Congestion-aware motion planning game with Markov decision process dynamics

Sarah H. Q. Li\textsuperscript{1}, Daniel Calderone\textsuperscript{1}, Behçet Açıkmeşe\textsuperscript{1}

Abstract—To model a competitive multi-agent motion planning scenario for heterogeneous players with distinct dynamics and objectives, we propose a novel atomic congestion game framework with Markov decision process (MDP) dynamics. We assume that all players share the same state-action space but have unique costs and transition dynamics. Each player’s cost is a function of the joint state-action density, and thus is implicitly coupled to the opponents’ policies. With proper cost design, the resulting Nash equilibrium avoids congesting the state-action space and each player optimally achieves its individual objective with respect to the presence of its opponents. For a class of player cost functions, we demonstrate equivalence between players’ Q-value functions and the KKT point of a potential minimization problem, and derive sufficient conditions for the existence of a unique Nash equilibrium. Finally, we outline a learning algorithm for finding the Nash equilibrium that extends single-agent MDP/reinforcement learning algorithms and has linear complexity in the number of players. Throughout the paper, we provide examples for multi-agent motion planning, accumulating in a multi-robot warehouse demonstration, in which robots autonomously retrieve and deliver packages while avoiding collisions.

I. INTRODUCTION

As autonomous motion planning algorithms become widely-adapted by aeronautical, robotics, and operational sectors [1], [2]. The common underlying assumption that the operating environment is stationary is no longer sufficient. More likely, autonomous players share the operating environment with other players who may have conflicting objectives. While the possibility for multi-agent conflicts has pushed single-agent motion planning algorithms towards greater emphasis on robustness and collision avoidance, we believe that the overarching goal should be to actively consider other players’ trajectories and achieve optimality with respect to the multi-agent dynamics.

On the other hand, recent progress in single agent autonomous motion planning is partially driven by successes in reinforcement learning (RL) [3]. The framework that enables the vast applicability of RL is MDP—a simple model that allows for non-linear dynamics and environmental uncertainty. MDPs provide a strong theoretical framework from which we gain insights into model-free approaches, ultimately leading to better algorithm design. The canonical framework for multi-agent reinforcement learning is a stochastic game. Yet unlike MDPs, stochastic games are less well-understood and are notoriously difficult to solve [4]. The lack of a well-understood model directly leads to a lack of convergence results for multi-agent reinforcement learning, despite its empirical success in applications including space [1], human-robot interaction [5], and autonomous driving [6].

We propose the MDP congestion game framework as a simpler alternative. By leveraging common congestion features in multi-agent motion planning, our key contribution is reducing the N-player coupled MDP problem to a single potential minimization problem. As a result, we can use optimization techniques to analyze the Nash equilibrium as well as apply gradient descent methods to compute it.

Contributions. To address the lack of competitive path planning game models for players with MDP dynamics, we propose a MDP congestion game with finite players and heterogeneous cost functions and dynamics. We define Bellman equation-type conditions for the Nash equilibrium, formulate a game potential function and provide a necessary and sufficient condition for its existence. When the game potential exists, we show equivalence between the Q-value function and the KKT points of the potential minimization problem, and provide sufficient conditions for a unique Nash equilibrium. Specifically towards motion planning, we show that identical player cost functions are likely to generate a set of Nash equilibria. Finally, we provide a parallelizable algorithm that converges to the Nash equilibrium and give its rates of convergence. We demonstrate our model and algorithm on a 2D autonomous warehouse problem where robots must retrieve and deliver target packages from the correct locations while sharing a common navigation space.

II. RELATED WORK

MDP congestion game [7] is a type of stochastic population game and is related to potential mean field games [8], [9] in discrete time and state-action space [10] and mean field games on graphs [11]. The model in this paper is the atomic heterogeneous version of our prior work [7] on continuous population potential games with MDP dynamics or of the classic nonatomic routing games [12] where routes have been replaced by policies. In the continuous population case, MDP congestion games have been analyzed for constraint satisfaction in [13] and sensitivity to infrastructure parameters in [14].

Model-based multi-agent motion planning is typically solved via graph-based searches [15] and mixed integer linear programming [16]. Recently, reinforcement learning has been introduced as a viable method for solving multi-agent motion planning [1], [3]. In most scenarios, the motion planning problem is modeled as an MDP [17], [18]. In particular, [18] adopts a stochastic game model for human-robot collision avoidance, but focuses more on algorithm development rather than game structure analysis.

