Tolling for Constraint Satisfaction in Markov Decision Process Congestion Games

Sarah H. Q. Li[1], Yue Yu[1], Daniel Calderone[2], Lillian Ratliff[2], Behçet Açıkmeşe[1]

Abstract— Markov Decision Process (MDP) congestion game is an extension of classic congestion games, where a continuous population of selfish agents each solves a Markov decision processes with congestion: the payoff of a strategy decreases as more population uses it. We draw parallels between key concepts from capacitated congestion games and MDPs. In particular, we show that the population mass constraints in MDP congestion games are equivalent to imposing tolls/incentives on the reward function, which can be utilized by a social planner to achieve auxiliary objectives. We demonstrate such methods on a simulated Seattle ride-share model, where tolls and incentives are enforced for two distinct objectives: to guarantee minimum driver density in downtown Seattle, and to shift the game equilibrium towards a maximum social output.

I. Introduction

We consider a class of non-cooperative games, *Markov decision process congestion games* (MDPCG) [1], [2], which combine features of classic *nonatomic routing games* [3]–[5]—i.e. games where a continuous population of agents each solve a shortest path problem—and *stochastic games* [6], [7]—i.e. games where each agent solves a Markov decision process (MDP). In MDP congestion games, similar to *mean field games with congestion effects* [8], [9], a continuous population of selfish agents each solve an MDP with congestion effects on its state-action rewards: the payoff of a strategy decreases as more population mass chooses it. An equilibrium concept for MDPCG's akin to the *Wardrop equilibrium* [3] for routing games was introduced in [1].

In this paper, we consider *modifying* MDPCG's game rewards to enforce artificial state constraints that may arise from a system level. For example, in a traffic network with selfish users, tolls can be used to lower the traffic in certain neighbourhoods to decrease ambient noise. Drawing on techniques from *capacitated routing games* [10], [11] and *constrained MDPs* [12], [13], we derive reward modification methods that shifts the game equilibrium mass distribution. Alternatively, constraints may arise in the following scenario: central agent, which we denote by a *social planner*, may enforce constraints to improve user performance as measured by an alternative objective. Equilibria of MDPCGs have been shown to exhibit similar inefficiencies to classic routing games [14], [15]. As in routing games, we show how reward adjustments can minimize the gap between the equilibrium distribution and the socially optimal distribution [16], [17].

Since MDPCG models selfish population behaviour under stochastic dynamics, our constraint enforcing methods can be considered as an incentive design framework. One practical application in particular is modifying the equilibrium behaviour of ride-sharing drivers competing in an urban setting. Ride-share has become a significant component of urban mobility in the past decade [18]. As data becomes more readily available and computation more automated, drivers will have the option of employing sophisticated strategies to optimize their profits—e.g. as indicated in popular media, there are a number of mechanisms available to support strategic decision-making by ride-sharing drivers [19]–[21]. This provides the need for game theoretic models of ride-sharing competition [22]: while rational drivers only seek to optimize their individual profits, ride-sharing companies may choose to incentivize driver behaviours that are motivated by other objectives, such as maintaining driver coverage over large urban areas with varied rider demand as well as increasing overall profits.

The rest of the paper is organized as follows. Section II provides a discussion of related work. In Section III, we introduce the optimization model of MDPCG's and highlight the relationship between the classical congestion game equilibrium—i.e. *Wardrop equilibrium*—and Q-value functions from MDP literature. Section IV-A shows how a social planner can shift the game equilibrium through reward adjustments. Section IV-B adopts the Frank-Wolfe numerical method [23] to solve the game equilibrium and provides an online interpretation of Frank-Wolfe in the context of MDPCG. Section V provides an illustrative application of MDPCG, in which agents repeatedly play a ride-share model game in the presence of population constraints as well as improving the social welfare. Section VI concludes and comments on future work.

II. Related work

Stochastic population games were first studied in the literature as *anonymous sequential games* [24]–[27]. Recent developments in stochastic population games has been in the *mean field game* [8], [28] community. Our work is related to potential mean field games [8], [29] in discrete time, discrete state space [30] and mean field games on graphs [9], [31].

Our work can also be thought of as a *continuous population potential game* [32] where the strategies are policies of an MDP or as a modification of classic nonatomic routing games [5] where routes have been replaced by policies.

Techniques for cost modification to satisfy capacity constraints in nonatomic routing games were developed in [10]. See also [5, Sec. 2.8.2] for a discussion of tolling to enforce

side-constraints and [5, Sec. 2.4] for a discussion of tolling to improve social welfare in routing games.

III. MDP CONGESTION GAMES

We consider a continuous population of selfish agents each solving a finite-horizon MDP with horizon of length T, finite state space $\mathcal{S}$, and finite action space $\mathcal{A}$. We use the notation $[T] = \{1, \ldots, T\}$ to denote the integer set of length T.

