Title
System Optimum Diversion of Congested Freeway Traffic

Authors
Laval, Jorge A.
Munoz, Juan Carlos

Publication Date
2002-06-01

Institute of Transportation Studies
University of California at Berkeley

System Optimum Diversion of Congested Freeway Traffic

Jorge A. Laval and Juan Carlos Munoz
Department of Civil and Environmental Engineering
Transportation Group
University of California, Berkeley

RESEARCH REPORT
UCB-ITS-RR-2002-6

June 2002
(Originally completed in 2001)
ISSN 0192 4095

System Optimum Diversion of Congested Freeway Traffic

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

Abstract

We study the system optimum dynamic traffic assignment (SO-DTA) in a network consisting of a freeway and neighboring city streets. There is only one bottleneck in the freeway and every destination is somewhere downstream of the bottleneck. Vehicles can be diverted through off-ramps leading to alternative local street routes. We formulate the problem and determine a graphical solution procedure based on Newell's cumulative plots, which yields the optimal diverted flow over time. On-ramps can be conveniently incorporated in this procedure yielding SO metering rates. The following variants are considered: capacitated and uncapacitated off-ramps, and deterministic and stochastic demand.

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 this alternative 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 and therefore is worthy of study.

In this paper we study the system optimum (SO) flow pattern over a freeway section with a single bottleneck and several off-ramps and on-ramps upstream (see Fig. 1). We consider the many-to-one case where trips originate on any on-ramp and are headed to a common destination downstream of the bottleneck. However, vehicles can also reach their destination by taking any of the off-ramps and bypassing the bottleneck via local streets (but not getting back to the freeway upstream of the bottleneck). Off-ramps have a fixed capacity and we assume that the diverted vehicles encounter freely-flowing conditions on local streets. Although we also assumed that queues don't occupy space (point queues without spillovers) we identified solutions where the limitations of this assumption were minimized.

Our goal is to determine which vehicles should stay on the freeway and which ones should take each of the upstream off-ramps and on-ramps so that the total time spent in the system is minimized. Initially, we assume that every demand curve is known, but later we explore the case where future demand is uncertain.

We show that for this type of network the 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. We utilize the optimality conditions from Ziliaskopoulos (2000), i.e. at every instant in time, the route with the least marginal cost should be used. The optimality condition is easy to implement since it uniquely determines the beginning and the end of the diverting period for each ramp.

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 one optimal SO solution (and the most appealing one) consists of diverting the traffic that the city streets can handle, i.e. no queues on the off-ramps.

In Newell (1980) one can find 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.

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 con-

ditions for this problem. Compared to the methods presented in this paper, Ziliaskopoulos' approach can handle more complicated networks. However, the number of variables involved in the model is proportional to the product of the number of cells in the network, the number of time steps and the number of origins. For moderate size problems involving a few miles and an acceptable time step (say, two seconds) the problem would become 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 hard to implement. The method proposed in this paper allows us to solve our problem with pencil and paper 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 is formalized using mathematical programming. This general problem is then solved in sections 5 and 6. In all the cases the arrival process is known a priori, except for section 7 where a stochastic arrival process is considered. In section 8 the results are discussed, analyzing practical implementations, and suggesting future developments.

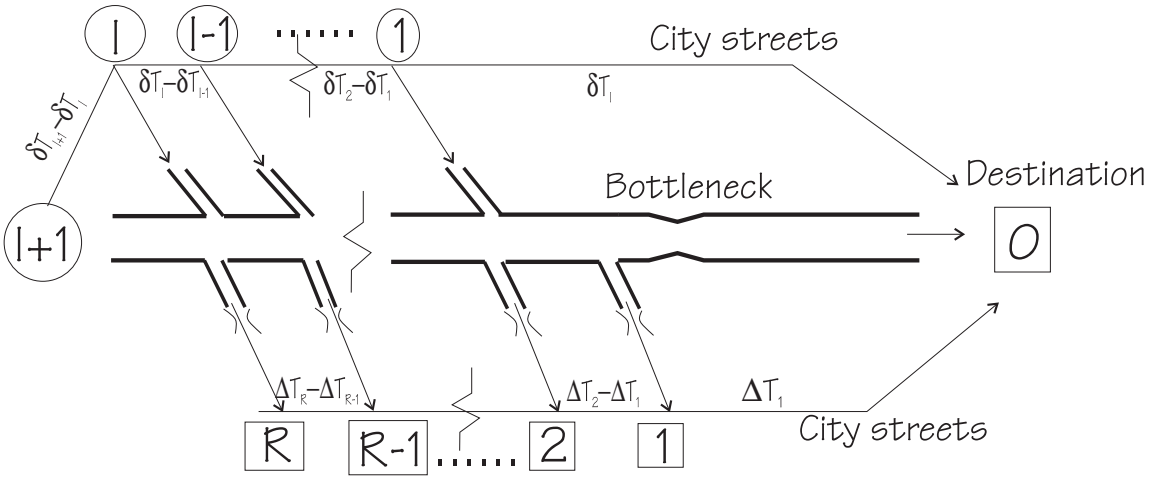

Figure 1: The Network. Times correspond to extra free flow trip times.

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 beginning and end of the freeway are denoted on-ramp $I+1$ and off-ramp 0, respectively. Off-ramp 1 is the closest to and R the furthest from the destination. Analogously, on-ramp 1 is the closest to and I the furthest from the destination. 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 to reach their final destination by driving along local streets. We assume that once a vehicle has exited the freeway it does not get back on. The capacity of off-ramp r is μ_r and is dictated by its discharge capacity to the city streets. On-ramp's capacity is assumed unlimited (or never reached). The freeway free-flow travel time between on-ramp $I+1$ and on-ramp i is called f_i, and the travel time between on-ramp $I+1$ and the freeway's bottleneck is called t_f. The trip time from this bottleneck to the destination is assumed fixed and constant for all travelers. Since we assume no congestion in the city streets, if a vehicle takes off-ramp r, it faces a fixed *extra* trip time of $\Delta T_r; \Delta T_r \geq \Delta T_{r-1}; \Delta T_0 = 0$. 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 $\delta T_i; \delta T_i \leq \delta T_{i+1}, i \in [1, I]$.

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

Later in the paper we formulate this problem using mathematical programming and provide the structure of an optimal solution and an algorithm to identify this solution. First, we state one necessary condition that all system optimum solutions must satisfy:

Condition: According with Ziliaskopoulos (2000), in SO-DTA drivers faced with a route choice must always choose the one with lower marginal cost. 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.

