Available online at www.sciencedirect.com
ScienceDirect
Transportation Research Part C 14 (2006) 244–262
TRANSPORTATION RESEARCH PART C
www.elsevier.com/locate/trc

Optimal freeway ramp metering using the asymmetric cell transmission model

Gabriel Gomes [a,*], Roberto Horowitz [b]

[a] *California PATH, 1357 S. 46th Street, Bldg. 452, Richmond, CA 94804-4648, USA*
[b] *Department of Mechanical Engineering, University of California at Berkeley, 5138 Etcheverry Hall, Mailstop 1740, Berkeley, CA 94720-1740, USA*

Received 27 January 2006; received in revised form 1 August 2006; accepted 2 August 2006

Abstract

The onramp metering control problem is posed using a cell transmission-like model called the asymmetric cell transmission model (ACTM). The problem formulation captures both freeflow and congested conditions, and includes upper bounds on the metering rates and on the onramp queue lengths. It is shown that a near-global solution to the resulting nonlinear optimization problem can be found by solving a single linear program, whenever certain conditions are met. The most restrictive of these conditions requires the congestion on the mainline not to back up onto the onramps whenever optimal metering is used. The technique is tested numerically using data from a severely congested stretch of freeway in southern California. Simulation results predict a 17.3% reduction in delay when queue constraints are enforced.

Keywords: Freeway modeling; Macroscopic model; Ramp metering; Optimization

1. Introduction

The increasing congestion on urban freeways is a fact that is not only obvious to most commuters, but also well documented. The 2005 edition of the Urban Mobility Report (Schrank and Lomax, 2005) states that per-traveler annual delay has increased from 16 h in 1982 to 47 h in 2003. The annual delay on freeways has gone from 0.7 billion hours to 3.7 billion hours, while the percentage of the total classified as either *severe* or *extreme* congestion has risen from 12% to 40%. These trends are countered by traffic engineers with a variety of measures, including infrastructure expansions, public transportation services, and several operational enhancements known collectively as Intelligent Transportation Systems (ITS). Among the operational strategies for improving freeway performance is *onramp metering*, in which the flow of vehicles allowed onto the freeway is regulated in order to avoid breakdown due to oversaturation.

* Corresponding author. Tel.: +1 510 665 3710; fax: +1 510 665 3537.
E-mail addresses: gomes@path.berkeley.edu (G. Gomes), horowitz@me.berkeley.edu (R. Horowitz).

0968-090X/$ - see front matter
doi:10.1016/j.trc.2006.08.001

The history of optimization-based ramp metering begins with the work of Wattleworth and Berry (1965). This first formulation of the problem used a static model of traffic behavior, whereby the flow at any cross-section in the system could be expressed as the sum of the flows entering the freeway upstream of that location, scaled by a known proportion of vehicles that did not exit at any upstream offramp. This density-less model lead to a linear program because it avoided the main nonlinearity in freeway traffic behavior: the relationship between flow and density known as the *fundamental diagram*. Many subsequent contributions built upon this approach, including Yuan and Kreer (1971) and Wang and May (1973). Later authors further extended the model to capture the entire *corridor*, which comprises the freeway and an alternative parallel route. Payne and Thompson (1974) considered ''Wardrop's first principle'' as dictating the selection of routes by drivers, coupled with an onramp control formulation similar to Wattleworth's, and solved it with a suboptimal dynamic programming algorithm. Iida et al. (1989) posed a similar problem, and employed a heuristic numerical method consisting of iterated solutions of two linear programs (control and assignment). A more recent enhancement has been the consideration of *dynamic models*. Most problem formulations using dynamic models have reverted to the simpler situation, where the effect of onramp control on route selection is not considered (Bellemans et al., 2003; Hegyi et al., 2002; Kotsialos et al., 2002). Papageorgiou (1980) posed a linear program using a dynamic model restricted to freeflow traffic conditions.

The task of generating optimal ramp metering plans is a delicate one. Zhang et al. (1996) concludes that freeways are best left uncontrolled (i.e. no improving controller exists) whenever they are either *uniformly congested* or *uniformly uncongested*, meaning that the state of congestion cannot be affected by onramp control. Even when the freeway is in a state of *mixed congestion*, and can therefore benefit from onramp control, there exist only a few mechanisms for reducing travel time. Banks (2000) identifies four: (1) increasing bottleneck flow, (2) diverting traffic to alternative routes, (3) preventing accidents, and (4) preventing the obstruction of offramps by congestion on the mainline. The second and third mechanisms are difficult phenomena to model and verify, and are not considered in most optimal control designs. Increasing bottleneck and offramp flow, both related to the avoidance of congestion, are left as the two principal mechanisms for reducing travel time. However, congestion can only be reduced by storing the surplus vehicles in the onramp queues, and this often conflicts with the limited storage space in the onramps. These can typically hold up to 30 vehicles each, which is a small number when compared to the number of vehicles on a congested freeway. The metering problem is thus recognized as one of careful management of onramp storage space and timely release of accumulated onramp queues. Given the small margins, the quality of the numerical solution becomes a very important factor, in addition to the validity of the model and its calibration.

The most commonly used models in freeway control design are first order models, such as the cell-transmission model (Daganzo, 1994; Daganzo, 1995), and second order models, such as Metanet (Messmer and Papageorgiou, 1990). Second order models have the distinct advantage over first order models that they can reproduce the *capacity drop*, which is the observed difference between the freeway capacity and the queue discharge rate. First order models, because they do not capture this phenomenon, are incapable of exploiting the benefits of increasing bottleneck flow (Banks' first mechanism). They can only reduce travel time by increasing offramp flow. The obvious disadvantage to second order models is that they lead to more complex optimization problems. To date, the optimization problems constructed using second order models have only been solved in the sense of local optimality (Bellemans et al., 2003; Hegyi et al., 2002; Kotsialos et al., 2002).

There are at least two scenarios in which the use of first order models is justified. First, when the bottleneck is closely preceded by an offramp. This situation is common, since bottlenecks are often caused by traffic merges immediately downstream of an offramp/onramp pair. In this case, the two mechanisms (capacity drop and offramp blockage) are triggered more or less simultaneously, and the optimal plans for first and second order models can be expected to be similar (both will seek to minimize congestion). Second, when the *duration* of the congestion period cannot be significantly altered by ramp metering, due to limitations in onramp storage space. Here travel time can only be reduced by managing the *length* of the mainline queue, such that offramp blockage is minimized. Again, this situation is probably quite common.

The technique developed here produces a global solution to a first order model. It is the only approach known to the authors to yield a global optimum without constraining the model to freeflow speeds. The approach derives from two facts. First, the model's only nonlinearity is the fundamental diagram $F(\rho)$, which is a concave function. The set defined by all values of flow below this function is therefore a convex set, as

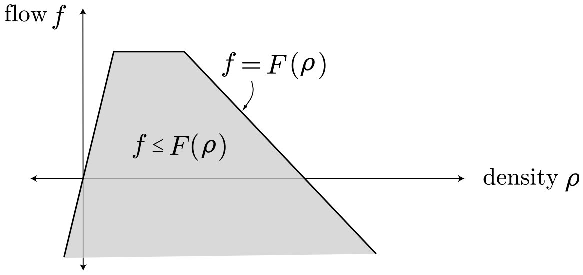

Fig. 1. Concave fundamental diagram.

illustrated in Fig. 1. The second fact is that minimizing travel time is equivalent to maximizing a weighted sum of flows. Because the travel time objective function favors larger flows, it is not unreasonable to expect the solution to the relaxed problem to "naturally" fall on the upper boundary, and therefore solve the nonlinear problem as well. This idea of relaxing the flow constraint has been suggested previously. Papageorgiou (1995) makes similar claims for a store-and-forward type freeway model. Ziliaskopoulos (2000) formulates a linear program for the dynamic traffic assignment problem, but does not require the solution to fall on the fundamental diagram.

The paper is organized as follows. Section 2 describes the freeway model. It is shown in Section 3 that negative flows and densities are not predicted by the model. Section 4 provides the formulation of the nonlinear problem and its linear relaxation, as well as proof of the main result. The technique is demonstrated with a numerical example in Section 5.

2. The asymmetric cell transmission model (ACTM)

The ACTM is a modified version of Daganzo's cell-transmission model (CTM) (Daganzo, 1994; Daganzo, 1995). The important difference between the two is in the treatment of traffic merges. In contrast with the CTM, merges in the ACTM are limited to asymmetric connections, such as onramp-to-mainline junctions, where a minor branch feeds into a major branch. An additional parameter (γ) is used to control the blending of the two flows.

To apply the ACTM, the freeway is divided into I sections, with each section containing at most one onramp and/or one offramp (Fig. 2). In sections with both an onramp and an offramp, the onramp must be upstream of the offramp. Freeway sections are numbered 0 through $I-1$, starting from the upstream-most section. Time is divided into K intervals of length Δt. The following are sets of section and time indices:

$\mathscr{I}$	set of all freeway sections	$\mathscr{I} = \{0 \ldots I-1\}$
$\mathscr{K}$	set of time intervals	$\mathscr{K} = \{0 \ldots K-1\}$
$\mathscr{E}$	set of sections with onramps	$\mathscr{E} \subseteq \mathscr{I}$
$\mathscr{E}^+$	set of sections with metered onramps	$\mathscr{E}^+ \subseteq \mathscr{E}$

All traffic variables are normalized to vehicle units. Flow variables $f_{i[k]}$, $r_{i[k]}$, $c_{i[k]}$, $d_{i[k]}$, and $s_{i[k]}$ are interpreted as a number of vehicles per time interval Δt. Density variables $n_{i[k]}$ and $l_{i[k]}$ represent the number of vehicles on the mainline and onramp portions of section i at time $k\Delta t$. Definitions for each of these quantities are given below.

