* 2.4, *and that* F *is as constructed above. Suppose also that* C *and* D *are both continuous. Then an equilibrium exists, and the set of equilibria is a non-empty subset of* $[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$.

Proof. For each (X,Y) in F, put $\Phi((X,Y)) = Proj_F((X,Y)+(S(Y)-C(X),D(Y)-T(X))$. Then, $\Phi: F\to F$ is continuous and F is closed, bounded and convex; so there is a fixed point of Φ by Brouwer's fixed point theorem. This point must be a zero of V. By Lemma 1, this point must belong to $[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$ and also satisfies the equilibrium conditions and so is an equilibrium in $\frac{1}{2}F\subset F$. By Lemma 2 there are no equilibria outside $[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$, and therefore the set of equilibria is a non-empty subset of $[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$. ■

Proof of existence using monotonicity

Theorem 2. *Suppose that* C *and* D *satisfy all the conditions listed in Section* 2.4, *and that* F *is as constructed above. Then an equilibrium exists, and the set of equilibria is a non-empty closed bounded convex subset of* $[0^N,\frac{1}{2}UX]\times[0^K,\frac{1}{2}UY]$.

Proof. We are given that C and $-D$ are both monotone and it follows that $(C(X)-S(Y),T(X)-D(Y))$ is a monotone function of (X,Y) on F.

Consider the search direction $\Delta(X,Y)=(\Delta^1(X,Y),\Delta^2(X,Y))$ where the ijr^{th} component of the flow direction $\Delta^1(X,Y)$ is for each ijr given by:

$$\Delta^1_{ijr}(X,Y)=(UX_{ijr}-X_{ijr})^2[Y_{ij}-C_{ijr}(X)]_+-X^2_{ijr}[C_{ijr}(X)-Y_{ij}]_+;$$

and the ij^{th} component of the cost direction $\Delta^2(X,Y)$ is for each ij given by:

$$\Delta^2_{ij}(X,Y) = (UY_{ij} - Y_{ij})^2[D_{ij}(Y) - T_{ij}(X)]_+ - Y^2_{ij}[T_{ij}(X) - D_{ij}(Y)]_+.$$

The whole search direction

$$\Delta(X,Y) = (\Delta^1(X,Y), \Delta^2(X,Y)) \tag{4}$$

for all $(X,Y) \in F$. (The direction (4) is precisely the algorithm (D) direction, introduced in [Smi84a] and [Smi84b], in this setting, utilising the function $(C(X) - S(Y), T(X) - D(Y))$ on the set $F = [0^N, UX] \times [0^K, UY]$.)

Since $(C(X) - S(Y), T(X) - D(Y))$ is a monotone continuously differentiable function of (X,Y) on F, it follows from [Smi84a] and [Smi84b] that $\Delta(X,Y)$ is a descent direction for objective V at (X,Y) (unless $V(X,Y) = 0$ in which case (X,Y) is an equilibrium) and also that there is a positive real number h such that for each $(X,Y) \in F$, $(X,Y) + t\Delta(X,Y) \in F$ for each t such that $0 \leq t \leq h$.

Thus, for any (X,Y) in F, $\Delta(X,Y)$ is a feasible descent direction for V at (X,Y); unless $V(X,Y) = 0$.

Now $V(X,Y)$ is a continuous function of (X,Y) on the compact set $F = [0^N, UX] \times [0^K, UY]$. Therefore V attains its greatest lower bound on this set at say (X^*, Y^*). If $V(X^*, Y^*) > 0, \Delta(X^*, Y^*)$ is a feasible descent direction for V at (X^*, Y^*) contradicting the definition of (X^*, Y^*).

This contradiction arises from the assumption that $V(X^*, Y^*) > 0$. It follows that $V(X^*, Y^*) = 0$ and an equilibrium does exist in F.

Moreover by a standard lemma (Minty's lemma) the set of equilibria is under the present conditions also convex, since $(C - S, T - D)$ is monotone. Since V is continuous the set $E = \{(X,Y) \in F; V(X,Y) = 0\}$ is also closed.

Thus there is a non-empty closed convex set E of equilibria such that $E \subset [0^N, \frac{1}{2}UX] \times [0^K, \frac{1}{2}UY]$. ■

4 An Equilibration Method

4.1 The basic condition

We now impose the following basic condition on all following work.

- *There is a fixed set of N routes joining K OD pairs.*
- *C and $-D$ are monotone and continuously differentiable.*
- *$F = [0^N, UX] \times [0^K, UY]$.*
- *The set of equilibria is a non-empty subset of $[0^N, \frac{1}{2}UX] \times [0^K, \frac{1}{2}UY]$.*

Any pair (X,Y) in F will be called feasible. A natural question is now: Given the two functions C and D satisfying the above conditions, how do we approximate or estimate a variable demand equilibrium (X,Y)?

In order to estimate an equilibrium (X,Y), we use an iterative scheme which approximates the equilibrium conditions more and more closely. A very general algorithm which includes almost all algorithms for solving this variable demand problem is to start anywhere in F and update (X,Y) as follows: (X^1,Y^1) is any feasible starting value for (X,Y) and following some rule or algorithm:

$$(X^1,Y^1) \to (X^2,Y^2) \to (X^3,Y^3) \to \cdots .$$

We shall now suppose that we have an algorithm which generates an infinite sequence such as that above, and that for any feasible start point all the succeeding pairs are also feasible in that they all belong to F.

4.2 Definition of convergence and a Lyapunov function

The sequence $(X^1,Y^1) \to (X^2,Y^2) \to (X^3,Y^3) \to \cdots$ is said to *converge* to the equilibrium set E if (i) the equilibrium set E is non-empty and (ii) the Euclidean distance $dist((X^n,Y^n),E)$ between (X^n,Y^n) and the equilibrium set E tends to zero as $n \to \infty$.

In [Smi84a] and [Smi84b], Smith shows that (given the basic condition above) there is a continuous real-valued function G defined for all $(X,Y) \in F$ such that

- $G(X,Y) > 0$ if $(X,Y) \in F \setminus E$;
- $G(X,Y) = 0$ if $(X,Y) \in E$; and
- $\Delta(X,Y) \cdot \nabla V(X,Y) \le -G(X,Y)$ for all $(X,Y) \in F$.

Thus V is a Lyapunov function for the dynamical system:

$$\begin{aligned} d((X(t),Y(t))/dt &= \Delta((X(t),Y(t)) \text{ for all } t \ge 0, \\ (X(0),Y(0)) &= (X^0,Y^0) \in F. \end{aligned}$$

Here (X^0,Y^0) is an arbitrary start point $\in F$. The steepness of the descent of the Lyapunov function V at (X,Y), as (X,Y) follows Δ, is estimated by G. That is:

$$\begin{aligned} dV(X(t),Y(t))/dt &= \nabla V(X(t),Y(t)) \cdot \Delta(X(t),Y(t)) \\ &\le -G(X(t),Y(t)) \text{ for all } t \ge 0. \end{aligned}$$

We may take the function G to be as follows:

$$\begin{aligned} G(X,Y) &= \\ &\sum_{ijr}\{X_{ijr}^3[C_{ijr}(X) - Y_{ij}]_+^3 + (UX_{ijr} - X_{ijr})^3[Y_{ij} - C_{ijr}(X)]_+^3\} \\ &+ \sum_{ij}\{Y_{ij}^3[T_{ij}(X) - D_{ij}(Y)]_+^3 + (UY_{ij} - Y_{ij})^3[D_{ij}(Y) - T_{ij}(X)]_+^3\}. \quad (5) \end{aligned}$$

4.3 The projection $Proj_F$

Throughout all algorithms described here, all (X, Y) generated will be feasible or belong to F. Any "tentative values" of (X, Y) generated by an equilibrating (or, later, optimising) algorithm which are not feasible will always be projected back onto F.

For each $(X, Y) \in \mathbb{R}^{N+K}$ the projection $Proj_F(X, Y)$ of (X, Y) onto the relevant feasible set (F here) is defined as follows, by projecting each co-ordinate of X and each co-ordinate of Y independently. Thus:

$$Proj^1_{ijr}(X) = \begin{cases} 0, & \text{if } X_{ijr} < 0, \\ X_{ijr}, & \text{if } 0 \leq X_{ijr} \leq UX_{ijr}, \\ UX_{ijr}, & \text{if } X_{ijr} > UX_{ijr}. \end{cases}$$

$$Proj^2_{ijr}(Y) = \begin{cases} 0, & \text{if } Y_{ij} < 0, \\ Y_{ij}, & \text{if } 0 \leq Y_{ij} \leq UY_{ij}, \\ UY_{ij}, & \text{if } Y_{ij} > UY_{ij}\ . \end{cases}$$

Then, for each $(X, Y) \in \mathrm{R}^{N+K}$, we put $Proj_F(X, Y) = (Proj^1(X), Proj^2(Y)) \in F$, and $Proj_F(X, Y)$ is the point of F closest to (X, Y).

5 Dynamic Armijo-Like Step Lengths

Suppose that the basic condition in Section 4.1 holds. We reduce V to zero by moving (X, Y) continually in the direction $\Delta(X, Y)$ specified in (4) in Section 3.2 above.

Given V, Δ, and G, to show descent to equilibrium over the whole trajectory generated by an equilibrium-seeking algorithm which follows Δ we need to specify step length choices in some detail. We follow a dynamic Armijo-like scheme very close to that described in [Smi84b].

Henceforth we will let z stand for (X, Y) and z^n stand for (X^n, Y^n).

5.1 A dynamic Armijo-like algorithm

Here we suppose that if we are at iteration n, at a non-equilibrium $(X^n, Y^n) = z_n$ where the search direction is $\Delta(z_n)$ and the step length actually used at z_n is u_n then our next z will be

$$z_{n+1} = Proj_F(z_n + u_n\Delta(z_n))$$

where $Proj_F$ denotes projection onto F. The real number $u_n > 0$ will be called a used step length and $t_n > 0$ will be called a step length.

For the purposes of most of this section we will suppose that if u_n is a used step length then $Proj_F(z_n + u_n\Delta(z_n)) = z_n + u_n\Delta(z_n)$. That is, we suppose that the projection operator does not actually do anything; so the boundary of the feasible set F is here having no effect. To determine u_n we specify a dynamic Armijo-like scheme based on a continuous function G such as that given above.

To motivate the scheme, it is clear that if the step length t at z were very small then the change in $V = V(z + t\Delta(z)) - V(z)$ would be more negative than $-\frac{3}{4}tG(z)$. So the slope of $V(z + t\Delta(z))$ against t would be at least as steep as $-\frac{3}{4}G(z)$ for small t. We do not wish to have such small steps t as the reduction in V might be very small, by virtue of the small step size t.

On the other hand if t is large the slope of $V(z+t\Delta(z))$ against t can be no steeper than $-\frac{1}{4}G(z)$ on average; for if the average gradient of $V(z + t\Delta(z))$ against t were $< -\frac{1}{4}G(z)$ for large t then $V(z + t\Delta(z))$ would sometimes be $< V(z) - \frac{1}{4}tG(z)$ and this would be negative for large t, which is impossible. (Of course, by its definition, $V \geq 0$ always.) We do not wish to have steps this large as the reduction in V might be very small, due to a shallow negative slope (or perhaps V may even increase, due to a positive slope).

So we seek step lengths which give rise to slopes between $-\frac{3}{4}G(z)$ and $-\frac{1}{4}G(z)$. Such step lengths have an Armijo property and allow convergence to equilibrium to be shown.

The dynamic Armijo (z_n, t_{n-1})-updating equilibration algorithm in detail

To be specific, we start at an arbitrary $z_1 \in F$ and $t_0 = 1$. This 0^{th} or initial step length t_0 is fairly arbitrary, but it must be positive.

If we are at a current non-equilibrium point $z_n \in F$, and the previous possible step length was t_{n-1}, then we are to update z_n and t_{n-1} (for $n \geq 1$) according to some fairly simple rules as follows.

Firstly, z_n is kept fixed and t_{n-1} is halved to obtain:

$$t_{n-1}, \tfrac{1}{2}t_{n-1}, (\tfrac{1}{2})^2 t_{n-1}, \cdots, (\tfrac{1}{2})^p t_{n-1},$$

where the halving ceases as soon as:

$$V[z_n + (\tfrac{1}{2})^p t_{n-1}\Delta(z_n)] - V(z_n) < -\tfrac{1}{8}(\tfrac{1}{2})^p t_{n-1}G(z_n)$$

for the first time. $p = 0$ is allowed here; it may be that

$$V[z_n + t_{n-1}\Delta(z_n)] - V(z_n) < -\tfrac{1}{8}t_{n-1}G(z_n)$$

already. The halving surely ceases by definition of G.

Secondly let $u_n = (\frac{1}{2})^p t_{n-1}$ for this p (this is to be the used step length at z_n) and let $z_{n+1} = z_n + u_n\Delta(z_n)$.
Finally update t_{n-1} as follows:

- if $V[z_n + u_n\Delta(z_n)] - V(z_n) < -\frac{3}{4}u_nG(z_n)$, put $t_n = 2u_n$;
- if $-\frac{3}{4}u_nG(z_n) \leq V[z_n + u_n\Delta(z_n)] - V(z_n) < -\frac{1}{4}u_nG(z_n)$, put $t_n = u_n$; and

- if $-\frac{1}{4}u_nG(z_n) \leq V[z_n + u_n\Delta(z_n)] - V(z_n) < -\frac{1}{8}u_nG(z_n)$, put $t_n = \frac{1}{2}u_n$.

These are three mutually exclusive possibilities and they together exhaust all eventualities since, by choice of $u_n = (\frac{1}{2})^p t_{n-1}$,

$$V[z_n + u_n\Delta(z_n)] - V(z_n) \geq -(\tfrac{1}{8})u_nG(z_n)$$

is not possible.

Clearly not all t_n are in fact used to move z. The used step lengths are the u_n. It is clear that

$$V[z_n + u_n\Delta(z_n)] - V(z_n) < -(\tfrac{1}{8})u_nG(z_n)$$

for all used step lengths u_n. (It would be natural to stop when $V(z_n)$ is less than some preassigned positive number.)

6 Convergence to Equilibrium

Suppose that the basic condition in Section 4.1 holds. For any starting point $z_1 \in F$ and with an initial "previous" step length $t_0 = 1$ we suppose that the dynamic Armijo-like algorithm specified above in Section 5.1 generates an infinite trajectory:

$$z_1, z_2, z_3, \cdots, z_n, \cdots$$

where each $z_n \in F$ and an infinite sequence

$$t_0, t_1, t_2, t_3, \cdots, t_n, \cdots$$

of possible non-negative step lengths. We assume that these sequences are infinite: so we never actually hit the equilibrium set E. (If the sequence hits an equilibrium we simply stop.)

To prove convergence, we use proof by contradiciton. Suppose that $\{z_n\}$ is an infinite sequence generated by our algorithm which does not converge to the equilibrium set $E = \{z \in F; V(z) = 0\}$. Then, since F is closed and bounded, $\{z_n\}$ must have a non-equilibrium limit point.

Let $w \in F$ be such a non-equilibrium limit point of the sequence $\{z_n\}$, so that $V(w) > 0$.

Now in the appendix we show that $V(z_n)$ is eventually less than $V(w)$ and therefore, since $\{V(z_n)\}$ is decreasing, $\{V(z_n)\}$ cannot have $V(w)$ as a limit point. Hence, as V is continuous, $\{z_n\}$ cannot in fact have w as a limit point. This provides the contradiction we seek.

It follows that the sequence $\{z_n\}$ has no non-equilibrium limit points. Hence all limit points of the sequence $\{z_n\}$ are equilibria and the sequence $\{z_n\}$ must converge to the set of equilibria or $dist(z_n, E) \to 0$.

7 Optimising Prices

Now we suppose that there is a specified smooth function Z which is regarded as a measure of total disbenefit. We will seek to minimise Z at an equilibrium by charging prices for traversing certain arcs or routes in the network, thus influencing the equilibrium traffic distribution. Of course if Z depends on the prices charged then charges will also influence Z directly.

7.1 Assumption concerning Z

We suppose throughout that Z is a continuously differentiable, non-constant function of the route-flow vector X, the OD cost vector Y and the new route-price vector P.

7.2 Adding an arc-price vector p or a route-price vector P

Suppose that for each arc a in the network a price p_a can be charged. Suppose also that the vector p of all the p_a is confined to some polyhedral closed bounded set of feasible arc-price vectors p. For many arcs the only feasible charge might be zero. For each feasible set of arc price vectors p route ijr (the r^{th} route joining node i to node j) will be subject to a corresponding charge P_{ijr} (the sum of the relevant p_a) and the vector P of all possible route prices P_{ijr} will be confined to some polyhedral closed bounded set F_{price} of feasible route-price vectors. Let $(X, Y) \in F$ and $P \in F_{price}$. Then the vector (X, Y, P) will be called a *user-equilibrium* if and only if for all i, j and r,

$$\begin{aligned}&Y_{ij} - C_{ijr}(X) - P_{ijr} \leq 0 \quad \text{and}\\ &Y_{ij} - C_{ijr}(X) - P_{ijr} < 0 \quad \text{implies} \quad X_{ijr} = 0, \quad \text{and} \quad D_{ij}(Y) - T_{ij}(X) = 0.\end{aligned}$$

Our basic condition given in Section 4.1 is now supposed to hold for each fixed $P \in F_{price}$. Further, we now suppose that the feasible set F is enlarged so that $\frac{1}{2}F$ contains *all* equilibria for *all* the given control vectors $P \in F_{price}$.

Given our smooth objective function $Z = Z(X, Y, P)$, we now wish to approximate an optimal (X, Y, P) or at least a stationary (X, Y, P) as follows: (X_1, Y_1, P_1) is any starting value for $(X, Y, P) \in F \times F_{price}$ and $(X_1, Y_1, P_1) \rightarrow (X_2, Y_2, P_2) \rightarrow (X_3, Y_3, P_3) \rightarrow \cdots$.

The previous V and Δ now involve P naturally. For all $(X, Y, P) \in F \times F_{price}$ we put:

$$\begin{aligned}V(X, Y, P) = \sum_{ijr} \{&X_{ijr}^2 [P_{ijr} + C_{ijr}(X) - Y_{ij}]_+^2 + (UX_{ijr} - X_{ijr})^2 \\ &\cdot [Y_{ij} - (P_{ijr} + C_{ijr}(X))]_+^2\} + \sum_{ij} Y_{ij}^2 [T_{ij}(X) - D_{ij}(Y)]_+^2 \\ &+ (UY_{ij} - Y_{ij})^2 [D_{ij}(Y) - T_{ij}(X)]_+^2;\end{aligned}$$

$$\begin{aligned}\Delta^1_{ijr}(X,Y,P) = {} & (UX_{ijr} - X_{ijr})^2[Y_{ij} - (P_{ijr} + C_{ijr}(X))]_+ \\ & -X^2_{ijr}[(P_{ijr} + C_{ijr}(X)) - Y_{ij}]_+,\end{aligned}$$

$$\Delta^2_{ij}(X,Y,P) = (UY_{ij} - Y_{ij})^2[D_{ij}(Y) - T_{ij}(X)]_+ - Y^2_{ij}[T_{ij}(X) - D_{ij}(Y)]_+,$$

and

$$\Delta(X,Y,P) = (\Delta^1(X,Y,P), \Delta^2(X,Y,P)).$$

This is the direction (4) suitably changed. Allowing for P, the previous equilibrium constraint

$$(X,Y) \in F \text{ and } V(X,Y) = 0$$

now becomes:

$$(X,Y,P) \in F \times F_{price} \text{ and } V(X,Y,P) = 0.$$

We will also write: $H = F \times F_{price}$ and suppose that

$$H = \{(X,Y,P); h_i(X,Y,P) \leq 0 \text{ for } i = 1,2,3,...,N_H\},$$

where we here also suppose that all the N_H functions h_i are linear. Nonegativity constraints are to be part of this set of linear constraints defining the set H of feasible (X,Y,P).

7.3 The price-enhanced basic condition and a constraint qualification

For the rest of the paper, we now assume that the following control-enhanced basic conditions hold.

- *There is a fixed set of N routes joining K OD pairs.*
- $F = [0^N, UX] \times [0^K, UY]$.
- *C and $-D$ are continuously differentiable monotone functions of X and Y, respectively.*
- *There is a closed bounded polyhedral set F_{price} of feasible route-price vectors.*
- *For each $P \in F_{price}$, the set E_P of equilibria is a non-empty subset of $\frac{1}{2}F = [0^N, \frac{1}{2}UX] \times [0^K, \frac{1}{2}UY]$.*

We have already defined the projection of a point onto a feasible set. Now we define the projection $Proj_S[v](x)$ of a vector v based at $x \in S$ onto the set S as follows.

$$Proj_S[v](x) = lim_{t \to 0+}[\tfrac{1}{t}(Proj_S(x + tv) - x)]$$

Suppose that $(X, Y, P) \in H$ is not an equilibrium so that $V(X, Y, P) > 0$. It follows that for this (X, Y, P), $\Delta(X, Y, P)$ is a feasible descent direction for V with P fixed and hence that:

$$Proj_F[-\nabla_{(X,Y)}V](X, Y, P) \neq 0 \text{ if } V(X, Y, P) > 0.$$

Here only (X, Y) varies in F and P is fixed in F_{price}.

It follows at once that:

$$Proj_H[-\nabla_{(X,Y,P)}V](X, Y, P) \neq 0 \text{ if } V(X, Y, P) > 0. \tag{6}$$

Here (X, Y, P) varies in H.

Replacing (X, Y, P) by just x, the equilibrium condition

$$(X, Y, P) \in F \times F_{price} \text{ and } V(X, Y, P) = 0$$

becomes: $x \in H$ and $V(x) = 0$.

Also (6) above may now be written:

$$Proj_H[-\nabla V](x) \neq 0 \text{ if } V(x) > 0. \tag{7}$$

Condition (7) may naturally be thought of as a constraint qualification applying to the set $E_\varepsilon = \{x \in H; V(x) \leq \varepsilon\}$ of approximate equilibria, for any positive ε. Suppose that condition (7) holds and suppose that x is any point in the set $H \cap E_\epsilon$ such that $V(x) = \varepsilon$ and $h_i(x) = 0$ for $i = 1, 2, 3, \cdots, k$ and no others. Then there is a direction $\delta = Proj_H[-\nabla V](x)$ such that $\delta \cdot descV(x) > 0$ and $\delta \cdot desch_i(x) \geq 0$ for $i = 1, 2, 3, \cdots, k$. That is: there is a feasible descent direction for V on the edge of any E_ε provided $\varepsilon > 0$.

