Shaping Large Population Agent Behaviors Through Entropy-Regularized Mean-Field Games

Yue Guan[1], Mi Zhou[2], Ali Pakniyat[3], and Panagiotis Tsiotras[4]

***Abstract*— Mean-field games (MFG) were introduced to efficiently analyze approximate Nash equilibria in large population settings. In this work, we consider entropy-regularized mean-field games with a finite state-action space in a discrete time setting. We show that entropy regularization provides the necessary regularity conditions, that are lacking in the standard finite mean field games. Such regularity conditions enable us to design fixed-point iteration algorithms to find the unique mean-field equilibrium (MFE). Furthermore, the reference policy used in the regularization provides an extra means, through which one can control the behavior of the population. We first formulate the problem as a stochastic game with a large population of N homogeneous agents. We establish conditions for the existence of a Nash equilibrium in the limiting case as N tends to infinity, and we demonstrate that the Nash equilibrium for the infinite population case is also an ϵ-Nash equilibrium for the N-agent regularized game, where the sub-optimality ϵ is of order $\mathcal{O}\big(1/\sqrt{N}\big)$. Finally, we verify the theoretical guarantees through a resource allocation example and demonstrate the efficacy of using a reference policy to control the behavior of a large population of agents.**

I. Introduction

Decision making in decentralized systems arises in many applications, ranging from multi-robot task allocation [1]–[3], swarm robotics [4]–[6], communication [7]–[9], finance [10]–[12], etc. The scalability of the solution to large populations is an important consideration in these settings, as the complexity of the system increases drastically with the number of agents.

To address scalability issues, the mean field approach was introduced in [13]–[16]. The mean-field game (MFG) formulation reduces the interactions among agents to a game between a *representative agent* and a population of infinitely many other agents. Such a population is often referred to as the *mean field*, and the solution in this limiting case is the mean-field equilibrium (MFE). In the continuous setting, the MFE is characterized by a Hamilton-Jacobi-Bellman equation (HJB) coupled with a transport equation. The HJB equation describes the optimality conditions for the policy of the representative agent, and the transport equation captures the evolution of the population distribution. Furthermore, the optimal policy computed by the representative agent constitutes an ϵ-Nash equilibrium when all the agents in the *finite* N-population deploy this policy, for some sufficiently large N. The existence and uniqueness of such an optimal policy have been established in [13].

Although the discretization of continuous MFG has been studied in prior works [17], direct analysis results for discrete-time and finite state-action space MFG are still relatively sparse. One of the challenges in the finite MFG is the absence of regularity conditions regarding the mean field [18]. That is, when the population mean field changes slightly, the corresponding optimal policy for the representative agent could change drastically [19]. Previous works have used Boltzmann policies [20] and projection to meshed probability measure spaces [18] to avoid such issues. More recent works directly introduced a relative entropy term to the reward structure to provide regularity conditions. The existence of stationary entropy-regularized MFE was examined in [21]. The authors in [19] studied transient MFGs with finite horizon. They used entropy-regularization to stabilize the iterative algorithm and reduced regularization over time to search for the equilibrium of the original MFG.

Different from these previous works, our work considers the reference policy in the entropy-regularization as an extra feature that allows us to control the behavior of a large population. Consider the situation, for instance, where a "coordinator" of a large population of agents desires to impose a certain group behavior, but it does not have access to the actual rewards. The agents are selfish and not concerned about the overall performance of the population, but they have access to their own actual rewards. If the coordinator designs a policy and forces the whole population to adopt it, the result could be undesirable, as such a policy will not be informed of the actual agent rewards. Without a reference policy, however, agents may fail to find the MFE, or they may find a MFE that does not induce a desirable group behavior. We argue that the entropy-regularized MFG is a good framework to model such scenarios. Through a resource allocation example, we show that by adjusting the multiplier of the regularization term, we can produce a tuneable behavior of the population that trades off between the desirable group behavior and the cumulative rewards each agent collects.

Contributions: In this work, we formulate a game of N-homogeneous agents and show that under pair-wise coupled rewards, the state of such a system can be *exactly* represented by a distribution over the state space. We then consider the limiting infinite population game and introduce entropy

[1] Yue Guan is a PhD student with the School of Aerospace Enginnering, Georgia Institute of Technology, Atlanta, GA, USA yguan44@gatech.edu

[2]Mi Zhou is a PhD student with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. mzhou91@gatech.edu

[3]Ali Pakniyat is an Assistant Professor with the department of Mechanical Engineering, University of Alabama, Tuscaloosa, AL, USA. apakniyat@ua.edu

[4]Panagiotis Tsiotras is the David & Andrew Lewis Chair Professor with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. tsiotras@gatech.edu

regularization to construct contractive operators to find the unique regularized MFE. We consider a special class of MFGs where the agents have coupled rewards but decoupled dynamics. For this class of MFGs, we streamline and simplify the convergence proof of [19]. Finally, we verify the theoretical results through a numerical example for a resource allocation problem and demonstrate that certain performance is not possible without entropy-regularization and a properly selected prior.

II. Problem Formulation

In this work we follow the notation established in [19]. Consider a large population game consisting of N homogeneous agents, where typically $N \gg 1$. We define the game through the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}^1, \dots, \mathcal{R}^N, N, T \rangle$. The game is over a discrete-time with a finite horizon T. In this formulation, we assume that all agents share the same finite state space $\mathcal{S}$ and the same finite action space $\mathcal{A}$. At time t, agent i takes an action $a_t^i \in \mathcal{A}$ and transitions from state s_t^i to s_{t+1}^i according to the dynamics $\mathcal{T}$, which we discuss in more detail later. As a consequence of its own action, as well as the actions of all other agents, agent i receives a reward $\mathcal{R}_t^i(s_t^i, a_t^i, s_t^{-i})$, where s_t^{-i} is a shorthand notation for $(s_t^1, \dots, s_t^{i-1}, s_t^{i+1}, \dots, s_t^N)$. Each agent follows a (time-varying) Markov policy $\pi^i = \{\pi_t^i\}_{t=0}^T$, such that at each time step t, this policy is a mapping $\pi_t^i : \mathcal{S} \times \mathcal{A} \to [0,1]$ that satisfies $\sum_a \pi_t^i(a|s^i) = 1$ for all $s^i \in \mathcal{S}$. We use Π to denote the space of admissible policies. Overloading the notation, we denote the set of discrete time steps as $T = \{0, \dots, T\}$.

a) Dynamics: We assume that all agents have the same decoupled dynamics $\mathcal{T} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$. The value of $\mathcal{T}(s_{t+1}|s_t, a_t)$ represents the probability of transitioning from state s_t to state s_{t+1} under action a_t. The function $\mathcal{T}$ satisfies $\sum_{s' \in \mathcal{S}} \mathcal{T}(s'|s,a) = 1$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. In the sequel, we use the notation $\mathcal{T}(\cdot|s_t^i, \pi_t^i)$ to denote the distribution of agent i's state at the next time step by following the policy π_t^i. Formally,

$$\mathcal{T}(s_{t+1}^i|s_t^i, \pi_t^i) = \sum_{a_t^i \in \mathcal{A}} \mathcal{T}(s_{t+1}^i|s_t^i, a_t^i)\, \pi_t^i(a_t^i|s_t^i). \quad (1)$$

