ITS UC DAVIS
INSTITUTE OF TRANSPORTATION STUDIES

Research Report – UCD-ITS-RR-10-42

The Morning Commute Problem with Coarse Toll and Nonidentical Commuters

May 2010

Feng Xiao
Zhen Qian
Michael Zhang

Institute of Transportation Studies ◦ University of California, Davis
One Shields Avenue ◦ Davis, California 95616
PHONE (530) 752-6548 ◦ FAX (530) 752-6572
www.its.ucdavis.edu

SUSTAINABLE TRANSPORTATION CENTER
of the Institute of Transportation Studies

Final Research Report R04-5

The Morning Commute Problem with Coarse Toll and Nonidentical Commuters

May 2010

Feng Xiao
Zhen Qian
Michael Zhang

Department of Civil and Environmental Engineering
University of California, Davis
One Shields Ave., Davis, CA 95616

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

The Morning Commute Problem with Coarse Toll and Non-identical Commuters

Feng Xiao, Zhen Qian and H. Michael Zhang*
Department of Civil & Environmental Engineering
One Shields Avenue, Davis, CA 95616

Abstract

This paper studies the morning commute problem under a flat peak-period toll within the context of heterogeneous commuters. All the possible queuing profiles resulting from different choices of toll level and charging time interval are examined. The optimal toll patterns were derived from minimizing the total travel cost of all commuters, excluding toll cost, and we proved that at the optimum there will be no queue or capacity waste at the bottleneck at both the starting and ending points of the charging time interval. Moreover, the optimal coarse toll scheme is pareto-improving. Different from the homogeneous case, which can be regarded as a special case of the heterogeneous case, price discrimination occurs when commuters have different values of time. We find that commuters in the middle pack of the value-of-time distribution are more easily to be hurt by higher toll charges and the optimal solution depends on the units in which the system cost is measured. Numerical examples are provided to demonstrate and compare the resulting equilibrium flow patterns under different objectives.

* Corresponding Author. Email: hmzhang@ucdavis.edu.

1. Introduction

The peak period coarse toll problem on a single bottleneck, where a flat toll is charged during part of the morning commuting peak period, has been studied in the literature in the context of homogeneous commuters. For the optimal toll charge and charging period, Arnott et al. (1990a) pointed out that the queue at the bottleneck at the starting and ending points of the charging period should be eliminated. Based on this important property, the optimal toll level, optimal starting and ending times of the tolling period, as well as the total system cost can all be easily calculated. The optimal toll level was found to be independent of the value-of-time attached to schedule-early and schedule-late delay. However, they did not provide details of how the queuing profile changes with respect to toll level and the choices of the starting and ending times of the tolling period, neither did they consider heterogeneous commuters. According to their numerical example, the one-step coarse toll is almost half efficient (not exactly half) as the fine (first-best) toll. Later they extended these results to a one-to-one bi-bottleneck parallel network (Arnott et al. 1990b). Bernstein and Elsanhouri (1994) corrected an error in that paper by announcing that if the demands are inter-dependent, the optimal toll level should depend on all the three value-of-time parameters attached to queuing, schedule-early and schedule-late delays; Laih (1994) reconsidered the single-bottleneck coarse toll problem by providing an easier way to calculate the optimal coarse toll without discussing the explicit evolution of the queue. Unfortunately, the methodology proposed in that paper is problematic. His analysis based on the assumption that the flat toll will not alter the trip price of each commuter, which we find is generally not true. In fact, from our analysis we can see that an arbitrarily high coarse toll may induce a period that no one departs from home or even a period that no one passes the bottleneck. As a result, the coarse toll does alter the trip cost of each commuter. Especially, when the coarse toll scheme is optimized, everyone will be benefited, even they have different values of time. Because of his wrong assumption, the conclusion that for a one-step toll scheme, "the optimal toll level is half of the maximum optimal time-varying toll and can at most eliminate half of the total queuing time" no longer holds.

The heterogeneity in commuters has also been addressed in the literature, yet the models vary due to different assumptions made. One way to classify these models is based on if the model is discrete (there are finite classes of commuters) or continuous (the number of the classes is infinite). Most previous studies assume finite multi-class users: Zijpp and Koolstra (2002) provided a generic algorithm that solves the departure time choice equilibrium given heterogeneous departure time preferences, arbitrary origin-bound and destination-bound rescheduling cost functions, and arbitrary queuing cost functions. Arnott et al. (1988) analytically solved the departure time choice of multiple groups of travelers who differ in one of the three aspects of heterogeneity. More recently, Ramadurai et al. (2008) developed a linear complementarity formulation for solving the single bottleneck problem in discrete time and user classes. Lindsey (2004) investigates the existence and uniqueness of departure-time user equilibrium in the bottleneck model with multi-user classes. Compared with the aim to solve the problem by numerical algorithms in most of the discrete models, Newell (1987) obtained certain analytical results from assuming a continuously distributed value of time (VOT). He graphically described the morning commute pattern for non-identical travelers. The queuing pattern was derived for a certain class of cost models and it was shown that a stable commuting pattern exists and is dictated by a certain fraction of travelers.

Another way to classify the models is by the definition of heterogeneity: Newell (1987) defined "non-identical" as different ratios of queuing time cost and schedule delay cost for each person. Since tolling was not considered, the departure time choice is only dependent on the distribution of this ratio and has nothing to do with the absolute VOT. Arnott et al. (1988) analyzed the departure time decisions of morning commuters who differ in three different ways: travel time and schedule delay costs, relative costs of schedule-early and late delay and work starting time. Huang (2000) dealt with pricing and modal split in a competitive mass transit/highway system with two groups of commuters that differ in their disutility from travel time, schedule-early delay and transit crowding. Instead of considering a coarse toll during the peak period on a bottleneck with fixed demand, more studies focus on a uniform toll throughout the whole rush hour. Since a uniform toll covering the whole time period will not influence the departure time choices of the commuters when the amount of commuters is fixed, people either assume an elastic demand (Braid 1989) or a parallel link or travel mode competing for the demand with the bottleneck (Tabuchi 1993, Braid 1996, Danielis and Marcucci 2002).

In this paper we first analytically solve the optimal coarse toll problem for a single bottleneck with identical commuters. We show how the profile evolves with respect to different choices of toll level and tolling period. Then the results are extended to heterogeneous commuters. In the latter case, the changes in each individual commuter's utility after the coarse toll was applied are also investigated. Moreover, we show how the optimal solution changes when the toll operator and the commuters make tradeoffs between cost and time. Finally, numerical examples are provided to compare the different cases.

2. **No-Toll Equilibrium with Identical Commuters**

In this section we first provide the no-toll equilibrium (NTE) at a single bottleneck with identical commuters during the morning rush hour. Although these results are well known, we include them here for completeness because it introduces the needed concepts and problem setting for our coarse toll problem. In all subsequent discussions, we assume all the morning commuters have the same work starting time, t^*. The cumulative number of arrivals at the bottleneck by time t is $A(t)$ and the waiting time in queue for any commuter who passes the bottleneck at time t is $w(t)$. The total number of people commuting during the morning peak is N and the passing rate (bottleneck capacity) at the bottleneck is s. Without loss of generality, we assume that there is no travel time cost other than the queuing time cost at the bottleneck. Thus a commuter arrives at the bottleneck as soon as s/he departs from home and arrives at work immediately after leaving the bottleneck. Firstly, the arrival rate at the bottleneck exceeds the passing rate and a queue builds up from time t_q; After the arrival rate goes down below the bottleneck capacity, the queue dissipates linearly till it disappears at time $t_{q'}$. From the definition, equilibrium is obtained when no individual has an incentive to change his/her departure time. The cumulative departure curve (which in the rest of this paper we briefly call "the profile"), is defined as the cumulative departures from home as a function of time. The profile at NTE is drawn in Figure 1.

