

Siriphong Lawphongpanich · Donald W. Hearn

An MPEC approach to second-best toll pricing*

Received: February 22, 2003 / Accepted: March 1, 2004

Published online: 7 July 2004 – © Springer-Verlag 2004

Abstract. This paper addresses two second-best toll pricing problems, one with fixed and the other with elastic travel demands, as mathematical programs with equilibrium constraints. Several equivalent nonlinear programming formulations for the two problems are discussed. One formulation leads to properties that are of interest to transportation economists. Another produces an algorithm that is capable of solving large problems and easy to implement with existing software for linear and nonlinear programming problems. Numerical results using transportation networks from the literature are also presented.

Key words. Congestion Pricing – Traffic Equilibrium – Mathematical Programming with Equilibrium Constraints

1. Introduction

Traffic congestion has become part of everyday life in heavily populated metropolitan areas. When it is not feasible, economically or otherwise, to increase the capacity of the transportation network, imposing appropriate tolls on roads can reduce traffic congestion because tolls can encourage travellers (or network users) to seek less direct routes or to travel during a less congested period.

In the literature, the problem of determining tolls to reduce congestion is often referred to as a toll or congestion pricing problem and the formulation of such a problem can be static or dynamic. This paper focuses on the former, for which the problem can be further classified as first and second best. The first-best toll pricing problem assumes that every road or arc in network can be tolled. In this case, transportation economists (see, e.g., Arnott and Small [1]) often prefer tolls that are based on marginal social cost pricing. However, other first-best tolls exist (see, Hearn and Ramana [20]) and several (e.g., Hearn and Ramana [20], Dial [7] and [8], and Hearn and Yildirim [21] and [22]) propose models and methodologies for calculating the first best tolls with various (secondary) objective functions.

On the other hand, the second-best toll pricing problem assumes that, for political and other reasons, some roads are not tollable. Because tolls with this and other types of restrictions do not generally yield the maximum benefit possible, they are referred to as ‘second-best’ (see, e.g., Johansson-Stenman and Sterner [23]). In the transportation literature, many (see, e.g., Yang and Lam [41], Labbé et al. [24], Larsson and Patriksson [25], Brotcorne et al. [6], Ferrari [10], Lim [27], Patriksson and Rockafellar [35], and

S. Lawphongpanich, D.W. Hearn: Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611-6595, USA. e-mail: {Lawhong, Hearn}@ise.ufl.edu

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

Verhoef [38]) model the second-best toll pricing problem either as a bilevel optimization problem (see, e.g., Shimizu et al. [37] and Bard [3]) or a mathematical program with equilibrium constraints (see, e.g., Luo et al. [28] and Outrata et al. [34]). Several (e.g., Yang and Lam [41], Labbé et al. [24], Brotcorne et al. [6], Ferrari [10], Patriksson and Rockafellar [35], and Verhoef [39]) have proposed algorithms to solve the second-best problem and some provide numerical results on small to medium sized networks.

In this paper, our focus is on using results from the MPEC literature to develop equivalent nonlinear programming formulations for the second-best problem with two goals. One is to establish properties for the second-best tolls of interest to transportation economists and the other is to consider an algorithm that is a natural consequence of one nonlinear programming formulation. Some (e.g., Bazaraa et al. [4]) refer to the algorithm as a cutting plane algorithm and others (e.g., Migdalas [33]) prefer to call it the Benders' scheme. In the transportation literature, Marcotte [30] proposes it as an algorithm for a network design problem. Although the example in the Appendix shows that its master problem does not satisfy the Mangasarian-Fromovitz constraint qualification [29], the algorithm is still appealing for several reasons. One is that it can be implemented with existing software for linear and nonlinear programs. The other is the fact that the algorithm with the modifications described herein converges to stationary points of the real-world problems we tested.

For the remainder, Section 2 presents a general MPEC formulation that includes the second-best problem as a special case. Section 3 discusses equivalent nonlinear programming formulations for the MPEC. From these equivalent formulations, Section 4 establishes properties of second-best toll vectors and Section 5 describes and investigates properties of the algorithm for solving the second-best problem. Section 6 presents computation results for two transportation networks from the literature—Sioux Falls, North Dakota, and Hull, Canada. Finally, Section 7 concludes the paper.

2. Formulation

Consider a mathematical program with equilibrium constraints of the form:

$$\begin{aligned}
 \text{P-VI:} \quad & \min_{(p, \pi)} f(p) \\
 & \text{s.t. } p \in P \\
 & \quad \pi \in \Pi \\
 & \quad (g(p) + \pi)^T (q - p) \geq 0, \quad \forall q \in P
 \end{aligned} \tag{1}$$

where $f(p)$ is a continuously differentiable function on a bounded polyhedron $P \subseteq R^n$, $g(p)$ is a continuous mapping from R^n to R^n , and $\Pi \subseteq R^n$ is a compact set. Equation (1) requires p to be a solution of the variational inequality (VI) defined by the mapping $g(p) + \pi$ and the set P , i.e., p solves $\text{VI}[g(p) + \pi, P]$.

Below are two instances of P-VI that are of interest. One is the second-best toll pricing (SBTP) problem with elastic travel demands and the other is the version with fixed demands.

2.1. Second-best toll pricing problem with elastic demands

To state the problem, let

- a be an index for arcs or links in the network.
- k be an index for OD pairs. It is also convenient to refer to an OD pair as (p, q) , where p and q denote the origin and destination nodes, respectively.
- K be the set containing indices of all OD pairs.
- x^k be the vector of link flows for the k^{th} OD pair.
- t be the travel demand vector whose element, t_k , is the demand for the k^{th} OD pair.
- E_k be a vector that defines the origin and destination nodes for the k^{th} OD pair. In particular, if p and q denote the origin and destination of the k^{th} OD pair, then $E_k = e_p - e_q$, where e_p and e_q are the p^{th} and q^{th} unit vectors, respectively.
- v be the aggregate flow vector (or traffic volume), i.e., $v = \sum_{k \in K} x^k$.
- $s(v)$ be the travel time or cost vector whose element, $s_a(v)$, denotes the travel time for link a . To ensure the existence of a user equilibrium flow, assume that each $s_a(v)$ is continuous. In addition, $\nabla s(v)$ represents the Jacobian of $s(v)$.
- $w(t)$ be the inverse demand vector whose element, $w_k(t_k)$, denotes the inverse travel demand function for the k^{th} OD pair. For the same reason as above, assume that each $w_k(t_k)$ is continuous.
- A be the node-arc incidence matrix for the network. To ensure feasibility, assume that the network contains at least one directed path from p to q for every OD pair (p, q) .
- β be a toll vector.
- Y be the set of arcs that cannot be tolled or the set of *non-tollable* arcs.

Using the above notation, the set of all feasible volume-demand vector, (v, t) , can be stated as follows:

$$V = \left\{ (v, t) : v = \sum_{k \in K} x^k, Ax^k = E_k t_k, x^k \geq 0, t_k \geq 0, \forall k \in K \right\}.$$

We assume that V is bounded. In practice, optimal solutions to the problems described herein are always bounded. (See also Florian and Hearn [13].)

When the objective function of the optimization below approaches ∞ as $\|(v, t)\| \rightarrow \infty$, V can be considered as a bounded polyhedron. Otherwise, we assume that V implicitly includes sufficiently large upper bounds for x_a^k and t_k for all $k \in K$ and a .

Given the above definitions and assumptions, below is one formulation of SBTP when the demand for each OD pair is elastic:

$$\begin{aligned} \text{ED-VI:} \quad & \min_{(v, t, \beta)} s(v)^T v - \sum_{k \in K} \int_0^{t_k} w_k(z) dz \\ \text{s.t.} \quad & (v, t) \in V, \\ & \beta_a \geq 0, \quad \forall a \notin Y \\ & \beta_a = 0, \quad \forall a \in Y \\ & (s(v) + \beta)^T (u - v) - w(t)^T (d - t) \geq 0, \quad \forall (u, d) \in V \end{aligned}$$

Consider the objective function of ED-VI. The expression $s(v)^T v$ in the first term is the total travel time or cost associated with making $\sum_{k \in K} t_k$ trips. For the second term, some

transportation economists (e.g., Verhoef [38]) view the expression $\sum_{k \in K} \int_0^{t_k} w_k(z) dz$ as the benefit gained from making the trips. Subtracting the cost from the benefit yields an expression for a net user benefit, whose value should be maximized. Equivalently, ED-VI minimizes the negative of the net user benefit instead.

For the constraints, the first ensures that the volume-demand vector is feasible. The second set requires tolls on tollable arcs to be nonnegative. The third forces tolls on non-tollable arcs to be zero. Finally, the last set guarantees that the volume-demand vector solves a VI that represents a tolled user equilibrium condition (see, e.g., Hearn and Yildirim [21]).