There are only two types of choices: i) enter an on-ramp or stay on local streets, and ii) take an off-ramp or stay in the freeway. Therefore, we must check for the condition at only the $I + R + 1$ decision points in the network.

Corollary 1: 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.

Corollary 2: For a single peak period, according to corollary 1, off-ramps should be activated in ascending order and deactivated in descending order.

2.2 Marginal cost analysis

Marginal cost analysis can be a powerful tool to identify system optimum conditions. The marginal cost of a route is the extra total cost if an additional driver is added to that route while flows on all other routes are known and remain unchanged. Thus, marginal costs are a characteristic of future network assignments and change with time. For each origin we define its system marginal cost as the minimum marginal cost of all routes emanating from this origin.

Let's now compare the marginal costs for the following two routes: exiting at off-ramp 1 (and never enter the freeway again), and staying on the freeway until the destination is reached. Fig. 2 shows the marginal cost versus time for the two alternatives and the system in three different scenarios. Fig. 2(a) shows the case where no vehicles are diverted to the off-ramp, thus the system and freeway marginal costs coincide. In this figure, t_0 and T_0 denote, respectively, the times when the first and last vehicle experiencing a queue in the freeway enter the system. Thus, the freeway queue will last for $T_0 - t_0$ units of time [2]. The marginal cost of the local streets route is constant and equal to the trip time through the route (because we assume no congestion there). In the freeway route, the marginal cost is lower for vehicles not involved in the queue (before t_0 and after T_0) and therefore the off-ramp should

[2]Notice that t_0 and T_0 will depend on how downstream off-ramps and on-ramps are used. However, the results derived here will be of value for later analysis

not be used during those periods. However, between t_0 and T_0 vehicles suffer a delay and cause an externality to other vehicles. Therefore the marginal cost on that period is higher.

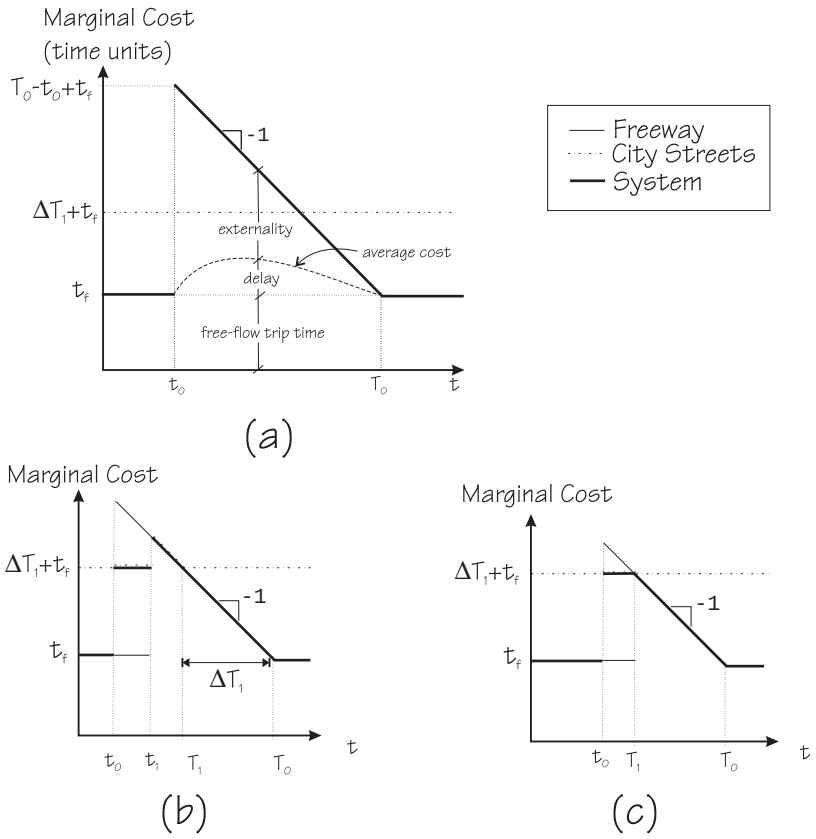

Figure 2: Marginal costs of two routes: taking off-ramp 1, and staying in the freeway. (a) No vehicle takes the off-ramp. (b) Optimal solution when off-ramp 1's capacity is reached. (c) Optimal solution when off-ramp 1's capacity is never reached.

If we call t_a and t_b the moments when a driver enters the system and passes through the bottleneck respectively, then the delay for the vehicle is $t_b - t_a - t_f$ and the externality is equal to $T_0 - t_b$. This is true since all vehicles to be queued coming behind $(\mu_0(T_0 - t_b))$ would arrive to the destination $1/\mu_0$ units of time earlier had our vehicle taken off-ramp r. The marginal cost is equal to $T_0 - t_a$ since it is the sum of the cost experienced by the driver and the externality. Thus, the marginal cost decreases with time at a slope of

-1 and presents a discontinuity at t_0. We have also highlighted the three components of the marginal cost: free-flow trip times, delay and externality. Notice that the average cost is the sum of the first two components. The externality decreases in time, is maximum at t_0 and zero at T_0.

In the case of Fig. 2(a) our solution would be improved if vehicles where to be diverted after t_0 (since the marginal cost through the freeway is higher than through the streets at that time). If vehicles are diverted and the capacity of the off-ramp is reached (let say at t_1), the marginal cost of the off-ramp route would then take a similar shape than the freeway route's marginal cost (but presenting the discontinuity at t_1 and then dropping at rate -1 until the uncongested marginal cost is reached again, let's say at T_1). In the optimal solution both marginal costs should be equal once both routes have reached capacity (between t_1 and T_1)[3]. Fig. 2(b) depicts this solution. This figure shows that $T_0 - T_1 = \Delta T_1$) indicating and optimal off-ramp management policy. This suggests that the last vehicle to exit the off-ramp and the last vehicle to be trapped in the freeway queue will reach the destination simultaneously.

This argument helps to understand the structure of our solution. In the above example we examined the local optimality conditions relating two paths. However this approach turns quite complicated when many paths are involved. In what follows we will develop a graphic approach that will allow us to identify all optimal solutions in just one graph. The above analysis will be used as a complement and a reinforcement. In the next section we will illustrate this graphical approach with the simplest problem we can think of.

3 Uncongested Off-ramps, no on-ramps

In this section we examine the case when upstream off-ramps 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 $R = 1$. For now, we will not consider on-ramps in the freeway, i.e. I=0.

