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 parameter, through which one can control the behavior of the population. We first consider 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 system, where the sub-optimality ϵ is of order $\mathcal{O}(1/\sqrt{N})$. 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.

I. INTRODUCTION

Decision making in decentralized systems arises in many applications, ranging from multi-robot task allocation [1]–[3], finance [4]–[6], 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 the scalability issues, the mean field approach was introduced in [7]–[9]. 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 [7].

Although the discretization of continuous MFG has been studied in prior works [10], 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 [11]. That is, when the population mean field changes slightly, the corresponding optimal policy for the representative agent could change drastically [12]. Previous works have used Boltzmann policies [13] and projection to meshed probability measure spaces [11] to avoid such issues. More recent works [14] directly introduced a relative entropy term to the reward structure to provide regularity conditions. The authors in [12] used entropy-regularization to stabilize the iterative algorithm and reduced regularization over time to learn the original MFE. The existence and uniqueness of the entropy-regularized MFE was also examined.

Different from these previous works, in this paper we explicitly consider 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 the 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 an 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 one can encode desired population behaviors into the reference policy. By adjusting the multiplier of the regularization term, one can further produce a tunable behavior of the population that balances the encoded behavior and the cumulative rewards.

Contributions: In this work, we formulate a game of N-homogeneous agents and show that under pairwise coupled rewards, the state of the system can be *exactly* represented by a distribution over the state space. We then consider the limiting infinite population game and introduce entropy regularization to construct contraction operators to find the unique regularized MFE. We consider a special class of MFGs where the agents have coupled rewards but decou-

[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

pled dynamics. For this class of MFGs, we streamline and simplify the convergence proof in [12]. 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 reference policy.

II. PROBLEM FORMULATION

Consider a large population game consisting of N homogeneous agents, where $N \gg 1$. We define the game through the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}^1, \ldots, \mathcal{R}^N, N, H \rangle$. The game is over a discrete-time with a finite horizon H. 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 will discuss in greater detail later on. As a consequence of its own action, 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 $(s_t^1, \ldots, s_t^{i-1}, s_t^{i+1}, \ldots, s_t^N)$. Each agent follows a (time-varying) Markov policy $\pi^i = \{\pi_t^i\}_{t=0}^H$, such that, at each time step t, this policy is a mapping $\pi_t^i : \mathcal{S} \times \mathcal{A} \to [0,1]$. We use Π to denote the space of admissible policies. For simplicity, we use $t \leq H$ to denote that $t \in \{0, \ldots, H\}$.

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.[1] In the sequel, we use the notation $\mathcal{T}(\cdot|\cdot, \pi_t^i)$ to denote the dynamics of agent i 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

$$s_{t+1}^i \sim \mathcal{T}\big(\cdot|s_t^i, \pi_t^i\big), \quad t = 0, \ldots, H-1, \qquad s_0^i \sim \mu_0. \quad (2)$$

b) Rewards: The reward of agent i at time t is

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

where $\Theta_t : \mathbb{R} \to \mathbb{R}$ and $L_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.

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 $t \leq H$, $|\Theta_t(x) - \Theta_t(y)| \leq K_\Theta |x - y|$.*

Assumption 2. *There exists a positive $L_{\max}$ such that, for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$ and $t \leq H$, $|L_t(s, a, s')| \leq L_{\max}$.*

The maximum magnitude of the reward is $\mathcal{R}_{\max} = \max_{|x| \leq L_{\max}, t \leq H} |\Theta_t(x)|$. Note that the reward 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

[1] We use $s^i \in \mathcal{S}$ to denote the state of the specific agent i, and $s \in \mathcal{S}$ to denote the state of a generic agent. Similar rules apply to a^i and a.

aggregation of the system state $(s_t^1, \ldots, s_t^N)$ to a 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 then defined as $\mu^N = \{\mu_t^N\}_{t=0}^H$. Note that $\sum_s \mu_t^N(s) = 1$, and thus 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, $\mathcal{M} = \big(\mathcal{P}(\mathcal{S})\big)^{H+1}$ is the space of the probability measure flows and $\mu^N \in \mathcal{M}$.

