System optimum dynamic traffic assignment graphical solution method for a congested freeway and one destination

Juan Carlos Muñoz a,*, Jorge A. Laval b

a Department of Transport Engineering, Pontificia Universidad Católica de Chile, C1asilla 306, Santiago 22, Código 105, Chile
b Institute of Transportation Studies, University of California, Berkeley, CA, USA

Received 7 May 2003; received in revised form 27 November 2004; accepted 4 January 2005

Abstract

This paper studies the system optimum dynamic traffic assignment in a network consisting of a hypothetical surface street grid and a congested freeway section. Vehicles can be diverted through off-ramps, and on-ramps can be metered. The family of solutions are identified graphically using Newell’s queueing diagrams. Because enforcing diversion is still a technological puzzle, these results provide a benchmark for future ITS applications, and a building-block for including both departure time choice and several destinations. It is also shown that pricing according to marginal cost would be difficult to implement in this case, that eliminating all queues from the freeway is always suboptimal, and that ramps near the bottleneck should be metered more severely.

© 2005 Elsevier Ltd. All rights reserved.

1. Introduction

Vehicles often remain trapped in queues caused by a freeway bottleneck even when there exists the possibility of bypassing it through local streets. Although taking a bypass may not be convenient for the users as individuals, it may be beneficial for the system as a whole.
In this paper we identify system optimum ramp metering and diversion strategies for the single destination network shown in Fig. 1. We show that for this type of network, the system optimum dynamic traffic assignment (SO-DTA) can be identified by using a very simple graphical method based on cumulative vehicle count curves, and is therefore consistent with vehicle conservation.

Few publications have approached the problem using a graphical solution method based on vehicle conservation. Al-Deek (1993) explored the user optimum (UO) solution for a similar network but focused on incident situations. In that work, it was argued that an SO solution would divert too much traffic to city streets. A UO solution may therefore be more suitable. Our results show that the most appealing SO solution consists of diverting the exact amount of traffic that the city streets can handle, i.e., without queues forming on the off-ramps. Newell (1980) provides an elegant analysis for the case of a freeway in an idealized rectangular grid network. He identifies the geographical location surrounding the freeway that should use the freeway under UO and SO static equilibrium. De Palma and Jehiel (1995) show that some queuing can be socially optimal for a small network where drivers can choose their departure times.

In a numerical counterpart to the graphical approach, Ziliaskopoulos (2000) presented the SO-DTA formulation for a single destination network as a linear program that encapsulates the cell transmission model (Daganzo, 1994), and discovered the corresponding necessary and sufficient optimality conditions. Compared to the methods presented in this paper, Ziliaskopoulos’ approach can handle more complicated networks, but for moderately sized problems involving distances of a few miles the problem becomes unmanageable on a regular personal computer. Additionally, his model assumes that the position of vehicles can be controlled at all times (holding), which makes the solution even more difficult to implement. The graphic method proposed in this paper reproduces the optimality conditions using calculus of variations, and its complexity is independent of its time and space dimensions.

The implementation of the proposed strategies is still a technological puzzle, mainly because diversion does not necessarily benefit diverted drivers and it is hard to enforce. However, the results presented here do provide a benchmark for future ITS applications and a building-block for more realistic cases; e.g., with several destinations or departure time choice.

1 For greater clarity, the on-ramps are shown on the left of the figure.
This paper is organized as follows: Section 2 introduces the problem and provides some insights into its solution. Section 3 illustrates our approach with a simple case. In Section 4 the problem without on-ramps is solved. Section 5 extends this analysis to a network with on-ramps. In Section 6 the results are discussed, practical implementations are analyzed and some suggestions are made for future improvements.

