Optimal Multi-Step Toll Design under General User Heterogeneity

Hongyu Chen¹, Yu (Marco) Nie *,¹, and Yafeng Yin²

¹Department of Civil and Environmental Engineering
Northwestern University
2145 Sheridan Road, Evanston, IL 60208, USA

²Department of Civil and Coastal Engineering
University of Florida
Gainesville, FL 32611, USA

June 30, 2015

Abstract

This paper studies the optimal multi-step toll design problem for the bottleneck model with general user heterogeneity. The design model is formulated as a mathematical program with equilibrium constraints (MPEC), which is NP-hard due to non-convexity in both the objective function and the feasible set. An analytical method is proposed to solve the MPEC by decomposing it into smaller and easier quadratic programs, each corresponding to a unique departure order of different user classes. The quadratic programs are defined on a polyhedral set, which makes it easier to identify a local optimum. Importantly, each quadratic program is constrained by a set of linear feasibility cuts that define the presence of each user class in the arrival window. We prove that the proposed method ensures global optimality provided that each quadratic program can be solved globally. To obviate enumerating all departure orders, a heuristic method is developed to navigate through the solution space by using the multipliers associated with the feasibility cuts. Numerical experiments are conducted on several small examples to validate the proposed methodology. These experiments show that the proposed heuristic method is effective in finding near-optimal solution within a relatively small number of iterations.

Keywords: step toll; general user heterogeneity; mathematical program with equilibrium constraint; quadratic program; bottleneck model

1 Introduction

The peak-time congestion pricing (Vickrey, 1969) has been widely studied in the context of traffic management. Despite its theoretical appeal, peak-time pricing remains unpopular among pub-
lic. Opponents frequently cite regressive redistributive effects as a main drawback (Small, 1983; Arnott et al., 1992; Evans, 1992). Such an argument hinges on the fact that pricing affects travelers unequally because they value time and schedule punctuality differently. Not surprisingly, the impact of such user heterogeneity on optimal toll design and welfare effects has attracted much attention in the past decades (Small, 1982; Cohen, 1987; Arnott et al., 1992, 1994; Lindsey, 2004; Small et al., 2005; van den Berg and Verhoef, 2011b; Liu and Nie, 2011; Hall, 2013). To maximize efficiency, Vickrey’s toll has to vary continuously with time. Yet, such a time-varying toll is rarely implemented in practice. Empirical evidence suggests that simpler pricing schemes may have a better chance to rally support. For example, according to a survey conducted in early 1990s (Higgins, 1994), travelers seem to dislike time varying tolls more than flat tolls. Indeed, existing congestion pricing schemes (e.g. London, Singapore and Stockholm, Lindsey et al., 2012) often use “step tolls” that keep toll rates constant in predefined discrete time windows.

This paper is focused on the optimal multi-step toll design problem under general user heterogeneity. In our context, user heterogeneity typically consists of the desired arrival time ($t^*$), the value of time ($\alpha$) and the values of schedule punctuality ($\beta$ for early arrival and $\gamma$ for late arrival). Previous studies often make assumptions about the relationship between $\alpha$, $\beta$ and $\gamma$ to simplify the analysis. The most restrictive assumptions require either that all three parameters vary proportionally (Vickrey, 1973; Xiao et al., 2011), or that only $\alpha$ (Arnott and Kraus, 1995; van den Berg and Verhoef, 2011a) or $\gamma$ (van den Berg, 2014) may vary; hereafter referred to as proportional heterogeneity, $\alpha$ heterogeneity and $\gamma$ heterogeneity, respectively. These assumptions effectively reduce a three-dimensional problem to a one-dimensional problem. A less restrictive structure of heterogeneity reduces the dimension of the problem to two by allowing $\alpha$ and $\beta$ to vary freely and assuming $\gamma$ as a function of $\beta$ (Cohen, 1987; Newell, 1987; Arnott et al., 1988, 1994; van den Berg and Verhoef, 2011b; Liu and Nie, 2011; Hall, 2013).

Welfare effects of congestion pricing seem to depend on the imposed heterogeneity structure. For example, with proportional heterogeneity, all users are better off or break even under Vickrey’s time-varying toll (Xiao et al., 2011; van den Berg and Verhoef, 2011b), and yet with $\alpha$ heterogeneity, all users are worse off or break even (van den Berg and Verhoef, 2011b). Notably, the gains from pricing rise with the value of time in these one-dimensional cases. However, such a monotonic relationship no longer holds with two-dimensional heterogeneity (van den Berg and Verhoef, 2011b). Clearly, restricting the heterogeneity structure, which has so far been considered necessary to maintain tractability, also limits the ability to appreciate and draw insights about the complex interactions at work.

Motivated by the above, we propose an analytical method to find optimal step tolls for the bottleneck model with users of general heterogeneity. By general heterogeneity, we mean that no arbitrary relationships are imposed on $\alpha$, $\beta$ and $\gamma$. $^1$ The proposed method aims to locate an exact solution by enumerating all possible combinations of user departure orders. The underlying idea behind the method is recently explored in Chen et al. (2014), which focuses on finding user equilibrium solutions for the bottleneck model with general heterogeneity and given step tolls. Chen et al. (2014) show that, once the departure order of different user classes is given, a linear equation system can be constructed to generate a candidate solution, which is then verified against

$^1$We still assume all users share the same desired arrival time $t^*$. However, different desired arrival times can be easily accommodated using the proposed method, as explained in conclusions (Section 6).
the equilibrium conditions. Since the number of possible departure orders is finite, the method guarantees finding the correct equilibrium solutions after exhausting all possible departure orders. This paper tackles the optimal toll design problem using the similar idea. Specifically, we construct and solve a toll design problem for each departure order, which is formulated as a quadratic program constrained by user equilibrium conditions associated with the departure order. The optimal step toll configuration can then be determined by comparing the objective functions of all feasible design problems. Note that the general toll design problem constrained by equilibrium conditions is a challenging network design problem known to be NP-hard, even without considering user heterogeneity. Indeed, because the number of possible departure orders grow exponentially with the problem size (the number of user classes and the number of step tolls), the exact method is impractical except for extremely small problems. In light of this limitation, an effective heuristic method will be developed to obviate the complete enumeration of departure orders.

Few analytical studies \(^2\) had considered a heterogeneity structure as general as pursued in this paper. Lindsey (2004) analyzes a similar model but his focus is to prove that it admits one and only one user equilibrium under mild conditions. His result is significant but does not prescribe solution and design methods. Recently, the user equilibrium problem for the step-tolled bottleneck model under general heterogeneity has been solved using a semi-analytical method (Liu et al., 2015) and an analytical method (Chen et al., 2014). However, neither study addresses the optimal design of step tolls.

The rest of this paper is organized as follows. Section 2 introduces the basic setting of the bottleneck model with general user heterogeneity and step tolls. Section 3 presents the formulations of the optimal step toll design problem as a mathematical program with equilibrium constraints. Development of analytical and heuristic solution methods are discussed in Section 4. Section 5 validates the proposed methodology using a few small numerical examples. Section 6 concludes the paper with suggestion for future directions of research.

### 2 Preliminaries

Consider a fixed number of individuals \((N)\) who commute from home to work during morning rush hour through a bottleneck with a capacity of \(s\). The en-route travel time of any individual who arrives at the workplace at time \(t\) contains two parts: a waiting time \(T(t)\) in the queue and a free flow travel time \(T_0\). Because the free flow travel time \(T_0\) will not affect the following analysis, we assume \(T_0 = 0\). It is also assumed that all commuters prefer to pass the bottleneck and arrive at work at \(t^* = 0\). Whenever an early or late arrival occurs, a schedule cost is incurred. At user equilibrium, each commuter chooses an arrival time \(t\) to minimize a general cost which includes travel delay \(T(t)\) and schedule cost. To model user heterogeneity, commuters are divided into \(n\) classes, each including \(N_i, (i = 1, ..., n)\) commuters who have identical unit cost of travel time \((\alpha_i)\) and unit cost of schedule delay \((\beta_i\) for early arrival and \(\gamma_i\) for late arrival). Note that the existence of equilibrium requires that \(\beta_i < \alpha_i, \forall i\) (Lindsey, 2004). To avoid potential degenerative solutions, we also assume \(\beta_i/\alpha_i \neq \beta_j/\alpha_j\) and \(\gamma_i/\alpha_i \neq \gamma_j/\alpha_j\) for any \(i \neq j\). For the convenience of

---

\(^2\)Examples of considering user heterogeneity in discretized bottleneck models may be found in Yang and Huang (1997); Yang and Meng (1998); Ramadurai et al. (2010); Doan et al. (2011).
the reader, Table 7 in Appendix lists main notations.

We divide the analysis period into two sub periods: the early arrival period \((-∞, t^∗]\) and the late arrival period \([t^∗, ∞]\). We use \(A\) to identify the arrival periods. Specifically, for the early arrival period \(A = E\) and for the late arrival period \(A = L\). Let \(p_{j}^{E}\) be the distance in time between \(t^∗\) and the beginning of the \(j\)th toll window in the early arrival period, and \(p_{j}^{L}\) be the distance in time between \(t^∗\) and the end of the \(j\)th toll window in the late arrival period. Moreover, \(m^A\) is the number of step toll windows in the arrival period \(A = \{E, L\}\). Note that the toll windows in each arrival period are numbered starting at 0 and from the far end (see Figure 1 for an illustration). That is, \(p_{0}^{A} = ∞ > p_{1}^{A} > · · · > p_{m^A}^{A} = 0, ∀ A = \{E, L\}\). Accordingly, the length of \(j\)th toll window in the arrival period \(A\) is \(T_{j}^{A} = p_{j}^{A} − p_{j+1}^{A}, j = 1, ... m^A, ∀ A = \{E, L\}\). For \(j = 0, · · · , m^A\), the amount of toll in each arrival period is denoted as \(π_{j}^{A}\). Note that \(π_{0}^{A} = 0, ∀ A = \{E, L\}\), which dictates that the furthest toll windows from \(t^∗\) on both sides are in fact no-toll windows. Also, \(π_{j}^{A} ≥ 0, ∀ j = 1, · · · , m^A, A = \{E, L\}\). In what follows, \(A_j\) will be used to identify the \(j\)th arrival window in the arrival period \(A\), where \(j = 0, 1, · · · , m^A\) and \(A = \{E, L\}\).