d) Metric spaces: We use total variation [15] as the metric for the probability measure space $\mathcal{P}(\mathcal{X})$. When $\mathcal{X}$ is finite, the total variation between $\nu, \nu' \in \mathcal{P}(\mathcal{X})$ is given by $\mathrm{d}_{\mathrm{TV}}(\nu, \nu') = \frac{1}{2}\sum_{x \in \mathcal{X}} |\nu(x) - \nu'(x)| = \frac{1}{2}\|\nu(x) - \nu'(x)\|_1$. We equip both $\mathcal{M}$ and Π with the supremum metric induced by the total variation. That is, for $\mu, \mu' \in \mathcal{M}$, we define $\mathrm{d}_\mathcal{M}(\mu, \mu') = \max_{t \leq H} \mathrm{d}_{\mathrm{TV}}(\mu_t, \mu_t')$, and for policies $\pi, \pi' \in \Pi$ $\mathrm{d}_\Pi(\pi, \pi') = \max_{t \leq H}\max_{s \in \mathcal{S}} \mathrm{d}_{\mathrm{TV}}(\pi_t(s), \pi_t'(s))$, 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 [15].

e) Distribution induced rewards: Due to symmetry, the state-coupled reward in (3) can be characterized through the empirical distribution. Overloading the notation, we also write

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

With the above definition, we further define the distribution-induced reward for each agent as

$$\mathcal{R}_t^i(s_t^i, a_t^i, \mu_t^N) \triangleq \Theta_t\Big(L_t(s_t^i, a_t^i, \mu_t^N)\Big). \quad (5)$$

Lemma 1. *The reward function $\mathcal{R}_t^i(s, a, \mu)$ in (5) is globally Lipschitz with respect to the probability measure $\mu \in \mathcal{P}(\mathcal{S})$, with Lipschitz constant $2K_\Theta L_{\max}$.*

Proof. Please see full version of the paper [16]. □

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

$$J^{i,N}(\pi^i, \pi^{-i}) = \mathbb{E}\Big[\sum_{t=0}^H \mathcal{R}_t^i(s_t^i, a_t^i, \mu_t^N)\Big], \quad (6)$$

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

$$\max_{\pi^i \in \Pi} J^{i,N}(\pi^i, \pi^{-i}), \qquad i = 1, \ldots, N. \tag{7}$$

One of the most common solution concepts for a game with such coupled optimization is the Nash equilibrium [17].

Definition 1. *A Nash equilibrium (NE) 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 $i = 1, \ldots, N$,*

$$J^{i,N}(\pi^i, \pi^{-i*}) \leq J^{i,N}(\pi^{i*}, \pi^{-i*}) + \epsilon, \quad \forall \pi^i \in \Pi. \tag{8}$$

In other words, any unilateral deviation from an ϵ-NE can improve an agent's performance by at most ϵ.

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

Assumption 3. *For all $i, j \in \{1, \ldots, N\}$, $\pi^i = \pi^j$.*

Assumption 3 leads, in general, to a loss in performance [18]. However, identical policy is a standard assumption in the literature of large scale systems for reasons of simplicity and robustness [19]. In light of Assumption 3, henceforth, we will drop the superscripts on the policies and denote the policy used by all agents as π.

III. MEAN FIELD APPROXIMATION

As 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. In general, the propagation of the distribution of μ_t^N is computationally expensive. On the other hand, under the identical policy π used by all (infinite in number) agents, the trajectory of the mean field is deterministic [14]. Furthermore, μ_t follows a simple propagation rule:

$$\mu_{t+1} = \mu_t \left[\mathcal{T}(\pi_t) \right], \tag{9}$$

where $[\mathcal{T}(\pi_t)]$ is a right stochastic matrix constructed based on (1). We refer to the time sequence $\mu = \{\mu_t\}_{t=0}^H \in \mathcal{M}$ 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 will show later, the empirical distribution converges to the mean field as the number of agents approaches infinity.