The population mass distribution, $y \in \mathbb{R}^{T \times |\mathcal{S}| \times |\mathcal{A}|}$, is defined for each time step $t \in [T]$, state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$. $y_{tsa} \in \mathbb{R}$ is the population mass in state s taking action a at time t, and $\sum_a y_{tsa}$ is the total population mass in state s at time t.

Let $P \in \mathbb{R}^{(T-1) \times |\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{A}|}$ be a stochastic transition tensor. $P_{ts'sa} \in \mathbb{R}$ defines the probability of y_{tsa} transitioning to state s' in stage $t+1$ when action a is chosen. The transition tensor P is defined such that $P_{ts'sa} \geq 0$ and $\sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{ts'sa} = 1, \forall s', s \in \mathcal{S}, a \in \mathcal{A}, t \in [T-1]$.

The population mass distribution obeys the stochastic mass propagation equation

$$\sum_{a \in \mathcal{A}} y_{1sa} = p_s, \quad \forall s \in \mathcal{S}$$
$$\sum_{a \in \mathcal{A}} y_{t+1,sa} = \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{tss'a} y_{ts'a}, \ \forall \ t \in [T-1]$$

where p_s is the initial population mass in state s.

The reward of each time-state-action triplet is given by a function $r_{tsa} : \mathbb{R}_{+}^{T \times |\mathcal{S}| \times |\mathcal{A}|} \to \mathbb{R}$. $r_{tsa}(y)$ is the reward for taking action a in state s at time t for given population distribution y. In particular, if $r_{tsa}(y)$ simply depends on y_{tsa}, then there exists functions $\ell_{tsa} : \mathbb{R}_{+} \to \mathbb{R}$ such that

$$r_{tsa}(y) = \ell_{tsa}(\delta_{tsa}^T y) \tag{1}$$

where δ_{tsa} is an indicator vector for (t, s, a) such that $\delta_{tsa}^T y = y_{tsa}$. We say the game is a *congestion game* if the rewards have the form of (1) and the functions $\ell_{tsa}(y_{tsa})$ satisfy the following assumption.

Assumption 1. $\ell_{tsa}(y_{tsa})$ *is a strictly decreasing continuous function of y_{tsa} for each t, s, a.*

Intuitively, the reward of each time-state-action triplet decreases as more members of the population choose that state-action pair at that time. We will use $r(y)$ or $\ell(y)$ to refer to the tensor of all reward functions in each case.

Each member of the population solves an MDP with population dependent rewards $r_{tsa}(y)$. As in the MDP literature, we define Q-value functions for each (t, s, a) pair as

$$Q_{tsa} = \begin{cases} r_{tsa}(y) + \sum_{s'} P_{ts'sa} \Big(\max_a Q_{t+1,s'a} \Big) & t \in [T-1] \\ r_{tsa}(y) & t = T \end{cases} \tag{2}$$

In the game context, Q-value function $Q_{tsa}(y)$ represents the distribution dependent payoff that the population receives when choosing action a at (t, s). The Q-value functions can be used to define an equilibrium akin to the Wardrop equilibrium of routing games [1].

Definition 1 (Wardrop Equilibrium [1]). *A population distribution over time-state-action triplets, $\{y_{tsa}^{\star}\}_{t \in [T], s \in \mathcal{S}, a \in \mathcal{A}}$ is an* MDP Wardrop equilbrium *for the corresponding MDPCG, if for any (s, t), $y_{tsa}^{\star} > 0$ implies $Q_{tsa} \geq Q_{tsa'} \quad \forall a' \neq a, \ a' \in \mathcal{A}$.*

Intuitively, definition 1 amounts to the fact that at every state and time, population members only choose actions that are optimal. When game rewards satisfy assumption 1, MDPCG can be characterized as a *potential game*.

Definition 2 (Potential Game [1], [32]). *We say that the MDPCG associated with rewards $r(y)$ is a* potential game *if there exists a continuously differentiable function F such that $\partial F / \partial y_{tsa} = r_{tsa}(y)$.*

In the specific case when the rewards have form (1), we can use the potential function

$$F(y) = \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \int_0^{y_{tsa}} \ell_{tsa}(x) \ dx \tag{3}$$

As shown in [1, Theorem 1.3] given a potential function $F(y)$, the equilibrium to the finite horizon MDPCG can be found by solving the following optimization problem for an initial population distribution p.

$$\max_y \ F(y)$$
$$\text{s.t.} \ \sum_{a \in \mathcal{A}} y_{t+1,sa} = \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{tss'a} y_{ts'a}, \ \forall \ t \in [T-1],$$
$$\sum_{a \in \mathcal{A}} y_{1sa} = p_s, \quad \forall \ s \in \mathcal{S},$$
$$y_{tsa} \geq 0, \quad \forall \ s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]$$

The proof that the optimizer of (4) is a Wardrop equilibrium relies on the fact that the Q-value functions (2) are encoded in the KKT optimality conditions of the problem. The equilibrium condition (Definition 1) is then specifically derived from the complementary slackness condition [1]. When $F(y)$ has form (3) and Assumption 1 is satisfied, $F(y)$ is strictly concave, and MDPCG (4) has a unique Wardrop equilibrium.

