Low-Complexity Ramp Metering for Freeway Congestion Control via Network Utility Maximization

Negar Mehr[1], Roberto Horowitz[2] and Ramtin Pedarsani[3]

***Abstract*— In this paper, we present a novel framework for freeway ramp metering that is based on maximizing the aggregate utility of onramp flows. We show how solving the dual problem of maximizing the network utility via a gradient projection algorithm synthesizes a low-complexity control law that is simple enough to be implemented on real platforms, while being robust to measurement noises. Our control algorithm can be partially distributed at each time step, every onramp selects a traffic flow to maximize its own benefit, and the network adjusts unit traffic flow prices for different onramps. We provide theoretical guarantees on the convergence of our algorithm under mild technical assumptions. We further demonstrate the practicality of our method in a case study where the state of the art controls fail (due to infeasibility) and introduce multiple interesting future directions.**

I. Introduction

As the expeditious growth in traffic congestion has led to a significant rise in fuel consumption, air pollution and delay, the key role of traffic management and traffic control is getting more prominent. For the task of control design, researchers divide traffic control into control of urban arterials and freeway traffic control, where the latter is the focus of this work. In order to ameliorate traffic congestion in freeways, it is demonstrated that ramp metering is an effectual strategy [4]. Ramp metering refers to dictating the input vehicular flows to the freeway mainlines through its onramps such that the appearance of congestion is evaded, or the throughput of the network is maximized.

In this paper, we define a novel methodology for controlling congestion in freeway networks. We employ the idea of network utility maximization, which is a well known and powerful congestion control scheme in communication networks [7], [12], [11], [16], [20], [19], for enhancing freeway traffic conditions. We model freeway through popular Asymmetric Cell Transmission [3] model, and formulate a convex optimization problem seeking for maximizing the network utility. We construct the dual problem and show that by solving the dual problem via a gradient projection algorithm at every time step, our algorithm can successfully balance network flows. Further, our optimization problem is formulated such that solving the dual problem leads to a control algorithm that can be partially distributed among freeway onramps making it appropriate for large–scale freeway traffic control. We also formally prove that under mild technical assumptions, using our control law, vehicular flows successfully converge to the optimal flows where the network utility is maximized. Our approach is different from Model Predictive Control in the sense that our control algorithm is a *reactive* control rather than having anticipatory action, which is actually the key for its simplicity. In particular, the nature of our algorithm is similar to that of [2]: If the arrivals are such that the network is undersaturated, the system converges to an equilibrium where the onramp queues disappear, and the freeway is not congested. However, the more interesting regime is when the network is oversaturated, and one needs to determine the onramp flows such that the freeway remains uncongested, while the network resources (freeway capacity) is shared *fairly* among the onramps.

In addition to the afforementioned properties, we show that our approach is robust to measurement noise which is an inseparable component of traffic sensors' data.

In summary, our contributions encompass the followings:

- We introduce a novel framework for traffic control in freeways which can be further extended to other type of traffic networks. Our proposed scheme that is based on maximizing network's utility can potentially introduce multiple future research directions in the context of traffic control and pricing.
- We provide theoretical guarantees on the convergence of our control algorithm under certain assumptions.
- We demonstrate how our algorithm can be distributed such that each onramp can update its own onramp flow while improving the overall utility of the network.
- We also prove the robustness of our algorithm to measurement noises.
- We showcase the effectiveness of our method in balancing freeway flows by simulating realistic freeway control scenarios; in particular, we show how our algorithm can allocate freeway space fairly to the onramps while the state of the art control fails in achieving that.

II. prior work

The very first instance of ramp metering controllers is fixed–time controllers [18] that only allow vehicular flow to enter freeway in a fixed proportion of cycle times, which requires determination of the green time a–priori, independent of traffic conditions. A popular widely-employed controller known as ALINEA [17] acts as a *local* feedback controller that regulates vehicular density downstream of each onramp around its critical density. In [6], ramp metering is viewed as

[1]Negar Mehr is with the Mechanical Engineering Department, University of California, Berkeley, 410 McLaughlin Hall, Berkeley, CA, USA negar.mehr@berkeley.edu

[2]Roberto Horowitz is with Mechanical Engineering Department of University of California, Berkeley, 6143 Etcheverry Hall, Berkeley, CA, USA horowitz@me.berkeley.edu

[3]Ramtin Pedarsani is with the Deparment of Electrical and Computer Engineering of University of California, Santa Barbara, 3159 Harold Frank Hall, Santa Barbara, CA, USA ramtin@ece.ucsb.edu

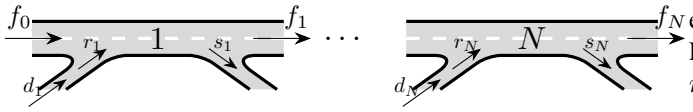


Fig. 1: Freeway Segments

tracking a desired output while rejecting disturbances assuming that a desired freeway trajectory is fed to the controller. There is also a wide range of model predictive controls optimizing for either a performance metric of interest (the expectation of the performance metric in the presence of random arrivals) or forcing the freeway to converge to an uncongested equilibrium [3], [5], [14], [9], [13]. Nonetheless, as freeway networks are normally large–scale, these MPC controllers are hard to implement in real time.

Recently, temporal logic tools have also been utilized for traffic control problems ranging from synthesis through Linear Temporal Logic Specifications [1] to Model Predictive Control with Signal Temporal Logic Specifications [21], [15]. Also, compositional synthesis of such controllers is addressed in [8]. However, such approaches require encoding desired properties of the system as temporal logic specifications. In addition to that, in online MPC designs with temporal specifications can only be encoded using mixed–integer constraints leading to complexity of the optimization problem Our approach is significantly different from the existing works in the literature not only in its simplicity but also in its capability of dealing with oversaturated traffic regimes which is actually the scenario when the MPC laws and temporal–logic–based controls fail as their formulated optimization problem cannot handle oversturated arrivals. In this paper, we describe how a simple ramp metering control, competent of capturing fairness among onramps, can be constructed.