In the transportation science literature, it is common to assume that each demand function has an inverse because it leads to a mathematical program whose solution satisfies a traffic equilibrium condition (see, e.g., Florian and Hearn [13]). There are also other interpretations for the objective function in ED-VI as well as other alternatives. Gartner [18] and Zhang and Ge [43] offer different interpretations for the objective function based on economic concepts such as user surplus, utility to consumers, private costs, social cost, etc. Yang and Bell [40] propose two alternative objective functions. One maximizes the total realized travel demand, i.e., $\sum_{k \in K} t_k$, and the other maximizes the consumer surplus, i.e., $\sum_{k \in K} \int_0^{t_k} w_k(z) dz - \sum_{k \in K} c_k t_k$, where c_k is the generalized travel cost for OD pair k . With respect to ED-VI, c_k represents the travel time plus the tolls along a utilized path between OD pair k . (See also Zhang and Ge [43] for other possibilities.)

Fisk and Boyce [15] offer an alternative variational inequality formulation for an elastic demand traffic equilibrium without assuming that each demand function has an inverse. However, it is not clear that this formulation leads to a mathematical program whose objective function has an economic interpretation similar to above.

When every arc is tollable ($Y = \emptyset$), ED-VI reduces to the first-best toll pricing problem, an easier problem to solve. In particular, consider the elastic demand traffic assignment problem with the system objective (see, e.g., Florian and Hearn [13]) or the *system problem* with elastic demands:

$$\text{S-OPT: } (v^S, t^S) = \arg \max_{(v,t)} \left\{ \sum_{k \in K} \int_0^{t_k} w_k(z) dz - s(v)^T v : (v, t) \in V \right\}.$$

When $\beta_{\text{MSCP}} = \nabla s(v^S)^T v^S$ is nonnegative, it is easy to show that the triplet $(v^S, t^S, \beta_{\text{MSCP}})$ is an optimal solution to ED-VI. Arnott and Small [1] refers to β_{MSCP} as the marginal social cost pricing (MSCP) toll vector. For other possible first-best toll vectors, see Hearn and Yildirim [21].

On the other hand, when no arc is tollable, i.e., $\beta_a = 0$ for every a , every solution to ED-VI, e.g., (v^U, t^U) , must satisfy the following VI which represents a user equilibrium condition:

$$\text{U-OPT: } s(v^U)^T (u - v^U) - w(t^U)^T (d - t^U) \geq 0, \quad \forall (u, d) \in V.$$

When $s(v)$ is a gradient vector of a function, the above VI problem is equivalent to an optimization problem known in the literature (see, e.g., Florian and Hearn [13]) as the elastic demand traffic assignment problem with the user objective or, simply, the *user problem* with elastic demands.

To establish a relationship between the three problems (ED-VI, S-OPT, and U-OPT), let (v^*, t^*, β^*) denote a global optimal solution to ED-VI and $f(v, t) = s(v)^t v - \sum_{k \in K} \int_0^{t^k} w_k(z) dz$. Then, the following holds:

$$f(v^U, t^U) \geq f(v^*, t^*) \geq f(v^S, t^S).$$

2.2. Second-best Toll Pricing Problem with fixed demands

When the travel demands are fixed, the set of all feasible traffic volume or flow vectors becomes

$$V = \left\{ v : v = \sum_k x^k, Ax^k = b_k, x^k \geq 0, \forall k \in K \right\},$$

where b_k is a vector compatible with A and has exactly two non-zero components, one equals to the demand, d_k , for OD pair k and the other equals $-d_k$. Then, SBTP with fixed demands can be written as follows:

$$\begin{aligned} \text{FD-VI:} \quad & \min_{(v, \beta)} s(v)^T v \\ & \text{s.t. } v \in V, \\ & \beta_a \geq 0, \quad \forall a \notin Y \\ & \beta_a = 0, \quad \forall a \in Y \\ & (s(v) + \beta)^T (u - v) \geq 0, \quad \forall u \in V \end{aligned}$$

As explained earlier, the objective function of FD-VI represents the total travel time and the constraints have similar interpretation as those in ED-VI.

As in the elastic demand case, below are the system and user problems with fixed demands (see, e.g., Florian and Hearn [13]):

$$\text{S-OPT:} \quad v^S = \arg \min_v \{s(v)^T v : v \in V\}$$

$$\text{U-OPT:} \quad s(v^U)^T (u - v^U) \geq 0, \quad \forall u \in V.$$

Similar to before, the following holds:

$$s(v^U)^T v^U \geq s(v^*)^T v^* \geq s(v^S)^T v^S$$

where (v^*, β^*) is a global optimal solution to FD-VI.

3. Equivalent formulations

To establish the results and motivate the algorithm in latter sections, this section presents three nonlinear programs equivalent to P-VI. Two of these follow from the development in Luo et al. [28] and the other relies on the extreme point representation of a bounded convex polyhedron. To state these nonlinear programs, assume that P is a bounded set of the form $\{p : Ap = b, p \geq 0\}$ and $g(p)$ is a strictly monotone mapping. These two assumptions imply that the VI in (1) has a solution for all π (see, e.g., Facchinei and Pang [9]).

To motivate the first equivalent nonlinear program, observe that any p feasible to P-VI must solve the VI in (1). This implies that there exist multipliers $\lambda \geq 0$ and ξ satisfying the following KKT conditions (see, e.g., Proposition 1.2.1 in Facchinei and Pang [9]):

$$\begin{aligned} (g(p) + \pi) - A^T \xi - \lambda &= 0 \\ p^T \lambda &= 0. \end{aligned}$$

Using the fact that both λ and p are nonnegative and $A^T p = b$, λ can be eliminated and the KKT conditions can be written more compactly as (see, e.g., Hearn and Ramana [20])

$$\begin{aligned} A^T \xi &\leq g(p) + \pi \\ b^T \xi &= (g(p) + \pi)^T p. \end{aligned}$$

Thus, using these two conditions to describe p that is feasible to (1), P-VI can be stated as

$$\begin{aligned} \text{P-KKT:} \quad & \min_{(p, \pi, \xi)} f(p) \\ & \text{s.t. } p \in P \\ & \pi \in \Pi \\ & A^T \xi \leq g(p) + \pi \\ & b^T \xi = (g(p) + \pi)^T p. \end{aligned}$$

Theoretically, the equivalence between P-KKT and P-VI follows from Theorem 1.3.5 in Luo et al. [28] and the fact that P-VI satisfies the sequentially bounded constraint qualification or SBCQ. To establish SBCQ for P-VI under the above assumptions, let $M(p, \pi)$ represents the set of all multiplier vectors ξ that satisfy the above KKT conditions at the point (p, π) , i.e.,

$$M(p, \pi) = \left\{ \xi : A^T \xi \leq g(p) + \pi, b^T \xi = (g(p) + \pi)^T p \right\}.$$

Then, the desired result follows from the following two theorems.

Theorem 1. *Let F denote the feasible region of P-VI. Then, the set $M(p, \pi) \neq \emptyset$ for all $(p, \pi) \in F$.*

Proof. The result follows immediately from the strong duality theorem and the fact that p solves VI $[g(p) + \pi, P]$ if and only if it solves the linear program $\min\{(g(p) + \pi)^T q : Aq = b, q \geq 0\}$. (See, e.g., Facchinei and Pang [9].) \square

Theorem 2. *For any sequence $\{p^n, \pi^n\} \subset F$, there exists a $\xi^n \in M(p^n, \pi^n)$ such that ξ^n is bounded for all n .*

Proof. To simplify our notation, we eliminate the superscript n from ξ^n , p^n , and π^n . By adding a slack vector, α , and replacing ξ with $(\xi^+ - \xi^-)$, the conditions defining $M(p, \pi)$ can be represented more compactly as follows:

$$\begin{aligned} Qr &= q \\ r &\geq 0, \end{aligned}$$

where $r = (\xi^+, \xi^-, \alpha)$, Q is a matrix representing the coefficients of (ξ^+, ξ^-, α) on the left hand side of the system defining $M(p, \pi)$, and q is a vector whose components are $(g(p) + \pi)$ and $(g(p) + \pi)^T p$.

Assume without loss of generality that Q has full row rank. Then, Theorem 1 ensures that the set $M(p, \pi) = \{q : Qr = q, r \geq 0\}$ is nonempty. Moreover, $M(p, \pi)$ has at least one extreme point solution of the form $r_B = Q_B^{-1}q$ and $r_N = 0$, where Q and r have been decomposed into $[Q_B : Q_N]$ and $[r_B : r_N]$ with Q_B being a nonsingular submatrix. (See, e.g., Theorem 2.6.5 in Bazaraa et al. [4].) Because $g(p)$ and π are bounded for all $(p, \pi) \in F$, the vectors q and, consequently, r are bounded. \square

Theorems 1 and 2 establish that SBCQ holds for P-VI. Consequently (see Theorem 1.3.5 in Luo et al. [28]), P-VI and P-KKT are equivalent in that a global optimal solution to one problem yields a global optimal solution to the other. In addition, Theorem 1.3.4 in Luo et al. [28] implies that the feasible region of P-VI is closed and global minimizers to both P-VI and P-KKT exist.