Now the proposed optimisation algorithm, where (X^1, Y^1, P^1) is any starting value for (X, Y, P) in H and

$$(X^1, Y^1, P^1) \to (X^2, Y^2, P^2) \to (X^3, Y^3, P^3) \to \cdots,$$

becomes: x^1 is any starting value for x in H and

$$x^1 \to x^2 \to x^3 \to \cdots.$$

7.4 The simplest simultaneous descent direction

We wish to minimise $Z(X, Y, P) = Z(x)$ subject to $V(X, Y, P) = V(x)$ being zero or small, so we need to optimise V and Z simultaneously in some way.

Initially we let x lie in the interior of H. We also need our control-enhanced basic condition above in Section 7.3. We make the further initial basic assumption that $\nabla Z(x) \neq 0$ for all $x \in H$.

Now, under our control-enhanced basic conditions in Section 7.3 above, $V(x) > 0$ implies $\nabla V(x) \neq 0$. So we may now let (for $x \in H$ and $V(x) > 0$):

$$\begin{aligned} descV(x) &= -\nabla V(x)/||\nabla V(x)||, \\ descZ(x) &= -\nabla Z(x)/||\nabla Z(x)|| \text{ and} \\ desc(Z,V)(x) &= \frac{1}{2}descV(x) + \frac{1}{2}descZ(x). \end{aligned} \tag{8}$$

This direction (8), if non-zero, reduces V and Z *simultaneously.* In any case this direction is never an ascent direction, for either V or Z.

However *descV* is not defined at any x for which $V(x) = 0$ and also changes sharply in the vicinity of any such point. So it is natural to change this direction (8) slightly so that it is defined everywhere and is smoothly varying. This may be done by enlarging the equilibrium set E. We are led to put (where $\varepsilon > 0$):

$$E_\varepsilon = \{x \in H : 0 \leq V(x) \leq \varepsilon\}$$

and for all $x \in H$:

$$\begin{aligned} \Delta_\varepsilon(x) =& [V(x)/\varepsilon]desc(Z,V)(x) + [1 - V(x)/\varepsilon]_+ descZ(x) \\ &+ [V(x)/\varepsilon - 1]_+ descV(x). \end{aligned} \tag{9}$$

This direction (9) is identical in form and similar in motivation to direction (2.3) in [CS01]; $desc(Z,V)(x)$ is there the projection of $descZ(x)$ onto the hyperplane of locally-constant V but here is the average of the two directions. The present direction (9), using the average, has been introduced so as to allow a more straightforward effective step length selection procedure to be designed. This is outlined below.

For any x in the interior of H the zeros of $\Delta(x)$ coincide exactly with points x in E_ε at which Z is stationary; or those points in E_ε for which there is no descent direction for Z which remains inside E_ε. We will now show this. We will also show later that an algorithm following direction (9) leads, under natural conditions, to the set of points x at which Z is stationary provided a dynamic Armijo-like step length rule is adopted.

In [CQW02], Cohen et al. are critical of [CS01]; however they only consider the discontinuous direction (2.1) in that paper and do not refer to the smoothed direction (2.3) which has motivated direction (9) above. Thus their comments do not impact the optimisation methods described in this current paper.

7.5 Optimality conditions in the interior of H

Definitions of ε-feasible descent and ε-linear optimality (for x in the interior of H)

Given Z and given $x \in intH$, the vector u will be called an *ε-feasible descent direction at x* (in the interior of H) if and only if $u{\cdot}[descZ(x)] > 0$ and either

$$\begin{aligned} &V(x) < \varepsilon \text{ or} \\ &V(x) = \varepsilon \text{ and } u{\cdot}[descV(x)] \geq 0. \end{aligned}$$

Then x^* (in the interior of H) is said to be *ε-linearly-optimal* if and only if $V(x^*) \leq \varepsilon$ and there is no ε-feasible Z-descent direction at x^*.

ε-linear-optimality conditions (in the interior of H)

In this section we derive optimality conditions for any x in the interior of H. These results are particularly useful if H happens to have been chosen correctly; so that the upper and lower bounds on $x = (X, Y, P)$ are not binding when this optimality condition holds.

For x in the interior of H, we let $\Delta_\varepsilon(x)$ be given by (9) above and show how this vector may be used to classify points x according to whether they are ε-linearly-optimal or not.

Observe that $desc(Z, V)(x) = \frac{1}{2}descV(x) + \frac{1}{2}descZ(x)$ (if non-zero) is a direction in which both V and Z decline; and is never a direction of increase for either V or Z. So if $x \in intH$ and $\frac{1}{2}descV(x) + \frac{1}{2}descZ(x)$ is non-zero then x is not ε-linearly-optimal and also $\Delta_\varepsilon(x)$ reduces both V and Z. This is the crux.

Here we show that at least for $x \in intH$, $\Delta_\varepsilon(x) = 0$ if and only if x is ε-linearly-optimal. To do this we consider the following four mutually exclusive cases (for x in the interior of H):

1. $V(x) > \varepsilon$,
2. $0 \leq V(x) < \varepsilon$,
3. $V(x) = \varepsilon$ and $desc(Z, V)(x) \neq 0$, and
4. $V(x) = \varepsilon$ and $desc(Z, V)(x) = 0$.

In each of the first three of these cases we show that x is not ε-linearly-optimal by showing that $\Delta_\varepsilon(x)$ is a non-zero direction in which either the degree of disequilibrium, V, declines (Case 1); or is a direction in which Z improves maintaining $V \leq \varepsilon$ (Cases 2 and 3). We also show that, in Case 4, x is ε-linearly-optimal and also that in this case $\Delta_\varepsilon(x) = 0$.

Case 1 ($V(x) > \varepsilon$): In this case $x \notin E_\varepsilon$ and also

$$\Delta_\varepsilon(x) = [V(x)/\varepsilon]desc(Z, V)(x) + [V(x)/\varepsilon - 1]descV(x)$$

is non-zero as $descV(x)$ is non-zero and $desc(Z, V)(x)$ is never a direction in which V ascends. Of course x is not ε-feasible and so is not ε-linearly-optimal. Here following Δ_ε(x) reduces V.

Case 2 ($0 \leq V(x) < \varepsilon$): In this case $x \in E_\varepsilon$ and

$$\Delta_\varepsilon(x) = [V(x)/\varepsilon]desc(Z, V)(x) + [1 - V(x)/\varepsilon]descZ(x)$$

is again non-zero as $descZ(x)$ is non-zero and $desc(Z, V)(x)$ is never a direction in which Z ascends. Here following $\Delta_\varepsilon(x)$ reduces Z while of course maintaining $V \leq \varepsilon$.

Case 3 ($V(x) = \varepsilon$ and $desc(Z,V)(x) \neq 0$): In this case $x \in E_\varepsilon$ and $\Delta(x) = desc(Z,V)(x)$ is non-zero and so is a simultaneous descent direction for both V and Z. Thus $\Delta_\varepsilon(x)$ is an ε-feasible Z-descent direction at x so x is not ε-linearly-optimal. Here again following $\Delta_\varepsilon(x)$ reduces Z maintaining $V \leq \varepsilon$.

Case 4 ($V(x) = \varepsilon$ and $desc(Z,V)(x) = 0$): In this case also $x \in E_\varepsilon$. Now $desc(Z,V)(x) = 0$ and so $descV(x) = -descZ(x)$. Consider any Z-descent direction u. Then $u\cdot[descZ(x)] > 0$ and so $u\cdot[descV(x)] = u\cdot[-descZ(x)] < 0$, and u is not an ε-feasible direction at x. Thus there is no ε-feasible Z-descent direction from x and also $V(x) \leq \varepsilon$, so x is ε-linearly-optimal. In this case $\Delta_\varepsilon(x) = 0$.

Conclusion: We have shown that (at least for $x \in intH$) zeros of $\Delta_\varepsilon(x)$ coincide with points $x \in E_\varepsilon$ at which there is no ε-feasible descent direction for Z at x. Such points are ε-linearly-optimal. We have also shown that $\Delta_\varepsilon(x)$ is an H-feasible descent direction for V if x is not in E_ε and that $\Delta_\varepsilon(x)$ is an ($H-$ and) ε-feasible descent direction for Z if x is in E_ε and so is not ε-linearly-optimal. Thus, at least for x in the interior of H, $\Delta_\varepsilon(x)$ is an excellent arbiter of ε-linear-optimality at x; and for those x which are not ε-linearly-optimal indicates a sensible non-zero direction for moving x.

7.6 A dynamic Armijo-like optimisation algorithm

Here we outline briefly the main changes in the previous equilibration algorithm needed to create an optimisation version.

Now we suppose that if we are at iteration n; at a non-ε-linearly-optimal $x_n \in H$ where the search direction is $\Delta_\varepsilon(x_n)$ and the step length actually used at x_n is u_n then our next x will be $x_{n+1} = Proj(x_n + u_n\Delta_\varepsilon(x_n))$ where $Proj$ now denotes projection onto H.

Initially we suppose that if u_n is any used step length then $Proj(x_n + u_n\Delta_\varepsilon v(x_n)) = x_n + u_n\Delta_\varepsilon(x_n)$. That is, we suppose that the projection operator does not actually do anything; and so the boundary of the feasible set H is here having no effect.

Let, for $x \in H$,

$$G_1(x) = -\nabla V(x)\cdot\Delta_\varepsilon(x) \text{ and } G_2(x) = -\nabla Z(x)\cdot\Delta_\varepsilon(x).$$

Unlike in the previous pure equilibration case where there is a formula for G, here $G_1(x)$ and $G_2(x)$ must both be estimated in an algorithm using a two-point estimation where the second point lies a short distance in the direction Δ_ε from the first.

For our purposes here we specify a simple dynamic Armijo-like optimisation scheme based on the previous dynamic Armijo-like equilibration scheme but now taking account of the two objective functions V and Z and the two continuous gap functions G_1 and G_2 above, instead of just the previous V and the previous G.

As before we update (x_n, t_{n-1}). To be specific, we start at an arbitrary $x_1 \in H$ and (arbitrarily as before) $t_0 = 1$.

If we are at a current non-ε-linearly-optimal point x_n and the previous possible step length was t_{n-1} then we are to update x_n and t_{n-1} (for $n \geq 1$) according to the following rules. These imitate the earlier equilibration rules. To be most general we will, in this algorithm statement, set

$$V[y] = V(Proj_H(y)) \text{ for all } y \in R^{N+K}.$$

The update of (x_n, t_{n-1}) depends on whether $V(x_n) > \varepsilon$ or $V(x_n) \leq \varepsilon$.

I. Suppose that $V(x_n) > \varepsilon$. In this case the algorithm is essentially the equilibration algorithm above. Firstly, x_n is kept fixed and t_{n-1} is halved as before to obtain the sequence $t_{n-1}, (\frac{1}{2})t_{n-1}, (\frac{1}{2})^2 t_{n-1}, \cdots, (\frac{1}{2})^p t_{n-1}$ where the halving ceases as soon as:

$$V[x_n + (\tfrac{1}{2})^p t_{n-1} \Delta_\varepsilon(x_n)] - V(x_n) < -(\tfrac{1}{8})(\tfrac{1}{2})^p t_{n-1} G_1(x_n)$$

for the first time. $p = 0$ is allowed here; it may be that

$$V[x_n + t_{n-1}\Delta_\varepsilon(x_n)] - V(x_n) < -(\tfrac{1}{8})t_{n-1}G_1(x_n)$$

already. The halving surely ceases by definition of G_1 since $\Delta_\varepsilon(x_n)$ is always a descent direction for V at x_n if $V(x_n) > \varepsilon$.

Then let $u_n = (\frac{1}{2})^p t_{n-1}$ for this p (this is to be the used step length at x_n) and let $x_{n+1} = x_n + u_n \Delta_\varepsilon(x_n)$.

Finally, update t_{n-1} as follows:

- If $V[x_n + u_n\Delta_\varepsilon(x_n)] - V(x_n) < -\frac{3}{4}u_n G_1(x_n)$, put $t_n = 2u_n$.
- If $-\frac{3}{4}u_n G_1(x_n) \leq V[x_n + u_n\Delta_\varepsilon(x_n)] - V(x_n) < -\frac{1}{4}u_n G_1(x_n)$, put $t_n = u_n$.
- If $-\frac{1}{4}u_n G_1(x_n) \leq V[x_n + u_n\Delta_\varepsilon(x_n)] - V(x_n) < -\frac{1}{8}u_n G_1(x_n)$, put $t_n = \frac{1}{2}u_n$.

These are three mutually exclusive possibilities and they together exhaust all eventualities since, by choice of $u_n = (\frac{1}{2})^p t_{n-1}$,

$$V[x_n + u_n\Delta(x_n)] - V(x_n) \geq -\tfrac{1}{8}u_n G_1(x_n)$$

is not possible.

II. Suppose that $V(x_n) \leq \varepsilon$. Firstly, x_n is kept fixed and t_{n-1} is halved to obtain the sequence t_{n-1}, $(\frac{1}{2})t_{n-1}$, $(\frac{1}{2})^2 t_{n-1}, \cdots$, $(\frac{1}{2})^p t_{n-1}$where the halving ceases as soon as:

$$Z[x_n + (\tfrac{1}{2})^p t_{n-1}\Delta_\varepsilon(x_n)] - Z(x_n) < -\tfrac{1}{8}(\tfrac{1}{2})^p t_{n-1} G_2(x_n)$$

and

$$V[x_n + (\tfrac{1}{2})^p t_{n-1} \varDelta_\varepsilon(x_n)] \leq \varepsilon$$

for the first time. We allow $p = 0$ here as before. The halving surely ceases by definition of G_2 and the Z-descent property of $\varDelta_\varepsilon$ maintaining $V \leq \varepsilon$, because x_n is not ε-linearly-optimal.

Then let $u_n = (\frac{1}{2})^p t_{n-1}$ for this p (this is to be the used step length at x_n) and let $x_{n+1} = x_n + u_n \varDelta_\varepsilon(x_n)$.

Finally update t_{n-1} as follows:

- If $Z[x_n + u_n \varDelta_\varepsilon(x_n)] - Z(x_n) < -\frac{3}{4} u_n G_2(z_n)$, put $t_n = 2u_n$.
- If $-\frac{3}{4} u_n G_2(x_n) \leq Z[x_n + u_n \varDelta_\varepsilon(x_n)] - Z(x_n) < -\frac{1}{4} u_n G_2(x_n)$, put $t_n = u_n$.
- If $-\frac{1}{4} u_n G_2(x_n) \leq Z[x_n + u_n \varDelta_\varepsilon(x_n)] - Z(x_n) < -\frac{1}{8} u_n G_2(x_n)$, put $t_n = \frac{1}{2} u_n$.

Again these are three mutually exclusive possibilities and they together exhaust all eventualities since, by choice of $u_n = (\frac{1}{2})^p t_{n-1}$,

$$Z[x_n + u_n \varDelta_\varepsilon(x_n)] - Z(x_n) \geq -(\tfrac{1}{8}) u_n G_2(x_n)$$

is not possible.

The algorithm is terminated as soon as $V(x_n) - \varepsilon$ and $G_2(x_n)$ are both less than preassigned tolerances.

8 Convergence to a Stationary Point

We assume that our price-enhanced basic condition, given in Section 7.3, holds.

8.1 Convergence preliminaries

For any starting point $x_1 \in H$ and with an initial step length $t_0 = 1$ we suppose that the algorithm specified above in Section 7.6 generates an infinite trajectory:

$$x_1, x_2, x_3, \cdots, x_n, \cdots$$

and an infinite sequence $t_0, t_1, t_2, t_3, \cdots, t_n, \cdots$ of possible step lengths.

To prove convergence in this case we will use proof by contradiction. So we suppose that $\{x_n\}$ is a given infinite sequence generated by our algorithm which does not converge to the set O_ε of ε-linearly-optimal points. Then not all limit points of the sequence $\{x_n\}$ belong to O_ε and since H is closed and bounded the sequence has a limit point w which is not in O_ε.

Thus we now make the basic assumption that w is not ε-linearly optimal and is also a limit point of $\{x_n\}$. We will show that this leads to a contradiction. This will show that, in fact, if the sequence $\{x_n\}$ is generated by the algorithm then all limit points of the sequence $\{x_n\}$ are in O_ε and so (as H is closed and bounded) $dist(x_n, O_\varepsilon) \to 0$ as $n \to \infty$.

8.2 Convergence proof

Theorem 3. *Let our objective function Z satisfy: $\nabla Z(x) \neq 0$ for all x in H. Let our price-enhanced basic condition in Section 7.3 hold. Suppose that the sequence $\{x_n\}$ is generated by the optimisation algorithm in Section 7.6 and also lies in the interior of H, let x^* be the limit of any subsequence of the above sequence $\{x_n\}$, and let $x^* \in intH$. Then x^* is ε-linearly-optimal.*

Proof. Suppose that x^* is the limit of a subsequence of the above sequence and that $x^* \in intH$. We will show that none of the following three alternatives can occur:

1. $V(x^*) > \varepsilon$,
2. $0 \leq V(x^*) < \varepsilon$,
3. $V(x^*) = \varepsilon$ and $desc(Z, V)(x) \neq 0$.

It will then follow that Case 4 in Section 7.5 above holds or that $V(x^*) = \varepsilon$ and $desc(Z, V)(x^*) = 0$. In this case x^* is ε-linearly optimal.

So let x^* be the limit of a subsequence (in the interior of H).

Ruling out Case 1: This is essentially as before with equilibration in Section 6.

Ruling out Case 2: Suppose that Case 2 does hold or that $V(x^*) < \varepsilon$. Then

$$\begin{aligned}\Delta_\varepsilon(x*) =& [V(x^*)/\varepsilon] desc(Z, V)(x^*) + [1 - V(x^*)/\varepsilon]_+ descZ(x^*)\\ &+ [V(x^*)/\varepsilon - 1] + descV(x^*)\\ =& [V(x^*)/\varepsilon] desc(Z, V)(x^*) + [1 - V(x^*)/\varepsilon] descZ(x^*) \neq 0.\end{aligned}$$

Now $[1 - V(x^*)/\varepsilon] descZ(x^*)$ is a descent direction for Z at x^* (and $desc(Z, V)(x^*)$ is never ascent) so $[\nabla Z(x^*)] \cdot \Delta_\varepsilon(x^*) < 0$.

A small enhancement of the equilibration argument in Section 6 above and the appendix below now works in this case too, but with Z instead of V, ruling out Case 2.

Ruling out Case 3: Suppose that Case 3 does hold or that $V(x^*) = \varepsilon$ and $desc(Z, V)(x^*)$ is non-zero. Then

$$\begin{aligned}\Delta_\varepsilon(x^*) =& [V(x^*)/\varepsilon] desc(Z, V)(x^*) + [1 - V(x^*)/\varepsilon]_+ descZ(x^*)\\ &+ [V(x^*)/\varepsilon - 1]_+ descV(x^*)\\ =& desc(Z, V)(x^*) \neq 0,\end{aligned}$$

and (being non-zero) is a descent direction for both V and Z at x^*.

An enhancement of the equilibration argument in Section 6 and the appendix now works in this case too, but this argument must in this case be applied to both V and Z. The radius r in this latter case is to be the minimum of two radii, one ensuring a suitable reduction in V (so that $V < V(x^*) = \varepsilon$ still, once the ball $B(x^*, r)$ is left) and the other ensuring a suitable reduction in Z (so that $Z < Z(x^*)$ once the ball $B(x^*, r)$ is left). This rules out (3) as then the sequence $\{x_n\}$ can never return to be close to x^* again, and x^* cannot be a limit point of the sequence $\{x_n\}$.

Conclusion: It follows from the arguments above that if x^* is any limit point in the interior of H, Case 4 in Section 7.5 must hold or:

$$V(x^*) = \varepsilon \text{ and } desc(Z, V)(x^*) = 0.$$

In this case x^* is ε-linearly-optimal, as we have seen from the ε-linearly-optimality conditions above in Section 7.5. ■

9 Allowing for the Boundary of *H*

We now take account of the boundary of H by re-designing the search direction Δ_ε. We change the directions

$$descV(x), descZ(x) \text{ and } desc(Z, V)(x)$$

for those x close to the boundary of H to new directions called:

$$desc_H V(x), desc_H Z(x) \text{ and } desc_H(Z, V)(x).$$

We also change Δ_ε to the new direction $\Delta_{H\varepsilon}$. Sometimes, on the boundary of H, these changes are achieved by a straightforward projection onto H. Near to the boundary of H the change is in general a convex combination of different projections onto different subsets of H.

In order to specify the new directions for all x in H we order the N_H constraint function values $h_k(x)$ at x in order of decreasing size. So we always have

$$0 \geq h_1(x) \geq h_2(x) \geq h_3(x) \geq h_4(x) \cdots \geq h_{N_H}(x).$$

Closeness: For the purposes of this paper we shall suppose that x in H is close to the boundary "$h_k = 0$" if $0 \geq h_k(x) \geq -1$. This specification may be subject to obvious variation, especially in the case that H is a box. In this case we might normalise the h_k so that each h_k takes the value 0 on one face and -100 (say) on the opposing face. Thus the "1" here is to be thought of as a number which is small compared to the separation (according to the h functions) of the faces defining H. If H is not a box then we might suppose that the "non-box" h functions are arranged to have a minimum value of about -100 in H.

9.1 x close to one of the constraints defining H

Suppose first that x is close to the boundary of just one of the constraints defining H: given by $h_1(x) \leq 0$. So let us agree that $0 \geq h_1(x) \geq -1$ and $\mathrm{h}_k(x) < -1$ for all $k > 1$. For any such x we define the new $desc_H V(x)$ by linearly interpolating between

1. $desc_0 V(x)$ (which is the "original" $descV(x)$) and
2. $desc_1 V(x)$ (obtained by projecting $descV(x)$ onto the single constraint $\{x; h_1(x) \geq 0\}$ *as if x were such that* $h_1(x) = 0$).

We need to specify precisely: $desc_0 V(x)$, $desc_1 V(x)$ and the linear interpolation. We also need to do the same for $descZ$ and $desc(Z, V)$.

$desc_0 V(x)$: Of course we let $desc_0 V(x) = descV(x)$. (It is the result of projecting $descV(x)$ onto zero constraints).

$desc_1 V(x)$: We define $desc_1 V(x)$, the "modified" $descV(x)$, to be the projection of $descV(x)$ onto the 1 constraint $h_1(x) \leq 0$ as if x were such that $h_1(x) = 0$. This projection may be calculated by solving the following minimisation problem in ν_1:

$$\begin{aligned} &\min \; ||descV(x) + \nu_1 desch_1(x)|| \\ &\text{s.t. } \; \nu_1 \geq 0. \end{aligned}$$

Let the solution be ν_1^*. Then, $desc_1 V(x) = descV(x) + \nu_1^* desch_1(x)$.

Interpolation ($desc_H V(x)$ when $0 \geq h_1(x) \geq -1$ and $h_i(x) < -1$ for $i > 1$): Here we combine $desc_0 V(x)$ and $desc_1 V(x)$ by using x-dependent weights $w_0 = 0 - h_1(x)$ and $w_1 = h_1(x) - (-1)$, and putting $desc_H V(x) = w_0 desc_0 V(x) + w_1 desc_1 V(x)$. Thus we have defined $desc_H V(x)$ for all x such that $0 \geq h_1(x) \geq -1$ and $h_i(\mathrm{x}) < -1$ if $i > 1$; by linear interpolation between $desc_0 V(x)$ and $desc_1 V(x)$.