III. Notation

Before we proceed, we need to describe the notation we adopt in this paper. We let $\mathbb{R}^n_+ = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x}_i \geq 0\}$ to be the set of n dimensional vectors with non–zero elements. Vectors are denoted by lower case bold letters. The i_{th} element of a vector $\mathbf{x}$ is shown by x_i. In order to distinguish matrices from vectors, we show matrices with upper case bold letters such as $\mathbf{A}$. We show random variables by upper case letters X, and random vectors by bold upper case letters $\mathbf{X}$.[1] Also, for any arbitrary vector $\mathbf{x}$, let $[\mathbf{x}]_S$ denote the convex projection of the vector $\mathbf{x}$ onto a set S. Furthermore, $[\mathbf{x}]_+$ represents the positive part of the elements of $\mathbf{x}$, and $[x]_a^b \triangleq \min\big(\max(x,a),b\big)$.

IV. Freeway Modeling

In order to model freeway dynamics, we use Asymmetric Cell Transmission Model (ACTM) [3], which is a first order model that integrates ramp queues and mainline densities. Consider the freeway to be divided into N segments such that each segment i has at most one onramp and one offramp (See Figure 1). For segment $i, \forall 1 \leq i \leq N$, its mainline density $n_i(k)$ and ramp queue $l_i(k)$ are the states of the segment i, defined as:

- $n_i(k)$: Number of vehicles stored in segment i at time step k.
- $l_i(k)$: Number of vehicles queued on the onramp of segment i at time step k.

The available control inputs at each time step k, are the onramp flows, $r_i(k)$.Once the onramp flows at time step k are decided, they will lead to mainline vehicular flows leaving each segment to its downstream segment denoted by $f_i(k)$. Note that a proportion of vehicular flows might leave a segment through its offramp. We let $s_i(k)$ denote the flow of vehicles leaving segment i through its offramp.

In order to describe how $r_i(k)$ is mapped to $f_i(k)$, we need to define the following network parameters:

- v_i : Normalized free–flow speed of vehicles in segment i.
- w_i : Normalized congestion wave speed in segment i.
- β_i : Split ratio of segment i defined as the fraction of vehicles leaving segment i through its offramp.
- $\bar{n}_i$: Jam density of segment i which is the maximum number of vehicles that segment i can accommodate.
- $\bar{f}_i$: Capacity of segment i, i.e. the maximum number of vehicles that can leave segment i.

Remark 1. *For the offramp–less sections, it is simply assumed that $\beta_i = 0$. Moreover, for the onramp–less sections, it is assumed that no flow is entering freeway through such onramps; hence, their corresponding onramp flow is always set to zero,*

Remark 2. *The model we described here assumes that for each segment, the onramp is located at the beginning of the segment, which is an appropriate assumption as one can always segmentize freeway such that onramps are located at the beginning of the segments.*

Assuming that the freeway is calibrated, implying that the above parameters are available, mainline flows are obtained by:

$$f_i(k) = \min\{\beta_i' v_i n_i(k), w_{i+1}(\bar{n}_{i+1} - n_{i+1}(k)), \bar{f}_i\}, \quad (1)$$

where $\beta_i' = 1 - \beta_i$. Equation (1) indicates that the flow of each segment is restricted by the number of vehicles available in segment i, the available downstream supply and the maximum possible flow. Note that the boundary segments of the freeway are exceptions. For the last segment where there is no downstream segment, we have

$$f_N(k) = \min\{\bar{\beta}_N v_N n_N(k), \bar{f}_N\}. \quad (2)$$

Also, for the very first upstream segment, $f_0(k)$ is the exogenous arrival entering the network. There exist exogenous arrivals (demands) to the onramps too. For each onramp i, its exogenous arrival is denoted by d_i as depicted in Figure 1.

[1]The difference between matrices and random vectors will be clear from the context.

Now, we can describe the update rule of the system states using the introduced quantities. The dynamics of the system is simply obtained by the mass conservation law:

$$n_i(k+1) = n_i(k) + f_{i-1}(k) + r_i(k) - f_i(k) - s_i(k), \quad (3)$$
$$l_i(k+1) = l_i(k) + d_i(k) - r_i(k). \quad (4)$$

Equations (3), (4) and (1) determine freeway dynamics. The nonlinearity of freeway dynamics arise from the $\min$ operator in Equation (1). In fact, the piecewise affine dynamics of freeway can be viewed as a hybrid system outlining the complication of freeway control design.

V. Control Synthesis

In this section, we provide a description of our ramp metering scheme for freeway congestion control. Before explaining the details, we first provide some intuition for the algorithm. Let $\mathbf{r} = \begin{bmatrix} r_1 & \cdots & r_N \end{bmatrix}^T$ represent the vector of onramp flows determined by the controller, and r_i be the flow of the i_{th} onramp. We consider the freeway as a network of links with certain capacities that is shared by a set of exogenous arrivals. Arrival i is characterized by a *utility function*, $U_i(r_i)$ that is a strictly concave and increasing function of the vehicular flow r_i. Our high level goal is to determine how the network resources should be fairly distributed among different arrivals. To this end, we want to maximize the sum of the utilities $\sum_i U_i(r_i)$ subject to capacity constraints such that the freeway is not congested. To synthesize a simple control algorithm, we propose to solve our optimization problem in a distributed manner using a gradient projection algorithm for the dual problem, so that one obtains a control policy with no complex coordination among different onramps while the algorithm is able to adapt to changes in time-varying network conditions. At a high level, the algorithm operates as follows. At each time step, the network calculates a price p_i for the arrival i for a unit of vehicular flow. The onramp controller then chooses a flow $r_i(p_i)$ that maximizes its own *benefit*:

$$r_i(p_i) = \arg\max_x U_i(x) - p_i x.$$

The algorithm iterates and converges to an optimal price vector that leads to both individual and social optimal solution for maximizing the sum of utilities.

