Approximate Equilibrium Computation for Discrete-Time Linear-Quadratic Mean-Field Games

Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar

***Abstract*— While the topic of mean-field games (MFGs) has a relatively long history, heretofore there has been limited work concerning algorithms for the computation of equilibrium control policies. In this paper, we develop a computable policy iteration algorithm for approximating the mean-field equilibrium in linear-quadratic MFGs with discounted cost. Given the mean-field, each agent faces a linear-quadratic tracking problem, the solution of which involves a dynamical system evolving in retrograde time. This makes the development of forward-in-time algorithm updates challenging. By identifying a structural property of the mean-field update operator, namely that it preserves sequences of a particular form, we develop a forward-in-time equilibrium computation algorithm. Bounds that quantify the accuracy of the computed mean-field equilibrium as a function of the algorithm's stopping condition are provided. The optimality of the computed equilibrium is validated numerically. In contrast to the most recent/concurrent results, our algorithm appears to be the first to study *infinite*-horizon MFGs with *non-stationary* mean-field equilibria, though with focus on the linear quadratic setting.**

I. Introduction

Recent years have witnessed the tremendous progress of operation, control, and learning in multi-agent systems [1]–[5], where multiple agents strategically interact with each other in a common environment, to optimize either a common or individual long-term return. Despite the substantial interest, most existing *algorithms* for multi-agent systems suffer from *scalability issues*, due to their complexity increasing exponentially with the number of agents involved. This issue has precluded the application of many algorithms to systems with even a moderate number of agents, let alone to real-world applications [6], [7].

One way to address the scalability issue is to view the problem in the context of *mean-field games* (MFGs), proposed in the seminal works of [8], [9] and, independently, [10]. Under the mean-field setting, the interactions among the agents are approximately represented by the distribution of all agents' states, termed the mean-field, where the influence of each agent on the system is assumed to be infinitesimal in the large population setting. In fact, the more agents are involved, the more accurate the mean-field approximation is, offering an effective tool for addressing the scalability issue. Moreover, following the so-termed *Nash certainty equivalence (NCE) principle* [8], the solution to an MFG, referred to as a *mean-field equilibrium (MFE)*, can be determined by each agent computing a best-response control policy to some mean-field that is consistent with the aggregate behavior of all agents. This principle decouples the process of finding the solution of the game into a computational procedure of determining the best-response to a fixed mean-field at the agent level, and an update of the mean-field for all agents. In particular, a straightforward routine for computing the MFE proceeds as follows: first, each agent calculates the optimal control, best-responding to some given mean-field, and then, after executing the control, the states are aggregated to update the mean-field. This routine is referred to as the *NCE-based approach*, which serves as the foundation for our algorithm.

Serving as a standard, but significant, benchmark for general MFGs, linear-quadratic MFGs (LQ-MFGs) [11]–[13] have been advocated in the literature. In particular, the cost function describing deviations in the state, from the mean-field, as well as the cost for a given control effort is assumed to be quadratic while the transition dynamics are assumed to be linear. Intuitively, the cost incentivizes each agent to *track* the collective behavior of the population, which, for any fixed mean-field, leads to a *linear-quadratic tracking* (LQT) subproblem for each agent. Though simple in form, equilibrium computation in LQ-MFGs (most naturally posed in continuous state-action spaces) inherits most of the challenges from equilibrium computation in general MFGs. While much work has been done in the continuous-time setting [11]–[13], the discrete-time counterpart has received considerably less attention. It appears that, only the work of [14] (which considered a model with *unreliable communication* with an average cost criterion) has studied a discrete-time version of the model proposed in [11]. The formulation of the discrete-time model of our paper, and the associated equilibrium analysis, are in a setting distinct from [14], and constitute one of the contributions of the present work.

There has been an increasing interest in developing (model-free) equilibrium-computation algorithms for certain MFGs [15]–[18]; see [19, Sec. 4] for more a detailed summary. The closest setting to ours is in the concurrent while independent work on learning for discrete-time LQ-MFGs [18]. However, given any fixed mean-field, [18] treats each agent's subproblem as a *linear quadratic regulator (LQR) with drift*, which is different from the continuous-time formulation [11]–[13]. This is made possible because they considered mean-field trajectories that are constant in time (also referred to as *stationary* mean-fields). This is in contrast to the LQT subproblems found in both the literature [11]–[13] and in our formulation. While the former admits a *forward-in-time* optimal control that can be obtained using policy iteration and standard reinforcement learning

The authors are affiliated with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign Urbana, IL 61801.