A user equilibrium (UE) is attained if no commuter could reduce his/her commute cost by changing the departure time. This notion of equilibrium may be illustrated graphically using an isocost curve (see e.g. Hendrickson and Kocur, 1981; Newell, 1987; Cohen, 1987; Arnott et al., 1994; Lindsey, 2004), i.e., at equilibrium, all commuters from the same class must be on the same isocost curve. Let \(L_{i}\), \(2004\), i.e., at equilibrium, all commuters from the same class must be on the same isocost curve.

Let \(t\) denote the time at which commuters depart from the bottleneck (corresponding to the arrival time at the workplace), the isocost curve with step toll is given as:

\[
\mu_{i}(t) = \begin{cases} 
T(t) + \frac{β_{i}(t^∗−t)}{α_{i}} + \frac{π_{j}^{E}}{α_{j}} & t^∗ − p_{j}^{E} < t ≤ t^∗ − p_{j+1}^{E} \quad j = 0, 1, · · · , m_{E} \quad \text{for early arrival} \\
T(t) + \frac{γ_{i}(t^∗−t)}{α_{i}} + \frac{π_{j}^{L}}{α_{j}} & t^∗ + p_{j}^{L} < t ≤ t^∗ + p_{j+1}^{L} \quad j = 0, 1, · · · , m_{L} \quad \text{for late arrival} 
\end{cases}
\]

where \(μ_{i}\) is the travel cost for class \(i\) measured in travel time and \(T(t)\) is the travel time corresponding to arrival time \(t\). Hence, \(dμ_{i}(t)/dt = 0\) implies that

\[
\frac{dT_{i}}{dt} = \begin{cases} 
\frac{β_{i}}{α_{i}}, & t ≤ t^∗ \quad \text{early arrival} \\
−\frac{γ_{i}}{α_{i}}, & t > t^∗ \quad \text{late arrival} 
\end{cases}
\]

which gives the slopes of the isocost curves. For the purpose of illustration, Figure 1 shows isocost curves of two user classes and multiple step tolls. Note that the isocost curve of each class actually represents the commuters’ willingness to pay for each time slot. At UE, an arrival time slot is always assigned to the class with highest willingness to pay. Graphically, this means that commuters should always stay on the upper envelope of all the isocost curves.

The primary complexity in step toll analysis arises from the discontinuity at the end of each toll window in the late arrival period. In order to compensate this discontinuity, the first commuter in the \(j\)th toll window must experience a higher travel time than the last commuter would do in the \(j + 1\)th toll window \(∀ j = 1, m_{L}^j − 1\). This calls for additional assumptions about how commuters behave at the end of each toll window. In the literature, the mass arrival assumption (Arnott et al., 1990) dictates that a group of users would arrive at the beginning of \(j\)th toll window to create a temporal queue and that they would all experience the same expected cost from that queue to offset the toll relief. The braking-induced idle assumption (Lindsey et al., 2012), on
the other hand, resolves the discontinuity by forcing the bottleneck to operate below capacity at the end of each toll window. Both assumptions will lead to a slightly different rush hour period and displace commuters’ arrival time compared to the no-toll UE, which complicates the analysis especially when \( m^A > 1 \), \( A = \{E, L\} \). In contrast, Laih (1994, 2004) shows that in the case of homogenous users, the separate-waiting (SW) assumption avoids such complexities while keeping the rush hour period and commuters’ arrival time intact after the step toll. showed This nice property of the SW assumption still holds with heterogeneous users (Chen et al., 2014). Since we expect that the analysis outcome is relatively insensitive to these behavioral assumptions (Nie, 2013), the SW assumption will be adopted here for its simplicity.

### 3 Optimal step toll design problem

In this section, we will give a general formulation for the step toll design problem. To avoid unnecessary complexities, we shall assume \( m^A \) is given for \( A = \{E, L\} \). In the homogeneous case, a larger value of \( m^A \) always leads to higher efficiency, since the continuously varying toll can eliminate all queuing delays (Vickrey, 1969). Yet, how many step toll windows can be implemented is often dictated by technology constraints and/or political processes. Thus, it seems reasonable to assume that the policy makers would choose \( m^A \) first, before fine-tuning other design parameters such as \( \pi^A_j \) and \( p^A_j \). The design formulation presented herein aims at finding optimal \( \pi^A_j \) and \( p^A_j \), \( \forall j, A \) for given \( m^A \).

The proposed formulation will be in the form of a mathematical program with equilibrium constraints (MPEC) (Harker and Pang, 1988; Luo et al., 1996). Therefore, we will first present the step-tolled equilibrium model, formulated as an asymmetric traffic assignment problem with side constraints, along the line of Liu et al. (2015) and Chen et al. (2014).

#### 3.1 Step-tolled equilibrium

According to Arnott et al. (1994) and Lindsey (2004), the departure orders of all user classes in any window \( A_j \) can be determined based on their values of \( \alpha, \beta \) and \( \gamma \). Namely, the higher the ratio \( \beta_i/\alpha_i \) (for early arrival) or \( \gamma_i/\alpha_i \) (for late arrival) is, the closer class \( i \)'s arrival time will be...
to \( t^* \). Let \( E(i) \) be the class ID of the class whose \( \beta_i / \alpha_i \) value ranks at the \( i^{th} \) place in all classes, and \( L(i) \) be the class ID of the class whose \( \gamma_i / \alpha_i \) value ranks at the \( i^{th} \) place in all classes. Then with the notation \( A = \{ E, L \} \), \( A(i) \) is the class ID for ratio ranks. Without loss of generality, we assume that the class IDs are already ranked by \( \beta_i / \alpha_i \) at the beginning, which implies

\[
E(i) = i.
\]

Let \( N_i^{A(j)} \) and \( C_i^{A(j)} \) be respectively the number of class \( i \) users and their equilibrium cost (measured in the units of travel delay) in the toll window \( A_j, j = 0, 1, \cdots, m^A, A = \{ E, L \} \). We use \( \mathbf{N} \) and \( \mathbf{C} \) to denote the corresponding vectors, and note that the length of these vectors is \( n (m^E + m^L + 2) \).

Similarly, \( \mathbf{b} = \{ \pi^A_i, j = 1, \cdots, m^A, A = \{ E, L \} \} \) and \( \mathbf{p} = \{ p_{ij}, j = 1, \cdots, m^A, A = \{ E, L \} \} \) are used to represent the vectors of toll and the boundaries of the toll windows, respectively. Note that the length of these two vectors is \( m^E + m^L \).

Using the properties of isocost curves, the general travel cost of class \( i \) in arrival window \( A_j (A = E, L) \), measured in the time unit, can be written as follows (The reader is referred to Liu et al. (2015) for details).

\[
C_{E_i}^{A(j)} = \sum_{k=1}^{i} \frac{\beta_k}{\alpha_k} N_k^{E_i} \frac{E_j}{s} + \frac{\beta_i}{\alpha_i} \sum_{k=1}^{n} \frac{N_k^{E_i}}{s} + \frac{\beta_i}{\alpha_i} N_i^{E_{i+1}} + \pi_j^{E_i}, \quad i = 1, 2, \ldots, n, j = 0, 1, \ldots, m^E; \tag{3a}
\]

\[
C_{L(i)}^{A(j)} = \sum_{k=1}^{i} \frac{\gamma_{L(k)}}{\alpha_{L(k)}} N_{L(i)}^{E_j} \frac{s}{s} + \frac{\gamma_{L(i)}}{\alpha_{L(i)}} \sum_{k=1}^{n} \frac{N_{L(i)}^{E_j}}{s} + \frac{\gamma_{L(i)}}{\alpha_{L(i)}} N_{i}^{L_{i+1}} + \pi_j^{L_i}, \quad i = 1, 2, \ldots, n, j = 0, 1, \ldots, m^L; \tag{3b}
\]

The flow conservation constraints and nonnegativity constraints are

\[
\sum_{j=0}^{m^E} N_i^{E_j} + \sum_{j=0}^{m^L} N_i^{L_j} = N_i, \quad i = 1, 2, \ldots, n; \tag{4}
\]

\[
N_i^{A(j)} \geq 0, \quad i = 1, 2, \ldots, n, j = 0, 1, \ldots, m^A, A = \{ E, L \}; \tag{5}
\]

Chen et al. (2014) noted that the capacity constraints are needed for all arrival windows that have a finite length, i.e.

\[
\sum_{i=1}^{n} N_i^{A(j)} \leq s (p_{ij}^A - p_{ij+1}^A) = s l_j^A, \forall j = 1, \ldots, m^A, A = \{ E, L \}. \tag{6}
\]

Finally, the dynamic user equilibrium conditions dictate that

\[
N_i^{A(j)} (C_i^{A(j)} + \lambda_j^A - \mu_i) = 0, \forall j = 0, \cdots, m^A, A = \{ E, L \}, i = 1, \cdots, n; \tag{7a}
\]

\[
C_i^{A(j)} + \lambda_j^A \geq \mu_i, \forall j = 0, \cdots, m^A, A = \{ E, L \}, i = 1, \cdots, n; \tag{7b}
\]

where \( \lambda_j^A \) is the multiplier associated with the capacity constraint for \( A_j \) (6), interpreted as an “additional delay”, and \( \mu_i \) is the UE cost of class \( i \). Note that all users in \( A_j \) share the same additional delay created by the capacity restriction (Chen et al., 2014).

The problem of finding \( \mathbf{N}^* \) and \( \mathbf{C}^* \) that satisfy (4-7) can be formulated as a traffic assignment problem defined on a network as depicted in Figure 2. The network has \( n \) origin-destination
Figure 2: Topology of transformed network corresponding to the bottleneck model (each link is associated with a toll window)

pairs. Each O-D pair \( i \) starts from \( i \) and ends at \( i + 1 \) in the network, corresponding to the user class \( i \) with a demand \( N_i \). There are \( m^E + m^L + 2 \) routes connecting each O-D pair, representing all possible arrival windows. Commuters of each class \( i \) split among these routes according to the route choice behavior defined by the Wardrop principle (Wardrop, 1952). The route costs are given by (3), which is non-separable and asymmetric.