IV. CONSTRAINED MDPCG

In this section, we analyze the problem of shifting the game equilibrium by augmenting players' reward functions. In Section IV-A, we show that introducing constraints cause the optimal population distribution to obey Wardrop equilibrium for a new set of Q-value functions. Section IV-B outlines the Frank Wolfe numerical method for solving a constrained MDPCG as well as provides a population behavioural interpretation for the numerical method.

A. Planning Perspective: Model and Constraints

The Wardrop equilibrium of an MDP congestion game is given by (4). The planner may use additional constraints to achieve auxiliary global objectives. For example, in a city's traffic network, certain roads may pass through residential neighbourhoods. A city planner may wish to artificially limit traffic levels to ensure residents' wellbeing.

1239

We consider the case where the social planner wants the equilibrium population distribution to satisfy constraints of the form

$$g^i(y) \geq 0 \quad \forall i \in \mathcal{I} \tag{5}$$

where g^i are continuously differentiable concave functions.

The social planner cannot explicitly constrain players' behaviour, but rather seeks to add incentive functions $\{f_{tsa}^i\}_{i \in \mathcal{I}}$ to the reward functions $\ell(y)$ in order to shift the equilibrium to be within the constrained set defined by (5). The modified rewards have form

$$\bar{r}_{tsa}(y) = r_{tsa}(y) + \sum_{i \in \mathcal{I}} f_{tsa}^i(y) \tag{6}$$

To determine the incentive functions, the social planner first solves the constrained optimization problem

$$\begin{aligned} \max_{y} \ & F(y) \\ \text{s.t.} \ & \sum_{a \in \mathcal{A}} y_{t+1,sa} = \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} P_{tss'a} y_{ts'a}, \ \forall \, t \in [T-1], \\ & \sum_{a \in \mathcal{A}} y_{1sa} = p_s, \ \forall \, s \in \mathcal{S} \\ & y_{tsa} \geq 0, \ \forall \, s \in \mathcal{S}, \ a \in \mathcal{A}, \ t \in [T] \\ & g_i(y) \geq 0, \ \forall \, i \in \mathcal{I} \end{aligned} \tag{7a}$$

and then computes the incentive functions as

$$f_{tsa}^i(y) = (\tau^i)^\star \frac{\partial g^i}{\partial y_{tsa}}(y) \tag{8}$$

where $\{(\tau^i)^\star \in \mathbb{R}_+\}_{i \in \mathcal{I}}$ are the optimal Lagrange multipliers associated with the additional constraints (5).

The following theorem shows that the Wardrop equilibrium of the MDPCG with modified rewards in (6) satisfies the new constraints in (7a).

Theorem 1. *Let the MDPCG* (4) *with rewards* $r(y)$ *be a potential game with a strictly concave potential function* $F(y)$. *If* $y^\star$ *is a Wardrop equilibrium for a modified MDPCG with reward functions* $\bar{r}_{tsa}(y) = r_{tsa}(y) + \sum_{i \in \mathcal{I}} (\tau^i)^\star \frac{\partial g^i}{\partial y_{tsa}}(y)$, *then* $y^\star$ *also solves* (7) *and thus satisfies the additional constraints* (5).

Proof. The Lagrangian of (7) is given by

$$\begin{aligned} L(y, \mu, V, \tau) = \ & F(y) - \sum_{tsa} \mu_{tsa} y_{tsa} + \sum_i \tau^i g^i(y) \\ & + \sum_{t=1}^{T-1} \sum_s \Big(\sum_{as'} P_{t,ss'a} y_{t,s'a} - \sum_a y_{t+1,sa} \Big) V_{t+1,s} \\ & + \sum_s \Big(p_s - \sum_a y_{1sa} \Big) V_{1s} \end{aligned} \tag{9}$$

and note that by strict concavity, $\sup_{y \geq 0} \inf_{\mu \geq 0, V, \tau \geq 0} L(y, \mu, V, \tau)$ has unique solution, which we denote by $(y^\star, \mu^\star, V^\star, \tau^\star)$. We then note that

$$\bar{F}(y) = F(y) + \sum_i (\tau^i)^\star g^i(y) \tag{10}$$

is a potential function for the MDPCG with desired modified rewards. Since $F(y)$ is strictly concave, $g^i(y)$ is concave, and $(\tau^i)^\star$ is positive, $\bar{F}(y)$ is strictly concave. The equilibrium

for the MDPCG with modified rewards can be computed by solving (7) with $\bar{F}(y)$ as the objective.