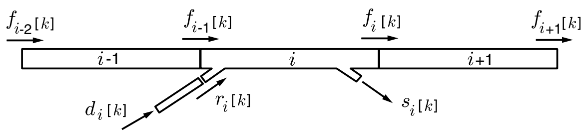

Fig. 2. Model variables.

$n_{i[k]}$	number of vehicles in section i at time $k\Delta t$.
$l_{i[k]}$	number of vehicles queueing in the onramp of section $i \in \mathscr{E}$ at time $k\Delta t$.
$f_{i[k]}$	number of vehicles moving from section i to $i+1$ during interval k.
$r_{i[k]}$	number of vehicles entering section $i \in \mathscr{E}$ from its onramp during interval k.
$c_{i[k]}$	metering rate for onramp $i \in \mathscr{E}^+$.
$d_{i[k]}$	demand for onramp $i \in \mathscr{E}$.
$s_{i[k]}$	number of vehicles using offramp i during interval k.

The parameters of the model are listed below. Their rough interpretation as parameters in a triangular fundamental diagram is illustrated in Fig. 3.

v_i	normalized freeflow speed	$\in [0,1]$
w_i	normalized congestion wave speed	$\in [0,1]$
ξ_i	onramp flow allocation parameter	$\in [0,1]$
$\bar{n}_i$	jam density	[veh]
$\bar{f}_i$	mainline capacity	[veh]
$\bar{s}_i$	offramp capacity	[veh]
γ	onramp flow blending coefficient	$\in [0,1]$
$\beta_{i[k]}$	dimensionless split ratio for offramp i	$\in [0,1]$

The model has five basic components: the mainline and onramp conservation equations, mainline and onramp flows, and offramp flows. Offramp flow is assumed to be related to the mainline flow by a known *split ratio* $\beta_{i[k]} \in [0,1]$:

Offramp flows $\forall i \in \mathscr{I},\ k \in \mathscr{K}$:

$$\begin{aligned} s_{i[k]} &= \beta_{i[k]}(s_{i[k]} + f_{i[k]}) \\ \therefore \quad s_{i[k]} &= \frac{\beta_{i[k]}}{1-\beta_{i[k]}} f_{i[k]} = \frac{\beta_{i[k]}}{\bar{\beta}_{i[k]}} f_{i[k]} \end{aligned} \tag{1}$$

where $\bar{\beta}_{i[k]} \triangleq 1 - \beta_{i[k]}$ has been defined to simplify the equations. Also for convenience, the split ratio is defined for every section and set to 0 if the section does not contain an offramp. The special case of $\beta_{i[k]} = 1$, in which the offramp flows cannot be determined from Eq. (1), is resolved with:

$$s_{i[k]} = \min\{v_i(n_{i[k]} + \gamma r_{i[k]}); \bar{s}_i\} \tag{2}$$

The assumption of given split ratios is common in freeway control design but not entirely correct, since these are actually functions of the control variable. The alternative is to assume known origin–destination information, however this has its own drawbacks. For example, the OD estimation problem is not uniquely solvable given only loop detector data. Also, segregating flows by destination introduces the problem of having to manage a FIFO queue on the onramps, which has been shown by Erera et al. (1999) to make the ramp metering

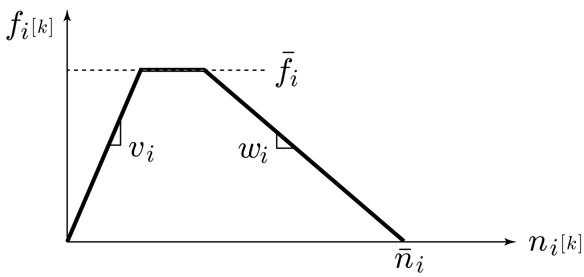

Fig. 3. Model parameters.

problem intractable. Zhang and Levinson (2004) provide convincing arguments in favor of the use of split ratios instead of origin–destination matrices.

Mainline flow is calculated, similarly to the CTM, as the minimum of what can be sent by the upstream section assuming maximum speed and what can be received by the downstream section. It is assigned the largest value of $f_{i[k]}$ that complies with:

$$f_{i[k]} + s_{i[k]} \leqslant v_i(n_{i[k]} + \gamma r_{i[k]}) \quad \ldots\text{freeflow term} \tag{3}$$
$$f_{i[k]} \leqslant w_{i+1}(\bar{n}_{i+1} - n_{i+1[k]} - \gamma r_{i+1[k]}) \quad \ldots\text{congestion term} \tag{4}$$
$$f_{i[k]} \leqslant \bar{f}_i \quad \ldots\text{mainline capacity} \tag{5}$$
$$s_{i[k]} \leqslant \bar{s}_i \quad \ldots\text{offramp capacity} \tag{6}$$

Eq. (3) limits the total flow that can leave section i during time interval k, assuming that traffic moves at the freeflow speed v_i. Eq. (4) ensures that the mainline flow does not exceed what can be accommodated by the downstream section. The right hand side of this equation is the portion w_{i+1} of the total unoccupied space in section $i+1$. Eqs. (5) and (6) are the mainline and offramp capacity limits.

The section densities of Eqs. (3) and (4) are intermediate values which include a portion γ of the onramp flow. This blending coefficient dictates how much of the onramp flow is added to the mainline stream before the mainline flow is computed.[1] Considering Eq. (1), this leads to the following expression for $f_{i[k]}$:

Mainline flows $\forall i \in \mathscr{I}$, $k \in \mathscr{K}$:

$$f_{i[k]} = \min\{\bar{\beta}_{i[k]} v_i(n_{i[k]} + \gamma r_{i[k]}); w_{i+1}(\bar{n}_{i+1} - n_{i+1[k]} - \gamma r_{i+1[k]}); F_{i[k]}\} \tag{7}$$

where $F_{i[k]} \triangleq \min\left\{\bar{f}_i; \frac{\bar{\beta}_{i[k]}}{\beta_{i[k]}}\bar{s}_i\right\}$. Note that Eq. (7) with $\gamma = 0$ is similar to the CTM equation for simple or diverging cell connections and specified turning percentages (Daganzo, 1995, Eqs. (4) and (9b)). Analogous to mainline flows, onramp flows are computed such that none of the following limits are exceeded:

$$r_{i[k]} \leqslant l_{i[k]} + d_{i[k]} \quad \ldots\text{demand} \tag{8}$$
$$r_{i[k]} \leqslant \xi_i(\bar{n}_i - n_{i[k]}) \quad \ldots\text{mainline space} \tag{9}$$
$$r_{i[k]} \leqslant c_{i[k]} \quad \ldots\text{ramp metering rate (for } i \in \mathscr{E}^+) \tag{10}$$

Eq. (9) is a restriction to $r_{i[k]}$ due to limited space on the mainline. The parameter ξ_i determines the allotment of available space for vehicles entering from the onramp. This leads to the following expression for $r_{i[k]}$:

Onramp flows $\forall i \in \mathscr{E}$, $k \in \mathscr{K}$:

$$r_{i[k]} = \begin{cases} \min\{l_{i[k]} + d_{i[k]}; \xi_i(\bar{n}_i - n_{i[k]})\} & \text{if } i \in \mathscr{E} \setminus \mathscr{E}^+ \\ \min\{l_{i[k]} + d_{i[k]}; \xi_i(\bar{n}_i - n_{i[k]}); c_{i[k]}\} & \text{if } i \in \mathscr{E}^+ \end{cases} \tag{11}$$

This onramp flow equation is similar in form to Eq. (7). It has been suggested previously in (Kotsialos et al., 2002). The number of vehicles in the onramp and on the mainline evolve according to conservation Eqs. (12) and (13).

Onramp conservation $\forall i \in \mathscr{E}$, $k \in \mathscr{K}$:

$$l_{i[k+1]} = l_{i[k]} + d_{i[k]} - r_{i[k]} \tag{12}$$

with initial condition $l_{i[0]}$ and boundary condition $d_{i[k]}$.

Mainline conservation $\forall i \in \mathscr{I}$, $k \in \mathscr{K}$:

$$\begin{aligned} n_{i[k+1]} &= n_{i[k]} + f_{i-1[k]} + r_{i[k]} - f_{i[k]} - s_{i[k]} \\ &= n_{i[k]} + f_{i-1[k]} + r_{i[k]} - f_{i[k]}/\bar{\beta}_{i[k]} \quad (\text{when } \beta_{i[k]} \neq 1) \end{aligned} \tag{13}$$

[1] The blending coefficient is considered uniform for notational purposes only. Different values of γ could be used for each section, and even for Eq. (3) versus (4), with only slight changes to Theorem A. Furthermore, the value of γ does not enter the discussion of Section 4 (beyond having to meet the requirements of Theorem A).

with initial condition $n_{i[0]}$. The boundary condition for this equation is the flow entering the first mainline section, $up_{[k]}$. It can be represented as either a prescribed mainline flow, i.e. $f_{-1[k]} = up_{[k]}$ or as a demand into a fictitious onramp, i.e. $d_{0[k]} = up_{[k]}$ and $f_{-1[k]} = 0$. The second method is preferred because it prevents the upstream section from overflowing (Theorem A).

Eqs. (7), (11)–(13) constitute the ACTM. The only significant departure from the CTM is in the calculation of merging flows. The approach used in the CTM is to allocate a portion of the available space in the downstream receiving cell, and to move as much of the demand as possible from the two sending cells into the common space. The ACTM on the other hand, makes separate allocations for each merging branch, w_i for the mainline and ξ_i for the onramp. The flows can then be calculated separately in the same way as simple cell connections: by taking the minimum of the demand, the allocated space, and the capacity (or ramp metering rate). Thus, the non-concave/non-convex mid{ } functions of the CTM are replaced with concave min{ } functions. This structural change is the basis for the optimization technique developed in Section 4.

