SO-DTA Graphical Solution Method for a Congested Freeway and One Destination

Juan Carlos Muñoz[1] and Jorge A. Laval
Department of Civil and Environmental Engineering
Transportation Group
University of California, Berkeley

Abstract

This paper studies the system optimum dynamic traffic assignment in a network consisting of a 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, several destinations and departure time choice. This paper also shows why pricing according to marginal cost cannot work; that eliminating all queues from the freeway is always suboptimal; and that ramps near the bottleneck should be metered more severely.

1 Introduction

Quite often vehicles are entrapped in queues caused by a freeway bottleneck despite the possibility of bypassing the bottleneck through local streets. Although taking a bypass may not be convenient for the users as individuals,

[1]Instructor on leave at the Pontificia Universidad Catlica de Chile, Ph.D. student at U.C. Berkeley

it might be beneficial for the system as a whole. The results suggest that, under congestion, keeping the freeway free of queues is always suboptimal. Additionally, the results provide a benchmark for future Intelligent Transportation Systems applications.

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, which yields the optimal flows in each path over time.

1.1 Literature review

There are few publications that approach the problem with a graphical solution method based on conservation of vehicles. Al-Deek (1993) explored the user optimum (UO) solution over a similar network but focused on incident situations. In that work, it was argued that a SO solution would divert too much traffic to city streets. Thus, a UO solution is more suitable. Our results show that the most appealing SO solution consists of diverting the traffic that the city streets can handle, i.e. no queues 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 (1994) show that some queuing can be socially optimal for a small network where drivers can choose their departure times.

In the numerical counterpart, Ziliaskopoulos (2000) presented the SO-DTA formulation for a single destination network as a linear program that encapsulates the cell transmission model (Daganzo, 1994). He found the necessary and sufficient optimality conditions for this problem. Compared to the methods presented in this paper Ziliaskopoulos' approach can handle more complicated networks, but for moderate size problems involving a few miles the problem becomes unmanageable for 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 harder to implement. The graphic method proposed in this paper recovers the optimality conditions using calculus of variations and its complexity is independent of its time and space dimensions.

This paper is organized as follows: Section 2 introduces the problem and

provides some insights of its solution. Section 3 illustrates our approach with the simplest case we can think. In section 4 the problem without on-ramps is solved. Section 5 extends this work for a network with on-ramps. In section 6 a stochastic arrival process (not known a priori) is considered. In section 7 the results are discussed, practical implementations are analyzed, and some future developments are suggested.

2 Problem definition

The network consists of I on-ramps (origins), R off-ramps and a single destination at the end of the freeway, see Fig. 1. The end of the freeway is denoted off-ramp 0. On- and off-ramps labelled 1 are closest to the destination (and on- and off-ramps I and R are the furthest). A bottleneck of capacity μ_0 is located immediately upstream from off-ramp 0. However, vehicles can bypass this bottleneck by taking any of the off-ramps heading to local streets[2]. We assume freely flowing traffic conditions in the local streets. The capacity of off-ramp r is μ_r and is dictated by its discharge capacity to the city streets. On-ramps' capacity is assumed unlimited (or never reached). Although we assume that queues don't occupy space (point queues without spillovers) we identify solutions where the limitations of this assumption are minimized. The freeway free-flow travel time between on-ramp 0 and on-ramp i is denoted f_i, and the travel time between on-ramp 0 and the freeway's bottleneck is t_f. The trip time from this bottleneck to the destination is assumed zero for all travellers. Since we assume no congestion in the city streets, if a vehicle takes off-ramp r, it faces a fixed *extra* trip time of $\Delta_r; \Delta_r \geq \Delta_{r-1}, r \in [1, R]; \Delta_0 = 0$. 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 $\delta_i; \delta_i \geq \delta_{i-1}, i \in [1, I]; \delta_0 = 0$.

For the case of ramp metering, we restrict the set of all possible routes to enter an on-ramp or stay on local streets; while for diversion the routes are take an off-ramp or stay on the freeway. In essence, we are avoiding routes the either re-enter the freeway or enters the freeway by another on-ramp. Our goal is to determine at each origin, which vehicles take each route.