\textsuperscript{1}Authors are with the William E. Boeing Department of Aeronautics and Astronautics, University of Washington, Seattle. sarahli@uw.edu djcai@uw.edu behchet@uw.edu
III. Heterogeneous MDP Congestion Game

Consider a finite number of players $|N| = \{1, \ldots, N\}$ with a shared finite state-action space given by $([S], [A])$ and common time interval $T = \{0, 1, \ldots, T\}$. Each player $i$ has individual time-dependent transition probabilities given by $P^i \in \mathbb{R}^{T \times S \times A}$, where at time $t$, $P^i_{tsa}$ is the transition probability from state $s$ to state $s'$ using action $a$, is element-wise non-negative and satisfies the simplex constraints.

$$\sum_{s'} P^i_{ts's'a} = 1, \ \forall (i, t, s, a) \in [N] \times T \times [S] \times [A]. \quad (1)$$

**State-action distribution.** Each player $i$ controls the state-action distribution $x^i \in \mathbb{R}^{(T+1) \times S \times A}$. At time $t$, $x^i_{tsa}$ is the joint probability of player $i$ being in state $s$ and taking action $a$. This is distinct from the policy—the conditional probability of taking action $a$ when in state $s$. We collectively denote player $i$’s feasible MDP distributions from the initial distribution $x^i_0 \in \mathbb{R}^S$ as the set $\mathcal{X}(P^i, x^i_0)$.

$$\mathcal{X}(P^i, x^i_0) := \left\{ \pi \in \mathbb{R}^{(T+1) \times S \times A} \mid \sum_{s,a} \pi_{0sa} = x^i_{0sa}, \forall s \in [S], \sum_{s',a} P^i_{(t+1)sa's'a} \pi_{tsa'a} = \sum_{s,a} \pi_{(t+1)sa}, \forall (t,s) \in T \times [S] \right\}. \quad (2)$$

The joint state-action distribution of all players is given by

$$x = [x^1, \ldots, x^N] \in \mathbb{R}^{N \times (T+1) \times S \times A}. \quad (3)$$

**MDP cost.** We generalize standard MDP state-action costs to be continuously differentiable functions of $x$: player $i$ incurs a cost $\ell^i_{tsa}(x)$ for taking action $a$ at state $s$ and time $t$.

$$\ell^i_{tsa} : \mathbb{R}^{N \times (T+1) \times S \times A} \to \mathbb{R}, \ \forall (i, t, s, a) \in [N] \times T \times [S] \times [A]. \quad (4)$$

Additionally, we denote the ordered vector form of $(\ell^1, \ldots, \ell^N)$ as $\xi : \mathbb{R}^{N \times (T+1) \times S \times A} \to \mathbb{R}^{N \times (T+1) \times S \times A}$,

$$\xi(x) = [\ell^1_{000}(x), \ell^1_{001}(x), \ldots, \ell^N_{(T+1)sa}(x)]. \quad (5)$$

**Example 1.** A good indicator of congestion in motion planning is the congestion distribution $y$, defined as

$$y = \sum_{j \in [N]} \alpha_j x^j \in \mathbb{R}^{(T+1) \times S \times A}, \ \alpha_j \geq 0, \ \forall j \in [N], \quad (6)$$

where $\alpha_j$ is player $j$’s congestion impact factor. If all players contribute to congestion equally, $\alpha_j = 1 \ \forall j \in [N]$. For heterogeneous players, different players impact congestion differently. For example, a semi-truck contributes significantly more to the congestion than a sedan, therefore it requires greater impact factor. For all $(i, t, s, a) \in [N] \times T \times [S] \times [A]$, a possible cost function for motion planning is

$$\ell^i_{tsa}(y, x^i) = \alpha_i f_{tsa} \left( \sum_{a'} y_{tsa'} \right) + \alpha_i g_{tsa} \left( y_{tsa} \right) + h^i_{tsa}(x^i_{tsa}) \quad (7)$$

where $f_{tsa}(\cdot)$ is a state-dependent congestion function, $g_{tsa}(\cdot)$ is a state-action-dependent congestion function, and $h^i_{tsa}(\cdot)$ is a player-specific congestion function.

**Coupled MDPs.** Given a initial distribution $z^i_0 \in \mathbb{R}^S$ and a fixed joint-state-action distribution $x$, player $i$’s MDP problem is given by

$$\min_{x^i} \sum_{t \in T} \sum_{s \in [S]} \sum_{a \in [A]} x^i_{tsa} \ell^i_{tsa}(x^i, x^1, \ldots, x^N) \quad \text{s.t.} \ x^i \in \mathcal{X}(P^i, z^i_0). \quad (8)$$

If $\ell^i_{tsa}(x)$ is independent of $x^i$ for all $(t, s, a) \in T \times [S] \times [A]$, player $i$’s solves a standard linear program MDP. Otherwise, player $i$ solves an MDP with non-linear costs.