To obtain the second equivalent nonlinear program, recall that P is a bounded polyhedron. Thus, it is natural to consider representing P as a convex combination of its extreme points, of which there are finitely many. In other words, let q^i denote the i^{th} extreme point of P . Then, for any $p \in P$, there exists $\lambda_i \in [0, 1]$ such that $\sum_{i=1}^n \lambda_i = 1$ and $p = \sum_{i=1}^n \lambda_i q^i$, where n is the number of extreme points for P . Consequently, (1) is equivalent to the following:

$$(g(p) + \pi)^T (q^i - p) \geq 0, \forall i = 1, \dots, n. \quad (2)$$

If (2) holds, then (1) also holds because every inequality in (1) can be represented as a convex combination of inequalities in (2). Conversely, if (1) holds for all $q \in P$, then it must hold for every extreme point of P , i.e., (2) also holds. Thus, using this extreme point representation, P-VI becomes

$$\begin{aligned} \text{P-EX:} \quad & \min_{(p, \pi)} f(p) \\ & \text{s.t. } p \in P \\ & \quad \pi \in \Pi \\ & \quad (g(p) + \pi)^T (q^i - p) \geq 0, \forall i = 1, \dots, n \end{aligned}$$

The last equivalent formulation uses the regularized gap function in the same manner as described in Marcotte and Zhu [31]. Consider the following regularized gap function (see, e.g., Auchmuty [2] and Fukushima [16]):

$$G(p, \pi) = \max_q \left\{ (g(p) + \pi)^T (p - q) - \frac{\alpha}{2} \|p - q\|^2 : q \in P \right\}$$

where $\alpha > 0$ is a fixed scalar. As defined, $G(p, \pi) \geq 0$, $\forall p \in P, \pi \in \Pi$ and $G(p, \pi)$ is continuously differentiable. In addition, $G(p, \pi) = 0$ if and only if p solves VI $[g(p) + \pi, P]$. Thus, P-VI can be equivalently written as

$$\begin{aligned} \text{P-Gap:} \quad & \min_{(p, \pi)} f(p) \\ & \text{s.t. } p \in P \\ & \quad \pi \in \Pi \\ & \quad G(p, \pi) \leq 0. \end{aligned}$$

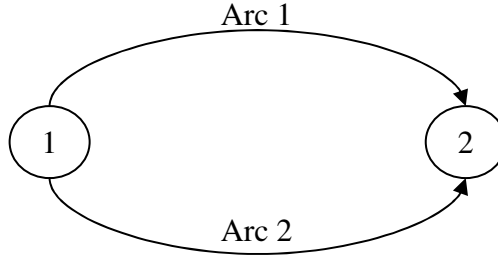


Fig. 1. A two-arc problem

Observe that the feasible region of P-Gap is compact because $G(p, \pi)$ is bounded from below. Thus, the continuity of $f(p)$ ensures that a global minimizer of P-Gap exists. In addition, other gap or merit functions in the literature (see, e.g., Fukushima [17] and Facchinei and Pang [9]) can also be used in place of the above regularized gap function.

4. Properties of second-best tolls

This section establishes two properties of interest in transportation economics using P-KKT. Both properties assume that each travel demand is elastic. The first relates the second-best tolls to the marginal social cost prices. The second property demonstrates that the total toll revenue must be constant. This is an extension of the result in Hearn and Yildirim [21] to the second-best problem.

When each travel demand is elastic, many transportation economists (see, e.g., McDonald [32] and Verhoef [38]) have observed in small examples that some of the marginal social costs from non-tollable arcs must be shifted to those that are tollable in order to maximize the net user benefit. For example, consider the two-arc problem in Figure 1, where Arc 1 is tollable and Arc 2 is not.

In addition, there is only one OD pair and its inverse demand function is $w(t)$. The travel cost functions are $s_1(v_1)$ and $s_2(v_2)$ where v_1 and v_2 represent the amount of flow on the two arcs.

Although an optimal solution for this problem can be expressed in a simpler form (see, Yildirim [42]), several authors (see, e.g., Verhoef [38]) prefer to express the optimal toll for Arc 1 as

$$\beta_1 = s'_1(v_1)v_1 + \frac{w'(t)}{s'_2(v_2) - w'(t)} s'_2(v_2)v_2. \quad (3)$$

This expression shows that part of the marginal social cost, $s'_2(v_2)v_2$, on the non-tollable Arc 2 must be shifted to Arc 1 in order to maximize the net user benefit. This notion of shifting marginal social costs from non-tollable arcs to the tollable ones is appealing to transportation economists who have long argued for the use of marginal social cost prices as congestion tolls (see, e.g., Arnott and Small [1]). Therefore, it is of interest to generalize (3) to larger networks. However, it would be unreasonable to expect that the second-best tolls can be expressed in a closed-form fashion and the shift as evident as in (3). Instead, the two results below provide a mechanism that illustrates these shifts

indirectly using the KKT multipliers. One result assumes that an equivalent formulation of the second-best problem satisfies the strong stationarity conditions in Scheel and Scholtes [36] and the other relies on the ‘tightened’ nonlinear programming formulation described in the same paper.

When applied to the second-best toll pricing problem with elastic demands, P-KKT with the complementarity constraints written explicitly can be stated as follows:

$$\begin{aligned}
 \text{ED-KKT:} \quad & \min_{(v,t,\beta,\rho)} s(v)^T v - \sum_{k \in K} \int_0^{t_k} w_k(z) dz \\
 \text{s.t.} \quad & (v, t) \in V \\
 & \beta_a \geq 0, \quad \forall a \notin Y \\
 & \beta_a = 0, \quad \forall a \in Y \\
 & (s(v) + \beta) - A^T \rho^k = \alpha^k, \quad \forall k \in K \\
 & -w_k(t_k) + E_k^T \rho^k = \pi_k, \quad \forall k \in K \\
 & (\alpha^k)^T x^k = 0, \quad \forall k \in K \\
 & \pi^T t = 0 \\
 & \alpha^k \geq 0, \pi_k \geq 0, \quad \forall k \in K.
 \end{aligned}$$

Scheel and Scholtes [36] show that the Mangasarian-Fromovitz constraint qualification [29] or MFCQ, does not hold for ED-KKT. On the other hand, it is well known that the complementarity constraints in ED-KKT can be equivalently replaced by $(\alpha^k)^T x^k \leq 0$ and $\pi^T t \leq 0$. When these are combined with other constraints, ED-KKT can be written more compactly as

$$\begin{aligned}
 \text{ED-KKT1:} \quad & \min_{(v,t,\beta,\rho)} s(v)^T v - \sum_{k \in K} \int_0^{t_k} w_k(z) dz \\
 \text{s.t.} \quad & (v, t) \in V \\
 & \beta_a \geq 0, \quad \forall a \notin Y \\
 & \beta_a = 0, \quad \forall a \in Y \\
 & (s(v) + \beta) \geq A^T \rho^k, \quad \forall k \in K \\
 & w_k(t_k) \leq E_k^T \rho^k, \quad \forall k \in K \\
 & (s(v) + \beta)^T v \leq w(t)^T t.
 \end{aligned}$$

Below, Theorem 3 establishes a relationship between the second-best tolls and marginal social cost prices assuming that the strong stationarity conditions in Scheel and Scholtes [36] hold for ED-KKT. As shown in Fletcher et al. [12], this in turn implies that there exists a set of KKT multipliers for ED-KKT1.

Theorem 3. *Let $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ be a global optimal solution to the ED-KKT that is strongly stationary. If the multiplier associated with the last constraint is positive, then $\bar{\beta}$ is well defined and, for any k , can be expressed as follows:*

$$\bar{\beta} = \frac{1}{\theta} \left[\delta^k - (1 + \theta)[s(\bar{v}) + \nabla s(\bar{v})^T \bar{v}] - A^T \lambda^k + \nabla s(\bar{v})^T \sum_{k \in K} \psi^k \right].$$

Proof. As shown in Fletcher et al. [12], the strong stationarity of $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ implies that there exist nonnegative $\psi^k, \xi_k, \delta^k, \sigma_k, \tau, \theta$ and unrestricted λ^k that satisfy the KKT conditions associated with ED-KKT1, i.e.,

$$\begin{aligned}
(1 + \theta)[s(\bar{v}) + \nabla s(\bar{v})^T \bar{v}] + A^T \lambda^k - \nabla s(\bar{v})^T \sum_{k \in K} \psi^k + \theta \bar{\beta} - \delta^k &= 0, \forall k \in K \\
-w_k(\bar{t}_k) - E_k^T \lambda^k + w'_k(\bar{t}_k) \xi_k - \theta[w_k(\bar{t}_k) + w'_k(\bar{t}_k) \bar{t}_k] - \sigma_k &= 0, \forall k \in K \\
\theta \bar{v} - \sum_{k \in K} \psi^k - \tau &= 0, \forall k \in K \\
A \psi^k - E_k \xi_k &= 0, \forall k \in K \\
[s(\bar{v}) + \bar{\beta} - A^T \bar{\rho}^k]^T \psi^k &= 0, \forall k \in K \\
[w_k(\bar{t}_k) - E_k^T \bar{\rho}^k] \xi_k &= 0, \forall k \in K \\
([s(v) + \beta]^T v - w(t)^T t) \theta &= 0 \\
(x^k)^T \delta^k &= 0, \forall k \in K \\
\bar{t}_k \sigma_k &= 0, \forall k \in K \\
\bar{\beta}^T \tau &= 0.
\end{aligned}$$