Assuming that all agents start with the same initial state distribution μ_0 and that each agent i deploys a policy π^i, we have N independent processes, where the i-th process follows the dynamics

$$\begin{cases} s_0^i \sim \mu_0 \\ s_{t+1}^i \sim \mathcal{T}\big(\,\cdot\,|s_t^i, \pi_t^i\big), \quad t = 0, \dots, T. \end{cases} \quad (2)$$

b) Rewards: We consider the following reward structure. For each agent i, its reward at time t is defined as

$$\mathcal{R}_t^i(s_t^i, a_t^i, s_t^{-i}) = \Theta_t\Big(\frac{1}{N}\sum_{k=1}^N L_t(s_t^i, a_t^i, s_t^k)\Big), \quad (3)$$

where $\Theta_t : \mathbb{R} \to \mathbb{R}$ is a real-valued function, and L_t is a pairwise state-coupled reward function.

We make the following assumptions on Θ_t and L_t.

Assumption 1. *The function Θ_t is uniformly globally Lipschitz continuous in t with Lipschitz constant K_Θ. That is, for all $x, y \in \mathbb{R}$, and for all $t \in T$, $|\Theta_t(x) - \Theta_t(y)| \le K_\Theta\,|x-y|$.*

Assumption 2. *The function L_t is bounded, that is, there exists $L_{\max}$ such that $|L_t(s,a,s')| \le L_{\max}$, for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$ and $t \in T$.*

We can then define the maximum magnitude of the reward as $\mathcal{R}_{\max} = \max_{|x| \le L_{\max}} |\Theta_t(x)|$.

Note that the reward structure in (3) is indifferent to the ordering of the agents. As a consequence, agent i's reward can be computed, given only the fraction of agents at each state. This observation motivates the *aggregation* of the state of the whole system $(s_t^1, \dots, s_t^N)$ to an empirical distribution of the agents' states.

c) Empirical distribution: For the N processes in (2), we define the empirical distribution at time t as

$$\mu_t^N(s) = \frac{1}{N}\sum_{k=1}^N \mathbb{1}_s(s_t^k), \quad s \in \mathcal{S}, \quad (4)$$

where $\mathbb{1}_x$ is the indicator function, i.e., $\mathbb{1}_x(y) = 1$ if $y = x$, and 0 otherwise. The empirical distribution flow is defined as $\mu^N = \{\mu_t^N\}_{t=0}^T$. Note that μ_t^N is a probability measure over $\mathcal{S}$. We denote the space of probability measures over $\mathcal{S}$ as $\mathcal{P}(\mathcal{S})$. Then, $\mathcal{M} = (\mathcal{P}(\mathcal{S}))^{T+1}$ is the space of the probability measure flows and $\mu^N \in \mathcal{M}$.

d) Metric spaces: To establish the convergence results later, we first present the metric spaces of the distribution flow $\mathcal{M}$ and the policy space Π.

We use total variation as the metric for the probability measure space $\mathcal{P}(\mathcal{X})$ [22]. When $\mathcal{X}$ is finite, the total variation between $\nu, \nu' \in \mathcal{P}(\mathcal{X})$ is given by

$$\mathrm{d_{TV}}(\nu, \nu') = \frac{1}{2}\sum_{x \in \mathcal{X}} |\nu(x) - \nu'(x)| = \frac{1}{2}\big\|\nu(x) - \nu'(x)\big\|_1.$$

We equip both $\mathcal{M}$ and Π with the sup metric induced by the total variation. That is, for $\mu, \mu' \in \mathcal{M}$, we define

$$\mathrm{d}_{\mathcal{M}}(\mu, \mu') = \max_{t \in T} \mathrm{d_{TV}}\ (\mu_t, \mu_t'), \quad (5)$$

and for policies $\pi, \pi' \in \Pi$

$$\mathrm{d}_\Pi(\pi, \pi') = \max_{t \in T} \max_{s \in \mathcal{S}}\ \mathrm{d_{TV}}\left(\pi_t(s), \pi_t'(s)\right), \quad (6)$$

where $\pi_t(s) \in \mathcal{P}(\mathcal{A})$ is the distribution the policy assign over actions when an agent is at state s.

It can be shown that both $(\mathcal{M}, \mathrm{d}_{\mathcal{M}})$ and (Π, d_Π) are complete metric spaces.

e) Distribution induced rewards: Due to symmetry, the state-coupled reward in (3) can be characterized through the

empirical distribution. That is,

$$\begin{aligned} L_t(s_t^i, a_t^i, \mu_t^N) &\triangleq \sum_{s' \in \mathcal{S}} L_t(s_t^i, a_t^i, s') \mu_t^N(s') \\ &= \sum_{s' \in \mathcal{S}} L_t(s_t^i, a_t^i, s') \Big(\frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{s'}(s_t^k) \Big) \\ &= \frac{1}{N} \sum_{k=1}^{N} \Big(\sum_{s' \in \mathcal{S}} L_t(s_t^i, a_t^i, s') \mathbb{1}_{s'}(s_t^k) \Big) \\ &= \frac{1}{N} \sum_{k=1}^{N} L_t(s_t^i, a_t^i, s_t^k). \end{aligned}$$

With this observation, we define the reward for each agent induced by the empirical distribution μ_t^N as

$$\mathcal{R}_t^i(s_t^i, a_t^i, \mu_t^N) \triangleq \Theta_t\Big(L_t(s_t^i, a_t^i, \mu_t^N) \Big). \tag{7}$$

Lemma 1. *The reward function $\mathcal{R}_t^i(s, a, \nu)$ in (7) is globally Lipschitz with respect to the probability measure $\nu \in \mathcal{P}(\mathcal{S})$, with Lipschitz constant $2K_\Theta\, L_{\max}$.*

Proof. One can verify that, for all $\nu, \nu' \in \mathcal{P}(\mathcal{S})$,

$$|L_t(s, a, \nu) - L_t(s, a, \nu')| \leq 2L_{\max} \mathrm{d}_{\mathrm{TV}}(\nu, \nu').$$

Then, through composition with Lipschitz function Θ_t, the desired Lipschitz constant for $\mathcal{R}_t^i(s, a, \nu)$ can be shown. □

f) Expected cumulative reward: The expected cumulative reward of agent i induced by the policies of the agents $(\pi^1, \ldots, \pi^N)$ is given by

$$J^{i,N}(\pi^i, \pi^{-i}) = \mathbb{E}_{\mu_0}\Big[\sum_{t=0}^{T} \mathcal{R}_t^i(s_t^i, a_t^i, \mu_t^N) \Big], \tag{8}$$

where the expectation is taken over the trajectories of the system where each agent i starts with the initial distribution μ_0 and follows policy π^i. Each agent's objective is to select a policy that maximizes its own expected cumulative reward. We therefore have the following N coupled optimization problems:

$$\max_{\pi^i \in \Pi} \; J^{i,N}(\pi^i, \pi^{-i}), \qquad i = 1, \ldots, N. \tag{9}$$

One of the most common solution concepts for games is the Nash equilibrium [23].