3. Implicit bounds

An important property of the original CTM is that it never predicts negative flows or densities, nor do the densities ever exceed the jam density. That is, the following constraints always hold:

$$n_i[k] \in [0, \bar{n}_i] \quad \text{and} \quad f_{i[k]} \geqslant 0 \tag{14}$$

These *implicit bounds* are a minimum requirement for any model to be considered a reasonable approximation of freeway traffic. In the case of the CTM, they are a consequence of the consistency of the model with the LWR theory, and of the particular rules used for merges and diverges.

It is well known that a major drawback of many higher order models is that they can predict backward moving traffic. The problem is typically dealt with by replacing the negative values with small positive values as the model is being integrated. However, such an artificial fix requires the model equations to be violated, which compromises its usefulness as a tool for understanding traffic behavior. In the context of optimal control design, a hard positivity constraint is usually imposed, but this only masks the underlying problem. The question that arises is whether the ACTM retains the property expressed by Eq. (14). The following theorem establishes conditions under which it does.

Theorem A. *Given initial and boundary conditions, ramp metering rates, and model parameters satisfying,*

$$\begin{array}{lll}
\textit{Initial conditions:} & n_{i[0]} \in [0, \bar{n}_i] & \forall i \in \mathscr{I} \\
 & l_{i[0]} \geqslant 0 & \forall i \in \mathscr{E} \\
\textit{Boundary conditions:} & d_{i[k]} \geqslant 0 & \forall i \in \mathscr{E}, k \in \mathscr{K} \\
 & f_{-1[k]} = 0 & \forall k \in \mathscr{K} \\
\textit{Onramp metering rates:} & c_{i[k]} \geqslant 0 & \forall i \in \mathscr{E}^+, k \in \mathscr{K} \\
\textit{Model parameters:} & v_i, w_i \in [0, 1] & \forall i \in \mathscr{I} \\
 & \xi_i \in \left[0, \frac{1-w_i}{1-\gamma w_i}\right] & \forall i \in \mathscr{E} \\
 & \gamma \in [0, 1] & \forall i \in \mathscr{E} \\
 & \bar{f}_i, \bar{s}_i \geqslant 0 & \forall i \in \mathscr{I} \\
 & \beta_{i[k]} \in [0, 1] & \forall i \in \mathscr{I}, k \in \mathscr{K}
\end{array}$$

The evolution of the ACTM is then bounded by

$$\begin{array}{lll}
n_{i[k]} \in [0, \bar{n}_i], & f_{i[k]} \geqslant 0 & \forall i \in \mathscr{I}, k \in \mathscr{K} \\
l_{i[k]} \geqslant 0, & r_{i[k]} \geqslant 0 & \forall i \in \mathscr{E}, k \in \mathscr{K}
\end{array}$$

A proof can be found in Appendix A. This theorem ensures that unrealistic behaviors such as backward moving traffic, negative densities, and densities exceeding the jam density are not predicted by the ACTM. Most of the conditions are covered by the physical definitions of the parameters and variables; e.g. v_i, $w_i \in [0,1]$, $d_{i[k]} \geqslant 0$, etc. The only two that are not necessarily satisfied are $f_{-1[k]} = 0$ and the upper bound on ξ_i. The first, $f_{-1[k]} = 0$, is met if the upstream mainline boundary flow is supplied through a fictitious onramp into section $i = 0$. The only restrictive condition is then $\xi_i \leqslant (1 - w_i)/(1 - \gamma w_i)$. However, w_i is usually no greater than 0.3 (a freeflow speed of 100 kph and a congestion wave speed of 25 kph yields $w_i < 0.25$). With $w_i\gamma \in [0,1]$, the bound is no more restrictive than $\xi_i \leqslant 0.7$. Realistic values of ξ_i are well within this bound.

4. Problem formulation and solution

Our goal is to find ramp metering rates that minimize the *total travel time* incurred by all users of the freeway system. This will be achieved by solving a nonlinear optimization problem. In addition to the constraints of the traffic model, the formulation also includes limits on the *metering rates* and *onramp queue lengths*

$$\underline{\text{Metering rate bounds:}} \quad c_{i[k]} \geqslant \underline{c}_i \quad \forall k \in \mathscr{K},\ i \in \mathscr{E}^+ \tag{15}$$

$$c_{i[k]} \leqslant \bar{c}_i \quad \forall k \in \mathscr{K},\ i \in \mathscr{E}^+ \tag{16}$$

$$\underline{\text{Queue length bounds:}} \quad l_{i[k]} \leqslant \bar{l}_i \quad \forall k \in \mathscr{K},\ i \in \mathscr{E} \tag{17}$$

where $\underline{c}_i$ $\bar{c}_i$, and $\bar{l}_i$ are given constants. The objective function to be minimized is a linear combination of total travel time (TTT) and total travel distance (TTD):

$$J \triangleq \mathtt{TTT} - \eta\, \mathtt{TTD} \tag{18}$$

with $\eta > 0$. TTT and TTD are defined as

$$\mathtt{TTT} \triangleq \sum_{i\in\mathscr{I}} \sum_{k\in\mathscr{K}} n_{i[k]} + \sum_{i\in\mathscr{E}} \sum_{k\in\mathscr{K}} l_{i[k]} \tag{19}$$

$$\mathtt{TTT} \triangleq \sum_{i\in\mathscr{I}} \sum_{k\in\mathscr{K}} f_{i[k]} + \sum_{i\in\mathscr{E}} \sum_{k\in\mathscr{K}} r_{i[k]} \tag{20}$$

This objective function favors larger travel distances with smaller, but not necessarily minimal travel times. However, it is shown in Appendix B that TTD is a prescribed constant, independent of the metering rates, whenever the split ratios are constant in time and the final condition is an empty freeway. Minimizing J therefore also minimizes TTT under these two conditions.

In order to approximate an empty final condition, a fictitious "cool down" period must be appended to the end of the optimization time window, in which all demands are set to zero. With η positive, it will always be advantageous to evacuate the freeway by maximizing onramp and mainline flows during the cool down period. A positive η is also required by lemma B to guarantee the equivalence of the nonlinear and linear problems. The significance of these assumptions is discussed at the end of this section.

Next we state three optimization problems. $\mathscr{N}$ is the full nonlinear problem whose solution is sought. $\mathscr{M}$ is a nonlinear simplification of $\mathscr{N}$. By lemma A, a solution to $\mathscr{M}$ can be used to construct a solution to $\mathscr{N}$, under certain conditions. $\mathscr{L}$ is a linear relaxation of $\mathscr{M}$. Lemma B states that $\mathscr{L}$ and $\mathscr{M}$ are equivalent given another set of conditions. These two results are combined to establish Theorem B.

Problem $\mathscr{N}$. Given conditions satisfying Theorem A,

minimize: TTT $- \eta$ TTD

subject to: Conservation equations: Eqs. (12) and (13),
Mainline and onramp flows: Eqs. (7) and (11),
Metering rate bounds: Eqs. (15) and (16),
Queue length bounds: Eq. (17)

Problem $\mathscr{M}$. Given conditions satisfying Theorem A,

minimize: $\mathtt{TTT} - \eta\,\mathtt{TTD}$

subject to: Conservation equations: Eqs. (12) and (13),

Mainline flows: Eq. (7),

Simplified onramp flows:

$$r_{i[k]} = d_{i[k]}, \quad k \in \mathscr{K}, i \in \mathscr{E} \setminus \mathscr{E}^+ \tag{21}$$

$$r_{i[k]} \leqslant l_{i[k]} + d_{i[k]}, \quad k \in \mathscr{K}, i \in \mathscr{E}^+ \tag{22}$$

$$r_{i[k]} \leqslant \bar{c}_i, \quad k \in \mathscr{K}, i \in \mathscr{E}^+ \tag{23}$$

$$r_{i[k]} \geqslant 0, \quad k \in \mathscr{K}, i \in \mathscr{E}^+ \tag{24}$$

Queue length bounds: Eq. (17)

Problem $\mathscr{L}$. Given conditions satisfying Theorem A,

minimize: $\mathtt{TTT} - \eta\,\mathtt{TTD}$

subject to: Conservation equations: Eqs. (12) and (13),

Relaxed mainline flows:

$$f_{i[k]} \leqslant \bar{\beta}_{i[k]} v_i (n_{i[k]} + \gamma r_{i[k]}), \quad k \in \mathscr{K}, i \in \mathscr{I} \tag{25}$$

$$f_{i[k]} \leqslant w_{i+1} (\bar{n}_{i+1} - n_{i+1[k]} - \gamma r_{i+1[k]}), \quad k \in \mathscr{K}, i \in \mathscr{I} \tag{26}$$

$$f_{i[k]} \leqslant F_{i[k]}, \quad k \in \mathscr{K}, i \in \mathscr{I} \tag{27}$$

Simplified onramp flows: Eqs. (21) through (24),

Queue length bounds: Eq. (17)

Problem $\mathscr{N}$ is a non-concave and non-convex problem, due to the mainline and onramp flow $\mathscr{E}$ constraints Eqs. (7) and (11). Eq. (11) is replaced in Problem $\mathscr{M}$ with linear equality and inequality constraints, Eqs. (21)–(24). Note that $\mathscr{M}$ does not include the onramp metering rates $c_{i[k]}$ as unknowns. Note also that the traffic model equations used in the problem statements do not cover the case $\beta_{i[k]} = 1$. All split ratios will be assumed less than 1 throughout this section. This assumption is made without loss of generality, given the added assumption of constant split ratios (from lemma B), since a constant split ratio of 1 effectively divides the freeway into independent portions which can be treated separately.