The first condition yields that, for any k ,

$$\bar{\beta} = \frac{1}{\theta} \left[\delta^k - (1 + \theta)[s(\bar{v}) + \nabla s(\bar{v})^T \bar{v}] - A^T \lambda^k + \nabla s(\bar{v})^T \sum_{k \in K} \psi^k \right].$$

By the hypothesis, $\theta > 0$. So, the above expression for $\bar{\beta}$ is well defined and involves MSCP toll vectors. \square

To illustrate Theorem 3, consider the two-arc problem in Figure 1 and let $s_1(v_1) = v_1$, $s_2(v_2) = v_2 + 2$, and $w(t) = 9 - t/2$. The set of all feasible volume-demand vectors reduces to $V = \{(v_1, v_2, t) : v_1 + v_2 = t; v_1, v_2, t \geq 0\}$ and the ED-KKT1 formulation for the problem is:

\max	$9t - \frac{t^2}{4} - v_1^2 - v_2^2 - 2v_2$		Multipliers
s.t.	$v_1 + v_2 - t$	$= 0$	λ
	$\rho - v_1 - \beta_1$	≤ 0	ψ_1
	$\rho - v_2 - 2$	≤ 0	ψ_2
	$-\rho + (9 - \frac{t}{2})$	≤ 0	ξ
	$(v_1 + \beta_1)v_1 + (v_2 + 2)v_2 - (9 - \frac{t}{2})t$	≤ 0	θ
	v_1, v_2, t, β_1	≥ 0	

Using a nonlinear programming software, an optimal solution to this problem is $(\bar{v}_1, \bar{v}_2, \bar{t}) = (3.3636, 3.5455, 6.9091)$, $\bar{\beta}_1 = 2.1818$, $\bar{\rho} = 5.5455$ and the net user benefit = 19.2727.

At $(\bar{v}_1, \bar{v}_2, \bar{t})$, the multipliers for the KKT conditions in Theorem 3 are not unique. In fact, the set of multipliers is unbounded because ED-KKT1 violates MFCQ. (See, e.g., Gauvin [19].) However, solving a linear program that maximizes θ over the set of feasible multipliers with appropriate bounds yields the following: $\lambda = -12.2727$, $\psi_1 = 3.3636$, $\psi_2 = 5.9091$, $\xi = 9.2727$, $\theta = 1.0$, and the remaining multipliers associated with nonnegativity constraints (i.e., $\delta_1, \delta_2, \sigma$, and τ) are all zero. Below is the expression for β_1 using these multipliers:

$$\begin{aligned}
\begin{bmatrix} \bar{\beta}_1 \\ \bar{\beta}_2 \end{bmatrix} &= \frac{1}{\theta} \left\{ \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} - (1 + \theta) \begin{bmatrix} s_1(\bar{v}_1) + s'_1(\bar{v}_1)^T \bar{v}_1 \\ s_2(\bar{v}_2) + s'_2(\bar{v}_2)^T \bar{v}_2 \end{bmatrix} - \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} s'_1(\bar{v}_1) \psi_1 \\ s'_2(\bar{v}_2) \psi_2 \end{bmatrix} \right\} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 2 \begin{bmatrix} 6.7272 \\ 9.0909 \end{bmatrix} + 12.2727 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 3.3636 \\ 5.9091 \end{bmatrix} = \begin{bmatrix} 2.1819 \\ 0 \end{bmatrix}. \quad (4)
\end{aligned}$$

Using (3),

$$\begin{aligned}\bar{\beta}_1 &= s'_1(\bar{v}_1)\bar{v}_1 + \frac{w'(\bar{t})}{s'_2(\bar{v}_2) - w'(\bar{t})} s'_2(\bar{v}_2)\bar{v}_2 \\ &= 3.3636 - \frac{0.5}{(1 + 0.5)} 3.5455 = 2.1818.\end{aligned}\quad (5)$$

Subject to slight round-off errors, the tolls in (4) and (5) are the same. As explained earlier, the shift in marginal social costs, $s'_i(\bar{v}_i)\bar{v}_i$, in (4) is not as evident as the one shown in (5). Instead, the shift in (4) occurs via the KKT multipliers.

As an alternative to the strong stationarity assumption, it is also possible to use the ‘tightened’ nonlinear programming (TNLP) in Scheel and Scholtes [36] to derive a similar expression for β . At a global optimal solution, $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$, define the following sets of indices:

- a) $\Omega(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) = \{a : s_a(\bar{v}) + \bar{\beta}_a - [A^T \bar{\rho}^k]_a > 0, \forall k\}$ and
 b) $\Gamma(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) = \{k : w_k(\bar{t}_k) - E_k^T \bar{\rho}^k < 0\}$.

In addition, let $\Omega^c(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ and $\Gamma^c(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ denote the complements of $\Omega(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ and $\Gamma(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ with respect to the sets of arcs and commodities, respectively. The theorem below follows from the Slater’s CQ for problems with equality and inequality constraints (see, e.g., Chapter 5 in Bazaraa et al. [4]) and a result in Scheel and Scholtes [36].

Theorem 4. *Let $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ be a global optimal solution to the ED-KKT. If $s_a(v)$ is concave for each $a \in \Omega(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ and linear for each $a \in \Omega^c(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ and $w_k(t_k)$ is convex for each $k \in \Gamma(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ and linear for $k \in \Gamma^c(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$, then $\bar{\beta}$ is well defined and, for any k , can be written as follows:*

$$\bar{\beta}_a = \left[\nabla s(\bar{v})^T \bar{v} + A^T (\lambda^k + \bar{\rho}^k) + \nabla s(\bar{v})^T \tau + \delta^k \right]_a + \Delta_a^k,$$

where $\Delta_a^k = s_a(\bar{v}) + \bar{\beta}_a - [A^T \bar{\rho}^k]_a$.

Proof. After eliminating the variables α^k and π_k , ED-KKT can be written as follows:

$$\begin{aligned}\text{ED-KKT2:} \quad & \min_{(v, t, \beta, \rho)} s(v)^T v - \sum_{k \in K} \int_0^{t_k} w_k(z) dz \\ & \text{s.t.} \quad (v, t) \in V \\ & \quad \beta_a \geq 0, \quad \forall a \notin Y \\ & \quad \beta_a = 0, \quad \forall a \in Y \\ & \quad s(v) + \beta - A^T \rho^k \geq 0, \quad \forall k \in K \\ & \quad w_k(t_k) - E_k^T \rho^k \leq 0, \quad \forall k \in K \\ & \quad (s(v) + \beta - A^T \rho^k)^T x^k = 0, \quad \forall k \in K \\ & \quad (w_k(t_k) - E_k^T \rho^k) t_k = 0. \quad \forall k \in K\end{aligned}$$

At $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$, define the following sets of (active) indices:

$$\begin{aligned}I^1(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) &= \{(a, k) : s_a(\bar{v}) + \bar{\beta}_a = [A^T \bar{\rho}^k]_a\}, \quad I^2(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) = \{(a, k) : \bar{x}_a^k = 0\} \\ I^3(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) &= \{k : w_k(\bar{t}_k) = E_k^T \bar{\rho}^k\}, \quad I^4(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) = \{k : \bar{t}_k = 0\}.\end{aligned}$$

Then, Scheel and Scholtes [36] define the TNLP formulation associated with ED-KKT2 as follows:

TNLP:		Multipliers
$\min_{(v, \bar{t}, \bar{\beta}, \bar{\rho})}$	$s(v)^T v - \sum_{k \in K} \int_0^{t_k} w_k(z) dz$	
s.t.	$Ax^k - E_k t_k = 0,$	$\forall k \in K \quad \lambda^k$
	$\beta_a \geq 0,$	$\forall a \notin Y \quad \tau_a \leq 0$
	$\beta_a = 0,$	$\forall a \in Y \quad \tau_a$
	$s_a(v) + \beta_a - [A^T \rho^k]_a = 0,$	$\forall (a, k) \in I^1(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \psi_a^k$
	$s_a(v) + \beta_a - [A^T \rho^k]_a \geq 0,$	$\forall (a, k) \notin I^1(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \psi_a^k \leq 0$
	$x_a^k = 0,$	$\forall (a, k) \in I^2(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \delta_a^k$
	$x_a^k \geq 0,$	$\forall (a, k) \notin I^2(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \delta_a^k \leq 0$
	$w_k(t_k) - E_k^T \rho^k = 0,$	$\forall k \in I^3(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \xi_k$
	$w_k(t_k) - E_k^T \rho^k \leq 0,$	$\forall k \notin I^3(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \xi_k \geq 0$
	$t_k = 0,$	$\forall k \in I^4(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \sigma_k$
	$t_k \geq 0,$	$\forall k \notin I^4(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho}) \quad \sigma_k \leq 0$
	$v = \sum_{k \in K} x^k.$	η