Research supported in part by AFOSR (FA9550-19-1-0353), in part by ARL (W911NF-17-2-0196),and in part by ARO (W911NF-16-1-0485).

(RL) algorithms [20]–[22], the latter leads to a *backward-in-time* optimal control problem, which, in general, has been recognized to be challenging to solve, especially in a model-free fashion [23], [24]. Most other RL algorithms for general MFGs are also restricted to the stationary mean-field setting [15], [16], which does not apply to the LQ-MFG problem here. Fortunately, by identifying a structural property of our policy iteration algorithm and employing an NCE-based equilibrium-computation approach, one can develop a computable algorithm that executes forward in time.

Contribution. Our contribution in this paper is three-fold: (**1**) We formally introduce the formulation of discrete-time LQ-MFGs with discounted cost, complementing the standard continuous-time formulation [9], [11], and the discrete-time average-cost setting of [14], together with existence and uniqueness guarantees for the MFE. (**2**) By identifying structural results of the NCE-based policy iteration update, we develop an equilibrium-computation algorithm, with convergence error analysis, that can be implemented *forward-in-time*. (**3**) We illustrate the quality of the computed MFE in terms of the algorithm's stopping condition and the number of agents. Our structural results and equilibrium-computation algorithm lay foundations for developing model-free RL algorithms, as our immediate future work.

Outline. The remainder of the paper proceeds as follows. In Section II, we introduce the linear-quadratic mean-field game model. Section III provides a background of relevant results from the literature on mean-field games as well as establishes a characterization of the mean-field equilibrium for our setting. Section IV outlines some properties of the computational process and presents the algorithm. Numerical results are presented in Section V. Concluding remarks and some future directions are presented in Section VI. Proofs of all results have been relegated to the Appendix.

II. Linear Quadratic Mean-Field Game Model

Consider a dynamic game with $N < \infty$ agents playing on an infinite time horizon. For each agent $n \in [N]$, let $z_t^n \in \mathbb{R}$ represent the current state and $u_t^n \in \mathbb{R}$ represent the current control. Each agent n's state is assumed to follow linear time-invariant (LTI) dynamics,

$$z_{t+1}^n = az_t^n + bu_t^n + w_t^n, \tag{1}$$

with constants $a \in \mathbb{R}$, $b \in \mathbb{R} \setminus \{0\}$, independent and identically distributed initial state z_0^n with mean ν_0 and variance σ_0^2, and independent identically distributed noise terms, $w_t^n \sim \mathcal{N}(0, \sigma_w^2)$, assumed to be independent of $z_0^{n'}$, $w_s^{n'}$ for all s and t, and for all $n' \neq n$.

At the beginning of each time step, each agent observes every other agent's state. Thus, assuming perfect recall, the information of agent n at time t is $i_t^n = \big((z_0^1, \ldots, z_0^N), u_0^n, \ldots, (z_{t-1}^1, \ldots, z_{t-1}^N), u_{t-1}^n, (z_t^1, \ldots, z_t^N)\big)$. A control policy for agent n at time t, denoted by η_t^n, maps its current information i_t^n to a control action $u_t^n \in \mathbb{R}$. The joint control policy is the collection of policies across agents, and is denoted by $\eta_t = (\eta_t^1, \ldots, \eta_t^N)$. The joint control law is the collection of joint control policies across time, denoted by $\eta = (\eta_0, \eta_1, \ldots)$.

The agents are coupled via their expected cost functions. The expected cost for agent n under joint policy η and the initial state distribution, denoted by $J^n(\eta)$, is defined as,

$$J^n(\eta) := \sum_{t=0}^{T} \gamma^t \mathbb{E}_\eta \left[c_z \big(z_t^n - \frac{1}{N-1} \sum_{n' \neq n} z_t^{n'} \big)^2 + c_u (u_t^n)^2 \right], \tag{2}$$

where $\gamma \in [0, 1)$ is the discount factor and $c_z, c_u > 0$ are cost weights for the state and control, respectively. The expectation is taken with respect to the randomness of all agents' state trajectories induced by the joint control law η and the initial state distribution.