We now explain the details of the algorithm. For each time step, we define the following:

$$\begin{aligned} \tilde{f}_1(k) &= (\bar{\beta}_1)(f_0 + r_1(k)), \\ \tilde{f}_2(k) &= (\bar{\beta}_2)(\tilde{f}_1(k) + r_2(k)), \\ &\vdots \\ \tilde{f}_N(k) &= (\bar{\beta}_N)(\tilde{f}_{N-1}(k) + r_N(k)). \end{aligned} \quad (5)$$

Using Equations 5, we want to solve the following optimization problem at every time step, k:

$$\begin{aligned} \underset{\mathbf{r}(k)}{\text{maximize}} \quad & \sum_{i=1}^{N} U_i(r_i(k)) \\ \text{subject to} \quad & \tilde{f}_i(k) \le \bar{f}_i, \ i = 1, \ldots, N \\ & \tilde{f}_i(k) \le w_{i+1}\left(\bar{n}_{i+1} - n_{i+1}(k)\right), \\ & \qquad i = 1, \ldots, N-1. \end{aligned} \quad (6)$$

where U_i, $1 \le i \le N$ is the increasing concave utility function that represents network's utility in sending traffic flow r_i to the freeway at ramp i. Examples of the utility functions include $\log(r_i)$ and r_i^c, $0 < c < 1$. Note that $\tilde{f}_i(k)$'s, indeed, encode the steady state relationship between onramp flows and mainline flows. In other words, if onramp flow vector $\mathbf{r}(k)$ were used in steady state of the freeway, their corresponding steady state mainline flows would have been $\tilde{f}_1(k), \tilde{f}_2(k), \cdots, \tilde{f}_N(k)$. The steady state relations of Equation 5 can be easily derived from Equation 3.

Due to the recursive construction of $\tilde{f}_i$'s, all constraints are solely formulated linearly in terms of the decision variable $\mathbf{r}(k)$ and known quantities (mainline arrival rate f_0, and measured densities, $n_i(k)$'s). As a result, the optimization problem (6) can be written as:

$$\begin{aligned} \underset{\mathbf{r}(k)}{\text{maximize}} \quad & \sum_{i=1}^{N} U_i(r_i(k)) \\ \text{subject to} \quad & \mathbf{A}\mathbf{r}(k) \le \mathbf{b}(k), \end{aligned} \quad (7)$$

where the matrix $\mathbf{A}_{2N-1 \times N}$ and vector $\mathbf{b}_{2N-1 \times 1}$ are defined as follows:

$$\mathbf{A} = \begin{bmatrix} \bar{\beta}_1 & 0 & \cdots & 0 & 0 \\ \bar{\beta}_1 & 0 & \cdots & 0 & 0 \\ \bar{\beta}_1\bar{\beta}_2 & \bar{\beta}_2 & \cdots & 0 & 0 \\ \bar{\beta}_1\bar{\beta}_2 & \bar{\beta}_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ \bar{\beta}_{N-1}\cdots\bar{\beta}_1 & \bar{\beta}_{N-1}\cdots\bar{\beta}_1 & \cdots & \bar{\beta}_{N-1} & 0 \\ \bar{\beta}_{N-1}\cdots\bar{\beta}_1 & \bar{\beta}_{N-1}\cdots\bar{\beta}_1 & \cdots & \bar{\beta}_{N-1} & 0 \\ \bar{\beta}_N\cdots\bar{\beta}_1 & \bar{\beta}_N\cdots\bar{\beta}_1 & \cdots & \bar{\beta}_{N-1}\bar{\beta}_N & \bar{\beta}_N \end{bmatrix},$$

$$\mathbf{b}(k) = \begin{bmatrix} \bar{f}_1 \\ w_2(\bar{n}_2 - n_2(k)) \\ \bar{f}_2 \\ w_3(\bar{n}_3 - n_3(k)) \\ \vdots \\ \bar{f}_{N-1} \\ w_N(\bar{n}_N - n_N(k)) \\ \bar{f}_N \end{bmatrix} - \begin{bmatrix} \bar{\beta}_1 f_0 \\ \bar{\beta}_1 f_0 \\ \bar{\beta}_2\bar{\beta}_1 f_0 \\ \bar{\beta}_2\bar{\beta}_1 f_0 \\ \vdots \\ \bar{\beta}_{N-1}\cdots\bar{\beta}_1 f_0 \\ \bar{\beta}_{N-1}\cdots\bar{\beta}_1 f_0 \\ \bar{\beta}_N\cdots\bar{\beta}_1 f_0 \end{bmatrix}.$$

Since the utility function $\sum_{i=1}^{N} U_i(r_i)$ is a concave function, and the constraints are linear in the decision variable, the optimization problem (7) is a convex problem. Thus, as a substitute, we can solve the dual problem to find the solution of its primal problem. In order to solve the dual, let's construct the Lagrangian:

$$L(\mathbf{r}, \boldsymbol{\alpha}, k) = \sum_i U_i(r_i(k)) - \boldsymbol{\alpha}^T(\mathbf{A}\mathbf{r}(k) - \mathbf{b}(k)), \quad (8)$$

where $\boldsymbol{\alpha}$ is a vector of shadow prices or dual variables corresponding to the linear inequality constraints $\mathbf{Ar} \leq \mathbf{b}$. In order to solve the dual problem, we need to maximize L over $\mathbf{r}$ and minimize it over $\boldsymbol{\alpha}$ while making sure that the shadow prices are always non–negative. We can use gradient descent to find the optimal solution iteratively. Let $\bar{r}$ be a constant denoting an upper bound on the traffic flow that can be sent to the freeway for all onramps. At every time step k, define $\mathcal{C}(k)$ to be the following set:

$$\mathcal{C}(k) = \{\mathbf{r} \in \mathbb{R}^N | \forall\, 1 \leq i \leq N, \\ 0 \leq r_i \leq l_i(k) + d_i(k)\,, r_i \leq \bar{r}\,\}, \quad (9)$$

i.e., $\mathcal{C}(k)$ is the set of all onramp flows which are less than or equal to the number of available vehicles on the onramps and the maximum possible onramp flows. As we will see, we use this set in order to make sure that we are not allowing more vehicles than the actual number of available cars.