It is well known (see e.g. Liu et al., 2015) that the flow vector \( N^* \) satisfies (4-7) if and only if it solves a variational inequalities problem VIP\((\Omega, C)\):

\[
\langle C(N^*), N - N^* \rangle \geq 0, \forall N \in \Omega
\]  

where \( C \) is defined in Equation (3) and

\[
\Omega = \{N|N \text{ satisfies Constraints (4) } - (6)\}.
\]

3.2 Optimal toll design

The objective of the toll design is to minimize the inefficiency in the system, which is measured by the total commute costs including both schedule and travel costs. Noting \( C \) defined by (3) includes the toll income, the total system cost can be written as follows:

\[
W = \sum_{i=1}^{n} \sum_{A=E,L} \left( \sum_{j=0}^{m^A} \alpha_i^A N_i^{A_j} (C_i^{A_j} + \lambda_j) - \sum_{j=1}^{m^A} \pi_j^A N_i^{A_j} \right), \tag{9}
\]

where the first term in (9) represents the total commuter cost and the second is the toll income. Since for any selected tolls scheme, the users must make their departure time choices according to the user equilibrium conditions, \( N_i^{A_j}, C_i^{A_j} \) and \( \lambda_j \) in (9) must be the solution to the VIP (8). Therefore, the toll design problem can be written as the following mathematical program with
equilibrium constraints (MPEC):

$$\min_{[p, \beta]} W$$

subject to

$${\bf N}, C$$ and $$\lambda$$ solve (8)

$$[p, \beta] \in \Gamma$$

where (10b) represents the equilibrium constraints and

$$\Gamma = \{[p, \beta]|p \geq 0, \pi \geq 0, p_j^A \geq p_{j+1}^A, j = 0, 1, \ldots, m^A, A = \{E, L\}\}.$$  (11)

The literature shows that, in the homogeneous and some special heterogeneity cases (proportional heterogeneity and $$\gamma$$ heterogeneity) discussed in (Arnott et al., 1990; Laih, 1994, 2004; Xiao et al., 2011, 2012; van den Berg and Verhoef, 2011a; van den Berg, 2014), the optimal toll window boundaries $$p$$ are closely related to the toll $$\pi$$. Specifically, for a given $$\pi$$, the optimal boundaries are always determined such that: (1) the bottleneck always operates at the capacity in all tolled arrival windows, and (2) the first (last) commuter is subject to no queuing delay in each of the early (late) tolled arrival window. If this property holds in the case of general heterogeneity, the capacity constraint (6) will always bind for an optimally designed toll, i.e.

$$\sum_{i=1}^{n} N_i^A = s l_i^A, j = 1, 2, \ldots, m^A, \forall A = \{E, L\}$$

(12)

Summing the above equations from $$j$$ to $$m^A$$ yields

$$p_j^A = \sum_{w=j+1}^{m^A} \sum_{k=1}^{n} \frac{N_k^{Aw}}{s} j = 1, 2, \ldots, m^A, \forall A = \{E, L\}$$

(13)

Accordingly, $$p_{j+1}^A$$ in Equation (3) can be replaced with the new expression above, i.e.

$$C_{E_i}^{E_j} = \sum_{k=1}^{i-1} \frac{\beta_k}{\alpha_k} \frac{N_k^{E_j}}{s} + \frac{\beta_{E_i}}{\alpha_i} \sum_{k=i}^{n} \frac{N_k^{E_j}}{s} + \frac{\gamma_j}{\alpha_j} \sum_{k=1}^{m^E} \sum_{w=j+1}^{m^E} \frac{N_k^{E_w}}{s} + \frac{\pi_j}{\alpha_j}, \quad i = 1, 2, \ldots, n, j = 0, 1, \ldots, m^E;$$

(14a)

$$C_{L(i)}^{L_j} = \sum_{k=1}^{i-1} \frac{\gamma_{L(k)}}{\alpha_{L(k)}} \frac{N_{L(j)}^{L(k)}}{s} + \frac{\gamma_{E(i)}}{\alpha_{L(i)}} \sum_{k=i}^{n} \frac{N_{L(j)}^{L(k)}}{s} + \frac{\gamma_{E(i)}}{\alpha_{L(i)}} \sum_{k=1}^{m^L} \sum_{w=j+1}^{m^L} \frac{N_k^{L_w}}{s} + \frac{\pi_j}{\alpha_{L(i)}}, \quad i = 1, 2, \ldots, n, j = 0, 1, \ldots, m^L.$$  

(14b)

Consequently, the optimal toll design problem can be significantly simplified by making use of the above relationship between $$p$$ and $$\beta$$. Not only is the number of decision variables in the upper level reduced by two thirds, but the capacity constraints in the lower level problem can also be eliminated. Since it remains an open question whether this relationship holds under general user heterogeneity, the above simplified formulation may be used as an approximation when a relatively crude solution is sufficient or when the computation overhead is a major concern. It is however not used in the numerical experiments presented later.
4 Solution method

Thanks to the equilibrium constraints, the MPEC formulation (10) is a non-convex optimization problem that is difficult to solve. In this section, an exact solution algorithm will be developed by converting the original problem into a series of much simpler problems, each corresponding to a unique departure order in all arrival windows. A heuristic method that aims to overcome the difficulty of enumerating all departure orders will also be presented.

4.1 Main idea

For each arrival window \( A_j \), \( \forall j = 0, \cdots, m^A, A = \{E, L\} \), the set of all user classes that pass the bottleneck in the window is called the Class Membership Set (CMS) of the window, denoted as \( \rho^{A_j} = (\rho^A_j, \rho^E_j, \cdots, \rho^n_j). \) \( \rho^A_j, \rho^E_j, \cdots, \rho^n_j \in \{0,1\}, (i = 0,1,\ldots,n) \) indicates whether a class \( i \) appears in windows \( A_j \). Note that both real user classes (Class 1 to \( n \)) and the dummy user class (Class 0) may appear in an arrival period. For \( A_0 \), an idle arrival period always exists, implying Class 0 should always be present in \( A_0 \). A vector \( \rho = \{\rho^{A_j}\} \) is called a Class Membership Set Realization (CMSR). We note that a CMSR corresponds to a unique route choice pattern in the transformed network shown in Figure 2. In other words, enumerating CMSR is effectively equal to enumerating all paths in the transformed network. Finally, \( \Sigma \) is used to denote the set of all possible CMSR.

Our idea is as follows. For each given \( \rho \), \( N \) and \( \mu \) can be solved from Equations (3-7), as functions of \([p, \bar{B}]\). Note that the user equilibrium conditions require the isocost curve of any classes present in an arrival window (in the transformed network, this means the class uses the route corresponding to the arrival window) is on the upper envelop in that window. It will become clear later that this requirement can be represented by a convex feasible set of \([p, \bar{B}]\) denoted as \( \Xi(\rho) \). For now, recalling that \([p, \bar{B}] \in \Gamma \) (\( \Gamma \) is defined in (11)), we can formulate the design problem corresponding to \( \rho \) as follows:

\[
\min_{[p, \bar{B}]} W_\rho(p, \bar{B}) \quad (15a)
\]

subject to

\[
[p, \bar{B}] \in \Gamma \cap \Xi(\rho) \quad (15b)
\]

where \( W_\rho(p, \bar{B}) \) is equivalent to \( W \) in equation (9) when a CMSR \( \rho \) is given. We shall show that Problem (15) is a quadratic program defined over a polyhedron, and that the optimal solution of the original toll design problem can be found by comparing the optimal solutions of these simpler quadratic programs. In what follows, we first define the objective function and the constraints for (15).

4.2 Objective function of the subproblem

We first show how \( W_\rho \) can be written as a quadratic function of \([p, \bar{B}]\), once \( \rho \) is given. Note that \( \rho^A_j = 1 \) if route \( A_j \) is used for the O-D pair \( i-i + 1 \) in Figure 2, and 0 otherwise. For \( i = 0 \), which corresponds the dummy user class, \( \rho^A_0 = 0(1) \) implies that the capacity constraint (6) is (not)
binding. Equations (3-7) can then be converted to the following linear equation system:

\[
\sum_{j=0}^{m^A} N_i^{A_j} \rho_i^{A_j} = N_i, \forall i = 1, \ldots, n \tag{16a}
\]

\[
C_i^{A_j} = \mu_i - \lambda_i^{A_j}, \forall i = 1, \ldots, n, \rho_i^{A_j} = 1, j = 0, \ldots, m^A, A = \{E, L\} \tag{16b}
\]

\[
N_i^{A_j} = 0, \forall i = 1, \ldots, n, \rho_i^{A_j} = 0, j = 0, \ldots, m^A, A = \{E, L\} \tag{16c}
\]

\[
\lambda_i^{A_j} = 0; \forall j = 0, \ldots, m^A, A = \{E, L\}, \rho_0^{A_j} = 1 \tag{16d}
\]

\[
\sum_{i=1}^n N_i^{A_j} = s_i^{A_j}, \forall j = 0, \ldots, m^A, A = \{E, L\}, \rho_0^{A_j} = 0 \tag{16e}
\]

where \( \lambda_i^{A_j} \) is the multiplier corresponding to the capacity constraint (6). The above system has \((2n + 1)(m^E + m^L + 2) + n\) unknowns \((N, C, \mu, \lambda)\), and \((n + 1)(m^E + m^L + 2) + n\) equations. Recall Equation (3) defines the relationship between \(N, C, \mu, \lambda\), which gives another \(n(m^E + m^L + 2)\) equations. Hence, we can write (through Gaussian elimination for example)

\[
N_i^{A_j} = \sum_{B = \{E, L\}} \sum_{k=1}^{m^B} \left( \theta_i^{A_j} p_i^{B_k} + \eta_i^{A_j} n_k^{B_k} \right) + \theta_i^{0} A_j, \forall i = 1, \ldots, n, \rho_i^{A_j} = 1, j = 0, \ldots, m^A, A = \{E, L\} \tag{17a}
\]

\[
\mu_i = \sum_{A = \{E, L\}} \sum_{j=1}^{m^A} \left( \sigma_i^{A_j} p_i^{A_j} + \phi_i^{A_j} n_i^{A_j} \right) + \sigma_i^{0} A_j, \forall i = 1, \ldots, n \tag{17b}
\]

where \( \theta_i^{A_j}, \eta_i^{A_j}, \theta_i^{0}, \sigma_i^{A_j}, \phi_i^{A_j} \) and \( \sigma_i^{0} \) are coefficients. Thus,

\[
W_p = \sum_i \left\{ \alpha_i N_i \left( \sum_{j=1}^{m^A} \left( \sigma_i^{A_j} p_i^{A_j} + \phi_i^{A_j} n_i^{A_j} \right) + \sigma_i^{0} \right) - \sum_{j=1}^{m^A} \pi_i^{A_j} \rho_i^{A_j} \left( \sum_{B = \{E, L\}} \sum_{k=1}^{m^B} \left( \theta_i^{A_j} p_i^{B_k} + \eta_i^{A_j} n_k^{B_k} \right) + \theta_i^{0} A_j \right) \right\}, \tag{18}
\]

which is a quadratic but potentially non-convex function of \([p, \pi]\).