Definition 1. *A Nash equilibrium is a tuple $(\pi^{1*}, \ldots, \pi^{N*})$ such that for all $i = 1, \ldots, N$,*

$$J^{i,N}(\pi^i, \pi^{-i*}) \leq J^{i,N}(\pi^{i*}, \pi^{-i*}), \quad \forall \pi^i \in \Pi.$$

Definition 2. *For $\epsilon \geq 0$, an ϵ-Nash equilibrium is a tuple $(\pi^{1*}, \ldots, \pi^{N*})$ such that for all $i = 1, \ldots, N$,*

$$J^{i,N}(\pi^i, \pi^{-i*}) \leq J^{i,N}(\pi^{i*}, \pi^{-i*}) + \epsilon, \quad \forall \pi^i \in \Pi. \tag{10}$$

In other words, when operating at an ϵ-Nash equilibrium, any unilateral deviation can improve an agent's performance by at most ϵ.

In this paper, we restrict our attention to identical policies for all agents.

Assumption 3. *$\pi_t^i = \pi_t^j$ for all $t \in T$ and $i, j \in \{1, \ldots, N\}$.*

This simplifying assumption leads, in general, to a loss in performance. Readers can refer to [24] for an example. However, identical policy is a standard assumption in the literature on large scale systems for reasons of simplicity and robustness [25]. In the mean field setting, such an assumption is needed to fully exploit the potential of the symmetric structure in the problem formulation.

In light of Assumption 3, henceforth, we will drop the superscripts on the policies and denote the policy used by all agents at time t as π_t.

III. Mean Field Approximation

When N approaches infinity, the limiting game constitutes the mean field game. The mean field is defined as the empirical distribution of the infinite population. We denote the mean field at time t as μ_t. Aside from describing the infinite population, the introduction of the mean field also has attractive computational benefits. Recall that the empirical distribution in (4) is a *random* vector. To properly evaluate the expected reward with the nonlinear function Θ_t in (3), one needs the distribution of μ_t^N at each time step. With a general dynamics $\mathcal{T}$, the propagation of the distribution of μ_t^N could be computationally expensive. On the other hand, under the identical policy π used by all agents, the trajectory of the mean field is deterministic [21]. Furthermore, μ_t follows a simple propagation rule:

$$\mu_{t+1} = \mu_t\left[\mathcal{T}(\pi_t)\right], \tag{11}$$

where $[\mathcal{T}(\pi_t)]$ is a right stochastic matrix constructed based on (1). We refer to the time sequence $\mu = \{\mu_t\}_{t=0}^T \in \mathcal{M}$ as the *mean field flow*.

It is tempting to approximate the empirical distribution of a finite N-population with the mean field. Indeed, as we show in Section IV, the empirical distribution converges to the mean field as the number of agents approaches infinity with rate $\mathcal{O}(1/\sqrt{N})$.

A. Representative Agent

Before tackling the large population game with N agents, we consider the limiting infinite population case via specifying the behaviour of the representative agent. Since the effect of dynamic uncertainties on all agents takes the same form, the mean field flow μ can be solely generated from the representative agent's dynamics and its policy. Assuming that the mean field flow μ is known and fixed, this yields a standard MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}_\mu \rangle$. The state space, action space and the transitions of the induced MDP come directly from the original game. The reward induced by the mean field μ is given by

$$\mathcal{R}_{\mu,t}(s, a) = \Theta_t\big(L_t(s, a, \mu_t)\big). \tag{12}$$

The representative agent can then maximize its expected cumulative reward given the mean field flow μ as follows:

$$J_\mu(\pi^*) = \max_{\pi \in \Pi} \; \mathbb{E}_{\mu_0}\left[\sum_{t=0}^{T} \mathcal{R}_{\mu,t}(s_t, a_t) \right]. \tag{13}$$

Note that the optimal policy obtained depends on μ. We use the operator $\mathcal{B}_{\rm opt} : \mathcal{M} \to \Pi$ to denote the mapping from the mean field flow to an optimal policy of the induced MDP[1]:

$$\pi^* = \mathcal{B}_{\rm opt}(\mu). \tag{14}$$

When all agents deploy the policy π of the representative agent, a new mean field flow is induced and can be propagated via (11) starting from μ_0. We use the operator $\mathcal{B}_{\rm prop} : \Pi \to \mathcal{M}$ to denote this propagation. That is,

$$\mu = \mathcal{B}_{\rm prop}(\pi). \tag{15}$$

The mean field equilibrium (MFE) is defined as a consistent pair of $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$ such that

$$\pi^* = \mathcal{B}_{\rm opt}(\mu^*), \quad \mu^* = \mathcal{B}_{\rm prop}(\pi^*). \tag{16}$$

The existence of such consistent pair can be established through a Brouwer's fixed point argument [26]. One may attempt to use fixed-point iterations to find a solution to (16). Unfortunately, the composed mean-field equilibrium operator $\Gamma = \mathcal{B}_{\rm prop} \circ \mathcal{B}_{\rm opt}$ is only non-expensive and not contractive, in general. One may refer to the examples and the proof of this statement in [19].

IV. Entropy-Regularized Mean Field Games

Entropy regularization techniques have been used extensively to stabilize learning algorithms and to reduce maximization bias [27]–[29]. The extra entropy cost introduced to the reward structure prevents abrupt policy changes between iterations. In the context of mean field games, entropy-regularization stabilizes the system such that a small change in the mean field flow does not cause an abrupt change in the optimal policy of the representative agent, thus inducing a contractive MFE operator.

Given a reference policy $\rho \in \Pi$ such that $\rho_t(a|s) > 0$ for all $t \in T, s \in \mathcal{S}$, and $a \in \mathcal{A}$, we introduce an entropy regularization term to (13) as follows:

$$J_\mu^{\rm KL}(\pi;\rho) = \mathbb{E}_{\mu_0}\left[\sum_{t=0}^{T}\Big(\mathcal{R}_{\mu,t}(s_t,a_t) - \frac{1}{\beta}\log\frac{\pi_t(a_t|s_t)}{\rho_t(a_t|s_t)}\Big)\right], \tag{17}$$

where $\beta > 0$ is the inverse temperature, and it is a design parameter. The reference policy can encode any preference one has about the population behavior. When no such information encoding is needed, one can simply use a uniform prior. When β is small, the regularization term in (17) is dominant, and π approaches the reference ρ. When β is large, the agent is allowed to diverge from the reference policy to increase the rewards collected. As a consequence, the optimal π approaches a greedy policy produced by $\mathcal{B}_{\rm opt}$ as in (14).