Initially, it is assumed that the (vehicle) arrival curve at each on-ramp is known, although we will attempt to solve the stochastic case later in this

[2]We assume that once a vehicle has exited the freeway it does not get back on.

paper. The goal is to determine the time dependant paths vehicles should follow so that the total time spent in the system is minimized. That is, at every on-ramp which vehicles should enter the freeway, and which ones should use the local streets; at every off-ramp which of the on-coming vehicles should be diverted.

2.1 Optimality conditions

The natural dynamic extension of the optimality condition for the static case is that, for all times, at all origins the route with the least marginal cost is chosen. The marginal cost on a given route corresponds to the extra delay caused to all upcoming vehicles when an additional vehicle uses the route (externality) plus the 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 of the off-ramps downstream are not at capacity and could serve those vehicles. This should be obvious since if the vehicle stays on the freeway, the extra trip time along local streets is saved and no other vehicle is impacted by this decision. Similarly, for a single peak period, off-ramps should be used for diversion in ascending order and this use should be stopped in descending order.

3 Uncongested off-ramps, no on-ramps

In this section we examine the case when off-ramps upstream of the bottleneck never get congested ($\mu_r = \infty, r \geq 1$). Notice that due to corollary 1, 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 on-ramps in the freeway, i.e. I=0.

Let's define the following counting processes started after the passage of a common reference vehicle:

$A^i(t)$ = cumulative number of vehicles that having entered via on-ramp i would like to arrive to the destination by time t, $\forall i \in [0 \dots I]$
$A^i_r(t)$ = cumulative number of vehicles that having entered via on-ramp i and wanting to arrive to the destination by time t, are diverted through off-ramp r, $\forall i \in [0 \dots I], r \in [0 \dots R]; A^i(t) = \sum_{r=0}^{R} A^i_r(t)$. If off-ramp r is upstream

from on-ramp i then $A_r^i(t) = 0$.
$D_r^i(t)$ = cumulative number of vehicles that reached the destination by time t after entering through on-ramp i and exiting through off-ramp r, $\forall i \in [0 \dots I], r \in [0 \dots R]$. If off-ramp r is upstream from on-ramp i then $D_r^i(t) = 0$.
$D(t)$ = cumulative number of vehicles that have reached the destination by time t; $D(t) = \sum_{i=0}^{I} \sum_{r=0}^{R} D_r^i(t)$.

In the reminder of this section the super index i will be dropped for clarity since only one origin is considered.

3.1 Single-peak demand

Consider the case of Fig. 2a where we have a single peak demand curve $A(t)$. More general arrival patterns will be considered at the end of the section. Let's call t_0 the time when the slope of $A(t)$ initially exceeds μ_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 whole period of analysis.

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. After the bottleneck becomes active (at $t = t_0$), $D_0(t)$ grows linearly at a rate μ_0. Then, if we assume N as given we can identify the moment when the queue will vanish, T_0, as the last moment such that $A(t) - D_0(t)$ is equal to N. We call $D_0(T_0) = N_0$. If N is fixed, the total time spent by diverted vehicles is constant (recall that the off-ramp has infinite capacity) and therefore we should 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 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 $A_0(t)$ can not 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 vehicle diverted 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 $A_1(t - \Delta_1)$. All these curves are shown in Fig. 2a.

To determine the optimal value of N, we note that, in the optimum, a small perturbation dN induces a variation on the freeway delay of $dN(T_0 - T_1)$ (shaded area in Fig. 2b) and of $-dN\Delta_1$ on the off-ramp delay. For N to be

optimal the sum of both quantities should be zero[3].

Thus, the optimality condition is simply:

$$T_0 - T_1 = \Delta_1 \tag{1}$$

This means that the duration of the queued episode 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 Δ_1 units of time later.

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 Δ_1, as shown in Fig. 3.

This queueing diagram also shows the externality that a trip causes to the rest of the users as $E(t^*)$ in the figure. This is true since all vehicles to be queued coming behind of the vehicle (after t^*) would arrive to the destination $1/\mu_0$ units of time earlier had the trip not been realized. 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 vanishes. The reader is referred to Laval and Muñoz (2002) for details.