The following lemma shows that $\mathscr{N}$-optimal metering rates can be derived from a solution to $\mathscr{M}$.

Lemma A. *A solution to $\mathscr{N}$ can be constructed from a solution to $\mathscr{M}$ whenever*:

1. *Each $\mathscr{M}$-optimal $r_{i[k]}$ is strictly less than $\xi_i(\bar{n}_i - n_{i[k]})$, and*
2. $\underline{c}_i = 0$.

Proof. Under the first assumption, $r_{i[k]}$ never equals the $\xi_i(\bar{n}_i - n_{i[k]})$ term in Eq. (11). Then,

$$r_{i[k]} = \begin{cases} l_{i[k]} + d_{i[k]} & \text{if } i \in \mathscr{E} \setminus \mathscr{E}^+ \\ \min\{l_{i[k]} + d_{i[k]}; c_{i[k]}\} & \text{if } i \in \mathscr{E}^+ \end{cases}$$

In the unmetered case, using Eq. (12) we find that onramp queues do not form ($l_{i[k]} = 0$). Hence, $r_{i[k]}$ equals the onramp demand $d_{i[k]}$ (except at $k = 0$ where $l_{i[0]}$ must be added). For metered onramps, using the second assumption, Eqs. (11), (15) and (16) become

$$r_{i[k]} = \min\{l_{i[k]} + d_{i[k]}; c_{i[k]}\} \tag{28}$$

$$c_{i[k]} \geqslant 0 \tag{29}$$

$$c_{i[k]} \leqslant \bar{c}_i \tag{30}$$

The metering rate $c_{i[k]}$ is a free parameter, constrained only by its lower and upper bounds 0 and $\bar{c}_i$. The onramp flow $r_{i[k]}$ is at most $l_{i[k]} + d_{i[k]}$, and less only when $c_{i[k]}$ is less than $l_{i[k]} + d_{i[k]}$. In $\mathscr{M}$, the onramp flows are restricted to

$$r_{i[k]} \leqslant l_{i[k]} + d_{i[k]} \tag{31}$$

$$r_{i[k]} \geqslant 0 \tag{32}$$

$$r_{i[k]} \leqslant \bar{c}_i \tag{33}$$

It can be easily verified that by defining $c_{i[k]} = r_{i[k]}$, with $r_{i[k]}$ conforming to (31)–(33), all of constraints (28)–(30) are satisfied. The optimal solution to $\mathscr{M}$ along with $c_{i[k]} = r_{i[k]}$ is therefore feasible for $\mathscr{N}$. It is also optimal since $c_{i[k]}$ does not appear elsewhere in $\mathscr{M}$. $\square$

The first requirement of lemma A states that no onramp flows should be restricted by a lack of space on the mainline, whenever the freeway is optimally metered. This rarely happens on *metered* onramps, where the onramp flow is limited by the maximum metering rate $\bar{c}_i$, which can usually be accommodated by the mainline. However the condition may disqualify some freeways with heavy *unmetered* onramps, such as freeway-to-freeway connectors.

Lemma B. *Problems $\mathscr{M}$ and $\mathscr{L}$ are equivalent, in the sense that their solution sets are identical, whenever*

1. *all split ratios are constant in time (denoted β_i), and*
2. *all offramp-less sections have $v_i < 1$ and $w_{i+1} < 1$.*

Proof. The two problems are considered equivalent if every $\mathscr{L}$-optimal solution is also $\mathscr{M}$-optimal, and vice-versa:

$$\{\psi \text{ solves } \mathscr{M}\} \Longleftrightarrow \{\psi \text{ solves } \mathscr{L}\}$$

We denote the feasibility sets for $\mathscr{L}$ and $\mathscr{M}$ respectively as $\Omega_{\mathscr{L}}$ and $\Omega_{\mathscr{M}}$. Note that $\mathscr{L}$ is a relaxation of $\mathscr{M}$, since $\Omega_{\mathscr{M}}$ is contained in $\Omega_{\mathscr{L}}$. Therefore, any solution of $\mathscr{L}$ that lies within $\Omega_{\mathscr{M}}$ must also solve $\mathscr{M}$. For the two problems to be equivalent, the entire set of solutions of $\mathscr{L}$ must be contained in $\Omega_{\mathscr{M}}$.

$$\{\psi \text{ solves } \mathscr{L}\} \;\Rightarrow\; \psi \in \Omega_{\mathscr{M}}$$

Conversely stated, the problems are equivalent if there are no solutions of $\mathscr{L}$ in the set $\Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$:

$$\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}} \;\Rightarrow\; \{\psi \text{ does not solve } \mathscr{L}\}$$

In more concrete terms, we seek to show that if a point $\psi = \{n_{i[k]}, l_{i[k]}, f_{i[k]}, r_{i[k]}\} \in \Omega_{\mathscr{L}}$ has some component $f_{i[k]}$ in the interior of the set defined by Eqs. (25)–(27), then ψ cannot be $\mathscr{L}$-optimal. A feasible point ψ can be shown *not* to solve $\mathscr{L}$ if there exists a perturbation $\mathbf{\Delta}$ that is both *feasible* and *improving*:

$$\mathbf{\Delta} \text{ is } \textit{feasible} \text{ if } \exists \epsilon > 0 \text{ such that:} \quad \psi + \epsilon\mathbf{\Delta} \in \Omega_{\mathscr{L}} \tag{34}$$

$$\mathbf{\Delta} \text{ is } \textit{improving} \text{ if } \exists \epsilon > 0 \text{ such that:} \quad J(\psi + \epsilon\mathbf{\Delta}) < J(\psi) \tag{35}$$

Due to the linearity of $J(\psi)$, Eq. (35) is equivalent to $J(\mathbf{\Delta}) < 0$. We will prove equivalence by finding a feasible and improving perturbation for every $\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$. The concept is illustrated in Fig. 4.

Every point $\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$ can be classified according to the section and time indices, i and k, of the component $f_{i[k]}$ that falls within the interior of Eqs. (25)–(27). This classification generates $I \times K$ categories or subsets Γ_{ik}. Every $\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$ belongs to at least one subset Γ_{ik}, since at least one of its component $f_{i[k]}$'s must lie beneath its upper boundary. The components of a point $\psi \in \Gamma_{\iota\kappa}$ satisfy the following:

$$n_{i[k+1]} = n_{i[k]} + f_{i-1[k]} + r_{i[k]} - f_{i[k]}/\bar{\beta}_{i[k]} \tag{36}$$

$$l_{i[k+1]} = l_{i[k]} + d_{i[k]} - r_{i[k]} \tag{37}$$

$$f_{i[k]} < \min\{\cdots\} \quad \text{if } i = \iota, k = \kappa \tag{38}$$

$$f_{i[k]} \leqslant \min\{\cdots\} \quad \text{otherwise} \tag{39}$$

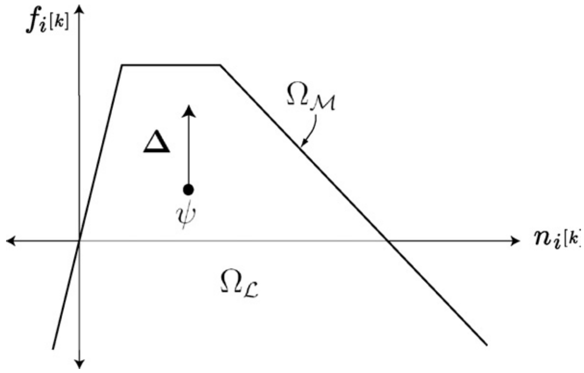

Fig. 4. Perturbation to $\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$.

$$r_{i[k]} = d_{i[k]}, \quad i \in \mathscr{E} \setminus \mathscr{E}^+ \tag{40}$$
$$r_{i[k]} \leqslant l_{i[k]} + d_{i[k]}, \quad i \in \mathscr{E}^+ \tag{41}$$
$$r_{i[k]} \leqslant \bar{c}_i, \quad i \in \mathscr{E}^+ \tag{42}$$
$$r_{i[k]} \geqslant 0, \quad i \in \mathscr{E}^+ \tag{43}$$

where $\min\{\cdots\}$ is shorthand for the right hand side of Eq. (7). For each subset $\Gamma_{\imath\kappa}$ we define a particular perturbation $\overline{\mathbf{\Delta}}_{\imath\kappa} = \{\Delta n_{i[k]}, \Delta l_{i[k]}, \Delta f_{i[k]}, \Delta r_{i[k]}\}$ as follows:

$$\Delta n_{i[k+1]} = \Delta n_{i[k]} + \Delta f_{i-1[k]} + \Delta r_{i[k]} - \Delta f_{i[k]}/\bar{\beta}_{i[k]} \quad \text{with } \Delta n_{i[0]} = 0 \tag{44}$$
$$\Delta f_{i[k]} = \begin{cases} 1 & \text{if } i = \imath, k = \kappa \\ \min\{\bar{\beta}_{i[k]} v_i \Delta n_{i[k]}; -w_{i+1}\Delta n_{i+1[k]}; 0\} & \text{otherwise} \end{cases} \tag{45}$$
$$\Delta l_{i[k]} = \Delta r_{i[k]} = 0 \tag{46}$$