[1]In general, $\mathcal{B}_{\rm opt}$ is a set-valued function, since the optimal policy of an MDP needs not be unique. One may refer to [19] for more details.

A. Optimization for the Regularized MDP

It can be shown that the *unique* optimal policy that solves $\max_\pi J_\mu^{\rm KL}(\pi;\rho)$ is given by the following (weighted) Boltzmann distribution [27]

$$\pi_{\mu,t}^{\rm KL}(a|s) = \frac{1}{Z_t(s)}\rho_t(a|s)\exp\left[\beta Q_{\mu,t}^{\rm KL}(s,a;\rho)\right], \tag{18}$$

where $Z_t(s)$ is the normalization factor, and the entropy-regularized state-action value function $Q_{\mu,t}^{\rm KL}$ can be obtained from the LogSumExp recursion:

$$Q_{\mu,t}^{\rm KL}(s,a;\rho) = \mathcal{R}_{\mu,t}(s,a) + \sum_{s'}\mathcal{T}(s'|s,a)\Big(\tfrac{1}{\beta}\log\sum_a \rho_t(a|s)\exp\left[\beta Q_{\mu,t+1}^{\rm KL}(s,a;\rho)\right]\Big),$$

with the boundary condition $Q_{\mu,T}^{\rm KL}(s,a;\rho) = \mathcal{R}_{\mu,T}(s,a)$. Let the entropy-regularized policy optimization be described by the operator $\mathcal{B}_{{\rm opt},\beta}^{\rm KL} : \mathcal{M} \to \Pi$. We can then define the entropy-regularized MFE (ER-MFE) operator $\Gamma_\beta^{\rm KL} : \mathcal{M} \to \mathcal{M}$ as

$$\Gamma_\beta^{\rm KL} = \mathcal{B}_{\rm prop} \circ \mathcal{B}_{{\rm opt},\beta}^{\rm KL}, \tag{19}$$

The regularized equilibrium is then defined as follows.

Definition 3. *The entropy-regularized mean field equilibrium (ER-MFE) is a consistent pair $(\pi^{{\rm KL}*}, \mu^{{\rm KL}*}) \in \Pi \times \mathcal{M}$ such that $\pi^{{\rm KL}*} = \mathcal{B}_{{\rm opt},\beta}^{\rm KL}(\mu^{{\rm KL}*})$ and $\mu^{{\rm KL}*} = \mathcal{B}_{\rm prop}(\pi^{{\rm KL}*})$.*

In the sequel, we establish the existence and uniqueness of the ER-MFE. The main goal is to show that the ER-MFE operator $\Gamma_\beta^{\rm KL}$ is contractive if the inverse temperature β is selected properly. The following derivation is more direct and easier to demonstrate than the one reported in [19]. The reason is that we are restricting ourselves to the case when the agents have decoupled dynamics. This case allows us to individually analyze the trajectory of each agent. As a consequence, the cross disturbance analysis for the ϵ-Nash argument is streamlined, since a single agent's deviation from the optimal policy does not directly impact the empirical distribution of the rest of the population. In addition, the decoupled dynamics gives us an expression for the Lipschitz constant of $\mathcal{B}_{\rm prop}$, while no such explicit formula for this constant is provied in [19].

B. Convergence Analysis

We first establish the Lipschitz continuity of the operators $\mathcal{B}_{\rm prop}$ and $\mathcal{B}_{{\rm opt},\beta}^{\rm KL}$ in the following Lemmas.

Lemma 2. *For all $\pi, \pi' \in \Pi$, we have that*

$$\mathrm{d}_{\mathcal{M}}(\mathcal{B}_{\rm prop}(\pi), \mathcal{B}_{\rm prop}(\pi')) \le K_{\rm prop}\ \mathrm{d}_\Pi(\pi,\pi'), \tag{20}$$

where

$$K_{\rm prop} = \frac{|S|(|S|^T - 1)}{|S| - 1}. \tag{21}$$

Proof. See the Appendix. □

The following two Lemmas are adopted from the results in [19].

Lemma 3 ([19]). *Under Assumptions 1 and 2, the entropy-regularized Q-function Q_μ^{KL} is Lipschitz with respect to μ, that is,*

$$\max_{t,s,a} \left| Q_{\mu,t}^{\mathrm{KL}}(s,a;\rho) - Q_{\mu',t}^{\mathrm{KL}}(s,a;\rho) \right| \leq K_Q^{\mathrm{KL}} \, \mathrm{d}_{\mathcal{M}}(\mu,\mu').$$

Furthermore, $K_Q^{\mathrm{KL}} = \max_{0\leq t\leq T} K_{Q,t}^{\mathrm{KL}}$, where $K_{Q,t}^{\mathrm{KL}}$ is defined via

$$K_{Q,t}^{\mathrm{KL}} = 2K_\Theta L_{\max} + \frac{\rho_{\max}\exp\left(2\beta_{\max}(T+1)\mathcal{R}_{\max}K_{Q,t+1}^{\mathrm{KL}}\right)}{\rho_{\min}}, \tag{22}$$

with boundary condition $K_{Q,T}^{\mathrm{KL}} = 2K_\Theta L_{\max}$, and $\rho_{\max} = \max_{t,s,a}\rho_t(a|s) > 0$, $\rho_{\min} = \min_{t,s,a}\rho_t(a|s) > 0$.

Lemma 4 ([19]). *Under Assumption 1, the entropy-regularized operator $\mathcal{B}_{\mathrm{opt}}^{\mathrm{KL}}$ is Lipschitz, that is,*

$$\mathrm{d}_\Pi\left(\mathcal{B}_{\mathrm{opt},\beta}^{\mathrm{KL}}(\mu), \mathcal{B}_{\mathrm{opt},\beta}^{\mathrm{KL}}(\mu')\right) \leq K_{\mathrm{opt},\beta}^{\mathrm{KL}} \, \mathrm{d}_{\mathcal{M}}(\mu,\mu'),$$

where,

$$K_{\mathrm{opt},\beta}^{\mathrm{KL}} = \frac{|\mathcal{A}|(|\mathcal{A}|-1)\beta\rho_{\max}^2}{2\rho_{\min}^2} K_Q^{\mathrm{KL}}. \tag{23}$$

The Lipschitz continuity in Lemma 4 guarantees that a small change in the mean field can only result in a small change in the optimal policy. With the Lipschitz constants of $\mathcal{B}_{\mathrm{prop}}$ and $\mathcal{B}_{\mathrm{opt},\beta}^{\mathrm{KL}}$, we arrive at the following result regarding the selection of β to ensure convergence.