Let's define the following counting processes:
$A(t)$ = cumulative number of vehicles that have entered the system by time t

[3]This indicates that there may be multiple optimal assignments during this period since both routes have identical marginal costs

$a_r(t)$ = cumulative number of vehicles that have entered the system by time t and will take off-ramp r, $r \in [0,1]$
$d_r(t)$ = cumulative number of vehicles that have reached the destination through off-ramp r by time t, $r \in [0,1]$.

Here and elsewhere in the paper, cumulative curves are started after the passage of a common reference vehicle.

3.1 Single-peak demand

Consider the case of Fig. 3(a) 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 first time when the slope of $A(t)$ exceeds μ_0. 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 $t_0 + t_f$), $d_0(t)$ grows linearly at a rate μ_0. Then, if we assume N as given we can identify T_0 as the moment when the queue clears in the freeway. That is the last moment such that $A(t - t_f) - d_0(t)$ is equal to N. We call $d_0(T_0) = N_0$. If N is fixed then the total time spent in the off-ramp route is constant (recall that it 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 $(T_0 - t_f, N_0)$. Thus, we draw $a_0(t)$ starting at $(T_0 - t_f, N_0)$ 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 $a_0(t) = max\{d_0(t + t_f), A(t) - N\}$. We will call T_1 the time at which the last vehicle diverted enters the off-ramp (i.e. when $d_0(t + t_f) = A(t) - N$ 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 $a_1(t - t_f - \Delta T_1)$. In Fig. 3(a) we show all these curves and highlight the three components of the marginal cost of a trip.

Notice how all this procedure simplifies if we shift $d_r(t), r \in [0,1]$ horizontally to the left by their respective free flow trip times as is shown in Fig. 3(b). This is equivalent to start all the clocks with the passage of the reference vehicle or to assume $v_f = \infty$. Let's identify the new curves with capital letters: $A_r(t)$ and $D_r(t), r \in [0,1]$. Now $A(t) = A_0(t) + A_1(t)$ and the

area between $A_0(t)$ and $D_0(t)$ represents the total delay suffered by drivers in the freeway. Analogously we will define $D(t) = D_0(t) + D_1(t)$. From now on we will refer to events as they happen in this shifted diagram (Figs. 3(b)).

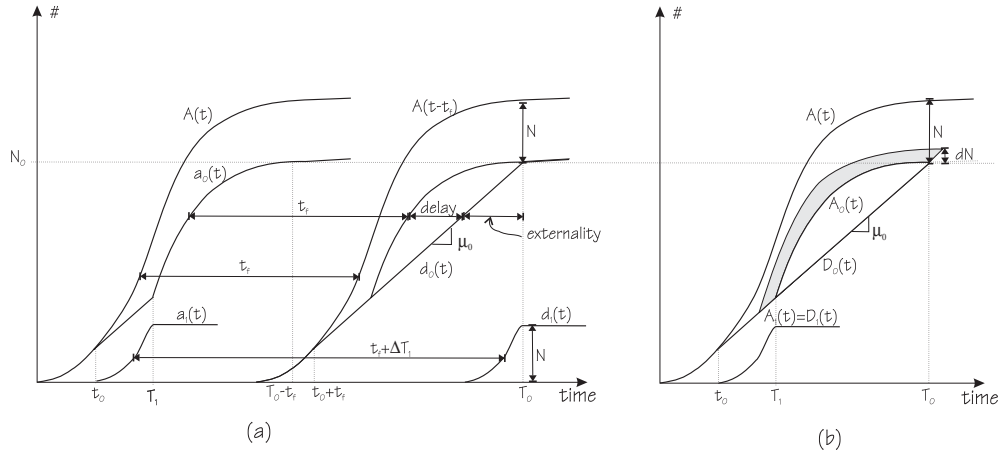

Figure 3: (a) Optimal solution for the single off-ramp case using cumulative plots. (b) Same solution decomposed by route after a time shift for all the cumulative arrival to destination curves.

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. 3) and of $-dN\Delta T_1$ on the off-ramp delay. For N to be optimal the sum of both quantities should be zero.

Thus, the optimality condition is simply:

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

This means that the duration of the queued episode on the freeway must equal the extra travel time using city streets. This result is equivalent to the result obtained in the previous section but assuming that the marginal cost of the off-ramp route is constant independent on its flow. The marginal cost diagram for this case is shown in Fig. 2(c). 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 ΔT_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 ΔT_1, as shown in Fig. 4.

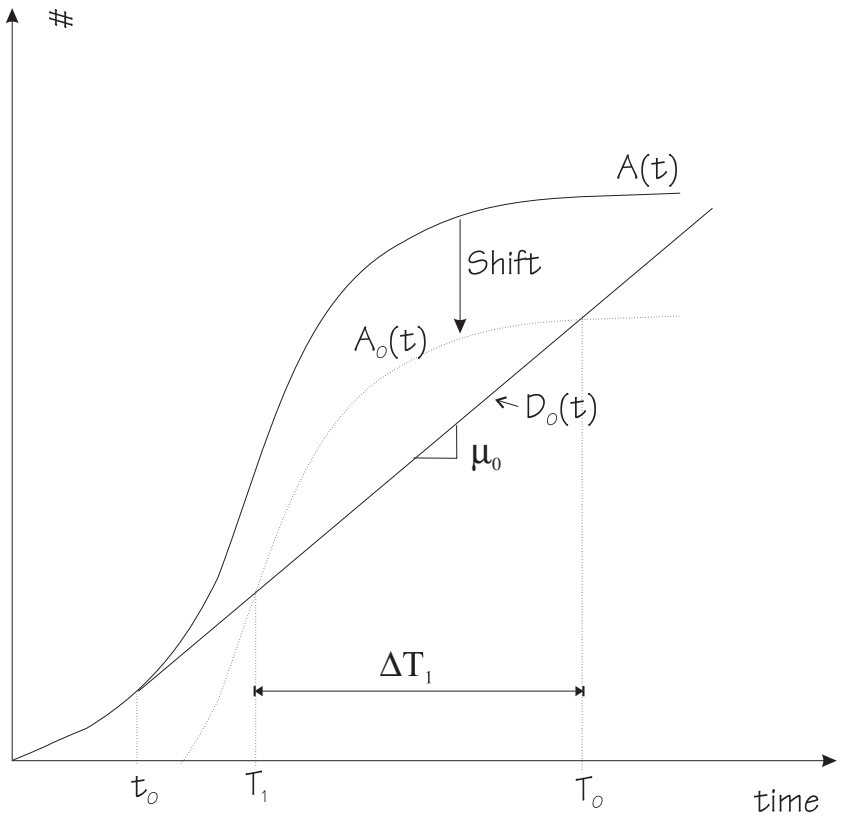

Figure 4: Single off-ramp solution for a single-peak demand case

3.2 Multiple-peaks 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. 5 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 ($\tau_3 - \tau_1 > \Delta T_1$). 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 at τ_2.