Now, we are ready to outline our metering scheme. Let $\{\gamma_k, k \geq 1\}$ be a positive decreasing sequence such that, (i) $\sum_{k=1}^{\infty} \gamma_k = \infty$ and (ii) $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$. Assume that the initial conditions $n_i(0)$ and $l_i(0)$ are given, we can update the shadow prices and onramp flows at every time step in the following manner:

1) Initialize $\mathbf{r}$ and $\boldsymbol{\alpha}$ with an arbitrary $\mathbf{r}(-1)$ and $\boldsymbol{\alpha}(-1) \geq 0$.
2) Measure $n_i(k)$'s and construct the vector $\mathbf{b}(k)$.
3) Update the shadow prices and metering rates by:

$$\boldsymbol{\alpha}(k) = [\boldsymbol{\alpha}(k-1) + \gamma_k(\mathbf{Ar}(k-1) - \mathbf{b}(k-1))]_+, \quad (10)$$

$$\mathbf{r}(k) = \underset{\mathbf{r}}{\operatorname{argmax}} \Big[\big(\sum_{i=1}^{N} U_i(r_i) - \\ (\boldsymbol{\alpha}(k-1))^T(\mathbf{Ar} - \mathbf{b}(k-1))\big) \Big]_{\mathcal{C}(k-1)}, \quad (11)$$

4) Apply the updated metering rates $\mathbf{r}(k)$, and let the model evolve to the next time step.

Interestingly, (11) can be written in a distributed way. Define

$$\tilde{r}_i(k) = \operatorname{argmax}_x\, U_i(x) - \big(\sum_{j=1}^{2N-1} \alpha_j(k)\mathbf{A}_{ji}x\big). \quad (12)$$

Note that the optimization in (11) and (12) is independent of $\mathbf{b}(k)$. So, precisely

$$\tilde{r}_i(k) = (U_i')^{-1}\big(\sum_j \alpha_j(k)\mathbf{A}_{ji}\big), \quad (13)$$

given that $\sum_j \alpha_j(k)\mathbf{A}_{ji}$ is in the range of $U_i(.)$; if not, one can simply project $\sum_j \alpha_j(k)\mathbf{A}_{ji}$ on the range of $U_i(.)$ as it is increasing.

Further, the projection on the convex set $\mathcal{C}$ in (11) can be done in a distributed way since the set is a hypercube, i.e. $r_i(k) = [\tilde{r}_i(k)]_{\mathcal{C}(k)}$. Thus, the onramp flow i can be updated independently from other onramps using the following closed-form formula:

$$r_i(k) = [(U_i')^{-1}\big(\sum_j \alpha_j \mathbf{A}_{ji}\big)]_0^{\min(l_i(k)+d_i(k),\bar{r}_i)} \quad (14)$$

The above update rule has the following economic intuition. Define $p_i \triangleq \sum_j \boldsymbol{\alpha}_j \mathbf{A}_{ji}$ to be the effective price for onramp i to send x units of flow to the freeway. This price is indirectly calculated as a measure of how much onramp i's flow contributes to the congestion of the freeway network. Then, based on its own utility, the onramp's optimal decision is to send $\arg\max_x U_i(x) - px$ (projected on the set $\mathcal{C}$ to obtain feasible flows) to the freeway. Thus, $r_i = [(U_i')^{-1}(p_i)]_{\mathcal{C}}$.

We now present the main theoretical result of the paper. Informally, the result states that our ramp metering algorithm generates a sequence of onramp flows and shadow prices that approach the optimal solution of (6) while the freeway is in the steady state. To formally prove the convergence result, we need the following technical assumptions:

Assumption 1. *The utility functions $U_i(.)$ are increasing, strictly concave, continuously differentiable and bounded in the interval $(0, \bar{r}]$.*

Assumption 2. *In the case of noisy measurements of the densities $n_i(k)$, the noise process, $Z_i(k)$ can be modeled as additive, L_2–bounded, zero-mean and independent noise, for all i, i.e.*

$$\hat{n}_i(k) = n_i(k) + Z_i(k), \quad (15)$$

where $\hat{n}_i(k)$ is the measured density of segment i at time k and $E[Z_i(k)] = 0$, $E[Z_i^2(k)] < \infty$.

Assumption 3. *$\mathbf{b}(k)$ can be modeled as a decomposition of a steady state vector and a bounded and vanishing error vector*

$$\mathbf{b}(k) = \mathbf{b}_s + \mathbf{e}(k), \quad (16)$$

where $\|\mathbf{e}(k)\| < \infty$ for all k, and $\|\mathbf{e}(k)\| \to 0$ as $k \to \infty$.

Assumption 4. *All the exogenous arrivals are constant and time–invariant*

Assumption 5. *The very upstream exogenous arrival f_0 is strictly less than $\bar{f}_1$. (This assumption is valid in most of the real world scenarios, congestions are propagated from downstream to upstream)*

Note that assumption 2 and 4 are required for theoretical proof of the algorithm practicality. It was verified in simulations that relaxing such assumptions does not lead to the degradation in the performance of the algorithm.

Theorem 1. *Suppose that Assumptions 1 through 5 hold.*

(i) Starting from any initial flows $\mathbf{r}(-1)$ and shadow prices $\boldsymbol{\alpha}(-1)$, every accumulation point $(\mathbf{r}^*, \boldsymbol{\alpha}^*)$ of the sequence $(\mathbf{r}(k), \boldsymbol{\alpha}(k))$ generated by the ramp metering algorithm is primal-dual optimal;

(ii) The control algorithm is robust to independent zero-mean measurement noise, i.e., convergence to

primal-dual optimal solution occurs for the control algorithm run by the noisy measurements in (15).