The Lagrangian for (7) with $\bar{F}(y)$ is given by $\bar{L}(y, \mu, V) = L(y, \mu, V, \tau^\star)$. Again by strict concavity,

$$\sup_{y \geq 0} \inf_{\mu \geq 0, V} \bar{L}(y, \mu, V) = \sup_{y \geq 0} \inf_{\mu \geq 0, V} L(y, \mu, V, \tau^\star)$$

has a unique solution which we denote as $(\bar{y}^\star, \bar{\mu}^\star, \bar{V}^\star)$. It follows that $\bar{y}^\star = y^\star$. Thus the game equilibrium with modified rewards, $\bar{y}^\star$ satisfies desired constraints. $\square$

For the social planner, Theorem 1 has the following interpretation: in order to impose constraints of form (5) on a MDPCG, the planner could solve the constrained game (7) for optimal dual variables $\tau^\star$ and offer incentives of form (8).

B. Population Perspective: Numerical Method

After the social planner has offered incentives, the population plays the Wardrop equilibrium defined by modified rewards (6); this equilibrium can be computed using the Frank Wolfe (FW) method [23], given in Algorithm 3, with known optimal variables $\{\tau_i^\star\}$.

FW is a numerical method for convex optimization problems with continuously differentiable objectives and compact feasible sets [33], including routing games. One advantage of this learning paradigm is that the population does not need to know the function $r(\cdot)$. Instead, they simply react to the realized rewards of previous game at each iteration. It also provides an interpretation for how a Wardrop equilibrium might be asymptotically reached by agents in MDPCG in an online fashion.

Assume that we have a repeated game play, where players execute a fixed strategy determined at the start of each game. At the end of each game k, rewards of game k based on y_{tsa}^k are revealed to all players. FW models the population as having two sub-types: *adventurous* and *conservative*. Upon receiving reward information $\ell_{tsa}(y_{tsa}^k)$, the adventurous population decides to change its strategy while the conservative population does not. To determine its new strategy, the adventurous population uses value iteration on the latest reward information—i.e. Algorithm 1—to compute a new optimal policy. Their resultant density trajectory is then computed using Algorithm 2. The step size at each iteration is equivalent to the fraction of total population who switches strategy. The stopping criteria for the FW algorithm is determined by the Wardrop equilibrium notion—that is, as the population iteratively gets closer to an optimal strategy, the marginal increase in potential decreases to zero.

Algorithm 1 Value Iteration Method

Input: r, P.
Output: $\{\pi_{ts}^\star\}_{t \in [T], s \in \mathcal{S}}$
    for $t = T, \dots, 1$ **do**
        $V_{ts} = \max_{a \in \mathcal{A}} Q_{tsa}$, $\pi_{ts}^\star = \operatorname*{argmax}_{a \in \mathcal{A}} Q_{tsa}$, $\forall \, s \in \mathcal{S}$ ▷
    Eqn. (2)
    end for

Algorithm 2 Retrieving density trajectory from a policy

Input: P, p, π.
Output: $\{d_{tsa}\}_{t\in[T],s\in\mathcal{S},a\in\mathcal{A}}$
$\quad d_{tsa} = 0, \ \forall \, t\in[T], s\in\mathcal{S}, a\in\mathcal{A}$
$\quad d_{1s\pi_{1s}} = p_s, \ \forall s\in\mathcal{S}$
$\quad$ **for** $t=2,\dots,T$ **do**
$\qquad d_{ts(\pi_{ts})} = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} P_{t-1,ss'a}d_{t-1,s'a}, \ \forall \, s\in\mathcal{S}$
$\quad$ **end for**

Algorithm 3 Frank Wolfe Method with Value Iteration

Input: $\bar{\ell}$, P, p_s, N, ϵ.
Output: $\{y^\star_{tsa}\}_{t\in[T],s\in\mathcal{S},a\in\mathcal{A}}$.
$\quad y^0 = 0 \in \mathbb{R}^{T\times|\mathcal{S}|\times|\mathcal{A}|}$
$\quad$ **for** $k=1,2,\dots,N$ **do**
$\qquad c^k_{tsa} = \bar{\ell}_{tsa}(y^k), \quad \forall \, t\in[T], s\in\mathcal{S}, a\in\mathcal{A}$
$\qquad \pi_{ts} = \text{VALUEITERATION}(c^k, P)$ $\qquad\qquad$ ▷ Alg. 1
$\qquad d^k = \text{RETRIEVEDENSITY}(P, p_s, \pi_{ts})$ $\qquad$ ▷ Alg. 2
$\qquad y^k = (1-\frac{2}{k+1})y^{k-1} + \frac{2}{k+1}d^k$
$\qquad$ Stop if
$\qquad\quad \sum_{t\in[T]}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(c^k_{tsa} - c^{k-1}_{tsa}\right)^2 \leq \epsilon$
$\quad$ **end for**

In contrast to implementations of FW in routing game literature, Algorithm 3's descent direction is determined by solving an MDP [34, Section 4.5] as opposed to a shortest path problem from origin to destination [5, Sec.4.1.3]. Algorithm 3 is guaranteed to converge to a Wardrop equilibrium if the predetermined step sizes decrease to zero as a harmonic series [23]—e.g., $\frac{2}{k+1}$. FW with predetermined step sizes has been shown to have sub-linear worst case convergence in routing games [33]. On the other hand, replacing fixed step sizes with optimal step sizes found by a line search method leads to a much better convergence rate.