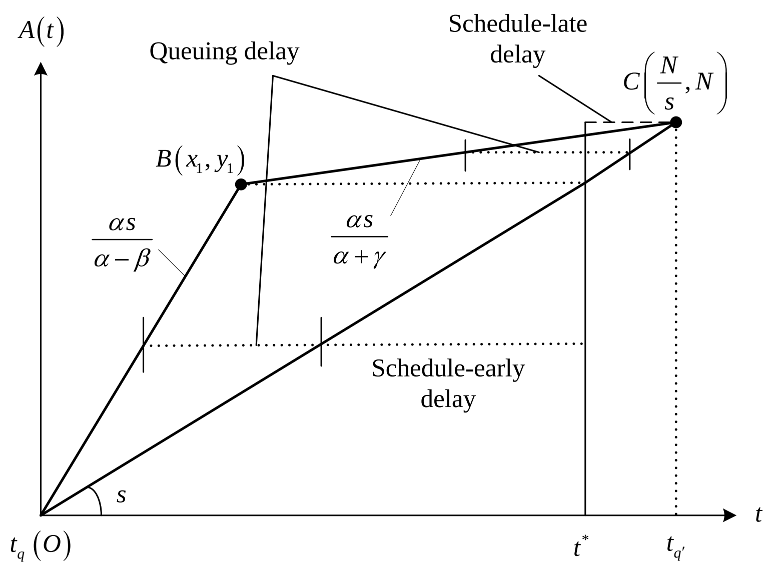

Figure 1. Profile of NTE

We assume α is the unit monetary value attached to queuing delay time, β is the unit monetary value of schedule-early delay and γ is the unit monetary value of schedule-late delay. In accordance with empirical evidences and for the existence and uniqueness of the equilibrium, the relation $\gamma > \alpha > \beta$ must hold. Here to locate the profile accurately, we provide each endpoint of the departure curve a coordinate. The queue starting time t_q is set to be the origin, i.e. $t_q = 0$. As described in Vickrey's paper (1969), at NTE the slope of $\overline{OB}$ is equal to $\alpha s/(\alpha - \beta)$ and the slope of $\overline{BC}$ is equal to $\alpha s/(\alpha + \gamma)$. Thus we can analytically solve the coordinates of point B

$$x_1 = \frac{\gamma(\alpha - \beta)}{\alpha(\gamma + \beta)} \frac{N}{s}. \tag{1}$$

$$y_1 = \frac{N\gamma}{\gamma + \beta}. \tag{2}$$

The work starting time t^* is equal to the x-axis coordinate of the point on line $\overline{OC}$ (the queue discharge curve) which has y_1 as the y-axis coordinate. Thus it's not hard to obtain that

$$t^* = \frac{y_1}{s} = \frac{\gamma}{\gamma + \beta} \frac{N}{s} \tag{3}$$

The total travel cost of an individual who departs at time t comprises two parts, the queuing delay and the schedule delay. A general expression of an individual's travel cost when s/he passes the bottleneck at time t is as below

$$C(t) = \alpha w(t) + \max\left(\beta\left(t^* - t\right), \gamma\left(t - t^*\right)\right) \tag{4}$$

Since the commuters are identical, at NTE every commuter will incur the same travel cost, which is equal to the schedule-early delay t^* of the first individual who experiences no queuing delay. Thus from eqn.(3), each commuter's travel cost at NTE can be calculated as

$$\bar{C} = \beta t^* = \frac{\beta\gamma}{\gamma + \beta} \frac{N}{s} \tag{5}$$

And the total system cost at NTE is

$$\overline{TC} = \frac{\beta\gamma}{\gamma + \beta} \frac{N^2}{s} \tag{6}$$

It's well known that an optimal dynamic toll can be found to totally eliminate the queuing delay in the system. Under this optimal dynamic toll, each commuter will pay the amount of toll equivalent to her queuing delay cost and thus the departure is evenly distributed throughout the time interval $\left(t_q, t_{q'}\right)$, at the rate of *s*, the bottleneck capacity.

3. The Optimal Coarse Toll with Identical Commuters

In real life it is impractical to implement the optimal dynamic toll because of the difficulty in collecting all the information needed for deriving the toll, and the confusion it could cause the public with its too frequent toll rate changes. On the other hand, an approximate form to such a toll, which divides the peak periods into several tolling intervals and a flat toll is charged in each tolling interval, can be implemented with ease. In this paper, we deal with a special case of such coarse tolls, one with only a single toll charge and tolling period. The solution of this coarse toll problem provides insights to more refined coarse tolls with multiple toll levels and tolling periods.

In this paper, a coarse toll is defined to be a flat charge ρ to the commuters passing the bottleneck within a time interval $[t^+, t^-]$. Since the coarse toll is defined as a rush hour tolling scheme, it's reasonable to assume that the toll is applied at $t^+ \in [t_q, t^*]$ and lifted at $t^- \in [t^*, t_{q'}]$. Thus every selection of the three parameters $\left(\rho, t^+, t^-\right)$ represents a tolling pattern, which also determines a unique departure profile.

Undoubtedly, compared with the no-toll case queue lengths at the bottleneck within time interval $[t^+, t^-]$ will be reduced after a flat toll is charged. Some of the commuters who travel inside $[t^+, t^-]$ will be forced to travel outside the time interval, either before t^+ or after t^-, yet the profile within $[t^+, t^-]$ is still similar with the original profile at NTE, because the commuters are identical and a uniform toll has no effect on the departure time choices. For those commuters who travel outside $[t^+, t^-]$, they also follow the same departure rate as NTE. However, it's not clear what the profile will be like around the starting and ending points of the tolling period. In fact, at the time when tolling starts, there could be two different situations:

i) When the toll is relatively low, there could be a while no one departs before t^+ but commuters pass the bottleneck all the time (See Figure 2). From the definition of equilibrium, the last person who passes the bottleneck without paying the toll should have the same travel cost as the first person who pays the toll. Thus the commuter departing at point B experiences a longer queuing time of $\overline{BD}$ than the commuter departing at point C who experiences a queuing time $\overline{CD}$ and an additional toll ρ . If we define $m = \overline{BC} - \overline{BD}$, in this case $m < 0$. And because the commuters departing at B and C experience the same schedule-early delay, the toll should be equal to the difference between the queuing delay cost, i.e.

$$\rho = \alpha \overline{BC} \tag{7}$$

ii) When the toll is set too high, there could be a while that no one travels through the bottleneck, which we refer to as "capacity waste" (See Figure 3). Now the last person who passes the bottleneck without paying the toll experiences a longer schedule-early delay than the first person who pays the toll. In this case $m \geq 0$. The first person who pays the toll will experience no queuing delay. The toll should be equal to the sum of the differences between the queuing and schedule-early delay costs, i.e.

$$\rho = \alpha \overline{BD} + \beta \overline{CD} \tag{8}$$

For both cases i) and ii), it's not hard to find that

$$\overline{BC} = y_1 / s + m - x_1 . \tag{9}$$

Similarly, at the time when the toll is lifted, the last person who pays the toll should have the same travel cost as the first person who passes the bottleneck after the toll. This could happen only when the rest of commuters depart together immediately after the toll is lifted. Since the position in the queue for these commuters is random, everyone has the same expectation of queuing and schedule-late delays. The expected total travel cost is ρ units higher than the total cost of the commuters traveling inside the time interval. From Figure 2, we observe that the expected travel cost for passing the bottleneck after t^- should be equal to that of the middle commuter who arrived simultaneously at the bottleneck after the toll is lifted. It can be easily calculated that the size of the commuters passing the bottleneck after the time interval is $2s\rho / (\alpha + \gamma)$. Different from the situation at t^+, at t^- there could be only one possibility: the queue is greater than or equal to 0. The situation of "capacity waste" will never happen here, since if it happens, commuters will be able to simultaneously reduce their queuing and schedule-late delays by simply departing earlier.

From the above observations, we need to consider two possible profiles in order to find the optimal coarse toll pattern, which are discussed separately in the following two sub sections.

3.1 The profile without capacity waste

The profile is drawn in Figure 2. In this profile, at the starting time of the coarse toll t^+, queue exists at the bottleneck.