The rest of this section provides the proof of the theorem. We first prove the noiseless case of the theorem. Let

$$D(\boldsymbol{\alpha}) = \max_{\mathbf{r}:0\le r_i\le \bar{r}} L(\mathbf{r},\boldsymbol{\alpha})$$

be the dual function for steady state vector $\mathbf{b}_s$. The dual program is then

$$\min_{\boldsymbol{\alpha}:\alpha_i\ge 0} D(\boldsymbol{\alpha}).$$

The following lemma is an immediate result of Assumption 1.

Lemma 1. *The dual function is convex, lower bounded, and* $\|\nabla D\| < \infty$.

Proof. First, note that the dual function is a pointwise maximum of affine functions so it is convex. Second, clearly the function is lower bounded as setting $\mathbf{r} = 0$ implies a finite lower bound on the function. Finally, since $r_i \le \bar{r}$, the gradient of the dual function is also bounded. □

Let $\boldsymbol{\alpha}^*$ be a minimizer of the dual program. Let $\Delta(k) = \frac{1}{2}\|\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*\|^2$.

Lemma 2. *For any* $\epsilon > 0$*, there exists* $k_0(\epsilon)$ *such that when* $\Delta(k) \ge \epsilon$ *and* $D(\boldsymbol{\alpha}(k)) - D(\boldsymbol{\alpha}^*) \ge \delta(\epsilon)$*, if* $k \ge k_0(\epsilon)$*,*

$$\Delta(k+1) \le \Delta(k) - \zeta(k),$$

for a positive sequence $(\zeta(k), k \ge k_0(\epsilon))$ *such that*

$$\sum_{k=k_0(\epsilon)}^{\infty} \zeta(k) = \infty.$$

Proof. We first expand $\Delta(k+1)$ as follows.

$$\Delta(k+1) = \frac{1}{2}\|[\boldsymbol{\alpha}(k) + \gamma_k(\mathbf{A}\mathbf{r}(k) - \mathbf{b}(k))]_+ - \boldsymbol{\alpha}^*\|^2 \quad (17)$$

$$= \frac{1}{2}\|[\boldsymbol{\alpha}(k) + \gamma_k(\mathbf{A}\mathbf{r}(k) - \mathbf{b}_s - \mathbf{e}(k))]_+ - \boldsymbol{\alpha}^*\|^2 \quad (18)$$

$$\le \frac{1}{2}\|\boldsymbol{\alpha}(k) + \gamma_k(\mathbf{A}\mathbf{r}(k) - \mathbf{b}_s - \mathbf{e}(k)) - \boldsymbol{\alpha}^*\|^2 \quad (19)$$

$$= \frac{1}{2}\|\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^* - \gamma_k\nabla D(\boldsymbol{\alpha}(k)) - \gamma_k\mathbf{e}(k))\|^2 \quad (20)$$

$$\le \Delta(k) + \frac{1}{2}\gamma_k^2 C + \gamma_k(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}(k))^T \times (\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k)), \quad (21)$$

where C is some finite positive constant. We now explain the steps. (17) is by the update rule (10), (18) is by Assumption 3, (19) is due to the fact that convex projection on the set of non-negative shadow prices is non-expansive, (20) is by rearranging the terms, and (21) is due to the fact that ∇D is L_2–bounded by Lemma 1, and $\mathbf{e}$ is L_2–bounded by Assumption 3. Now let

$$\zeta'(k) = \gamma_k(\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*)^T(\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k)) - \frac{1}{2}\gamma_k^2 C.$$

By convexity of the dual function,

$$(\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*)^T(\nabla D(\boldsymbol{\alpha}(k)) \ge D(\boldsymbol{\alpha}(k)) - D(\boldsymbol{\alpha}^*).$$

Further, if $\Delta(k) \ge \epsilon$ and $D(\boldsymbol{\alpha}(k)) - D(\boldsymbol{\alpha}^*) \ge \delta(\epsilon)$, we have

$$\zeta'(k) \ge \zeta(k) \triangleq \gamma_k\delta(\epsilon) - \frac{1}{2}\gamma_k^2 C - \gamma_k\|\mathbf{e}(k)\|\|\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*\|.$$

Since $\gamma_k \to 0$, one can pick a sufficiently large $k_0(\epsilon)$ such that $\zeta(k) \ge 0$ for $k \ge k_0(\epsilon)$. Further, since $\|\mathbf{e}(k)\| \to 0$ as $k \to \infty$ and $\|\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*\|$ is bounded,the step size of the algorithm can be chosen such that $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$ and

$$\sum_{k=1}^{\infty}\gamma_k\|\mathbf{e}(k)\|\|\boldsymbol{\alpha}(k) - \boldsymbol{\alpha}^*\| < \infty,$$

while $\sum_{k=1}^{\infty}\gamma_k = \infty$.[2] Then,

$$\sum_{k=k_0}^{\infty}\zeta(k) \ge \delta(\epsilon)\sum_{k=k_0}^{\infty}\gamma_k - K = \infty,$$

for some finite $K > 0$, which completes the proof of the lemma. □

Lemma 3. *If* $\Delta(k) \le \epsilon$*, there exists* $k_1(\epsilon)$ *such that* $\Delta(k+1) \le 3\epsilon$ *for* $k \ge k_1(\epsilon)$*.*

Proof. First note the norm inequality;

$$\|\mathbf{x} + \mathbf{y}\|^2 \le 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2).$$

Thus,

$$\Delta(k+1) \le 2\Delta(k) + \gamma_k^2\|\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k)\|^2.$$

Since $\|\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k)\|^2$ is bounded and γ_k converges to 0, one can pick a sufficiently large $k_1(\epsilon)$ such that $\gamma_k^2\|\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k)\|^2 \le \epsilon$ for $k \ge k_1(\epsilon)$, which implies that $\Delta(k+1) \le 2\Delta(k) + \epsilon$. Thus, if $\Delta(k) \le \epsilon$, for sufficiently large k, $\Delta(k+1) \le 3\epsilon$, which completes the proof of the lemma. □

Gathering the results from Lemmas 2 and 3, we now complete the proof of the noiseless case of the theorem. It follows from Lemma 2 that there exists a sufficiently large $k_2(\epsilon)$ such that for some minimizer of the dual function, $\tilde{\boldsymbol{\alpha}}$, $\|\boldsymbol{\alpha}(k_2) - \tilde{\boldsymbol{\alpha}}\|^2 \le 2\epsilon$. We choose

$$k_3(\epsilon) = \max(k_1(\epsilon), k_2(\epsilon)).$$

Then, due to Lemma 3, $\|\boldsymbol{\alpha}(k) - \tilde{\boldsymbol{\alpha}}\|^2 \le 6\epsilon$ for all $k \ge k_3(\epsilon)$. Since this is true for arbitrary $\epsilon > 0$, it proves part (i) of the theorem for noiseless measurements.