V. NUMERICAL EXAMPLE

In this section, we apply the techniques developed in Section IV to model competition among ride-sharing drivers in metro Seattle. Using the set up described in Section V-A, we demonstrate how a ride-share company takes on the role of social planner and shifts the equilibrium of the driver game in the following two scenarios: 1) Ensuring minimum driver density in various neighborhoods (Section V-B), 2) Improving the social welfare (Section V-C).

A. Ride-sharing Model

Consider a ride-share scenario in metro Seattle, where rational ride-sharing drivers seek to optimize their profits while repeatedly working Friday nights. Assume that the demand for riders is constant for each game play. The *time step* is taken to be 15 minutes, i.e. the average time for a ride, after which the driver needs to take a new action.

We model Seattle's individual neighbourhoods as an abstract set of *states*, $s \in \mathcal{S}$, as shown in Fig 1. Adjacent neighbourhoods are connected by edges. The following states
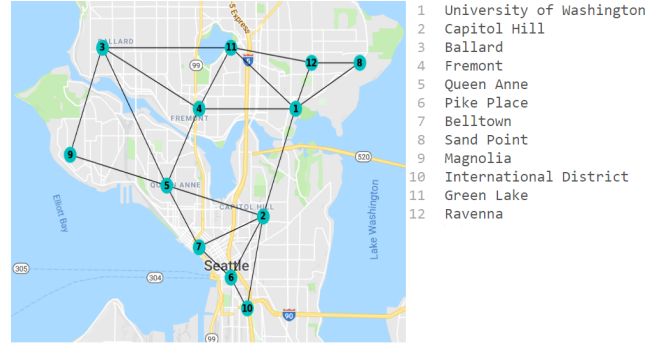

Fig. 1: State representation of metro Seattle.

are characterized as residential: 'Ballard' (3), 'Fremont' (4), 'Sand Point' (8), 'Magnolia' (9), 'Green Lake' (11), 'Ravenna' (12). Assume drivers have equal probabilities of starting from any of the residential neighbourhoods.

Because drivers cannot see a rider's destination until after accepting a ride, the game has MDP dynamics. At each state s, drivers can choose from two actions. a_r, wait for a rider in s, or a_{s_j}, transition to an adjacent state s_j. When choosing a_r, we assume the driver will eventually pick up a rider, though it may take more time if there are many drivers waiting for riders in that neighborhood. Longer wait times decrease the driver's reward for choosing a_r.

On the other hand, there are two possible scenarios when drivers choose a_{s_j}. The driver either drives to s_j and pays the travel costs without receiving a fare, or picks up a rider in s_i. We allow the second scenario with a small probability to model the possibility of drivers deviating from their predetermined strategy during game play.

The *probability of transition* for each action at state s_i are given below, where N_i denotes the set of neighbouring states, and $|N_i|$ the number of neighbouring states for state s_i.

$$P(s,a,s_i) = \begin{cases} \frac{1}{|N_i|+1}, & \text{if } s\in N_i, a=a_r \\ \frac{1}{|N_i|+1}, & \text{if } s=s_i, a=a_r \\ \frac{0.1}{|N_i|}, & \text{if } s\in N_i, s\neq s_j, a=a_{s_j} \\ 0.9, & \text{if } s\in N_i, s=s_j, a=a_{s_j} \\ 0, & \text{otherwise} \end{cases}$$

The *reward function* for taking each action is given by

$$\ell_{tsa}(y_{tsa}) = \mathbb{E}_{s'}\left[m_{ts's} - c^{\text{trav}}_{ts's}\right] - c^{\text{wait}}_t \cdot y_{tsa}$$
$$= \sum_{s'} P_{ts'sa}\left[m_{ts's} - c^{\text{trav}}_{ts's}\right] - c^{\text{wait}}_t \cdot y_{tsa}$$

where $m_{ts's}$ is the monetary cost for transitioning from state s to s', $c^{\text{trav}}_{ts's}$ is the travel cost from state s to s', c^{wait}_t is the coefficient of the cost of waiting for a rider. We compute these various parameters as