3.2 Multiple-peak demand

Single-peak arrival curves are typical of the morning and evening commute. However, 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. 4 for an illustration). In this case, we would have three points where the arrival curve, after a shift of N_s, touches $D_0(t)$. Let's call these points τ_1, τ_2 and τ_3 ($\tau_3 - \tau_1 > \Delta_1$). In this case identifying the optimal solution requires distinguishing four cases. In all of them the optimal number of vehicles to divert will be at least N_s, thus there will be no queue in the freeway before τ_1, at τ_2, and after τ_3.

[3]Second order terms are ignored.

1. $\tau_2 - \tau_1 < \Delta_1$ and $\tau_3 - \tau_2 < \Delta_1$: Then N_s corresponds to the optimal number of drivers to divert since further shifting would end in queues shorter than Δ_1.

2. $\tau_2 - \tau_1 > \Delta_1$ and $\tau_3 - \tau_2 < \Delta_1$: For the interval $[\tau_1, \tau_2]$ the solution is as in the single peak case, i.e. keep shifting down the portion of $A(t)$ in $[\tau_1, \tau_2]$ until the new intersection points define a distance of Δ_1; for the interval $[\tau_2, \tau_3]$ no further shifting is necessary.

3. $\tau_2 - \tau_1 < \Delta_1$ and $\tau_3 - \tau_2 > \Delta_1$: analogous to case 2.

4. $\tau_2 - \tau_1 > \Delta_1$ and $\tau_3 - \tau_2 > \Delta_1$: Both intervals can be further shifted independently until the single-peak SO solution is found for each interval.

Assuming that off-ramps have an infinite capacity is unrealistic. However, its solution provides valuable insights for the more realistic case of finite capacities, analyzed in the following 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 their bottlenecks are located at the end of each off-ramp. To this end, let us add the following notation:

N_r = Total number of vehicles diverted through off-ramp r, $r \in [0 \dots R]$;
$N = \sum_{r=1}^{R} N_r$
T_r = Time when the last driver diverted to off-ramp r leaves the off-ramp, $r \in [0 \dots R]$
t_r = The time when the slope of $A(t)$ initially exceeds $\sum_{j=0}^{r} \mu_j$, $r \in [0 \dots R]$.

4.1 Single off-ramp

We will first derive the optimality conditions for a freeway with only one off-ramp upstream from the bottleneck. Let assume that N vehicles are diverted to off-ramp 1 during the whole 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. According to corollary 1, if a queue grows in

the off-ramp and the bottleneck the queue in the off-ramp will vanish 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. Then, diversion stops and the freeway bottleneck works at capacity until the queue vanishes.

The top half of Fig. 5(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 curve at the off-ramp.

The optimality condition for N can be obtained after we perturb it by a small dN (shifting dN vehicles from the off-ramp to the freeway at time t^*). This perturbation induces a variation in the freeway delay of $dN(T_0 - t^*)$ and on off-ramp 1's delay of $dN(T_1 - t^* + \Delta_1)$. In the SO optimum, these marginal delays must be identical. Thus, a necessary optimality condition for this case reads simply:

$$T_0 = T_1 + \Delta_1 \tag{2}$$

This means that the queue in the freeway should clear Δ_1 time units after it vanishes on the off-ramp. Note that the solution does not define how we should allocate vehicles arriving between t_1 and T_1. It only stipulates that during that period both servers should work at capacity. We conclude that the optimal solution is not unique. Indeed, Figs. 5(a) and 5(b) represent two extreme optimal solutions. In Fig. 5(a), the freeway has been kept as empty as possible, while in Fig. 5(b) no queues were allowed to grow in the ramp. Notice that 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) to verify that the additional optimality condition is $T_1 = T_2 + \Delta_2 - \Delta_1$ when $R = 2$, independent of the partition. This is illustrated in figure 6 where the unique optimal $D(t)$ is displayed. Notice 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 are not.