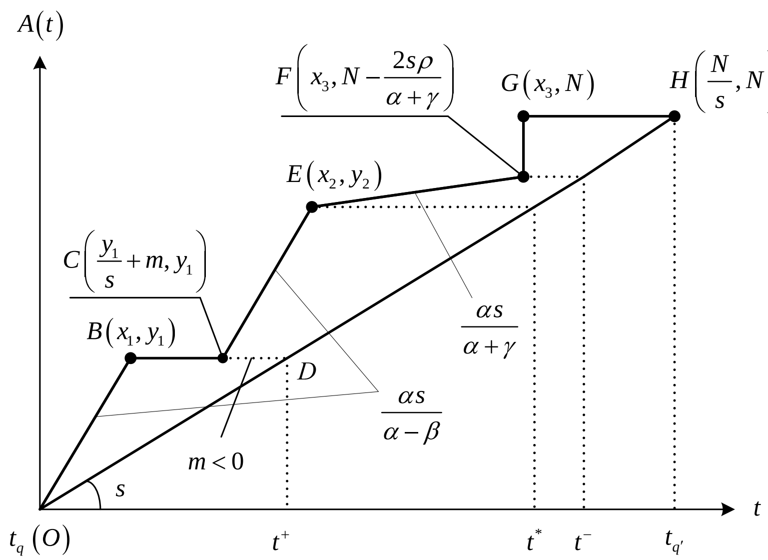

Figure 2. Profile without capacity waste (homogeneous case)

When the toll level and tolling interval $\left(\rho, t^+, t^-\right)$ are such that this profile holds, one can find the optimal toll pattern by minimizing the total system cost, subject to the constraint that this profile holds. Obviously the degree of freedom of this problem is three. The resulting nonlinear constrained optimization problem is stated as follows:

$$\min_{\rho, y_1, y_2 \geq 0} TC = \frac{\beta N}{s} y_2 - \rho \cdot \left(N - \frac{2s\rho}{\alpha+\gamma} - y_1 \right) \tag{10}$$

s.t.

$$\frac{y_1}{x_1} = \frac{\alpha s}{\alpha - \beta} \tag{11}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{\alpha s}{\alpha - \beta} \tag{12}$$

$$\frac{N-\frac{2s\rho}{\alpha+\gamma}-y_2}{x_3-x_2}=\frac{\alpha s}{\alpha+\gamma} \tag{13}$$

$$\frac{y_1}{s}+m-x_1=\frac{\rho}{\alpha} \tag{14}$$

$$m\le 0 \tag{15}$$

$$\frac{N-\frac{2s\rho}{\alpha+\gamma}}{x_3}\ge s \tag{16}$$

The objective function (10) calculates the total system cost, which is equal to the total delay cost (including queuing delay and schedule delay) minus the toll revenue. Constraints (11), (12) and (13) guarantee the departure rates satisfying the equilibrium. Constraints (14) is from eqns.(7) and (9). Constraints (16) ensures that the queue at time t^- is greater than or equal to 0. In other words, constraints (11)-(16) ensures that the toll pattern produces a profile consistent with the one shown in Figure 2.

From some straightforward algebraic manipulations, the above optimization problem can be simplified as

$$\min_{\rho,x_3,m} TC=\frac{N^2\beta(\alpha+\gamma)}{s(\beta+\gamma)}-\frac{\alpha\beta N}{\beta+\gamma}x_3+\left(\frac{2s}{\alpha+\gamma}+\frac{s}{\beta}\right)\rho^2-\frac{2\beta+\gamma}{\beta+\gamma}N\rho-\frac{\alpha\rho s}{\beta}m \tag{17}$$

s.t.

$$x_3\le\frac{N}{s}-\frac{2\rho}{\alpha+\gamma} \tag{18}$$

$$m\le 0 \tag{19}$$

We call this simplified version **Problem I**. Obviously, Problem I still has three variables (x_3,m,ρ) to be solved for.

3.2 The profile with capacity waste

When the toll charge is high enough, there will be an early time interval in the tolling period that no one departs from home until a commuter’s schedule-early delay decreases sufficiently to compensate for the toll charge, which results in the profile shown in Figure 3. In this profile, there is capacity waste during the tolling period $[t^+, t^-]$.

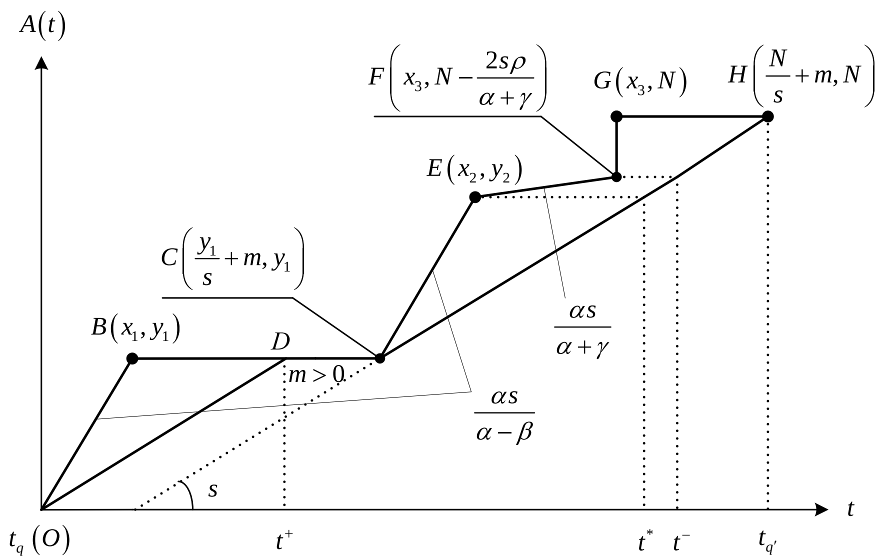

Figure 3. Profile with capacity waste (homogeneous case)

Similarly, one can find the optimal toll pattern for this profile by solving the following nonlinear optimization problem

$$\min_{\rho, y_1, y_2, m\geq 0} TC = \beta N\left(\frac{y_2}{s}+m\right) - \rho \cdot \left(N - \frac{2s\rho}{\alpha+\gamma} - y_1\right) \tag{20}$$

s.t.

$$\frac{y_1}{x_1} = \frac{\alpha s}{\alpha-\beta} \tag{21}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{\alpha s}{\alpha-\beta} \tag{22}$$

$$\frac{N - \frac{2s\rho}{\alpha+\gamma} - y_2}{x_3 - x_2} = \frac{\alpha s}{\alpha+\gamma} \tag{23}$$

$$\frac{y_1}{s} + \frac{\beta}{\alpha} m - x_1 = \frac{\rho}{\alpha} \tag{24}$$

$$m \geq 0 \tag{25}$$

$$\frac{N - \frac{2s\rho}{\alpha+\gamma}}{x_3 - m} \geq s \tag{26}$$

From eqns.(21) and (24) we can solve for y_1, which is expressed by m and ρ

$$y_1 = s\left(\frac{\rho}{\beta} - m\right) \tag{27}$$

Since y_1 has to be nonnegative, together with constraint (25) we have

$$0 \le m \le \frac{\rho}{\beta} \tag{28}$$

We can eliminate y_1 in (20)-(26) and obtain the following simplified problem, which we refer as **Problem II.**

$$\begin{aligned}\min_{\rho, x_3, m} TC = & \frac{N^2\beta(\alpha+\gamma)}{s(\beta+\gamma)} - \frac{\alpha\beta N}{\beta+\gamma} x_3 + \left(\frac{2s}{\alpha+\gamma} + \frac{s}{\beta}\right)\rho^2 - \frac{2\beta+\gamma}{\beta+\gamma} N\rho \\ & + \beta s\left(\left(\frac{N}{s} - \frac{\rho}{\beta}\right) + \frac{N}{s}\frac{\alpha-\beta}{\beta+\gamma}\right)m\end{aligned} \tag{29}$$

s.t.

$$x_3 \le m + \frac{N}{s} - \frac{2\rho}{\alpha+\gamma} \tag{30}$$

$$0 \le m \le \frac{\rho}{\beta} \tag{31}$$

3.3 The unified problem

In Problems I and II, it is easy to show from the first-order necessary conditions that both constraints (18) and (30) are binding at the optimum. This means that at the optimum, there will be no queue at the ending points of the toll charge period. After substituting the binding constraints (18) and (30) into the objective functions, we are able to combine the two profiles together to form a unified optimization problem below. The benefit is that constraints (15) and (25) related to m will be removed. The optimal coarse toll scheme for a bottleneck with identical commuters can thus be solved by the following nonlinear optimization problem

$$\min_{m,\rho} TC = \left(\frac{2s}{\alpha+\gamma} + \frac{s}{\beta}\right)\rho^2 - \frac{N\gamma(\alpha+2\beta+\gamma)}{(\beta+\gamma)(\alpha+\gamma)}\rho + \frac{N^2\beta\gamma}{s(\beta+\gamma)} + \lambda|m| \tag{32}$$

s.t.

$$\lambda = \begin{cases} \dfrac{\alpha}{\beta}\rho s, & m \le 0 \\ s\left(\dfrac{N}{s}\dfrac{\beta\gamma}{\beta+\gamma} - \rho\right), & 0 < m \le \dfrac{\rho}{\beta} \end{cases} \tag{33}$$

By solving this problem (See Appendix A) we obtain $m = 0$, which means the first commuter who pays the toll will experience no queuing delay. Thus we conclude that

when the system cost is minimized with optimal coarse toll level ρ_s and optimal tolling time interval $\left(t_s^+, t_s^-\right)$, there will be no queue or capacity waste at time t_s^+ and t_s^-.