$$m_{ts's} = (\text{Rate})\cdot(\text{Dist}) \tag{11a}$$
$$c^{\text{trav}}_{ts's} = \tau \underbrace{(\text{Dist})}_{\text{mi}}\underbrace{(\text{Vel})^{-1}}_{\text{hr/mi}} + \underbrace{\left(\begin{smallmatrix}\text{Fuel}\\\text{Price}\end{smallmatrix}\right)}_{\text{\$/gal}}\underbrace{\left(\begin{smallmatrix}\text{Fuel}\\\text{Eff}\end{smallmatrix}\right)^{-1}}_{\text{gal/mi}}\underbrace{(\text{Dist})}_{\text{mi}} \tag{11b}$$
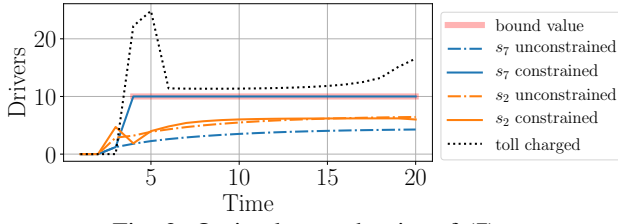
Fig. 2: Optimal state density of (7).

$$c_{ta}^{\text{wait}} = \begin{cases} \tau \cdot \left(\underbrace{\frac{\text{Customer}}{\text{Demand Rate}}}_{\text{rides/hr}} \right)^{-1}, & \text{if } a = a_r \\[2em] \epsilon_{tsa_{s'}}, & \text{if } a = a'_s \end{cases} \quad (11c)$$

where $\epsilon_{tsa_{s'}}$ is the congestion effect from drivers who all decide to traverse from s to s', and τ is a time-money tradeoff parameter, computed as $\left(\frac{\text{Rate} \cdot D_{ave}}{\text{Time Step}} \right)$, where the average trip length, D_{ave}, is equivalent to the average distance between neighbouring states. The values independent of specific transitions are listed in the Tab. I.

Rate	Velocity	Fuel Price	Fuel Eff	τ	D_{ave}
$6 /mi	8 mph	$2.5/gal	20 mi/gal	$27 /hr	1.25 mi

TABLE I: Parameters for the driver reward function.

B. Ensuring Minimum Driver Density

To ensure rider satisfaction, the ride-share company aims for a minimum driver coverage of 10 drivers in 'Belltown', $s = 7$, a neighborhood with highly variable rider demand. To this end, they solve the optimization problem in (7) where (7a) for $t \in \{3, \ldots, T\}$, $s = 7$, take on form $g_i(y) = \sum_a y_{tsa} - 10$. The modified rewards from Theorem 1 are given by $\bar{r}_{tsa}(y) = \ell_{tsa}(y_{tsa}) + \tau_{ts}^\star$, where each $\tau_{ts}^\star$ is the optimal dual variable corresponding to each new constraint.

The optimal population distribution in 'Belltown' (state 7) and an adjacent neighbourhood, 'Capitol Hill' (state 2), are shown in Fig. 2. The imposed constraints also affect optimal population distribution of adjacent states, as shown by the population distribution of Capitol Hill. Note that the incentive $\tau_{ts}^\star$ is applied to all actions of state s. Furthermore, if the solution to the unconstrained problem is feasible for the constrained problem, then $\tau_{ts}^\star = 0$—i.e. no incentive is offered. We simulate drivers' behaviour with Algorithm 3, as a function of decreasing termination tolerance ϵ. In Fig. 3, the result shows that the optimal population distribution from the FW algorithm converge to Wardrop equilibrium as the approximation tolerance ϵ decreases.

C. Increasing Social Welfare

In most networks with congestion effects, the population does not achieve the maximum *social welfare*, which can be achieved by optimizing (4) with objective $J(y) = \sum_{t \in [T]} \sum_{s \in S} \sum_{a \in A} y_{tsa} r_{tsa}(y)$. In general, a gap exists between $J(x^\star)$ and $J(y^\star)$, where $y^\star = \{y_{tsa}^\star\}$ is the optimal
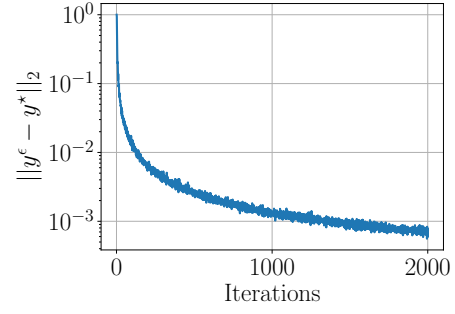

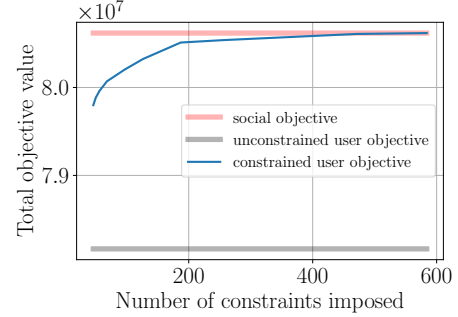Fig. 3: Convergence of y^ϵ to $y^\star$ normalized by $\|y^\star\|_2$.

Fig. 4: Constrained user optimal as a function of ϵ.

solution to (4), and $x^\star = \{x_{tsa}^\star\}$ is the optimal solution to (4) with objective $J(y)$.