**Remark 1.** The cost coupling in $\ell^i_{tsa}(x)$ is distinct from stochastic games, where costs couple over the joint policy space. By enforcing coupling on the state-action density, we avoid artificially inflating $\ell^i_{tsa}$ when the opponent has policy $a$ in state $s$, but also has a low probability of being in $s$.

**Bellman iteration.** For a fixed joint-state-action distribution $x$, player $i$’s optimal cost-to-go in $\ell^i_{tsa}(x)$ can be recursively defined via Q-value functions $Q^i_{x} : \mathbb{R}^{N \times (T+1) \times S \times A} \to \mathbb{R}$ as

$$Q^i_{x} = \ell^i_{tsa}(x) + \sum_{s',a} P^i_{tsa's'a} \min_{a'} Q^i_{x}(\ell^i_{tsa}, x^1, \ldots, x^N, a'), \ \forall t \in [T] \quad (9)$$

We emphasize that (9) is the optimal cost-to-go given a fixed distribution $x$. If $Q^i_{x}$ causes player $i$ to change $x^i$, then $\ell^i_{tsa}(x)$ in (8) changes. Therefore, (9) is not the optimal solution to (8). However, (9) can be utilized to define a stable equilibrium for unilateral optimality among the players.

**Definition 1 (Nash Equilibrium).** The joint state-action distribution $x = [x^1, \ldots, x^N]$ is a Nash equilibrium if for $i = 1, \ldots, N$, $x^i$ satisfies

$$x^i_{tsa} > 0 \Rightarrow Q^i_{x}(x^1, \ldots, x^N) = \min_{a' \in [A]} Q^i_{x}(x^1, \ldots, x^N, a'), \ \forall t \in [T], z^i_0, x^1 \in \mathcal{X}(P^i, z^i_0). \quad (10)$$

I.e., player $i$ choosing action $a$ at $(t,s)$ with positive probability only if $Q^i_{x}(x^1, \ldots, x^N, a')$ is optimal for the choice of $a' \in [A]$ and $x^i \in \mathcal{X}(P^i, z^i_0)$.

Intuitively, Nash equilibrium implies that each player’s state-action distribution $x^i$ is simultaneously optimal and stable with respect to the joint state-action distribution $x$.

**Remark 2.** Definition 1 is similar to MDP Wardrop equilibrium [20]. However, since individuals no longer correspond to infinitesimal density changes, the equilibrium must be stable over the $\mathcal{X}(P^i, z^i_0)$ space as well as the action space.

**A. Potential optimization form**

So far, $\ell^i_{tsa}(x)$ describes a finite player, general-sum game with MDP dynamics whose Nash equilibrium can be NP-hard to find [4]. We are specifically interested in games of form (8) that can be reduced to a minimization problem given by

$$\min_{x^i} F(x) \quad \text{s.t.} \ x^i \in \mathcal{X}(P^i, P^i), \ \forall i \in [N]. \quad (11)$$
where $F$ is the potential function of the corresponding game.

**Definition 2** (Potential function). We say an MDP congestion game with cost functions $\{f^i\}_{i \in [N]}$ has a potential function $F: \mathbb{R}^{N(T+1)SA} \rightarrow \mathbb{R}$ if $F$ satisfies

$$\frac{\partial F(x)}{\partial x_{tsa}} = f_{i_{tsa}}(x), \forall (i, t, s, a) \in [N] \times T \times [S] \times [A].$$ (12)

Additionally, the following assumption on $\{f^i\}_{i \in [N]}$ is necessary and sufficient for the existence of $F$ [12, Eqn.2.44].

**Assumption 1.** For all $(i, t, s, a), (i', t', s', a') \in [N] \times T \times [S] \times [A]$, the set of player cost functions $\{f^i\}_{i \in [N]}$ satisfies

$$\frac{\partial f_{i_{tsa}}(x)}{\partial x_{tsa}} = \frac{\partial f_{i'_{tsa}}(x)}{\partial x_{tsa}}. \quad (13)$$

**Remark 3.** Assumption 1 implies that $F$ is conservative:

$$\exists x_1, x_2 \in \{x_{tsa} \mid (i, t, s, a) \in [N] \times T \times [S] \times [A]\},$$

$$\frac{\partial^2 F(x)}{\partial x_1 \partial x_2} = \frac{\partial^2 F(x)}{\partial x_2 \partial x_1}. \quad (14)$$

Or equivalently, the Jacobian of $\xi$ must be symmetrical.