$\overline{\mathbf{\Delta}}_{\imath\kappa}$ will be shown to be a feasible and improving perturbation for every $\psi \in \Gamma_{\imath\kappa}$. To show feasibility, we verify that the components of $\psi + \epsilon\overline{\mathbf{\Delta}}_{\imath\kappa}$ satisfy each of the equations that define $\Omega_{\mathscr{L}}$, for some $\epsilon > 0$. For example, adding Eq. (36) to ϵ times Eq. (44),

$$(n_{i[k+1]} + \epsilon\Delta n_{i[k+1]}) = (n_{i[k]} + \epsilon\Delta n_{i[k]}) + (f_{i-1[k]} + \epsilon\Delta f_{i-1[k]}) + (r_{i[k]} + \epsilon\Delta r_{i[k]}) - (f_{i[k]} + \epsilon\Delta f_{i[k]})/\bar{\beta}_{i[k]}$$

we find that the components of $\psi + \epsilon\overline{\mathbf{\Delta}}_{\imath\kappa}$ satisfy Eq. (13) for any ϵ. Eq. (12) is verified similarly. Eqs. (21)–(24) and (17) are trivially satisfied since $\Delta l_{i[k]} = \Delta r_{i[k]} = 0$. The three relaxed mainline flow Eqs. (25)–(27), have two cases: $[i = \imath,\ k = \kappa]$ and [otherwise]. In the first case we have $\Delta f_{\imath[\kappa]} = 1$ and $f_{\imath[\kappa]} < \min\{\cdots\}$. As illustrated in Fig. 5, $\Delta n_{\imath[\kappa]} = \Delta n_{\imath+1[\kappa]} = 0$. Then, using Eq. (38):

$$f_{\imath[\kappa]} < \min\{\bar{\beta}_{\imath[\kappa]} v_\imath((n_{\imath[\kappa]} + \epsilon\Delta n_{\imath[\kappa]}) + \gamma(r_{\imath[\kappa]} + \epsilon\Delta r_{\imath[\kappa]})); w_{\imath+1}(\bar{n}_{\imath+1} - (n_{\imath+1[\kappa]} + \epsilon\Delta n_{\imath+1[\kappa]}) \\ - \gamma(r_{\imath+1[\kappa]} + \epsilon\Delta r_{\imath+1[\kappa]})); F_{\imath[\kappa]}\}$$

for any ϵ. Focusing on the first term in the min{ }, it is always possible to find some $\epsilon > 0$ such that:

$$f_{\imath[\kappa]} + \epsilon\Delta f_{\imath[\kappa]} \leqslant \bar{\beta}_{\imath[\kappa]} v_\imath((n_{\imath[\kappa]} + \epsilon\Delta n_{\imath[\kappa]}) + \gamma(r_{\imath[\kappa]} + \epsilon\Delta r_{\imath[\kappa]}))$$

This verifies Eq. (25) in the case $[i = \imath,\ k = \kappa]$. Eqs. (26) and (27) are done similarly. In the [otherwise] case we have

$$f_{i[k]} \leqslant \min\{\cdots\}$$
$$\Delta f_{i[k]} = \min\{\bar{\beta}_{i[k]} v_i \Delta n_{i[k]}; -w_{i+1}\Delta n_{i+1}[k]; 0\}$$

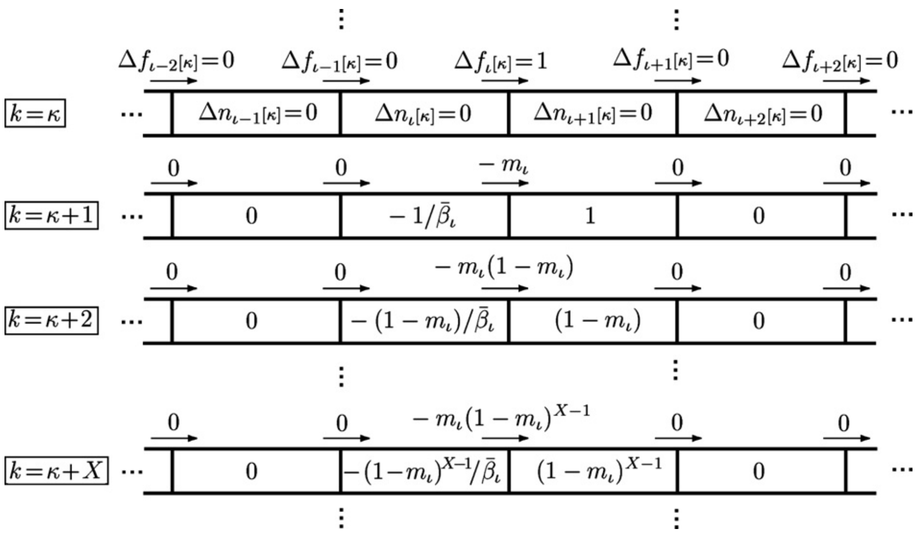

Fig. 5. Evolution of $\overline{\mathbf{\Delta}}_{\imath\kappa}$.

Taking the first term in each min{ },

$$
\begin{array}{rcll}
f_{i[k]} & \leqslant & \bar{\beta}_{i[k]} v_i (n_{i[k]} + \gamma r_{i[k]}) & \\
\Delta f_{i[k]} & \leqslant & \bar{\beta}_{i[k]} v_i \Delta n_{i[k]} & \times\, \epsilon \\
\hline
f_{i[k]} + \epsilon \Delta f_{i[k]} & \leqslant & \bar{\beta}_{i[k]} v_i ((n_{i[k]} + \epsilon \Delta n_{i[k]}) + \gamma (r_{i[k]} + \epsilon \Delta r_{i[k]})) &
\end{array}
$$

The same can be done for the remaining terms. We conclude that $\psi + \epsilon \overline{\mathbf{\Delta}}_{\imath\kappa} \in \Omega_{\mathscr{L}}$. $\overline{\mathbf{\Delta}}_{\imath\kappa}$ is therefore a feasible perturbation for every $\psi \in \Gamma_{\imath\kappa}$.

We will next show that $\overline{\mathbf{\Delta}}_{\imath\kappa}$ is also improving. Fig. 5 illustrates the propagation of $\overline{\mathbf{\Delta}}_{\imath\kappa}$ in space and time. At time $k = \kappa$, all Δn's are zero. The unit increase in $f_{\imath[\kappa]}$ produces at time $\kappa + 1$ an increase in density downstream and a decrease in density upstream. It is shown in Appendix C that the effect of this initial pulse does not propagate downstream beyond section $\imath + 1$ or upstream beyond section $\imath$. It is also shown in Appendix C that the non-zero components of $\overline{\mathbf{\Delta}}_{\imath\kappa}$ have the following closed forms whenever the split ratios are constant in time:

$$
\Delta f_{\imath[\kappa+X]} = \begin{cases} 1 & X = 0 \\ -m_\imath (1 - m_\imath)^{X-1} & X > 0 \end{cases} \tag{47}
$$

$$
\Delta n_{\imath[\kappa+X]} = -\frac{1}{\bar{\beta}_\imath} (1 - m_\imath)^{X-1}, \quad X > 0 \tag{48}
$$

$$
\Delta n_{\imath+1[\kappa+X]} = (1 - m_\imath)^{X-1}, \quad X > 0 \tag{49}
$$

where $m_\imath \triangleq \max\{v_\imath; w_{\imath+1}\}$. Eq. (47)–(49) can be used to calculate $J(\overline{\mathbf{\Delta}}_{\imath\kappa})$:

$$
\begin{aligned}
J(\overline{\mathbf{\Delta}}_{\imath\kappa}) &= \sum_{X=1}^{K-\kappa} (\Delta n_{\imath[\kappa+X]} + \Delta n_{\imath+1[\kappa+X]}) - \eta \sum_{X=0}^{K-\kappa-1} \Delta f_{\imath[\kappa+X]} \\
&= \sum_{X=1}^{K-\kappa} \left(-\frac{1}{\bar{\beta}_\imath} (1 - m_\imath)^{X-1} + (1 - m_\imath)^{X-1} \right) - \eta \left(1 + \sum_{X=1}^{K+\kappa-1} (-m_\imath)(1 - m_\imath)^{X-1} \right) \\
&= \left(1 - \frac{1}{\bar{\beta}_\imath} \right) \sum_{X=1}^{K-\kappa} (1 - m_\imath)^{X-1} - \eta \left(1 - m_\imath \sum_{X=1}^{K+\kappa-1} (1 - m_\imath)^{X-1} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(1-\frac{1}{\overline{\beta}_\imath}\right)\frac{1-(1-m_\imath)^{K-\kappa}}{1-(1-m_\imath)} - \eta\left(1 - m_\imath\frac{1-(1-m_\imath)^{K-\kappa-1}}{1-(1-m_\imath)}\right)\\
&= \left(1-\frac{1}{\overline{\beta}_\imath}\right)\frac{1-(1-m_\imath)^{K-\kappa}}{m_\imath} - \eta\left(1-(1-(1-m_\imath)^{K-\kappa-1})\right)\\
&= \underbrace{-\frac{\beta_\imath}{1-\beta_\imath}\frac{1-(1-m_\imath)^{K-\kappa}}{m_\imath}}_{<0 \text{ whenever } \beta_\imath>0}\underbrace{-\eta(1-m_\imath)^{K-\kappa-1}}_{<0 \text{ whenever } m_\imath<1}
\end{aligned} \tag{50}
$$

$J(\overline{\mathbf{\Delta}}_{\imath\kappa})$ is strictly negative for all sections $\imath$ and time periods κ, whenever every section has either $\beta_\imath > 0$ (i.e. is an offramp) or $m_\imath < 1$ (i.e. $v_\imath < 1$ and $w_{\imath+1} < 1$) . Under these conditions we have found a feasible and improving perturbation for all $\psi \in \Omega_{\mathscr{L}} \setminus \Omega_{\mathscr{M}}$, and thus shown that any solution to $\mathscr{L}$ must lie on $\Omega_{\mathscr{M}}$, and therefore also solve $\mathscr{M}$. □

The role of the total travel distance in the objective function can be appreciated in Eq. (50): the positive η provides an incentive for the solution of $\mathscr{L}$ to seek the upper boundary of $\Omega_{\mathscr{L}}$ in offramp-less sections. Total travel time alone is not sufficient, as is demonstrated in the following discussion. Fig. 6 shows a freeway divided into three regions: A is upstream of an offramp, B is between the offramp and a bottleneck, and C is downstream of the bottleneck. In each case we will consider how the solution to $\mathscr{L}$ might behave when $\eta = 0$, that is, when minimizing travel time is the only objective.