The typical approach to closing the social welfare gap is to impose mass dependent incentives. An alternative method, perhaps under-explored, is to impose constraints. As opposed to congestion dependent taxation methods for improving social welfare [4], [16], constraint generated tolls are congestion independent.

We can compare the two distributions and generate upper/lower bound constraints with an ϵ threshold—see Algorithm 4 for constraint selection method. The number of constraints increases with decreasing ϵ. Since the objective function in (4) is continuous in y_{tsa}, as ϵ approaches zero, the objective will also approach the social optimal. In Fig. 4, we

---

Algorithm 4 Constraint Generation

Input: $x^\star$, $y^\star$.
Output: $\mathcal{U} = \{(u_i, t, s, a) \in \mathbb{R} \times [T] \times S \times A\}$
$\qquad \mathcal{L} = \{(l_i, t, s, a) \in \mathbb{R} \times [T] \times S \times A\}$
$\quad$ **for each** $s \in S$, $a \in A$, $t \in [T]$ **do**
$\qquad$ **if** $y_{tsa}^\star - x_{tsa}^\star > \epsilon$ **then**
$\qquad\quad$ $(x_{tsa}^\star, t, s, a) \to \mathcal{U}$
$\qquad$ **else if** $y_{tsa}^\star - x_{tsa}^\star < \epsilon$ **then**
$\qquad\quad$ $(x_{tsa}^\star, t, s, a) \to \mathcal{L}$
$\qquad$ **end if**
$\quad$ **end for**

---

compare the optimal social welfare to the social welfare at Wardrop equilibrium of the unconstrained congestion game, modeled in Section (V-A), as a function of the population size. We use CVXPY [35] to solve the optimization problem.
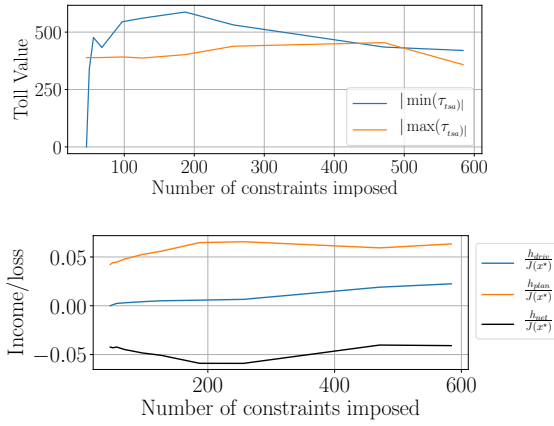
Fig. 5: Maximum and minimum tolls as a function of number of constraints (top). The income/loss required to increase social welfare as a function of constraints imposed (bottom).

We utilize Algorithm 4 to generate incentives for the congestion game. Then, we simulate (7) and compare the game output to the social objective in Fig. 4. For a population size of 3500, there is a discernible gap between the social and user-selected optimal values. Note that with only 200 (t, s, a) constraints, the gap between the social optimal and the user-selected equilibrium is already less than 5%.

An interesting question is how much of the total market worth is affected by the incentives. In Fig. 5, we demonstrate how payouts vary based on the number of constraints imposed. Let $(\cdot)_- = \min\{0, \cdot\}$ and $(\cdot)_+ = \max\{0, \cdot\}$. The total payout from the drivers to the social planner and vice versa are given by $h_{\text{driv}} = \sum_{tsa} y_{tsa} |(\tau_{tsa})_-|$ and $h_{\text{plan}} = \sum_{tsa} y_{tsa} (\tau_{tsa})_+$. The net revenue the social planner receives from tolls is $h_{\text{net}} = \sum_{tsa} y_{tsa} (\tau_{tsa}) = h_{\text{plan}} - h_{\text{driv}}$. Fig. 5(b) shows how these quantities change as the total number of constraints is increased.

VI. Conclusions

We presented a method for adjusting the reward functions of a MDPCG in order to shift the Wardrop equilibrium to satisfy population mass constraints. Applications of this constraint framework have been demonstrated in a ride-share example in which a social planner aims to constrain state densities or to maximize overall social gain without explicitly constraining the population. Future work include developing online methods that updates incentives corresponding to constraints while the game population adjusts its strategy.

References

[1] D. Calderone and S. S. Sastry, "Markov decision process routing games," in *Proc. Int. Conf. Cyber-Physical Syst.* ACM, 2017, pp. 273–279.

[2] D. Calderone and S. Shankar, "Infinite-horizon average-cost markov decision process routing games," in *Proc. Intell. Transp. Syst.* IEEE, 2017, pp. 1–6.

[3] J. G. Wardrop, "Some theoretical aspects of road traffic research," in *Inst. Civil Engineers Proc. London/UK/*, 1952.

[4] M. Beckmann, "A continuous model of transportation," *Econometrica*, pp. 643–660, 1952.