2. Problem definition

The network consists of $I$ on-ramps (origins), $R$ off-ramps, and a single destination at the end of the freeway denoted off-ramp 0 (see Fig. 1). On-ramps and off-ramps are labelled in ascending order, so that on-ramp 1 and off-ramp 1 are the closest to the destination and on-ramp $I$ and off-ramp $R$ are the furthest from it. Note that $I \geq 1$ since the upstream end of the freeway section is modeled as an on-ramp. A bottleneck of capacity $\mu_0$ is located immediately upstream of off-ramp 0. However, vehicles can bypass this bottleneck by taking any of the off-ramps leading into local streets, where traffic conditions are assumed to be freely flowing. The capacity of off-ramp $r$ is $\mu_r$, and is dictated by its discharge capacity to these streets, while the capacity of the on-ramps is assumed to be unlimited (or never reached). Although we assume that queues do not spill back to upstream ramps, we identify solutions where the limitations imposed by this assumption are minimized. The freeway free-flow travel time between on-ramp 0 and on-ramp $i$ is denoted $f_i$, and the free-flow travel time between on-ramp 0 and the freeway’s bottleneck is $t_b$. The trip time from this bottleneck to the destination is assumed to be zero for all travellers. Since we assume no congestion on the city streets, if a vehicle takes off-ramp $r$, it faces a fixed extra trip time of $D_r$; $D_r \geq D_{r-1}$, $r \in [1, R]$; $A_0 = 0$. Notice that this extra trip time is independent of the trip’s origin. Similarly, a vehicle wishing to enter at on-ramp $i$ that is diverted to local streets faces a fixed extra trip time of $d_i$; $d_i \geq d_{i-1}$, $i \in [1, I]$; $d_0 = 0$.

It is assumed that the (vehicle) arrival curve at each on-ramp is known. The goal is to determine the dependant paths vehicles should follow so that the total time spent in the system is minimized. That is, at every on-ramp we ask which vehicles should enter the freeway, and which ones should use the local streets; at every off-ramp, which of the approaching vehicles should be diverted.

2.1. Optimality conditions

The natural dynamic extension of the optimality condition for the static case indicates that for all times and all origins, the route with the least marginal cost is chosen (Ziliaskopoulos, 2000). The marginal cost for a given route corresponds to the extra delay imposed by an additional vehicle on all following vehicles (externality) plus the additional vehicle’s trip time. Thus, marginal costs are a function of future route flows.

As a consequence, a vehicle should never be diverted to an off-ramp if any downstream off-ramp is undersaturated and could be used by those vehicles. This should be obvious since if

---

2 We assume that once a vehicle has exited the freeway it does not return to it.
the vehicle stays on the freeway, the extra trip time along local streets is saved and no other vehicle is impacted. Similarly, for a single-peak period, off-ramps should start diverting vehicles in ascending order and stop diverting in descending order.

3. Uncongested off-ramps, no on-ramps

In this section we examine the case where off-ramps upstream of the bottleneck never become congested \((\mu_0 = \infty, r \geq 1)\). Note that due to the optimality conditions just stated, in this uncongested case only off-ramp 1 carries flow in the SO solution. Therefore, we only need to consider \(R = 1\). In this section we will not consider freeway on-ramps, i.e., \(I = 1\).

We begin by defining the following counting processes, which are initiated upon the passage of a common reference vehicle:

- \(A_i(t)\) = cumulative number of vehicles entered via on-ramp \(i\) that wish to arrive at the destination by time \(t\), \(\forall i \in \{1, \ldots, I\}\).
- \(A_r\) = cumulative number of vehicles entering via on-ramp \(i\) wishing to arrive at the destination by time \(t\), that are diverted through off-ramp \(r\), \(\forall i \in \{1, \ldots, I\}, r \in \{0, \ldots, R\}\); notice that \(A_r(t) = \sum_{r=0}^{R} A_r(t)\). If off-ramp \(r\) is upstream of on-ramp \(i\) then \(A_r(t) = 0\).
- \(D_r(t)\) = cumulative number of vehicles that reached the destination by time \(t\) after entering through on-ramp \(i\) and exiting through off-ramp \(r\), \(\forall i \in \{1, \ldots, I\}, r \in \{0, \ldots, R\}\). If off-ramp \(r\) is upstream of on-ramp \(i\) then \(D_r(t) = 0\).
- \(D(t)\) = cumulative number of vehicles that have reached the destination by time \(t\); notice that \(D(t) = \sum_{i=1}^{I} \sum_{r=0}^{R} D_r(t)\).