We define $desc_H Z(x)$ similarly and also we define $desc_H(Z, V)(x)$ similarly too as follows.

$desc_0(Z, V)(x)$: We put $desc_0(Z, V)(x) = desc(Z, V)(x) = \frac{1}{2} descV(x) + \frac{1}{2} descZ(x)$. This expression also arises by solving the following minimisation problem in λ, μ:

$$\begin{aligned} &\min \; ||\lambda descV(x) + \mu descZ(x)|| \\ &\text{s.t. } \; \lambda + \mu = 1, \\ &\qquad \lambda \geq 0, \mu \geq 0. \end{aligned}$$

If the solution is λ^* and μ^* then of course $\lambda^* = \frac{1}{2}, \mu^* = \frac{1}{2}$.

$desc_1(Z,V)(x)$: In this case, we need to project $descV$ and $descZ$ " simultaneously" onto the single constraint $h_1(x) \leq 0$, as if $h_1(x)$ were $= 0$. So now we solve the following minimisation problem in λ, μ and ν_1:

$$\begin{aligned} &\min ||\lambda descV(x) + \mu descZ(x) + \nu_1 desch_1(x)|| \\ &\text{s.t. } \lambda + \mu = 1, \\ &\qquad \lambda, \mu, \nu_1 \geq 0. \end{aligned}$$

If the solution is λ^*, μ^* and ν_1^*, then $desc_1(Z,V)(x) = \lambda^* descV(x) + \mu^* descZ(x) + \nu_1^* desch_1(x)$.

Interpolation ($desc_H(Z,V)(x)$ when $0 \geq h_1(x) \geq -1$ and $h_i(x) < -1$ for $i > 1$): Here we combine naturally $desc_0(Z,V)(x)$ and $desc_1(Z,V)(x)$ by using weights $w_0 = 0 - h_1(x)$ and $w_1 = h_1(x) - (-1)$ and putting $desc_H(Z,V)(x) = w_0 desc_0(Z,V)(x) + w_1 desc_1(Z,V)(x)$. Thus we have defined $desc_H(Z,V)(x)$ for all x such that $0 \geq h_1(x) \geq -1$; by linear interpolation between $desc_0(Z,V)(x)$ and $desc_1(Z,V)(x)$.

Comment: These modified vectors all have the character of a reduced gradient; the reduction being embodied in terms like $w_1 \nu_1^* desch_1(x)$ which, for $0 \geq h_1(x) \geq -1$, have the effect of reducing those components of the original $descV(x)$ or $descZ(x)$ or $desc(Z,V)(x)$ pointing towards the edge of H. If x is actually on the edge of H then these modified vectors will have zero components pointing outwards. For example, if $h_1(x) = 0$ and $descV(x) \cdot \nabla h_1(x) > 0$ (so that $descV(x)$ points out of H) then $desc_H V(x) \cdot \nabla h_1(x) = 0$ (so that the modification $desc_H V(x)$ does not point out of H).

9.2 x close to two of the constraints defining H

To show how to extend the above ideas we consider just the two-constraint modification of $desc(Z,V)$. So now suppose that $0 \geq h_1(x) \geq h_2(x) \geq -1$ and $h_i(x) < -1$ if $i > 2$. In this case we need to project $descV$ and $descZ$ "simultaneously" onto the two constraints $h_1(x) \leq 0$ and $h_2(x) \leq 0$ as if $h_1(x) = 0$ and $h_2(x) = 0$. More precisely, we now solve the minimisation problem in λ, μ, ν_1, ν_2:

$$\begin{aligned} &\min ||\lambda descV(x) + \mu descZ(x) + \nu_1 desch_1(x) + \nu_2 desch_2(x)|| \\ &\text{s.t. } \lambda + \mu = 1, \\ &\qquad \lambda, \mu, \nu_1, \nu_2 \geq 0. \end{aligned}$$

If the solution is λ^*, μ^*, ν_1^* and ν_2^*, then $desc_2(Z,V)(x) = \lambda^* descV(x) + \mu^* descZ(x) + \nu_1^* desch_1(x) + \nu_2^* desch_2(x)$. The suffix "2" indicates that there are two constraint functions, h_1 and h_2, involved in the projection.

Interpolation ($desc_H(Z,V)(x)$ when $0 \geq h_1(x) \geq h_2(x) \geq -1$ and $h_i(x) < -1$ for $i > 2$): We combine $desc_0(Z,V)(x), desc_1(Z,V)(x)$ and $desc_2(Z,V)(x)$, just defined by using weights $w_0 = 0 - h_1(x)$, $w_1 = h_1(x) - h_2(x)$, and $w_2 = h_2(x) - (-1)$ as follows:

$$desc_H(Z,V)(x) = w_0 desc_0(Z,V)(x) + w_1 desc_1(Z,V)(x) + w_2 desc_2(Z,V)(x).$$

Thus we have defined $desc_H(Z,V)(x)$ for all x such that $0 \geq h_1(x) \geq h_2(x) \geq -1$ and $h_i(x) < -1$ for $i > 2$; by linear interpolation between the values $desc_0(Z,V)(x), desc_1(Z,V)(x)$ and $desc_2(Z,V)(x)$.

9.3 x close to three of the constraints defining H

Here we follow exactly the same lines as shown above to obtain just $desc_H$ $(Z,V)(x)$ for an x such that $0 \geq h_1(x) \geq h_2(x) \geq h_3(x) \geq -1$ and $h_i(x) < -1$ for $i > 3$. We focus just on the interpolation stage.

We combine $desc_0(Z,V)(x)$, $desc_1(Z,V)(x)$, $desc_2(Z,V)(x)$, and $desc_3$ $(Z,V)(x)$, by using weights $w_0 = 0 - h_1(x)$, $w_1 = h_1(x) - h_2(x)$, $w_2 = h_2(x) - h_3(x)$, and $w_3 = h_3(x) - (-1)$ as follows:

$$\begin{aligned} desc_H(Z,V)(x) &= w_0 desc_0(Z,V)(x) + w_1 desc_1(Z,V)(x) \\ &+ w_2 desc_2(Z,V)(x) + w_3 desc_3(Z,V)(x). \end{aligned} \tag{10}$$

9.4 Generalisation: x close to several constraint boundaries

For any x in H close to the boundary of H, there is $k(x) \geq 1$ such that

$$0 \geq h_1(x) \geq h_2(x) \geq h_3(x) \geq h_4(x) \geq \cdots \geq h_{k(x)}(x) \geq -1$$

and $h_i(x) < -1$ for $i > k(x)$. ($k(x)$ is the number of the H-constraint boundaries x is close to.)

The above special specifications may now be generalised naturally to give $desc_H V(x)$, $desc_H Z(x)$ and $desc_H(Z,V)(x)$ at any such x in H, however many constraint boundaries this x is close to; that is, whatever $k(x)$ may be. Here we just outline the general extrapolation procedure for $desc_H(Z,V)$ where $k(x) > 0$.

For each j satisfying $1 \leq j \leq k(x)$, we solve the problem:

$$\begin{aligned} \min \; & ||\lambda descV(x) + \mu descZ(x) + \sum_{1 \leq i \leq k(x)} \nu_i desch_i(x)|| \\ \text{s.t. } & \lambda + \mu = 1, \\ & \lambda, \mu, \nu_i \geq 0 \text{ for } 1 \leq i \leq j \\ & \nu_i = 0 \text{ for } i > j. \end{aligned}$$

The sum is over $k(x) \geq 1$ terms. If the solution is $\lambda^j, \mu^j, \nu_1^j, \nu_2^j, \nu_3^j, \cdots, \nu_k^j$ then ($\nu_i^j = 0$ for $i > j$ and) $desc_j(Z,V)(x)$ is defined by:

$$desc_j(Z,V)(x) = \lambda^j descV(x) + \mu^j descZ(x) + \sum_{1 \le i \le k(x)} \nu_i^j desch_i(x).$$

Then these special $desc_j(Z,V)(x)$ (for $j = 1, 2, \cdots, k(x)$) are combined with $desc_0(Z,V)(x)$ to form the whole vector $desc_H(Z,V)(x)$ as follows:

$$desc_H(Z,V)(x) = \sum_{j=0}^{k(x)} w_j desc_j(Z,V)(x)$$

where

$$w_0 = 0 - h_1(x), w_1 = h_1(x) - h_2(x), \cdots, w_{k-1} = h_{k(x)-1}(x) - h_{k(x)}(x),$$

and $w_{k(x)} = h_{k(x)}(x) - (-1)$.

There are $k(x) + 1$ terms here including the 0^{th} term $desc_0(Z,V)(x) = desc(Z,V)(x)$, which is calculated as if there are no constraints. For any x in H satisfying $h_i(x) < -1$ for all i, so that x is not close to any of the boundaries of H,

$$desc_H(Z,V)(x) = desc_0(Z,V)(x) = \frac{1}{2} descZ(x) + \frac{1}{2} descV(x).$$

Reduced direction at x: As a consequence of the above specifications we are now able to amend the definition of $\Delta_\varepsilon(x)$ for x in H, reducing the degree to which this vector points toward nearby boundaries of H. We obtain:

$$\begin{aligned} \Delta_{H\varepsilon}(x) = {} & [V(x)/\varepsilon] desc_H(Z,V)(x) \\ & + [V(x)/\varepsilon - 1]_+ desc_H V(x) + [1 - V(x)/\varepsilon]_+ desc_H Z(x). \end{aligned}$$

This is a continuous function of x in H. It is also, if $V(x) = \varepsilon$, an H-feasible simultaneous descent direction, for both Z and V, provided such a direction exists.

9.5 Simultaneous descent allowing for the boundary of H

Here we generalise the earlier "interior to H" approach to the case where points generated may be close to or on the boundary of H, using $\Delta_{H\varepsilon}$.

Definition of an ε-feasible Z-descent vector at $x \in H \cap E_\varepsilon$ (for x possibly on the boundary of H): The vector u is an ε-feasible Z-descent vector at $x \in H \cap E_\varepsilon$ if:

$$\begin{aligned} & u \cdot descZ(x) > 0, \\ & u \cdot desch_i(x) \ge 0 \text{ if } h_i(x) = 0 \text{ (for all } i\text{) and} \\ & u \cdot descV(x) \ge 0 \text{ if } V(x) = \varepsilon. \end{aligned}$$

Definition of ε-linear-optimality (for x possibly on the boundary of H): x is ε-linearly-optimal if $x \in H \cap E_\varepsilon$, and there is no ε-feasible Z-descent direction from x.

ε-linear-optimality conditions (valid also at points on the boundary of H)

Let x belong to H and let $\varepsilon > 0$. Now we consider five cases:

1. $V(x) > \varepsilon$,
2. $0 \leq V(x) < \varepsilon$ and $desc_H Z(x) \neq 0$,
3. $V(x) = \varepsilon$ and $desc_H(Z, V)(x) \neq 0$,
4. $0 \leq V(x) < \varepsilon$ and $desc_H Z(x) = 0$, and
5. $V(x) = \varepsilon$ and $desc_H(Z, V)(x) = 0$.

In each of the first three of these cases we show that x is not ε-linearly-optimal and that $\Delta_{H\varepsilon}$ is then a direction in which either the degree of disequilibrium, V, improves (Case 1); or is a direction in which Z improves maintaining $V \leq \varepsilon$ (Cases 2 and 3). We also show that, in Cases 4 and 5, x is ε-linearly-optimal.

Case 1 ($V(x) > \varepsilon$): In this case $\Delta_{H\varepsilon} = [V(x)/\varepsilon] desc_H(Z, V)(x) + [V(x)/\varepsilon - 1] desc_H V(x)$ is non-zero as $desc_H V(x)$ is non-zero (as, for example, the algorithm (D) direction is an H-feasible descent direction) and of course x is not feasible and so is not ε-linearly-optimal. Here following $\Delta_{H\varepsilon}$ reduces V.

Case 2 ($0 \leq V(x) < \varepsilon$ and $desc_H Z(x) \neq 0$): In this case $\Delta_{H\varepsilon} = [V(x)/\varepsilon] desc_H(Z, V)(x) + [1 - V(x)/\varepsilon] desc_H Z(x)$ is again non-zero as $desc_H$ $Z(x)$ is non-zero and is a feasible descent direction for Z at x so x is not ε-linearly-optimal. Here following $\Delta_{H\varepsilon}$ reduces Z maintaining $V \leq \varepsilon$.

Case 3 ($V(x) = \varepsilon$ and $desc_H(Z, V)(x) \neq 0$): In this case $\Delta_{H\varepsilon}(x) = desc_H(Z, V)(x) \neq 0$, and is a simultaneous descent direction for both V and Z. Thus $\Delta_{H\varepsilon}(x)$ is non-zero and is a feasible descent direction at x so x is not ε-linearly-optimal. Here following $\Delta_{H\varepsilon}$ reduces Z maintaining $V \leq \varepsilon$.

Case 4 ($0 \leq V(x) < \varepsilon$ and $desc_H Z(x) = 0$): In this case there is clearly no feasible descent direction for Z at x so x is ε-linearly-optimal. $\Delta_{H\varepsilon}(x) = 0$.

Case 5 ($V(x) = \varepsilon$ and $desc_H(Z, V)(x) = 0$): In this case $\Delta_{H\varepsilon}(x) = desc_H(Z, V)(x) = 0$. Consider any H-feasible Z-descent direction u from x. Then (by the following lemma) $u \cdot descV(x) < 0$, and u is not an ε-feasible direction. Thus there is no ε-feasible descent direction and x is ε-linearly-optimal. $\Delta_{H\varepsilon}(x) = 0$.

Lemma 3. *Suppose that $x \in H \cap E_\varepsilon$, $V(x) = \varepsilon$, $desc_H(Z,V)(x) = 0$ and $h_i(x) = 0$ for just those $i = 1,2,3,\cdots k$. Let u be an H-feasible Z-descent direction at x, so that $u \cdot descZ(x) > 0$ and $u \cdot desch_i(x) \geq 0$ for $i = 1,2,3,\cdots k$. Then $u \cdot descV(x) < 0$ and u is not an ε-feasible direction.*

Proof. Suppose that $x \in H \cap E_\varepsilon, V(x) = \varepsilon$, $h_i(x) = 0$ for just those $i = 1,2,3,\cdots,k$ and $desc_H(Z,V)(x) = 0$. Suppose now also that u satisfies:

$$u \cdot descZ(x) > 0; u \cdot desch_i(x) \geq 0 \text{ for } i = 1,2,3,\cdots,k; u \cdot descV(x) \geq 0.$$

By condition (7) in Section 7.3, if x is any point in the set $H \cap E_\varepsilon$ such that $V(x) = \varepsilon$ and $h_i(x) = 0$ for $i = 1,2,3,\cdots,k$ then there is a direction δ such that

$$\delta \cdot desch_i(x) \geq 0 \text{ for } i = 1,2,3,\cdots,k \text{ and } \delta \cdot descV(x) > 0.$$

This δ is an H-feasible descent direction for V at x. Then for a sufficiently small $\theta > 0$:

$$\begin{aligned}&(u + \theta\delta) \cdot descZ(x) > 0,\\&(u + \theta\delta) \cdot desch_i(x) \geq 0 \text{ for } i = 1,2,\cdots,k \text{ and}\\&(u + \theta\delta) \cdot descV(x) > 0.\end{aligned}$$

This direction $(u + \theta\delta)$ reduces both Z and V and is H-feasible and so $desc_H(Z,V)(x) \neq 0$, which is not so. It follows that if the Z-descent vector u satisfies $u \cdot descZ(x) > 0$, and $u \cdot desch_i(x) \geq 0$ for $i = 1,2,3,\cdots,k$ then $u \cdot descV(x) < 0$, as required. This result is connected to the lemma due to Farkas. ■

10 Convergence to a Stationary Point in *H*

The general idea is as before.

10.1 Convergence proof

Theorem 4. *Suppose now that the sequence $\{x_n\}$ is generated by the algorithm described above in Section 7.6 with Δ_ε there replaced by the new direction $\Delta_{H\varepsilon}$; as before we start at any x_1 in H and any $t_0 > 0$. Suppose that our price-enhanced basic condition in Section 7.3 holds. Let $x^* \in H$ be the limit of any subsequence of the above sequence $\{x_n\}$. Then x^* is ε-linearly-optimal.*

Proof. Suppose that $x^* \in H$ is the limit of a subsequence of the sequence $\{x_n\}$. Consider the following alternatives:

1. $V(x^*) > \varepsilon$,
2. $0 \leq V(x^*) < \varepsilon$ and $desc_H Z(x^*) \neq 0$,
3. $V(x^*) = \varepsilon$ and $desc_H(Z,V)(x^*) \neq 0$,
4. $0 \leq V(x^*) < \varepsilon$ and $desc_H Z(x^*) = 0$ or
5. $V(x^*) = \varepsilon$ and $desc_H(Z,V)(x^*) = 0$.

Rather as before, ruling out Case 1 ensures that x^* is in E_ε, and ruling out Cases 2 and 3 then ensures that Case 4 or 5 must hold. It then follows from the previous work that $x^* \in E_\varepsilon$ and is ε-linearly-optimal.

Cases 1, 2, and 3 are ruled out by arguments similar to those already utilised in the proof of Theorem 3 in Section 8.2, but using the new direction. These arguments work as the new direction is continuous. The argument in Section 6 needs to be extended and utilised in this case, as it was in Theorem 3. Some of the detail here is omitted and is given in [Smi05b].

Thus for any limit x^*, of any subsequence, either Case 4 or 5 must hold and in each of these two cases x^* is ε-linearly-optimal, as we have seen from the ε-linear-optimality conditions in Section 9.5 above. ■

11 Optimisation in the Payne-Thompson Model

The basic structure exploited above is as follows: there is a function Φ such that $-\Phi(x,p)$ is a smooth monotone function of the non-control vector x for each fixed control vector p. Thus this same optimisation approach may, at first sight, be applied with different interpretations of Φ, x and p. These different interpretations correspond to different equilibrium models and different control parameters.

It is in fact possible to weaken these condition somewhat in various directions. However the above structure may already be applied to a variety of models with price variables present.

For example, the same approach, utilising just the above monotone structure, is applicable if prices are included within:

- Stochastic variable demand models,
- The Evans [Gravity + Wardrop] model (see [Eva76],
- Variable or fixed demand explicit queueing models,
- A variable demand explicit queueing model which allows a special responsive control policy; and
- A variable demand explicit queueing model where prices and signal controls are optimised.

Here we just consider the last of the list above. In this case there are constraints involving both state variables and control variables. This introduces complications in the previous approach which we do not address here.

11.1 Introduction to the Payne-Thompson model

In order to study ramp control on freeways, [PT75] introduced an equilibrium model with queueing delays at bottlenecks. In their model as long as a bottleneck is saturated the queueing delay at the bottleneck is independent of flow, and is represented by an independent non-negative variable determined by the equilibrium conditions; but if the bottleneck is unsaturated then the bottleneck delay must be zero. Bringing in the delays at bottlenecks in this way, as separate variables, allows ramp metering to be modelled sensibly. The "new" independent bottleneck delays are here added to costs arising from a more standard cost-flow function which may be thought of as applying to the rest of the arc.

We here consider the Payne-Thompson model with capacity constraints and explicit queueing delays. We insert signal green-times as in [Smi87].

11.2 The Payne-Thompson model with prices and controls

Let v_a denote the traffic flow along arc a, let s_a be the saturation flow at the exit of arc a (both in vehicles per minute), let b_a be the bottleneck delay or cost at the exit of arc a (in minutes per vehicle) and let q_a be the proportion of time that arc a is green.

We assume given non-decreasing arc cost functions $c_a : \mathbb{R}_+ \to \mathbb{R}_+$ (typically with $c_a(v_a)$, in minutes per vehicle, defined for all $v_a \geq 0$). Then the whole cost (in minutes/vehicle) of traversing arc a is to be $c_a(v_a) + b_a$ where v_a, q_a and b_a must together satisfy the "delay-equilibrium" condition:

$$v_a \leq q_a s_a \text{ and if } v_a < q_a s_a \text{ then } b_a = 0$$

or

$$(v, q) \text{ is supply-feasible and unsaturated bottlenecks cause no delay.}$$

Then we make the following definitions:

$$\begin{aligned} v_a(X) &= \text{ flow along arc } a = \sum_{\text{relevant } ijr} X_{ijr}. \\ C_{ijr}(X) &= \text{ non-bottleneck cost of travel along route } ijr = \sum_{\text{relevant } a} c_a(v_a). \\ B_{ijr}(b) &= \text{ bottleneck delay or cost on route } ijr = \sum_{\text{relevant } a} b_a. \end{aligned}$$

The equilibrium conditions already introduced now become (for a fixed price vector P and a fixed arc green-time vector q): for all i, j, r, a,

$$Y_{ij} - C_{ijr}(X) - B_{ijr}(b) - P_{ijr} \leq 0 \text{ and}$$
$$Y_{ij} - C_{ijr}(X) - B_{ijr}(b) - P_{ijr} < 0 \text{ implies } X_{ijr} = 0;$$
$$D_{ij}(Y) - T_{ij}(X) = 0; \text{ and}$$
$$v_a(X) - s_a q_a \leq 0 \text{ and}$$
$$v_a(X) - s_a q_a < 0 \text{ implies } b_a = 0.$$

(All variables satisfy non-negativity constraints.)

This is the equilibrium model introduced by Payne and Thompson, with the unvarying route-price vector P and now the arc-green-time vector q added.

Suppose as before that the vector P of all route prices P_{ijr} is confined to some polyhedral closed bounded set F_{price} of feasible route-price vectors.

Suppose now that for each signal stage k a green-time Q_k is awarded. Suppose also that the vector Q of all the Q_k is confined to some polyhedral closed bounded set of feasible Q. Then some arcs a will be subject to a corresponding green-time q_a (the sum of the relevant Q_k) and the vector q of all arc green-times will be confined to some polyhedral closed bounded set F_{green} of feasible green-time vectors.

It is easy to check that, for each (P, q), - (the left hand side above) is a monotone function of (X, Y, b). This strongly suggests that the optimisation method described previously, suitably developed, may be utilised in this rather different equilibrium setting.

12 Conclusion

This paper has specified a method for approximately solving bilevel optimisation problems, in which a stationary point of a given objective Z is sought subject to the flow pattern being an approximate variable demand equilibrium. It has been shown that if the cost function and - (the demand function) are monotone and smooth then there are limit points and any interior limit point generated by the suggested algorithm is an approximate equilibrium (at which $V \leq \varepsilon$) which is stationary for the objective function Z. The method utilises the "simultaneous descent" direction; this is obtained under certain (interior) conditions simply by bisecting the angle between $-\nabla V$ and $-\nabla Z$; in this direction V and Z simultaneously decline.