The number of commuters who pass through the bottleneck outside $[t^+, t^-]$ is

$$V = \frac{2s\rho}{\alpha+\gamma} + \frac{s\rho}{\beta} = \frac{\gamma\left(\gamma+\alpha+2\beta\right)}{2\left(\alpha+\gamma\right)\left(\beta+\gamma\right)}N > \frac{N}{2} \tag{34}$$

which means more commuters choose to pass the bottleneck outside $\left[t_s^+, t_s^-\right]$ to avoid the toll. In other words, to minimize the system cost, the tolling time window cannot be longer than half of the entire morning commute period. Solving for the optimal toll gives

$$\rho^s = \frac{\beta\gamma}{2\left(\beta+\gamma\right)}\left(\frac{N}{s}\right) \tag{35}$$

3.4 Numerical solution and sensitivity analysis

We use a numerical example to show the change in system cost with respect to the combination choices of ρ and m. To easily verify the results, we adopt the same values of the commuters attached to the queue and schedule delays as the example provided in Arnott et al's paper (1990a): $\alpha = 6.4$, $\beta = 3.9$ and $\gamma = 15.21$. To compare the total costs under different pairs of $\left(\rho, m\right)$, we assume the demand $N = 100$ and the passing rate $s = 50$.

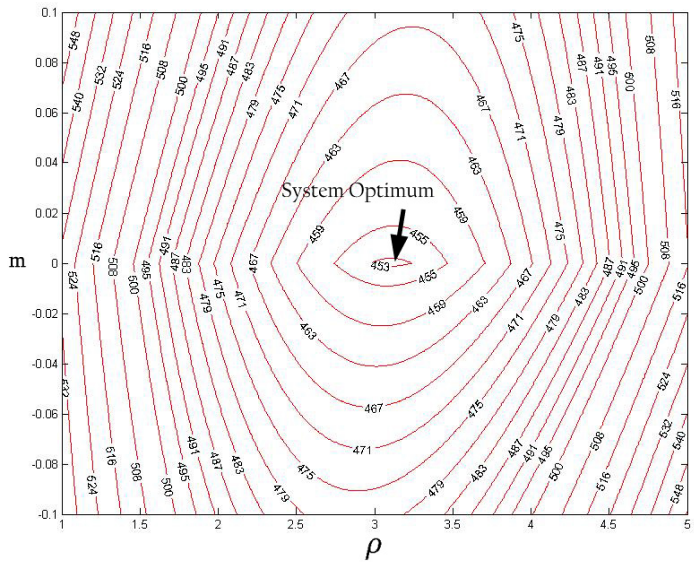

Figure 4. System cost w.r.t. ρ and m

From Figure 4 we notice that no matter how much the coarse toll is, the system cost is always minimized at $m = 0$. The lesson learned from this result is that once a toll level is decided, the tolling time interval has to be carefully chosen: when the time interval is too long, the capacity of the bottleneck cannot be fully utilized, while if the time interval is too short, an additional amount of deadweight loss of queuing delay is induced. The optimum is obtained at a toll level of \$3.1, which is consistent with the result in Arnott et al's paper (1990a). The total travel cost saved by the optimal coarse toll is 27.08%.

4. The Optimal Coarse Toll with Non-Identical Commuters

In real life, the income level of an individual determines how much the individual's value of time, α, could be, yet the relative value α/β depends mainly on how flexible one's work schedule is. There could be no certain relationship between income level and the flexibility of the job: sometimes highly paid white-collar workers have more flexible work hours compared with blue-collar workers; yet sometimes due to the job's specific nature, high income commuters may still have rigid work schedules, while low income commuters could have flexible work schedules. Nevertheless, it's still reasonable to believe that people who have a higher valuation of schedule delay will also have a higher valuation of the time spent waiting at the bottleneck. Thus to reasonably simplify the problem, we assume that everyone has the same work flexibility, but a different VOT, i.e. $\alpha/\beta = const, > 1$, and α follows a distribution

$$F(\omega) = \Pr\{\alpha \le \omega\} \tag{36}$$

The widely used assumption that β/γ is constant throughout the whole population is also adopted here. Then the individual cost of the $v\,th$ person is

$$C(v,t) = \alpha(v)\Big(w(t) + \max\big(\eta_1(t^* - t), \eta_2(t - t^*)\big)\Big) \tag{37}$$

where $\beta = \alpha\eta_1, \gamma = \alpha\eta_2$. And to guarantee the existence and uniqueness of the equilibrium, we assume $0 < \eta_1 < 1 < \eta_2$. For convenience, we arrange the commuters in increasing order of α. $\alpha(v)$ gives the $v\,th$ person's value of queuing delay. From this definition, $\alpha(v)$ is monotone and increasing.

After a coarse toll is imposed, only those with high value of time will travel within $[t^+, t^-]$. If we assume that the $V\,th$ commuter is the commuter with the lowest VOT among those traveling within $[t^+, t^-]$, then this commuter will have the same VOT $\alpha(V)$ with the highest VOT among those people who travels outside, i.e. there's no difference to the $V\,th$ commuter whether he/she chooses to travel inside or outside. The proof is straightforward: given a toll level, ρ, the difference of delay costs between traveling inside and outside is $\rho/f(\alpha(V))$, where $f(\alpha(v))$ is an increasing function with respect to v. For the commuter with VOT $\alpha(V')$, where $V' < V$, s/he will not choose to travel inside, because the equivalent delay cost caused by toll is $\rho/f(\alpha(V')) > \rho/f(\alpha(V))$, reversely.

For convenience, we define

$$A_1 = \int_0^V \alpha(v)dv \tag{38}$$

$$A_2 = \int_V^N \alpha(v)dv \tag{39}$$

$$K = A_1 + A_2 = \int_0^N \alpha(v)dv \tag{40}$$

Similar to the discussion in the homogeneous case, there are two possible profiles. But compared with the homogeneous case, we have one more parameter, V, and one more constraint that the integration A_2 should be equal to the number of commuters traveling within the time interval $\left[t^{+}, t^{-}\right]$. Thus the problem for heterogeneous case still has the same degree of freedom as in the homogeneous case.

4.1 The profile without capacity waste

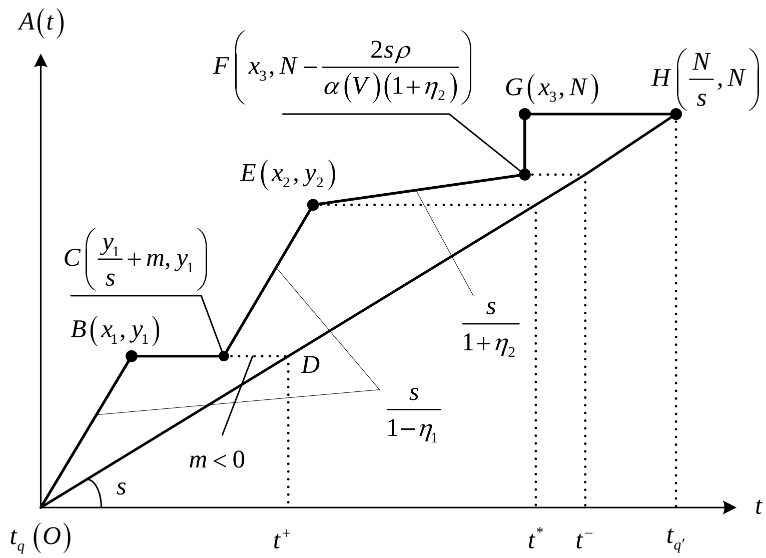

Figure 5. Profile without capacity waste (heterogeneous case)

The nonlinear optimization problem with multiple constraints will be:

$$\min_{V, x_2, y_2} TC = \frac{\eta_1 y_2}{s} A_1 + \left(\frac{y_2}{s} - x_2\right) A_2 = \frac{\eta_1 A_1 + A_2}{s} y_2 - A_2 x_2 \tag{41}$$

s.t.

$$\frac{y_1}{x_1} = \frac{s}{1-\eta_1} \tag{42}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{s}{1-\eta_1} \tag{43}$$

$$\frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)} - y_2}{x_3 - x_2} = \frac{s}{1+\eta_2} \tag{44}$$

$$\frac{y_1}{s} + m - x_1 = \frac{\rho}{\alpha(V)} \tag{45}$$

$$m \le 0 \tag{46}$$

$$\frac{N-\dfrac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3}\geq s \tag{47}$$

$$V=\frac{2s\rho}{\alpha(V)(1+\eta_2)}+y_1 \tag{48}$$

The total system cost is calculated by summing up the total delay costs of the commuters passing the bottleneck within and outside the tolling interval. Constraint (48) ensures that the total amount of commuters whose VOT are greater than $\alpha(V)$ should be equal to the number of commuters traveling within the time interval $\left[t^+,t^-\right]$.