In the remainder of this section the superscript \(i\) will be dropped for clarity since there is only one origin.

3.1. Single-peak demand

Consider the case of Fig. 2 where we have a single-peak demand curve \(A(t)\). More general arrival patterns will be considered at the end of the section. Let \(t_0\) be the time when the slope of \(A(t)\) initially exceeds \(\mu_0\); \(T_0\) the moment when the freeway queue vanishes and \(T_1\) the moment when the last diverted vehicle leaves the off-ramp. Also, let \(N\) be the total number of drivers diverted through off-ramp 1 during the entire analysis period.

The optimal solution can be obtained using the following argument. Drivers should not be diverted to the off-ramp when the bottleneck is not active.\(^3\) Once the bottleneck becomes active (at \(t = t_0\)), \(D_0(t)\) grows linearly at a rate \(\mu_0\). If we then assume \(N\) as given we can identify the moment when the queue vanishes, \(T_0\), as the last moment such that \(A(t) - D_0(t)\) is equal to \(N\). Let \(D_0(T_0) = N_0\). If \(N\) is fixed, the total time spent by diverted vehicles is constant (recall that the off-ramp has infinite capacity) and we should therefore only minimize the delay on the freeway (the area between \(A_0(t)\) and \(D_0(t)\)). Now \(D_0(t)\) is already drawn and we know that \(A_0(t)\) passes

\(^3\) A bottleneck is said to be active when a queue is observed only upstream of it.
through \((t_0, A(t_0))\) and \((T_0, N_0)\). Thus, we draw \(A_0(t)\) starting at \((T_0, N_0)\) and proceeding backwards in time with the steepest possible curve subject to the constraint that the slope of \(A_0(t)\) cannot exceed that of \(A(t)\). It follows that \(A_0(t) = \max\{D_0(t), A(t) - N\}\). This defines \(T_1\), the moment when the last diverted vehicle leaves the off-ramp (i.e., when \(D_0(t) = A(t) - N\) for the first time). Once \(A_0(t)\) is identified, \(A_1(t)\) can be drawn as \(A(t) - A_0(t)\) and \(D_1(t)\) as \(A_1(t - A_1)\). All these curves are shown in Fig. 2.

To determine the optimal value of \(N\), we use calculus of variations. At the optimum, a small perturbation \(dN\) induces a variation in the freeway delay of \(dN(T_0 - T_1)\) (shaded area in Fig. 2) and of \(-dNA_1\) in the off-ramp delay. For \(N\) to be optimal, the sum of both quantities should be zero.\(^4\)

Thus, the optimality condition is given by

\[
T_0 - T_1 = A_1
\]

This means that the duration of the queue on the freeway must equal the extra travel time using city streets. Therefore, the unique SO solution for uncongested off-ramps consists in allowing only capacity flow on the freeway (diverting everybody else) until \(T_1\), then stop diverting. The freeway queue will vanish \(A_1\) units of time later. Thus, it should be clear that if a queue shorter than the optimal must be held, the longer the better; if a queue longer than the optimal must be held, the shorter the better.

Graphically, \(A_0(t)\) can be determined by shifting the demand curve \(A(t)\) down vertically until the horizontal distance between the intersection points with \(D_0(t)\) (distance \(T_0 - T_1\)) equals \(A_1\), as shown in Fig. 3.