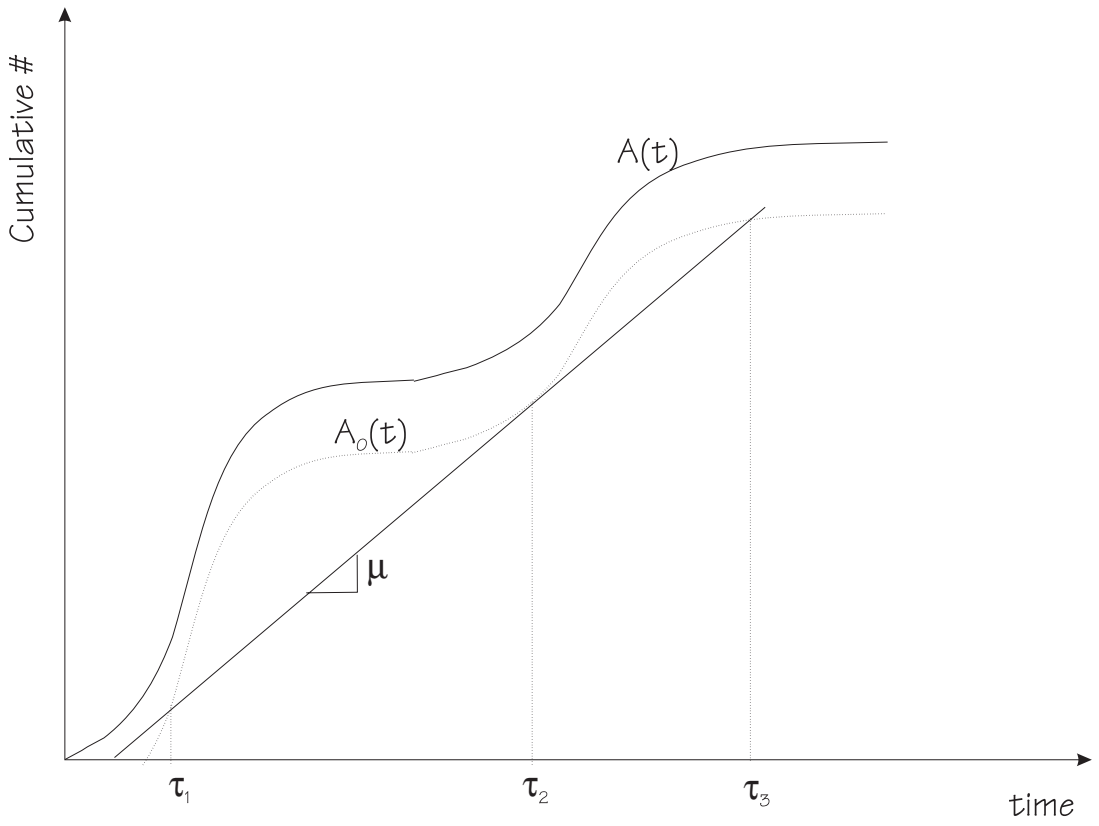

Figure 5: Case of an arrival curve with multiple peaks

1. $\tau_2 - \tau_1 < \Delta T_1$ and $\tau_3 - \tau_2 < \Delta T_1$: Then N_s corresponds to the optimal number of drivers to divert since further shifting would end in queues shorter than ΔT_1.

2. $\tau_2 - \tau_1 > \Delta T_1$ and $\tau_3 - \tau_2 < \Delta T_1$: 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 ΔT_1 in time; for the interval $[\tau_2, \tau_3]$ no further shifting is necessary.

3. $\tau_2 - \tau_1 < \Delta T_1$ and $\tau_3 - \tau_2 > \Delta T_1$: analogous to case 2.

4. $\tau_2 - \tau_1 > \Delta T_1$ and $\tau_3 - \tau_2 > \Delta T_1$: Both intervals can be shifted

further independently until the single-peak SO solution is found for each interval.

Assuming that off-ramps have an infinite capacity is certainly an unrealistic assumption. However, it's solution provides valuable insights for the more realistic case of finite capacities analyzed later in this paper. Now that the reader has been familiarized with the problem and our approach, we will provide a mathematical formulation for the general case. This step will provide new insights about the optimal solution for our problem.

4 Formulation

In this section we formulate the general problem with I on-ramps and R capacitated off-ramps. The nomenclature and formulation will be important for attempting this general problem later on.

4.1 Definitions

Let the following quantities be the cumulative number of vehicles that, by time t, have:
$a^i(t)$ = entered the system through on-ramp i, $\forall i \in [0 \dots I]$
$a_r^i(t)$ = entered the system through on-ramp i that will exit through off-ramp r, $\forall i \in [0 \dots I], r \in [0 \dots R]; a^i(t) = \sum_{r=0}^{R} a_r^i(t)$. If off-ramp r is upstream from on-ramp i then $a_r^i(t) = 0$.
$d_r^i(t)$ = reached the destination 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 $d_r^i(t) = 0$.
$d(t)$ = reached the destination ($d(t) = \sum_{i=0}^{I} \sum_{r=0}^{R} d_r^i(t)$).

We assume that $a^i(t)$ are given while all other curves need to be determined. Notice that the horizontal separation between $a_r^i(t)$ and $d_r^i(t)$ reflects the travel time (free-flow plus delay) of a vehicle over the path connecting on-ramp i and the destination through off-ramp r. Notice as well that vehicles departing from an on-ramp at the same time might take different paths and therefore arrive at the destination at different moments.

Although the problem can be formulated using the above notation, its formulation and its graphical solution are considerably simpler if we shift all

curves to the left by the freeway free flow trip time between on-ramp 0 and the place where each curve is counted (as was done in the previous section). Thus, we do the following linear transformation:
$A^i(t) = a^i(t + f_i), \forall i \in [0 \dots I]$
$A^i_r(t) = a^i_r(t + f_i), \forall i \in [0 \dots I], r \in [0 \dots R]$
$D^i_r(t) = d^i_r(t + t_f + \Delta T_r), \forall i \in [0 \dots I], r \in [0 \dots R]$
$D(t) = \sum_{i=0}^{I} \sum_{r=0}^{R} D^i_r(t)$

In the case of $A^i_r(t)$ and $A^i(t)$ the magnitude of the shift corresponds to the free flow travel time between on-ramp 0 and the respective on-ramp while in the case of $D^i_r(t)$, it corresponds to the free flow travel time between on-ramp 0 and off-ramp 0 through ramp r. Notice that now the horizontal separation between $A^i_r(t)$ and $D^i_r(t)$ corresponds only to the delay of a vehicle (excludes the free-flow travel time).