In addition, they show that $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ solves TNLP at least locally. Under the hypothesis, TNLP satisfies Slater's CQ. So, the multipliers exist and the KKT conditions for TNLP at $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ yield the following expression for any k :

$$\bar{\beta}_a = [\nabla s(\bar{v})^T \bar{v} + A^T (\lambda^k + \bar{\rho}^k) + \nabla s(\bar{v})^T \tau + \delta^k]_a + \Delta_a^k,$$

where $\Delta_a^k = s_a(\bar{v}) + \bar{\beta}_a - [A^T \bar{\rho}^k]_a$. Note that $\Delta_a^k = 0 \forall (a, k) \in I^1(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$. \square

The second property addresses the toll revenue collected from the transportation system. Below, Theorem 5 is an extension in Hearn and Yildirim [21] and shows that this revenue must be constant. When $(\bar{v}, \bar{t}, \bar{\beta}, \bar{\rho})$ solves ED-KKT, $(\bar{\beta}, \bar{\rho})$ must satisfy the following linear system:

$$\begin{aligned} s(\bar{v}) + \beta &\geq A^T \rho, & \forall k \\ w_k(\bar{t}_k) &\leq E_k^T \rho, & \forall k \\ (s(\bar{v}) + \beta)^T \bar{v} &= w(\bar{t})^T \bar{t} \\ \beta_a &= 0, & \forall a \in Y. \end{aligned}$$

Let $W(\bar{v}, \bar{t})$ denote the set of all possible solutions to this system, i.e., $W(\bar{v}, \bar{t})$ is the toll set associated with (\bar{v}, \bar{t}) . Generally, $W(\bar{v}, \bar{t})$ is not a singleton set and the theorem below demonstrates that every toll vector in $W(\bar{v}, \bar{t})$ generates the same revenue.

Theorem 5. $\beta^T \bar{v} = w(\bar{t})^T \bar{t} - s(\bar{v})^T \bar{v}, \forall (\beta, \rho) \in W(\bar{v}, \bar{t})$.

Proof. As in Hearn and Yildirim [21], the expression follows from the third equation of the above system and is constant with respect to (\bar{v}, \bar{t}) . \square

5. Cutting constraint algorithm for P-EX

Consider problem P-EX. Although finite, the number of extreme points for the feasible region of the second-best problem with fixed or elastic demands is extremely large. Therefore, it is natural to generate these extreme points one at a time, each of which produces a constraint that cuts away part of the region not feasible to the original problem. Some (see, e.g., Bazaraa et al. [4]) refer to this type of algorithms as the cutting plane algorithm when the constraints are linear and others (see, e.g., Migdalas [33]) refer to it as the Benders' scheme when the constraints are nonlinear. Marcotte [30] also proposed an algorithm using this scheme for solving a network design problem formulated as an MPEC. When applied to P-EX, this extreme point generation idea leads to the following algorithm:

Cutting constraint algorithm for P-EX

Step 0: Let $q^1 = \arg \min\{g(0)^T q : q \in P\}$. Set $n = 1$ and go to Step 1

Step 1: Solve the following (master) problem:

$$\begin{aligned} (p^n, \pi^n) = \arg \min_{(p, \pi)} f(p) \\ \text{s.t. } p \in P \\ \pi \in \Pi \\ (g(p) + \pi)^T (q^i - p) \geq 0, \forall i = 1, \dots, n \end{aligned}$$

Step 2: Solve the (sub)problem: $q^{n+1} = \arg \min\{(g(p^n) + \pi^n)^T q : q \in P\}$. If $(g(p^n) + \pi^n)^T (q^{n+1} - p^n) \geq 0$, stop and (p^n, π^n) is a solution to P-EX. Otherwise, set $n = n + 1$ and go to Step 1.

Observe that the problems in Steps 0 and 2 are linear because P is a bounded polyhedron. In Step 1, (p^n, π^n) is not feasible to the constraint subsequently generated in Step 2, i.e.,

$$(g(p) + \pi)^T (q^{n+1} - p) \geq 0.$$

In other words, the above constraint cuts away (p^n, π^n) and it is easy to see from this observation that Step 2 generates distinct extreme points. Therefore, the algorithm must stop after a finite number of iterations.

Unfortunately, the master problem in Step 1 at some iterations and, more specifically, P-EX do not satisfy MFCQ. (See a counterexample in the Appendix.) To make it more amenable to standard nonlinear programming software in the next section, we heuristically replace the problem in Step 1 with the following perturbed master problem:

$$\begin{aligned} (p^n, \pi^n) = \arg \min_{(p, \pi)} f(p) \\ \text{s.t. } p \in P \\ \pi \in \Pi \\ (g(p) + \pi)^T (q^i - p) \geq -\epsilon, \forall i = 1, \dots, n \end{aligned}$$

The motivation for this heuristic follows from the following observation. For any $\pi \in \Pi$, let p^π solves $\text{VI}[g(p) + \pi, V]$. Then, $p^\pi \in P$ and $(g(p^\pi) + \pi)^T (q^i - p^\pi) \geq$

$0 > -\epsilon$ for all $i = 1, \dots, n$, i.e., none of the cutting constraints are binding at (p^π, π) . When they are binding, these constraints cause MFCQ to fail. (See the Appendix.)

Moreover, the perturbation need not be done explicitly because all nonlinear programming software allow for some feasibility tolerance. However, the default feasibility tolerance may not be appropriate. Based on our preliminary experiments with practical problems (see the next section), the default feasibility tolerance of, e.g., $10E-6$ in MINOS is too small. In many cases, MINOS would terminate because it cannot find a feasible solution to the master problem. By relaxing the feasibility tolerance to $10E-4$, MINOS was able to solve the perturbed master problem in all cases.

6. Numerical results

To investigate its effectiveness, we implemented the cutting constraint algorithm (CCA) using the algebraic modelling system GAMS [5] on a 300 MHz IBM SP2 computer with 512 MB of RAM. We used MINOS Version 5.51 (as explained above, with $10E-4$ instead of $10E-6$ as the feasibility tolerance) and CPLEX Version 8.1 to solve the master and subproblems, respectively.

Consider first the second-best problem with fixed demands. In this case, the master and subproblems in Steps 1 and 2 of CCA become:

$$\begin{aligned} \text{Master: } \quad & \min_{(v, \beta)} s(v)^T v \\ \text{s.t. } \quad & v \in V = \left\{ v : v = \sum_k x^k, Ax^k = b^k, x^k \geq 0, \forall k \in K \right\} \\ & \beta_a \geq 0, \forall a \notin Y \\ & \beta_a = 0, \forall a \in Y \\ & (s(v) + \beta)^T (u^i - v) \geq 0, \forall i = 1, \dots, n. \end{aligned}$$

$$\begin{aligned} \text{Subproblem : } \quad & u^{n+1} = \arg \min_{(u, x)} (s(v^n) + \beta^n)^T u \\ \text{s.t. } \quad & u = \sum_{k \in K} x^k \\ & Ax^k = b^k, \quad \forall k \in K \\ & x^k \geq 0, \quad \forall k \in K. \end{aligned}$$

Observe that the subproblem decomposes into K problems and each one is solvable as a shortest path problem because there is no arc capacity.

In Step 0 of CCA, our experience indicates that it is more efficient to choose q^1 (or, in the present context, u^1) to be an optimal solution S-OPT, i.e., we set $u^1 = v^S$. Doing so yields in the first iteration the following master problem:

$$\begin{aligned} \text{Master1: } \quad & \min_{(v, \beta)} s(v)^T v \\ \text{s.t. } \quad & v \in V \\ & \beta_a \geq 0, \quad \forall a \notin Y \\ & \beta_a = 0, \quad \forall a \in Y \\ & (s(v) + \beta)^T (v^S - v) \geq 0. \end{aligned}$$

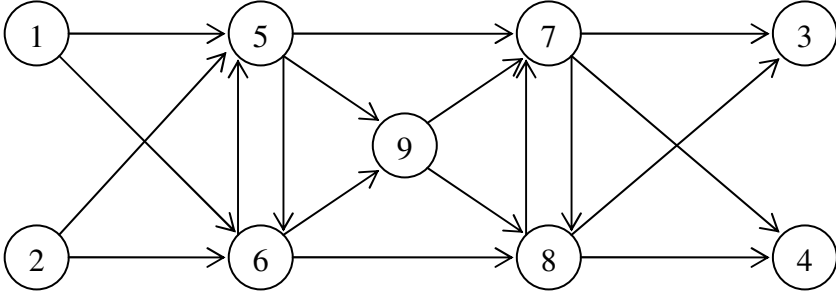


Fig. 2. Nine-Node Network

Note that the S-OPT solution v^S is feasible to Master1. Furthermore, it also yields the smallest objective value (travel delay) because it is a solution to S-OPT. When $v = v^S$, the left hand side of the last constraint in Master1 is zero for any β . Thus, (v^S, β) , for any $\beta \in \{\beta : \beta_a = 0, \forall a \in Y, \beta_a \geq 0, \forall a \notin Y\}$, is an optimal solution to Master1. One obvious choice is to set β to zero. Instead, we solve a capacitated user problem, i.e., we solve U-OPT with v_a^S as a capacity for arc a if it is tollable, and use the optimal dual variables associated with the capacity constraints to form an optimal solution to Master1.