The above approach requires reducing a linear equation system with \((2n + 1)(m^E + m^L + 2) + n\) unknowns, which can be a considerable computational burden. The approach first proposed by Chen et al. (2014) provides a more efficient alternative. For narrative convenience, the approach is described in what follows. Note that the toll design parameters are treated as inputs in Chen et al. (2014), whereas they are considered solution variables in this paper. Therefore, while the formulas may seem similar, they are interpreted differently, and perhaps more important, used differently in guiding the design of the heuristic algorithm.

Let us first write the isocost curve of class \(i\) in the arrival window \(A_j\) as

\[
y = b_i^{A_j} + a_i^{A_j} t_r \tag{19}
\]

where \(b_i^{A_j} = \mu_i - \frac{\pi_i^{A_j}}{s_i} \) and \(a_i^{A_j} = \frac{\mu_i}{s_i} \) if \(A = E\) and \(a_i^{A_j} = -\frac{\mu_i}{s_i} \) if \(A = L\).

For the given \(\rho\), let \(n_D^{A_j}\) be the number of classes in \(A_j\), and \(D^{A_j}(i)\) be the ratio rank ID of the class whose time flexibility ranks at the \(i^{th}\) place among all \(n_D^{A_j}\) presenting class in the arrival period \(A_j\), then the class ID of \(i^{th}\) class in \(A_j\) is given by \(A(D^{A_j}(i))\).
classes out of five appear in Figure 3: Illustration of isocost curves in the arrival window. The same analytical solutions will simplify the reduction of the linear system significantly. However, both approaches will lead to a linear equation system with the only unknowns being $\mu_i$, the general class cost measured in time unit. Clearly, the new system is much smaller compared to (16), which will simplify the reduction of the linear system significantly. However, both approaches will lead to the same analytical solutions.

Figure 3: Illustration of isocost curves in the arrival window $E_j$ from Chen et al. (2014): three classes out of five appear in $E_j$ for the given $\rho$.

We define

$$x_{E_j}^{A_{(D^j_i)}} = \begin{cases} 
-p_{j}^{E_j}, & i = 1 \\
-\frac{b_{j}^{E_j}}{a_{j}^{E_j}} - \frac{b_{j+1}^{E_j}}{a_{j+1}^{E_j}}, & i = 2, \ldots, n_{D}^{E_j} + 1 \\
-p_{j+1}^{E_j}, & i = n_{D}^{E_j} + 1 
\end{cases};

x_{A_{(D^j_i)}}^{L_j} = \begin{cases} 
-p_{j}^{L_j}, & i = 1 \\
-\frac{b_{j}^{L_j}}{a_{j}^{L_j}} - \frac{b_{j+1}^{L_j}}{a_{j+1}^{L_j}}, & i = 2, \ldots, n_{D}^{L_j} + 1 \\
p_{j+1}^{L_j}, & i = n_{D}^{L_j} + 1 
\end{cases};

(20)

where $x_{A_{(D^j_i)}}^{A_{(D^j_i)}}$ and $x_{A_{(D^j_i)}}^{A_{(D^j_i)}}$ ($A = E, L$) represent the boundaries of the arrival window $A_{ji}$, and $x_{A_{(D^j_i)}}^{A_{(D^j_i)}}$ is the $t$ coordinate of the point where the isocost curves of $A(D^j_i(i - 1))$ and $A(D^j_i(i))$ intersect. In Figure 3, an example is provided in which three out of five user classes appear in $E_j$. The figure clearly shows that, once the above coordinates are located, the amount of flows for each class of commuters in $A_{ji}$ can be computed as

$$N_{A_{(D^j_i)}}^{A_{(D^j_i)}} = s(x_{A_{(D^j_i)}}^{A_{(D^j_i)}} - x_{A_{(D^j_i)}}^{A_{(D^j_i)}}), i = 1, 2, \ldots, n_{D}^{A_{(D^j_i)}}.$$

Evidently, those classes who do not appear in the window according to $\rho$ would have zero flow. Using this approach, all class flows in any window $A_{ji}$ are expressed as a function of $\mu_i$ and the step toll parameters $(p, b)$ in a closed form. Invoking the flow conservation conditions (5) then leads to an $n$ by $n$ linear equation system with the only unknowns being $\mu_i$, the general class cost measured in time unit. Clearly, the new system is much smaller compared to (16), which will simplify the reduction of the linear system significantly. However, both approaches will lead to the same analytical solutions.
4.3 Feasibility set defined by isocost curves

As shown in Figure 3, an arbitrarily determined \( \rho \) may not ensure every class thought to be present in an arrival window has its isocost curve stay on the upper envelope, which is an inherent requirement of the dynamic equilibrium conditions. To enforce this requirement, additional feasibility cuts must be added to form the aforementioned feasible set \( \Xi(\rho) \) (Constraint (15b)) for \([\mathbf{p}, \pi]\).

First, for any class that does appear in \( A_j \), its flow defined in (21) must be nonnegative, which leads to the following constraints:

\[
x^A_j A(D^A_j(i) + 1) \geq x^A_j A(D^A_1(i)), i = 1, 2, \ldots, \, n^A_j. (22)
\]

Second, for any class that does not appear in \( A_j \), we must ensure that its isocost curve is “dominated” by those of the classes that do appear. We classify these “missing classes” into three categories.

For those whose \(|a^A_k|\) (the absolute value of the slope of the isocost curve) are smaller than \(|a^A_{A(D^A_j(1))}|\), the \( y \) intercept of their isocost curves at the far end of the arrival window must be lower than that of \( A(D^A_j(1)) \), i.e.

\[
\begin{align*}
\left\{ \begin{array}{l}
\rho^A_j A(D^A_j(1)) - a^A_j A(D^A_j(1)) p^A_j \geq b^A_j - a^A_j p^A_j, \quad A = E \\
\rho^A_j A(D^A_j(1)) + a^A_j A(D^A_j(1)) p^A_j \geq b^A_j + a^A_j p^A_j, \quad A = L,
\end{array} \right. \quad k = 0, A(1), \ldots, A(D^A_j(1) - 1). (23)
\end{align*}
\]

For the missing classes whose \(|a^A_k|\) satisfies the following condition

\[
|a^A_{A(D^A_j(1))}| < |a^A_k| < |a^A_{A(D^A_j(1) + 1)}|,
\]

the intersection of its isocost curve with that of \( A(D^A_j(i + 1)) \) must lie further away from \( t^* \) compared to the intersection between isocost curves of \( A(D^A_j(i)) \) and \( A(D^A_j(i + 1)) \). Mathematically this requirement can be written as follows:

\[
\begin{align*}
\left\{ \begin{array}{l}
x^A_j A(D^A_j(i + 1)) \left( a^A_j A(D^A_j(i)) - a^A_j \right) \geq b^A_j A(D^A_j(i)) - b^A_j, \quad A = E \\
x^A_j A(D^A_j(i + 1)) \left( a^A_j A(D^A_j(i)) - a^A_j \right) \leq b^A_j A(D^A_j(i)) - b^A_j, \quad A = L
\end{array} \right. \quad i = 1, 2, \ldots, n^A_j - 1, k = A(D^A_j(i + 1)), \ldots, A(D^A_j(i + 1) - 1). (24)
\end{align*}
\]

Finally, for all the missing classes whose \(|a^A_k| > |a^A_{A(D^A_j(n^A_j))}|\), the \( y \) intercept of its isocost curve with the boundary of the arrival windows near to \( t^* \) must be lower than that of \( A(D^A_j(n^A_j)) \):

\[
\begin{align*}
\left\{ \begin{array}{l}
\rho^A_j A(D^A_j(n^A_j)) - a^A_j A(D^A_j(n^A_j)) p^A_{j+1} \geq b^A_j A(D^A_j(n^A_j)) - a^A_j p^A_{j+1}, \quad A = E \\
\rho^A_j A(D^A_j(n^A_j)) + a^A_j A(D^A_j(n^A_j)) p^A_{j+1} \leq b^A_j A(D^A_j(n^A_j)) + a^A_j p^A_{j+1}, \quad A = L,
\end{array} \right. \quad k = A(D^A_j(n^A_j)), \ldots, n. (25)
\end{align*}
\]
Combining Equations (22)-(25), there are in total \((n + 1)(m^E + m^L + 2)\) linear constraints, each corresponding to exactly one class (including the dummy user) in each arrival window. We thus formally define

\[
\Xi(\rho) = \{ [p, \pi] | [p, \pi] \text{ satisfies Equations (22) – (25)} \}.
\]

Since Equations (22)-(25) are all linear in \(p\) and \(\pi\), \(\Xi(\rho)\) is a polyhedral set. Because \(\Gamma\) is also polyhedral, Problem (15) is defined on a polyhedral set.

### 4.4 Optimality and the exact algorithm

We now formally state and prove the following result, which is at the core of the proposed solution method.

**Proposition 1 (Optimality).** Let \(W^*\) be the optimal solution to the MPEC formulation of the optimal toll design problem (10).

\[
W^* = \min_{\rho \in \Sigma} \left( \min_{[p, \theta] \in \Gamma \cap \Xi(\rho)} W_\rho(p, \theta) \right)
\]

**Proof:** We first show that \(\bigcup_{\rho \in \Sigma} \Xi(\rho)\) covers the entire \(\Gamma\), the feasible set of \([p, \theta]\). Note that, for any given \([p, \theta] \in \Gamma\), Lindsey (2004) proves that a unique tolled equilibrium always exist. Since \(\Sigma\) enumerates all possible CMSR, the existence of equilibrium implies that we must be able to find a \(\rho\) such that \([p, \theta] \in \Xi(\rho)\). In other words, every point in \(\Gamma\) must also be in a \(\Xi(\rho)\). Second, we show that for \(\rho_1 \neq \rho_2\), the intersection \(\Xi(\rho_1) \cap \Xi(\rho_2) = \emptyset\). Suppose that the intersection is not empty, then any \([p, \theta] \in \Xi(\rho)\) in the intersection corresponds to two different UE solutions, which contradicts to the fact that such solution is unique given \([p, \theta]\) is given. Therefore, solving the subproblem (15) for each \(\rho \in \Sigma\) would search every part of \(\Gamma\) once (and only once) for the minimum system cost, while satisfying the equilibrium conditions (imposed through the objective function and \(\Xi(\rho)\)). This completes the proof. \(\square\)

For clarity, a detailed procedure of the enumeration method is summarized in Algorithm 1.