The same figure also shows the externality that a given vehicle experiencing the queue imposes on the rest of the users, given by \(E(t')\). This is the case because all other vehicles coming behind

\(^4\) If we assume that the slope of \(A(t)\) is higher than \(\mu_0\) at \(T_1\) and and lower than it at \(T_0\), then the second-order terms \(O(dN^2)\) can be ignored.
that given vehicle (after $t'$) would have arrived at the destination $1/\mu_0$ units of time earlier had the latter not taken that trip. Thus, the marginal cost of the trip is equal to $t_f + T_0 - t'$ since it is the sum of the cost experienced by the driver and the externality. Notice that the marginal cost of trips before the queue is triggered and after the queue vanishes is constant and equal to $t_f$. Once the freeway queue is triggered, the marginal cost jumps to its highest level and then decreases with time at a slope of $-1$ until the queue disappears. Fig. 4 displays the marginal cost and its three components (free flow trip time, delay and externalities) for vehicles approaching the off-ramp at different times. From the figure it is clear that pricing according to marginal cost would be hard to implement given that once the queue is triggered, the earlier one arrives at the bottleneck the more expensive the toll would be.

![Fig. 3. Single off-ramp solution for a single-peak demand case.](image)

![Fig. 4. Marginal cost of vehicles approaching off-ramp 1 at different times.](image)
3.2. Multiple-peak demand

Single-peak arrival curves are typical of the morning and evening commuter rush. If $A(t)$ has several peaks, the system optimal solution can still be obtained by shifting the arrival curve vertically. However, a third intersection point might appear before the optimality condition described above is satisfied (see Fig. 5). In this case, we would have three points where the arrival curve, after a shift of $N_s$, touches $D_0(t)$ before optimality condition (1) is met. Let us call the time coordinates of these three points $\tau_1$, $\tau_2$ and $\tau_3$ (note that $\tau_3 - \tau_1 > A_1$, otherwise the optimality condition would have been met). To find the optimal solution we must first identify four different cases. In all four of them the optimal number of diverted vehicles will be at least $N_s$, thus there will be no queue on the freeway before $\tau_1$, at $\tau_2$, and after $\tau_3$. The four cases are

1. $\tau_2 - \tau_1 < A_1$ and $\tau_3 - \tau_2 < A_1$: In this case no further vehicles should be diverted, and $N_s$ corresponds to the optimal number of drivers to be diverted. The first freeway queue will therefore vanish at $\tau_2$ and then immediately begin to grow again. Note that the duration of each queue is shorter than $A_1$. The optimality of this solution can be checked by calculus of variations as follows.

First consider the diversion of one more vehicle at $t$.

If $t \leq \tau_1$ then the diversion has no impact on the freeway queue since the queue would be triggered at the same time. However, the diversion imposes an extra delay of $A_1$ on the diverted vehicle.

If $\tau_1 \leq t \leq \tau_2$ then the vehicles upstream from the diverted vehicle arriving to the bottleneck before $\tau_2$ would benefit from a shorter queue since it would vanish $1/\mu_0$ units of time earlier. Additionally, the diverted vehicle would save its freeway delay. These two benefits add up to exactly $\tau_2 - t$. However, the diverted vehicle would suffer an extra delay of $A_1$. Since
\( \tau_2 - \tau_1 < A_1 \), diverting this vehicle is not optimal. Since at \( \tau_2 \) the freeway queue vanishes, diverting the vehicle does not affect the formation of the second queue (lasting from \( \tau_2 \) to \( \tau_3 \)).

If \( t \leq \tau_2 \leq t \leq \tau_3 \), the argument is similar. The savings on the freeway would be at most \( \tau_3 - \tau_2 \), and since this is not greater than \( A_1 \), it should not be diverted.

Now consider the diversion of one fewer vehicle at \( t \). Since in this solution vehicles are diverted before \( \tau_1 \), the additional vehicle on the freeway would make the queue a bit longer and therefore the two queues would merge into a non-dissipating one in the interval \([\tau_1, \tau_3]\). The extra freeway delay would thus be \( \tau_3 - t \), an amount greater than \( \tau_3 - \tau_1 \) which in turn exceeds \( A_1 \). Therefore, this perturbation is also non-optimal.

2. \( \tau_2 - \tau_1 > A_1 \) and \( \tau_3 - \tau_2 < A_1 \): For the interval \([\tau_1, \tau_2]\) the solution is as in the single-peak case; i.e., the portion of \( A(t) \) in \([\tau_1, \tau_2]\) is shifted down until the new intersection points define a distance of \( A_1 \); for the interval \([\tau_2, \tau_3]\) no further shifting is necessary.