The paper has also shown how the simultaneous descent direction may be modified close to the boundary of the feasible region. Now, under our conditions, (1) the search direction obtained is continuous and (2) the linear H-feasibility constraints, together with the nonlinear equilibrium constraint $V \leq \varepsilon$, satisfy a constraint qualification. This has allowed the approach also to work when the hard feasibility constraints are active during the optimisation process or at a limit point.

Finally the paper has shown the direction of an extension so as to optimise signal controls and prices; using a network equilibrium model with explicit queueing delays introduced by Payne and Thompson. It would be possible to introduce signal controls without moving to the Payne-Thompson model; but this model is a "natural" for signal control.

Of course the optimisation problem here is non-convex so a variety of start points should be taken and the optimisation procedure followed from each.

References

[AM83] Aashtiani, H., Magnanti, T.: Equilibrium on a congested transport network. SIAM Journal of Algebraic and Discrete Methods, **2**, 213–216 (1983)

[AL79] Abdulaal, M., Leblanc, L.: Continuous network design problems. Transportation Research, **13B**, 19–32 (1979)

[All74] Allsop, R.E.: Some possibilities for using traffic control to influence trip distribution and route choice. Proceedings of the 7th International Symposium on Transportation and Traffic Theory, 345–374 (1974)

[Bar02] Bar-Gera, H. Origin-based algorithms for the traffic assignment problem. Transportation Science, **36(4)**, 398–417 (2002)

[BB03] Bar-Gera, H., Boyce, D.: Origin-based algorithms for combined travel forecasting models. Transportation Science, **37B(5)**, 405–422 (2003)

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

[CC61] Charnes, A., Cooper, W.W.: Multicopy traffic network models. Proceedings of the Symposium on the Theory of Traffic Flow, held at the General Motors Research Laboratories, 1958, Elsevier, Amsterdam (1961)

[Chi97] Chiou, S-W.: Optimisation of area traffic control subject to user equilibrium traffic assignment. Proceedings of the 25th European Transport Forum, Seminar F, Volume II, 53–64 (1997)

[CW02] Clark, S.D., Watling, D.P.: Sensitivity analysis of the probit-based stochastic user equilibrium assignment model. Transportation Research, **36B**, 617–635 (2002)

[CS98] Clegg, J., Smith, M.J.: Bilevel optimisation of transportation networks. In: Mathematics in Transport Planning and Control, the Proceedings of the Third International IMA Conference on Mathematics in Transport Planning and Control, Pergamon, 29–36 (1998)

[CSXY01] Clegg, J., Smith, M.J., Xiang, Y., Yarrow, R.: Bilevel programming applied to optimising urban transportation. Transportation Research, **35B**, 41–70 (2001)

[CS01] Clegg, J., Smith, M.J.: Cone projection versus half-space projection for the bilevel optimisation of transportation networks. Transportation Research, **35B**, 71–82 (2001)

[CSX99] Clune, A., Smith, M., Xiang, Y.: A Theoretical Basis for Implementation of a Quantitative Decision Support System Using Bilevel Optimisation. In: Ceder, A. (ed.) Proceedings of the Fourteenth International Symposium on Transportation and Traffic Theory, Jerusalem, Pergamon, 489–514 (1999)

[CQW02] Cohen, G., Quadrat, J-P., Wynter, L.: On the convergence of the algorithm for bilevel programming problems by Clegg and Smith. Transportation Research, **36B**, 939–944 (2002)

[COM96] COMSIS: Incorporating feedback in Travel Forecasting Methods. Pitfalls and Common Concerns, Travel Model Improvement Program, Report for the US Department of Transportation (1996)

[Dav94] Davis, G.A.: Exact Local Solution of the Continuous Network Design Problem via Stochastic User Equilibrium Assignment. Transportation Research, **28B**, 61–75 (1994)

[DofE98] Department of the Environment, Transport and the Regions: A New Deal for Transport: Better for Everyone. The Stationery Office (1998)

[Eva76] Evans, S.P.: Derivation and Analysis of some Models for Combining Trip Distribution and Assignment. Transportation Research, **10(1)**, 37–57 (1976)

[Fis84] Fisk, C.S.: Optimal signal controls on congested networks. In: Volmuller, J., Hammerslag, R. (eds.) Proceedings of the Ninth International Symposium on Transportation and Traffic Theory, Delft, VNU Science Press, Utrecht, 197–216 (1984)

[FL00] Fletcher, R., Leyffer, S: Nonlinear programming without a penalty function. University of Dundee Numerical Analysis report NA 171 (2000)

[Gar80] Gartner, N.H.: Optimal traffic assignment with elastic demands: A review. Part II: Algorithmic approaches. Transportation Science, **14**, 192–208 (1980)

[GS94] Gauvin, J., Savard, G.: The steepest descent direction for the nonlinear bilevel programming problem. Operations Research Letters, **15**, 265–272 (1994)

[LPR96] Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical programs with equilibrium constraints. Cambridge University Press (1996)

[Mar83] Marcotte, P.: Network Optimisation with Continuous Control Parameters, Transportation Science, **17**, 181–197 (1983)

[Mar86] Marcotte, P.: Network Design Problem with Congestion Effects: A Case of Bilevel Programming. Mathematical Programming, **34**, 142-162 (1986)

[Mig95] Migdalas, A.: Bilevel Programming in Traffic Planning: Models, Methods and Challenge. Journal of Global Optimization, **7**, 381–405 (1995)

[OZ95] Outrata, J., Zowe, J.: A numerical approach to optimization problems with variational inequality constraints. Mathematical Programming, **68**, 105–130 (1995)

[PR02] Patriksson, M., Rockafellar, R.T.: A Mathematical model and Descent Algorithm for Bilevel Traffic Management. Transportation Science, **36**, 271–291 (2002)

[PT75] Payne, H.J., Thompson, W.A.: Traffic assignment on transportation networks with capacity constraints and queueing. Paper presented at the 47th National ORSA/TIMS North American Meeting (1975)

[RM04] Rodrigues, H.S., Monteiro, M.T.: Solving mathematical programs with complementarity constraints (MPCC) with Nonlinear Solvers. Poster at the 12th French-German-Spanish Conference on Optimization, Avignon (2004)

[SAC99] SACTRA: Transport and the Economy, The Stationery Office (1999)

[Smi84a] Smith, M.J.: A descent algorithm for solving a variety of monotone equilibrium problems. Proceedings of the Ninth International Symposium on Transportation and Traffic Theory, The Netherlands, VNU Science Press, Utrecht, 273 - 297 (1984a)

[Smi84b] Smith, M.J.: A Descent Method for Solving Monotone Variational Inequalities and Monotone Complementarity Problems. Journal of Optimization Theory and Applications, **44**, 485–496 (1984b)

[Smi87] Smith, M.J.: Traffic control and traffic assignment in a signal-controlled network with queueing. In: Gartner, N., Wilson, N.H.M. (eds.) Proceedings of the Tenth International Symposium on Transportation and Traffic Theory, MIT, 319–338 (1987)

[Smi05a] Smith, M.J.: Bilevel optimisation of prices in a variey of transportation models. In: Mahmassani, H.S. (ed.) Proceedings of the Sixteenth International Symposium on Transportation and Traffic Theory, University of Maryland, 1–21 (2005a)

[Smi05b] Smith, M.J.: Simultaneous descent: some details. Working paper available from the University of York (2005b)

[SXY97] Smith, M.J., Xiang, Y., Yarrow, R.: Bilevel optimisation of signal timings and road prices on urban road networks. Preprints of the IFAC/IFIP/IFORS Symposium, Crete, 628–633 (1997) (available from the University of York)

[SXY98] Smith, M.J., Xiang, Y., Yarrow, R.: Descent Methods of Calculating Locally Optimal Signal Controls and Prices in Multi-Modal and Dynamic Transportation Networks. In: Bell, M.G.H. (ed.) Selected Proceedings of the 4th EURO Transportation Meeting, University of Newcastle, 9–34 (1998)

[SXYG98] Smith, M.J., Xiang, Y., Yarrow, R., Ghali, M.O.: Bilevel and Other modelling Approaches to Urban Traffic Management and Control. Paper presented at the 25th Birthday of the Centre de Reserche sur les Transports, University of Montreal (1996), In: Marcotte. P., Nguyen, S. (eds.) Equilibrium and Advanced Transportation Modelling. Kluwer Academic Publishers, Massachussetts, 283–325 (1998)

[TGA79] Tan, H.N., Gershwin, S.B., Athans, M.: Hybrid optimization in urban transport networks. Laboratory for information and Decision Systems, Technical Report DOT-TSC-RSPA-79-7; published by Massachusetts Institute of Technology, Cambridge Massachusetts, USA (1979)

[TF88] Tobin, R.L., Friesz, T.L.: Sensitivity analysis for equilibrium network flow. Transportation Science, **22**, 242–250 (1988)

[War52] Wardrop, J.G.: Some Theoretical Aspects of Road Traffic Research. Proceedings, Institution of Civil Engineers II, **1**, 235–278 (1952)

[YY94] Yang, H., Yagar, S.: Traffic assignment and traffic control in general freeway-arterial corridor systems. Transportation Research, **28B**, 463–486 (1994)

[Yan96a] Yang, H. Sensitivity analysis for queueing equilibrium network flow and application to traffic control. Mathematical and Computer Modelling, **22**, 247–258 (1996a)

[Yan96b] Yang, H.: Sensitivity analysis for the elastic-demand network equilibrium problem with applications, Transportation Research, **31B**, 55–70 (1996b)

[Yan96c] Yang, H.: Equilibrium network traffic signal setting under conditions of queueing and congestion. In: Stephanedes, Y.J., Filippi, F. (eds.) Applications of Advanced Technologies in transportation Engineering. Proceedings of the 4th International Conference, American Society of Civil Engineers, 578–582 (1996c)

Appendix

This appendix justifies the argument in Section 6.

Descent of $\{V(z_n)\}$ below $V(w)$ if w is a limit point of the sequence $\{z_n\}$ and $V(w) > 0$

Suppose that our basic condition in Section 4.1 holds.

Suppose that the sequence $\{(z_n, t_{n-1})\}$ has been derived using the equilibration algorithm above (in Section 5.1) so that $z_{n+1} = z_n + u_n \Delta_\varepsilon(z_n)$ for all $n = 1, 2, \ldots$.

Also let $V(w) > 0$ and suppose that w is a limit point of $\{z_n\}$.

Initially we assume that w lies in the interior of F.

In this appendix we show that, given such a sequence $\{z_n\}$ and such a w, there is a suffix k such that $V(z_k) < V(w)$. Now $V(z_n)$ is strictly decreasing and so it further follows that

$$V(z_n) < V(z_k) < V(w)$$

if $n > k$ and thus $V(w)$ cannot be a limit point of the sequence $\{V(z_n)\}$. However V is continuous and it therefore also follows that the w cannot be a limit point of the sequence $\{z_n\}$. This is the contradiction required in Section 6.

To show that this contradiction does in fact arise from our assumptions above, suppose now that Δ_ε and G are as in Sections 3.2 and 4.2 above.

Let $r > 0$ and $B(w, r)$ be the *closed* ball of radius $r > 0$ centered at $w \in intF$. Suppose that r is so small that $B(w, r)$ lies entirely inside $intF$.

Since V is continuous and $V(w) > 0, V(z) > 0$ for z sufficiently close to w and so there is a possibly even smaller $r > 0$ such that

$$V(z) > 0 \text{ if } z \in B(w, r) \subset intF.$$

The function G is positive and continuous on the compact set $B(w, r)$ and so assumes its least value $g = g(w, r) > 0$ at some point of $B(w, r)$. Therefore for such $r > 0$:

$$\nabla V(z) \cdot \Delta_\varepsilon(z) \leq -G(z) \leq -g < 0 \text{ for all } z \in B(w, r).$$

Consider now

$$\nabla(V(z + t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z)$$

for all (z, t) in $F \times [0, \eta]$ where $\eta = \eta(w, r) > 0$ is chosen so that $z + t\Delta_\varepsilon(z) \in intF$ if $z \in B(w, r)$ and $0 \leq t \leq \eta$. This latter product set allows, via the second co-ordinate space $[0, \eta]$, for all points on a closed partial ray $[z, z + \eta\Delta_\varepsilon(z)]$ emanating from z in direction $\Delta_\varepsilon(z)$.

We already know from above that

$$\nabla V(z + 0\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z) = \nabla V(z) \cdot \Delta_\varepsilon(z) \leq -G(z) \leq -g$$

for all $z \in B(w,r)$. Also $\nabla V(z+t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z)$ is continuous on the compact set $B(w,r) \times [0,\eta]$ and so $\nabla V(z+t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z)$ is uniformly continuous on $B(w,r) \times [0,\eta]$. Hence there is a real number h such that $0 < h \leq \eta$ and

$$|\nabla V(z+t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z) - \nabla V(x+\tau\Delta_\varepsilon(x)) \cdot \Delta_\varepsilon(x)| < \tfrac{1}{4}g$$

if (z,t) and (x,τ) are two points in $B(w,r) \times [0,\eta]$ less than or equal to a distance h apart.

Now consider, in the above inequality, letting $x = z \in B(w,r)$ and $\tau = 0$. We deduce that, for $z \in B(w,r)$ and $0 \leq t \leq h \leq \eta$:

$$|\nabla V(z+t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z) - \nabla V(z) \cdot \Delta_\varepsilon(z)| < \tfrac{1}{4}g.$$

Hence, for an arbitrary $z \in B(w,r)$ and for all $0 \leq t \leq h$,

$$\begin{aligned}\nabla V(z+t\Delta_\varepsilon(z)) \cdot \Delta_\varepsilon(z) &< \nabla V(z) \cdot \Delta_\varepsilon(z) + \frac{1}{4}g \\ &\leq -G(z) + \frac{1}{4}g \\ &\leq -G(z) + \frac{1}{4}G(z) \\ &= -\frac{3}{4}G(z) \leq -\frac{3}{4}g.\end{aligned}$$

Let $z \in B(w,r)$ and let $0 \leq t \leq h$. Then, for this $z \in B(w,r)$ and t, integrating the above inequality from 0 to t:

$$V(z+t\Delta_\varepsilon(z)) - V(z) \leq -\tfrac{3}{4}G(z)t \leq -\tfrac{3}{4}gt.$$

Let $t = u_n \leq h$ in the above inequality. Then by the statement of the algorithm,

$$z_{n+1} = z_n + u_n\Delta_\varepsilon(z_n) \text{ and } t_n = 2u_n.$$

Consider z_{n+1} and u_{n+1}. If $z_{n+1} \in B(w,r)$ still then integrating as before:

$$V(z_{n+1} + \tfrac{1}{2}t_n\Delta_\epsilon(z)) - V(z) \leq -\tfrac{3}{4}G(z_{n+1})(\tfrac{1}{2}t_n) \leq -\tfrac{1}{8}G(z_{n+1})(\tfrac{1}{2}t_n)$$

as $\frac{1}{2}t_n = u_n \leq h$. It follows that t_n is halved no more than once to obtain u_{n+1} and so

$$u_{n+1} = t_n = 2u_n \text{ or } u_{n+1} = \tfrac{1}{2}t_n = u_n.$$

Hence

$$u_{n+1} \geq u_n \text{ if } u_n \leq h.$$

It follows that for any given $(z_i, u_i) \in B(w,r)$ there is a constant $\alpha = \alpha(u_i) > 0$ such that if

$$j > i \text{ and } z_i, z_{i+1}, z_{i+2}, \ldots, z_j \in B(w,r)$$

then

$$u_n \geq \alpha(u_i) \text{ for } n = i, i+1, i+2, \ldots, j.$$

Now let $i \leq n \leq j$. Then

$$V(z_n + u_n \Delta(z_n)) - V(z_n) \leq -\tfrac{1}{8} g u_n \leq -\tfrac{1}{8} g \alpha(u_i)$$

Hence

$$V(z_{n+1}) - V(z_n) \leq -\tfrac{1}{8} g \alpha(u_i)$$

and so adding over $n = i, i+1, i+2, \ldots, j$,

$$V(z_{j+1}) - V(z_i) \leq -(j-i)\tfrac{1}{8} g \alpha(u_i).$$

Since $V(z_{j+1}) \geq 0$, it follows that

$$j - i \leq \tfrac{V(z_i)}{\frac{1}{8} g \alpha(u_i)}.$$

Hence:

$$j \leq i + \frac{V(z_i)}{\frac{1}{8} g \alpha(u_i)}.$$

So the sequence $\{z_n\}$ certainly exits the ball $B(w, r)$ whenever it enters this ball.

Now w is not an equilibrium (so $V(w) > 0$) and is also a limit point of the sequence $\{z_n\}$. Hence, for any real number a satisfying $0 < a < 1$, it now follows that there are two natural numbers i and j, depending on a, such that $1 < i < j$,

$$\begin{aligned} &z_i \in B(w, ar), \\ &z_{i+1} \in B(w,r), z_{i+2} \in B(w,r), z_{i+3} \in B(w,r), \ldots, z_j \in B(w,r), \\ &z_{j+1} \notin B(w,r). \end{aligned}$$

Now let $M = M(w,r) = sup\{||\Delta_\varepsilon(x)||; z \in B(w,r)\}$. Then, by the definition of i and j:

$$\begin{aligned} (1-a)r &\leq ||z_{j+1} - z_i|| \\ &\leq ||\sum_i^j u_n \Delta(z_i)|| \\ &\leq M[u_i + u_{i+1} + \ldots + u_j]. \end{aligned} \tag{11}$$

By the algorithm statement,

$$V(z_n + u_n \Delta(z_n)) - V(z_n) < -\tfrac{1}{8} u_n G(z_n)$$

for all n and so it follows that

$$
\begin{aligned}
V(z_{j+1}) - V(z_j) &< -\frac{1}{8}gu_j,\\
V(z_j) - V(z_{j-1}) &< -\frac{1}{8}gu_{j-1},\\
&\vdots\\
V(z_{i+1}) - V(z_i) &< -\frac{1}{8}gu_i.
\end{aligned}
$$

Adding these:

$$
\begin{aligned}
V(z_{j+1}) - V(z_i) &= \sum_i^j \{V(z_{n+1}) - V(z_n)\}\\
&< \sum_i^j \{-\frac{1}{8}gu_n\}\\
&= -\frac{1}{8}g\sum_i^j \{u_n\}\\
&= -\frac{1}{8}g[u_i + u_{i+1} + \cdots + u_j]\\
&\leq -\frac{1}{8}g(1-a)r/M
\end{aligned}
$$

by (11).

Therefore:

$$
\begin{aligned}
V(z_{j+1}) - V(w) &= V(z_{j+1}) - V(z_i) + V(z_i) - V(w)\\
&\leq -\frac{1}{8}g(1-a)r/M + Mar\\
&= -\frac{1}{8}gr/M + [\frac{1}{8}gr/M + Mr]a\\
&< 0
\end{aligned}
$$

if a is sufficiently small.

We have now shown as desired that there is a suffix $k = j + 1$ such that $V(z_k) < V(w)$. Now $V(z_n)$ is strictly decreasing and so it further follows that

$$V(z_n) < V(z_k) < V(w)$$

if $n > k$ and thus $V(w)$ cannot be a limit point of the sequence $\{V(z_n)\}$. However, as we remarked above, V is continuous and it therefore also follows that the w cannot be a limit point of the sequence $\{z_n\}$. This is the contradiction required in Section 6.

A very similar argument applies if we suppose that w is a non-equilibrium limit point on the boundary of F. There are no major changes as the direction $\Delta_\varepsilon(z)$ is feasible even when z lies on the boundary of F. One change to the above proof is to replace $B(w, r)$ by $F \cap B(w, r)$.

Minimal Revenue Network Tolling: System Optimisation under Stochastic Assignment.

Kathryn Stewart[1] and Mike Maher[2]

[1] Transport Research Institute and School of the Built Environment, Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, Scotland, k.stewart@napier.ac.uk

[2] Transport Research Institute and School of the Built Environment, Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, Scotland, m.maher@napier.ac.uk

Summary. The classical road tolling problem is to toll network links such that, under the principles of Wardropian User Equilibrium (UE) assignment, a System Optimising (SO) flow pattern is obtained. Such toll sets are however non-unique, and further optimisation is possible: for example, *minimal revenue* tolls create the desired SO flow pattern at minimal additional cost to the users. In the case of deterministic assignment, the minimal revenue toll problem is capable of solution by various methods, such as linear programming [BHR97] and heuristically by reduction to a multi-commodity max-flow problem [Dia00]. However, it is generally accepted that deterministic models are less realistic than stochastic, and thus it is of interest to investigate the principles of tolling under stochastic modelling conditions. This paper develops methodologies to examine the minimal revenue toll problem in the case of Stochastic User Equilibrium. Tolling solutions for both 'true' System Optimum and Stochastic System Optimum under SUE are derived, using both logit and probit assignment methods.

Key words: Traffic assignment; Stochastic user equilibrium; Probit model; Logit model; Optimal tolls; Marginal social costs.

1 Introduction

1.1 General Background to Road User Charging

Road Tolling is a commonly used term, but can be used to describe different situations. For example there are many instances of 'toll roads' particularly in continental Europe, whereby a charge is made for travel along usually a section of high quality trunk road. Similarly a charge may be made to use a short length of road, primarily in the case of a tunnel or a bridge as is common in the UK. Such charges are usually either fixed or related to distance travelled,

Mathematical and Computational Models for Congestion Charging, pp. 201-218

and payment is made at a toll-booth at one end of the charged section, either electronically, or by actual payment at the booth.

Congestion charging by means of implementing road user tolls, has been much discussed, but has been implemented in relatively few cities. Toll rings exist and are operational in Oslo and Bergen in Norway, and area-charging schemes exist in Singapore and now in London. These operational road user charging schemes have used a cordon system, which has the benefit of being transparent and easy to implement, and acts to discourage drivers from entering the controlled area, but once the driver is within the cordon, there is no additional incentive to choose a route that would be beneficial to the system as a whole. Intuitively it would seem logical that if road tolls are to be implemented, they should in some way be optimal; that is they should be as effective as possible with regard to specified criteria. It may be a political objective to maximise revenue, within limits of political acceptability, whilst not seeking particularly to discourage road users or to lessen congestion, which would lead to relatively cheap tolls. If instead congestion reduction were the primary objective, tolls would be set very high to discourage usage, an extreme case of which would be to completely restrict traffic and impose high fines for violation. If optimality is desired however, suitable criteria must first be defined. Theoretically this is often considered by fixing network demand, and then considering how that traffic may be assigned throughout the network such that the overall network cost is minimised.