Now we can generalize the optimality conditions for this problem as:

1. If an off-ramp will be used then it starts being used as soon as it is needed and is used at capacity during its diversion period. Therefore, if $A_r(T) \geq 0$, then $\forall r \in [1, R]$:
   (a) $\dot{A}_r(t) = \dot{D}_r(t) = \dot{A}(t) - \sum_{j=0}^{r-1} \mu_j$, $\forall t \in [t_{r-1}, t_r]$.
   (b) $\dot{D}_r(t) = \mu_r$, $\forall t \in [t_r, T_r]$.
2. All the demand must be served: $\sum_{r=0}^{R} A_r(t) = A(t)$, $\forall t$
3. Arrival curves are nondecreasing functions:
4. $\dot{A}_r(t) \geq 0$, $\forall t$, $\forall r \in [0, R]$
5. After the queue on off-ramp r is cleared, no on-coming vehicle will take it: $\dot{A}_r(t) = \dot{D}_r(t) = 0$, $\forall t > T_r$, $\forall r \in [1, R]$
6. The queue in ramp r ends $\Delta_r - \Delta_{r-1}$ time units earlier than in $r-1$: $T_r = T_{r-1} - (\Delta_r - \Delta_{r-1})$, $\forall r \in [1, R]$

Graphically, the solution is quite intuitive. Let's 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. If off-ramp 1 is used, then the slope of $D(t)$ is μ_0 in the interval $[T_0 - \Delta_1, T_0]$. If off-ramp 2 is used, then the slope of $D(t)$ is $\mu_0 + \mu_1$ in the interval $[T_0 - \Delta_2, T_0 - \Delta_1]$ and so on. Therefore, $D(t)$ will be piece-wise linear with at most p pieces. Fig. 6 provides an illustration of the shape of $D(t)$ for the case $p = 3$. Each piece i of this curve can be described as a straight segment:

$$D_i(t) = a_i + b_i(t - \tau_i) \quad \begin{array}{ll} \forall t \in [-\infty, \tau_1] & \text{if } i = 1 \\ \forall t \in [\tau_{i-1}, \tau_i] & \text{if } i > 1 \end{array}$$

where $\tau_i = \Delta_p - \Delta_{p-i}$, $b_i = \sum_{j=0}^{p-i} \mu_j$, and $a_i = \sum_{k=1}^{i} (\tau_k - \tau_{k-1}) = a_{i-1} + (\tau_i - \tau_{i-1}) b_i$. The border conditions are $\tau_0 = 0, a_0 = 0$.

Then the game is to define a very late T_0 and build $D(t)$ from there[4]. A good choice of T_0 is the time when the queue would clear without diversion.

[4]Build D(t) from there means drawing the following function: $D(t + \Delta_p - T_0) + A(T_0) - D(\Delta_p)$

Fig. 7(a) shows a good example of an initial T_0. Then we should reduce T_0 and move $D(t)$ along $A(t)$ until the two curves first touch. Let's call this time $t = \tau$.

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

$$\dot{D}(\tau^+) \leq \dot{A}(\tau) \leq \dot{D}(\tau^-)$$

or b) a point $\tau = t_i$ where the slope of $D(t)$ and $A(t)$ coincide. The optimal solution for Fig. 7(a) is shown in Fig. 7(b).

Note that with this procedure T_r and $N_r, \forall r \in [1 \dots R]$ 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 gives flexibility as to where to send the diverted traffic, which is useful when we have finite storage space.

5 Off-ramps and On-ramps

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

5.1 No On-ramps, no off-ramps

The simplest case we can envision is when $I = 0, R = 0$. Then, in the SO solution vehicles would be diverted to prevent a queue in the freeway lasting longer than δ_0. As soon as the queue in the freeway will last shorter than δ_0, diversion should be stopped. Therefore, this problem is equivalent to the single uncongested off-ramp case seen in section 3.1.

5.2 Single On-ramp, no off-ramps

If $I = 1, R = 0$, then we can solve the problem using the tools developed in the previous section after we apply two simple modelling tricks. First

we assume that all vehicles that would arrive at on-ramp 1 instead arrive to on-ramp 2 f_1 units of time earlier. Next we model on-ramp 1 as an off-ramp with a capacity equal to the (time dependent) on-ramp's demand rate. Then every vehicle taking the off-ramp represents a vehicle that never took the on-ramp in the original problem. Analogously, vehicles not taking the off-ramp represents vehicles entering the freeway through the on-ramp. Now the problem has shifted from $I = 1, R = 0$ to a $I = 0, R = 1$ with a time dependent capacity off-ramp, $\mu_1(t)$ given by the on-ramp's demand rate.