Finding a potential function $F$ is non-trivial and not always possible. However, we can easily formulate a potential function for cost functions $\{f^i\}_{i \in [N]}$.

**Example 2.** For a congestion distribution $y$ and cost functions $\{f^i\}_{i \in [N]}$, their game potential is given by

$$F(x) = \sum_{t,s} \int_0^1 x_{tsa} f_{tsa}(u) \frac{\partial u}{\partial u} + \sum_{t,s,a} \int_0^1 g_{tsa}(u) \frac{\partial u}{\partial u} + \sum_{i,t,s,a} \int_0^1 h(u) \frac{\partial u}{\partial u}. \quad (15)$$

We can check that $F$ satisfies Assumption 1. In particular, the partial derivatives $\frac{\partial f_{i_{tsa}}}{\partial x_{tsa}}$ and $\frac{\partial g_{tsa}}{\partial x_{tsa}}$ are nonzero only if $(t, s, a) = (t', s', a')$, and the partial derivatives $f_{t_{tsa}}$ and $f_{i_{tsa}}$ are nonzero only if $(t, s) = (t', s').$

**Effect of weighted congestion impact.** The congestion impact factor $\alpha_i$ scales both the player impact on $y$ and the player sensitivity to congestion through $\ell^i$. When $\alpha_i < 1$, player $i$ impacts congestion less and cares less about the congestion. Similarly, when $\alpha_i > 1$, player $i$ impacts congestion more and cares more about the congestion. In the semi-truck vs sedan scenario from Example 1 the sedan contributes less to congestion than the truck and may be more willing to squeeze in beside a semi-truck in the same state. On the other hand, semi-trucks are less maneuverable and less willing to share a state with a sedan.

**B. Existence and uniqueness of Nash equilibrium**

When a potential function $F$ exists, we can show that the KKT points of (11) correspond to Q-value functions of a state-action distribution [20].

**Theorem 1** (Existence). Let $x$ be a joint state-action distribution for a game with continuously differentiable cost functions $\{\ell^1, \ldots, \ell^N\}$. For $(i, t, s, a) \in [N] \times T \times [S] \times [A]$, let $V_{tsa}^i, \mu_{tsa}^i$ be defined via $Q_{tsa}^i(x)$ as

$$V_{tsa}^i = \min_{a'} Q_{tsa}^i(x), \quad \mu_{tsa}^i = Q_{tsa}^i(x) - V_{tsa}^i. \quad (16)$$

If a potential function $F$ exists for $\{\ell^1, \ldots, \ell^N\}$, then $(x, V, \mu)$ is a KKT point of (11) if and only if $Q_{tsa}^i(x) = V^i - \mu^i$ is $i$-s Q-value function for every $i \in [N]$.

**Proof.** We prove the theorem statement by showing equivalence between the KKT conditions [21, Prop.3.3.7] of (11) and Q-value functions (9). For player $i$, we assign the dual variable $\nu^i = [\nu^i_0, \ldots, \nu^i_1] \in \mathbb{R}^{T+1}$ to the first and second constraints in $\mathcal{C}_t(P_i, x_{t_{tsa}}^i)$ (3), and $\mu^i \in \mathbb{R}^{T(S+1)SA}$ where $\mu_{tsa}^i \geq 0$ to the non-negative constraint on each variable $x_{tsa}^i$. The Lagrangian of (11) is given by

$$L(x^1, \ldots, x^N, \nu^0, \ldots, \nu^N, \mu^0, \ldots, \mu^N) = F(x) - \sum_{i,t,s,a} \mu_{tsa}^i x_{tsa}^i + \sum_{t,s} \nu_{ts} \left( x_{tsa}^i - \sum_a a_{tsa}^i \right) + \sum_{t,s} \nu_{tsa} \left( x_{tsa}^i - \sum_a a_{tsa}^i \right). \quad (17)$$

For all $t \neq T$, the gradient of the Lagrangian is $dL/dx_{tsa}^i = \ell_{tsa}^i(x) + \sum_{s,a} P_{tsa}^i (t+1)x_{tsa}^i - \mu_{tsa}^i - \nu_{tsa}^i$. When $t = T$, $dL/dx_{Tsa}^i = \ell_{Tsa}^i(x) - \mu_{Tsa}^i$. KKT points satisfy the gradient condition $dL/dx_{tsa}^i = 0$ for all $(i, t, s, a) \in [N] \times T \times [S] \times [A]$. We can then derive the following recursive relation for $v_{tsa}^i$

$$\ell_{tsa}^i(x) + \sum_{s,a} P_{tsa}^i (t+1)x_{tsa}^i - \mu_{tsa}^i - \nu_{tsa}^i = 0 \quad t \neq T \quad (18)$$

$$\ell_{Tsa}^i(x) - \mu_{Tsa}^i = 0 \quad t = T.$$ (18)