The situation in region C is simple. Vehicles downstream of the bottleneck must travel as fast as possible in order to reach the downstream boundary quickly and register the smallest possible travel time. Minimizing total travel time is sufficient in this case for the solution to rise to the upper boundary of $\Omega_{\mathscr{L}}$.

Vehicles in B however gain nothing by moving quickly, since their release rate from the bottleneck is determined by the bottleneck capacity. A vehicle in region B may choose to slow down temporarily, allow a gap to open, and then catch up without impacting the total travel time, as long as the downstream discharge and upstream offramp flows are not affected. The additional incentive needed to prevent this behavior is supplied by the total travel distance.

Vehicles in region A, as in C, minimize their travel times by exiting the freeway as quickly as possible. However, if we now allow the split ratios to vary with time, this no longer translates into faster speeds in region A. A situation could then arise in which it is beneficial for vehicles to slow down in order to catch a larger split ratio at the offramp. If the increase in the split ratio is large, then a large additional incentive (η) will be needed to motivate drivers to travel at their maximum speed. Although it was not shown here, the requirement of constant split ratios in Lemma B can be replaced with split ratios whose rate of change is bounded by a term that depends on η.

How can adding a constant TTD to the objective function affect the optimum? The answer to this question is that TTD is not a predetermined constant while there remain vehicles on the freeway. It converges asymptotically to the value calculated in Appendix B as the density on the freeway approaches zero. But under the assumption that $\beta_i < 1$ or $v_i < 1$, an absolutely empty freeway can never be attained. Hence, the sequence of solutions obtained by solving $\mathscr{L}$ with progressively longer cool-down periods, and therefore emptier final conditions, approximates the global minimizer of travel time.

The following theorem summarizes the conclusions of this section.

Theorem B. *A solution to $\mathscr{N}$ can be found by solving $\mathscr{L}$ whenever*:

1. *Each $\mathscr{L}$-optimal $r_{i[k]}$ is less than $\xi_i(\bar{n}_i - n_i[k])$*,
2. $\underline{c}_i = 0$,

Fig. 6. Freeway section.

3. *the split ratios are constant in time, and*
4. *all offramp-less sections have* $v_i < 1$ *and* $w_{i+1} < 1$.

Furthermore, the solution approaches a global minimizer of total travel time as the final condition approaches an empty freeway.

5. Numerical experiments

Theorem B suggests an efficient method for generating onramp metering plans that are near-global minimizers of total travel time, according to the ACTM. The method was tested numerically using the geometric layout and traffic demands from a 22 km stretch of Interstate 210 in Pasadena, California. This site contains 20 metered onramps and a single uncontrolled freeway connector from I-605, and has been studied extensively in (Muñoz et al., 2004; Gomes et al., 2004). The parameters of the ACTM $(v_i, w_i, \xi_i, \bar{f}_i, \bar{s}_i, \bar{n}_i, \gamma)$ were adjusted to produce a congestion pattern that resembles the morning peak period on I-210, between 5 a.m. and 10 a.m. Fig. 7 shows the simulated speed contour plot without onramp metering. The speed variable was calculated with:

$$\text{speed}_{i[k]} \triangleq \frac{f_{i[k]}/\bar{\beta}_{i[k]}}{n_{i[k]} + \gamma r_{i[k]}} \left(\frac{L_i}{\Delta t}\right)$$

This formula produces $\text{speed}_{i[k]} = 100$ kph when the freeway is free flowing. The two darker shades indicate speeds below 85 kph and 65 kph.

Problem $\mathscr{L}$ was solved for 10 time horizons ranging from 30 min to 5 h. In all cases an additional 1-h cooling period was appended. The commercial LP solver MOSEK 3.0 was used to generate the solutions. Each of the 10 time horizons was solved with and without onramp queue length constraints, for a total of 20 exper-

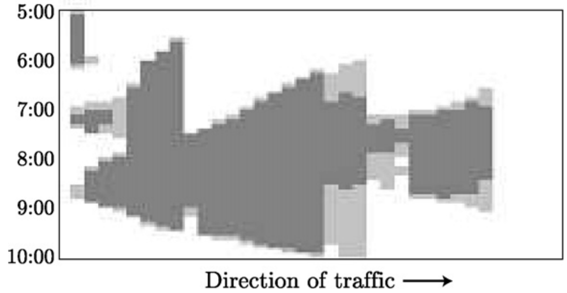

Fig. 7. Uncontrolled speed contour map.

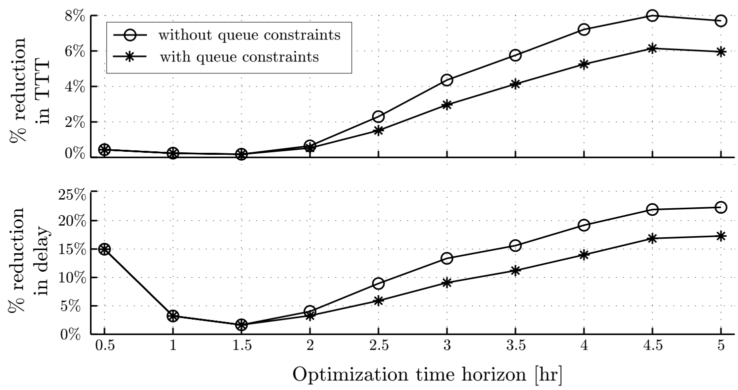

Fig. 8. Travel time delay reductions.

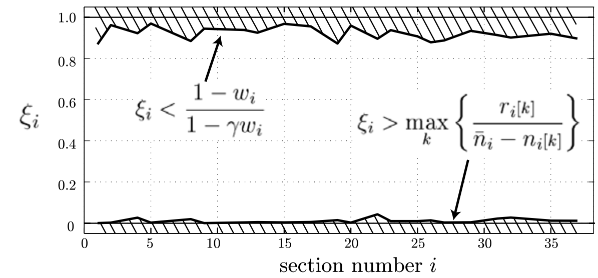

Fig. 9. Range of valid ξ_i values.

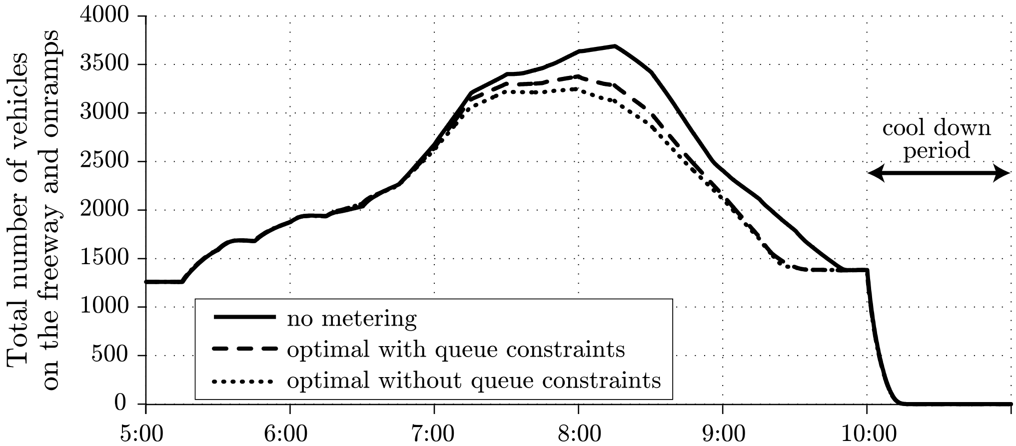

Fig. 10. Total freeway density.

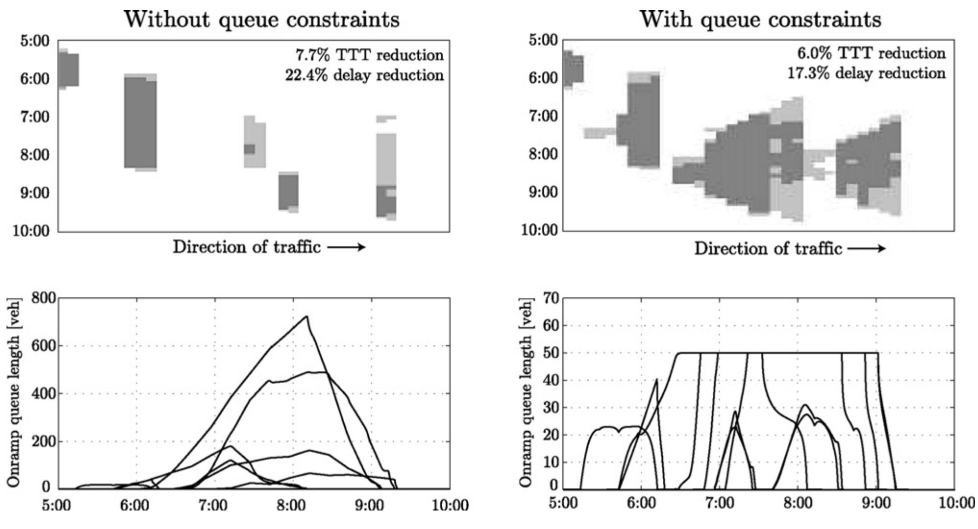

Fig. 11. Optimized speeds and queue lengths.

iments. The size of the LP ranged from 85,860 constraints and 64,800 variables for the 30-minute problem to 343,440 constraints and 259,200 variables for the 5-h problem. Resulting percent reductions in TTT and delay (time spent in congestion) are plotted in Fig. 8.