### 4.5 Heuristics

The exact algorithm above decomposes the network design problem into a very large number of quadratic programs. To be precise, the total number of these programs is on the order of \(2^n(m^E + m^L + 2)^n\), which is astronomically large even for relatively small \(n\) and \(m^A\). This is not a surprise since the original problem is known to be NP-hard. Moreover, while the quadratic program is defined on a polyhedral set, its objective function may not be convex, and therefore the problem of finding its global optimal is NP-hard itself. Thus, solving the toll design problem for any non-trivial \(n\) and \(m^A\) would have to involve some kind of heuristics.

The idea pursued here is related to the heuristic algorithm developed by Chen et al. (2014) for solving step-tolled user equilibrium for the same problem. The algorithm starts from a pre-defined \(\rho\), and finds a corresponding UE solution. It then identifies in each arrival windows the classes whose isocost curves stay on the upper envelope, and use that information to assemble
Algorithm 1 Exact method for the toll design problem (10)
1: Output: $W^*, [p^*, \pi^*]$.
2: Initialize:
3: Create the set $\Sigma = \{\rho^1, \rho^2, ..., \rho^K\}$, where $K = |\Sigma|$. Set iteration index $k = 1$, and $W^* = +\infty$.
4: while $k < K$ do
5: Set a linear equation system with respect to $\mu$ using Equations (20), (21) and (6) based on $\rho^k$;
6: if the linear equation system is solvable then
7: Solve the linear equation system to find the coefficients $\sigma^A_j$, $\phi^A_j$ and $\sigma_0^A_j$ as in Equation (17b);
8: Find coefficients $\theta^A_j$, $\eta^A_j$, $\theta_0^A_j$ as in Equation (17a), based on Equations (21).
9: Create the quadratic objective function $W_{\rho^k}$ based on the coefficients.
10: Set $k = k + 1$.
11: end if
12: Set $W^*_{\rho^k} = +\infty$.
13: while $W^*_{\rho^k} < W^*$ do
14: Find coefficients $\theta^A_j$, $\eta^A_j$, $\theta_0^A_j$ as in Equation (17a), based on Equations (21).
15: Create the quadratic objective function $W_{\rho^k}$ based on the coefficients.
16: Solve the quadratic program (15) to find the global optimum $W^*_{\rho^k}$ and $[p^*, \pi^*]$.
17: end while

a new $\rho$, which is then examined in the next iteration. Their preliminary computational experiments indicate that this simple idea consistently locates optimal solutions after only checking a very small subset of $\Sigma$. However, this method cannot be directly applied to the toll design problem, since the upper envelope is not readily available, but rather consists of linear functions of $[p, \pi]$.

The proposed heuristic method is built on two conjectures. First, the overall solution quality depends more on finding the right $\rho$ than solving each quadratic program to its global optimality. While the objective function is non-convex in general, it may well be convex within the feasible set for a given $\rho$. Even if it is indeed non-convex, being trapped at a local optimum may still be inconsequential as long as the multipliers are effective in steering the search for the right $\rho$. Second, for a given $\rho$, the multipliers associated with the constraints in $\Xi(\rho)$ will offer useful information about which of the $\rho \in \Sigma$ should be examined in the next round. Let $\kappa^A_i$, $i = 0, \ldots, n, j = 1, \ldots, m^A, A = \{E, L\}$ be these multipliers, obtained after solving the quadratic program (15) to a local optimum. Note that a non-zero multiplier implies that the corresponding constraint is binding, and the objective function may be improved should the constraint be relaxed. Since each of the constraints in $\Xi(\rho)$ defines whether a class $i$ should appear in $A_j$ or not, relaxing a constraint essentially means flipping the appearance status of a class in a given arrival window.

The question is: when there are more than one non-zero multipliers, which ones should be

---

This is not always true, since the objective function may not be differentiable at the boundary. The reason is that crossing these boundaries may lead to a different CMSR, and hence a different objective function.
chosen to flip the appearance status of the associated classes? This choice is potentially critical to the overall performance of the heuristic method. In this study, the procedure described in Algorithm 2 is used to make this choice. The algorithm basically flips the appearance status of the class that has the maximum non-zero multiplier in each arrival time, subject to the requirement that any class must appear in at least one time window in the new $\rho$ (Lines 9 - 11 in Algorithm 2 provide this protection).

**Algorithm 2** Method for determining the next $\rho$ from $\lambda$

1: **Input:** multipliers $\kappa$ and the current CMSR $\rho$.
2: **Output:** the next CMSR $\rho'$
3: **Initialize:**
4: Set $z_i = \sum_j \sum_k A_{ij}^{Ak}$, the number of times class $i$ appears in all windows. Set $\rho' = \rho$.
5: for all $A_{ij}$ do
6: Set $I = \arg\max_i \{x_i^j | \kappa_i^j > 0\}$.
7: if $I$ is not empty then
8: if $\rho_i^{A_{ij}} = 1$ then
9: if $z_i > 0$ then
10: $\rho_i^{A_{ij}} = 0, z_i = z_i - 1$
11: end if
12: else
13: $\rho_i^{A_{ij}} = 1, z_i = z_i + 1$.
14: end if
15: end if
16: end for

It follows from Algorithm 2 that the heuristic method should be terminated if all multipliers in the constraints of $\Xi(\rho)$ are zero, because no new $\rho$ can be discovered. However, the optimal solution may be on the boundary of $\Xi(\rho)$, where multipliers can be greater than zero. Such a boundary solution may arise, for example, when the appearance status of one class in a window affects neither the optimal toll design nor the total system cost. To avoid potential cycling over such boundary solutions endlessly, an additional termination criterion should be introduced. The strategy adopted in this paper is to keep track of recent solutions so that such cycles can be detected.

Algorithm 3 describes the heuristic method in details. The heuristic method differs from the exact method in two aspects. First, the heuristic method only solves the quadratic program (15) to a local optimum. Second, it does not enumerate the space of $\Sigma$. Rather, it starts from a randomly selected $\rho$ and terminates when (1) all constraints in $\Xi(\rho)$ have zero multipliers (leading to $\rho^{k*} = \rho^k$, Line 14 in Algorithm 3); or (2) a cycle is detected (Line 17 - 25 in Algorithm 3).

A couple of other remarks are also in order here. First, note that if the linear system created for solving $\mu$ is degenerate, a new random point will be picked to restart the search (Line 18 in Algorithm 3). Second, when the termination condition (Line 14 in Algorithm 3) is met, to avoid being trapped at a local optimum, one may attempt to resolve the current quadratic program with different starting points, and reevaluate the termination condition. This extension to Algorithm
3 is not considered here, but its implementation should be straightforward.

5 Numerical experiments

5.1 A two-class with a single-step-toll example

To better illustrate the proposed methodology, a simple two-class with a single-step-toll example is presented here. For convenience, we set $\alpha_1 = \alpha_2$. Since there is only a single step toll, we have $m^E = m^L = 1$ and $\pi^E_1 = \pi^L_1$. For simplicity, we define $\pi_1 = \pi^E_1 = \pi^L_1$. The capacity of the bottleneck ($s$) is 1000 vph, and the detailed information of user classes is given in Table 1.

| Class No. | Demand | $a_i$ ($$/hour) | $\beta_i$ ($$/hour) | $\gamma_i$ ($$/hour) | $\beta_i \div a_i$ | $\alpha_i$ |
|-----------|--------|-----------------|--------------------|----------------------|-------------------|---------|
| 1         | 1000   | 6               | 3                  | 12                   | 1/2               | 2       |
| 2         | 1500   | 6               | 4                  | 24                   | 2/3               | 4       |

When $\pi_1 = 0$, the costs of the first and second classes at user equilibrium are $6$ and $96$, respectively, which correspond to a total social cost $W_{UE} = 17077$. Details of solving no-toll equilibrium is omitted here for brevity; the reader is referred to Chen et al. (2014) for details.

It is easy to verify that, even in such a trivial example, there are in total 784 CMSRs ($|\Sigma| = 784$), which makes it almost impossible to examine all scenarios analytically. For the purpose of illustration, we will focus on the following CMSR, which happens to give the global optimum:

$$\rho^0 = \{\rho^{E_0}, \rho^{E_1}, \rho^{L_1}, \rho^{L_0}\} = \{(1,1,1), (0,0,1), (0,1,1), (1,1,0)\}.$$ 

Note that $E_0$ is the no-toll early arrival window, $E_1$ is the tolled early arrival window, $L_1$ is the tolled late arrival window and $L_0$ is the no-toll late arrival window. Basically, $\rho^*$ place both classes, as well as the dummy class in $E_0$, class 2 in $E_1$, both classes in $L_1$, and class 1 and the dummy class in $L_0$.