The problem can be further simplified by eliminating some of the dependent variables (See Appendix B)

$$\begin{aligned}\min_{V,x_3,\rho} TC=&\frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)}+\left(\frac{N}{s}-\frac{2\rho}{\alpha(V)(1+\eta_2)}\right)-\frac{1}{1+\eta_2}x_3\right)\\&-\left(\frac{K}{1-\eta_1}-A_1\right)\frac{\rho}{\alpha(V)}\end{aligned} \tag{49}$$

s.t.

$$x_3\leq\frac{N}{s}-\frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{50}$$

$$m=\frac{\rho}{\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2}-\frac{\eta_1 V}{s} \tag{51}$$

$$m\leq 0 \tag{52}$$

Again from the first-order optimality conditions one can show that constraint (50) should be binding. Therefore the problem can be further simplified by removing this constraint.

$$\min_{V,x_2,y_2} TC=\frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s}+\frac{\rho}{\alpha(V)}\left(A_1-\frac{(1+2\eta_1+\eta_2)\eta_2 K}{(\eta_1+\eta_2)(1+\eta_2)}\right) \tag{53}$$

s.t.

$$m=\frac{\rho}{\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2}-\frac{\eta_1 V}{s} \tag{54}$$

$$m\leq 0 \tag{55}$$

From eqn.(54) we have

$$\frac{\rho}{\alpha(V)}=\left(m+\frac{\eta_1 V}{s}\right)\frac{1+\eta_2}{1+2\eta_1+\eta_2} \tag{56}$$

Substituting (56) into (53) the problem becomes

$$\min_{V,x_2,y_2} TC = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} + \left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)m + \eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{57}$$

s.t.

$$m \le 0 \tag{58}$$

We call this **Problem III**.

4.2 The profile with capacity waste

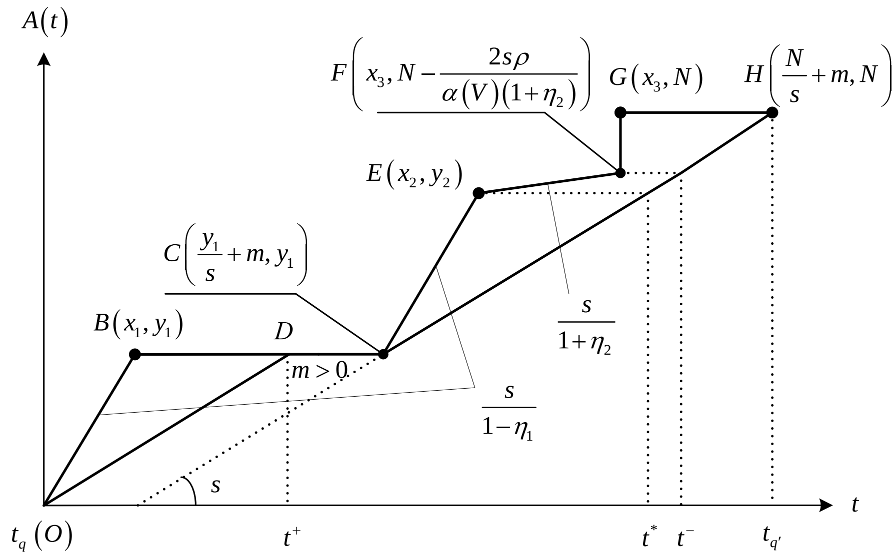

Figure 6. Profile with capacity waste (heterogeneous case)

From Figure 6, the nonlinear optimization problem with multiple constraints will be:

$$\min_{V,x_2,y_2} TC = \eta_1\left(\frac{y_2}{s}+m\right)A_1 + \left(\frac{y_2}{s}+m-x_2\right)A_2 = \frac{\eta_1 A_1 + A_2}{s}y_2 - A_2 x_2 + (\eta_1 A_1 + A_2)m \tag{59}$$

s.t.

$$\frac{y_1}{x_1} = \frac{s}{1-\eta_1} \tag{60}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{s}{1-\eta_1} \tag{61}$$

$$\frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)} - y_2}{x_3 - x_2} = \frac{s}{1+\eta_2} \tag{62}$$

$$\frac{y_1}{s} + \eta_1 m - x_1 = \frac{\rho}{\alpha(V)} \tag{63}$$

$$m \geq 0 \tag{64}$$

$$\frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3 - m} \geq s \tag{65}$$

$$V = \frac{2s\rho}{\alpha(V)(1+\eta_2)} + y_1 \tag{66}$$

The problem can be further simplified (See Appendix C for the proof)

$$\begin{aligned} \min_{V,x_2,y_2} TC = & \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3 + m\right) \\ & -\left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \end{aligned} \tag{67}$$

s.t.

$$x_3 \leq \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{s} + m \tag{68}$$

$$m = \frac{\rho}{\eta_1 \alpha(V)} \frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{69}$$

$$0 \leq m \leq \frac{\rho}{\eta_1 \alpha(V)} \tag{70}$$

Again, from first-order optimality conditions one finds that constraint (68) is binding. By replacement of variables, the optimal coarse toll problem with capacity waste can be further simplified

$$\min_{V,\rho} TC = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} + \frac{\rho}{\alpha(V)}\left(A_1 + \frac{(1+2\eta_1+\eta_2)K}{(\eta_1+\eta_2)}\right) - \frac{\eta_1(1+2\eta_2)K}{\eta_1+\eta_2}\frac{V}{s} \tag{71}$$

s.t.

$$m=\frac{\rho}{\eta_1\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2}-\frac{V}{s} \tag{72}$$

$$0\le m\le\frac{\rho}{\eta_1\alpha(V)} \tag{73}$$

From eqn.(72) we have

$$\frac{\rho}{\alpha(V)}=\left(m+\frac{V}{s}\right)\frac{\eta_1(1+\eta_2)}{1+2\eta_1+\eta_2} \tag{74}$$

Substituting (74) into (71), the problem becomes

$$\begin{aligned}\min_{V,\rho} TC=&\frac{\eta_1\eta_2K}{\eta_1+\eta_2}\frac{N}{s}+\left(\frac{\eta_1(1+\eta_2)}{1+2\eta_1+\eta_2}A_1+\frac{\eta_1(1+\eta_2)K}{\eta_1+\eta_2}\right)m\\&+\eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1-\frac{\eta_2K}{\eta_1+\eta_2}\right)\frac{V}{s}\end{aligned} \tag{75}$$

s.t.

$$0\le m\le\frac{\rho}{\eta_1\alpha(V)} \tag{76}$$

We call this **Problem IV**.