Theorem 1. *The entropy-regularized mean-field equilibrium (ER-MFE) operator $\Gamma_\beta^{\mathrm{KL}} = \mathcal{B}_{\mathrm{prop}} \circ \mathcal{B}_{\mathrm{opt},\beta}^{\mathrm{KL}}$ is contractive for*

$$\beta < \min\left\{\beta_{\max}, \frac{2\rho_{\min}^2}{\rho_{\max}^2|\mathcal{A}|(|\mathcal{A}|-1)} \frac{1}{K_Q^{\mathrm{KL}} K_{\mathrm{prop}}}\right\}. \tag{24}$$

Proof. Choosing β as in (24) $K_{\mathrm{prop}}K_{\mathrm{opt},\beta}^{\mathrm{KL}} < 1$. Consequently, the ER-MFE operator $\Gamma_\beta^{\mathrm{KL}}$ in (19) is contractive. □

C. Error Bounds on the Mean Field Approximations

To motivate the proof of error bounds on the mean field approximation, we first present the following lemma, which characterizes the asymptotic convergence of the empirical distribution flow μ^N to the mean field flow μ as the number of agents N approaches infinity.

Lemma 5. *Suppose a mean field flow μ is induced by the representative agent with policy π. Let μ^N denote the empirical distribution of an N-agent system, where all agents deploy the same policy π. Then, for all $t \in T$,*

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(\mu_t^N, \mu_t\right)\right] = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \tag{25}$$

Proof. Since the dynamics is decoupled and all agents apply the same policy, the N processes in (2) are i.i.d. The convergence rate is then a result from the L^2 weak law of large numbers [30]. For details see the Appendix. □

Next, we show that the ER-MFE π^* for the infinite population game is an ϵ-Nash equilibrium for a finite N-agent regularized game.

Theorem 2. *Consider an ER-MFE (π^*, μ^*). Then, for all $\widetilde{\pi} \in \Pi$, we have*

$$J_{\mathrm{KL}}^{i,N}(\widetilde{\pi}, \pi^*) \leq J_{\mathrm{KL}}^{i,N}(\pi^*, \pi^*) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right), \tag{26}$$

where $J^{i,N}(\widetilde{\pi}, \pi^*)$ is the value induced when agent i applies policy $\widetilde{\pi}$ and all other agents apply policy π^*.

Before we present the proof for Theorem 2, we first establish the convergence of the deviated empirical distribution to the optimal mean field.

The N-agent trajectory under the optimal policy π^* is given by $s_{t+1}^i \sim \mathcal{T}\big(\cdot\,|s_t^i, \pi_t^*(s_t^i)\big)$ for $i = 1,\ldots,N$. Without loss of generality, we let agent 1 deviate from π^* by selecting the policy $\widetilde{\pi}$. The trajectory of the deviated N-agent system with agent 1 following $\widetilde{\pi}$ and all other agents following π^* is given by $\widetilde{s}_{t+1}^1 \sim \mathcal{T}\big(\cdot\,|\widetilde{s}_t^1, \widetilde{\pi}_t(\widetilde{s}_t^1)\big)$, and $\widetilde{s}_{t+1}^i \sim \mathcal{T}\big(\cdot\,|\widetilde{s}_t^i, \pi_t^*(\widetilde{s}_t^i)\big)$, for $i = 2,\ldots,N$. For both the optimal system and the deviated system, the agents' initial distributions are μ_0.

Lemma 6. *Let $\widetilde{\mu}^N$ denote the empirical distribution flow induced by agent 1 deviating to $\widetilde{\pi}$ while all other N-1 agents following the optimal policy π^*. Then, $\widetilde{\mu}_t^N$ converges to the optimal mean field μ_t^*. Furthermore, for all $t \in T$,*

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(\tilde{\mu}_t^N, \mu_t^*\right)\right] = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$$

Proof. The deviated empirical distribution is given by

$$\widetilde{\mu}_t^N(s) = \frac{1}{N}\sum_{k=1}^N \mathbb{1}_s(\tilde{s}_t^k) = \frac{1}{N}\mathbb{1}_s(\tilde{s}_t^1) + \frac{1}{N}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k).$$

Due to the decoupled dynamics, $\tilde{s}_t^k$ follows the same distribution as s_t^k for $k = 2,\ldots,N$. Then, the expected difference between the deviated empirical distribution and the original optimal mean field can be bounded as

$$\begin{aligned}
&\mathbb{E}\left|\tilde{\mu}_t^N(s) - \mu_t^*(s)\right| \\
&\leq \mathbb{E}\left|\frac{1}{N}\mathbb{1}_s(\tilde{s}_t^1)\right| + \mathbb{E}\left|\frac{1}{N}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k) - \mu_t^*(s)\right| \\
&\leq \frac{1}{N} + \mathbb{E}\left|\frac{1}{N-1}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k) - \mu_t^*(s)\right| \qquad (27)\\
&\qquad\qquad + \mathbb{E}\left|\frac{1}{N(N-1)}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k)\right| \\
&\leq \frac{1}{N} + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \frac{1}{N} = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \qquad (28)
\end{aligned}$$

The second term in (27) corresponds to the scenario of $N-1$ agents all applying the optimal policy π^*. By Lemma 5, we obtain the presented convergence rate. Finally, from (28), we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(\tilde{\mu}_t^N, \mu_t^*\right)\right] &= \mathbb{E}\sum_{s\in\mathcal{S}}|\tilde{\mu}_t^N(s), \mu_t^*(s)| \\
&= \sum_{s\in\mathcal{S}}\mathbb{E}\,|\tilde{\mu}_t^N(s) - \mu_t^*(s)| = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \qquad (29)
\end{aligned}$$

which yields the desired result. □

Now, we are ready to provide a proof for Theorem 2 using cross disturbance analysis similar to [13].

Proof of Theorem 2. For the N-agent system, the value of agent 1 induced by its policy deviation is bounded as

$$J_{\mathrm{KL}}^{1,N}(\widetilde{\pi},\pi^*) = \mathbb{E}\sum_{t=0}^{T}\left[\mathcal{R}_t\big(\widetilde{s}_t^1,\widetilde{\pi}_t,\widetilde{\mu}_t^N\big) - \frac{1}{\beta}\log\frac{\widetilde{\pi}_t(a_t|\widetilde{s}_t^1)}{\rho_t(a_t|\widetilde{s}_t^1)}\right]$$

$$\leq \mathbb{E}\sum_{t=0}^{T}\left[\mathcal{R}_t\big(\widetilde{s}_t^1,\widetilde{\pi}_t,\mu_t^*\big) - \frac{1}{\beta}\log\frac{\widetilde{\pi}_t(a_t|\widetilde{s}_t^1)}{\rho_t(a_t|\widetilde{s}_t^1)}\right] + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \tag{30}$$

$$\leq \mathbb{E}\sum_{t=0}^{T}\left[\mathcal{R}_t\big(s_t^1,\pi_t^*,\mu_t^*\big) - \frac{1}{\beta}\log\frac{\pi_t^*(a_t|s_t^1)}{\rho_t(a_t|s_t^1)}\right] + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \tag{31}$$

$$= J_{\mathrm{KL}}(\pi^*,\pi^*) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \tag{32}$$

In (30), we used the convergence result in Lemma 6 and the Lipschitz continuity of $\mathcal{R}_t$ to replace $\widetilde{\mu}_t^N$ with μ_t^*. To arrive at (31), we used the optimality of π^* for the regularized MDP induced by μ^*.