Erera et al (2000) showed that identifying the optimal ramp metering in a general network was an NP-hard problem because choosing the order in which vehicles should depart from the on-ramps (depending on the path they would follow) forced a combinatorial problem. However, since our network has a single destination, there is no need to break the FIFO rule in the on-ramps; e.g. all the vehicles are identical for our purposes. Therefore, we can formulate our problem as a mathematical programming problem:

$$\min \sum_{i=0}^{I} \sum_{r=0}^{R} \int_{t=0}^{T} [A^i_r(t) - D^i_r(t)]dt + \sum_{i=0}^{I} \sum_{r=0}^{R} A^i_r(T)\Delta T_r \tag{2a}$$

subject to

$$D^i_r(t) \leq A^i_r(t) \qquad \forall r \in [0 \dots R], \forall i \in [0 \dots I], \forall t \in [0 \dots T] \tag{2b}$$

$$\sum_{i=0}^{I} \dot{D}^i_r(t) \leq \mu_r \qquad \forall r \in [0 \dots R], \forall t \in [0 \dots T] \tag{2c}$$

$$\sum_{r=0}^{R} A^i_r(t) = A^i(t) \qquad \forall i \in [0 \dots I], \forall t \in [0 \dots T] \tag{2d}$$

$$\dot{A}^i_r(t), \dot{D}^i_r(t) \geq 0 \qquad \forall i \in [0 \dots I], \forall r \in [1 \dots R], \forall t \in [0 \dots T] \tag{2e}$$

The objective function corresponds to the total delay due to congestion and diversion. Its first term function corresponds to the extra time spent

by vehicles because their path is congested. The second term corresponds to the extra free-flow time spent by vehicles that are diverted. Constraints (2b) and (2c) ensure that vehicles are served neither before they arrive nor faster than the bottlenecks' capacities while constraints (2d) ensures that all vehicles are served. Finally, the derivative with respect to time (denoted by a dot in the top) of arrival and departure curves must be non-negative[4].

Notice that our problem is linear and therefore simple to solve using mathematical programming tools after discretizing. However, in this paper we will derive the optimal solution graphically to obtain insights about its structure.

5 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, $r = 0 \dots R$
T_r = Time when the last driver diverted to off-ramp r leaves the off-ramp, $r = 0 \dots R$
t_r = The first time when the slope of $A(t)$ exceeds $\sum_{j=0}^{r} \mu_j$, $r = 1 \dots R$.

Let's consider first the $R = 1$ case.

5.1 Single off-ramp

We will first derive the optimality conditions for a freeway with only one off-ramp upstream from the bottleneck. If we assume that N vehicles are diverted to off-ramp 1 during the whole period, then our objective is to minimize the total off-ramps delays (the extra trip time along local streets would be fixed), that is, the area between $A(t)$ and $D(t)$. According to corollary 1 if a queue grows in both off-ramps the queue in off-ramp 1 will vanish before the queue in off-ramp 0. 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 bottleneck works at capacity until the queue vanishes.

[4]Notice that derivatives are linear functions.

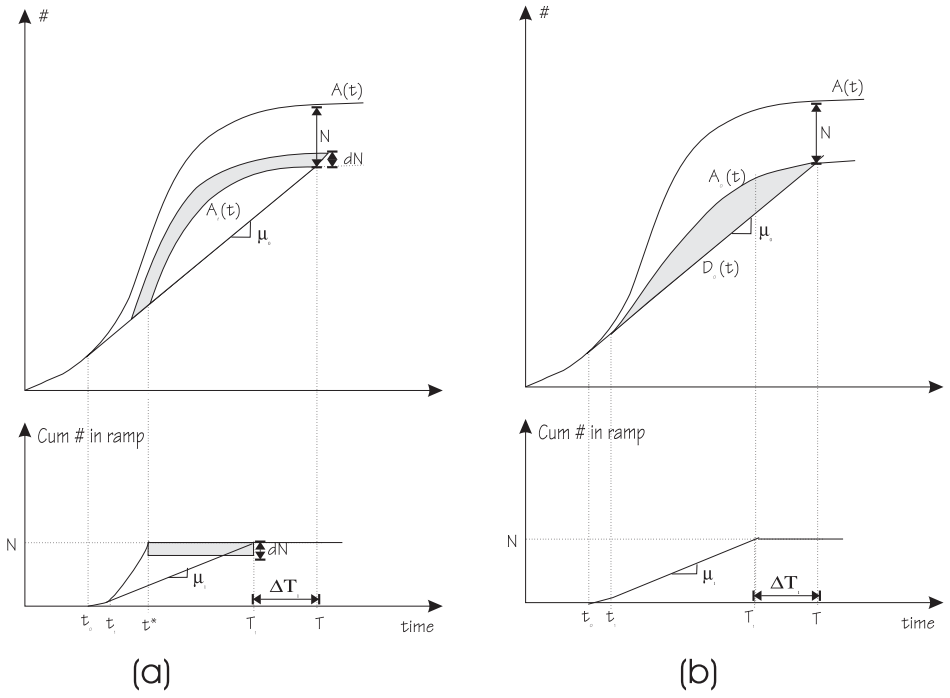

Figure 6: Optimal solutions for the case with a single capacitated off-ramp. (a) The freeway's queue starts as late as possible. (b) No queue grows in the off-ramp

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 correspondent arrival and departure curve at the off-ramp.

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

$$T_0 = T_1 + \Delta T_1 \tag{3}$$

This means that the queue in the freeway should clear ΔT_1 time units after it vanishes on the off-ramp path which is consistent with our analysis

in previous sections. Note that the solution does not define how should we allocate vehicles arriving between t_1 and T_1. It only requests that during that period both servers should work at capacity. We might conclude that the optimal solution is not unique. Indeed, Figs. 6(a) and 6(b) represent two extreme optimal solutions. In Fig. 6(a), the freeway has been kept as empty as possible, while in Fig. 6(b) no queues were allowed to grow in the ramp. Notice that since the problem was shown to be linear, any linear combination of these two solutions is also optimal.