A. Representative Agent

Before tackling the large population game with N agents, we consider the limiting infinite population case by 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 dynamics and its policy. Assuming that the mean field flow μ is known and fixed, this yields a standard Markov Decision Process (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

$$\mathcal{R}_{\mu,t}(s, a) = \Theta_t \big(L_t(s, a, \mu_t) \big). \tag{10}$$

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

$$J_\mu(\pi^*) = \max_{\pi \in \Pi} \mathbb{E}_{\mu_0} \left[\sum_{t=0}^H \mathcal{R}_{\mu,t}(s_t, a_t) \right]. \tag{11}$$

Note that the optimal policy depends on μ. We use the operator $\mathcal{B}_{\text{opt}} : \mathcal{M} \to \Pi$ to denote the mapping from the mean field flow to an optimal policy of the induced MDP:[2]

$$\pi^* = \mathcal{B}_{\text{opt}}(\mu). \tag{12}$$

When all agents employ the policy π of the representative agent, a new mean field flow μ is induced and can be propagated via (9) starting from μ_0. We use the operator $\mathcal{B}_{\text{prop}} : \Pi \to \mathcal{M}$ to denote this propagation. That is,

$$\mu = \mathcal{B}_{\text{prop}}(\pi). \tag{13}$$

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

$$\pi^* = \mathcal{B}_{\text{opt}}(\mu^*), \quad \mu^* = \mathcal{B}_{\text{prop}}(\pi^*). \tag{14}$$

The existence of such consistent pair can be established through a Brouwer's fixed point argument [20].

One may attempt to use fixed-point iterations to find a solution to (14). Unfortunately, the composed mean-field equilibrium (MFE) operator $\Gamma = \mathcal{B}_{\text{prop}} \circ \mathcal{B}_{\text{opt}}$ is only non-expansive and not a contraction, in general. The non-contractiveness of the MFE operator Γ comes from the *hard* maximization within the MDP optimization $\mathcal{B}_{\text{opt}}$, where slightly different induced rewards may lead to significantly different optimal policies. Consequently, even if two initial mean-field flows are close, they may induce totally different optimal policies, which then leads to two new mean-field flows that are far apart. For a more detailed discussion, one may refer to [12].

---

[2]In general, $\mathcal{B}_{\text{opt}}$ is a set-valued function, since the optimal policy of an MDP need not be unique.

IV. Entropy-Regularized Mean Field Games

Entropy regularization is a technique used extensively to stabilize learning algorithms and to reduce the maximization bias [21]. The extra entropy cost introduced to the reward structure prevents abrupt policy changes between iterations.

To address the issue of non-contractiveness of the operator Γ, we introduce an entropy-regularization term and replace the hard maximization in $\mathcal{B}_{\text{opt}}$ with a soft maximization. With the soft maximization, a small change in the mean field flow will not result in 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 $t \leq H, s \in \mathcal{S}$, and $a \in \mathcal{A}$, we introduce an entropy regularization term to (11) as follows:

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

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.[3] When β is small, the regularization term 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 collected rewards. As a consequence, the optimal π approaches a greedy policy produced by $\mathcal{B}_{\text{opt}}$ as in (12).

A. Optimization for the Regularized MDP

It can be shown that the *unique* optimal policy that solves $\max_\pi J_\mu^{\text{KL}}(\pi; \rho)$ is given by the following (weighted) Boltzmann distribution [21]

$$\pi_{\mu,t}^{\text{KL}}(a|s) = \frac{1}{Z_t(s)}\rho_t(a|s)\exp\left[\beta Q_{\mu,t}^{\text{KL}}(s, a; \rho)\right], \quad (15)$$

where $Z_t(s)$ is a normalization factor, and the entropy-regularized state-action value function $Q_{\mu,t}^{\text{KL}}$ is computed as:

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

with the boundary condition $Q_{\mu,H}^{\text{KL}}(s, a; \rho) = \mathcal{R}_{\mu,H}(s, a)$.

As the entropy-regularized optimal policy is induced by the given mean field, we denote the operation performed in (15) using the operator $\mathcal{B}_{\text{opt},\beta}^{\text{KL}} : \mathcal{M} \rightarrow \Pi$. We can then define the entropy-regularized MFE (ER-MFE) operator $\Gamma_\beta^{\text{KL}} : \mathcal{M} \rightarrow \mathcal{M}$ as

$$\Gamma_\beta^{\text{KL}} = \mathcal{B}_{\text{prop}} \circ \mathcal{B}_{\text{opt},\beta}^{\text{KL}}, \quad (16)$$

The regularized equilibrium is then defined as follows.

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

In the sequel, we establish the existence and uniqueness of the ER-MFE. The main goal is to show that the ER-MFE operator is a contraction if β is selected properly. The

following derivation is more direct and easier to demonstrate than the one reported in [12]. The reason is that we are restricting ourselves to the case where the agents have decoupled dynamics. As a consequence, a single agent deviation from the optimal policy does not directly impact the distribution of the rest of the population.