It was confirmed in every case that the solution satisfies the equations of the traffic model to a high degree of precision ($\psi \in \Omega_{\mathscr{M}}$). It was also verified that the optimal onramp flows never exceeded

$\xi_i(\bar{n}_i - n_{i[k]})$ for the chosen values of ξ_i. Fig. 9 shows the upper and lower limits on the acceptable values of ξ_i stemming from the requirements of theorems A and B. The final density on the freeway, after the cooling period, was found to be extremely small (see Fig. 10), indicating that the solution is very close to the global optimum.

Optimized speed contour plots and onramp queues for the two 5-h trials are shown in Fig. 11. Note that the optimal strategy without queue constraints is to hold a large number of vehicles (over 700 in one case) on a few onramps, in order to keep the freeway almost completely uncongested. Delay is reduced by 22.4%, but at the expense of the drivers using those onramps.

The 5-h trial with queue constraints demonstrates that it is not possible to maintain freeflow conditions when the onramp queues are limited to 50 vehicles each. Any optimization technique that assumes freeflow conditions while constraining the onramp queues would have failed in this scenario. Travel time can only be reduced by shortening, but not eliminating, the period of time during which offramps are obstructed by mainline congestion. The task of the optimizer is therefore to distribute the control burden among several onramps, and to coordinate the accumulation and release of the onramp queues so as to minimize congestion. Despite this added complication, the optimizer is able to reduce delay by 17.3%.

6. Conclusions

The goal of this paper has been to develop an efficient method for computing optimal ramp metering plans for congested freeways. The starting point was the intuition that a search of the convex region lying below the fundamental diagram might yield a global optimum. The ACTM was developed to pursue this notion. This model is similar to the cell transmission model, except in the case of merging flows where two allocation parameters are used instead of one. Theorem A showed that this modification does not destroy an essential property of the model.

Theorem B provided sufficient conditions under which the optimal ramp metering problem can indeed be solved with a linear program. One of the conditions of the theorem requires that the inlet flows should not be obstructed by congestion on the mainline whenever the freeway is optimally metered. This "after the fact" requirement might seem at first to render the theorem useless; once the LP is solved, it is just as easy to check the conclusion of the theorem as the conditions. The theorem is useful nevertheless because it identifies the main reasons why the two problems are sometimes not equivalent. One of these is the tradeoff between onramp and mainline flows that arises when congestion backs into the onramps. In this situation, a positive perturbation to mainline flow produces a decrease in onramp flow, which in turn causes the onramp queue to grow. This perturbation is not feasible if the onramp queue is already full. The numerical example showed that this requirement did not disqualify a very congested test site. The values of ξ_i used in the example ranged from 0.14 to 0.18, and were sufficiently large to avoid the mainline/onramp conflict.

The second condition of the theorem, $\underline{c}_i = 0$, implies that the controller must be allowed to completely shut down the onramps. This is truly unrealistic, most ramp metering systems maintain a minimum metering rate of around 240 vph (1 vehicle every 15 s). In order to implement the optimal plan, the rates must be increased to at least 240 vph. It was found that about 42% of the optimal rates in the 5-h experiment with queue constraints were less than 240 vph. An "implementable" plan in which these values were simply replaced with 240 vph was simulated, and found to reduce delay by 12.3%, a sacrifice of 5% points. Future research will focus on finding better strategies for generating an implementable plan. Other future directions include the use of the technique in a rolling horizon framework, subject to uncertainties, its performance compared to other methods, and the applicability of the basic ideas to higher order freeway models.

Acknowledgement

This research was funded by the California Partners for Advanced Transit and Highways (PATH) under Task Order 4136, "Design, Field Implementation and Evaluation of Adaptive Ramp Metering Algorithms".

Appendix A. Proof of Theorem A

The proof is by induction. Assuming that $n_{i[k]} \in [0, \bar{n}_i]$ and $l_{i[k]} \geqslant 0$ holds for some k and all i, we show that $f_{i[k]} \geqslant 0$ and $r_{i[k]} \geqslant 0$. We then show that this implies $n_{i[k+1]} \in [0, \bar{n}_i]$ and $l_{i[k+1]} \geqslant 0$. Because $n_{i[k]} \in [0, \bar{n}_i]$ and $l_{i[k]} \geqslant 0$ holds for $k = 0$, the result follows.

First, from Eq. (11), with $l_{i[k]} \geqslant 0, d_{i[k]} \geqslant 0, \xi_i \geqslant 0, n_{i[k]} \leqslant \bar{n}_i; c_{i[k]} \geqslant 0$, it follows that $r_{i[k]} \geqslant 0$. To show $f_{i[k]} \geqslant 0$, we need to check that each of the four terms in Eq. (7) is positive. The only non-obvious one is the second. However, since both ξ_{i+1} and $\gamma \in [0, 1]$:

$$
\begin{aligned}
\gamma\xi_{i+1} \leqslant 1 &\Rightarrow \gamma\xi_{i+1}(\bar{n}_{i+1} - n_{i+1[k]}) \leqslant (\bar{n}_{i+1} - n_{i+1[k]}) \\
&\Rightarrow \gamma r_{i+1[k]} \leqslant (\bar{n}_{i+1} - n_{i+1[k]}) \quad \ldots \text{from Eq. (9)} \\
&\Rightarrow 0 \leqslant w_{i+1}(\bar{n}_{i+1} - n_{i+1[k]} - \gamma r_{i+1[k]})
\end{aligned}
$$

Therefore, $f_{i[k]} \geqslant 0 \ \forall i \in \mathscr{I}$. Using the above, we can deduce $l_{i[k+1]} \geqslant 0$ and $n_{i[k+1]} \in [0, \bar{n}_i]$:

$$
\begin{aligned}
l_{i[k+1]} &= l_{i[k]} + d_{i[k]} - r_{i[k]} \\
&\geqslant l_{i[k]} + d_{i[k]} - (l_{i[k]} + d_{i[k]}) \quad \ldots \text{from Eq. (8)} \\
&\geqslant 0
\end{aligned}
$$

$$
\begin{aligned}
n_{i[k+1]} &= n_{i[k]} + f_{i-1[k]} - f_{i[k]}/\bar{\beta}_{i[k]} + r_{i[k]} \\
&\geqslant n_{i[k]} - f_{i[k]}/\bar{\beta}_{i[k]} + r_{i[k]} \\
&\geqslant n_{i[k]} - \bar{\beta}_{i[k]} v_i (n_{i[k]} + \gamma r_{i[k]})/\bar{\beta}_{i[k]} + r_{i[k]} \quad \ldots \text{from Eq. (7)} \\
&\geqslant (1 - v_i) n_{i[k]} + (1 - \gamma v_i) r_{i[k]} \\
&\geqslant 0
\end{aligned}
$$

$$
\begin{aligned}
n_{i[k+1]} &= n_{i[k]} + f_{i-1[k]} - f_{i[k]}/\bar{\beta}_{i[k]} + r_{i[k]} \\
&\leqslant n_{i[k]} + f_{i-1[k]} + r_{i[k]} \\
&\leqslant n_{i[k]} + w_i(\bar{n}_i - n_{i[k]} - \gamma r_{i[k]}) + r_{i[k]} \quad \ldots \text{from Eq. (7)} \\
&\leqslant (1 - w_i) n_{i[k]} + r_{i[k]}(1 - \gamma w_i) + w_i \bar{n}_i \\
&\leqslant (1 - w_i) n_{i[k]} + \xi_i(\bar{n}_i - n_{i[k]})(1 - \gamma w_i) + w_i \bar{n}_i \quad \ldots \text{from Eq. (9)} \\
&\leqslant \begin{cases} (1 - w_i) n_{i[k]} + w_i \bar{n}_i & \text{if } i \notin \mathscr{E} \\ (1 - \bar{w}_i) n_{i[k]} + \bar{w}_i \bar{n}_i & \text{if } i \in \mathscr{E} \end{cases} \\
&\leqslant \bar{n}_i
\end{aligned}
$$

where $\bar{w}_i \triangleq w_i + \xi_i(1 - \gamma w_i)$. The last line holds since, by assumption, both w_i and $\bar{w}_i$ are in $[0, 1]$. $\left(\bar{w}_i \in [0, 1]\right.$ follows from $\left.\xi_i \in \left[0, \frac{1-w_i}{1-\gamma w_i}\right]\right)$.