For subsequent iterations, it is often difficult for general nonlinear programming software such as MINOS to find a feasible solution to the master problem, especially for large networks. In our implementation, the pair (\hat{v}, β^n) , where \hat{v} solves $\text{VI}[s(v) + \beta^n, V]$ and β^n is the current toll vector, works well as an initial feasible solution. In addition, the solution to the following ‘line search’ problem is often optimal to the master problem during the early iterations:

$$\begin{aligned}
 \text{Line search:} \quad & \min_{(v, \beta, \lambda)} s(v)^T v \\
 \text{s.t.} \quad & v = \lambda v^{(n-1)} + (1 - \lambda) \hat{v} \\
 & \beta_a \geq 0, \quad \forall a \notin Y \\
 & \beta_a = 0, \quad \forall a \in Y \\
 & (s(v) + \beta)^T (u^i - v) \geq 0, \quad \forall i = 1, \dots, n, \\
 & 0 \leq \lambda \leq 1.
 \end{aligned}$$

In our GAMS implementation, it is convenient to use MINOS to solve the above line search problem.

To illustrate a typical behavior of CCA, consider the nine-node network (shown in Figure 2) from Hearn and Ramana [20]. As the name suggested, this network has 9 nodes, 18 arcs, and 4 OD pairs.

Recall from Section 2.2 that v^S and v^U denote optimal solutions to the system and user problems, respectively. In addition, the objective function $f(v)$ denotes the total travel time or travel delay. For the nine-node problem, $f(v^S) = 2253.9179$ and $f(v^U) = 2455.8699$.

Table 1 displays iterates from CCA in which the only tollable arcs are (7, 3) and (7, 4). CCA requires nine iterations to solve the second-best problem. The column labeled

Table 1. Iterates from the Cutting Constraint Algorithm for the nine-node network with (7, 3) and (7, 4) as the only tollable arcs

Iter.	Master Objective	Relative Gap (%)	Tolled User Equilibrium Travel Delay
1	2253.9179	25.34	2477.1075
2	2281.1793	17.43	2455.8699
3	2306.1945	16.21	2641.3556
4	2341.8896	12.54	2721.1162
5	2362.3970	3.80	2455.9379
6	2379.5725	7.20	2464.3484
7	2412.7163	1.76	2459.5505
8	2420.5433	3.13	2455.5761
9	2451.0617	0.00	2451.0617

‘Relative Gap (%)’ reports the following value at the end of each iteration:

$$\text{Relative Gap} = 100 \times \frac{|(s(v^n) + \beta^n)^T (u^{n+1} - v^n)|}{|f(v^n)|}.$$

In the above ratio, the numerator represents the value of the gap function (see, e.g., Facchinei and Pang [9]) associated with $\text{VI}[s(v) + \beta, V]$ evaluated at the point (v^n, β^n) and measures how well the current solution satisfies the tolled user equilibrium condition in (1). Observe that the master objective values increase as we add more cuts to the problem. For the nine-node problem, the increase is coincidentally monotonic. In general, the master problem is a non-convex problem. Unless the solutions to the master problems are globally optimal, the master objective values may not increase in a monotonic fashion.

Because the master problem is a relaxation of the second-best problem, the objective value of the master problem underestimates the travel delay associated with the current toll vector during each iteration. To provide more accurate travel delays, Table 1 also reports in the last column the travel delay associated with the tolled user equilibrium solution using the current toll vector, β^n . In other words, the last column reports the ‘tolled user equilibrium travel delay.’

As discussed previously, the pair (v^S, β) , where v^S is a solution to S-OPT and $\beta \in \{\beta : \beta_a = 0, \forall a \in Y, \beta_a \geq 0, \forall a \notin Y\}$, is an optimal solution to the master problem in the first iteration. Thus, the master objective value in the first iteration must be the same as $f(v^S)$ and Table 1 reflects this fact in the first row. If β is chosen to be zero, then the tolled user equilibrium travel delay in the first iteration would be equal to $f(v^U) = 2455.8699$, the travel delay at the user solution. In our implementation, we solve the capacitated user problem and let β to be the optimal dual variable associated with the capacity constraints on arcs (7, 3) and (7, 4) instead. In the first iteration, the dual variables for arcs (7, 3) and (7, 4) are 6.1208 and 2.1208. These tolls yield a tolled user equilibrium travel delay of 2477.1075, a slightly higher delay than $f(v^U)$. When the algorithm terminates in iteration 9, the master objective value agrees with the tolled user equilibrium travel delay.

To evaluate the benefit of tolling, we compare the travel delay with and without the tolls. For the nine-node problem, the delay without toll is $f(v^U)$ or 2455.8699 and the one with tolls on arcs (7, 3) and (7, 4) is 2451.0617, a 0.20% reduction in travel delay. For

Table 2. Network Attributes

Network	# of Links	# of Nodes	# of OD Pairs
Sioux Falls	76	24	528
Hull	798	501	158

comparison, the maximum possible reduction in travel delay for the nine-node problem is 8.22%, the percent difference between the travel delays associated with the system and user solutions. In Hearn and Ramana [20], solutions to two tolling pricing problems, one minimizes the number of required toll booths and the other minimizes the toll revenue collected, require tolling 5 and 14 arcs, respectively, to achieve the minimum travel delay of 2253.9179.

To demonstrate CCA's potential for solving realistic problems, we also solved the second-best problem using the two street networks from the transportation science literature, one network is from Sioux Falls, North Dakota (see, LeBlanc et al. [26]) and the other is Hull, Canada (see, e.g., Florian et al. [14]). Table 2 provides the attributes of these two networks.

For Sioux Falls, the travel cost (or delay) function for every arc is of the form: $s_a(v_a) = T_a[1 + q_a(\frac{v_a}{c_a})^4]$. Our initial analysis indicates that the Hull network is not congested and tolling does not lead to any significant improvement in the travel delay. To make the network more meaningful for our study, we modified the travel cost functions for Hull slightly. For example, the travel cost functions for some arcs in the Hull network have the same form as those in Sioux Falls network while others may have linear or constant travel time. For our experiments, we made the travel cost function for every arc to have the same form as those for Sioux Falls. When necessary, the values for T_a , q_a , and c_a were randomly chosen from the intervals, $[0.15, 2.0]$, $[0.5, 1.0]$, and $[100, 1000]$, respectively. The remaining data for Hull are the same as the original.

To select arcs for tolling, we first solve the system and user problems to obtain the system, v^S , and user, v^U , optimal flows, respectively. We then designate arc a as tollable if v_a^U exceeds v_a^S by a given percentage. A different 'excess' percentage usually yields a different number of tollable arcs. We also terminate CCA when the relative gap is less than or equal to one percent.

Tables 3 and 4 summarize the results for Sioux Falls and Hull. For each excess percentage used, these two tables report the resulting number of tollable arcs, the objective value of the master problem at termination, the relative gap achieved, the number of iterations and the amount of CPU time required, and the final tolled user equilibrium travel delay. For Sioux Falls, the travel delays associated with the system and user solutions are 71.9426 and 74.8023, respectively. From Section 2.2, these two delays form the lower and upper bounds for the second-best problem. Observe in Table 3 that both the master objectives and the tolled user equilibrium travel delays are within these two bounds. Because smaller numbers of tollable arcs generally mean less control, the delays in Table 3 increase as the number of tollable arcs decreases. For Hull, the travel delays associated with the system and user solutions are 179.0629 and 186.7203, respectively. In Table 4, the master objectives and the tolled user equilibrium travel delays for Hull exhibit properties similar to those associated with Sioux Falls.