From the complementarity condition, $\mu_{tsa}^i = 0$ when $x_{tsa}^i > 0$, else $\mu_{tsa}^i \geq 0$.

$$v_{tsa}^i = \ell_{tsa}^i(x) + \sum_{s,a} P_{tsa}^i (t+1)x_{tsa}^i - \mu_{tsa}^i - \nu_{tsa}^i \quad t \neq T$$

$$v_{Tsa}^i = \ell_{Tsa}^i(x) - \mu_{Tsa}^i \quad t = T.$$ (19)

If we let $v_{tsa}^i = v_{tsa}^i$, then $\mu_{tsa}^i = \mu_{tsa}^i$ follows from (18). We conclude that the KKT point $(x, V, \mu)$ generates the Q-value functions (9) for $x$.

Next we show that at each fixed $x$, $(x, V, \mu)$ satisfy primal feasibility, dual feasibility, and complementarity conditions of (11). By definition of $x$ (3), primal feasibility is satisfied. We define the dual variables $V_{tsa}^i$ and $\mu_{tsa}^i$ as in (16), where $\mu$ automatically satisfies the complementarity conditions as

$$\mu_{tsa}^i = \left\{ \begin{array}{ll} 0 & x_{tsa}^i = 0 \\ Q_{tsa}^i(x) - V_{tsa}^i & x_{tsa}^i > 0 \end{array} \right. \quad \text{if } \ell_{tsa}^i(x) \geq 0,$$ (20)

By definition, $\mu_{tsa}^i \geq 0$ for all $(i, t, s, a) \in [N] \times T \times [S] \times [A]$. Therefore $(V, \mu)$ both satisfy dual feasibility conditions, and $(x, V, \mu)$ is a KKT point. \hfill \square
Theorem 2 shows that the KKT conditions are necessary conditions for Nash equilibrium. However, a unilateral change of $x^i$ could shift the KKT variable $x$ to another with a lower $Q$-value. We show next that a unique KKT point is a sufficient condition for $x$ being the Nash equilibrium.

**Theorem 2 (Uniqueness).** For a set of player costs $\ell^i(\cdot)$ where a potential function $F$ exists, if $\xi(x)$ satisfies that $\partial_\xi(x) > 0$, then the corresponding MDP congestion game has a unique Nash equilibrium.

**Proof.** When $\partial_\xi(x) > 0$, the optimization problem (11) optimizes a strictly convex function over linear constraints. Therefore, there exists a unique solution that minimizes the potential function and satisfies the KKT condition. We note that (19) holds globally for all feasible joint state-action distributions, therefore, $x$ is a Nash equilibrium.

**Corollary 1.** When all players’ cost functions are of form (7), a unique Nash equilibrium exists if $f_{tsa}(\cdot)$, $g_{tsa}(\cdot)$, and $h_{tsa}(\cdot)$ are strictly increasing for all $(i, t, s, a) \in [N] \times T \times [S] \times [A]$.

**Proof.** From (7), the inputs to $f_{tsa}, g_{tsa}, h_{tsa}$ are linear transformations of $x$ (3). Let $N \in \mathbb{R}^{(T+1) \times N \cdot (T+1)}$ and $M \in \mathbb{R}^{(T+1) \times S \cdot (T+1) \times A}$ be the corresponding linear matrices, such that $[N]_{(t,s)}$ is the input to $f_{tsa}$ and $[M]_{(t,s,a)}$ is the input to $g_{tsa}$. Define $f(Nx), g(Mx), h(x)$ to be component-wise equal to $f_{tsa}, g_{tsa}, h_{tsa}$, respectively. We can then write $\xi(x)$ as $\xi(x) = N^T f(Nx) + M^T g(Mx) + h(x)$. The Jacobian of $\xi$ is given by $\partial_\xi(x) = N^T \partial f(Nx) + M^T \partial g(Mx) + \partial h(x)$. For element-wise strictly increasing functions $f(\cdot), g(\cdot), h(\cdot)$, $\partial f/\partial x > 0, \partial g/\partial x > 0, \partial h/\partial x > 0$ for all $x$. Furthermore, since $\partial_\xi(x)$ is diagonal and full rank, $\partial_\xi(x) > 0$.

**Remark 4.** Corollary 1 implies that a strictly increasing $h^i$ is crucial to ensuring a unique Nash equilibrium. Without $h^i$, players have identical costs and are substitutable within the game. Therefore, $h^i$ can be interpreted as a regularization term to ensure uniqueness of Nash equilibrium.