5.2 Multiple off-ramps

In the case of several upstream off-ramps, the optimum value of N must be such that a small perturbation dN will produce a delay variation on the freeway equal to the sum of the induced delays on each ramp plus their additional free flow delay, regardless of what proportion of dN is assigned to each ramp. Let's assume that a vehicle is shifted from the freeway to the off-ramps at time t^*. If we consider the case of Fig. 7 where $R = 2$ and we let $\alpha(t), t \geq t^*$ be the proportion of dN assigned to ramp 1 at instant t, we have that:

$$\begin{aligned} dN(T_0 - t^*) = &\int_{t^*}^{T_1} \alpha(t) dN dt + dN\alpha(T_1)\Delta T_1 + \\ &+ \int_{t^*}^{T_2} (1 - \alpha(t)) dN dt + dN(1 - \alpha(T_2))\Delta T_2 \end{aligned} \tag{4}$$

where t^* is the first moment when a queue grows in the freeway. According with corollary 1, $T_2 \leq T_1$, thus $\alpha(T_2) = \alpha(t) \forall t \in [T_2, T_1]$. Then,

$$T_0 - t^* = T_2 - t^* + \int_{T_2}^{T_1} \alpha(t) dt + \alpha(T_1)\Delta T_1 + (1 - \alpha(T_1))\Delta T_2 \tag{5}$$

$$T_0 = T_2 + \alpha(T_1)(T_1 - T_2 + \Delta T_1 - \Delta T_2) + \Delta T_2 \tag{6}$$

Notice that for the special case where $T_2 < t^* < T_1$ we get equation 3. Thus:

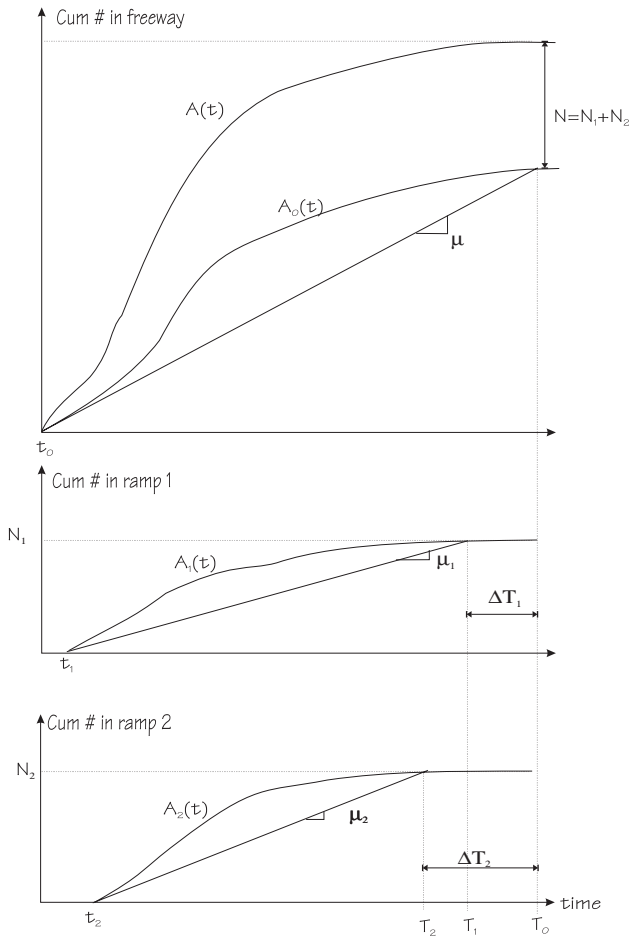

Figure 7: General solution with two upstream off-ramps

$$T_1 - T_2 + \Delta T_1 - \Delta T_2 = \alpha(T_1)(T_1 - T_2 + \Delta T_1 - \Delta T_2) \tag{7}$$

and our additional optimality condition is $T_1 = T_2 + \Delta T_2 - \Delta T_1$. Therefore, the optimality conditions are independent of the partition $\alpha(t)$.

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

1. If a 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:

(a) $\dot{A}_r(t) = \dot{D}_r(t) = \dot{A}(t) - \sum_{j=0}^{r-1} \mu_j$, $\forall t \in [t_{r-1}, t_r]$, $\forall r$.

(b) $\dot{D}_r(t) = \mu_r$, $\forall t \in [t_r, T_r]$, $\forall r$.

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. $\dot{A}_r(t) \geq 0$, $\forall t$, $\forall r$

5. After the queue on off-ramp r is cleared, no on-coming vehicle will take it: $\dot{A}_r(t) = \dot{D}_r(t) = 0$, $\forall t > T_r$, $\forall r$

6. The queue in ramp r ends $\Delta T_r - \Delta T_{r-1}$ time units earlier than in $r-1$: $T_r = T_{r-1} - (\Delta T_r - \Delta T_{r-1})$, $\forall r$

Graphically, the solution is quite intuitive. Let's assume that we know N (and therefore T_0). It follows that $D(t)$ goes through the point $(T_0, A(T_0))$ allowing us to draw $D(t)$ backward in time. If off-ramp 1 is used, then the slope of $D(t)$ is μ_0 in the interval $[T_0 - \Delta T_1, T_0]$. If off-ramp 2 is used, then the slope of $D(t)$ is $\mu_0 + \mu_1$ in the interval $[T_0 - \Delta T_2, T_0 - \Delta T_1]$ and so on. Therefore, $D(t)$ will be piece-wise linear with at most p pieces. Fig. 8(b) provides an illustration of the shape of $D(t)$. 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:

$$\begin{aligned} \tau_i &= \Delta T_p - \Delta T_{p-i} \\ b_i &= \sum_{j=0}^{p-i} \mu_j \\ a_i &= \sum_{k=1}^{i} (\tau_k - \tau_{k-1}) \\ &= a_{i-1} + (\tau_i - \tau_{i-1}) b_i \end{aligned}$$