The results for Sioux Falls suggest that problems with a smaller number of tollable arcs are easier to solve. However, the results from Hull indicate otherwise. The CPU

Table 3. Results for Sioux Falls

Excess %	# of Tollable Arcs	Master Objective	Relative Gap (%)	Iter. Req.	CPU Time (sec)	Tolled User Equilibrium Travel Delay
5 %	18	72.1036	0.9354	49	1743.38	72.6238
10 %	12	72.1861	0.9024	36	887.56	72.6293
15 %	4	73.0681	0.7764	14	184.78	73.8787
25 %	2	73.4916	0.4992	10	109.63	74.3196

Table 4. Results for Hull

Excess %	# of Tollable Arcs	Master Objective	Relative Gap (%)	Iter. Req.	CPU Time (sec)	Tolled User Equilibrium Travel Delay
5 %	179	179117.0	≤ 0.0001	16	6384.32	179117.0
10 %	135	179420.4	≤ 0.0001	8	1141.64	179420.4
15 %	93	179987.5	≤ 0.0001	7	1726.50	179987.5
25 %	58	180628.6	≤ 0.0001	7	2060.59	180628.6
50 %	21	181092.0	≤ 0.0001	10	4014.25	181092.0
75 %	12	181315.2	≤ 0.0001	8	3340.12	181315.2
100 %	10	181326.0	≤ 0.0001	11	5333.84	181326.0
200 %	8	181329.3	≤ 0.0001	11	5682.20	181329.3

time for solving the Hull network with only eight tollable arcs is similar to the one with 179 tollable arcs. For every excess percentage, observe also that CCA is able to achieve a nearly zero percent relative gap for Hull. We surmise that this phenomenon may be due in part to the fact that the traffic is still relatively uncongested in the Hull network with our randomly generated data. (Our attempt to double the travel demands for Hull gave similar results.)

We also solved the second-best problem with elastic demands for Sioux Falls and Hull. In the literature, the travel demands for both networks are fixed. For our testing, we assume that each inverse demand function is linear, i.e., $w(t) = a + bt$, where a and b are randomly generated. To determine these two numbers, let t_1 denote the free-flow travel time between an OD pair and t_2 denote the travel time between the same OD pair when the traffic reaches its (fixed demand) user equilibrium. Also, let d_k represent the (fixed) travel demand from the literature. Then, a and b are the intercept and slopes of the line that passes through two points, (t_1, ud_k) and (t_2, d_k) , where u is a uniform random number between 2 and 3.

Using the above random inverse demand functions, the net user benefits at the system and user solutions (i.e., $f(v^S, t^S)$ and $f(v^U, t^U)$) for Sioux Falls are 3093.87 and 2055.54, respectively, and the 34% excess percentage generates 49 tollable arcs. After approximately 6790 seconds, CCA generates a second-best toll vector with a net user benefit of 2968.62, a solution with less than 10E-4 relative gap and approximately 96% of the maximum net user benefit. Similarly, the two net user benefits for Hull are 5075.74 and 3637.46 and the 46% excess percentage generates 117 tollable arcs. After approximately 62606 seconds, CCA generates a second-best toll vector with a net user benefit

of 4059.27, a solution with less than $10E-4$ relative gap and 80% of the maximum net user benefit.

7. Conclusion

In this paper, we formulate the second-best toll pricing problem as a mathematical program with an equilibrium constraint expressed as a variational inequality. To investigate the properties associated with and derive an algorithm for the problem, we present three equivalent nonlinear programming formulations. These formulations differ in their representation of the tolled user equilibrium condition. The first formulation uses the KKT conditions to state the equilibrium condition, the second represents the feasible region as convex combinations of its extreme points, and the last relies on the regularized gap function to ensure that the equilibrium condition is satisfied. These equivalent formulations lead to two main results. One relates the second-best tolls to the marginal social cost prices via KKT multipliers and another yields a cutting constraint algorithm. To demonstrate its potential, we implemented the algorithm using commercially available software for linear and nonlinear programs and solved the second-best problems for two cities, Sioux Falls and Hull. Our numerical results suggest that the cutting constraint algorithm is capable of solving realistic second-best problems. As a topic for future research, an algorithm that better exploits the structure of the master problem would enhance the efficiency of the algorithm.

Acknowledgements. The authors would like to thank the two anonymous referees for their insightful comments and helpful suggestions on earlier versions of this paper.

References

1. Arnott, R., Small, K.: The economics of traffic congestion. *Am. Sci.* **20** (2), 123–127 (1994)
2. Auchmuty, G.: Variational principles for variational inequalities. *Numer. Func. Anal. Optim.* **10**, 863–874 (1989)
3. Bard, J.F.: *Practical Bilevel Optimization: Algorithms and Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands 1998
4. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: *Nonlinear Programming: Theory and Algorithms*. Second Edition, John Wiley & Sons, New York, New York, 1993
5. Brooke, A., Kendrick, D., Meeraus, A.: *GAMS: A User's Guide*. The Scientific Press, South San Francisco, California, 1992
6. Brotcorne, L., Labbé, M., Marcotte, P., Savard, G.: A bilevel model for toll optimization on a multicommodity transportation network. *Trans. Sci.* **35** (4), 345–358 (2001)
7. Dial, R.: Minimum revenue congestion pricing Part I: A fast algorithm for the single-origin case. *Trans. Res. B* **33** (3), 189–202 (1999)
8. Dial, R.: Minimum revenue congestion pricing Part II: A fast algorithm for the general case. *Trans. Res. B* **34** (8), 645–665 (2000)
9. Facchinei, F., Pang, J.-S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Vol. I and II. Springer, New York, 2003
10. Ferrari, P.: Road network toll pricing and social welfare. *Trans. Res. B* **36** (5), 471–483 (2002)
11. Fletcher, R., Leyffer, S.: Numerical experience with solving MPECs as NLPs. Numerical Analysis Report NA/210, Department of Mathematics, University of Dundee, 2002
12. Fletcher, R., Leyffer, S., Ralph, D., Scholtes, S.: Local convergence of SQP methods for mathematical programs with equilibrium constraints. Numerical Analysis Report NA/209, Department of Mathematics, University of Dundee, 2002

13. Florian, M., Hearn, D.W.: Network equilibrium models and algorithms. Chapter 6 of *Handbooks in Operations Research and Management Science*. In: Network Routing, M.O. Ball, T.L. Magnanti, C.L. Monma, G.L. Nemhauser (eds.), Volume 8, North-Holland, New York, 1995
14. 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)
15. Fisk, C.S., Boyce, D.E.: Alternative variational inequality formulations of the network equilibrium-travel choice problem. *Trans. Sci.* **17** (4), 454–463 (1983)
16. Fukushima, M.: Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Math. Program.* **53**, 99–110 (1992)
17. Fukushima, M.: Merit functions for variational inequality and complementarity problems. In: *Nonlinear Optimization and Applications*, G. Di Pillo, F. Giannessi (eds.), Plenum Publishing Corporation, New York, 1996
18. Gartner, N.H.: Optimal traffic assignment with elastic demands: A Review, Part I. Analysis Framework. *Trans. Sci.* **14** (2), 174–191 (1980)
19. Gauvin, J.: A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Math. Program.* **12**, 136–138 (1977)
20. Hearn, D.W., Ramana, M.V.: Solving congestion toll pricing models. In: *Equilibrium and Advanced Transportation Modeling*, P. Marcotte, S. Nguyen (eds.), Kluwer Academic Publishers, Boston, 1998, pp. 109–124
21. Hearn, D.W., Yildirim, M.B.: A toll pricing framework for traffic assignment problems with elastic demands. In: *Current Trends in Transportation and Network Analysis: Miscellanea in Honor of Michael Florian*, M. Gendreau, P. Marcotte (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001
22. Hearn, D.W., Yildirim, M.B.: A first-best toll pricing framework for variable demand traffic assignment problems. *Trans. Res. B*. To appear 2004
23. Johansson-Stenman, O., Sterner, T.: What is the scope for environmental road pricing? In: *Road pricing, Traffic Congestion and Environment*, K.J. Button, E.T. Verhoef (eds.), Edward Elgar Publishing Limited, London, England, 1998
24. Labbé, M., Marcotte, P., Savard, G.: A bilevel model of taxation and its application to optimal highway pricing. *Manage. Sci.* **44** (12), 1608–1622 (1998)
25. Larsson, T., Patriksson, M.: Side constrained traffic equilibrium models—traffic management through link tolls. In: *Equilibrium and Advanced Transportation Modelling*, P. Marcotte, S. Nguyen (eds.), Kluwer Academic Publishers, New York, 1998, pp. 125–151
26. LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P.: An efficient approach to solving the road network equilibrium traffic assignment problem. *Trans. Res.* **9**, 309–318 (1975)
27. Lim, A.: Transportation network design problems: An MPEC approach. Ph.D. Dissertation, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland, 2002
28. Luo, Z.-Q., Pang, J.-S., Ralph, D.: *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, New York, New York, 1996
29. Mangasarian, O.L., Fromovitz, S.: The Fritz John optimal necessary optimality conditions in the presence of equality and inequality constraints. *J. Math. Anal. Appl.* **17**, 37–47 (1967)
30. Marcotte, P.: Network optimization with continuous control parameters. *Trans. Sci.* **17** (2), 181–197 (1983)
31. Marcotte, P., Zhu, D.L.: Exact and inexact penalty methods for the generalized bilevel programming problem. *Math. Program. A* **74** (2), 141–157 (1996)
32. McDonald, J.F.: Urban highway congestion: An analysis of second-best tolls. *Trans.* **22**, 353–369 (1995)
33. Migdalas, A.: Bilevel programming in traffic planning: models, methods and challenge. *J. Global Optim.* **7**, 381–405 (1995)
34. Outrata, J.V., Kocvara, M., Zowe, J.: *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998
35. Patriksson, M., Rockafellar, R.T.: A mathematical model and descent algorithm for bilevel traffic management. *Trans. Sci.* **36** (3), 271–291 (2002)
36. Scheel, H., Scholtes, S.: Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. *Math. Oper. Res.* **25** (1), 1–22 (2000)
37. Shimizu, K., Ishizuka, Y., Bard, J.F.: *Nondifferentiable and Two-Level Mathematical Programming*. Kluwer Academic Publishers, Boston, 1997
38. Verhoef, E.T.: Second-best congestion pricing in general static transportation networks with elastic demands. *Region. Sci. Urban Econ.* **32** (3), 281–310 (2002)
39. Verhoef, E.T.: Second-best congestion pricing in general networks: Algorithms for finding second-best optimal toll levels and toll points. *Trans. Res. B* **36** (8), 707–729 (2002)
40. Yang, H., Bell, M.G.H.: Traffic restraint, road pricing and network equilibrium. *Trans. Res. B* **33** (4), 303–314 (1997)