**C. Frank-Wolfe learning dynamics**

MDPs are hugely popular partly due to its rich set of model-based and model-free solution methods [19], [22]. For MDP congestion games, we propose a learning dynamic that builds on the single-agent algorithms and guides all players towards the Nash equilibrium.

In Algorithm 1 each player can access an oracle which returns the cost for a given joint state-action distribution. Each player also has an algorithm of choice to solve the standard MDP problem on line 5. See [13, Alg.1] for a value iteration algorithm with compatible notations. Any model-based/model-free algorithm can be utilized, provided that the final policy is $\delta$ suboptimal and satisfies

$$V_{\pi_i}(x^k) \leq \min_\pi V_{\pi_i}(x^k) + \delta \quad (21)$$

Players then retrieve the corresponding state-action density $b^k$, take a convex combination of $b^k$ with the current state-action density $x^k$ to derive the next joint state-action density. All steps within lines 4 to 7 are parallelizable.

**Algorithm 1 Frank-Wolfe with dynamic programming/RL**

**Input:** $\{\ell^i\}_{i \in [N]}, \{P^i\}_{i \in [N]}, \{z^0_i\}_{i \in [N]}, N, \delta$.

**Output:** $x^k \in \mathcal{X}(P^i, z^0_i) \in \mathbb{R}^{(T+1) \times S \cdot (T+1) \times A}, \forall i \in [N]$.

1. $x^0 \in \mathcal{X}(P^i, z^0_i) \in \mathbb{R}^{(T+1) \times S \cdot (T+1) \times A}, \forall i \in [N]$.
2. for $k = 1, 2, \ldots$
3. for $i = 1, \ldots, N$
4. $q^i = \ell^i(x^1, \ldots, x^N)$
5. $\pi^i = \text{MDP}(q^i, P^i, [S], [A], T, \delta)$
6. $b^k = \text{RETRIEVE_DENSITY}(P^i, z^0_i, \pi^i) \triangleright \text{Alg. 2}$
7. $x^i(k+1) = (1 - \frac{2}{k+1})x^i + \frac{2}{k+1}b^k$
8. end for
9. end for

**Theorem 3.** If the set of player costs $\{\ell^i\}_{i \in [N]}$ and its vectored form $\xi(x)$ satisfy $d\xi(x)/dx \geq \alpha I$, $\alpha > 0$ for all $x = (x^1, \ldots, x^N)$ where $x^i \in \mathcal{X}(P^i, z^0_i)$, then the iterates of Algorithm 1 converge towards the Nash equilibrium as

$$\frac{1}{2} \sum_{i \in [N]} \|x^i - x^*\|^2 \leq C_F^2 \frac{2}{\alpha} + 2C_F^2 \delta \quad (22)$$

where $C_F$ is the potential function $F$’s curvature constant given by

$$C_F = \sup_{x^i, s^i \in \mathcal{X}(P^i, z^0_i)} \frac{2}{\gamma^2} (F(y) - F(x) - \sum_{i \in [N]} (x^i - y^i)^\top \ell^i(x)). \quad \gamma \in [0, 1]

y^i = x^i + \gamma(s^i - x^i)$$

**Proof.** Algorithm 1 is a straight-forward implementation of [23, Alg.2]. The only modification we introduce is changing the error term of the constrained inner problem from $\frac{1}{2 \cdot (k+2)} \delta C_{\pi_i}$ to $\delta > 0$. By assuming $d\xi(\dot{x})/dx \geq \alpha I$, we ensure that the potential function is strongly convex and therefore satisfies $\frac{1}{2} \sum_{i \in [N]} \|x^i - x^*\|^2 \leq F(x^*) - F(x^i)$. The convergence guarantee (22) then follows directly from [23, Thm.1] after applying our modifications.

**Remark 5 (Scalability).** Algorithm 4 has linear complexity in the number of players.

**IV. MULTI-AGENT PATH FINDING**

We apply our congestion game model to a multi-agent pick up and delivery scenario and evaluate the framework’s efficacy in collision avoidance and task completion. As shown in Figure 1, $N$ players navigate a 2D grid world to transport packages from pick up chutes located on the
bottom row to the drop off chutes located at the top row while avoiding collision with one another.

A. Stationary MDP Model

The physical operation space is a two dimensional grid world with 5 rows and 10 columns. In addition to capturing location, each state also dictates whether the robot is in pick up or delivery mode. The total state space is given by

\[ |S| = \{(x, y, m) \mid 1 \leq x \leq 5, 1 \leq y \leq 10, m \in \{P, D\}\}. \]

At each state, available actions are \[ |A| = \{u, d, r, l, s\} \], corresponding to up, down, right, left, stay. Player transition dynamics and rewards are stationary in time for each state-action. The transition probability of each state \((x, y, m)\) extends the location-based transition probabilities \(P^0\).