4.3 The unified problem

We further combine Problems III and IV together to form a unified problem, as we have done for the homogeneous case:

$$\min_{V} TC=\frac{\eta_1\eta_2K}{\eta_1+\eta_2}\frac{N}{s}+\lambda|m|+\eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1-\frac{\eta_2K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{77}$$

s.t.

$$\lambda=\begin{cases}\dfrac{\eta_2K}{\eta_1+\eta_2}-\dfrac{1+\eta_2}{1+2\eta_1+\eta_2}A_1, & m\le 0\\ \dfrac{\eta_1(1+\eta_2)}{1+2\eta_1+\eta_2}A_1+\dfrac{\eta_1(1+\eta_2)K}{\eta_1+\eta_2}, & 0<m\le\dfrac{\rho}{\eta_1\alpha(V)}\end{cases} \tag{78}$$

Because $0<\eta_1<1<\eta_2$, the term

$$\frac{(1+2\eta_1+\eta_2)\eta_2}{(\eta_1+\eta_2)(1+\eta_2)}=\frac{\eta_2+2\eta_1\eta_2+\eta_2^2}{\eta_1+\eta_2+\eta_1\eta_2+\eta_2^2}=1+\frac{\eta_1(\eta_2-1)}{\eta_1+\eta_2+\eta_1\eta_2+\eta_2^2}>1 \tag{79}$$

And because $A_1 \le K$, we have

$$\frac{\eta_2 K}{\eta_1+\eta_2}-\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1=\frac{1+\eta_2}{1+2\eta_1+\eta_2}\left(\frac{(1+2\eta_1+\eta_2)\eta_2 K}{(\eta_1+\eta_2)(1+\eta_2)}-A_1\right)>0 \tag{80}$$

Since the coefficient of m is always greater than zero, to reduce the system cost m has to be 0. The two profiles then reduce to the same profile without capacity waste or queue at the starting time of the toll and the minimized total system cost is the same.

From all the above discussions, we are able to conclude that for **heterogeneous commuters, where heterogeneity is defined by assumption (37)**, **there will be no queue or capacity waste at times t_s^+ and t_s^-, when the total system cost is minimized with optimal coarse toll level ρ_s and optimal tolling window (t_s^+, t_s^-), the same conclusion reached for the homogeneous case.**

We can calculate the optimal coarse toll level when the system cost is measured in money

$$\rho_m^s=\frac{(1+\eta_2)\eta_1}{1+2\eta_1+\eta_2}\frac{V}{s}\alpha(V) \tag{81}$$

The minimum total system cost is

$$TC_{\min}=\frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s}-V^2\alpha(V)\frac{\eta_1(1+\eta_2)}{(1+2\eta_1+\eta_2)s} \tag{82}$$

Substituting (81) into (53) we have

$$TC(V)=\frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s}+\eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1-\frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{83}$$

It can be easily proven that $\frac{\partial TC(V)}{\partial V}|_{V=0}<0$, $\frac{\partial TC(V)}{\partial V}|_{V=N}>0$ and $\frac{\partial^2 TC(V)}{\partial V^2}>0$, which ensures optimal V falls in $[0,N]$. By taking $\frac{\partial TC(V)}{\partial V}=0$ we obtain the equation with only V as variable:

$$\int_0^V \alpha(v)dv+V\alpha(V)=\frac{\eta_2(1+2\eta_1+\eta_2)}{(\eta_1+\eta_2)(1+\eta_2)}\int_0^N \alpha(v)dv \tag{84}$$

Eqn.(84) provides a method to calculate the amount of commuters V who are traveling outside the tolling period after toll is imposed, as long as we know the VOT distribution of the commuter population.

Generally we cannot get the analytical expression of V except for simple forms of VOT distribution. For instance, if VOT follows a uniform distribution, i.e. $\alpha(v) = b + av,\ a > 0, b > 0$, we can obtain

$$V = \frac{\sqrt{4b^2 + 6a\frac{\eta_2(1+2\eta_1+\eta_2)}{(\eta_1+\eta_2)(1+\eta_2)}\left(bN + \frac{1}{2}aN^2\right)} - 2b}{3a} \tag{85}$$

We know that the total system cost under user equilibrium is

$$\overline{TC} = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} \tag{86}$$

By comparing eqns.(82) and (86), we find the coarse toll scheme reduces the system cost. But since everyone has a different value of time, naturally we will ask the question: does the coarse toll scheme reduce everyone's travel cost in the monetary unit? We will answer this question in Section 5.

It is known that in a static transportation network, the optimal flow pattern derived from minimizing the total system cost measured in time could be different from that derived by minimizing the total system cost measured in money. We note that the same phenomenon can be found here. Under our assumption of heterogeneity, if we minimize the total system cost in time unit using the coarse toll, the optimal solution will be identical with what we get from the homogeneous case. The optimal toll for minimizing the total generalized travel time can be obtained from eqns.(35) and (34):

$$\rho_t^s = \frac{\eta_1\eta_2 N}{2s(\eta_1+\eta_2)}\alpha\left(\frac{\eta_2(\eta_2+1+2\eta_1)}{2(1+\eta_2)(\eta_1+\eta_2)}N\right) \tag{87}$$

which is generally different from the optimal toll level ρ_m^s corresponding to minimal total cost measured in monetary unit.

Another important observation can be made by assuming that the parameter values attached to queuing and schedule delays in the homogeneous case are the expectations of the VOT distributions in the heterogeneous case, represented by $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$. From eqn.(35) the optimal toll for homogeneous case becomes

$$\rho^s = \frac{\eta_1 \eta_2}{2(\eta_1 + \eta_2)} \left(\frac{N}{s} \right) \bar{\alpha} \tag{88}$$

Noting that $\int_0^V \alpha(v) dv \leq V \alpha(V)$ ("=" obtained when $\alpha(V)$ is a constant), from eqn.(84) we have

$$V \alpha(V) \geq \frac{\eta_2 (1 + 2\eta_1 + \eta_2)}{2(\eta_1 + \eta_2)(1 + \eta_2)} \int_0^N \alpha(v) dv \tag{89}$$

Substituing eqns.(81) and (88) into (89), we have

$$\frac{\rho_m^s}{\rho^s} \geq \frac{\int_0^N \alpha(v) dv}{N} \cdot \bar{\alpha} = 1 \tag{90}$$

Eqn.(90) implies that **using the mean of the VOTs instead of explicitely considering the heterogeneity will lead to an under-estimation of the optimal toll level**.

5. The Pareto-Improving Property of the Coarse Toll

The total system cost including toll revenue can be reduced by the coarse toll, but the toll scheme may still increase the travel costs of some individuals. In this section we examine the performance the coarse toll from the perspective of individual commuters.

For the homogeneous case, since all the commuters are identical, to see if the coarse toll is pareto-improving, we only need to examine whether the total delay cost under the optimal coarse toll, TC_{delay}, is reduced. We show without proof that

$$TC_{delay} = \left(\frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{s} \right) \left(1 - \frac{(\gamma - \alpha)\beta}{2(\beta + \gamma)(\gamma + \alpha)} \right) < \overline{TC} \tag{91}$$

Thus for the homogeneous case, the optimal coarse toll is pareto-improving.

For heterogeneous case, before tolling the $v\,th$ commuter's travel cost is

$$C(v)=\frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s}\alpha(v) \tag{92}$$

After the optimal coarse toll is imposed, the commuters traveling outside $[t^+,t^-]$ experience the same reduction of generalized travel time. Thus we only need to observe the cost of the first commuter who experienced only schedule early delay y_2/s. By solving constraints (42), (43), (44), (45) and (47) with $m=0$ (See Appendix C for details), we have the generalized travel time for the commuters traveling outside $[t^+,t^-]$ be

$$T_{out}=\eta_1\frac{y_2}{s}=\frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s}-\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s} \tag{93}$$

Thus the change of travel cost for the $v\,th$ commuter who travels outside the tolling window is

$$\Delta C_{out}=-\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V\alpha(v)}{s}<0 \tag{94}$$

At equilibrium, the $V\ th$ commuter will incur the same travel cost whether he chooses to travel outside or inside the tolling window $[t^+,t^-]$, and since a uniform toll will not influence the departure time choices for the commuters traveling inside the tolling window, we obtain the generalized travel time for the commuters traveling inside the tolling window by examining the $V\ th$ commuter

$$T_{in}=\frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s}-\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s}-\frac{\rho}{\alpha(V)} \tag{95}$$

Thus for the $v\,th$ commuter who travels inside the tolling window, the change of travel cost can be calculated as

$$\begin{aligned}\Delta C_{in}&=\left(\frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s}-\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s}-\frac{\rho}{\alpha(V)}\right)\alpha(v)+\rho\\&=-\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V\alpha(v)}{s}-\rho\left(\frac{\alpha(v)}{\alpha(V)}-1\right)<0\end{aligned} \tag{96}$$

To sum up eqns.(94) and (96), the changes of travel cost for non-identical commuters across the population are given as a piece-wise linear function:

$$\Delta C=\begin{cases}-\dfrac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\dfrac{V}{s}\alpha(v), & v\le V\\[2ex] -\dfrac{\eta_1\eta_2}{\eta_1+\eta_2}\dfrac{V}{s}\alpha(v)+\rho, & V<v\le N\end{cases} \tag{97}$$

We can see from eqn.(97) that **the optimal coarse toll is still pareto-improving in the heterogeneous case, except that for the commuters traveling inside the tolling period**

$[t^+,t^-]$**, their benefits from the coarse toll increase more quickly as their VOTs become higher**. The latter aspect is shown in Figure 7 for the special case that VOT follows a uniform distribution.