B. Convergence Analysis

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

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

$$\text{d}_{\mathcal{M}}(\mathcal{B}_{\text{prop}}(\pi), \mathcal{B}_{\text{prop}}(\pi')) \leq K_{\text{prop}}\,\text{d}_{\Pi}(\pi, \pi'), \quad (17)$$

where

$$K_{\text{prop}} = \frac{|S|(|S|^H - 1)}{|S| - 1}. \quad (18)$$

Proof. See the the full version of the paper [16]. $\square$

The following two Lemmas are adopted from [12].

Lemma 3 ([12]). *Under Assumptions 1 and 2, the entropy-regularized Q-function Q_μ^{KL} is Lipschitz with respect to μ for arbitrary $\beta > \beta_{\max}$ and $\beta_{\max} > 0$. That is,*

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

where $K_Q^{\text{KL}} = \max_{t \leq H} K_{Q,t}^{\text{KL}}$, and $K_{Q,t}^{\text{KL}}$ is defined as

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

with boundary condition $K_{Q,H}^{\text{KL}} = 2K_\Theta L_{\max}$, 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 ([12]). *Under Assumptions 1 and 2, the entropy-regularized operator $\mathcal{B}_{\text{opt}}^{\text{KL}}$ is Lipschitz, that is,*

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

where,

$$K_{\text{opt},\beta}^{\text{KL}} = \frac{|\mathcal{A}|(|\mathcal{A}| - 1)\beta\rho_{\max}^2}{2\rho_{\min}^2}K_Q^{\text{KL}}. \quad (19)$$

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 $\mathcal{B}_{\text{prop}}$ and $\mathcal{B}_{\text{opt},\beta}^{\text{KL}}$, we arrive at the following result regarding the selection of β to ensure that Γ_β^{KL} is a contraction.

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

$$\beta < \min\left\{\beta_{\max}, \frac{2\rho_{\min}^2}{\rho_{\max}^2|\mathcal{A}|(|\mathcal{A}| - 1)}\frac{1}{K_Q^{\text{KL}}K_{\text{prop}}}\right\}. \quad (20)$$

Proof. The β selection in (20) guarantees that $K_{\text{prop}}K_{\text{opt},\beta}^{\text{KL}} < 1$. Consequently, the ER-MFE operator Γ_β^{KL} is a contraction. $\square$

C. Error Bounds on the Mean Field Approximations

With the theoretical results established for the infinite-population equilibrium, we now examine the performance guarantee for applying the infinite-population equilibrium to a finite N-agent system. We first present the following lemma, which characterizes the asymptotic convergence of the empirical distribution flow μ^N to the mean field flow μ.

Lemma 5. *Suppose a mean field flow μ (infinite population) is induced by the 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 $t \leq H$,*

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(\mu_t^N, \mu_t\right)\right] = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right). \tag{21}$$

Proof. See the appendix. □

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

Theorem 2. *Consider an ER-MFE (π^*, μ^*).[4] Then, for all $\widetilde{\pi} \in \Pi$, we have*

$$J_{\mathrm{KL}}^{i,N}(\widetilde{\pi}, \pi^*) \leq J_{\mathrm{KL}}^{i,N}(\pi^*, \pi^*) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right), \tag{22}$$

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

Note that the N-agent trajectory under the optimal policy π^* is given by $s_{t+1}^i \sim \mathcal{T}\left(\cdot \mid s_t^i, \pi_t^*(s_t^i)\right)$ for $i = 1, \ldots, N$. Without loss of generality, we let agent 1 deviate and select some other policy $\widetilde{\pi}$. For both the optimal system and the deviated system, the agents' initial distributions are μ_0. Then, the trajectory of the deviated N-agent system is given by

$$\begin{aligned}
\widetilde{s}_{t+1}^1 &\sim \mathcal{T}\left(\cdot \mid \widetilde{s}_t^1, \widetilde{\pi}_t(\widetilde{s}_t^1)\right), \\
\widetilde{s}_{t+1}^i &\sim \mathcal{T}\left(\cdot \mid \widetilde{s}_t^i, \pi_t^*(\widetilde{s}_t^i)\right), \text{ for } i = 2, \ldots, N.
\end{aligned} \tag{23}$$

The following lemma characterizes the convergence of the deviated empirical distribution to the optimal mean field.