We now prove part (ii) of the theorem. The intuition for the robustness of the algorithm to measurement noise is as follows. Our control policy is based on gradient projection; thus, with noisy measurements, the gradient projection algorithm is essentially replaced by a *stochastic* gradient projection algorithm, which still converges to the optimal solution under reasonable technical assumptions on the noise

[2]The details of how to construct the sequence of step sizes to satisfy this property are removed due to lack of space, and will show up in the extended version of the paper.

process. The proof technique for the convergence of the stochastic gradient projection algorithm uses the L_2–bounded martingale convergence theorem, and builds on the proof of part (i).

First note that matrix $\mathbf{A}$ is not a function of the densities; thus, it remains unchanged despite having measurement noise. Further, the measured vector $\hat{\mathbf{b}}(k)$ can be written as

$$\hat{\mathbf{b}}(k) = \mathbf{b}(k) + \mathbf{V}(k), \tag{22}$$

where

$$\mathbf{V}(k) = [0, -w_2 Z_2(k), 0, -w_3 Z_3(k), \cdots, 0, -w_N Z_N(k), 0]^T.$$

By Assumption 2 and since w_i's are bounded, the random vector $\mathbf{V}(k) \in \mathbb{R}^{2N-1}$ is again zero-mean and L_2–bounded. The control policy with noisy measurements remain the same as the one explained by Equations (10) and (11), with the difference that $\mathbf{b}(k)$ in (10) is measured with noise and replaced by $\hat{\mathbf{b}}(k)$. We now prove that Lemmas 2 and 3 sill hold for the control algorithm with noisy measurements. For Lemma 2, similarly to (17)–(21), one can show that

$$\Delta(k+1) \leq \Delta(k) + \frac{1}{2}\gamma_k^2 C + \gamma_k(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}(k))^T \times (\nabla D(\boldsymbol{\alpha}(k)) + \mathbf{e}(k) + \mathbf{V}(k)). \tag{23}$$

As one can see, the only difference between (21) and (23) is the extra term corresponding to $\mathbf{V}(k)$. Thus, what we need to show to prove the lemma in the noisy case is that $\mathbf{Y}(m) \triangleq \sum_{k=1}^{m} \gamma_k \mathbf{V}(k)$ converges to an L_2–bounded random vector as m goes to infinity, so that $\sum_k \zeta(k) = \infty$ still holds. Note that

$$E[\mathbf{Y}(m+1)|\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}(m)] = 0.$$

Thus, the process $\mathbf{Y}(m)$ is a martingale, and by martingale convergence theorem, $\mathbf{Y}(m)$ converges to an L_2–bounded random vector. This proves that Lemma 2 still holds for the noisy control algorithm. For Lemma 3, since the noise process is bounded, it is implied that $\|\nabla D(\alpha(k)) + \mathbf{e}(k) + \mathbf{V}(k)\|^2$ is bounded. Thus, as γ_k is vanishing, Lemma 3 still holds for the noisy algorithm. The rest of the proof remains identical to the proof of part (i) of the theorem, which completes the proof of Theorem 1. ■

The following corollary is an immediate result of Theorem 1.

Corollary 1. *In the case of feasible arrivals, the closed-loop system is stable.*

Proof. By Theorem 1, the onramp flows converge to the primal-dual optimal solution. Since the utility functions are increasing, and $\mathbf{r}$ remains in the feasible set $\mathcal{C}$, the optimal solution is $r_i = d_i$ that implies stability by the feasibility assumption. □

A. Discussion

A freeway encounters two regimes of arrivals: undersaturated or oversaturated. In the case of constant arrivals, understurated refers to a set of arrivals for which there exists an equilibrium point with the steady state flows being all feasible (They lie in the interval between zero and segments' capacities). In other words, for undersaturated arrivals, even without metering, the freeway will achieve its equilibrium when onramp queues are empty in steady state. Thus, for such regime of arrivals, even under no control conditions, onramp queues are discharged and mainline densities are stabilized. As a result, controllers are required in the presence of undersaturated arrivals to ensure that a certain transient behavior is obeyed by the system. For such arrival rates, our control strategy will simply let the freeway converge to its equilibrium without considering how the transient behavior of the system might look like. If one wishes to steer the freeway such that it converges to a different equilibrium, the previously proposed controllers [3], [22], [9] can be utilized.

Nonetheless, it is important to decide on the onramp flows in presence of oversaturated arrivals as they are pretty common during rush hours. However, the state of the art ramp metering controls fail in such scenarios as they hit infeasibility. When oversaturated arrivals are faced, no matter how the onramps are controlled, queues will grow for the period of infeasible arrivals; however, the interesting question to answer is how to allocate freeway capacity fairly to the onramps while avoiding or reducing mainline congestion. This is the case where the strength of our algorithm becomes evident as it can reduce congestion and delay significantly with low computational cost. The effectiveness of our method for such scenarios is paramount due to the fact that well–known control designs such as [3] fail to perform under these conditions.