Up to this point, we have shown that the difference between the value of the deviated N-agent system and the optimal value of the *infinite population system* is bounded by $\mathcal{O}(1/\sqrt{N})$. Next, we show that the value of the finite N-agent system under the identical optimal policy π^* for all agents is also $\mathcal{O}(1/\sqrt{N})$-close to the optimal value of the infinite population system.

$$J_{\mathrm{KL}}^{1,N}(\pi^*,\pi^*) = \mathbb{E}\sum_{t=0}^{T}\left[\mathcal{R}_t\big(s_t^1,\pi_t^*,\mu_t^N\big) - \frac{1}{\beta}\log\frac{\pi_t^*(a_t|s_t^1)}{\rho_t(a_t|s_t^1)}\right]$$

$$\leq \mathbb{E}\sum_{t=0}^{T}\left[\mathcal{R}_t\big(s_t^1,\pi_t^*,\mu_t^*\big) - \frac{1}{\beta}\log\frac{\pi_t^*(a_t|s_t^1)}{\rho_t(a_t|s_t^1)}\right] + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$$

$$= J_{\mathrm{KL}}(\pi^*,\pi^*) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big),$$

where the inequality is a result of the Lipschitz continuity of $\mathcal{R}_t$ and the convergence rate in Lemma 5. One can also lower bound $J_{\mathrm{KL}}^{1,N}(\pi^*,\pi^*)$ and get

$$\Big|\, J_{\mathrm{KL}}^{1,N}(\pi^*,\pi^*) - J_{\mathrm{KL}}(\pi^*,\pi^*)\Big| = \mathcal{O}\big(\frac{1}{\sqrt{N}}\big). \tag{33}$$

Combining (32) and (33), we can show that

$$J_{\mathrm{KL}}^{1,N}(\widetilde{\pi},\pi^*) \leq J_{\mathrm{KL}}^{1,N}(\pi^*,\pi^*) + \mathcal{O}\big(\frac{1}{\sqrt{N}}\big).$$

Since all agents are homogeneous, the same result applies to all agents, thus completing the proof. □

V. Numerical Example

In this section, a resource allocation problem is formulated as a mean field game to verify the previous theoretical results. Consider a resource allocation problem over the graph $\langle \mathcal{S},\mathcal{E}\rangle$ shown in Fig. 1. We use $\mathcal{S}$ and $\mathcal{E}$ to denote the set of nodes and edges of the graph, respectively. A large group of agents

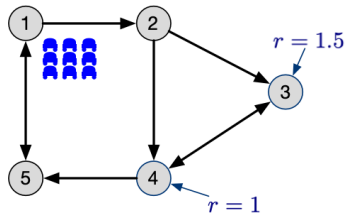


Fig. 1. Graph for a resource allocation problem. All nodes contain self-loops, i.e., the agent can always choose to stay at its current node.

traverses through the graph to collect the rewards assigned at the terminal time step T, and no running reward is assigned. At time T, if an agent is at state 3, then it receives a reward of 1.5. If it is at state 4, it receives a reward of 1. Otherwise, the agent receives no reward. At the same time, the agents are penalized for staying at a node with a large population density. In summary, we have for the state-coupled rewards

$$L_T(s^i,a^i,s^k) = \underbrace{1.5\mathbb{1}_{s_3}(s^i) + \mathbb{1}_{s_4}(s^i)}_{\text{rewards at states 3 and 4}} - \underbrace{\mathbb{1}_{s^i}(s^k)}_{\substack{\text{penalty of sharing}\\ \text{node with agent } k}},$$

$$L_t(s^i,a^i,s^k) = 0, \qquad \text{for all } t = 0,\dots,T-1.$$

We set the nonlinear function in (3) to $\Theta_T(x) = x^2$. Each agent at state s can choose one of the adjacent states (graph node) s' to visit at the next time step. That is, the action space $\mathcal{A}(s)$ at state s is all the states s' such that $(s,s') \in \mathcal{E}$.

If we directly use fixed-point iterations without entropy-regularization to solve this mean field game, the algorithm fails to converge. For the first iteration, the agents use a policy that concentrates the whole population at node 4 for the extra reward. At the next iteration, the penalty for staying at node 4 is high for the representative agent based on the mean field flow from the previous iteration. The representative agent then constructs a policy to visit node 3. In summary, the policy found by the $\mathcal{B}_{\mathrm{opt}}$ oscillates between reaching node 3 and reaching node 4 at T.

Suppose now that a coordinator can send a command to the group of agents, and it decides that both states 3 and 4 need to be occupied by some agents, but she does not have access to the actual rewards available to the agents at these two nodes. Consequently, the coordinator can, at most, provide a reference policy to *guide* the agents to node 3 and 4, but the decision of which node is more rewarding to occupy can only be made by the agents themselves. We constructed a reference policy ρ that commands the agents to move to nodes 3 and 4. For example, $\rho(s_4|s_2) = \rho(s_3|s_2) = 0.5$ and $\rho(s_5|s_4) = 0.01$. This reference policy promotes the agents to move from state 2 to states 3 and 4, while discourages agents to move from state 4 to state 5. If the reference policy is directly applied by the agents, then the final distribution at nodes 3 and 4 are roughly the same, which is uninformed, as it does not reflect the difference in the rewards.

We now use the constructed reference policy to form the entropy-regularized MFG and solve it using the operator $\Gamma_\beta^{\mathrm{KL}}$ in (19) for two different values of β. The algorithm converges and the population distribution over the nodes is depicted in Fig. 2. Recall that a larger β means less

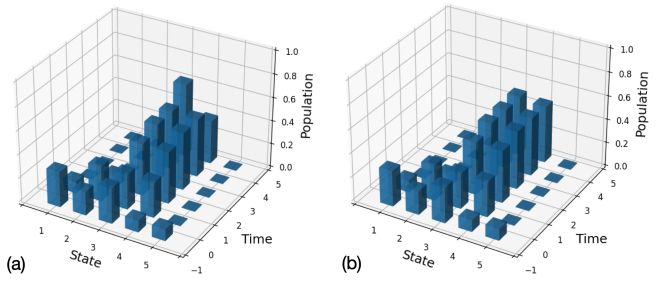


Fig. 2. Population distribution over time. Scenario in (a) has the inverse temperature of $\beta = 3$, and scenario in (b) has $\beta = 0.1$.

regularization in the reward structure. In Fig. 2(a), with a large β, the agents chase mainly the rewards, and the reference policy from the coordinator has little effect. In this scenario, the agents concentrate at node 3 upto the point when an additional number of agents at the same node would result in a penalty that diminishes the reward advantage that node 3 has over node 4. In Fig. 2(b), the value of β is small and the reference policy dominates. The agents start to ignore the reward advantage that node 3 has, and follow the reference policy instead. The parameter β enables us to generalize the behavior beyond these two extremes and to cover a continuous spectrum of population behavior.