Lemma 6. *Let $\widetilde{\mu}^N$ denote the empirical distribution flow resulted from the system in (23). Then, for all $t \leq H$,*

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

Proof. See the appendix. □

We now present a proof for Theorem 2 using cross disturbance analysis similar to [7].

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

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

$$\leq \mathbb{E}\sum_{t=0}^{H}\left[\mathcal{R}_t\left(\widetilde{s}_t^1, \widetilde{\pi}_t, \mu_t^*\right) - \frac{1}{\beta}\log\frac{\widetilde{\pi}_t(a_t|\widetilde{s}_t^1)}{\rho_t(a_t|\widetilde{s}_t^1)}\right] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{24}$$

---

[4]We drop the superscript KL for notation simplicity. Unless specified otherwise, π^* and μ^* refer to the entropy-regularized optimal solutions.

$$\leq \mathbb{E}\sum_{t=0}^{H}\left[\mathcal{R}_t\left(s_t^1, \pi_t^*, \mu_t^*\right) - \frac{1}{\beta}\log\frac{\pi_t^*(a_t|s_t^1)}{\rho_t(a_t|s_t^1)}\right] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{25}$$

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

In (24), 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 (25), we used the optimality of π^* in the regularized MDP induced by μ^*.

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. To see this, note that

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

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

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

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

$$\left|J_{\mathrm{KL}}^{1,N}(\pi^*, \pi^*) - J_{\mathrm{KL}}(\pi^*, \pi^*)\right| = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right). \tag{27}$$

Combining (26) and (27), it follows that

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

Since the population is 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 dynamic resource allocation problem [3] over a directed 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, respectively. A large group of agents traverse through the graph and collect the rewards assigned at the terminal time step H, and no running reward is assigned. At time H, if an agent is at state (node) 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 high population density. In summary, we have the following state-coupled rewards

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

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

We set the nonlinear function in (3) to $\Theta_H(x) = x^2$. Each agent at a state s (graph node) can choose one of the adjacent states 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 $(s, s') \in \mathcal{E}$.
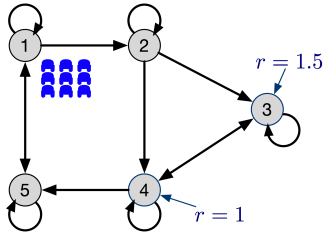

Fig. 1.   Graph for a resource allocation problem with self-loops.

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 3 for the extra reward. At the next iteration, the penalty for staying at node 3 is high given the mean field flow from the previous iteration. The representative agent then constructs a policy to visit node 4. In summary, the policy found by the $\mathcal{B}_{\text{opt}}$ without regularization oscillates between concentrating on node 3 and concentrating on node 4 at H.

Suppose now that a coordinator can send a command to the group of agents, and it decides that both nodes 3 and 4 need to be occupied by some agents, but s/he does not have access to the actual rewards at the 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.[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 does not reflect the difference in the rewards.

We now use the constructed reference policy to form the entropy-regularized MFG and solve the regularized game using the operator Γ_β^{KL} in (16) for two different β values. The algorithm converges and the population distribution over the nodes is depicted in Fig. 2. Recall that a larger β means less 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 up to 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.
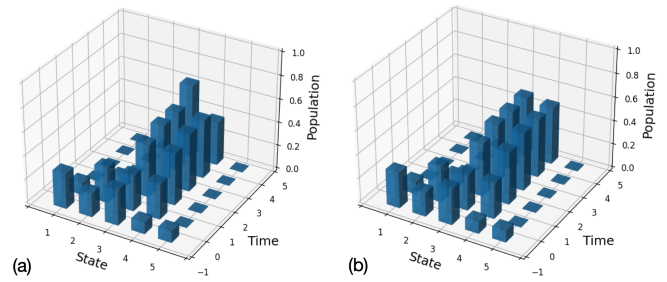
Fig. 2.   Population distribution over time. Scenario in (a) has the inverse temperature of $\beta = 3$, and scenario in (b) has $\beta = 0.1$.