3. \( \tau_2 - \tau_1 < A_1 \) and \( \tau_3 - \tau_2 > A_1 \): Analogous to case 2.

4. \( \tau_2 - \tau_1 > A_1 \) and \( \tau_3 - \tau_2 > A_1 \): Both intervals can be further shifted independently until the single-peak SO solution is found for each one.

Our assumption that off-ramps have an infinite capacity is, of course, unrealistic. But this solution provides valuable insights for the more realistic case of finite capacities, which is analyzed in the next sections.

4. Capacitated off-ramps, no on-ramps

In this section, we consider the case where off-ramps have limited capacity to handle diverted vehicles so that bottlenecks with capacity \( \mu_r \) are located at the end of each off-ramp. We add the following notation:

- \( N_r \) total number of vehicles diverted through off-ramp \( r \), \( r \in [0, \ldots, R] \); \( N = \sum_{r=1}^{R} N_r \)
- \( T_r \) time at which the last driver diverted to off-ramp \( r \) leaves the off-ramp, \( r \in [0, \ldots, R] \)
- \( t_r \) the time at which the slope of \( A(t) \) initially exceeds \( \sum_{j=0}^{R} \mu_j \), \( r \in [0, \ldots, R] \)

4.1. Single off-ramp

We begin by deriving the optimality conditions for a freeway with only one off-ramp upstream of the bottleneck. Let us assume that \( N \) vehicles are diverted to off-ramp 1 during the entire period. Thus, the objective is to minimize the total delay (the area between \( A(t) \) and \( D(t) \)) since the extra trip time along local streets would be fixed. We already know from the previous section that in a SO solution, if queues grow on the off-ramp and at the freeway bottleneck, the queue on the off-ramp will disappear earlier. Thus, in an optimal solution the system should process vehicles as fast as possible until \( N \) vehicles have been diverted to off-ramp 1. Diversion then stops and the freeway bottleneck works at capacity until the queue vanishes.

The top part of Fig. 6(a) shows the construction of a curve \( D_0(t) \) satisfying this condition. The bottom part of the figure shows the construction of the corresponding arrival and departure curves at the off-ramp.
The optimality condition for \( N \) can be obtained by calculus of variation. Shifting \( dN \) vehicles from the off-ramp to the freeway at time \( t^* \) induces a variation in the freeway delay of \( dN(T_0 - t^*) \) and in the off-ramp 1 delay of \( dN(T_1 - t^* + \Delta_1) \). In the SO, these marginal delays must be identical. Thus, a necessary optimality condition for this case is identical to the uncongested case:

\[
T_0 = T_1 + \Delta_1
\]  

Now the interpretation is that the queue on the freeway should clear \( \Delta_1 \) time units after it disappears from the off-ramp. Note that the solution does not define how we should allocate vehicles arriving between \( t_1 \) and \( T_1 \), it merely stipulates that during that period both bottlenecks should be working at capacity. We therefore conclude that the optimal solution is not unique. Indeed, Fig. 6(a) and (b) represent two extreme optimal solutions. In Fig. 6(a), the freeway is kept as empty as possible, while in Fig. 6(b) no queues are allowed to grow on the ramp. Clearly, since the number of vehicles taking each route and the total delay for the two solutions are identical, their total cost are the same as well. Also, since the problem is linear (see Laval and Muñoz (2002) for a formal formulation), any linear combination of these two solutions is also optimal.
4.2. Multiple off-ramps

As before, the optimum value of $N$ must be such that a small perturbation $dN$ will cancel out delays on the freeway and savings on the off-ramps, regardless of what proportion of $dN$ is assigned to each ramp. The reader is referred to Laval and Muñoz (2002) where it is demonstrated that when $R = 2$ the additional optimality condition is as follows independent of the partition:

$$T_1 = T_2 + A_2 - A_1$$

(3)

Notice that (2) is similar to (3) since $A_2 - A_1$ represents the extra free flow trip time through local streets between ramps 1 and 2. This is illustrated here in Fig. 7 where the unique optimal $D(t)$ is displayed. Note that its associated $D_i(t), i \in 0, 1, 2$ can also be uniquely identified. However, different $A_i(t), i \in 0, 1, 2$ can be built from this figure. That is, the intervals when off-ramps are used are defined uniquely, but the allocation of the queues is not.