Finally, to verify Theorem 2, we fixed the last $N-1$ agents' policy to the ER-MFE, and we computed the distribution of the random vector μ^N. We then let the first agent optimize the entropy-regularized MDP and compared the difference between its newly-optimized performance and the performance should it adopted the ER-MFE. A log-log plot of performance gain vs. number of agents is presented in Fig. 3. The performance gain trend is bounded by the reference line with a slope of -0.5, which verifies our claim of the $\mathcal{O}(1/\sqrt{N})$ convergence rate.

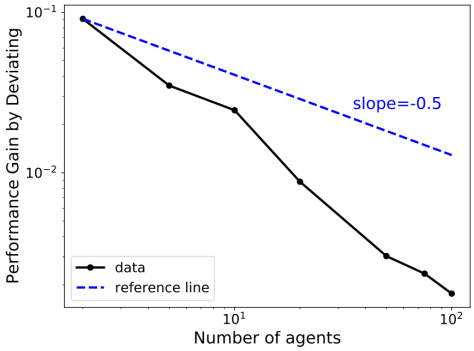


Fig. 3. Log-log plot of performance gain by an agent unilaterally deviating in a finite population.

VI. Conclusion

In this article, an entropy-regularized mean field-game with finite state-action space in a discrete time setting was formulated and analyzed. We demonstrated that entropy-regularization provides the regularity conditions that the standard MFG lacks. The condition for a contractive entropy-regularized mean-field equilibrium operator is presented. Furthermore, we provided a streamlined proof for the performance bound of the entropy-regularized MFE in a finite N-agent game. Through a resource allocation example, we demonstrated that the reference policy can be used to control the behavior of a large population, and the parameter β allows us to cover a continuous spectrum of population behaviors. Future work will involve extending the approach to the case of two large teams of agents competing against each other, modeled as a zero-sum game while the dynamics of agents within each team evolves as a mean field game.

Appendix

Proof of Lemma 2. Consider two policies $\pi, \pi' \in \Pi$ and the corresponding propagated mean field flows $\mu = \mathcal{B}_{\text{prop}}(\pi)$ and $\mu' = \mathcal{B}_{\text{prop}}(\pi')$. Then, at time step $t+1$, we have

$$\begin{aligned}
&\Big|\mu_{t+1}(s') - \mu'_{t+1}(s')\Big| \\
&= \Big|\sum_s \mu_t(s)\mathcal{T}(s'|s,\pi_t) - \sum_s \mu'_t(s)\mathcal{T}(s'|s,\pi'_t)\Big| \\
&\le \Big|\sum_s \mu_t(s)\mathcal{T}(s'|s,\pi_t) - \sum_s \mu_t(s)\mathcal{T}(s'|s,\pi'_t)\Big| \qquad \text{(A)} \\
&+ \Big|\sum_s \mu_t(s)\mathcal{T}(s'|s,\pi'_t) - \sum_s \mu'_t(s)\mathcal{T}(s'|s,\pi'_t)\Big|. \qquad \text{(B)}
\end{aligned}$$

For (A), we have

$$\begin{aligned}
\text{(A)} &= \Big|\sum_s \mu_t(s)\sum_a \mathcal{T}(s'|s,a)\big(\pi_t(a|s) - \pi'_t(a|s)\big)\Big| \\
&\le \sum_s \mu_t(s)\sum_a \mathcal{T}(s'|s,a)\Big|\big(\pi_t(a|s) - \pi'_t(a|s)\big)\Big| \\
&\le \sum_s \mu_t(s)\sum_a \Big|\big(\pi_t(a|s) - \pi'_t(a|s)\big)\Big| \\
&= 2\sum_s \mu_t(s)\mathrm{d}_{\text{TV}}(\pi_t(s),\pi'_t(s)) \le 2\mathrm{d}_\Pi(\pi,\pi').
\end{aligned}$$

For (B), we have

$$\begin{aligned}
\text{(B)} &\le \sum_s \mathcal{T}(s'|s,\pi'_t)\Big|\mu_t(s) - \mu'_t(s)\Big| \\
&\le \sum_s \Big|\mu_t(s) - \mu'_t(s)\Big| = 2\mathrm{d}_{\text{TV}}(\mu_t,\mu'_t).
\end{aligned}$$

Combining (A) and (B), we have

$$\begin{aligned}
\mathrm{d}_{\text{TV}}(\mu_{t+1},\mu'_{t+1}) &= \frac{1}{2}\sum_{s'\in\mathcal{S}} |\mu_{t+1}(s') - \mu'_{t+1}(s')| \\
&\le |\mathcal{S}|\big(\mathrm{d}_\Pi(\pi,\pi') + \mathrm{d}_{\text{TV}}(\mu_t,\mu'_t)\big).
\end{aligned}$$

For time step $t=0$, we assumed that $\mu_0 = \mu'_0$. Consequently, $\mathrm{d}_{\text{TV}}(\mu_0,\mu'_0) = 0$. Through induction, one can show that

$$\mathrm{d}_{\text{TV}}(\mu_t,\mu'_t) \le \frac{|\mathcal{S}|(|\mathcal{S}|^t - 1)}{|\mathcal{S}| - 1}\mathrm{d}_\Pi(\pi,\pi'), \qquad (34)$$

Since $|\mathcal{S}| > 1$, it follows that (34) is an increasing sequence of t. Consequently, we have

$$\begin{aligned}
\mathrm{d}_{\mathcal{M}}(\mathcal{B}_{\text{prop}}(\pi),\mathcal{B}_{\text{prop}}(\pi')) &= \mathrm{d}_{\mathcal{M}}(\mu,\mu') = \max_{t\in T}\mathrm{d}_{\text{TV}}(\mu_t,\mu'_t) \\
&\le \frac{|\mathcal{S}|(|\mathcal{S}|^T - 1)}{|\mathcal{S}| - 1}\mathrm{d}_\Pi(\pi,\pi').
\end{aligned}$$

□

Proof of Lemma 5. Let the time step $t \in T$ and the state $s \in \mathcal{S}$ be fixed. Recall that for each state s, the fraction of the agent population in that state is given by

$$\mu_t^N(s) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_s(s_t^i). \tag{35}$$