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 let the first agent optimize the entropy-regularized MDP induced by μ^N. We then compared the difference between its newly-optimized performance and the performance should the agent adopt 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.
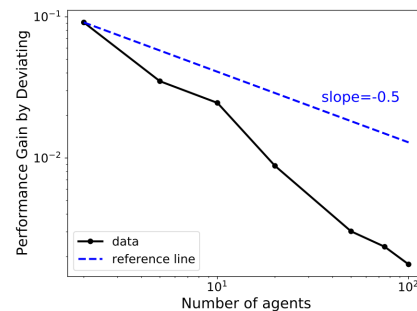

Fig. 3.   Log-log plot of performance gain by an agent unilaterally deviating in a finite population.

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 the entropy-regularization provides the regularity conditions that the standard MFG lacks. The conditions 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 verified the theoretical results and demonstrated that the reference policy for the entropy regularization 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.

Proof of Lemma 5. The empirical distribution is defined as

$$\mu_t^N(s) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_s(s_t^i). \qquad (28)$$

Let $X_s^i = \mathbb{1}_s(s_t^i)$. Since the dynamics are decoupled and all agents use the same policy, with the same initial distribution, the random variables X_s^i are i.i.d. 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 [15], we have $\mu_t^N(s) \xrightarrow{a.s.} \mathbb{E}\left[X_s^i\right]$ as $N \to \infty$. Thus, 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,

$$\mathbb{E}\big\|\mu_t^N - \mu_t\big\|_2^2 = \mathbb{E}\sum_{s \in \mathcal{S}} |\mu_t^N(s) - \mu_t(s)|^2$$

$$= \mathbb{E}\sum_{s \in \mathcal{S}} \Big| \frac{1}{N}\sum_{i=1}^N \big(X_s^i - \mathbb{E}\left[X_s^i\right]\big)\Big|^2$$

$$\leq \frac{1}{N}\sum_{s \in \mathcal{S}} \mathrm{Var}(X_s^i) = \frac{1}{N}\sum_{s \in \mathcal{S}} \mu_t(s)(1 - \mu_t(s))$$

$$= \frac{1}{N}(1 - \|\mu_t\|_2^2) \leq \frac{1}{N}.$$

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

$$\mathbb{E}[\mathrm{d}_{\mathrm{TV}}(\mu_t^N - \mu_t)] = \mathbb{E}\Big[\frac{1}{2}\sum_s |\mu_t^N(s) - \mu_t^*(s)|\Big]$$

$$= \frac{1}{2}\mathbb{E}\|\mu - \mu'\|_1 = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

$\square$

Proof of Lemma 6. 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

$$\mathbb{E}\left|\tilde{\mu}_t^N(s) - \mu_t^*(s)\right|$$

$$\leq \mathbb{E}\left|\frac{1}{N}\mathbb{1}_s(\tilde{s}_t^1)\right| + \mathbb{E}\left|\frac{1}{N}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k) - \mu_t^*(s)\right|$$

$$\leq \frac{1}{N} + \mathbb{E}\left|\frac{1}{N-1}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k) - \mu_t^*(s)\right| \qquad (29)$$

$$\qquad\qquad + \mathbb{E}\left|\frac{1}{N(N-1)}\sum_{k=2}^N \mathbb{1}_s(\tilde{s}_t^k)\right|$$

$$\leq \frac{1}{N} + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \frac{1}{N} = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \qquad (30)$$

The second term in (29) corresponds to the scenario of $N-1$ agents all applying the optimal policy π^*. By Lemma 5, we obtain the convergence rate. Finally, from (30), we have

$$\mathbb{E}\left[\mathrm{d}_{\mathrm{TV}}\left(\tilde{\mu}_t^N, \mu_t^*\right)\right] = \frac{1}{2}\mathbb{E}\sum_{s \in \mathcal{S}} |\tilde{\mu}_t^N(s) - \mu_t^*(s)| = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

$\square$

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] 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.

[3] D. Shishika, Y. Guan, M. Dorothy, and V. Kumar, "Dynamic defender-attacker blotto game," *arXiv preprint arXiv:2112.09890*, 2021.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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.

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

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

[16] Y. Guan, M. Zhou, A. Pakniyat, and P. Tsiotras, "Shaping large population agent behaviors through entropy-regularized mean-field games," *arXiv preprint arXiv:2110.07469*, 2021.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.