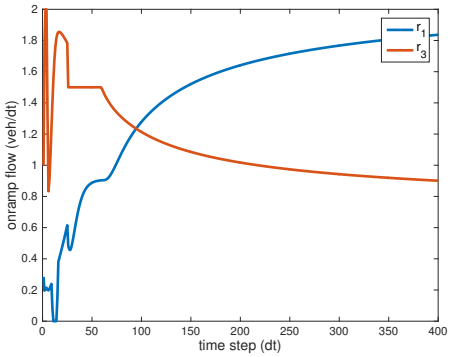
VI. Simulation Results

A. An Example

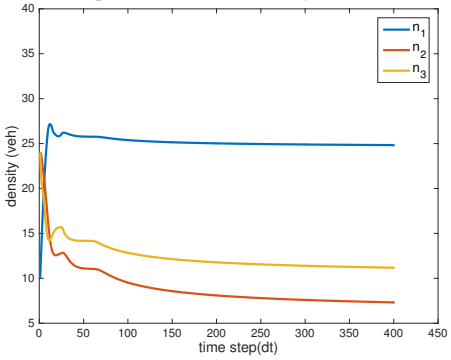
As an illustration of how our ramp metering algorithm performs, consider 3 segments of a freeway such that the first and third segments have onramps, while the first and second segments have offramps. Assuming three-second-long time steps, for each segment i, $1 \leq i \leq 3$, we have the following parameters: $v_i = 0.6$, $w_i = 0.2$, $\bar{f}_i = 4$, $\bar{n}_i = 26.7$. For the segments containing offramps, $i = 1, 2$, we have $\beta_i = 0.2$. Also, for the segments containing onramps, the arrivals entering the onramps are $d_i = 1.5, i = 1, 3$. The maximum possible onramp flow is $\bar{r}_i = 2, i = 1, 3$. The upstream mainline arrival, f_0, is 3. We assume that the initial condition of the freeway is $[10, 24, 24]$ for the freeway mainline densities and $[3, 3]$ for the onramp queues.

These values of arrivals correspond to an oversaturated arrival profile, this implies that the network is not capable of accommodating all these arrivals regardless of control strategy. In this setting, the MPC introduced in [3] fails as its introduced relaxations do not hold for these arrivals and initial conditions.

We use the $\log(.)$ function as our utility function and run our algorithm for a 20–minute time interval. Figure 2 shows the control inputs decided by the controller and the resulting mainline vehicular densities. Figure 3 displays the same quantities in the case where there is no control. In the absence of control, any empty space in the mainline is filled with

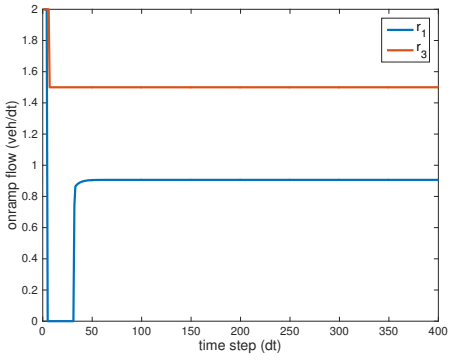


(a) Onramp flows determined by the controller.

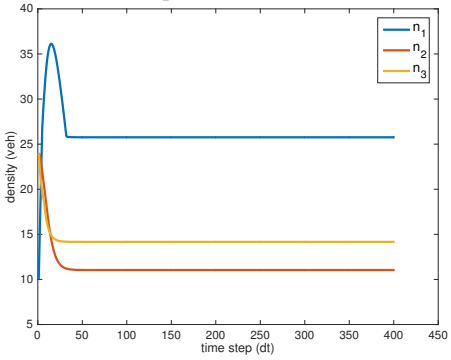


(b) Mainline densities resulting from the onramp flows determined by the controller.

Fig. 2: Trajectories resulting from the proposed ramp metering control.



(a) Onramp flows without control



(b) Mainline densities with no ramp metering control.

Fig. 3: Freeway trajectories in absence of ramp metering.

the vehicles available on the onramps. As Figure 3 shows, r_3 converges to the actual value of the onramp arrival, 1.5. On the other hand, our proposed controller is able to learn that the congestion in downstream segments can propagate to upstream segments; thus, it reduces the onramp flow in the third segment so that upstream segments can be less congested and discharge more flows from their offramps; while, it still does not shut down the onramp belonging to the third segment, being fair to people waiting on the third onramp.

In order to have a quantitative comparison of these two scenarios, we use the metrics introduced and defined in [2]. We compute total waiting time of all vehicles in all origin queues (which includes onramps and the very first upstream segment for the freeway example) and the total travel time of all vehicles in the freeway mainlines. Total waiting time in all origin queues reduces by 15.7% using our iterative control compared to the scenario where there exits no control. Moreover, total travel time of all vehicles in the freeway mainlines reduces by 16.4% by utilizing our control law. The intuition behind this reduction is that oversaturation of arrivals leads to growth in onramp queues. Nonetheless, the overall growth rate of the onramp queues is reduced using our algorithm Moreover, the total value of mainline densities is reduced using our ramp metering scheme since it tries to avoid mainline congestion Note that oversaturated arrivals are present for a finite amount of time, once arrivals become saturated, the onramps queues start to discharge. Another important quantity is the actual utility of the network that is a measure of how fairly network resources are shared among different arrivals. In the controlled case, the utility of the network converges to $\sum_i \log(r_i) = 0.5034$, but the utility of the network in the no–control case is 0.3070.

Remark 3. *A common assumption in freeway modeling is that the mainline arrivals enter the freeway through a fictitious onramp [4] which performs like a buffer such that it can accommodate the excess number of vehicles in the first mainline segment. This explains why vehicular densities are allowed to go beyond jam density in the first segment of freeway.*

B. Case Study

Now that we have demonstrated how our algorithm works in a toy example, we showcase its effect in a case study. We consider the I–210 East freeway in southern California near Los Angeles with 134 segments, 27 onramps and 23 offramps. We assume that all onramps are actuated (metered). The freeway is calibrated; its parameters such as jam densities and capacities are available to the controller. Assuming 3–second–long time steps, we run our control algorithm under a constant oversaturated rush hour arrival for 20 minutes using log function as our utility function. We consider the initial condition to be 50 for all the mainline segments and 10 for all onramp queues. We also use a constant learning factor of 1 for our algorithm. We compare the results our control algorithm to the case where ALINEA

is used in all onramps [17] (the only control law which can still perform in the oversaturated scenario).