Define $X_s^i = \mathbb{1}_s(s_t^i)$. Since the dynamics are decoupled and all agents use the same policy, with the same initial distribution μ_0, X_s^i are i.i.d. random variables with mean $\mathbb{E}\left[X_s^i\right]$. As $\mu_t^N(s)$ is the sample mean of X_s^i, from the strong law of large numbers [30], we have $\mu_t^N(s) \xrightarrow{a.s.} \mathbb{E}\left[X_s^i\right]$ as $N \to \infty$. Consequently, we have that the mean field satisfies $\mathbb{P}\left\{\mu_t(s) - \mathbb{E}\left[X_s^i\right] \neq 0\right\} = 0$. The variance of X_s^i is then $\mathrm{Var}(X_s^i) = \mathbb{E}\left[X_s^i\right] - \left(\mathbb{E}\left[X_s^i\right]\right)^2 = \mu_t(s)\left(1-\mu_t(s)\right)$. Here, we regarded $\mu_t(s)$ as a deterministic number and use the property $\mathbb{E}\left[(X_s^i)^2\right] = \mathbb{E}\left[X_s^i\right]$ as X_s^i is an indicator function. Furthermore,

$$\begin{aligned}
\mathbb{E}\left\|\mu_t^N - \mu_t\right\|_2^2 &= \mathbb{E}\sum_{s\in\mathcal{S}} |\mu_t^N(s) - \mu_t(s)|^2 \\
&= \mathbb{E}\sum_{s\in\mathcal{S}} \Big| \frac{1}{N}\sum_{i=1}^{N}\left(X_s^i - \mathbb{E}\left[X_s^i\right]\right)\Big|^2 \\
&\leq \frac{1}{N}\sum_{s\in\mathcal{S}} \mathrm{Var}(X_s^i) = \frac{1}{N}\sum_{s\in\mathcal{S}} \mu_t(s)(1-\mu_t(s)) \\
&= \frac{1}{N}(1 - \|\mu_t\|_2^2) \leq \frac{1}{N}.
\end{aligned}$$

By Jensen's inequality, we have $\mathbb{E}\left\|\mu_t^N - \mu_t\right\|_2 \leq 1/\sqrt{N}$. Finally, and since $\|\mu_t^N - \mu_t\|_1 \leq \sqrt{|\mathcal{S}|}\ \|\mu_t^N - \mu_t\|_2$, it follows that

$$\begin{aligned}
\mathbb{E}[\mathrm{d_{TV}}(\mu_t^N - \mu_t)] &= \mathbb{E}\Big[\frac{1}{2}\sum_s |\mu_t^N(s) - \mu_t^*(s)|\Big] \\
&= \frac{1}{2}\mathbb{E}\|\mu - \mu'\|_1 = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big),
\end{aligned}$$

which yields the desired result. □

REFERENCES

[1] S. Berman, A. Halász, M. A. Hsieh, and V. Kumar, "Optimized stochastic policies for task allocation in swarms of robots," *IEEE Transactions on Robotics*, vol. 25, no. 4, pp. 927–937, 2009.

[2] A. Prorok, M. A. Hsieh, and V. Kumar, "The impact of diversity on optimal control policies for heterogeneous robot swarms," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 346–358, 2017.

[3] G. A. Korsah, A. Stentz, and M. B. Dias, "A comprehensive taxonomy for multi-robot task allocation," *The International Journal of Robotics Research*, vol. 32, no. 12, pp. 1495–1512, 2013.

[4] K. Elamvazhuthi and S. Berman, "Mean-field models in swarm robotics: A survey," *Bioinspiration & Biomimetics*, vol. 15, no. 1, p. 015001, 2019.

[5] D. H. Kim, H. Wang, and S. Shin, "Decentralized control of autonomous swarm systems using artificial potential functions: Analytical design guidelines," *Journal of Intelligent and Robotic Systems*, vol. 45, no. 4, pp. 369–394, 2006.

[6] E. Bonabeau, G. Theraulaz, and M. Dorigo, *Swarm Intelligence*. Springer, 1999.

[7] M. Aziz and P. E. Caines, "A mean field game computational methodology for decentralized cellular network optimization," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 2, pp. 563–576, 2016.

[8] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and Z. Han, "Distributed interference and energy-aware power control for ultra-dense D2D networks: A mean field game," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1205–1217, 2016.

[9] C. Yang, J. Li, M. Sheng, A. Anpalagan, and J. Xiao, "Mean field game-theoretic framework for interference and energy-aware control in 5G ultra-dense networks," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 114–121, 2017.

[10] C.-A. Lehalle and C. Mouzouni, "A mean field game of portfolio trading and its consequences on perceived correlations," *arXiv preprint arXiv:1902.09606*, 2019.

[11] G. Fu, P. Graewe, U. Horst, and A. Popier, "A mean field game of optimal portfolio liquidation," *Mathematics of Operations Research*, 2021.

[12] A. Lachapelle, J.-M. Lasry, C.-A. Lehalle, and P.-L. Lions, "Efficiency of the price formation process in presence of high frequency participants: a mean field game analysis," *Mathematics and Financial Economics*, vol. 10, no. 3, pp. 223–262, 2016.

[13] M. Huang, R. Malhamé, and P. Caines, "Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle," *Communications in Information & Systems*, vol. 6, no. 3, pp. 221–252, 2006.

[14] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1560–1571, 2007.

[15] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[16] J.-M. Lasry and P.-L. Lions, "Jeux à champ moyen. ii–horizon fini et contrôle optimal," *Comptes Rendus Mathématique*, vol. 343, no. 10, pp. 679–684, 2006.

[17] Y. Achdou and I. Capuzzo-Dolcetta, "Mean field games: numerical methods," *SIAM Journal on Numerical Analysis*, vol. 48, no. 3, pp. 1136–1162, 2010.

[18] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," in *NeurIPS*, 2019.

[19] K. Cui and H. Koeppl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 1909–1917, PMLR, 2021.

[20] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International Conference on Machine Learning*, pp. 5571–5580, PMLR, 2018.

[21] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Q-learning in regularized mean-field games," *arXiv preprint arXiv:2003.12151*, 2020.

[22] S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow, *Stochastic Processes*, vol. 2. Wiley New York, 1996.

[23] G. Owen, *Game Theory*. Academic Press, 1982.

[24] J. Arabneydi and A. Mahajan, "Team optimal control of coupled subsystems with mean-field sharing," in *53rd IEEE Conference on Decision and Control*, pp. 1669–1674, IEEE, 2014.

[25] Z. Shi, J. Tu, Q. Zhang, L. Liu, and J. Wei, "A survey of swarm robotics system," in *International Conference in Swarm Intelligence*, pp. 564–572, Springer, 2012.

[26] N. Saldi, T. Basar, and M. Raginsky, "Markov–Nash equilibria in mean-field games with discounted cost," *SIAM Journal on Control and Optimization*, vol. 56, no. 6, pp. 4256–4287, 2018.

[27] R. Fox, A. Pakman, and N. Tishby, "Taming the noise in reinforcement learning via soft updates," in *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pp. 202–211, 2016.

[28] Y. Guan, Q. Zhang, and P. Tsiotras, "Learning Nash equilibria in zero-sum stochastic games via entropy-regularized policy approximation," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2462–2468, 2021.

[29] J. Peters, K. Mulling, and Y. Altun, "Relative entropy policy search," in *24th AAAI Conference on Artificial Intelligence*, 2010.

[30] G. Grimmett and D. Stirzaker, *Probability and random processes*. Oxford university press, 2001.