**Location-based transition.** Let \(u = (x, y)\) denote the location component of the state space. At each location, each action either points to a feasible target \(u_{targ}(a)\) or is infeasible. The set of all feasible targets from \(u\) is \(N(u)\). When a target exists, players have \(1 > q > 0\) chance of reaching it and \(1-q\) chance of reaching other states in \(N(u)\).

\[
P^0_{u'\mid u, a} = \begin{cases} q & u' = u_{targ}(a), \frac{1-q}{|N(u)|} & u' \in N(u)/\{u_{targ}(a)\}, \ 0 & \text{otherwise.} \end{cases}
\]

When the target location is infeasible, the player transitions into a neighboring state \(u' \in N(u)\) at random.

\[
P^0_{u'\mid u, a} = \begin{cases} \frac{1}{|N(u)|} & u' \in N(u), \ 0 & \text{otherwise.} \end{cases}
\]

**Full transition dynamics.** The full transition dynamics extend the location-based transition as follows: within each mode, players transition between locations via dynamics \(P^0\). Player modes may change at the following specific locations.

1) **Pick up chutes.** Each player \(i\) has unique pick up locations \(t^i = (x, y)\) located along the bottom row of Figure 1. We denote \(P^0\)‘s row vector whose index corresponds to \(t^i\) as \(c^i \in R^{1 \times S_A}\), and denote the matrix that’s zero everywhere and equal to \(1\) in the corresponding to \(t^i\) as \(C^i \in R^{S \times S_A}\).

2) **Drop off chutes.** All players share the drop off chutes in the top row locations \(R = \{(0,0), \ldots, (0,10)\}\) of Figure 1. We denote the \(P^0\) row vectors that correspond to a location in \(R\) as \(d \in R^{10 \times S_A}\), and denote the matrix that is zero everywhere and equal to \(d\) in the rows indexed by an element of \(R\) as \(D \in R^{S \times S_A}\).

Thus, each player’s full transition probability \(P^i\) is given by

\[
P^i = \begin{bmatrix} P^0 - (1 - r^i)C^i \\ r^iC^i \end{bmatrix} D \in R^{S \times S_A}, \quad (25)
\]

Here, \(r^i \in R\) denotes the probability of package arrival when player \(i\) is in \(t^i\). Modeled as an independent Poisson process with rate \(\lambda_i\) and interval \(\Delta t = 1s\), \(r^i\) is given by \(r^i = \exp(-\lambda_i \Delta t)\). When player \(i\) is in pick up mode and enters its pick up chute, it has \(r^i\) probability of seeing a package and entering delivery mode. When the player is in delivery mode and enters a drop off chute, it enters pick up mode with probability one.

B. Rewards

Player \(i\)‘s reward is of form (7): \(\ell^i = c^i_{tsa} + f_{ts}(y)\), where \(y = \sum_t x^t\) is the congestion distribution. The player-specific portion, \(c^i_{tsa}\), is defined identically for \(t \in T\) as

\[
c^i_{tsa} = \begin{cases} 1 & (x, y) = t^i, m = P, \ 0 & \text{otherwise}. \end{cases}
\]

The congestion penalty is an exponential function of the congestion distribution \(y\), defined as

\[
f_{ts}(y) = -\exp(\alpha_{tsa} \left( \sum_{m \in \{P,D\}} \sum_{a' \in A} y_{t,s,a'} - 1 \right)), \quad (28)
\]

where \(\alpha_{tsa} > 0\) for all \((t, s, a) \in T \times |S| \times |A|\).

C. Simulation results

We simulate the MDP congestion game for \(N = 3\) players with the parameters given in Table I below where the parameters given hold for all \(i \in [N]\) and \((t, s, a) \in T \times |S| \times |A|\). Each player’s pick up target locations are \(t^i = \{(x, y) \mid y_i \in [2, 5, 9]\}\). At \(t = 0\), each player is initialized to a random position in the drop off row with mode \(P\).

| \(N\) | \(q\) | \(\gamma\) | \(\lambda\) | \(\alpha_{tsa}\) | \(\Delta t\) | \(T\) |
|---|---|---|---|---|---|---|
| 3 | 0.98 | 0.99 | 0.5 | 0.5 | 1s | 60s |

**TABLE I:**

PARAMETERS FOR SIMULATION ENVIRONMENT.