The total waiting time of origin queues reduces by 78.4% using our control law compared to the case when ALINEA is used, which is a paramount amount of reduction (such huge changes in system performances for arterials were also reported in [2]). We also get a 68.4% improvement in mainline wait times by means of maximizing the network utility. The reason for the significant difference between the performance of the two controllers is that with ALINEA, the first upstream mainline density will increase due to the propagation of congestion upwards. In other words, there are periods of time when there is no space in the very first segment for the mainline arrival to even enter the freeway leading to increasing queues in the fictitious onramp of first segment. It is crucial to mention that in these simulation environments, the assumption is that there is no upper bound for the onramp queues which will affect the performance of both our control and ALINEA; hence, lower improvements might be attained in practice. Nonetheless, one can define the utility of each queue in our formulation inversely proportional to the bound on queue length in that segment to ensure that once queues are hitting their upper bounds, their weight and priority is increased. When running our simulations when noisy measurements, we got similar results (convergence and improvement in performance); however, due to space limitations, results with noisy measurements are not presented.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new ramp metering algorithm using an optimization framework for network utility maximization. We showed how constructing the dual problem and solving it via gradient projection algorithm leads to obtaining a simple robust (to measurement noise) control law which can significantly reduce congestion in the network. Further, we demonstrated that the control algorithm can be distributed between onramps.

There can be a number of future directions based on our framework in the context of traffic control. We are particularly interested in investigating the performance of our control algorithm under time–varying arrivals. Moreover, making the ramp metering control robust to segment capacities is of great importance due to the capacity drop phenomenon which is an inevitable consequence of lane changing. The applications of this framework is not limited to freeway networks; network utility maximization can be potentially employed for signal control in urban arterials and other traffic network settings. Additionally, the characterization of the arrivals that we presented suggests that proposing hybrid controls for freeways where the objective of the control is set based off of the arrival regime such as the similar work done for arterials in [10] will be of importance;

ACKNOWLEDGMENTS

The authors thank Gabriel Gomes for providing the I–210 East data. This work is supported by the National Science Foundation under Grant No. CPS 1446145 and the startup grant for Ramtin Pedarsani.

REFERENCES

[1] S. Coogan, E. Aydin Gol, M. Arcak, and C. Belta. Traffic network control from temporal logic specifications. 2014.

[2] C. Diakaki, M. Papageorgiou, and K. Aboudolas. A multivariable regulator approach to traffic-responsive network-wide signal control. *Control Engineering Practice*, 10(2):183–195, 2002.

[3] G. Gomes and R. Horowitz. Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C: Emerging Technologies*, 14(4):244–262, 2006.

[4] G. Gomes, R. Horowitz, A. A. Kurzhanskiy, P. Varaiya, and J. Kwon. Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies*, 16(4):485–513, 2008.

[5] A. Hegyi, B. De Schutter, H. Hellendoorn, and T. Van Den Boom. Optimal coordination of ramp metering and variable speed control-an mpc approach. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, volume 5, pages 3600–3605. IEEE, 2002.

[6] Z. Hou, J.-X. Xu, and J. Yan. An iterative learning approach for density control of freeway traffic flow via ramp metering. *Transportation Research Part C: Emerging Technologies*, 16(1):71–97, 2008.

[7] F. P. Kelly, A. K. Maulloo, and D. K. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[8] E. S. Kim, M. Arcak, and S. A. Seshia. Compositional controller synthesis for vehicular traffic networks. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 6165–6171. IEEE, 2015.

[9] S. Koehler, N. Mehr, R. Horowitz, and F. Borrelli. Stable hybrid model predictive control for ramp metering. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 1083–1088. IEEE, 2016.

[10] A. Kouvelas, K. Aboudolas, M. Papageorgiou, and E. B. Kosmatopoulos. A hybrid strategy for real-time traffic signal control of urban road networks. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):884–894, 2011.

[11] X. Lin, N. B. Shroff, and R. Srikant. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected areas in Communications*, 24(8):1452–1463, 2006.

[12] S. H. Low and D. E. Lapsley. Optimization flow controli: basic algorithm and convergence. *IEEE/ACM Transactions on Networking (TON)*, 7(6):861–874, 1999.

[13] X.-Y. Lu, T. Z. Qiu, P. Varaiya, R. Horowitz, and S. E. Shladover. Combining variable speed limits with ramp metering for freeway traffic control. In *Proceedings of the 2010 American Control Conference*, pages 2266–2271. IEEE, 2010.

[14] N. Mehr and R. Horowitz. Probabilistic freeway ramp metering. *rN*, 1(f1):1, 2016.

[15] N. Mehr, D. Sadigh, and R. Horowitz. Probabilistic controller synthesis for freeway traffic networks. In *American Control Conference (ACC), 2016*, pages 880–880. IEEE, 2016.

[16] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking (ToN)*, 8(5):556–567, 2000.

[17] M. Papageorgiou, H. Hadj-Salem, and J.-M. Blosseville. Alinea: A local feedback control law for on-ramp metering. *Transportation Research Record*, (1320), 1991.

[18] M. Papageorgiou and A. Kotsialos. Freeway ramp metering: An overview. In *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*, pages 228–239. IEEE, 2000.

[19] R. Pedarsani. Robust scheduling for queueing networks. 2015.

[20] R. Pedarsani, J. Walrand, and Y. Zhong. Robust scheduling and congestion control for flexible queueing networks. In *Computing, Networking and Communications (ICNC), 2014 International Conference on*, pages 467–471. IEEE, 2014.

[21] S. Sadraddini and C. Belta. Model predictive control of urban traffic networks with temporal logic constraints. In *2016 American Control Conference (ACC)*, pages 881–881. IEEE, 2016.

[22] S. Sadraddini and C. Belta. A provably correct mpc approach to safety control of urban traffic networks. *arXiv preprint arXiv:1602.01028*, 2016.