The procedure to solve this problem is identical to the constant capacity case solved in 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 $\dot{A}(t_1) = \mu_0 + \mu_1(t_1)$. 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 $I = I', R = 0$ can be solved as a straightforward extension of the previous case considering the optimal procedure for the $I = 0, R = I'$ case. As before, all on-ramps are modeled as off-ramps using the arrival rates as capacities. Their arrival curves are shifted to on-ramp 0 by their respective free-flow time. Now, the procedure outlined at the end of section 4.2 can be applied to this problem where $D(t)$ would still have pieces but is no longer linear. As before, if we count the pieces starting from the later one, piece i would be $\delta_i - \delta_{i-1}$ units of time long. But now its slope would be $\mu_0 + \sum_{j=1}^{i-1} \dot{A}^j(t)$. 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. 8 as an illustration of an optimal solution.

5.4 Multiple On-ramps, multiple off-ramps

In the previous case, we model on-ramps as off-ramps. Now we add more off-ramps to that model. Therefore, the slopes of the pieces of $D(t)$ will be the sum of some off-ramp capacities and some on-ramp arrival rates. Notice that since we activate first the closest off-ramp (or on-ramp) from the bottleneck and then sequentially move to those upstream, the solution obtained will still be feasible since vehicles arriving to an on-ramp will never exit through an off-ramp upstream.

If the capacity of on-ramps is likely to be reached, then the problem could be handled approximately by solving a 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.

In the following section we extend these results to a more realistic situation when $A(t)$ is a random process. We will see that among the multiple optimal solutions of the deterministic case, we would rather use one particular solution over the others.

6 Uncertain Demand

The solutions presented so far may not be easy to implement in real situations where the arrival curves are not deterministic but random processes. We will first examine the single capacitated off-ramp case (no on-ramps). The extension to multiple off-ramps and on-ramps will be straightforward and outlined towards the end of this section.

When the arrival of vehicles is unknown, the operator must guess the best moments to start and end diverting vehicles through each off-ramp. In this single off-ramp case the operator needs to determine when to start and end using the off-ramp (t_0 and T_1, respectively, see Fig. 5).

In the single off-ramp case, the ramp should only be used if the freeway queue is expected to last longer than Δ_1 units of time. Vehicles should start being diverted as soon as the arrival rate exceeds μ_0 (avoiding the queue in the freeway), that is t_0. Identifying the moment when the last vehicle diverted should leave the off-ramp (T_1 in this case) is less straightforward.

Let's assume that we have already a queue of length q_0 in the freeway and the off-ramp is working at capacity but no queue has grown on it (see Fig. 9). We want to decide if we should stop diverting now or later. To make this decision we will assume that the future arrival process responds to some distribution of arrival curves Z and we will call one realization of that distribution A_ζ. Fig. 10 represents the case when vehicles will arrive according with A_ζ and we decide to stop diverting at $t = 0$. Notice that in this case the queue will vanish at $t = T_\zeta$ (before at $t = \Delta_1$). Therefore, we should have stopped diverting vehicles earlier.

If we stop diverting vehicles through the off-ramp ε times units later then

the future cost would be equal to the area (O, q_o, e, c) which is the total time in queue, plus the area (O, h, b, a): the future extra cost of sending $\mu_1\varepsilon$ vehicles through the off-ramp. This cost can also be expressed as the area $(O, q_o, f) + (a, b, g, f) - (O, h, c) + (e, f, g)$. Note that the last two quantities are of order ε^2 and therefore neglectable. Since T_ζ is the time when the queue vanishes given the realization $A_\zeta(t)$, we can say that the total cost is:

$$C(\varepsilon, \zeta) \approx \int_0^{T_\zeta} \left[A_\zeta(t) - \mu t\right] dt + \left[\Delta_1 - T_\zeta\right] \mu_1 \varepsilon \tag{3}$$

and clearly:

$$\frac{\partial C(\varepsilon, \zeta)}{\partial \varepsilon} = (\Delta_1 - T_\zeta)\mu_1 \tag{4}$$