We can now generalize the optimality conditions for this problem in the following manner:

1. If an off-ramp is to be used (all downstream bottlenecks are working at capacity), it is used at maximum throughput. Therefore,
   (a) $\dot{A}_r(t) = \dot{D}_r(t) = \dot{A}(t) - \sum_{j=0}^{r-1} \mu_j, \ \forall t \in [t_{r-1}, t_r], \ \forall r \in [1, R]$.
   (b) $\dot{D}_r(t) = \mu_r, \ \forall t \in [t_r, T_r], \ \forall r \in [1, R]$.
2. All demand must be served: $\sum_{r=0}^{R} A_r(t) = A(t), \ \forall t, \forall r \in [1, R]$.

![Graph showing cumulative number in freeway](image-url)
3. Arrival curves are non-decreasing functions: \( \dot{A}_r(t) \geq 0, \forall t, \forall r \in [0, R] \).
4. Once the queue on off-ramp \( r \) is cleared, no on-coming vehicle will take it: \( \dot{A}_r(t) = \dot{D}_r(t) = 0, \forall t > T_r, \forall r \in [1, R] \).
5. The queue on ramp \( r \) ends \( A_r - A_{r-1} \) time units earlier than on \( r - 1 \): \( T_r = T_{r-1} - (A_r - A_{r-1}) \), \( \forall r \in [1, R] \).

Graphically, the solution is quite intuitive. Let us assume that we know \( T_0 \). It follows that \( D(t) \) goes through the point \((T_0, A(T_0))\) allowing us to construct \( D(t) \) backward in time. We are interested in identifying the shape of \( D(t) \) when a queue is triggered at the bottleneck. Whenever the bottleneck is not active, \( D(t) = A(t) \). The shape of \( D(t) \) will depend on how many off-ramps upstream of the bottleneck are used in the SO solution. If upstream off-ramps are never used, then \( D(t) \) is equal to \( A(t) \) except for a period just before \( T_0 \) and shorter than \( A_1 \), during which \( D(t) \) grows at slope \( \mu_0 \). If only one upstream off-ramp is used, then \( D(t) \) is equal to \( A(t) \) except for \( A_1 \) units of time before \( T_0 \) where \( D(t) \) grows at a rate \( \mu_0 \), and a period shorter than \( A_2 \) immediately before the period during which \( D(t) \) grows at rate \( \mu_0 + \mu_1 \). Thus, if the SO solution includes flow through \( p \) upstream off-ramps, then while a queue exists in the system, \( D(t) \) will be piecewise linear with \( p + 1 \) segments. Fig. 7 provides an illustration of the shape of \( D(t) \) for the case \( p = 2 \). The piecewise section of this curve lasting from \( t = t_p \) to \( t = T_0 \), can be described as

\[
D(t) = A(T_0) - \sum_{j=0}^{i} \mu_j(T_0 - t) + \sum_{j=1}^{i} \Delta t_j \left\{ \begin{array}{ll}
\forall t \in [t_p, T_0 - A_p] & \text{if } i = p \\
\forall t \in [T_0 - A_k, T_0 - A_{k-1}] & \text{if } i = k - 1
\end{array} \right. 
\]

However, the actual values of \( T_0, p \) and \( t_p \) remain to be identified. The problem is solved by assuming that \( T_0 \) takes a value large enough for the queue to clear without diversion, that \( p \) is the highest number of off-ramps that could possibly be used to divert vehicles, and that \( t_p = -\infty \). Under these assumptions, \( D(t) \) and \( A(t) \) should only intersect at \( t = T_0 \). Fig. 8(a) shows a good example of an initial \( T_0 \) for the case where \( p = 3 \). To find the optimal \( D(t) \), we reduce \( T_0 \) and move \( D(t) \) along \( A(t) \) until the two curves first touch, say at \( t = \tau \). We redefine \( p \) as the number of segments along \( D(t) \) between \( \tau \) and \( T_0 \), minus one, and \( t_p \) as \( \tau \).