41. Yang, H., Lam, W.H.K.: Optimal road tolls under conditions of queuing and congestion. *Trans. Res. A* **30** (5), 319–332 (1996)
42. Yildirim, M.B.: Congestion toll pricing models and methods for variable demand networks. Dissertation, Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida, 2001
43. 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

This appendix provides an example demonstrating that P-EX and P-Gap do not satisfy MFCQ and discusses the consequences of this fact.

Consider the two-arc problem (see Figure 1) in Section 4, where Arc 1 is tollable and Arc 2 is not. Recall that $s_1(v_1) = v_1$ and $s_2(v_2) = v_2 + 2$. Instead of allowing the demand to be elastic, the travel demand from node 1 to node 2 is seven. In addition, the maximum amount of toll on Arc 1 is eight.

The system problem associated with this two-arc problem is

$$\begin{aligned} \text{S-OPT: } \quad & \min v_1^2 + v_2^2 + 2v_2 \\ & \text{s.t. } v_1 + v_2 = 7 \\ & v_1, v_2 \geq 0. \end{aligned}$$

The optimal solution to S-OPT is $(v_1, v_2) = (4, 3)$ and has an objective value of 31.

The set of all feasible flow vectors, V , has two extreme points, $(7, 0)$ and $(0, 7)$, and the extreme point formulation of the second-best problem becomes

$$\begin{aligned} \text{FD-EX : } \quad & \min v_1^2 + v_2^2 + 2v_2 \\ & \text{s.t. } v_1 + v_2 = 7 \\ & (v_1 + \beta)(7 - v_1) + (v_2 + 2)(0 - v_2) \geq 0 \\ & (v_1 + \beta)(0 - v_1) + (v_2 + 2)(7 - v_2) \geq 0 \\ & 0 \leq \beta \leq 8 \\ & v_1, v_2 \geq 0. \end{aligned}$$

The optimal solution to FD-EX is $(v_1, v_2, \beta) = (4, 3, 1)$ and has an objective value of 31, the same value as S-OPT. The fact that FD-EX has the same objective value as S-OPT also verifies that $(v_1, v_2, \beta) = (4, 3, 1)$ is globally optimal.

By solving the associated tolled user equilibrium problem, a feasible solution to FD-EX must be of the form: $v(\beta) = (v_1(\beta), v_2(\beta)) = (\frac{9-\beta}{2}, \frac{5+\beta}{2})$, where $\beta \in [0, 8]$. At $(v(\beta), \beta)$, where $\beta \in [0, 8]$, the first three constraints are always binding. Based on the gradients of these three binding constraints, MFCQ requires that there exists a $d \in R^3$ satisfying the following conditions:

$$\begin{aligned} & d_1 + d_2 = 0 \\ & -2d_1 - (7 + \beta)d_2 + \frac{(5+\beta)}{2}d_3 > 0 \\ & -9d_1 - \beta d_2 - \frac{(9-\beta)}{2}d_3 > 0. \end{aligned}$$

The first condition requires $d_2 = -d_1$. After substituting $-d_1$ for d_2 , the second and third conditions imply that $d_1 > -d_3/2$ and $d_1 < -d_3/2$, respectively. Because these

two inequalities are contradictory, MFCQ is not satisfied at any feasible point $(v(\beta), \beta)$, where $\beta \in [0, 8]$.

The failure of MFCQ can pose numerical difficulties to NLP algorithms. (See, e.g., the discussion in Fletcher and Leyffer [11] and Fletcher et al. [12]) In particular, the failure of MFCQ implies that the set of multipliers satisfying the KKT conditions associated with FD-EX at any feasible point $(v(\beta), \beta)$ must be empty or unbounded. (See Gauvin [19].) To illustrate, the KKT conditions for FD-EX at $(v(\beta), \beta)$ where $\beta \in (0, 8)$ reduce to

$$\begin{aligned}\mu + 2\lambda_1 + 9\lambda_2 &= -(9 - \beta) \\ \mu + (7 + \beta)\lambda_1 + \beta\lambda_2 &= -(7 + \beta) \\ -\frac{(5+\beta)}{2}\lambda_1 + \frac{(9-\beta)}{2}\lambda_2 &= 0 \\ \lambda_1, \lambda_2 &\geq 0.\end{aligned}$$

The last equality implies that $\lambda_2 = \frac{(5+\beta)}{(9-\beta)}\lambda_1$. After substituting this expression for λ_2 , the first two equations become

$$\begin{aligned}\mu + \frac{(63+7\beta)}{(9-\beta)}\lambda_1 &= -(9 - \beta) \\ \mu + \frac{(63+7\beta)}{(9-\beta)}\lambda_1 &= -(7 + \beta).\end{aligned}$$

Thus, when $\beta \in (0, 8)$ and $\beta \neq 1$, the above system has no solution because $9 - \beta \neq 7 + \beta$ and there is no multiplier that satisfies the KKT conditions.

When $\beta = 1$, the optimal toll amount, the above KKT conditions reduce to

$$\begin{aligned}\mu + \frac{70}{8}\lambda_1 &= -8 \\ \mu + \frac{70}{8}\lambda_1 &= -8\end{aligned}$$

and $(\mu, \lambda_1) = (-8 - \frac{70}{8}\lambda_1, \lambda_1)$ is a solution for all $\lambda_1 \geq 0$. Thus, the set of KKT multipliers is unbounded.

The KKT conditions at $\beta = 0$ and 8 are slightly different. However, the same conclusions still hold, i.e., set of KKT multipliers is unbounded at $\beta = 0$ and empty at $\beta = 8$.

The regularized gap function associated with the above second-best problem can be written as follows:

$$\begin{aligned}G(v, \beta) &= \max (v_1 + \beta)(v_1 - y_1) + (v_2 + 2)(v_2 - y_2) - \frac{1}{2}\|y - v\|^2 \\ \text{s.t. } & y_1 + y_2 = 7 \\ & y_1, y_2 \geq 0.\end{aligned}$$

The optimal solution to the above problem is $(y_1, y_2) = \left(\frac{9-\beta}{2}, \frac{5+\beta}{2}\right)$ and the gap function can be written as

$$\begin{aligned}G(v, \beta) &= (v_1 + \beta)\left(v_1 - \frac{9-\beta}{2}\right) + (v_2 + 2)\left(v_2 - \frac{5+\beta}{2}\right) \\ &\quad - \frac{1}{2}\left(\frac{9-\beta}{2} - v_1\right)^2 - \frac{1}{2}\left(\frac{5+\beta}{2} - v_2\right)^2.\end{aligned}$$

Then, FD-Gap can be written as follows:

$$\begin{aligned} \min \quad & v_1^2 + v_2^2 + 2v_2 \\ \text{s.t.} \quad & v_1 + v_2 = 7 \\ & G(v, \beta) \leq 0 \\ & 0 \leq \beta \leq 8 \\ & v_1, v_2 \geq 0. \end{aligned}$$

At any feasible point $(v(\beta), \beta)$, where $\beta \in [0, 8]$, the first two constraints are always binding and MFCQ requires that there exists a $d \in R^3$ satisfying the following conditions:

$$\begin{aligned} d_1 + d_2 &= 0 \\ \frac{(9+\beta)}{2}d_1 + \frac{(9+\beta)}{2}d_2 &< 0. \end{aligned}$$

Because these two equations are inconsistent, MFCQ does not hold. Similar to before, this implies that the set of feasible KKT multipliers for FD-Gap is empty when $\beta \in [0, 8]$ and $\beta \neq 1$ and unbounded when $\beta = 1$.