Note that this corresponds to the shaded area in Fig. 10. Its expected value along all curves A_ζ in Z is:

$$E_\zeta \left[\frac{\partial C(\varepsilon, \zeta)}{\partial \varepsilon}\right] = (\Delta_1 - E_\zeta \left[T_\zeta\right])\mu_1 \tag{5}$$

By setting (5) equal to zero we see that, as expected, the optimal stopping time is such that the expected queue clearance time equals the extra free-flow travel time by the city streets, i.e.:

$$E_\zeta \left[T_\zeta\right] = \Delta_1 \tag{6}$$

If we assume that the queue is governed by a Brownian motion with negative drift $\lambda - \mu$ we would find that we should stop sending people by the off-ramp when the queue is:

$$q_o = \Delta_1(\mu - \lambda) \tag{7}$$

Note that (7) does not depend on the index of dispersion of the process. To be more general, we can consider the rate of the arrival process, Λ, as a random variable with mean λ and variance σ^2. Then the queue behaves as a conditional Brownian motion process because conditional on $\Lambda = \lambda$ the queue becomes a Brownian motion with negative drift $\lambda - \mu$. In this case (5) becomes:

$$E_{\zeta,\Lambda} \left[\frac{\partial C(\varepsilon, \zeta, \Lambda)}{\partial \varepsilon}\right] = (\Delta_1 - E_{\zeta,\Lambda} \left[T_{\zeta,\Lambda}\right])\mu_1$$

with

$$\begin{aligned} E_{\zeta,\Lambda}\left[T_{\zeta,\Lambda}\right] &= E_{\Lambda}\left[E_{\zeta}\left[T_{\zeta,\Lambda}|\Lambda\right]\right] \\ &= E_{\Lambda}\left[\frac{q_o}{\mu-\Lambda}\right] \end{aligned} \tag{8}$$

To compute (8) we can expand the term in brackets (call it $T(\Lambda)$) in a power series around $\Lambda = \lambda$. Thus

$$T(\Lambda) = T(\lambda) + (\Lambda-\lambda)T'(\lambda) + \frac{1}{2}(\Lambda-\lambda)^2 T''(\lambda) + \cdots \tag{9}$$

where $+\cdots$ represents higher order terms that may be neglected. Therefore:

$$\begin{aligned} E_{\Lambda}\left[\frac{q_o}{\mu-\Lambda}\right] &= E_{\Lambda}\left[T(\lambda) + (\Lambda-\lambda)T'(\lambda) + \frac{1}{2}(\Lambda-\lambda)^2 T''(\lambda) + \cdots\right] \\ &= T(\lambda) + \frac{1}{2}\sigma^2 T''(\lambda) + \cdots \\ &= \frac{q_o}{\mu-\lambda} + \sigma^2\frac{q_o}{(\mu-\lambda)^3} + \cdots \\ &= \frac{q_o}{\mu-\lambda}\left[1 + (\frac{\sigma}{\mu-\lambda})^2\right] + \cdots \end{aligned} \tag{10}$$

so that our optimality condition to determine when to stop diverting, (6), becomes:

$$q_o = \frac{\Delta_1(\mu-\lambda)}{1+(\frac{\sigma}{\mu-\lambda})^2} \tag{11}$$

which is smaller than the Brownian motion with deterministic drift found in (7).

Thus far in this section we have assumed that no queues develop at the off-ramp, i.e. at the time we stop diverting, the last diverted driver is being served by the off-ramp. Fortunately, in previous sections we observed that one optimal solution satisfied this condition. This is why at this point we can say that we prefer solutions with as little queue on the off-ramps as possible. This, of course, will be limited by the storage space of the freeway in order to avoid spillovers.

It is easy now to extrapolate this in the case of several (R) off-ramps: once we have our R off-ramps operating at capacity we stop diverting drivers

to the one most upstream (the R^{th}) when a condition analogous to (11) is satisfied. If we let $M_i = \sum_{k=1}^{i} \mu_k$ be the total off-ramp capacity when i off-ramps are operating, the analogous of (11) reads:

$$q_o^i = \frac{\Delta_i (M_{i-1} - \lambda)}{1 + (\frac{\sigma}{M_{i-1} - \lambda})^2} \tag{12}$$

Then we proceed sequentially until the first off-ramp is no longer needed. Note that at each stage we should have different (and hopefully better) estimates for the first two moments of the slope of $A(t)$.