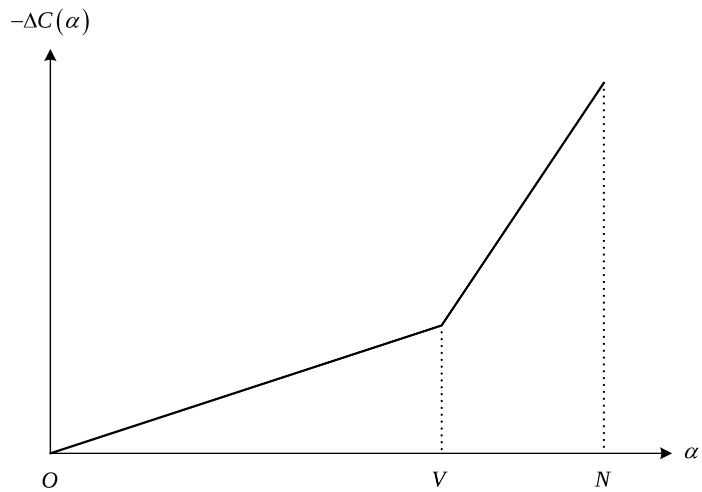

Figure 7. Utility (=-cost) change with VOT across the population

6. A numerical example

In this section a numerical example is used to compare the optimal solutions under different problem settings. We assume the demand $N=100$ and the passing rate $s=50$. The value attached to queuing delay $\alpha(v)=0.128v$, and $\eta_1=0.609$, $\eta_2=2.377$. Thus α is uniformly distributed within $[0,12.8]$ and we can calculate the means $\bar{\alpha}=6.4$, $\bar{\beta}=3.9$ and $\bar{\gamma}=15.21$, which are equal to the values used in the example for homogeneous case. We list the optimal solutions when total system cost is measured in time and monetary units in Table 1. For comparison, the results for the homogeneous case are also listed.

Table 1 Optimal solutions

	Homogeneous	**Heterogeneous (in time)**	**Heterogeneous (in money)**
Total saving	27.08%	27.08%	40.06%
Commuters travel inside	45.84%	45.84%	39.91%

Length of time interval	1.40	1.40	1.34
Optimal toll level	3.10	3.36	4.14
Work staring time	1.526	1.526	1.518
Toll revenue	142.3	154.1	165.2

From Table 1, if we measure the system cost in money, we find the heterogeneous case has a shorter tolling period and higher toll charge. The total percentage saving of the heterogeneous case is around 40%, greater than that calculated under the homogeneous case, which implies that the benefit of coarse toll will be underestimated if heterogeneity is not considered.

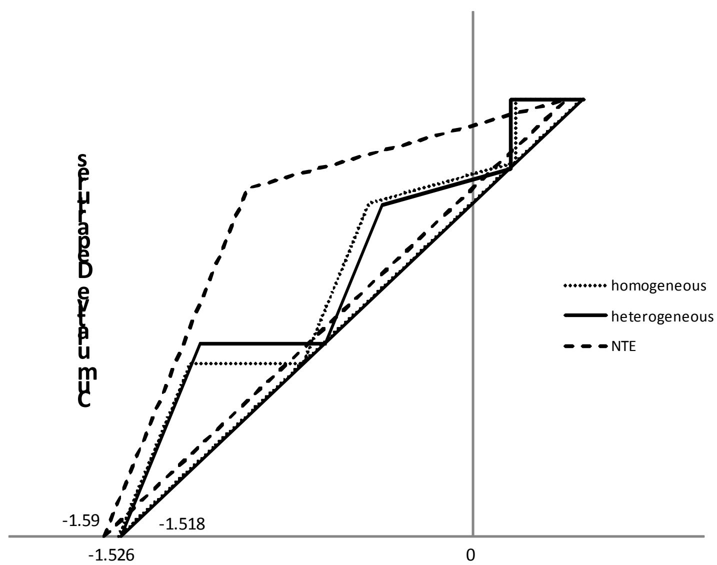

Figure 8. Profiles for numerical examples

Figure 8 shows the difference between the two optimal profiles. Without loss of generality, the work starting time is set to be the origin point. We observe that for both cases the departure time of the first person will off-set to the right compared with NTE, which implies that both coarse-toll schemes are cost-efficient. Less commuters are charged in the peak period and the total percentage saving is greater for the heterogeneous case.

When minimizing the total cost in time unit, we get the same profile with the homogeneous case (See columns 1 and 2 in Table 1). In this special example, the two ends of the population (the poor and the rich) will be benefited if the system performance is measured by money instead of time. Only the middle class will become worse off because the higher toll will force them to change their commute from inside to outside the tolling period and the exemption of toll cannot counteract the increase in delay cost.

7. Conclusions

This study investigates the properties of a single-step coarse toll scheme for a single bottleneck by considering the diverse values of time among the commuters. Though this coarse toll cannot completely eliminate the queue, it has the dual advantages of simplicity and congestion relief: the flat toll is easy to implement and a suitably chosen toll level and tolling window can make every commuter better off than before such a toll is levied.

There could be a range of optimal coarse toll schemes, since the utility functions of the commuters can be weighed differently between money and time. We cannot tell if the optimal toll charged is higher or lower when changing the weights, because it also depends on the form of VOT distribution. But we know that a higher toll charge will narrow the tolling window and for those who are forced to transfer from inside to outside of the tolling window, they will have an increase in their generalized cost.

The problem considered here can be extended in several ways, and here we mention a few of them: Some future extensions can be made by relaxing the assumptions in this study:

1) The assumption of proportionality in characterizing heterogeneity can be relaxed. We can assume that the $\alpha,\ \beta$ and γ parameters have separate distributions and are independent with each other. This, however, will complicate the problem and makes it harder, if possible, to obtain analytical results, and even to define what one means by optimum;

2) Multi-step coarse toll may be considered as a more general case of the one-step coarse toll. In fact when each step becomes infinitely small, the multi-step toll approaches to a fine toll. Clearly, more efficiency gain can be achieved at the price of identifying many more possible profiles, thanks to the increase of degrees of freedom in the optimization.

3) Instead of a single bottleneck, we may consider a corridor with multiple bottlenecks and multiple OD pairs. The problem becomes to determine not only how much and when to charge the toll, but also where to charge the toll. We can also take modal split into consideration so that each individual's cost will be changed in a different way.

Appendices

Appendix A. Solving the unified problem for homogeneous case

To solve the unified problem, we have to discuss the two situations below, respectively:

i) $m \le 0$, which corresponds to profile without capacity waste.

There could be two possibilities

a) When $\rho = 0$

The coefficient λ is equal to zero. We already know from eqn.(6) that the total system cost without toll is just the total system cost for NTE.

$$TC\big|_{\rho=0} = \overline{TC} = \frac{\beta\gamma}{\beta+\gamma}\left(\frac{N^2}{s}\right) \tag{98}$$

b) When $\rho > 0$

λ is always positive. Thus m has to be zero, which means at the starting point of coarse toll, there is no queue and after the starting point of coarse toll, there is no capacity waste. We can calculate the minimized total system cost by optimizing the toll level ρ

$$\begin{aligned} TC_{\min}\big|_{\rho>0} &= \left(\frac{3}{4} - \frac{(\gamma-\alpha)\beta}{4(\beta+\gamma)(\gamma+\alpha)}\right)\frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s} \\ &= \left(\frac{3}{4} - \frac{(\gamma-\alpha)\beta}{4(\beta+\gamma)(\gamma+\alpha)}\right)TC\big|_{\rho=0} < TC\big|_{\rho=0} \end{aligned} \tag{99}$$

Combining a) and b) we obtain that the optimal coarse toll level has to be positive and $m = 0$. The optimal toll level can be calculated by solving the objective function with $\lambda|m| = 0$

$$\rho^s = \frac{\beta\gamma}{2(\beta+\gamma)}\left(\frac{N}{s}\right) \tag{100}$$

ii) $0 < m < \dfrac{\rho}{\beta}$, which corresponds to the profile with capacity waste.