The reader can verify that invoking Equation (21) will yield

\[
N_{1E_0} = 8000\mu_1 - 6000\mu_2, \\
N_{1L_1} = 500\mu_1 - 500\mu_2 + 1000p_1^L, \\
N_{2E_0} = -6000\mu_1 + 6000\mu_2 - 1000p_1^E, \\
N_{2L_1} = -500\mu_1 + 500\mu_2,
\]

Using the flow conservation conditions the following equation system of $\mu_1$ and $\mu_2$ can be constructed

\[
N_1 = 1000 = N_{1E_0} + N_{1E_1} + N_{1L_1} + N_{1L_0} = 9000\mu_1 - 6500\mu_2; \\
N_2 = 1500 = N_{2E_0} + N_{2E_1} + N_{2L_1} + N_{2L_0} = -6500\mu_1 + 6500\mu_2.
\]

from which $\mu_1$ and $\mu_2$ can be solved as

$$ (\mu_1, \mu_2) = \left( 1, \frac{16}{13} \right). $$

16
Algorithm 3 Heuristic method for the toll design problem (10)

1: **Input:** Maximum number of main iteration $M$.
2: **Output:** $W^*, [p^*, \beta^*]$.
3: **Initialize:**
4: Set iteration index $k = 0$ and choose the initial CMSR $\rho^k$. Set $F = 0, W = -1, \Delta = \emptyset$.
5: **while** $k < M$ and $F = 0$ **do**
6: Set a linear equation system with respect to $\mu$ using Equations (20), (21) and (6) based on $\rho^k$.
7: if the linear equation system is solvable then
8: Solve the linear equation system to find the coefficients $\sigma^A_{i,j}$, $\phi^A_{i,j}$ and $\sigma^A_0$ as in Equation (17b).
9: Find coefficients $\eta^{A_i} \eta^{A_j} \theta^A_{i,j}$ as in Equation (17a), based on Equations (21).
10: Create the quadratic objective function $W_{\rho^k}$ based on the coefficients.
11: Specify $\Gamma \cap \Xi(\rho^k)$, where $\Gamma$ and $\Xi(\rho^k)$ are defined in Equations (11) and (26), respectively.
12: Solve the quadratic program (15) to find a local optimum $W^*_{\rho^k}, [p^{k*}, \pi^{k*}]$ and multipliers $\kappa^{A_{i,j}}$.
13: Call Algorithm 2 to get the next $\rho$, denoted as $\rho^{k*}$.
14: if $\rho^{k*} = \rho^k$ then
15: Set $[p^*, \beta^*] = [p^{k*}, \pi^{k*}], W^* = W_{\rho^k}$ and $F = 1$.
16: else
17: if $W_{\rho^k} = W$ then
18: if $\rho^k \in \Delta$ then
19: A cycle detected. Set $[p^*, \beta^*] = [p^{k*}, \pi^{k*}], W^* = W_{\rho^k}$ and $F = 1$.
20: else
21: Set $\Delta = \Delta \cup \rho^k$.
22: end if
23: else
24: $W = W_{\rho^k}, \Delta = \emptyset$.
25: end if
26: end if
27: else
28: Set $\rho^{k+1}$ as a randomly selected CMSR (restart from a new point).
29: end if
30: Set $\rho^{k+1} = \rho^{k*}$. Set $k = k + 1$.
31: **end while**
It is an interesting coincidence that the class travel cost \( \mu \) is constant, and not a function of any step toll parameters in this case. This is not the case in general, as shown later. With all \( \mu_i \) and \( N_i \) defined, the quadratic program associated with \( \rho^0 \) can now be written as:

\[
\begin{align*}
\min_{(\pi_1, p_1^E, p_1^L)} & \quad W = -1000\pi_1 p_1^E - 1000\pi_1 p_1^L + \frac{222000}{13} \\
\text{subject to:} & \\
\pi_1 + \frac{2}{3} p_1^E & \leq \frac{16}{13} \quad \text{(Constraint for Class 0 in } E_1); \\
\pi_1 + 2p_1^L & \leq 1 \quad \text{(Constraint for Class 0 in } L_1); \\
\pi_1 & \geq 0; \\
0 & \leq p_1^E \leq \frac{18}{13} \quad \text{(Constraints for Class 2 in } E_0 \text{ and Class 1 in } E_1); \\
\frac{3}{26} & \leq p_1^L \leq \frac{1}{2} \quad \text{(Constraints for Class 1 in } L_1 \text{ and } L_0). 
\end{align*}
\] (31a)

The other constraints are omitted because they are automatically satisfied after \( \mu \) is plugged in.

The optimal solution for this program is

\[
\begin{align*}
W^* & = 12949; \\
(\pi_1, p_1^E, p_1^L)^* & = \left( \frac{183}{52}, \frac{201}{208}, \frac{43}{208} \right). 
\end{align*}
\] (32a)

Figure 4 plots the isocost curves as well as the cumulative departure/arrival curves for the user equilibrium solution with the above single-step optimal toll scheme. Note that the first commuter of class 2 in \( E_1 \) and the class commuter of class 1 in \( L_1 \) have no travel delays, and there is no idle periods in these two arrival windows (i.e. the dummy class is not present).

The multipliers associated with the constraints (31c) and (31d) are \( 5278.8 \) and \( 1759.6 \), respectively, which suggests that placing class 0, i.e. the dummy class, in either \( E_1 \) or \( L_1 \) may improve the system cost. We examine these solutions in the following, which are \( \rho_1 - \rho_3 \) shown in Table

| Table 2: Cyclic optimal solutions in the two-class with a single-step-toll example |
|-----------------|-----------------|-----------------|-----------------|
| \( \rho^E_0 \)  | \( \rho^E_1 \)  | \( \rho^L_1 \)  | \( \rho^L_0 \)  |
| \( \rho_0 \)    | 1, 1, 1         | 0, 0, 1         | 0, 1, 1         | 1, 1, 0         |
| \( \rho_1 \)    | 1, 1, 1         | 1, 0, 1         | 0, 1, 1         | 1, 1, 0         |
| \( \rho_2 \)    | 1, 1, 1         | 1, 0, 1         | 1, 1, 1         | 1, 1, 0         |
| \( \rho_3 \)    | 1, 1, 1         | 0, 0, 1         | 1, 1, 1         | 1, 1, 0         |

For \( \rho_1 \) in Table 2, following the above procedure will yield

\[
(\mu_1, \mu_2) = \left( \frac{13}{238} \pi_1 + \frac{26}{119} p_1^E + \frac{71}{119} \pi_1 + \frac{9}{119} \pi_1 + \frac{36}{119} p_1^E + \frac{80}{119} \right)
\]

Note that for this \( \rho, \mu \) is indeed the function of the step toll parameters. The quadratic program
in this case reads

$$
\min_{(\pi_1, p_1^E, p_1^L)} W^1 = \frac{16250}{119} \pi_1^2 - \frac{54000}{119} \pi_1 p_1^E - 1000 \pi_1 p_1^L + \frac{480000}{119} p_1^E + \frac{1146000}{119} \pi_1
$$

subject to:

$$
\begin{align*}
-\pi_1 - 4p_1^E + 44 &\geq 0 \quad \text{(Constraint for Class 1 in } E_0); \\
15\pi_1 - 59p_1^E + 54 &\geq 0 \quad \text{(Constraint for Class 2 in } E_0); \\
13\pi_1 + 52p_1^E - 96 &\geq 0 \quad \text{(Constraint for Class 0 in } E_1); \\
125\pi_1 + 24p_1^E - 264 &\geq 0 \quad \text{(Constraint for Class 1 in } E_1); \\
-65\pi_1 + 216p_1^E + 480 &\geq 0 \quad \text{(Constraint for Class 2 in } E_1); \\
-\frac{40}{119}\pi_1 + \frac{78}{119} p_1^E - 6p_1^L + \frac{213}{119} &\geq 0 \quad \text{(Constraint for Class 0 in } L_1); \\
-\frac{15}{119}\pi_1 + \frac{60}{119} p_1^E + 12p_1^L - \frac{54}{119} &\geq 0 \quad \text{(Constraint for both Class 1 in } L_1 \text{ and Class 2 in } L_0); \\
\pi_1 &\geq 0; \quad p_1^E \geq 0; \quad p_1^L \geq 0.
\end{align*}
$$

Again, other constraints are automatically satisfied with the definition of $\mu$. Interestingly, the optimal solution for this problem is identical to what is given in (32), despite the quadratic program appears to be completely different. While dummy class is placed in $E_1$, it does not affect the solution because the optimality requires reducing the length of the idle period to zero. However, the multipliers associated with Constraints (33e) and (33h) are 2052.2 and 384.5,
respectively.

We will leave it to the reader to verify that for $\rho_2$

$$ (\mu_1, \mu_2) = \left( \frac{11}{162} \pi_1 + \frac{26}{135} p_E^1 + \frac{32}{135} p_L^1 + \frac{71}{135} \pi_1 + \frac{7}{81} \pi_1, \frac{38}{135} p_E^1 + \frac{26}{135} p_L^1 + \frac{83}{135} \right), $$

and for $\rho_3$,

$$ (\mu_1, \mu_2) = \left( \frac{1}{36} \pi_1 + \frac{1}{3} p_E^1 + \frac{5}{6} \pi_1 + \frac{1}{3} p_L^1 + \frac{83}{78} \right). $$

Again, both $\rho_2$ and $\rho_3$ lead to the exactly same optimal solution as given in (32), although they present different quadratic programs. Clearly, the multipliers will guide the solution process to cycle around the $\rho$ defined in Table 2, without changing the actual solution. This problem is precisely the reason why a cycle detection mechanism is introduced in Algorithm 3.

What other insights can one learn from the above observation? The fact that only the dummy class and the tolled windows are involved here is quite intriguing. Recall that the existence of a dummy class in a tolled window implies that the bottleneck would have an idle period. Removing such an idle period can improve the system cost, which explains why the multiplier is positive when there a dummy class in the window (see the case for $\rho_1$ above). On the other hand, when the dummy class is not in the window, the first (last for late arrival) appearing class must have a $y$ intercept at the far end of the window that is larger than or equal to zero (see Figure 3). A positive $y$ means that the first (last) commuter arriving in an early (late) arrival window has a non-zero travel delay. Accordingly, the system cost may be reduced by eliminating such delays, which can be achieved by introducing a dummy class. The final solution settles at the boundary point where the first (last) commuter arriving in an early (late) arrival window has a non-zero travel delay, and the idle period corresponding to the appearance of the dummy class is also zero (see Figure 4). We note that (1) these properties are consistent with those known for the system optimal flow patterns under special user heterogeneity (see e.g. Xiao et al., 2011, 2012), and (2) the cyclic solution pattern revealed here may be caused by the inherent requirement of these properties.

### 5.2 Impact of design flexibility

We now allow the tolls in the periods $E_1$ and $L_1$ to take different values, which essentially turns a single-step toll design problem into a multi-step toll design problem. The purpose is to examine how this increased design flexibility may affect the outcomes.