7 Discussion

We have been able to identify the SO-DTA for a simple but commonplace network. Our approach seems to be more appealing for the evening commute problem or for incident management, since we have assumed that drivers can not change their departure time in accordance with the travel time they expect.

Although our assumption of no congestion on local streets is not very realistic, our results should still be helpful for practitioners. We have stated the periods when vehicles should be forced to divert at each upstream off-ramp and when on-ramp metering rates should be activated. If vehicles face congestion in the city streets, then the diverting period for an on-ramp should still start at the time suggested here but should end earlier. Then, since the Δ's would be larger, we expect that fewer ramps should be used for diversion.

When there is congestion on city streets the problem is more complicated since (i) external users are affected by diverting vehicles; and (ii) travel times on off-ramp routes would be affected by flows on other off-ramps. If (i) is not relevant, then an iterative approach between a conservative traffic flow model in the local streets and the methodology proposed in this paper is being explored.

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 come from non-metered on-ramps and there is no queue in the exit off-ramps (otherwise we would need to distinguish vehicles according to destination in on-ramps and off-ramps, respectively). Fortunately, the solution with no queue in 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 may be challenging. Clearly, SO solutions are not obtained spontaneously since they are not user optimum. Diverted vehicles are individually better off staying in the freeway and the simplest way to implementing these solutions seems to be enforcement. Unfortunately, SO tolls are hard to implement, mainly because of the discontinuities in the marginal cost on freeway routes in every t_i. Additionally, in alternative rationing systems based on license plates our approach would not work since at the on-ramps we would need to distinguish a subset of drivers from the rest. As is shown in Erera et al (2000), the problem is NP-hard.

Also, our results suggest that during the peak period the best thing to do is to shut down on-ramps close to the bottleneck. We understand that the authorities would not be willing to implement such a drastic policy. In this case, the lowest acceptable ramp metering rate should be deployed. Then in our problem we should subtract this maximum metering rate from the previous capacity of the off-ramp (previously on-ramp).

8 References

Al-Deek, H. (1993), "The Role of Advanced Traveler Information Systems in Incident Management," Dissertation Abstract, Transportation Science,Vol. 27 No. 2, May 1993

Daganzo, C.F. (1994), "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory", Trans. Res. 28B (4), 269-287.

De Palma,A. and Jehiel,P. (1995), "Queuing May Be First-Best Efficient", Papers 9520, Paris X - Nanterre, U.F.R. de Sc. Ec. Gest. Maths Infor.

Erera, A.L., Daganzo, C.F., Lovell, D.J., "The access control problem on capacitated FIFO networks with unique O-D paths is hard", forthcoming in Operations Research, presented at the 79th Annual Meeting of the Transportation Research Board (January, 2000).

Laval, J.A., Muñoz, J.C. (2002), "System Optimum Diversion of Congested Freeway Traffic", Institute of Transportation Studies Research Report UCB-ITS-RR-2002-06, University of California, Berkeley

Newell, G.F. (1980), "Traffic Flow on Transportation Networks", MIT Press, Cambridge, MA.

Ziliaskopoulos, A.K. (2000), "A linear programming model for the single destination system optimum dynamic traffic assignment problem," Transportation Science, Vol. 34, No.1, pp. 1-4.

Figure 1: The Network (Times correspond to extra free flow trip times).

Figure 2: Single off-ramp case (a) Optimal solution for a given N, (b) Marginal increments analysis to identify optimal N.

Figure 3: Single off-ramp solution for a single-peak demand case

Figure 4: Case of an arrival curve with multiple peaks

Figure 5: Optimal solutions for the case with a single capacitated off-ramp. (a) The freeway's queue starts as late as possible. (b) No queue grows in the off-ramp

Figure 6: Two on-ramp case.

Figure 7: General problem with many off-ramps, no on-ramps (a) Initial step to identify an optimal solution (b) Optimal solution

Figure 8: Solution for a general problem with many off-ramps and on-ramps

Figure 9: Global depiction of a solution with one upstream off-ramp

Figure 10: Sensitivity analysis for the end of diversion time