Two types of intersection points \( t_p \) may exist: either (a) a corner of \( D(t) \) such that \( t_p + A_p = T_0 \) and the slope of \( A(t) \) at \( t_p \) is between the slope of \( D(t) \) immediately before and immediately after \( t_p \), that is:

\[
\dot{D}(t_p^+) \leq \dot{A}(t_p) \leq \dot{D}(t_p^-)
\]

or (b) a point \( t_p \) where the slope of \( D(t) \) and \( A(t) \) coincide as in Fig. 8.

Note that with this procedure, \( T_r \) and \( N_r, \forall r \in [1, \ldots, R] \) are uniquely identified for any optimal solution. However, as was illustrated for \( R = 1 \), there are multiple solutions of \( A_r(t) \) \( \forall r \) that yield the same optimal total cost. Therefore, it is not important which driver goes to which ramp, given that the number of diverted drivers for each ramp is fixed and that the optimality conditions specified earlier in this section are satisfied. This provides flexibility as to where to send the diverted traffic, which is useful when there is finite storage space.
5. Off-ramps and on-ramps

In this section we explain how to incorporate on-ramps with their own cumulative demand curves into the analysis. In this case, we cannot only divert vehicles through off-ramps but also restrict on-ramp entrance to certain vehicles, diverting them through local streets.

Fig. 8. General problem with many off-ramps, no on-ramps: (a) Initial step to identify an optimal solution. (b) Optimal solution.
5.1. No on-ramps, no off-ramps

The simplest case we can envision is where $I = 1$ and $R = 0$; i.e., an OD pair connected by a homogeneous freeway section and city streets. In the SO solution, vehicles would be diverted to prevent a queue on the freeway lasting longer than $\delta_0$. As soon as the queueing time on the freeway falls below $\delta_0$, diversion should be stopped. This case is thus equivalent to the single uncongested off-ramp case in Section 3.1.

5.2. Single on-ramp, no off-ramps

If $I = 2$ and $R = 0$, we can find the solution using the tools developed in the previous section after applying two simple modelling techniques. First, we assume that all vehicles that would arrive at on-ramp 1 instead arrive at on-ramp 2, $f_1$ units of time earlier. Next, we model on-ramp 1 as an “artificial” off-ramp with a capacity equal to the (time dependent) on-ramp’s demand rate. Then, every vehicle taking the artificial off-ramp represents a vehicle that never took the on-ramp in the original problem. Analogously, vehicles not taking the off-ramp represent vehicles entering the freeway through the on-ramp. Now the problem has shifted from $I = 2$, $R = 0$ to $I = 1$, $R = 1$ with a time dependent off-ramp capacity, $\mu_1(t)$, given by the on-ramp’s demand rate.

The procedure for solving this problem is identical to the constant capacity case solved in Section 4.1. The procedure to determine $D_0(t)$ and the sensitivity analysis to deduce (2) are still valid. However, the ramp will now start working at capacity from $t_1$, the earliest time satisfying $A(t_1) = \mu_0 + \mu_1(t_1)$. Note that now $D(t)$ and $D_1(t)$ are no longer linear during the period $[t_1, T_1]$.

5.3. Multiple on-ramps, no off-ramps

The case where $I = I'$, and $R = 0$ can be solved as a straightforward extension of the preceding case (Section 5.2) using the optimal procedure for the $I = 1$, $R = I' - 1$ case, as given in Section 4.2. As before, all on-ramps are modeled as artificial off-ramps using the arrival rates as capacities. Their arrival curves are shifted to on-ramp 0 by their respective free-flow trip times. Now the procedure outlined at the end of Section 4.2 can be applied to this problem where $D(t)$ would still consist of segments which may no longer be linear. As before, if we count the segments starting from the later one, segment $i$ would be $\delta_i - \delta_{i-1}$ units of time long. But now its slope would be $\mu_0 + \sum_{j=1}^{i-1} \dot{A}(t)$. Notice that each time $D(t)$ is shifted to the left, the rates $\mu_i(t)$ change. Thus $D(t)$ must be recomputed accordingly; see Fig. 9.