Whilst operational schemes are cordon based, trials have however been carried out in which tolling schemes have been tested with road pricing measures such as: distance travelled, time spent travelling and congestion caused (Cambridge study [MM00], [Iso98]), which demonstrate that the technology to implement a path or link based tolling system for urban areas does exist, and so such schemes may be feasible for actual implementation in the future. There is also current political interest in the UK regarding more developed tolling schemes: The Commission for Integrated Transport recently published a report 'Paying for Road Use' (CfIT, 2002), which suggests the introduction of nationwide road user charging. The report is of particular interest in that it suggests the use of marginal social cost pricing on all roads (i.e. motorways, A Roads, minor roads, city centres etc), balanced by a reduction (or abolition), of Vehicle Excise Duty, combined with a reduction in fuel duty, so that the result desired would be fiscal neutrality. Such a scheme would rely on charging for travel along a link, rather than passing across a cordon, and would therefore require similar technology to implement as would be required for a minimal revenue toll scheme.

1.2 Modelling the Effect of Road Tolls

Traffic assignment models seek to replicate the traffic pattern that is created when drivers choose their routes across a network from their origin to their destination.

In the case of deterministic assignment, it is assumed that drivers, with perfect network knowledge, will act selfishly to minimise their personal travel cost, resulting in the Wardropian User Equilibrium (UE) flow pattern. This occurs when the objective function (1) is minimised, (link flows $\boldsymbol{x}$ and link costs $\boldsymbol{c}$).

$$z_{UE}(\mathrm{x}) = \sum_a \int_0^{x_a} c_a(\omega) d\omega \tag{1}$$

Tolls may then be imposed to 'force' a UE assignment to result in an alternative desired flow pattern. The Social (or System) Optimum (SO) is one such desired flow pattern, where the Total Network Travel Cost (TNTC) is minimised (2), and occurs when all used routes between any OD pair have equal marginal cost.

$$z_{SO}(\mathrm{x}) = \sum_a x_a c_a(x_a) \tag{2}$$

The flow patterns that minimise the functions in (1) and (2) satisfy Wardrop's first and second equilibrium principles, respectively, and they are hereafter referred to as the UE and SO solutions. In the context of tolling the flow pattern that minimises the function (3) below satisfies Wardrop's first principle in the presence of tolls or the tolled user-equilibrium principle.

$$z_{TUE}(\mathrm{x}) = \sum_a \int_0^{x_a} [c_a(\omega) + T_a] d\omega \tag{3}$$

When $T_a = c^{'}_a(x^*_a)x^*_a$, where x^* is the SO solution and $c^{'}_a(x^*_a)$ denotes the first derivative of $c_a$at x^* , it is well known that the SO solution also minimises (3). In the literature, $c^{'}_a(x^*_a)x^*_a$ is referred to as the Marginal Social Cost Price (MSCP) toll. Such toll sets are however not unique and other toll sets exist which also minimise (3), e.g. [HR98] formulate the problem (Minimum Revenue or MinRev) of finding tolls as a linear program with the objective of minimising the revenue collected. When tolls are allowed to be negative they may be considered to be usage subsidies; in this case it is of interest to require that the toll revenue collected should equal the usage subsidies given out i.e. where fiscal neutrality is achieved.

The Minimal Revenue toll problem has, in the case of deterministic assignment, been solved such that the System Optimal solution is obtained, by various methods: for example, Linear Programming [BHR97], reduction to a multi-commodity max-flow problem [Dia00] and simplex method via CPLEX [HR98].

Route spreading, which is an observed phenomenon in traffic assignment, can be modelled by applying cost-flow relations to simulate congestion as in the case of deterministic assignment. Stochastic assignment methods however assume that instead of drivers having a 'perfect' knowledge of the varying OD costs of a network, they have a variable perception of these costs. Stochastic user equilibrium (SUE) assignment is based on the premise that each driver will act to minimise their perceived route cost, which follows a distribution such as those given in the logit or probit models. Traditionally deterministic assignment has been used to model congested urban networks. If the same methods are applied though to un-congested inter-urban networks they tend to result in an All-or-Nothing type solution that is unrealistic in practice. Stochastic methods may be used to successfully model inter-urban networks, but it is desirable to have a single method that will be capable of modelling both extremes (and the middle ground). Thus Stochastic User Equilibrium (SUE) methods have been developed [MH97a, MH97b]. It would seem logical that drivers do perceive costs differently from each other, either because of different levels of network knowledge or different priorities (e.g. avoidance of right turns or roundabouts, minimising distance or time), and so the use of a stochastic method would seem to be more realistic and thus it is useful to extend the concept of tolling to the stochastic case.

This paper therefore develops methodologies to examine the minimal revenue toll problem in the case of Stochastic User Equilibrium. A discussion of stochastic assignment methods is given in section 2.

In examining the case of Stochastic User Equilibrium the 'desired flow pattern' to be created must first be determined. The classical economics solution of replacing cost flow functions with marginal cost flow functions, does not generally result in the total network cost being minimised in the stochastic case [Yan99]. Thus tolls which are analogous to Marginal Social Cost Pricing (MSCP) in the deterministic case do not give the Deterministic System Optimal flow solution.

If the 'true' system optimal flow pattern is desired, it may be possible to derive tolls that are unrelated to MSCP. It is not obvious if such tolls exist or under which conditions they may exist and, if they *are* found to exist, if they are unique. If toll sets exist which are not unique, then as in the case of UE, it would be possible to impose additional constraints, and to search for (for example) minimal revenue tolls. Tolling methodologies to approach the SO solution under SUE are developed in section 3.

However, it may be more desirable in the stochastic case to produce instead a 'Stochastic System Optimum' (SSO) where the *perceived* total network cost is minimised, i.e. the SSO solution is that flow pattern which minimises the total of the travel costs perceived by drivers. This SSO solution may also be characterised as that which maximises consumer surplus [Yan99]. Tolling to achieve the SSO solution is the subject of section 4.

2 Stochastic Assignment Models

Stochastic methods are based on the assumption that a driver minimises their perceived cost, or chooses the alternative that gives the highest utility. Utility functions U_k may be expressed as the sum of a deterministic component V_k and a random error component ξ_k, where k is a member of the set of alternatives. That is,

$$U_k = V_k + \xi_k \quad \forall k \tag{4}$$

The probability that an alternative is chosen is the same as the probability that that alternative has highest utility in the choice set. Whilst not being entirely exhaustive [She85], the most commonly used stochastic method models assume either a Normal distribution (probit models), or the Gumbel distribution (logit models), for the drivers' perception error ξ_k .

The logit model is based on the use of the logistic function, which is a choice function used to choose between two or many alternatives.

It may be written:

$$p_i = \frac{\exp(-\theta.C_i)}{\sum\limits_j \exp(-\theta.C_j)} \tag{5}$$

where p_i is the probability of choosing alternative i, C_i is the cost associated with route i and θ is a dispersion parameter; the lower the value of θ, the higher the level of uncertainty, conversely a high value of θ would correspond to drivers having an accurate view of actual route costs, i.e. the deterministic case.

The logit formulation has the advantage of mathematical tractability, and has been used initially for that reason, but logit based loadings have a significant disadvantage in that they do not account for overlapping paths in a satisfactory manner. For example three completely distinct paths would have flow assigned in the same way as a single path together with two paths including a significant overlap. If each path had around equal cost, then each path would be assigned around one third of the traffic irrespective of any overlap. In addition the logit method assigns traffic based on an absolute difference in cost (time), for example a five minute difference in journey time will produce the same route choice proportions whether the difference relates to route times of 5 and 10 mins or route times of 200 and 205 mins. In the first case one route takes twice as long as another, whilst in the second, the five-minute difference may well not be perceived as 'any difference at all'. It would seem reasonable to require a model to account for the difference in journey time in relation to the total journey time when assigning traffic.

The probit model assumes that the random error term is normally distributed, and that the joint density function of the errors ξ_k, is Multivariate Normal (MVN). Thus the probability distribution of cost for each link is Normal, with mean μ being the value of the link cost flow relation, and variance σ^2 assumed to be proportional to the mean.

$$\beta = Variance/Mean = \sigma^2\mu \Rightarrow \sigma^2 = \mu\beta \tag{6}$$

$$c_a \sim N(c_a(x_a), \beta c_a(x_a)) \quad \forall a \tag{7}$$

At link level costs are often assumed to be independent, but in general it may be assumed that link costs also follow a Multivariate Normal distribution. The probit model solves the problem of overlapping paths by the use of correlations between the path cost perception errors.

There are various methods of solution for probit-SUE, such as:

1. Numerical integration of a multiple integral. Feasible numerical integration approaches now exist [RM02] which can be used for networks with up to around twenty alternative routes.
2. The Stochastic Assignment Method SAM [MH97a, MH97b], a heuristic based on 'Clarke's Method' [Cla61], where a successive approximation method is used; the maximum of two normally distributed random variables being approximated by another Normal variable.
3. Monte Carlo simulation, whereby a random value representing the perceived travel time of a link is sampled from the density function for that link, and an All-or-Nothing assignment is carried out based on the set of sampled perceived travel times across all network links. The process of sampling and assignment is repeated (multiple times) and averaged to give the final flow pattern.

The methodology developed within this paper however does not depend on any particular stochastic assignment method being used.

3 System Optimal Road Tolls

3.1 Path-based methodology

If it is desired that an SUE assignment using the original cost-flow functions with the addition of a toll, should produce the SO flow pattern that is obtained under deterministic assignment, where the TNTC is minimised, then using logit-based SUE this may be formulated as below:

$$X_i = D\frac{\exp -\theta(C_i + T_i)}{\sum\limits_j \exp -\theta(C_j + T_j)} \tag{8}$$

where D is the OD demand, C_i and X_i are the path costs and flows at the SO solution which may be found using deterministic assignment methods, and T_i are the desired path tolls to be determined.

The 'toll difference' between pairs of path tolls for each OD pair may then be found by the division of pairs of equations, thus:

$$T_j - T_i = \frac{1}{\theta}\ln(X_i/X_j) + (C_i - C_j) \tag{9}$$

A resulting order of magnitude of path tolls (for each OD pair) may be deduced, and assuming tolls to be non-negative, and seeking minimal revenue tolls, the smallest toll path (for each OD pair) may be set as zero, and the remaining path tolls calculated.

It would appear that the use of the logistic function to determine path differences requires all path flows to be non-zero, as zero flows would clearly result in infinite tolls (9). (It should also be noted here that this problem would also apply in the case of the probit model). There are consequently difficulties to be encountered when dealing with more complex networks, Smith et al. [SEL94] where there will generally exist technically feasible paths which have zero flow at SO.

It is however possible to divide the set of feasible paths into two sets, $\Omega^0$for zero-flow paths and Ω^1 for non-zero flow paths as defined below;

Let $\Omega^0 = \{k : X_k = 0\}$ and $\Omega^1 = \{i : X_i > 0\}$

Then $\forall\, i, j \in \Omega^1$, let T_i and T_j satisfy (9) and $\forall\, k \in \Omega^0$, let $T_k = M$
Then, the path flows, $X_i(T)$, associated with the above tolls are:

$$X_i(T) = D\frac{\exp -\theta(C_i + T_i)}{\sum\limits_{i\in\Omega^1} \exp -\theta(C_i + T_i) + \sum\limits_{i\in\Omega^0} \exp -\theta(C_i + M)}, \quad \forall i \in \Omega^1 \qquad (10)$$

$$X_k(T) = D\frac{\exp -\theta(C_k + M)}{\sum\limits_{i\in\Omega^1} \exp -\theta(C_i + T_i) + \sum\limits_{i\in\Omega^0} \exp -\theta(C_i + M)}, \quad \forall k \in \Omega^0 \qquad (11)$$

It is clear that $X_i(T) \to X_i$ and $X_k(T) \to 0$ as $M \to \infty$.

Hence for any $\varepsilon > 0$, there exists a sufficiently large M such that $X_i - X_i(T) \leq \varepsilon$, $\forall\, i \in \Omega^1$, and $X_k(T) \leq \varepsilon$, $\forall\, k \in \Omega^0$.

Thus it is possible to determine viable path toll sets, which will create a flow pattern approaching the true SO flow pattern, as closely as is desired under logit SUE. However in the limiting case ($\varepsilon \to 0$), M (the toll on zero-flow paths) will tend to infinity, and so an appropriate degree of closeness to the SO solution would need to be determined.

This is illustrated using the 9-node network with 2 origins and 2 destinations as shown in Figure 1 below. This network has been frequently used in the literature [BHR97] and [Dia00]; a modified version is used here (with 4 vertical links (5↔6, 7↔8) carrying zero flow removed), to render the network acyclic, and thus limit the path enumeration matrix so that 24 viable paths are obtained (six paths between each of the four OD pairs).

The link cost functions are of BPR type as shown where ($c_a^{(0)}$,Y_a) for each link are given on the diagram, x_a is link flow, c_a is link cost, $c_a^{(0)}$ is free flow link cost and Y_a is link capacity.

For the above network, the minimum TNTC = 2253.92. All links have non-zero flow at the Wardropian SO solution but this is not true of all paths: in a Wardropian SO assignment using 100 iterations of the Method of Successive Averages [She85], 3 paths were completely unused. Assigning a toll M to the

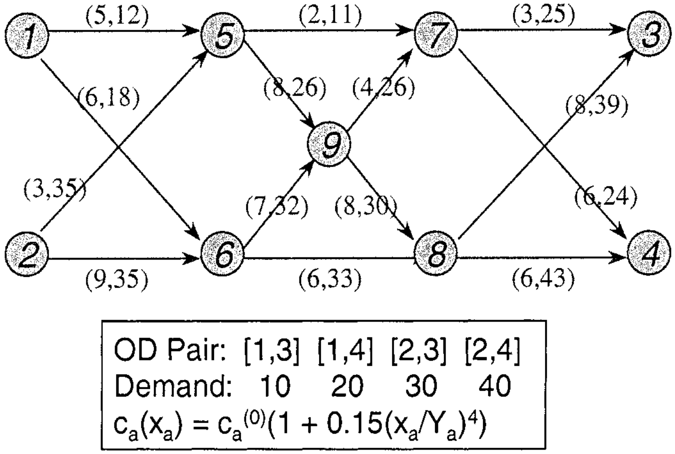


Fig. 1. Nine-node network diagram, showing OD demand and link cost-flow relations

zero-flow paths, a viable toll set is given below for $\theta = 0.1$. When M=50, TNTC = 2253.99.

OD pair	[1,3]	[1,4]	[2,3]	[2,4]
T=	12 9 M 0 M 2	15 16 18 16 12 0	11 1 45 5 M 0	16 11 7 28 19 0

The path tolls correspond to the paths given in the path-link incidence matrix A (Appendix1), with 6 paths for each OD pair.

A difficulty with this method is that although viable path toll sets can be determined, it is not necessarily possible to derive consistent link-based toll sets. This inconsistency may be demonstrated by combining sets of paths as is given in Figure 2 below.

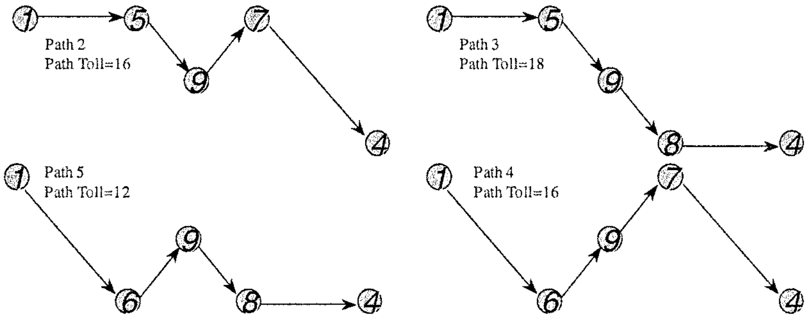


Fig. 2. Path combinations from Bergendorff's nine-node network

Considering OD pair [1,4], paths 2 and 5 together contain the same links as paths 3 and 4 together; therefore for consistent link tolls the total toll on paths

2 and 5 (total=28) should equal the total toll on paths 3 and 4 (total=34). This is clearly not the case (similar examples of inconsistency may also be demonstrated); so consistent link tolls may not be determined. Also whilst it is not problematic to assign large tolls (tending to infinity) on paths with zero flow, it would clearly not be desirable to assign such tolls to links.

Thus it must be considered whether a path-based tolling methodology would be sensible for implementation, even if developing efficient path-based assignment methods [Ros01], could be utilised. A tolling solution where the same link has a different cost depending on the overall route travelled would intuitively not appear to be equitable or practicable. Also the technology required to implement a path-based tolling scheme, would require vehicle tracking (at least in an urban context), which although technically feasible would result in concern over privacy issues.

Further, the path difference equations used require the use of logit assignment, which does not model networks with overlapping paths as well as probit. If a network with only two links is used, it is possible to solve for tolls algebraically in the probit case, but this is not the case for any more complex network. Consequently this method is not considered to be of potential use for practical implementation.

3.2 Link-Based Methodology

A link-based methodology to derive tolls that would create a flow pattern approaching the SO is therefore desirable. It was assumed from the previous results, that link-based tolls might not be sufficient to replicate the desired SO flow pattern in the limiting case, but that good sub-optimality would be acceptable for practical purposes.

The objective is still to minimise the total network travel cost, and this is attempted by seeking a link flow pattern that approaches the flows obtained under deterministic SO assignment. Thus links where the flow is higher than that desired have link costs progressively increased by the addition of a toll until the desired flow pattern is approached, as in the heuristic procedure given below:

Step 1: Find the SO solution and let $\mathbf{F}_{SO}$, $\mathbf{C}_{SO}$ and TNTC$_{SO}$ denote the corresponding flow pattern, link cost, and total network travel cost.
Step 2: Link toll vector set to zero: $\mathbf{T}_0 = \mathbf{0}$
Step 3: Set $n = 0$
Step 4: Perform SUE assignment: $\mathbf{C}_n$ and $\mathbf{F}_n$ obtained
Step 5: Calculate: $\mathbf{P} = (\mathbf{F}_n - \mathbf{F}_{SO})(|\mathbf{C}_n - \mathbf{C}_{SO}|)$
Step 6: Determine link j where $P(j)$ is greatest.
Step 7: [3] Perform iteration to calculate $t(j)$ s.t $F(j) = F_{SO}(j)$ to required degree of accuracy.

[3] The internal iteration in step 7 only regards the output flow for the single link where$P(j)$ is greatest as per step 6, and results in the link tolls shown in Table 1.

7a: Set $t(j_0) = \left|C_{j_0} - C_{SO_j}\right|$ where $C_{j_0}$is the current cost on link j (as per step 4)
7b: Set m=1
7c: Perform SUE assignment, calculate $\left|C_{j_m} - C_{SO_j}\right|$
7d: Set $t(j_m) = t(j_{m-1}) + \left|C_{j_m} - C_{SO_j}\right|$
7e: Calculate $P(j_m)$: Stop if sufficiently close to zero and let $t(j_m) = t(j)$, or set
$m = m + 1$ and repeat from step 7c.
Step 8: $\mathbf{T}_{n+1} = \mathbf{T}_n + \mathbf{t}$; where $t(i) = t(j)$ when $i = j$ and $t(i) = 0$ otherwise
Step 9: Calculate TNTC: Stop if TNTC sufficiently close to TNTC$_{SO}$ or set $n = n + 1$ and repeat from Step 4.

This method is illustrated using the previously used 9-node network (see Figure 1), with logit SUE where $\theta = 0.1$.

In Step 1 a deterministic SO assignment determines the desired link flow set, and the link costs are calculated for these flows. The minimum value of the TNTC is also recorded (for this example TNTC$_{SO}$ = 2253.9). An initial toll vector is then set with all tolls being zero (Step 2). An SUE assignment is then completed, and the link costs, link flows and the TNTC compared with those desired.

The toll set is constructed in a step-wise process, where only a single link toll is considered in each iteration; thus the link to be tolled in that iteration must be chosen. Steps 4 and 5 determine which link is chosen: choosing simply the link where the flow was most in excess of the desired SO flow for that link would not take into account the relative costs, and so a product of flow and cost difference is used here, although this may be refined in future work. As only non-negative tolls are being imposed, the absolute value of the cost difference is used, so that the chosen link, where the value of the product is greatest has a flow strictly greater than that desired.

Table 1 below shows the stepwise construction of a toll set.

It can be seen from Table 1 and from the graph in Figure 3 below, that the first few iterations are by far the most significant, and no great benefit is gained from continuing to approach the TNTC$_{SO}$ for many iterations. Further if it is desirable to keep as many links toll free as possible, it is not then sensible to continue to add small tolls on additional links, to reduce the TNTC only by tiny amounts.

The link toll set resulting from the 12 iterations given above, is shown in Figure 4 below, where link width is proportional to the size of the link toll. The TNTC achieved after 12 iterations is only 0.02% greater than the Minimum TNTC. However if the process was stopped after only 4 iterations, the TNTC achieved is still only 0.6% greater than TNTC$_{SO}$ and 4 links that could be tolled, would remain toll-free.

A more efficient interpolation procedure is being refined for the internal iteration for use in larger networks.

Table 1. Iterative building of 'Optimising' toll set

Iteration	0	1	2	3	4	5	6	7	8	9	10	11	12
t_1 (1-5)	-	-	-	-	-	-	-	-	-	-	**0.9**	0.9	0.9
t_2(5-7)	-	**7.2**	7.2	7.2	7.2	7.2	7.2	**8**	8	**8.9**	8.9	8.9	8.9
t_3(7-3)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_4 (1-6)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_5(2-5)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_6 (5-9)	-	-	-	-	-	-	-	-	**0.6**	0.6	0.6	0.6	0.6
t_7 (9-7)	-	-	-	-	-	-	**1.4**	1.4	1.4	1.4	1.4	1.4	1.4
t_8 (6-9)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_9 (9-8)	-	-	-	**13**	13	13	13	13	13	13	13	**13.8**	13.8
t_{10} (7-4)	-	-	**7.9**	7.9	**12.9**	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9
t_{11}(8-3)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_{12}(2-6)	-	-	-	-	-	**3.6**	3.6	3.6	3.6	3.6	3.6	3.6	**3.8**
t_{13}(6-8)	-	-	-	-	-	-	-	-	-	-	-	-	-
t_{14}(8-4)	-	-	-	-	-	-	-	-	-	-	-	-	-
TNTC	2441	2385	2337	2285	2268	2262	2259	2258	2257	2256	2255	2255	2254
REV	0	154	307	449	568	705	746	759	777	797	813	819	822

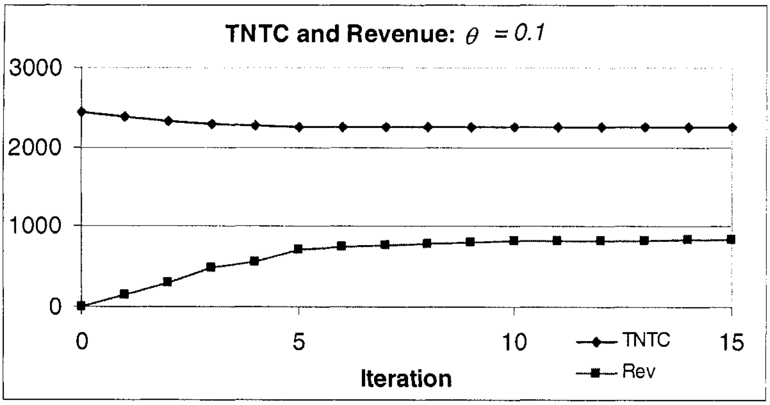
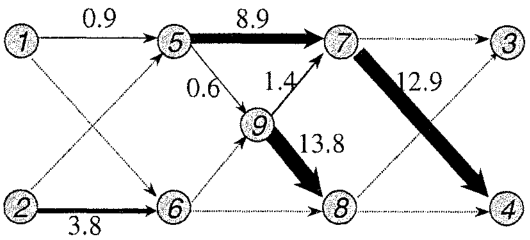


Fig. 3. Total Toll Revenue required for reduction in TNTC



$TNTC_{SO}$	= 2253.9
TNTC	= 2254.4

Fig. 4. logit toll-set for Bergendorff's network ($\theta = 0.1$) – 12 iterations

Whilst this methodology has been demonstrated using logit assignment, it may equally be used for other stochastic assignment models. Figure 5 below, shows the reduction in TNTC achieved for different values of the dispersion parameter θ using logit assignment, and for the variability parameter $\beta = 0.5$ using probit assignment. (The Stochastic Assignment Method SAM [MH97a, MH97b] was used here to obtain the probit results).