There still could be two possibilities:

a) When $\rho < \dfrac{\beta\gamma}{\beta+\gamma}\left(\dfrac{N}{s}\right)$.

λ is positive. To minimize the total travel cost, m has to be zero. The optimal toll level and minimum total travel cost are the same with those obtained in case i);

b) When $\rho \geq \dfrac{\beta\gamma}{\beta+\gamma}\left(\dfrac{N}{s}\right)$.

Since λ now becomes negative, to minimize the total travel cost, we have to take the upper-bound of m, ρ/β. The problem becomes

$$\min_{\rho} TC = \left(\frac{2s}{\alpha+\gamma}\right)\rho^2 - \frac{2N\beta\gamma}{(\beta+\gamma)(\alpha+\gamma)}\rho + \frac{N^2\beta\gamma}{s(\beta+\gamma)} \tag{101}$$

s.t.

$$\rho \geq \frac{\beta\gamma}{\beta+\gamma}\left(\frac{N}{s}\right) \tag{102}$$

The constraint is binding when the total cost is minimized. And we find the total cost $TC|_{\rho=\frac{\beta\gamma}{\beta+\gamma}\left(\frac{N}{s}\right)} > TC|_{\rho^s}$.

Appendix B. Simplifying the problem corresponding to heterogeneous case profile 1

From eqns.(42) and (45) we have

$$y_1 = \frac{s}{\eta_1}\left(\frac{\rho}{\alpha(V)} - m\right) \tag{103}$$

From eqns.(43) and (103) we have

$$y_2 = \frac{s}{1-\eta_1}\left(x_2 - \frac{\rho}{\alpha(V)}\right) \tag{104}$$

Substituting (104) into the objective function (41) we have

$$\min_{V,x_2,\rho} TC = \frac{\eta_1 K}{1-\eta_1}x_2 - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \tag{105}$$

From eqns.(44) and we have

$$x_2 = \frac{1}{\frac{s}{1-\eta_1} - \frac{s}{1+\eta_2}} \left(\frac{s\rho}{\alpha(V)(1-\eta_1)} + \left(N - \frac{2s\rho}{\alpha(V)(1+\eta_2)} \right) - \frac{s}{1+\eta_2} x_3 \right) \tag{106}$$

From eqns.(48) and (103) we also obtain that

$$m = \frac{\rho}{\alpha(V)} \frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{107}$$

Combining (105), (106) and (107) we can simplify the problem as follows

$$\min_{V,x_3,\rho} TC = \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2} \left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \right) - \frac{1}{1+\eta_2} x_3 \right) - \left(\frac{K}{1-\eta_1} - A_1 \right) \frac{\rho}{\alpha(V)} \tag{108}$$

s.t.

$$x_3 \le \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{109}$$

$$m = \frac{\rho}{\alpha(V)} \frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{110}$$

$$m \le 0 \tag{111}$$

Combining the discussions in both i) and ii) above we are able to conclude that the total system cost is minimized when $m = 0$

Appendix C. Simplify the problem corresponding to heterogeneous case profile 2

From eqns.(60) and (63) we have

$$y_1 = \frac{s\rho}{\eta_1 \alpha(V)} - ms \tag{112}$$

Since y_1 has to be nonnegative, we have

$$m \le \frac{\rho}{\eta_1 \alpha(V)} \tag{113}$$

From eqns.(61) and (112) we have

$$y_2 = \frac{s}{1-\eta_1} \left(x_2 - \frac{\rho}{\alpha(V)} \right) - ms \tag{114}$$

Substituting (114) into (59) we have

$$\min_{V,x_2,y_2} TC = \frac{\eta_1 K}{1-\eta_1} x_2 - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \tag{115}$$

We find that the objective function follows exactly the same form of (105).

From eqns.(62) and (114) we have

$$x_2 = \frac{1}{\frac{s}{1-\eta_1} - \frac{s}{1+\eta_2}}\left(\frac{s\rho}{\alpha(V)(1-\eta_1)} + \left(N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{s}{1+\eta_2}x_3 + ms\right) \tag{116}$$

From eqns. (112) and (66) we also obtain that

$$m = \frac{\rho}{\eta_1 \alpha(V)} \frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{117}$$

Combining (115)-(117), we can simplify the problem as follows

$$\begin{aligned}\min_{V,x_2,y_2} TC = &\frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3 + m\right) \\ &-\left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)}\end{aligned} \tag{118}$$

s.t.

$$x_3 \le \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{s} + m \tag{119}$$

$$m = \frac{\rho}{\eta_1 \alpha(V)} \frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{120}$$

$$0 \le m \le \frac{\rho}{\eta_1 \alpha(V)} \tag{121}$$

Appendix D. Calculation of travel time reduction in the heterogeneous case

To solve y_2 we only need to solve the following group of equations:

$$\begin{cases} \dfrac{y_1}{x_1} = \dfrac{s}{1-\eta_1} \\ \dfrac{y_1}{s} - x_1 = \dfrac{\rho}{\alpha(V)} \\ \dfrac{N - \dfrac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3} = s \\ \dfrac{y_2 - y_1}{x_2 - \dfrac{y_1}{s}} = \dfrac{s}{1-\eta_1} \\ \dfrac{N - \dfrac{2s\rho}{\alpha(V)(1+\eta_2)} - y_2}{x_3 - x_2} = \dfrac{s}{1+\eta_2} \end{cases} \tag{122}$$

There are five unknown independent variables and five independent equations. By the first two equations we have

$$y_1 = \frac{s\rho}{\eta_1 \alpha(V)} \tag{123}$$

Substituting the third equation and (123) into the last two equations we have

$$y_2 = \frac{\eta_2 N}{\eta_1 + \eta_2} - \frac{s\rho(\eta_2 - 1)}{\alpha(V)(1+\eta_2)(\eta_1 + \eta_2)} \tag{124}$$

We have already known the toll level expressed by eqn.(81), thus we have

$$y_2 = \frac{\eta_2 N}{\eta_1 + \eta_2} - \frac{(\eta_2 - 1)\eta_1 V}{(\eta_1 + \eta_2)(1 + 2\eta_1 + \eta_2)} \tag{125}$$

References

Arnott, R., A. Depalma, et al. (1990a). "Departure Time and Route Choice for the Morning Commute." Transportation Research Part B-Methodological **24**(3): 209-228.

Arnott, R., A. Depalma, et al. (1990b). "Economics of a Bottleneck." Journal of Urban Economics **27**(1): 111-130.

Bernstein, D. and I. Elsanhouri (1994). "A Note on Departure Time and Route Choice for the Morning Commute." Transportation Research Part B-Methodological **28**(5): 391-394.

Braid, R. M. (1989). "Uniform Versus Peak-Load Pricing of a Bottleneck with Elastic Demand." Journal of Urban Economics **26**(3): 320-327.

Braid, R. M. (1996). "Peak-load pricing of a transportation route with an unpriced substitute." Journal of Urban Economics **40**(2): 179-197.

Danielis, R. and E. Marcucci (2002). "Bottleneck road congestion pricing with a competing railroad service." Transportation Research Part E-Logistics and Transportation Review **38**(5): 379-388.

Huang, H. J. (2000). "Fares and tolls in a competitive system with transit and highway: the case with two groups of commuters." Transportation Research Part E-Logistics and Transportation Review **36**(4): 267-284.

Laih, C. H. (1994). "Queuing at a Bottleneck with Single-Step and Multistep Tolls." Transportation Research Part a-Policy and Practice **28**(3): 197-208.

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.

Newell, G. F. (1987). "The Morning Commute for Nonidentical Travelers." Transportation Science **21**(2): 74-88.

Tabuchi, T. (1993). "Bottleneck Congestion and Modal Split." Journal of Urban Economics **34**(3): 414-431.

Van der Zijpp, N. and K. Koolstra (2002). "Multiclass continuous-time equilibrium model for departure time choice on single-bottleneck network." Transportation Network Modeling 2002(1783): 134-141.

Vickrey, W. S. (1969). "Congestion Theory and Transport Investment." American Economic Review **59**(2): 251-260.

Ramadurai, G., Ukkusuri, S., Zhao, J. & Pang, J.-S. (2008), "Dynamic equilibrium in multi-user class single bottleneck models: A complementarity formulation", Proceedings of The 87th Annual Meeting of Transportation Research Board.

Lindsey, R. (2004), "Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes", Transportation Science **38**: 293-314.