We run Algorithm 1 for 100 time steps to find the Nash equilibrium. Each inner MDP problem is solved via value
and provided a gradient descent algorithm that guides single
player averages around 9. Therefore, player 0’s shortest path in a single agent environment is between 9.66 and 23.66 steps. In comparison, each player averages around 14 steps for a collision free path and has a worst waiting time of 13, 14, and 17, respectively.

In terms of collision, player have the highest chance of collision at initialization, when they tend to be huddled together on the top row. However, as the algorithm progresses, players space out and end up in more collision-free positions.

V. CONCLUSIONS

We derived a class of potential games for an $N$ player coupled MDP with heterogeneous rewards transition dynamics for multi-agent motion planning. For these games, we derive equivalence between players’ Q-value functions to the set of KKT points for a potential minimization problem and provided a gradient descent algorithm that guides single agent MDP algorithms towards Nash equilibria. Future work includes demonstrating integrating reinforcement learning methods into the gradient descent algorithm.

REFERENCES

[1] K. Yun, C. Choi, R. Alimo, A. Davis, L. Forster, A. Rahmani, M. Adil, and R. Madani, “Multi-agent motion planning using deep learning for space applications,” in ASCEND 2020, 2020, p. 4233.
[2] J. Ota, “Multi-agent robot systems as distributed autonomous systems,” Advanced engineering informatics, vol. 20, no. 1, pp. 59–70, 2006.
[3] S. H. Semnani, H. Liu, M. Everett, A. De Ruiter, and J. P. How, “Multi-agent motion planning for dense and dynamic environments via deep reinforcement learning,” IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 3221–3226, 2020.
[4] K. Chatterjee, R. Majumdar, and M. Jurdziński, “On nash equilibria in stochastic games,” in International workshop on computer science logic. Springer, 2004, pp. 26–40.
[5] M. Everett, Y. F. Chen, and J. P. How, “Motion planning among dynamic, decision-making agents with deep reinforcement learning,” in Int. Conf. Intell. Robots Syst. (IROS). IEEE, 2018, pp. 3052–3059.
[6] S. Shalev-Shwartz, S. Shammah, and A. Shashua, “Safe, multi-agent, reinforcement learning for autonomous driving,” arXiv preprint arXiv:1610.03295, 2016.
[7] D. Calderone and S. Shankar, “Infinite-horizon average-cost markov decision process routing games,” in Proc. Intell. Transp. Syst. IEEE, 2017, pp. 1–6.
[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, “From infinity to one: The reduction of some mean field space mean field games,” J. Math. Pures Appl., vol. 93, no. 3, pp. 308–328, 2010.
[10] 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.
[11] M. Patriksson, The traffic assignment problem: models and methods. Courier Dover Publications, 2015.
[12] S. H. Li, Y. Yu, D. Calderone, L. Ratliff, and B. Açıklame, “Tolling for constraint satisfaction in markov decision process congestion games,” in Amer. Control Conf. (ACC). IEEE, 2019, pp. 1238–1243.
[13] S. H. Li, D. Calderone, L. Ratliff, and B. Açıklame, “Sensitivity analysis for markov decision process congestion games,” in Conf. Decision Control (CDC). IEEE, 2019, pp. 1301–1306.
[14] L. Cohen, T. Uras, T. S. Kumar, and S. Koenig, “Optimal and bounded-suboptimal multi-agent motion planning,” in Twelfth Annual Symposium on Combinatorial Search, 2019.
[15] J. Chen, J. Li, C. Fan, and B. Williams, “Scalable and safe multi-agent motion planning with nonlinear dynamics and bounded disturbances,” arXiv preprint arXiv:2012.09052, 2020.
[16] H. Bayerlein, M. Theile, M. Caccamo, and D. Gesbert, “Multi-uav path planning for wireless data harvesting with deep reinforcement learning,” IEEE Open Journal of the Communications Society, vol. 2, pp. 1171–1187, 2021.
[17] S. Y. Lo, B. Fernandez, P. Stone, and A. L. Thomaz, “Towards safe motion planning in human workspaces: A robust multi-agent approach,” in Int. Conf. Robot. Autom. (ICRA). IEEE, 2021, pp. 7929–7935.
[18] M. L. Puterman, Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014.
[19] D. Calderone and S. S. Sastry, “Markov decision process routing games,” in Int. Conf. Cyber-Physical Syst. (ICCP). IEEE, 2017, pp. 273–280.
[20] D. P. Bertsekas, Nonlinear programming. Athena Scientific Belmont, 1999.
[21] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
[22] M. Jaggi, “Revisiting frank-wolfe: Projection-free sparse convex optimization,” in Int. Conf. Mach. Learning. PMLR, 2013, pp. 427–435.