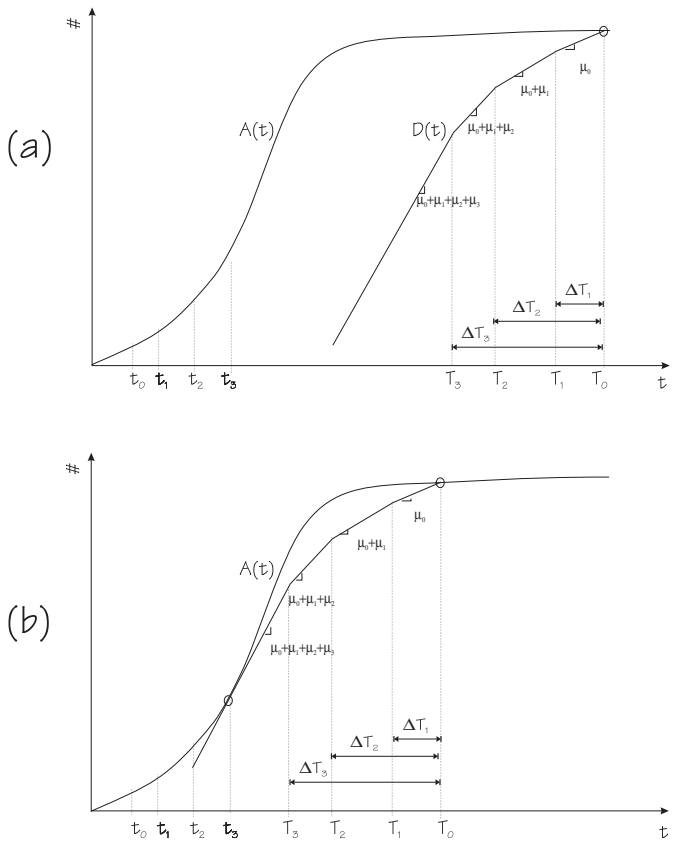

Figure 8: General problem with many off-ramps, no on-ramps (a) Initial step to identify an optimal solution (b) Optimal solution

with the border conditions $\tau_0 = 0, a_0 = 0$.

Then the game is to define a very late T_0 and build $D(t)$ from there[5]. A good choice of T_0 is the time when the queue would clear without diversion. Fig. 8(a) shows a good example of an initial T_0. Then we should reduce T_0 and move $D(t)$ along $A(t)$ until $D(t)$ first touches $A(t)$. Let's call this time $t = \tau$.

Notice that we may get two types of intersection points τ: a) either a corner of $D(t)$ such that the slope of $A(t)$ at $t = \tau$ is in between the slope of

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

$D(t)$ 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. 8(a) is shown in Fig. 8(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 a lot of flexibility as to where to send the diverted traffic, which is extremely useful when we have finite storage space.

6 Off-ramps and On-ramps

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

6.1 No On-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 δT_1. As soon as the queue in the freeway will last shorter than δT_1, diversion should be stopped. Therefore, this problem is equivalent to the single uncongested off-ramp case seen in section 3.1.

6.2 Single On-ramp

If $I = 1, R = 0$ then we can solve the problem using the tools developed in the previous section after we apply two simple modeling tricks. First we assume that all vehicles arriving at on-ramp 1 arrive to on-ramp 2 instead but 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 5.1. The procedure to determine $D_0(t)$ and the sensitivity analysis to deduce (3) are still valid. However, the ramp will now start working at capacity from t_1 which now corresponds to 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 $[t_1, T_1]$.

6.3 Multiple On-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 while their arrival curves are shifted to on-ramp $I + 1$ by their respective free-flow time. Now, the procedure outlined at the end of section 5.2 can be applied to this problem where $D(t)$ would still have pieces but no longer linear. As before, if we count the pieces starting from the later one, piece i would be $\delta T_i - \delta T_{i-1}$ units of time long ($\delta T_0 = 0$), 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. 9 as an illustration of an optimal solution.

6.4 Multiple On-ramps, multiple off-ramps

In the previous case, we model on-ramps as off-ramps. Now we just add more off-ramps to that model. Therefore, the slope of the pieces of $D(t)$ will be the sum of some off-ramps capacities and some on-ramps arrival rates. Notice that since we activate first the closest off-ramp (or on-ramp) from the bottleneck to then sequentially move to those upstream, the solution obtained will still be feasible since vehicles arriving to an on-ramp will never be exited 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 5.1. In this way the delays on the on-ramps are taken into account. The resulting optimal

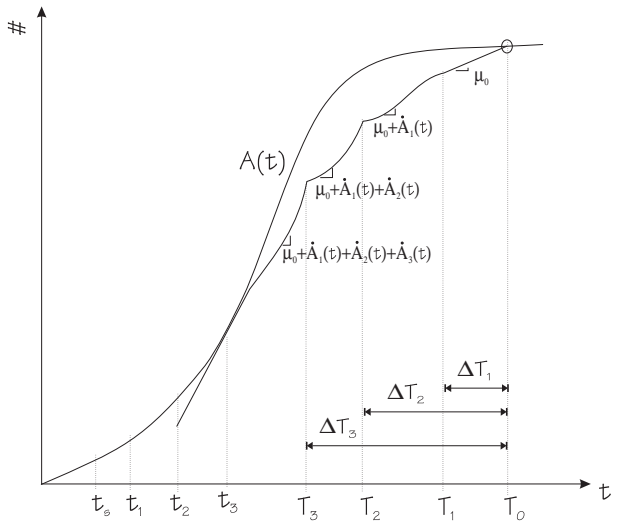

Figure 9: Solution for a general problem with many off-ramps and on-ramps

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, i.e., 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.

7 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 game of the operator is to guess when is the best moment 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. 6).

The off-ramp should only be used if the freeway queue is expected to last longer than ΔT_1 units of time. In this case vehicles should start being

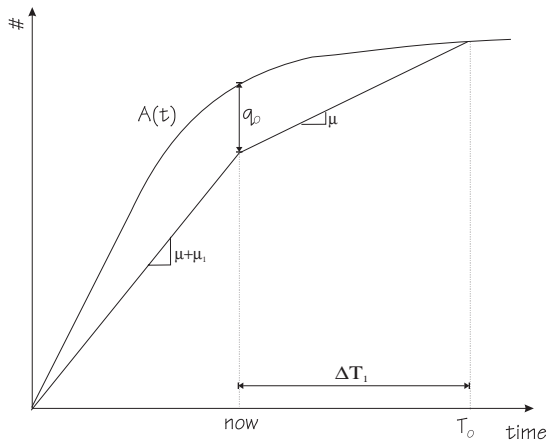

Figure 10: Global depiction of a solution with one upstream off-ramp

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. 10). We want to decide if we should stop diverting now or later. To take 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. 11 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 $t = \Delta T_1$), 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:

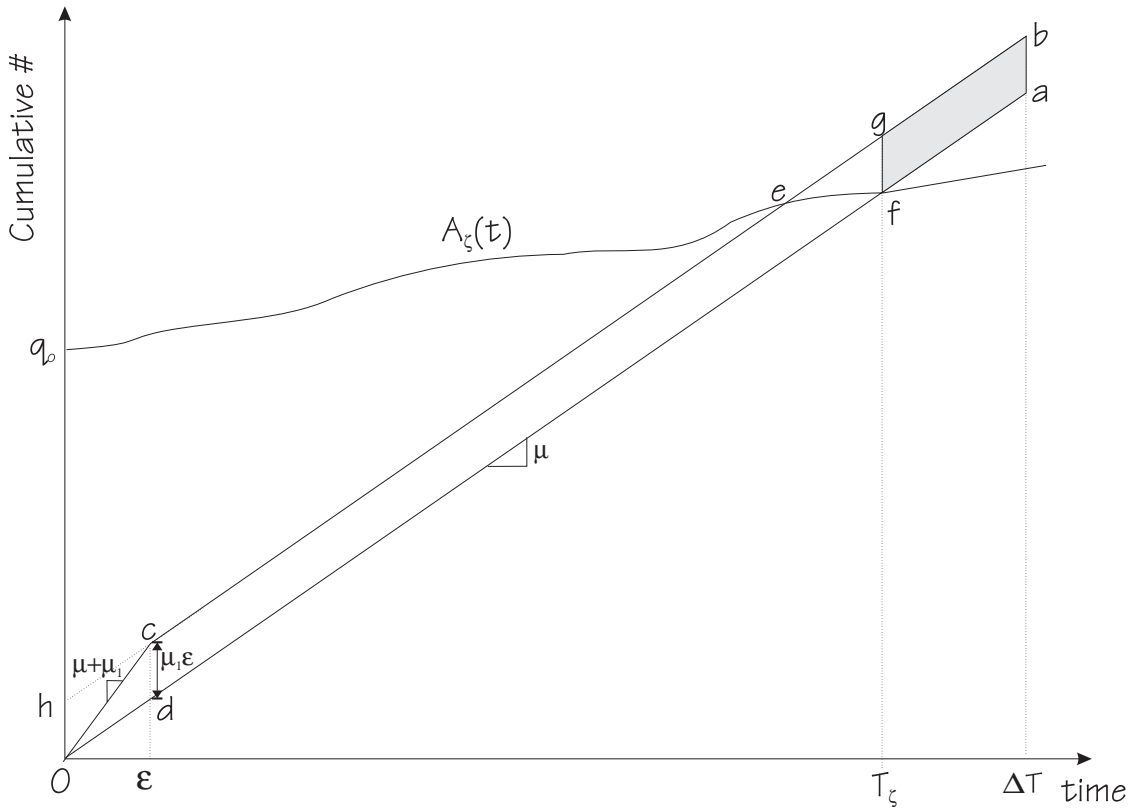

Figure 11: Sensitivity analysis for the end of diversion time

$$C(\varepsilon,\zeta) \approx \int_0^{T_\zeta} [A_\zeta(t) - \mu t]\, dt + [\Delta T_1 - T_\zeta]\, \mu_1 \varepsilon \tag{13}$$

and clearly:

$$\frac{\partial C(\varepsilon,\zeta)}{\partial \varepsilon} = (\Delta T_1 - T_\zeta)\mu_1 \tag{14}$$

Note that this corresponds to the shaded area in Fig. 11. Its expected value along all curves A_ζ in Z is:

$$E_\zeta\left[\frac{\partial C(\varepsilon,\zeta)}{\partial \varepsilon}\right] = (\Delta T_1 - E_\zeta\,[T_\zeta])\mu_1 \tag{15}$$

By setting (15) 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.:

$$E_\zeta [T_\zeta] = \Delta T_1 \tag{16}$$

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:

$$q_o = \Delta T_1 (\mu - \lambda) \tag{17}$$

Note that (17) 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 (15) becomes:

$$E_{\zeta,\Lambda}\left[\frac{\partial C(\varepsilon,\zeta,\Lambda)}{\partial \varepsilon}\right] = (\Delta T_1 - E_{\zeta,\Lambda}[T_{\zeta,\Lambda}])\mu_1$$

with

$$\begin{aligned} E_{\zeta,\Lambda}[T_{\zeta,\Lambda}] &= E_\Lambda [E_\zeta [T_{\zeta,\Lambda}|\Lambda]] \\ &= E_\Lambda \left[\frac{q_o}{\mu - \Lambda}\right] \end{aligned} \tag{18}$$

To compute (18) we can expand the term in brackets (call it $T(\Lambda)$) in a power series around $\Lambda = \lambda$. Thus

$$T(\Lambda) = T(\lambda) + (\Lambda - \lambda)T'(\lambda) + \frac{1}{2}(\Lambda - \lambda)^2 T''(\lambda) + \cdots \tag{19}$$

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

$$
\begin{aligned}
E_\Lambda\left[\frac{q_o}{\mu-\Lambda}\right] &= E_\Lambda\left[T(\lambda)+(\Lambda-\lambda)T'(\lambda)+\frac{1}{2}(\Lambda-\lambda)^2T''(\lambda)+\cdots\right] \\
&= T(\lambda)+\frac{1}{2}\sigma^2T''(\lambda)+\cdots \\
&= \frac{q_o}{\mu-\lambda}+\sigma^2\frac{q_o}{(\mu-\lambda)^3}+\cdots \\
&= \frac{q_o}{\mu-\lambda}\left[1+(\frac{\sigma}{\mu-\lambda})^2\right]+\cdots \qquad (20)
\end{aligned}
$$

so that our optimality condition to determine when to stop diverting, (16), becomes:

$$
q_o=\frac{\Delta T_1(\mu-\lambda)}{1+(\frac{\sigma}{\mu-\lambda})^2} \qquad (21)
$$

which is smaller than the Brownian motion with deterministic drift found in (17).

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 (21) 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 (21) reads:

$$
q_o^i=\frac{\Delta T_i(M_{i-1}-\lambda)}{1+(\frac{\sigma}{M_{i-1}-\lambda})^2} \qquad (22)
$$

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 $A(t)$.

8 Discussion

We have been able to identify the SO-DTA for a simple but commonly arising network. Our approach seems to be more appealing for the evening commute problem or an incidents management policy since we have assumed that drivers can not change their departure time according with the travel time they expect.

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

When there is congestion on city streets the problem is more complicated since then 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 an assignment model in the local streets and the methodology proposed in this paper could be explored.

Our model also assumes no flows 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 a great challenge. 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 discontinuous 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 and as is shown in Erera et al (2000) the problem turns to be NP-hard.

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

We have developed a general formulation for *many* on-ramps and off-ramps. However, since often the capacity upstream of the bottleneck is not dramatically greater than the capacity of the bottleneck, we shouldn't expect to use many off-ramps except in incident situations.

9 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.

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).

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.