It must be noted that the method used does not result in the TNTC strictly decreasing at every iteration, although the overall trend is that it does reduce as the desired flow pattern is approached. The internal iteration at Step 7, has in these examples been used to reduce the flow on a particular link so that it is very close to the desired flow value for that link at SO. During this internal iteration process, at some point the value of the difference product P will be greatest for a new link, after this point, the overall TNTC may no longer decrease. It is possible to amend this internal iteration, so that the link toll is determined at the minimal value for the TNTC that can be achieved by just varying the toll on this link. However it appears in practice that as this will generally give a smaller toll being added at each iteration, that it causes a greater number of the main iterations to be required. Consequently, the objective at each internal iteration is that the flow difference on that link should be reduced to (approximately) zero.

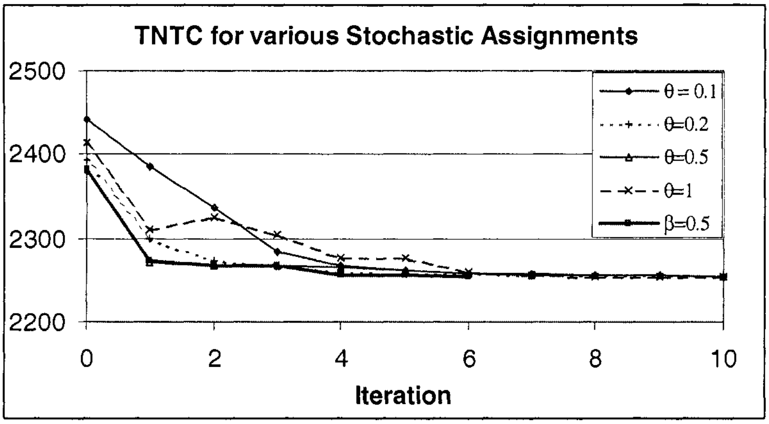


Fig. 5. TNTC with increasing iterations for various Stochastic Assignments

It of interest to note that if the logit and probit models are to be compared using the relation:

$$Var(U_k) = \pi^2/(6\theta^2) \qquad (12)$$

Cascetta [Cas90], where U_k is the probit utility function as in equation (1), then despite the link variances in the probit case obviously being different for each link, an approximate correspondence can be found in this case between probit $\beta = 0.5$, and a logit sensitivity parameter of $\theta \approx 0.5$. It can be observed in Figure 5 above, that the graphs for $\beta = 0.5$ and $\theta = 0.5$ predominantly coincide.

4 Stochastic Social Optimum Road Tolls

In the case of stochastic user equilibrium, it could be argued that it is not the 'actual' or deterministic total travel cost that should be minimised, but rather the perceived total network travel cost.

In the case of deterministic assignment, it is well known that that the Total Network Travel Cost is minimised and the System Optimal flow pattern is obtained, when cost-flow functions are replaced by marginal cost-flow functions. Recent work by Maher et al. [MSR05] has shown that the analogous case is true under stochastic assignment. Thus MSCP tolls may be easily found using existing link-based assignment methods.

The minimal revenue toll problem is thus similar to that in the deterministic case, and may be solved by linear programming. For comparative purposes a numerical example is included below.

4.1 An illustrative example

As in the deterministic case equally optimal toll sets exist for this network, and so further optimisation is possible. It is of interest to obtain as many links with zero tolls as possible, but even with this provision, in this example, there were four equally-optimal toll sets for each value of θ. A possible Min-Rev toll-set is given below in Table 2 for various θ. The links corresponding to the zero-flow paths are highlighted. Other zero-toll links may be observed, although it must be remembered that there are other equally optimal solutions that are not shown. Despite the existence of three distinct zero-toll trees for varying values of the sensitivity parameter, the change in individual link-toll values as θ varies appears to be reasonably smooth.

The zero-toll trees are shown in Figure 6 below; the zero-toll links being represented by the bold print arrows. As θ increases the driver's assumed perceived knowledge of network costs increases, so that as θ tends to infinity, the logit stochastic assignment tends towards a deterministic assignment, and the final zero-toll tree (θ=5) is indeed the same as that obtained by deterministic methods.

5 Summary

In attempting to approach the 'true' SO flow pattern through tolling, the algebraic logit formulation derived path-tolls that could not then be separated

Table 2. Minimal revenue toll sets as θ varies

Toll	θ						Deterministic
	0.01	0.1	0.2	0.5	1	5	
t1 (1-5)	12.8	2.9	1.0	0.2	0.1	0.0	0.0
t2 (5-7)	0.0	5.4	6.9	7.7	7.9	8.0	8
t3 (7-3)	0.1	2.6	3.7	5.1	5.9	6.9	7.2
t4 (1-6)	0.0	0.0	0.0	0.0	0.0	0.0	0
t5 (2-5)	6.0	2.4	2.2	2.9	3.4	3.9	4
t6 (5-9)	5.3	0.0	0.0	0.0	0.0	0.0	0.0
t7 (9-7)	0.0	0.0	0.0	0.0	0.0	0.0	0
t8 (6-9)	9.1	0.7	0.0	0.0	0.0	0.0	0.0
t9 (9-8)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
t10 (7-4)	6.4	3.9	3.3	3.1	3.1	3.2	3.2
t11 (8-3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
t12 (2-6)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
t13 (6-8)	0.0	0.0	1.9	4.4	5.6	6.8	7.2
t14 (8-4)	0.0	0.0	0.0	0.0	0.0	0.0	0

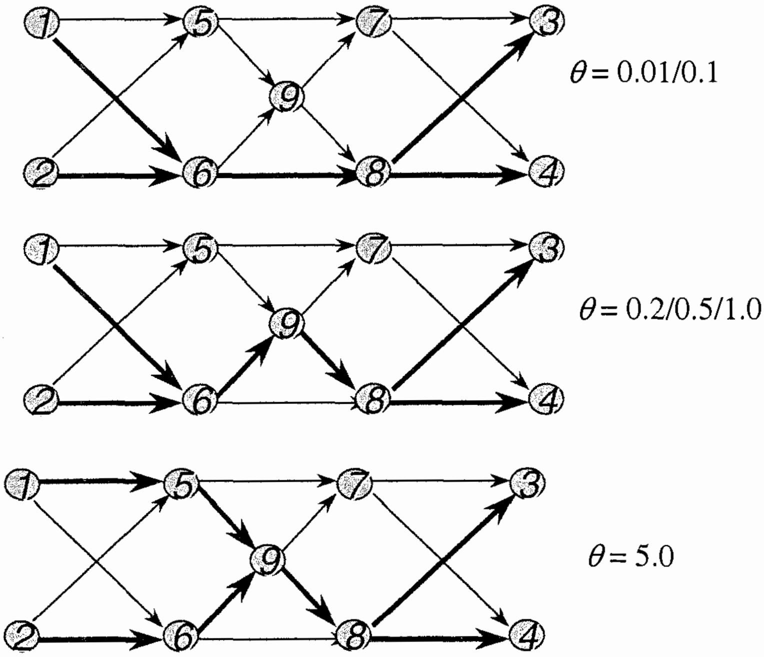


Fig. 6. Zero-toll trees as θ varies.

into consistent link tolls. Also the issue of small path flows encountered in the logit case would result in unreasonably large tolls on some routes. Further, algebraic methods are untenable in the probit case, and so such methods were not felt to be desirable, and instead an iterative heuristic link based method has been derived.

For the toy-network used here for illustration the desired SO flow pattern where TNTC was minimised could be closely approached, within a small number of iterations. This method does however require extension to examine larger networks to see how close to the TNTC it is possible to get in general. A sensible trade off between the cost imposed upon the drivers to achieve the reduction in TNTC, and the actual reduction obtained would need to be established for practical purposes. In addition it may be desirable to require certain links to be zero-tolled, and this could be included in this type of process.

In attempting to achieve the Stochastic Social Optimum flow pattern, by use of minimal-revenue tolling, the marginal social cost price tolls, known to create the desired flows, were used as a starting point. Path enumeration was then required to use these to derive minimal revenue path-based tolls and from these, link-based tolls. The minimal-revenue toll problem in this case is analogous to that for deterministic assignment, but with the stochastic nature of the assignment causing all used paths not to have a common cost. It would be possible here to use an iterative method similar to that used in seeking to approach the 'true' SO, but if possible it would be more desirable to utilise the easily established MSCP tolls as a starting point, but to derive a fully link based procedure. This is an area of ongoing work.

The desired flow pattern to be achieved in the stochastic case remains though an issue to be resolved. Is it more desirable in the stochastic case to minimise 'real' or 'perceived' costs throughout a network?

6 Future Work

This paper has been based on the assumption of a fixed demand stochastic equilibrium model. It is clear that imposing tolls on a network, will directly affect demand as well as being able to influence route choice. Elastic demand may be readily included in stochastic equilibrium models [MH97a, MH97b], and in the SSO case, MSCP tolls may be derived by using marginal cost functions in an SUEED algorithm. However for all other feasible toll sets, such as to seek minimal revenue tolls, additional work will be required. It has been shown that in the deterministic case with elastic demand, that all valid tolls generate the same toll revenue [HY02], and further work is required to determine whether this result extends to tolling to achieve SSO under SUEED. The heuristic to approach the 'true SO' which has been developed in this paper presupposes that the desired flow pattern is fixed, and may be determined. In the case of elastic demand, further iteration will be required to account for the change in the 'desired flow pattern' as each link toll is increased. This is the subject of future work.

Acknowledgment. The authors are grateful to two referees who provided useful comments and suggestions for improving the clarity of the arguments presented in the paper, and particularly to one who suggested a clearer and more rigorous approach to the arguments in Section 3.

References

[BHR97] Bergendorff, P., Hearn, D.W., Ramana, M.V.: Congestion Toll Pricing of Traffic Networks, Network Optimization. In: Pardalos, P.M., Hearn, D.W., Hager, W.W. (Eds) Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, **450**, 51–71 (1997)

[Cas90] Cascetta, E.: Metodi Quantitativi per la Pianificazione dei Sistemi di Trasporto. Cedam. P71 (1990)

[site1] Commission for Integrated Transport, Paying for Road Use, www.cfit.gov.uk (2002)

[Cla61] Clark, C.E.: The greatest of a finite set of random variables. Operations Research, **9**, 145–162 (1961)

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

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

[HY02] Hearn, D.W, Yildirim, M.B.: A Toll Pricing Framework for Traffic Assignment Problems with Elastic Demand. In : Gendreau, M., Marcotte, P. (eds.) Current Trends in Transportation and Network Analysis. Papers in honor of Michael Florian. Kluwer Academic Publishers, 135–145 (2002)

[HR98] Hearn, D.W, Ramana, M.V.: Solving Congestion Toll Pricing Models. In: Marcotte P, Nguyen (eds.) Equilibrium and Advanced Transportation Modeling. Kluwer Academic Publishers, 109–124 (1998)

[Iso98] Ison, S.: A concept in the right place at the wrong time: congestion metering in the city of Cambridge. Transport Policy, **5**, 139–146 (1998)

[MH97a] Maher M.J, Hughes P.C.: A probit-based stochastic user equilibrium assignment model. Transportation Research, **31B**, 341–355 (1997)

[MH97b] Maher, M.J, Hughes, P.C.: An Algorithm for SUEED Stochastic User Equilibrium with elastic demand. Presented at the 8^{th} IFAC Symposium on Transportation Systems, Chania, Crete (1997)

[MSR05] Maher, M.J, Stewart, K, Rosa, A.: Stochastic Social Optimum Traffic Assignment. Transportation Research, **39B**, 753–767 (2005)

[MM00] May, A.D., Milne, D.S.: Effects of alternative road pricing systems on network performance. Transportation Research, **34A**, 407–436 (2000)

[Ros01] Rosa, A.: Path-based Traffic Assignment with Probit Analytical Methods. Proc. 33^{rd} annual Universities Transport Study Group conference, Oxford University (2001)

[RM02] Rosa, A., Maher, M.J.: Algorithms for solving the probit path-based stochastic user equilibrium traffic assignment problem with one or more user classes. In: Taylor, M.A.P. (ed.) Transportation and Traffic Theory in the 21^{st} Century. Proceedings of the 15^{th} International Symposium on Transportation and Traffic Theory, Pergamon Press, 371–392 (2002)

[She85] Sheffi, Y.: Urban Transportation networks: Equilibrium Analysis with Mathematical Programming Methods, Prentice-Hall (1985)

[SEL94] Smith T.E., Eriksson E.A., Lindberg P.O.: Existence of Optimal Tolls under Conditions of Stochastic User-Equilibria. In: Johansson B., Mattsson L.G. (eds.) Road Pricing: Empirical Assessment and Policy. Kluwer Academic Publishers, Dordrecht, The Netherlands, 65–87 (1994)

[War52] Wardrop, J.G.: Some theoretical aspects on road traffic research. Proc. Inst. Civil Engineers, **11**, 325–378 (1952)

[Yan99] Yang, H.: System Optimum, Stochastic User Equilibrium, and Optimal Link Tolls. Transportation Science, **33**, 354–360 (1999)

Appendix

$$A = \begin{array}{|cccccccccccccc|c}
(1,5) & (5,7) & (7,3) & (1,6) & (2,5) & (5,9) & (9,7) & (6,9) & (9,8) & (7,4) & (8,3) & (2,6) & (6,8) & (8,4) & \\
\hline
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & [1,3] \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & [1,4] \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & [2,3] \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & [2,4] \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & \\
\hline
\end{array}$$

An Optimal Toll Design Problem with Improved Behavioural Equilibrium Model: The Case of the Probit Model

Agachai Sumalee[1], Richard Connors[2], and David Watling[3]

[1] Institute for Transport Studies, University of Leeds, 38 Woodhouse Lane, Leeds, LS2 9JT, United Kingdom, asumalee@its.leeds.ac.uk
[2] Institute for Transport Studies, University of Leeds, 38 Woodhouse Lane, Leeds, LS2 9JT, United Kingdom, rconnors@its.leeds.ac.uk
[3] Institute for Transport Studies, University of Leeds, 38 Woodhouse Lane, Leeds, LS2 9JT, United Kingdom, dwatling@its.leeds.ac.uk

Summary. This paper considers the optimal toll design problem that uses the Probit model to determine travellers' route-choices. Under probit, the route flow solution to the resulting stochastic user equilibrium (SUE) is unique and can be stated implicitly as a function of tolls. This reduces the toll design problem to an optimization problem with only nonnegativity constraints. Additionally, the gradient of the objective function can be approximated using the chain rule and the first order Taylor approximation of the equilibrium condition. To determine SUE, this paper considers two techniques. One uses Monte-Carlo simulation to estimate route choice probabilities and the method of successive averages with its prescribed step length. The other relies on the Clark approximation and computes an optimal step length. Although both are effective at solving the toll design problem, numerical experiments show that the technique with the Clark approximation is more robust on a small network.

Key words: Network Design Problem, Probit SUE, Optimal toll, Sensitivity Analysis

1 Introduction

Transport can be considered as an economic market where travellers are economics agents with the aim of maximising (or minimising) their utility (or disutility). With the cross-effect of one user's strategy on another through the congestion in the network, the concept of Nash's equilibrium can be invoked to define the converged travellers' strategies (e.g. route, mode, departure time, or destination choices). The Nash equilibrium occurs when no individual (traveller) can change their strategy to decrease their own disutility. However, it is

Mathematical and Computational Models for Congestion Charging, pp. 219-240

well known that under the assumption of individual utility maximization, the converged equilibrium point of the transport system may not be the optimal travel pattern for the overall system, nor for other aggregated objectives of the traffic system manager (e.g. total travel time, environmental impact, or social welfare).

Road pricing has been proposed as the means to direct the traffic equilibrium condition to a more desirable state ([Kni24], [Wal61]). Early developments of the theory of road pricing have been mainly associated with the concept of deterministic user equilibrium, namely Wardrop's user equilibrium (UE) principle [War52]. UE is a special case of Nash's equilibrium condition and has been widely adopted as the modelling assumption for representing travellers' behaviour. The key assumption of UE is that the traveller has perfect information regarding their travel choices and the alternatives. Despite questions about the realism of the assumption, the UE model has played a major role in the analysis of road pricing, in which a number of researchers over the years have focussed on deriving optimal toll patterns under the UE condition (e.g. [YH98], [SNR01], [Ver02]; [MLSS02], [SS04], [Sum04]).

The key element of microeconomic theory lies in understanding the consumer's behaviour. The concept of a random utility model (RUM) has been developed to better represent the individual's choice making process. RUM may be integrated with the traffic equilibrium model by representing the pay-off function, or disutility, as a random utility term. This random disutility of travel is widely referred to as the *perceived* disutility/cost of travel. The equilibrium point can then be defined as the situation where no traveller can switch his/her strategy to improve his or her *perceived* cost of travel. With this setting, we obtain the concept of Stochastic User Equilibrium (SUE). Apart from the enhanced realism of the behavioural model underlying the SUE model, the algorithmic advantage of using an SUE model in optimal toll design has also been previously implied (e.g. [Dav94], [PR03]). This issue will be discussed later on in the paper.

Many error structures have been proposed for SUE. They include the commonly used independent Weibull and multivariate normal that lead to the logit and probit models respectively [She85], as well as more general cross-nested logit models [PB99], mixed error component models [NDF02] and gamma link component distributions [CB02].

Among the logit and probit models, the former is more popular because of its closed form expression for the choice probabilities. Several researchers (e.g., [SEL94], [AK5], [Yan99]) have used the logit model to study toll pricing under SUE. However, the underlying assumption for the logit model is rather restrictive. In particular, it assumes that travel alternatives are uncorrelated and have no overlapping structure. Generally, this is referred to as the 'independence of irrelevant alternatives' assumption or IIA. On the other hand, despite its complexity the probit model can overcome the IIA issue of the overlapping routes. Thus, the probit SUE will be adopted as the model for travellers' behaviour in this paper.

The paper is organised into five further sections. The next section presents the formulation of the optimal toll design problem with SUE and the definition of SUE. Then, section 3 explains the treatment of variable demand (elastic demand) with the probit SUE and the different computational methods adopted for solving the probit SUE. Section 4 reformulates the optimal toll design with SUE in the form of an implicit program, and the algorithm for solving this problem is presented. Section 5 provides numerical results using a test network. Finally, section 6 concludes the paper.

2 Problem Formulation of Optimal Toll Design with Stochastic User Equilibrium

The problem discussed in this paper is the optimal toll design problem where the response from the users to the toll imposed is assumed to follow a random utility model. We focus on the case of an automobile network with a single mode, single user class, and single time period. The underlying network is a directed graph with N nodes and a set of links denoted A. The demand matrix q has entries q_{rs}, representing the travel demand from origin r to destination s, where $r, s = 1, \dots, N$. The vector of link flows is $\mathbf{x}$, with link costs $\mathbf{t}(\mathbf{x})$, so that $t_a(x_a)$ is the cost (without toll) of travelling along link $a \in A$ when the link flow is x_a. Let β_a denote the toll level of link $a \in A$. Then the generalised travel cost on link a is $t_a(x_a) + \beta_a$. In addition, let K_{rs} be the set of routes connecting node r to node s. Associated with K_{rs} is the link-route incidence matrix, Δ^{rs}, whose element, $\delta^{rs}_{a,k}$, equals 1 if link a is on route k that connects node r to node s. An assignment of flows to all routes is denoted by the vector $\mathbf{f}$, with $f^{rs}_k \geq 0 \ \forall k, r, s$. The assignment $\mathbf{f}$ is *feasible* for demand $\mathbf{q}$ if and only if

$$\sum_{k \in K_{rs}} f^{rs}_k = q_{rs} \ \forall r, s,$$

and the (convex) set of feasible route flows is denoted F. For any $\mathbf{f} \in F$, $\mathbf{c}(\mathbf{f})$ denotes the associated vector of route costs where

$$c^{rs}_k(\mathbf{f}) = \sum_{a \in A} \left(t_a(\mathbf{x}(\mathbf{f})) + \beta_a\right) \ \delta^{rs}_{a,k}.$$

Travellers are allowed to respond to the toll imposed by changing their routes or deciding not to travel (the precise mechanism for achieving this is described in section 3). The responses of the travellers are assumed to follow the Stochastic User Equilibrium condition (SUE). Let Φ be a mapping from $\Re^{|\beta|} \rightarrow \Re^{|\kappa|}$ that gives the vector $\mathbf{f}$ of feasible route flows satisfying the SUE condition, given a toll vector β. Let Z($\mathbf{f}$, β) be the objective function that we wish to optimise. We can then formulate the optimization problem for determining the optimal toll as:

$$\begin{aligned}&\max_{(\mathbf{f},\beta)} \mathrm{Z}(\mathbf{f},\beta)\\ &\text{s.t. } \mathbf{f}=\Phi(\beta)\\ &\qquad \beta\geq\mathbf{0}.\end{aligned}$$

Note that this problem can be considered as a mathematical program with equilibrium constraints (MPEC). As noted previously by many authors, this formulation can also be applied to the UE case, but with a mapping between the link flow vector and the toll vector, since the route flow in UE is not unique.

This paper assumes that the route choice behaviour follows a random utility model. In particular, the perceived cost of the k-th route is a random variable of the form:

$$C_k = c_k + \varepsilon_k$$

where $c_k = c_k(\mathbf{f})$ is the mean perceived route cost and the random errors $(\varepsilon_1, \varepsilon_2, ...)$ follow some joint probability density function with zero mean vector. These random error terms represent the fact that individual drivers have their own assessment of both network conditions and of the cost of taking different routes (including their personal preferences for some routes over others).