It is found (details omitted here) that the new design problem has the exactly the same optimal CMSR as the single-step-toll problem, but produces a different optimal toll scheme. As shown in Table 3, the new design slightly expands the toll window in the late arrival period, while reducing the amount of toll. It does exactly the opposite to the early arrival window. As expected, the system cost is improved as the flexibility increases, although the improvement is almost negligible in this case.
Table 3: Optimal solution comparison for single-step-toll and two-step-toll

| Number of Step Tolls | Optimal CMSR | Optimal Toll Scheme | Social Cost |
|----------------------|--------------|---------------------|-------------|
|                      | $\rho^*$     | $\pi_E^T$ ($)       | $\pi_L^T$ ($) | $p_E^T$ (hour) | $p_L^T$ (hour) | $W^*$ |
| 1                    | \{(1,1,1), (0,0,1), (0,1,1), (1,1,0)\} | 3.5192 | 3.5192 | 0.9663 | 0.2067 | 12949 |
| 2                    | \{(1,1,1), (0,0,1), (0,1,1), (1,1,0)\} | 3.6923 | 3.0000 | 0.9231 | 0.2500 | 12919 |

5.3 General user heterogeneity

The previous examples assume $\alpha_1 = \alpha_2$ for simplicity. We now consider a different example that has more general heterogeneity, defined in Table 4.

Table 4: Class information in the example with more general heterogeneity

| Class No. | Demand $N_i$ | Values of Time $\alpha_i$ ($/\text{hour}$) | $\beta_i$ ($/\text{hour}$) | $\gamma_i$ ($/\text{hour}$) | $\hat{\beta}_i$ | $\hat{\gamma}_i$ |
|-----------|-------------|------------------------------------------|---------------------------|---------------------------|----------------|----------------|
| 1         | 1000        | 10                                       | 3                         | 15                        | 3/10           | 3/2            |
| 2         | 700         | 8                                        | 7                         | 32                        | 7/8            | 4              |

With the help of the heuristic method, we identified three solutions that satisfy the termination criteria for a local optimum. The existence of multiple local optima is indeed expected, given the non-convex nature of the problem. Table 5 compares these local optima with the no-toll user equilibrium solution, and Figure 5 plots their isocost curves. In all three cases shown in the figure, there is a jump in the isocost curve at $t^*$, which is created by the jump in the optimal toll before and after $t^*$. We note that for the homogeneous case, the optimal toll price would not have such a jump around time $t^*$ since the optimal solution will set the toll unchanged at time $t^*$. These upward jumps, which imply that the first user in the next arrival window will indeed depart earlier than the last user in the previous arrival window, violate the FIFO principle around the times when the toll is reduced. However, since the SW assumption provides a separate waiting lane for users willing to arrive in the next arrival window and pay a lower toll, the FIFO principle is preserved within each lane, which is exactly how the SW assumption deals with the discontinuity problem.

Table 5: Optimal solution comparison for the case with more general heterogeneity

| Schemes | $\rho^*$ | $\pi_E^T$ ($) | $\pi_L^T$ ($) | $p_E^T$ (h) | $p_L^T$ (h) | $W^*$ | ($\mu_1, \mu_2$) |
|---------|----------|--------------|--------------|-------------|-------------|-------|-----------------|
| No-toll | \{(1,1,1), (1,1,1)\} | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 8462.5 | (0.425, 0.752) |
| Toll-1  | \{(1,1,1), (0,0,1), (0,1,1), (1,1,0)\} | 3.0473 | 2.0508 | 0.4353 | 0.1466 | 6889.0 | (0.425, 0.762) |
| Toll-2  | \{(1,1,1), (0,1,1), (0,1,1), (1,1,0)\} | 2.7528 | 2.0167 | 0.4991 | 0.1489 | 6889.2 | (0.425, 0.770) |
| Toll-3  | \{(1,1,0), (0,1,1), (0,1,1), (1,1,0)\} | 1.9543 | 1.9287 | 0.7652 | 0.1548 | 6941.5 | (0.425, 0.801) |

The three local optima correspond to different CMSR and toll schemes, but generate similar total system costs. In fact, the first two solutions almost yield the same system cost. The three solutions also have similar welfare effects, namely, class 1 breaks even while class 2 is slightly worse off with the step toll. This result indicates that the commuters with stronger schedule inflexibility are likely to suffer from congestion pricing, which agrees with the findings in literature.
Figure 5: Equilibrium solutions corresponding to different local optimal toll schemes for the example with more general heterogeneity
(e.g. van den Berg and Verhoef, 2011b). Finally, the fact that these local optima have similar total system costs is good news, because it suggests that the efficiency of the step tolls may not be very sensitive to design parameters and that the potential loss suffered from “being trapped” at a local optimum may be relatively small.

5.4 Convergence performance of the heuristic algorithm

As explained before, the exact algorithm will have to check 784 scenarios in order to solve the above two examples correctly. However, our experiments indicate that the heuristic method can solve these problems with much better efficiency. Table 6 summarizes the convergence performance of the heuristic method in solving the two problems reported in Sections 5.1 and 5.2. Note that for both problems, only four iterations are needed to find the optimum solution, and another three to confirm it (i.e. checking all cyclic solutions). Importantly, such a desirable performance seems insensitive to the initial solution. In fact, even when the initially selected CMSR leads to an infeasible quadratic program, the multipliers obtained from solving that infeasible program still successfully guide the remaining solution process.

While not tabulated here, we note that the performance of the heuristic algorithm in solving the problem in Section 5.3 is even more impressive. For almost all initial points tested, the algorithm needs only one iteration to find the optimum and another three for confirmation.

| Table 6: Convergence performance of the heuristic algorithm for the simple case |
|---|---|---|---|---|
| Feasible Starting CMSR | Infeasible Starting CMSR |
| No. Iteration | CMSR $\rho^k$ | Cost $W^k$ | CMSR $\rho^k$ | Cost $W^k$ |
| 1 | $\{(1,1,0),(1,1,1),(1,1,1),(1,0,0)\}$ | 14237 | $\{(1,0,0),(0,0,1),(0,1,1),(1,1,0)\}$ | N/A |
| 2 | $\{(1,1,0),(0,1,1),(1,1,1),(1,1,0)\}$ | 13882 | $\{(1,1,0),(0,1,1),(0,1,1),(1,1,0)\}$ | 13882 |
| 3 | $\{(1,1,1),(1,1,1),(0,1,1),(1,1,0)\}$ | 13882 | $\{(1,1,1),(1,1,1),(1,1,1),(1,1,0)\}$ | 13882 |
| 4 | $\{(1,1,1),(1,0,1),(1,1,1),(1,1,0)\}$ | 12949 | $\{(1,1,1),(1,0,1),(0,1,1),(1,1,0)\}$ | 12949 |
| 5 | $\{(1,1,1),(0,0,1),(0,1,1),(1,1,0)\}$ | 12949 | $\{(1,1,1),(0,0,1),(1,1,1),(1,1,0)\}$ | 12949 |
| 6 | $\{(1,1,1),(1,0,1),(0,1,1),(1,1,0)\}$ | 12949 | $\{(1,1,1),(1,0,1),(1,1,1),(1,1,0)\}$ | 12949 |
| 7 | $\{(1,1,1),(0,0,1),(1,1,1),(1,1,0)\}$ | 12949 | $\{(1,1,1),(0,0,1),(0,1,1),(1,1,0)\}$ | 12949 |

| Two-class Two-step-toll |
|---|---|---|---|---|
| Feasible Starting CMSR | Infeasible Starting CMSR |
| No. Iteration | CMSR $\rho^k$ | Cost $W^k$ | CMSR $\rho^k$ | Cost $W^k$ |
| 1 | $\{(1,1,1),(1,1,1),(1,1,1),(1,1,1)\}$ | 13988 | $\{(1,0,0),(0,0,1),(0,1,1),(1,1,0)\}$ | N/A |
| 2 | $\{(1,1,1),(0,1,1),(0,1,1),(1,1,1)\}$ | 13988 | $\{(1,1,0),(0,1,1),(0,1,1),(1,1,0)\}$ | 13771 |
| 3 | $\{(1,1,1),(0,0,1),(1,1,1),(1,1,1)\}$ | 13136 | $\{(1,1,1),(1,1,1),(1,1,1),(1,1,0)\}$ | 13771 |
| 4 | $\{(1,1,1),(1,0,1),(0,1,1),(1,1,0)\}$ | 12919 | $\{(1,1,1),(1,0,1),(0,1,1),(1,1,0)\}$ | 12919 |
| 5 | $\{(1,1,1),(0,0,1),(1,1,1),(1,1,0)\}$ | 12919 | $\{(1,1,1),(0,0,1),(1,1,1),(1,1,0)\}$ | 12919 |
| 6 | $\{(1,1,1),(1,0,1),(1,1,1),(1,1,0)\}$ | 12919 | $\{(1,1,1),(1,0,1),(1,1,1),(1,1,0)\}$ | 12919 |
| 7 | $\{(1,1,1),(0,0,1),(0,1,1),(1,1,0)\}$ | 12919 | $\{(1,1,1),(0,0,1),(0,1,1),(1,1,0)\}$ | 12919 |
6 Conclusions

We have proposed an analytical method to solve the optimal multi-step toll design problem for the bottleneck model with general user heterogeneity. The underlying design model is formulated as a mathematical program with equilibrium constraint, which is non-convex in terms of both the objective function and the feasible set. The proposed method decomposes the original model into a set of smaller and easier quadratic programs, each corresponding to a unique departure order of different user classes. While these quadratic programs may still have non-convex objective functions, they are all defined over a polyhedral set, which makes it easy to identify at least a local optimum. Importantly, a set of linear feasibility cuts is proposed for the quadratic program to define the appearance status of user classes in each arrival window. We prove that the proposed method ensures global optimality provided that each quadratic program can be solved globally. A heuristic method is then developed to obviate the impossible enumeration of all departure orders, which increases exponentially with respect to the number of user classes and step tolls. The heuristic method uses the multipliers associated with the feasibility cuts to navigate through the vast space of the departure order vector (so-called class membership set realization, or $\rho$). It terminates either when no positive multiplies can be found at all (an interior solution) or when the process begins to cycle around several essentially identical corner solutions.

The proposed methodology is validated using several small examples. The main findings from the numerical results are summarized below.

1. The results support the conjecture that, even under the general user heterogeneity, the system optimum solution should still have the following known properties (1) the bottleneck always operates at the capacity during the entire rush hour, and (2) the first (last) commuter in each early (late) tolled arrival window should be subject to no queuing delays.

2. In all tested experiments, the proposed heuristic method consistently locates local optimal solutions quickly, after searching only a small part of the feasible space, and the performance seems insensitive to the initial solution. Further studies are needed, however, to confirm the scalability of the method with large-scale experiments.