[5] M. Patriksson, *The traffic assignment problem: models and methods.* Courier Dover Publications, 2015.

[6] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci.*, vol. 39, no. 10, pp. 1095–1100, 1953.

[7] J.-F. Mertens and A. Neyman, "Stochastic games," *Int. J. Game Theory*, vol. 10, no. 2, pp. 53–66, 1981.

[8] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japan J. Math.*, vol. 2, no. 1, pp. 229–260, 2007.

[9] O. Guéant, "Existence and uniqueness result for mean field games with congestion effect on graphs," *Appl. Math. Optim.*, vol. 72, no. 2, pp. 291–303, 2015.

[10] T. Larsson and M. Patriksson, "An augmented lagrangean dual algorithm for link capacity side constrained traffic assignment problems," *Transport. Res. B*, vol. 29, no. 6, pp. 433–455, 1995.

[11] D. Hearn, "Bounding flows in traffic assignment models," *Research Report*, pp. 80–4, 1980.

[12] E. Altman, *Constrained Markov decision processes.* CRC Press, 1999, vol. 7.

[13] M. El Chamie, Y. Yu, B. Acikmese, and M. Ono, "Controlled markov processes with safety state constraints," *IEEE Transa. Autom. Control*, 2018.

[14] T. Roughgarden, *Selfish routing and the price of anarchy.* MIT press Cambridge, 2005, vol. 174.

[15] D. Calderone, "Models of competition for intelligent transportation infrastructure: Parking, ridesharing, and external factors in routing decisions," Ph.D. dissertation, U.C. Berkeley, 5 2017.

[16] A. Pigou, *The economics of welfare.* Routledge, 2017.

[17] R. Cole, Y. Dodis, and T. Roughgarden, "How much can taxes help selfish routing?" *J. Comput. Syst. Sci.*, vol. 72, no. 3, pp. 444–467, 2006.

[18] M. Furuhata, M. Dessouky, F. Ordóñez, M.-E. Brunet, X. Wang, and S. Koenig, "Ridesharing: The state-of-the-art and future directions," *Transport. Res.*, vol. 57, pp. 28–46, 2013.

[19] J. Knope. Download These 12 Best Apps For Uber Drivers & Lyft Drivers. [Online]. Available: https://therideshareguy.com/12-must-have-apps-for-rideshare-drivers

[20] M. Katz. This App Let's Drivers Juggle Competing Uber and Lyft Rides. [Online]. Available: https://www.wired.com/story/this-app-lets-drivers-juggle-competing-uber-and-lyft-rides/

[21] P. Solman. How Uber drivers game the app and force surge pricing. [Online]. Available: https://www.pbs.org/newshour/economy/uber-drivers-game-app-force-surge-pricing

[22] A. Ahmed, P. Varakantham, and S.-F. Cheng, "Uncertain congestion games with assorted human agent populations," *arXiv preprint arXiv:1210.4848 [cs.GT]*, 2012.

[23] R. M. Freund and P. Grigas, "New analysis and results for the frank–wolfe method," *Math. Program.*, vol. 155, no. 1-2, pp. 199–230, 2016.

[24] B. Jovanovic and R. W. Rosenthal, "Anonymous sequential games," *J. Math. Econ.*, vol. 17, no. 1, pp. 77–87, 1988.

[25] J. Bergin and D. Bernhardt, "Anonymous sequential games with aggregate uncertainty," *J. Math. Econ.*, vol. 21, no. 6, pp. 543–562, 1992.

[26] ——, "Anonymous sequential games: existence and characterization of equilibria," *Econ. Theory*, vol. 5, no. 3, pp. 461–489, 1995.

[27] P. Wiecek and E. Altman, "Stationary anonymous sequential games with undiscounted rewards," *J. Optim. Theory Appl.*, vol. 166, no. 2, pp. 686–710, 2015.

[28] O. Guéant, J.-M. Lasry, and P.-L. Lions, "Mean field games and applications," in *Paris-Princeton lectures on mathematical finance 2010.* Springer, 2011, pp. 205–266.

[29] O. Guéant, "From infinity to one: The reduction of some mean field games to a global control problem," *arXiv preprint arXiv:1110.3441 [math.OC]*, 2011.

[30] D. A. Gomes, J. Mohr, and R. R. Souza, "Discrete time, finite state space mean field games," *J. Math. Pures Appl.*, vol. 93, no. 3, pp. 308–328, 2010.

[31] O. Guéant, "Mean field games on graphs," in *NETCO 2014*, 2014.

[32] W. H. Sandholm, "Potential Games with Continuous Player Sets," *J. Econ. Theory*, vol. 97, no. 1, pp. 81–108, 2001.

[33] W. B. Powell and Y. Sheffi, "The convergence of equilibrium algorithms with predetermined step sizes," *Transport. Sci.*, vol. 16, no. 1, pp. 45–55, 1982.

[34] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[35] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2909–2913, 2016.