5.4. Multiple on-ramps, multiple off-ramps

Here we simply add real off-ramps to the model of the previous section. The slopes of the segments of $D(t)$ will therefore be the sum of some off-ramp capacities and some on-ramp arrival rates. Given that we activate first the closest off-ramp (or on-ramp) to the bottleneck, and then move sequentially to those further upstream, the solution obtained will still be feasible since vehicles arriving at an on-ramp will never exit through an upstream off-ramp.

If the capacity of on-ramps is likely to be reached, the problem could be handled in approximate fashion by solving an SO-DTA for the system defined by the on-ramp and its city street
alternative only, as in Section 4.1. In this way the delays on the on-ramps are taken into account. The resulting optimal cumulative departure curve from the on-ramp becomes its demand curve for the purpose of the above analysis.

6. Discussion

We have been able to identify the SO-DTA for a simple but nevertheless very common network. Our approach would appear to be more appealing for the evening commute problem or for incident management, since we have assumed that drivers cannot change their departure time in accordance with their expected travel time.

Although our assumption of no congestion on local streets is not very realistic, our results should still be helpful for practitioners and provide a benchmark for future Intelligent Transportation Systems applications. As long as demand remains unchanged, time savings by any ITS application will be bounded by those identified in this paper.

We have derived the periods when vehicles should be diverted at each upstream off-ramp, and when on-ramp metering rates should be activated. If vehicles face congestion in local streets, the diverting period for an on-ramp should still start at the time suggested here but should end earlier. Then, since the $\Delta$'s would be larger, we expect that fewer ramps should be used for diversion.

When city streets are congested the problem is more complicated because (i) external users are affected by diverted vehicles, and (ii) travel times on off-ramp routes would be affected by flows on other off-ramps.
Another practical limitation is that in general, the freeway and local street networks are owned and operated by different agencies, which may have different priorities. In this case, the agencies should work in coordination with each other in order to minimize the expected total delay in the system (e.g., traffic lights near freeways coordinated with flow diversion).

Our model also assumes that no flows are attracted by destinations close to the off-ramps (local flow). However, we can incorporate these local flows as long as they originate from non-metered on-ramps and there is no queue on the exit off-ramps (otherwise we would need to distinguish vehicles according to destination at on-ramps and off-ramps). Fortunately, the solution with no queue on the off-ramps takes care of the second condition. If in addition the first condition is valid, the capacity of each off-ramp should be reduced by the (time dependant) local flow.

Implementing the suggested policies is likely to be challenging. Clearly, SO solutions are not obtained spontaneously since they are not user optimum. Diverted vehicles are individually better off staying on the freeway, so the simplest way to implement these solutions seems to be through enforcement. Unfortunately, SO tolls are hard to implement, mainly because of the discontinuities in the marginal cost on freeway routes in every $t_l$. Additionally, in alternative rationing systems based on license plates our approach would not work since we would need to distinguish a subset of drivers from the rest at the on-ramps. As is shown in Erera et al. (2002), the problem is NP-hard.

Our results also suggest that during peak periods the best course of action is to close on-ramps located near the bottleneck. Since highway authorities would likely be unwilling to implement such a drastic policy, the lowest acceptable ramp metering rate should be deployed. In our model we would then subtract this maximum metering rate from the previous capacity of the artificial off-ramp.

The framework proposed in this paper can be extended to include (i) demand uncertainty, (ii) departure time choice and (iii) congestion on city streets. Ongoing research shows that graphical solutions can still be obtained when (i) and (ii) are included. Numerical methods may be necessary for taking into account (iii).

References