Appendix B. TTD is independent of the metering rates

The following equations result from summing the mainline and onramp conservation Eqs. (12) and (13) over time:

$$
\sum_{k=0}^{K-1} (f_{i-1[k]} + r_{i[k]} - s_{i[k]} - f_{i[k]}) = n_{i[K]} - n_{i[0]}
$$

$$
\sum_{k=0}^{K-1} (d_{i[k]} - r_{i[k]}) = l_{i[K]} - l_{i[0]}
$$

Using $n_{i[K]} = l_{i[K]} = 0$ and $\beta_{i[k]}$ constant, these become

$$\sum f_{i-1[k]} + \sum r_{i[k]} - \frac{1}{\bar{\beta}_i} \sum f_{i[k]} + n_{i[0]} = 0$$
$$\sum d_{i[k]} - \sum r_{i[k]} + l_{i[0]} = 0$$

where $\sum$ denotes a sum over all time intervals. Then,

$$\sum f_{i-1[k]} = \frac{1}{\bar{\beta}_i} \sum f_{i[k]} - p_i$$
$$\sum r_{i[k]} = \sum d_{i[k]} + l_{i[0]}$$

where $p_i \triangleq \sum d_{i[k]} + l_{i[0]} + n_{i[0]}$. The sequence $\sum f_{i[k]}$ can be solved by using the boundary condition $\sum f_{-1[k]} = 0$:

$$\sum f_{i[k]} = \sum_{q=0}^{i} \left(p_q \prod_{r=q}^{i} \bar{\beta}_r \right)$$

TTD can then be computed using only given data:

$$\mathtt{TTD} = \sum_{i=0}^{I-1} \left[\sum_{q=0}^{i} \left(p_q \prod_{r=q}^{i} \bar{\beta}_r \right) + \sum_{k=0}^{K-1} d_{i[k]} + l_{i[0]} \right]$$

Appendix C. Closed form for $\overline{\mathbf{\Delta}}_{\imath\kappa}$

We wish to show that the components of $\overline{\mathbf{\Delta}}_{\imath\kappa}$ are given by Eqs. (47)–(49). For times up to $\kappa, \overline{\mathbf{\Delta}}_{\imath\kappa}$ evolves identically to the ACTM with zero initial conditions, zero demands, and $\bar{n}_i = F_{i[k]} = 0$. Thus, all of its components prior to κ are zero. At time κ, $\Delta f_\imath[\kappa] = 1$ is introduced, which affects densities in sections $\imath$ and $\imath + 1$ at time $\kappa + 1$:

$$\Delta n_{\imath[\kappa+1]} = \Delta n_{\imath[\kappa]} + \Delta f_{\imath-1[\kappa]} - \frac{1}{\bar{\beta}_\imath} \Delta f_{\imath[\kappa]} = -\frac{1}{\bar{\beta}_\imath}$$
$$\Delta n_{\imath+1[\kappa+1]} = \Delta n_{\imath+1[\kappa]} + \Delta f_{\imath[\kappa]} - \frac{1}{\bar{\beta}_\imath + 1} \Delta f_{\imath+1[\kappa]} = 1$$

Then,

$$\begin{aligned} \Delta f_{\imath[\kappa+1]} &= \min\{\bar{\beta}_\imath v_\imath \Delta n_{\imath[\kappa+1]}; -w_{\imath+1} \Delta n_{\imath+1[\kappa+1]}; 0\} \\ &= \min\{\bar{\beta}_\imath v_\imath (-1/\bar{\beta}_\imath); -w_{\imath+1}(1); 0\} \\ &= -\max\{v_\imath; w_{\imath+1}\} \\ &= -m_\imath \end{aligned}$$

$$\begin{aligned} \Delta f_{\imath-1[\kappa+1]} &= \min\{\bar{\beta}_{\imath-1} v_{\imath-1} \Delta n_{\imath-1[\kappa+1]}; -w_\imath \Delta n_{\imath+1[\kappa]}; 0\} \\ &= \min\{0; -w_\imath(-1/\bar{\beta}_\imath); 0\} \\ &= 0 \end{aligned}$$

$$\begin{aligned} \Delta f_{\imath+1[\kappa+1]} &= \min\{\bar{\beta}_{\imath+1} v_{\imath+1} \Delta n_{\imath+1[\kappa+1]}; -w_{\imath+2} \Delta n_{\imath+2[\kappa+1]}; 0\} \\ &= \min\{\bar{\beta}_{\imath+1} v_{\imath+1}(1); 0; 0\} \\ &= 0 \end{aligned}$$

We have verified Eqs. (47)–(49) with $X=1$. Also that $\Delta f_{\imath-1[\kappa+X]}=\Delta f_{\imath+1[\kappa+X]}=0$, with $X=1$. The proof is completed by induction.

$$\begin{aligned}
\Delta n_{\imath[\kappa+X+1]} &= \Delta n_{\imath[\kappa+X]} + \Delta f_{\imath-1[\kappa+X]} - \frac{1}{\bar{\beta}_\imath}\Delta f_{\imath[\kappa+X]} \\
&= -\frac{1}{\bar{\beta}_\imath}(1-m_\imath)^{X-1} + 0 + \frac{1}{\bar{\beta}_\imath} m_\imath (1-m_\imath)^{X-1} \\
&= -\frac{1}{\bar{\beta}_\imath}(1-m_\imath)^{X}
\end{aligned}$$

$$\begin{aligned}
\Delta n_{\imath+1[\kappa+X+1]} &= \Delta n_{\imath+1[\kappa+X]} + \Delta f_{\imath[\kappa+X]} - \frac{1}{\bar{\beta}_{\imath+1}}\Delta f_{\imath+1[\kappa+X]} \\
&= (1-m_\imath)^{X-1} - m_\imath(1-m_\imath)^{X-1} - 0 \\
&= (1-m_\imath)^{X}
\end{aligned}$$

$$\begin{aligned}
\Delta f_{\imath[\kappa+X+1]} &= \min\{\bar{\beta}_\imath v_\imath \Delta n_{\imath[\kappa+X+1]}; -w_{\imath+1}\Delta n_{\imath+1[\kappa+X+1]}; 0\} \\
&= \min\{-\bar{\beta}_\imath v_\imath (1-m_\imath)^X/\bar{\beta}_\imath; -w_{\imath+1}(1-m_\imath)^X; 0\} \\
&= -(1-m_\imath)^X \max\{v_\imath; w_{\imath+1}\} \\
&= -m_\imath(1-m_\imath)^X
\end{aligned}$$

$$\begin{aligned}
\Delta f_{\imath-1[\kappa+X+1]} &= \min\{\bar{\beta}_{\imath-1} v_{\imath-1} \Delta n_{\imath-1[\kappa+X+1]}; -w_\imath \Delta n_{\imath[\kappa+X+1]}; 0\} \\
&= \min\{0; w_\imath (1-m_\imath)^X/\bar{\beta}_\imath; 0\} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\Delta f_{\imath+1[\kappa+X+1]} &= \min\{\bar{\beta}_{\imath+1} v_{\imath+1} \Delta n_{\imath+1[\kappa+X+1]}; -w_{\imath+2} \Delta n_{\imath+2[\kappa+X+1]}; 0\} \\
&= \min\{\bar{\beta}_{\imath+1} v_{\imath+1}(1-m_\imath)^X; 0; 0\} \\
&= 0
\end{aligned}$$

References

Banks, J., 2000. Are minimization of delay and minimization of freeway congestion compatible ramp metering objectives? Transportation Research Record (1727).

Bellemans, T., De Schutter, B., De Moor, B., 2003. Anticipative model predictive control for ramp metering in freeway networks. In: Proceedings of the American Control Conference.

Daganzo, C., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B 28 (4).

Daganzo, C., 1995. The cell transmission model, Part II: Network traffic. Transportation Research Part B 29 (2).

Erera, A., Daganzo, C., Lovell, D., 1999. The access control problem on capacitated FIFO networks with unique OD paths is hard. California PATH Program. Institute of Transportation Studies, UCB-ITS-PRR-99-35.

Gomes, G., May, A.D., Horowitz, R., 2004. Congested freeway microsimulation model using Vissim. Transportation Research Record: Journal of the Transportation Research Board, No. 1876.

Hegyi, A., De Schutter, B., Hellendoorn, H., Van Den Boom, T., 2002. Optimal coordination of ramp metering and variable speed control – an MPC approach. In: Proceedings of the American Control Conference.

Iida, Y., Hasegawa, T., Asakura, Y., Shao, C., 1989. A formulation of on-ramp traffic control system with route guidance for urban expressway. Communications in Transportation.

Kotsialos, A., Papageorgiou, M., Mangeas, M., Hadj-Salem, H., 2002. Coordinated and integrated control of motorway networks via nonlinear optimal control. Transportation Research Part C 10 (1).

Messmer, A., Papageorgiou, M., 1990. METANET: a macroscopic simulation program for motorway networks. Traffic Engineering and Control (31).

Muñoz, L., Sun, X., Sun, D., Gomes, G., Horowitz, R., Alvarez, L., 2004. Methodological calibration of the cell transmission model. In: Proceedings of the American Control Conference.

Papageorgiou, M., 1980. A new approach to time-of-day control based on a dynamic freeway traffic model. Transportation Research Part C 14B (4), 349–360.

Papageorgiou, M., 1995. An integrated control approach for traffic corridors. Transportation Research Part C 3 (1).

Payne, H., Thompson, W., 1974. Allocation of freeway ramp metering volumes to optimize corridor performance. IEEE Transactions on Automatic Control (19).

Schrank, D., Lomax, T., 2005. The 2005 Urban Mobility Report. Texas Transportation Institute.

Wang, J., May, A.D., 1973. Computer model for optimal freeway on-ramp control. Highway Research Record (469).

Wattleworth, J., Berry, D., 1965. Peak-period control of a freeway system – some theoretical investigations. Highway Research Record (89).

Yuan, L., Kreer, J., 1971. Adjustment of freeway ramp metering rates to balance entrance ramp queues. Transportation Research (5).

Zhang, L., Levinson, D., 2004. Optimal freeway ramp control without origin–destination information. Transportation Research Part B 38 (10).

Zhang, M., Ritchie, S., Recker, W., 1996. Some general results on the optimal ramp control problem. Transportation Research Part C 4 (2).

Ziliaskopoulos, A., 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. Transportation Science 34 (1).