3. The optimal step-toll design problem can have multiple local optima. Yet, the local optimal solution identified in our example are similar in terms of their overall efficiency, which suggests that accepting a local optimal design may be satisfactory in practice.

A theoretical issue that has yet to be addressed is a proper assessment of the solution quality provided by the heuristic method, which may be accomplished by developing a good global lower bound for each of the quadratic programs. Such efforts will likely help improve the effectiveness and efficiency of the heuristic method. Another theoretical question has to do with proving/disproving the desirable properties of the system optimum solution under general user heterogeneity (cf. the finding 1 above). Utilizing these properties could significantly simplify the optimal toll design problem. A first step along this direction can be gathering empirical evidence through numerical experiments. This paper does not consider heterogeneity in desired arrival times. Yet, incorporating this feature is relatively straightforward, since a new desired arrival
time would simply break existing arrival windows into more pieces. While the problem size may hence increase quickly with the number of desired arrival times, the isocost curve for each class can still be analytically represented in a similar fashion. Hence, the nature of the formulation would not change. Last but not least, the methodology proposed in this paper provides a new tool that will enable the analysts to fully understand the impact of general user heterogeneity on the design and welfare effects of step tolls. Thus, various applications of this tool will be an important component of future studies.

Acknowledgement

The work was supported by National Science Foundation under the award numbers CMMI-1256021 and CMMI-1256106. Constructive comments of three anonymous reviewers are greatly appreciated. The mistakes are those of the authors alone.

References

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56–57.

Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. Journal of urban economics 27 (1), 111–130.

Arnott, R., de Palma, A., Lindsey, R., 1992. Route choice with heterogeneous drivers and group-specific congestion costs. Regional Science and Urban Economics 22 (1), 71–102.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28 (2), 139–161.

Arnott, R., Kraus, M., 1995. Financing capacity in the bottleneck model. Journal of Urban Economics 38 (3), 272–290.

Chen, H., Liu, Y., Nie, Y., 2014. Solving the step-tolled bottleneck model with general user heterogeneity. Transportation Research Part B, under review.

Cohen, Y., 1987. Commuter welfare under peak-period congestion tolls: Who gains and who loses? International Journal of Transport Economics= Rivista Internazionale de Economia dei Trasporti 14 (3), 238–266.

Doan, K., Ukkusuri, S., Han, L., 2011. On the existence of pricing strategies in the discrete time heterogeneous single bottleneck model. Transportation Research Part B 45 (9), 1483–1500.

Evans, A. W., 1992. Road congestion pricing: when is it a good policy? Journal of Transport Economics and Policy 26 (3), 213–244.

Hall, J. D., 2013. Pareto improvements from lexus lanes: the case for pricing a portion of the lanes on congested highways. In: Kuhmo NECTAR Conference.
Harker, P. T., Pang, J.-S., 1988. On the existence of optimal solutions to mathematical programs with equilibrium constraints. Operations Research Letters 7, 61–64.

Hendrickson, C., Kocur, G., 1981. Schedule delay and departure time decisions in a deterministic model. Transportation Science 15 (1), 62–77.

Higgins, T. J., 1994. Congestion pricing: Implementation considerations. Transportation Quarterly 48 (3), 287–298.

Laih, C.-H., 1994. Queueing at a bottleneck with single-and multi-step tolls. Transportation Research Part A: Policy and Practice 28 (3), 197–208.

Laih, C.-H., 2004. Effects of the optimal step toll scheme on equilibrium commuter behaviour. Applied Economics 36 (1), 59–81.

Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation science 38 (3), 293–314.

Lindsey, R., van den Berg, V. A., Verhoef, E. T., 2012. Step tolling with bottleneck queuing congestion. Journal of Urban Economics 72 (1), 46–59.

Liu, Y., Nie, Y., 2011. Morning commute problem considering route choice, user heterogeneity and alternative system optima. Transportation Research Part B 45 (4), 619–642.

Liu, Y., Nie, Y., Hall, J., 2015. A semi-analytical approach for solving the bottleneck model with general user heterogeneity. Transportation Research Part B 71, 56–70.

Luo, Z.-Q., Pang, J.-S., Ralph, D., 1996. Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge, U.K.

Newell, G. F., 1987. The morning commute for nonidentical travelers. Transportation Science 21 (2), 74–88.

Nie, Y., 2013. A new tradable credit scheme for the morning commute problem. Networks and Spatial Economics, Available at http://dx.doi.org/10.1007/s11067-013-9192-8.

Ramadurai, G., Ukkusuri, S., Z., Pang, J.-S., 2010. Linear complementarity formulation for single bottleneck model with heterogeneous commuters. Transportation Research Part B 44 (2), 193–214.

Small, K., 1982. The scheduling of consumer activities: work trips. The American Economic Review 72 (3), 467–479.

Small, K., 1983. The incidence of congestion tolls on urban highways. Journal of Urban Economics 13 (1), 90–111.

Small, K., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists’ preferences for travel time and reliability. Econometrica 73 (4), 1367–1382.
van den Berg, V. A., 2014. Coarse tolling with heterogeneous preferences. Transportation Research Part B 64, 1–23.

van den Berg, V. A., Verhoef, E. T., 2011a. Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B 45 (1), 60–78.

van den Berg, V. A., Verhoef, E. T., 2011b. Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics 95 (7-8), 983–992.

Vickrey, W., 1969. Congestion theory and transport investment. The American Economic Review 59 (2), 251–260.

Vickrey, W., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record 476 (476), 36–48.

Wardrop, J. G., 1952. Road paper. some theoretical aspects of road traffic research. In: ICE Proceedings: Engineering Divisions. Vol. 1. Thomas Telford, pp. 325–362.

Xiao, F., Qian, Z. S., Zhang, H. M., 2011. The morning commute problem with coarse toll and nonidentical commuters. Networks and Spatial Economics 11 (2), 343–369.

Xiao, F., Shen, W., Zhang, H. M., 2012. The morning commute under flat toll and tactical waiting. Transportation Research Part B 46 (10), 1346–1359.

Yang, H., Huang, H.-J., 1997. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. Transportation Research Part B 31 (6), 425–440.

Yang, H., Meng, Q., 1998. Departure time, route choice and congestion toll in a queuing network with elastic demand. Transportation Research Part B 32 (4), 247–260.

A Notations
### Table 7: Description of notations used in the paper

| Variable | Description | Unit | Range |
|----------|-------------|------|-------|
| $s$      | capacity of the bottleneck | vehicle/hour |       |
| $n$      | number of user classes | - |       |
| $t^*$    | desired arrival time | - | 0     |
| $N_i$    | demand for class $i$ | vehicle | $i = 1, 2, ..., n$ |
| $N$      | total demand | vehicle | $i = 1, 2, ..., n$ |
| $a_i$    | unit cost of travel time of class $i$ | $$/hour | $i = 1, 2, ..., n$ |
| $\beta_i$ | unit cost early arrival of class $i$ | $$/hour | $i = 1, 2, ..., n$ |
| $\gamma_i$ | unit cost of late arrival of class $i$ | $$/hour | $i = 1, 2, ..., n$ |
| $A$      | arrival windows | - | $E$ (early), $L$ (late) |
| $m^A$    | number of step tolls in arrival window $A$ | - | $A = \{E, L\}$ |
| $A_j$    | $j$th arrival window in arrival window $A$ | - | $j = 1, ..., m^A, A = \{E, L\}$ |
| $p^A_j$  | distance between the far end of the arrival window $A_j$ and $t^*$ | - | $j = 1, ..., m^A, A = \{E, L\}$ |
| $\pi^A_j$ | toll in the arrival window $A_j$ | $/hour | $j = 0, 1, ..., m^A, A = \{E, L\}$ |
| $\beta^A_j$ | vector of $\pi^A_j$ | - |       |
| $E(i)$   | class ID of the class whose $\frac{\beta^A_j}{a_i}$ value ranks at the $i$th place in all classes | - | $i = 1, 2, ..., n$ |
| $L(i)$   | class ID of the class whose $\frac{\beta^A_j}{a_i}$ value ranks at the $i$th place in all classes | - | $i = 1, 2, ..., n$ |
| $A(i)$   | the generalization of $E(i)$ and $L(i)$ | - | $i = 1, 2, ..., n, A = \{E, L\}$ |
| $n^A_{ij}$ | number of classes that present in arrival window $A_j$ | - | $j = 1, ..., m^A, A = \{E, L\}$ |
| $D^{A_i}(i)$ | ratio rank ID of the class whose time flexibility ranks at the $i$th place among all presenting classes in arrival window $A_j$ | - | $i = 1, 2, ..., n^A_{ij}, j = 1, ..., m^A, A = \{E, L\}$ |
| $N^A_{ij}$ | number of commuters of class $i$ in the arrival window $A_j$ | vehicle | $i = 1, 2, ..., n, j = 0, 1, ..., m^A, A = \{E, L\}$ |
| $N$      | vector of all $N^A_{ij}$ | vehicle |       |
| $C^A_i$  | commute cost excluding the capacity-related penalty of class $i$ in the arrival window $A_j$ | hour | $i = 1, 2, ..., n, j = 0, 1, ..., m^A, A = \{E, L\}$ |
| $C_i$    | vector of all $C^A_i$ | hour |       |
| $\mu_i$  | user equilibrium cost of class $i$ | hour | $i = 1, 2, ..., n$ |
| $\rho_i$ | class membership set realization (indicating whether class $i$ is present in $A_j$) | $/hour | $i = 1, 2, ..., n, j = 0, 1, ..., m^A, A = \{E, L\}$ |
| $\rho$   | vector of $\rho_i$ | - |       |
| $W$      | system cost of the toll design problem | $/hour |       |
| $W_\rho$ | system cost of the quadratic program corresponding to $\rho$ | $/hour |       |
| $\Omega$ | feasible set of the tolled equilibrium problem | $/hour |       |
| $\Gamma$ | feasible set of toll parameters applied to $\rho$ | $/hour |       |
| $\Xi(\rho)$ | additional feasibility requirements for toll parameters, defined by isocost curves corresponding to $\rho$ | $/hour |       |
| $\lambda^A_j$ | capacity-related penalty travel delay in the arrival window $A_j$ | hour | $j = 0, 1, ..., m^A, A = \{E, L\}$ |
| $\kappa$ | vector of multipliers associated with constraints in $\Xi(\rho)$ | - |       |