Given the route cost vector $\mathbf{c}$, $P_k^{rs}(\mathbf{c})$ denotes the proportion of drivers who perceive route k to be the cheapest route from r to s, i.e.

$$\begin{aligned}P_k^{rs} &= \Pr\left(C_k^{rs} \leq C_j^{rs}\ \forall j\in K_{rs}, j\neq k\right)\\ &= \Pr\left(\varepsilon_k^{rs}+c_k^{rs}\ \leq \varepsilon_j^{rs}+c_j^{rs}\forall j\in K_{rs}, j\neq k\right),\end{aligned}$$

where Pr(.) denotes probability. Then, the stochastic user equilibrium (SUE) can be stated as follows:

At SUE, no driver can improve their perceived travel cost by unilaterally changing route.

The SUE route flow assignment (for $\mathbf{f}\in F$) is, therefore, the solution to the following fixed-point problem:

$$f_k^{rs} = q_{rs}P_k^{rs}(\mathbf{c}(\mathbf{f}))\ \forall k\in K_{rs},\ \forall r,s.$$

This states that, for a given OD pair, the flow on the k-th route consists of those drivers who perceive this to be the best route. Since $\mathbf{f}$ is defined to be a feasible set of flows, the total number of drivers on all routes connecting r to s matches the total travel demand from this origin to this destination. A network route flow vector satisfying SUE will be denoted $\mathbf{f}^*$. This fixed-point condition defines the mapping Φ between the SUE flows and the toll vector.

With the SUE, several properties that UE does not possess can be gained. Consider first a simplified network structure in which the only routes are non-overlapping and consist of single links, and that for given tolls the vector of link travel cost functions is continuous and strictly increasing in the vector of link flows. In this case, for given link tolls, there are unique UE link flows and route flows (see e.g. [Smi79]).

In UE, a route will be used if and only if the travel cost on this route is the minimum O-D travel cost (compared to all other routes connecting the same O-D pair). This can be represented as a complementarity condition: $0 \leq f_k^{rs} \perp (C_k^{rs} - C^{rs*}) \geq 0$, where C^{rs*} denotes the minimum travel cost from origin r to destination s and $x \perp y \equiv x \cdot y = 0$. This complementarity condition is non-differentiable when $f_k^{rs} = (C_k^{rs} - C^{rs*}) = 0$. Thus, when including this condition into the optimal toll design problem, one may face a non-differentiable optimization problem. This is an example of a wider phenomenon arising from the *complementarity condition* as constraints to optimization problems ([PR02], [LPR96]).

In general network structures, while the set of link flow solutions to the UE model at given tolls is a singleton under the assumption that the vector of link travel cost functions is continuous and strictly monotonic [Smi79], it is well known that the UE *route* flow solutions are typically non-unique. Therefore, route-based solution strategies are commonly faced with an additional hurdle of selecting a single UE route flow solution from a convex set, for example by an arbitrary choice of extreme point (e.g. [TF88]) or by an additional model selecting the 'most likely' route flows (e.g. [LLPR01]). Still, establishing desirable properties of a sequence of such 'unique' UE route flow solutions, as the tolls are altered, may be extremely problematic.

For problems with continuous and strictly monotone link cost functions as above, under mild conditions on the choice probability model, SUE is known to give rise to solutions (a) in which *all* routes are active, at least in theory, and (b) that are unique in the *route* flow domain (e.g. [CC95]). Therefore, it is natural to ask, is solving the optimal toll problem with an SUE network model actually *easier* than with a UE? At the same time, one is adopting a model that, from a behavioural perspective, is arguably superior in terms of its representation of the uncertainty and heterogeneity that surely exists in traveller decisions.

3 Probit Equilibrium with Variable Demand: Formulation and Solution Algorithm

The SUE model in section 2 assumes that travel demands are fixed. In this section, we allow demands to vary. Maher et al. [MHK99] assume that the demand for OD pair (r,s) is a function of the expected minimum travel time between the origin and destination, i.e. q_{rs} depends on $E\left[\min\left\{C_k^{rs} : k \in K_{rs}\right\}\right]$. When the logit route-choice is used, the demand function resulting from the assumption can be mathematically expressed in a closed form (see, e.g. [BDK86], [GP01]) but this is not the case for probit.

To make our model more manageable under probit, we add to the original network a pseudo-link (r,s) for each OD pair. The amount of flow on pseudo link (r,s) represents the number of drivers who decide not to travel from r to s. The perceived travel cost on each pseudo link (or link zero) is $c_0^{rs} + \varepsilon_0^{rs}$, where c_0^{rs} represents the deterministic disutility of not travelling and ε_0^{rs} is the associated random error in accordance with the probit model. Then, the proportion of drivers who decide not to travel is given by the following expressions:

$$\begin{aligned} P_0^{rs} &= \Pr\left(c_0^{rs} + \varepsilon_0^{rs} \leq c_k^{rs} + \varepsilon_k^{rs}\ \forall k \in K_{rs}\right) \\ &= \Pr\left(c_0^{rs} + \varepsilon_0^{rs} \leq \min_{k \in K_{rs}} \{c_k^{rs} + \varepsilon_k^{rs}\}\right), \end{aligned}$$

and the condition for SUE can be written in the same manner for those with fixed demand:

$$f_k^{rs} = q_{rs} P_k^{rs}(\mathbf{c}(\mathbf{f}))\ \forall k \in K_{rs}^0,\ \forall r, s,$$

where $K_{rs}^0 = K_{rs} \cup \{0\}$, with f_0^{rs} the number of drivers electing to not travel. Moreover, q_{rs}now represents the number of potential drivers, some of whom choose the pseudo link, i.e. decide not to travel.

The probit model assumes that perceived route costs are derived from normally distributed perceived link costs:

$$C_k^{rs} = \sum_a T_a \delta_{a,k}^{rs} \quad \forall k, r, s$$

with $T_a \sim N\left(t_a, \sigma_a^2\right)$, with σ_a^2 constant. In this paper we assume that the perceived link costs, $\{T_a\}$, are independent. The distribution of perceived route costs is therefore multivariate normal, $C \sim MVN\left(c, \Sigma\right)$, centred on the deterministic route costs. This results in a variance-covariance matrix, Σ, where the perceived costs of routes that have links in common are correlated.

To determine a solution that satisfies the above equilibrium condition, any algorithm that solves a probit-based SUE problem with fixed demand can be used. In Section 5, we consider the following algorithms:

- The method of successive averages (MSA) algorithm (see [She85]) with probit choice fractions estimated by a Monte Carlo (MC) simulation.
- A step-length algorithm recently proposed by Maher and Hughes [MH97] that uses the equivalent optimization formulation of SUE [DS77] with the Clark approximation ([Cla61], [HSD82]) for computing probit choice probabilities.

4 Implicit Programming Approach to Optimal Toll Design

Assume that the travel cost of link a, $t_a(x)$ is continuous for each a and the travel cost vector, $\mathbf{t}(\mathbf{x})$, is strictly monotone. Then, the route-flow solution of the probit-based SUE problem is unique (see e.g. [CC95]) and the optimal toll design problem can be formulated as follows:

$$\max_{\beta}\left\{\mathrm{Z}\left(\mathbf{x}^*\left(\beta\right),\beta\right):\beta\geq\mathbf{0}\right\}$$

where $\mathbf{x}^*\left(\beta\right)$denotes a link flow solution to the probit SUE problem at toll vector β.

As stated above, the optimal toll design problem is an optimization problem with simple bounds. Many algorithms for such a problem typically require, at minimum, calculating the gradient of the objective function at the current solution. When Z is relatively simple, its gradient can be approximated. To illustrate, consider the revenue function, i.e. $\mathrm{Z}\left(\mathbf{x}^*\left(\beta\right),\beta\right)=\beta^T\cdot\mathbf{x}^*\left(\beta\right)$. In this case,

$$\nabla_\beta\mathrm{Z}\left(\mathbf{x}^*\left(\beta\right),\beta\right)=\mathbf{x}^*\left(\beta\right)+\beta^T\cdot\nabla_\beta\mathbf{x}^*\left(\beta\right),$$

where $\nabla_\beta\mathbf{x}^*\left(\beta\right)$ denotes the Jacobian of $\mathbf{x}^*$ at β.

From the relationship between link and route flow, we can define the Jacobian of $\mathbf{x}^*$at βas:

$$\nabla_\beta\mathbf{x}^*\left(\beta\right)=\Delta\cdot\nabla_\beta\mathbf{f}^*\left(\beta\right),$$

where $\mathbf{f}^*\left(\beta\right)$ is a vector of SUE route flow solution at β, Δ is the link-route incidence matrix whose element, $\delta_{a,k}$, equals 1 if link a is on route k, and $\nabla_\beta\mathbf{f}^*\left(\beta\right)$ denotes the Jacobian of $\mathbf{f}^*$ at β. To approximate $\nabla_\beta\mathbf{f}^*\left(\beta\right)$, consider the 'gap' function:

$$\Psi\left(\mathbf{f},\beta\right)=\mathbf{f}-\mathbf{q}\cdot\mathbf{P}(\mathbf{c}(\mathbf{f},\beta)),$$

where $\mathbf{P}$ is the route-choice probability operator as defined in Section 2. Assuming all functions are differentiable, the first order Taylor approximation of $\Psi\left(\mathbf{f}^*\left(\beta\right),\beta\right)$ at $\left(\mathbf{f},\beta\right)=\left(\mathbf{f}^*\left(\beta_0\right),\beta_0\right)$ is:

$$\Psi\left(\mathbf{f},\beta\right)\approx\Psi\left(\mathbf{f}^*\left(\beta_0\right),\beta_0\right)+J_1\left(\mathbf{f}-\mathbf{f}^*\left(\beta_0\right)\right)+J_2\left(\beta-\beta_0\right),$$

where J_1 and J_2 are the Jacobians of Ψ evaluated at $\left(\mathbf{f}^*\left(\beta_0\right),\beta_0\right)$ with respect to $\mathbf{f}$ at β, respectively, i.e., $J_1=\nabla_{\mathbf{f}}\Psi\left(\mathbf{f}^*\left(\beta_0\right),\beta_0\right)$ and $J_2=\nabla_\beta\Psi\left(\mathbf{f}^*\left(\beta_0\right),\beta_0\right)$. (See [BI97], [Dag79], and [CW02] for the calculation of J_1 and J_2). Because $\Psi\left(\mathbf{f}^*\left(\beta\right),\beta\right)=0$ for all β, the above reduces to

$$0\approx0+J_1\left(\mathbf{f}-\mathbf{f}^*\left(\beta_0\right)\right)+J_2\left(\beta-\beta_0\right).$$

When J_1 is non-singular, the above implies that $-J_1^{-1}J_2$ is an approximation of the Jacobian of $\mathbf{f}^*(\beta)$ at β_0, i.e.,

$$\mathbf{f}^*(\beta) - \mathbf{f}^*(\beta_0) \approx -J_1^{-1}J_2(\beta - \beta_0)$$

or

$$\lim_{\beta \to \beta_0} \frac{\mathbf{f}^*(\beta) - \mathbf{f}^*(\beta_0) + J_1^{-1}J_2(\beta - \beta_0)}{\|\beta - \beta_0\|} \approx 0.$$

For the above example, $\nabla_\beta Z(\mathbf{x}^*(\beta), \beta) = \mathbf{x}^*(\beta) - \Delta J_1^{-1}J_2\beta$.

5 Numerical Experiments

5.1 Definition of the test network

The network adopted for the test has seven nodes connected by 18 links, with six pseudo-links representing the no-travel options for each OD movement (as required in the variable demand probit SUE model). Figure 1 shows the topology of the network. There are six OD pairs: (1, 5), (1, 7), (5, 1), (5, 7), (7, 1), and (7, 5). Table A.1 in the Appendix gives the origin-destination 'potential demand' matrix. The link cost functions are based on the BPR function $t_i(x_i) = a_i + b_i\left(\frac{x_i}{\kappa_i}\right)^{n_i}$ where a_i, b_i, n_i, and κ_i are given in Table A.2 in the Appendix. Note that for the 'no travel' or 'pseudo' links, there is only a constant parameter associated with the disutility of not conducting a trip, i.e. b_i= 0 for such links. As in [She85] the probit link error terms are independent and normally distributed with zero mean and standard deviations as listed in Table A.2 in the Appendix.

Tolls are implemented by adding the tolls to the free flow costs. When an additional cost is added to the free flow parameter for a pseudo-link, this can be thought of as representing an increase in the no-travel cost (for that OD movement) representing the increase in the utility of conducting a trip. There are 36 routes among the six OD pairs.

For the variance-covariance matrix, we have adopted the common approach used in the probit model [She85] of assuming the path cost covariance matrix is derived from independent Normal link cost error distributions. For these distributions, the variance for link j (excluding pseudo-links) is assumed to be $\sigma_j^2 = \alpha \cdot a_j^2$ where α is a link-independent scaling factor and a_j is the free flow travel cost for link j, and for the pseudo-links the variances are set to the mean of the variances on the real links.

Different values of αcan be used to define different levels of the perception error of the travellers on the travel time/cost resulting in different behavioural models. Several values α are adopted to investigate effect of the behavioural model on the route choice behaviour, the resulting flows, and the optimal toll levels (α = 0, 0.3, 1, 3).

The objective function adopted in the test is a combination of the revenue and the actual total travel time. The revenue, R, is simply calculated by summing the tolls multiplied by the relevant link flows for the tolled links. The total travel time, TTT, is calculated from only the real links (since the flow on the pseudo-links does not travel); it is the sum of the link flows multiplied by the link travel times (without the toll included).

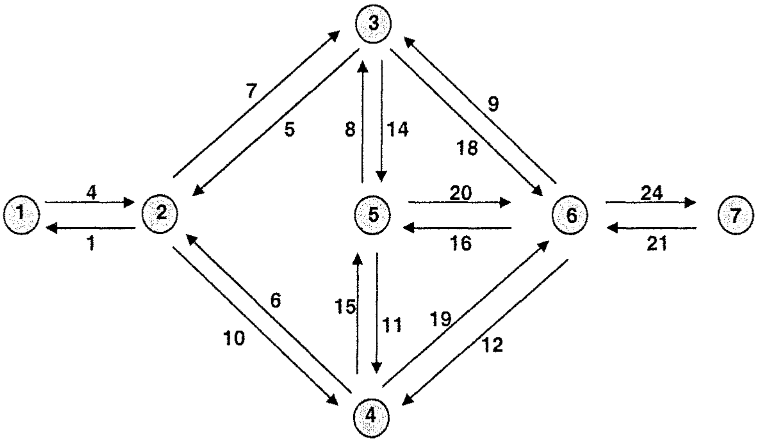


Fig. 1. The topology of the test network (without the pseudo-links)

The objective function is $Z = \mu R + (1 - \mu)(-TTT)$ where μ is a weighting factor with $0 \leq \mu \leq 1$. The gradient of the objective function with respect to tolls can be derived as follows:

$$\begin{aligned}\nabla_\beta Z \approx\ & \mu\left(\mathbf{x}^*(\beta) - \Delta J_1^{-1} J_2 \beta\right) \\ & + (1-\mu)\,\Delta J_1^{-1} J_2 \left\{\mathbf{t}\left(\mathbf{x}^*(\beta)\right) - \operatorname{diag}\left[\mathbf{x}^*(\beta)\left(\nabla_{\mathbf{x}} \mathbf{t}\left(\mathbf{x}^*(\beta)\right)\right)^T\right]\right\},\end{aligned}$$

where $\nabla_{\mathbf{x}}\mathbf{t}$ denotes the Jacobian of the travel cost with respect to flows, and J_1 and J_2 are as defined in Section 4.

To demonstrate the behaviour of the test network, the revenue generated and total travel time for different toll levels applied to each link in turn are shown in Figure 2 below. In these tests the covariance scaling factor, α, is set to 1. From the figures, the revenue levels generated are most sensitive to tolls on links 1,4, 21 and 24. The network diagram above shows that these are the links that cannot be avoided (by the relevant OD movements); the only alternative "route" is the no-travel option. Thus, it is no surprise to observe that these links can generate the highest revenues. For the other links in the network, travellers can avoid the tolled link by changing route. For the total travel time, tolling on certain links (e.g. link 8) increases the total travel time as we increase the toll. For other links (e.g. link 4) the opposite occurs. With the weighting factor μ= 0.5 the objective function values as each link is tolled individually are shown in Figure 3.

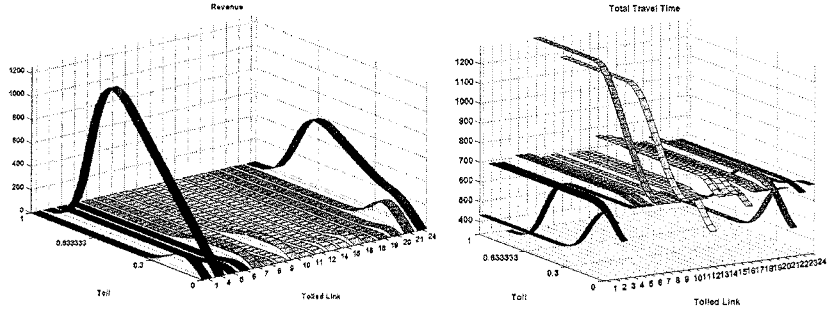


Fig. 2. Revenue and total travel time for different toll levels on each link

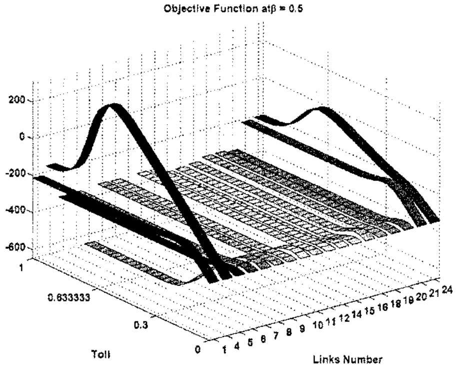


Fig. 3. Objective function levels for different toll levels applied to each link in turn

5.2 Comparison of different SUE solution algorithms

In this section, the two alternative algorithms proposed for solving the SUE problem (described in section 3.2) are tested. We consider the case of tolling links 14, 15, and 16 simultaneous with a uniform toll. For this one-dimensional problem the gradient of the objective function at each toll level can be plotted as shown in Figures 4 and 5. Three different levels of α are adopted for the test (α = 0.3, 1, and 3). Six curves are plotted, three for each method with different α in each figure. Figure 4 compares the gradient of the objective as calculated by 'numerical differencing' (a finite difference approximation) and by sensitivity analysis (see Section 4) in which the SUE flows are calculated by the first method (MSA + MC-estimated choice probabilities). Figure 5 shows the same comparison but the SUE flows are calculated by the second method (Clark approximation + optimal step length). In both figures, the curves with the bold line are the gradients calculated from numerical differencing and the broken lines are the gradients from the sensitivity analysis.

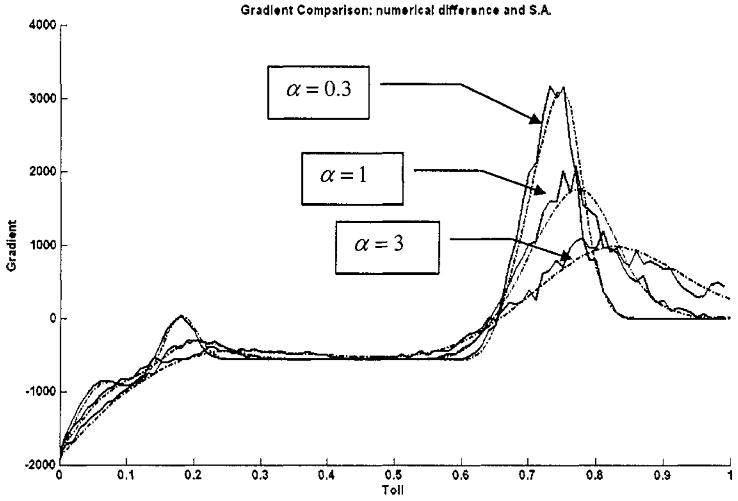


Fig. 4. Gradients of the objective at different toll levels calculated from the numerical differencing (solid line) and sensitivity analysis (broken line). MSA calculations using the MC simulation and predefined step-length.

In both cases, the gradients calculated by the sensitivity analysis method are reasonably smooth. In Figure 4, the numerical differencing produces a non-smooth gradient that is caused by the non-smooth objective function as calculated from the MC simulation and pre-defined step length. Although the Clark approximation does have disadvantages (in terms of where this approximation is valid) the resulting link flows (and hence objective function values) are much smoother than the corresponding values calculated on the basis of the MC simulation. The gradients calculated by numerical differencing of the SUE flows resulting from the Clark approximation based approach (bold line in Figure 5) are visually as smooth as the gradients calculated via sensitivity analysis in the same figure (broken line).

Obviously, different methods significantly influence the smoothness of the objective function. The MC based method does suffer from the unpredictability of the random trial process which may not guarantee the same SUE flows/route choice probabilities with different runs. On the other hand, the benefit of the MC based method is that with a high number of the trials the

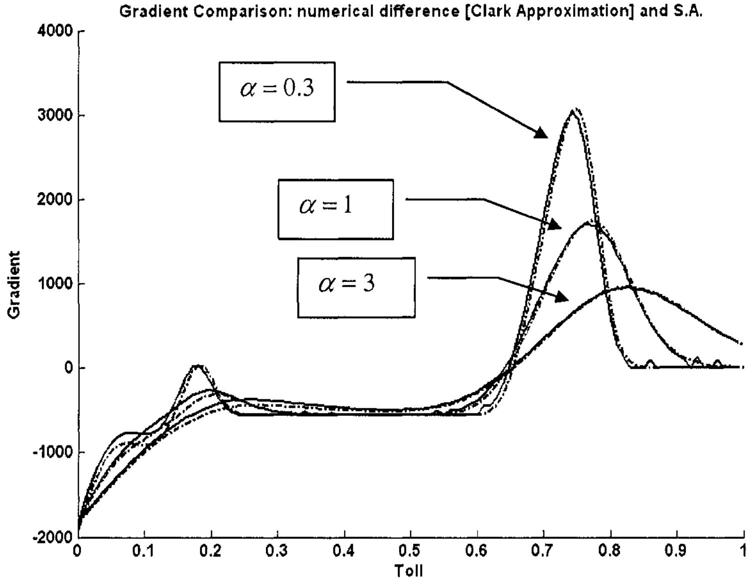


Fig. 5. Gradients of the objective at different toll levels calculated from the numerical differencing (solid line) and sensitivity analysis (broken line). MSA calculations using the Clark approximation and optimal step-length.

accuracy of the estimation of the route choice probability may be improved, but one can never be sure what constitutes a sufficient number of trials. The Clark approximation, despite its possible drawback on the accuracy of the approximation, does produce very good results in terms of the smoothness of the objective function. Nevertheless, in both cases the sensitivity analysis method can eventually define a smooth trend of the gradient reflecting the real property of the problem. The reason is that the sensitivity analysis method estimates the gradient based on a single point (see previous section). Thus, it does not suffer from the poor convergence of the SUE flows from one toll level to another whereas the numerical differencing, which uses two points of SUE flows, suffers from this error.

Based on this comparison, we decided to adopt the second approach (Clark approximation + optimal step length) for the tests in the following sections.

5.3 Effect of probit variances on the optimal toll policy

This section presents some numerical results using the optimization approach explained in Section 4 to find the optimal tolls for different cases. The sequential quadratic programming (SQP) algorithm in MATLAB ('fmincon' solver) is adopted to solve the problem, with the Jacobian of the objective function supplied (using the approach described in Section 4). Before applying the optimization algorithm to the test, we explore the effect of the behavioural model parameters on the objective function. Three different sets of tests are conducted. In the first set of tests, we apply the uniform toll level on link 8, 11, 14, and 15 with four different values of α. Similarly, the second set of tests involves imposing the uniform tolls on link 14, 15, and 16 making a pricing cordon around node 5. The third set of tests is to put the toll on link 4 only.

For all tests, we provide the plots the corresponding objective function values (see Figures 6, 7, and 8 below). Different values of the scaling parameter α show the influence of the behavioural model on the objective function profile. The first observation is the smoothing effect of the α parameter on the objective function. When α= 0 (UE case), non-smoothness of the objective function is apparent. This property of the MPEC with UE is well documented where the objective function can be non-differentiable at some point.

On the other hand, the objective function curves with $\alpha > 0$ appear to be smooth. As the probit variances increase (with α), so drivers become less reactive to changes due to the toll and there is non-zero probability for each route to be used. This property of the SUE model contributes to the smoothness of the objective function with respect to the toll. As mentioned earlier, although the main incentive of introducing the probit SUE in place of UE is to increase the realism of the lower level model for the optimal toll problem, the SUE model may also make the optimal toll problem become easier to deal with. The other observation is the possible change of the optimal toll solution for the different values of α. With all tests, the value of the optimal toll levels do change according to the level of α.

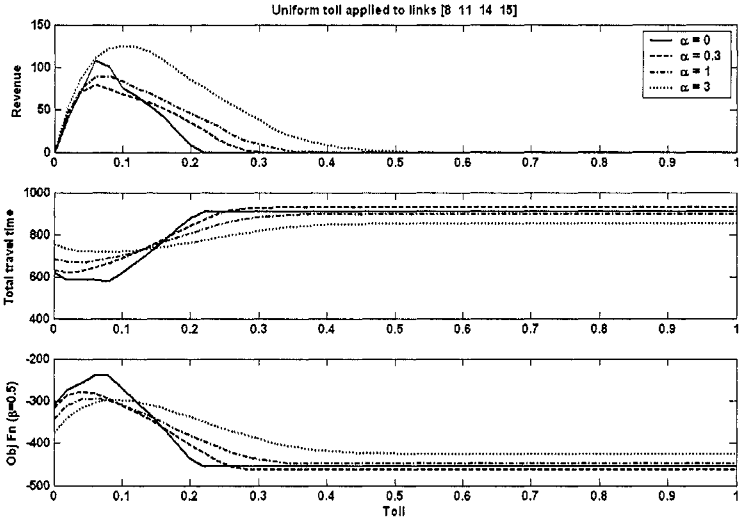


Fig. 6. Revenue, total travel time, and objective function curves with different values of α and different uniform toll levels on link 8, 11, 14, and 15

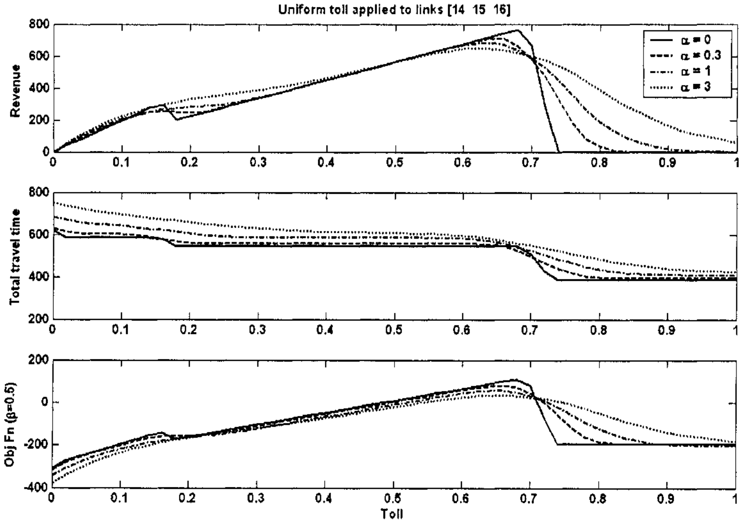


Fig. 7. Revenue, total travel time, and objective function curves with different values of αand different uniform toll levels on link 14, 15, and 16

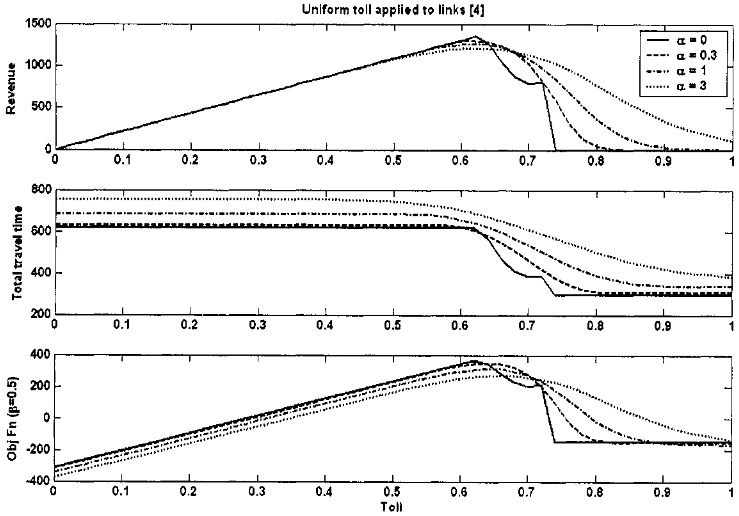


Fig. 8. Revenue, total travel time, and objective function curves with different values of αand different toll levels on link 4

Table 1 shows the results from applying the optimization algorithm to find the optimal uniform toll applied to links 14, 15, and 16 with different values of α.

Table 1. Optimal toll on links 14, 15, 16 with differentα found by the optimization algorithm

α	**Optimal Toll**	**Objective at optimal toll**
0.0001	0.68867	114.7479
0.3	0.65346	80.7625
1	0.64773	55.7007
3	0.6566	33.6732
10	0.7	25.6591

Figure 7 can be used to verify that the optimization algorithm can find the real optimal toll level for each case. Again, as mentioned the optimal toll levels change with the levels of α. Unfortunately, we cannot observe any clear relationship between the optimal toll and the level of α from the results.

The optimization algorithm is also applied to the find the optimal toll level on all links (except the pseudo links) simultaneously and the optimal toll level on each link in turn, again with different levels of α. Table 2 shows the result with the optimal toll on each link simultaneously and Table 3 shows the results with the toll on each link in turn.

Note that the column 'objective function at optimal toll' shows the absolute value of the objective function at that toll level. The objective function adopted here, as explained, is a weighted sum of the revenue and negative total travel time. Therefore, it is possible that the objective function may become negative even at the optimal toll. This does not mean the optimal toll generate dis-benefit, since the objective at the no toll scenario is a negative figure as well. Column 'benefit' in both tables presents the relative improvement of the objective of each toll policy compared with the no-toll situation. The optimization algorithm successfully solved all the scenarios reported here.

For the case with the tolls on all links, the improvement of the overall objective function increases as α increases. When all links are tolled, the links with the highest toll levels are links 4, 14, and 15. However, when each link is tolled individually, the links with the highest optimal tolls are links 1, 4, 21, and 24. Imposing the tolls on one of these link individually is actually equivalent to imposing the toll on all of the demand for some OD movement since these links are the feeding links of the demand from different OD pairs to the network (hence there is no alternative routes that avoid the tolls). The link generating the highest objective is link 4. The result may be that link 4 imposes the toll directly to a significant level of the demand in the network (the level of the demand coming from node 1 is highest compared to the other origin nodes, see Table A1 in the Appendix).

Table 2. Results from optimizing all link tolls simultaneously

α	Link Number	Optimal toll	Objective function	Benefit
	Link Number	(for this link)	at optimal toll	
0.3	1	0.0621	570.7513	886.44
	4	0.2431		
	5	0.0688		
	6	0.0660		
	7	0.2001		
	8	0.0815		
	9	0.0688		
	10	0.1930		
	11	0.0839		
	12	0.0659		
	14	0.2149		
	15	0.2171		
	16	0.1028		
	18	0.1228		
	19	0.1242		
	20	0.0952		
	21	0.0430		
	24	0.0449		
1	1	0.0379	554.0486	896.30
	4	0.2224		
	5	0.0527		
	6	0.0551		
	7	0.2003		
	8	0.0696		
	9	0.0617		
	10	0.2053		
	11	0.0726		
	12	0.0646		
	14	0.2256		
	15	0.2231		
	16	0.1177		
	18	0.1494		
	19	0.1468		
	20	0.0935		
	21	0.0395		
	24	0.0492		
3	1	0.0572	564.7796	933.9598
	4	0.2412		
	5	0.0723		
	6	0.0749		
	7	0.2185		
	8	0.0794		
	9	0.0827		
	10	0.1902		
	11	0.0791		
	12	0.0846		
	14	0.1627		
	15	0.1913		
	16	0.1134		
	18	0.1510		
	19	0.1798		
	20	0.0577		
	21	0.0543		
	24	0.0649		

Table 3. Results from optimizing each tolled link individually

α	Link Number Link Number	Optimal toll (for this link)	Objective function at optimal toll	Benefit
0.3	1	0.0621	570.7513	886.44
	4	0.2431		
	5	0.0688		
	6	0.0660		
	7	0.2001		
	8	0.0815		
	9	0.0688		
	10	0.1930		
	11	0.0839		
	12	0.0659		
	14	0.2149		
	15	0.2171		
	16	0.1028		
	18	0.1228		
	19	0.1242		
	20	0.0952		
	21	0.0430		
	24	0.0449		
1	1	0.0379	554.0486	896.30
	4	0.2224		
	5	0.0527		
	6	0.0551		
	7	0.2003		
	8	0.0696		
	9	0.0617		
	10	0.2053		
	11	0.0726		
	12	0.0646		
	14	0.2256		
	15	0.2231		
	16	0.1177		
	18	0.1494		
	19	0.1468		
	20	0.0935		
	21	0.0395		
	24	0.0492		
3	1	0.0572	564.7796	933.9598
	4	0.2412		
	5	0.0723		
	6	0.0749		
	7	0.2185		
	8	0.0794		
	9	0.0827		
	10	0.1902		
	11	0.0791		
	12	0.0846		
	14	0.1627		
	15	0.1913		
	16	0.1134		
	18	0.1510		
	19	0.1798		
	20	0.0577		
	21	0.0543		
	24	0.0649		

6 Conclusions

The traditional assumption of travellers' response to a road toll is the deterministic user equilibrium model. We have argued in this paper that a better representation of travellers' responses may be achieved through an improved behavioural model following random utility theory, as achieved through the probit SUE model. Optimal toll design with the probit SUE is then formulated, with the probit SUE framework extended in a novel way to include variable demand, by adding pseudo links to the network. The optimal toll problem with probit SUE can be categorised as a MPEC. However, the uniqueness and smoothness of the route choice probabilities in probit SUE, given a toll vector, help us in developing an optimization algorithm for tackling this problem, by reformulating the MPEC as an implicit programming problem. The key element in developing an algorithm to solve the reformulated optimal toll problem is the Jacobian of the objective function with respect to the tolls, which can be estimated in practice by applying the sensitivity analysis method.

In particular, we used the Sequential Quadratic Programming (SQP) algorithm in MATLAB to solve the optimal toll problems. The algorithm was applied to a test network (with 18 links and six OD pairs). Firstly, we tested the accuracy of two different algorithms for solving the probit SUE, one combining MSA with MC-based choice probabilities, and a second using Clark approximation method with optimal step length computation. The results show the instability of the MC based method. This is thought to be due to the lack of consistency in the convergence properties of the MC method at 'adjacent' (very similar) tolls. Clark approximation, on the other hand, produces a smoother objective function. However, there exists some uncertainty regarding the accuracy of the Clark approximation in estimating the probit route choice probabilities. Nevertheless, with both methods the sensitivity analysis can produce a reasonably smooth gradient due to the fact that in deriving the gradient of the objective, the sensitivity analysis method is only based on a single point of solution, hence reducing the uncertainty of the converged solution between two toll levels.

The second test concerned the influence of the behavioural parameters on the optimal toll solution. Different scaling parameters, which determine the magnitude of terms in the variance-covariance matrix of the probit model, were tested. The results showed some changes of the objective function curves with different scaling parameters, resulting in changes to the optimal toll solution. This result highlights the importance of calibrating the behavioural model in order to accurately determine the optimal toll policy. The last set of tests applied the optimization algorithm to the test (tolls on all links simultaneously and tolls on each link individually). The optimization algorithm successfully solved all test problems.

Despite encouraging results from these tests, further research is still required in order to make the algorithm work efficiently with a large scale application. Firstly, although the theory of the probit model suggests that all routes will always be used, in practice some routes may have a very small probability of being used, and these routes will be eliminated from the choice set due to the limitation of machine precision. In the current algorithm, we assume a fixed set of predetermined used routes, even when the toll is varied. This assumption can be relaxed easily within the iterative procedure to allow the set of used routes to be changed dynamically with the toll level, updating the route set at each iteration. The second issue is concerned with the computational burden of the calculation of the probit SUE. A more efficient algorithm exploiting other estimation techniques of the multi-dimensional integral is being investigated in order to increase the efficiency of the algorithm in solving a large scale SUE problem. Last but not least, we wish to explore the development of the optimization algorithm itself, aiming to improve it by better exploiting the structure of the problem, or through alternative reformulations of the problem.

Acknowledgements. This research has been supported by the UK Engineering and Physical Sciences Research Council grant GR/R53876/01. We also would like to thank Mike Maher and Andrea Rosa for useful discussions and suggestions on this work, as well as the two anonymous referees for their constructive comments and suggestions.

References

[AK5] Akamatsu, T., Kuwahara, M.: Optimal Toll Pattern on a Road Network under Stochastic User Equilibrium with Elastic Demand. Proceeding of the 5th WCTR Volume 1, 259–273 (1989)

[BMW56] Beckmann M.J., McGuire C.B., Winsten C.B.: Studies in the Economics of Transportation. Yale University Press, New Haven, Conn. (1956)

[BI97] Bell, M.G.H., Iida, Y.: Transportation Network Analysis. John Wiley & Sons, Chichester, England (1997)

[BDK86] Ben-Akiva, M., De Palma, A., Kanaroglu, P.: Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates. Transportation Science, **20(2)**, 164–181 (1986)

[CB02] Cantarella, G.E., Binetti, M.G.: Stochastic Assignment with Gammit Path Choice Models. In: Patriksson, M., Labbé, M. (eds) Transportation Planning: State of the Art. Kluwer, Dordrecht, Netherlands, 53–68 (2002)

[CC95] Cantarella, G.E., Cascetta, E.: Dynamic Processes and Equilibrium in Transportation Networks: Towards a Unifying Theory. Transportation Science, **29(4)**, 305–329 (1995)

[Cla61] Clark, C.E.: The greatest of a finite set of random variables. Operations Research, **9**, 145–162 (1961)

[CW02] Clark, S.D., Watling, D.P.: Sensitivity analysis of the probit-based stochastic user equilibrium assignment model. Transportation Research, **36B**, 617–635 (2002)

[Dag79] Daganzo, C.: Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press Inc, New York (1979)

[DS77] Daganzo, C.F., Sheffi, Y.: On stochastic models of traffic assignment. Transportation Science, **11(3)**, 253–274 (1977)

[Dav94] Davis, G.A.: Exact local solution of the continuous network design problem via stochastic user equilibrium. Transportation Research, **28B**, 61–75 (1994)

[Fis80] Fisk, C.: Some Developments in Equilibrium Traffic assignment. Transportation Research, **14B(3)**, 243–255 (1980)

[GP01] Gentile, G., Papola, N.: Network design through sensitivity analysis and singular value decomposition. Paper presented at TRISTAN IV, San Miguel, Azores, June 13^{th}–19^{th} (2001)

[HSD82] Horowitz, J.L., Sparmann, J.M., Daganzo, C.F.: An investigation of the accuracy of the Clark approximation for the multinomial probit model. Transportation Science, **16(3)**, 382–401 (1982)

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

[Lan84] Langdon, M.G.: Improved algorithms for estimating choice probabilities in the multinomial probit model. Transportation Science, **18(3)**, 267–299 (1984)

[LLPR01] Larsson, T., Lundgren, J.T., Patriksson, M., Rydergren, C.: Most likely traffic equilibrium route flows - analysis and computation. Equilibrium Problems & Variational Methods: International Workshop in Memory of Marino De Luca, Taormina, Italy, December (1998).

[LPR96] Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press (1996)

[MH97] Maher, M.J., Hughes, P.C.: A probit-based stochastic user equilibrium assignment model. Transportation Research, **31B**, 341–355 (1997)

[MHK99] Maher, M.J., Hughes, P.C., Kim, K.S.: New algorithms for the solution of the stochastic user equilibrium assignment problem with elastic demand. Proceedings of the 14^{th} International Symposium on Transportation and Traffic Theory, Jerusalem, Israel (1999)

[MLSS02] May, A.D., Liu, R., Shepherd, S.P., Sumalee, A.: The impact of cordon design on the performance of road pricing schemes. Transport Policy, **9**, 209–220 (2002)

[NDF02] Nielsen, O.A., Daly, A., Frederiksen, R.D.: A Stochastic Route Choice Model for Car Travellers in the Copenhagen Region. Networks and Spatial Economics, **2(4)**, 327–346 (2002)

[PR02] Patriksson, M., Rockafellar, R.T.: A Mathematical Model and Descent Algorithm for Bilevel Traffic Management. Transportation Science, **36(3)**, 271–291 (2002)

[PR03] Patriksson, M., Rockafellar, R.T.: Sensitivity Analysis of Aggregated Variational Inequality Problems, with Application to Traffic Equilibria. Transportation Science, **37(1)**, 56–68 (2003)

[PB99] Prashker, J.N., Bekhor, S.: Stochastic User-Equilibrium Formulations for Extended-Logit Assignment Models. Transportation Research Record, **1676**, 145–151 (1999)

[SNR01] Santos, G., Newbery, D., Rojey, L.: Static Versus Demand-Sensitive Models and Estimation of Second-Best Cordon Tolls: An Exercise for Eight English Towns. Transportation Research Record, **1747** (2001)

[She85] Sheffi, Y.: Urban Transportation Networks. Prentice Hall, New Jersey (1985)

[SP81] Sheffi, Y., Powell, W.B.: A Comparison of Stochastic and Deterministic Traffic Assignment over Congested Networks. Transportation Research, **15B(1)**, 53–64 (1981)

[SS04] Shepherd, S.P., Sumalee, A.: A Genetic Algorithm Based Approach to Optimal Toll Level and Location Problems. Networks and Spatial Economics, **4**, 161–179 (2004)

[Smi79] Smith, M.J.: The Existence, Uniqueness and Stability of Traffic Equilibria. Transportation Research, **13B**, 295–304 (1979)

[SEL94] Smith, T.E., Eriksson, E.A., Lindberg, P.O.: Existence of Optimal Tolls under Conditions of Stochastic User-equilibria. In: Johansson, B., Mattsson, L.G. (eds.) Road Pricing: Theory, Empirical Assessment and Policy. Kluwer Academic Publisherrs, 65–87 (1994)

[Sum04] Sumalee, A.: Optimal Road User Charging Cordon Design: A Heuristic Optimisation Approach. Computer-Aided Civil and Infrastructure Engineering, **19**, 377–392 (2004)

[TF88] Tobin, R.L., Friesz, T.L.: Sensitivity Analysis for Equilibrium Network Flow. Transportation Science, **22(4)**, 242–250 (1988)

[Ver02] Verhoef, E.T.: Second-best congestion pricing in general networks. Heuristic algorithms for finding second-best optimal toll levels and toll points. Transportation Research, **36B**, 707–729 (2002) .

[Wal61] Walters, A.A.: The Theory and Measurement of Private and Social Cost of Highway Congestion. Econometrica: Journal of the Econometric Society, **29(4)**, 676–699 (1961)

[War52] Wardrop, J.: Some theoretical aspects of road traffic research. Proc. of the Institute of Civil Engineers, **1(2)** (1952)

[Yan99] Yang, H.: System optimum, stochastic user equilibrium and optimal link tolls. Transportation Science, **33(4)**, 354–360 (1999)

[YH98] Yang, H., Huang, H.J.: Principle of marginal-cost pricing: how does it work in a general road network? Transportation Research, **32A(1)**, 45–54 (1998)

Appendix

Table 4. OD potential demand matrix for the test network

O/D	1	5	7
1	-	1125	1050
5	675	-	850
7	1050	850	-

Table 5. Link travel time parameters for the test network

Link Number i	a_i	b_i	c_i	n_i	σ_i
1	0.0125	0.0026515	1800	4.5	0.0125
2	0.16	0	1	1	0.041498
3	0.25	0	1	1	0.041498
4	0.0125	0.0026515	1800	4.5	0.0125
5	0.03	0.03	1100	3	0.03
6	0.033333	0.033333	1100	3.1	0.033333
7	0.03	0.03	1100	3	0.03
8	0.025	0.025	1100	3.2	0.025
9	0.075	0.015909	1100	3.5	0.075
10	0.033333	0.033333	1100	3.1	0.033333
11	0.026667	0.026667	1100	3.1	0.026667
12	0.07625	0.016174	1100	3	0.07625
13	0.8	0	1	1	0.041498
14	0.025	0.025	1100	3.2	0.025
15	0.026667	0.026667	1100	3.1	0.026667
16	0.02	0.02	1100	3.1	0.02
17	0.2	0	1	1	0.041498
18	0.075	0.015909	1100	3.5	0.075
19	0.07625	0.016174	1100	3	0.07625
20	0.02	0.02	1100	3.1	0.02
21	0.0125	0.0026515	1800	4.5	0.0125
22	0.8	0	1	1	0.041498
23	0.2	0	1	1	0.041498
24	0.0125	0.0026515	1800	4.5	0.0125

Links 13 and 22 are the pseudo links for O-D 1-5 and 1-7 respectively. links 2 and 23 are the pseudo links for O-D 5-1 and 5-7 respectively. Links 3 and 17 are the pseudo links for O-D 7-1 and 7-5 respectively.