In the finite-agent system described above, each agent is assumed to fully observe all other agents' states. As N grows, determining a policy that is a best-response to all other agents' policies becomes computationally intractable, precluding computation of a Nash equilibrium [25]. Fortunately, since the coupling between agents manifests itself as an average of all agent's states, one can approximate the finite agent game by an infinite population game in which a *generic agent* interacts with the mass behavior of all agents. The empirical average of all agents' states becomes the mean state process (*i.e.*, the *mean-field*), decoupling the agents and yielding a stochastic control problem. The infinite population game is termed a *mean-field game* [8]. In this paper, we focus on *linear-quadratic* MFGs in which the generic agents' dynamics are linear and its costs are quadratic.

The state process of the generic agent is identical to (1), that is,

$$z_{t+1} = az_t + bu_t + \omega_t, \tag{3}$$

where z_0 is distributed with mean ν_0 and variance σ_0^2, and ω_t is an i.i.d. noise process generated according to the distribution $\mathcal{N}(0, \sigma_w^2)$, assumed to be independent of the mean-field and the agent's state.

The generic agent's control policy at time t, denoted by μ_t, translates the available information at time t, denoted by $i_t = (z_0, u_0, \ldots, z_{t-1}, u_{t-1}, z_t)$, to a control action $u_t \in \mathbb{R}$. The collection of control policies across time is referred to as a control law and is denoted by $\mu = (\mu_0, \mu_1, \ldots) \in \mathcal{M}$ where $\mathcal{M}$ is the space of admissible control laws. The generic agent's expected cost under control law μ is defined as,

$$J(\mu, \bar{z}) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\mu \left[c_z (z_t - \bar{z}_t)^2 + c_u u_t^2 \right], \tag{4}$$

where $\bar{z}_t = \mathbb{E}[z_t]$ represents the mean-field at time t. The mean-field trajectory $\bar{z} := (\bar{z}_0, \bar{z}_1, \ldots)$ is assumed to belong to the space of bounded sequences, that is, $\bar{z} \in \mathcal{Z}$ where $\mathcal{Z} := \ell^\infty = \{x = (x_0, x_1, \ldots) \mid \sup_{t \geq 0} |x_t| < \infty\}$.

To define a mean-field equilibrium, first define the operator $\Lambda : \mathcal{M} \to \mathcal{Z}$ as a mapping from the space of admissible control laws $\mathcal{M}$ to the space of mean-field trajectories $\mathcal{Z}$.

Due to the information structure of the problem, the policy at any time only depends upon the current state [14]. It is defined as follows: given $\mu \in \mathcal{M}$, the mean-field $\bar{z} := \Lambda(\mu)$ is constructed recursively as

$$\bar{z}_{t+1} := A\bar{z}_t + B\mu_t(\bar{z}_t), \quad \bar{z}_0 = \nu_0. \tag{5}$$

Similarly, define an operator $\Phi : \mathcal{Z} \to \mathcal{M}$ as a mapping from a mean-field trajectory to its optimal control law,

$$\Phi(\bar{z}) := \operatorname{argmin}_{\mu} J(\mu, \bar{z}). \tag{6}$$

A mean-field equilibrium can now be defined.

Definition 1 ([26])**.** *The tuple $(\mu^*, \bar{z}^*) \in \mathcal{M} \times \mathcal{Z}$ is an MFE if $\mu^* = \Phi(\bar{z}^*)$ and $\bar{z}^* = \Lambda(\mu^*)$.*

The power of mean-field analysis is the fact that the equilibrium policies obtained in the infinite-population game are good approximations to the equilibrium policies in the finite-population game [8]–[10]. The focus of the current paper is on approximate equilibrium computation and, while we do not derive explicit bounds for finite N, we offer empirical results in Section V illustrating the effectiveness of the mean-field approximation.

III. Background: MFE Characterization

This section establishes some properties of mean-field equilibria. The results are complementary to those of [27], [8], and [14]. Note that while [14] constructs a discrete-time analogue of [8], the model of [14] considers an average-cost criterion, whereas here we consider a discounted-cost criterion, as in [26].

Recall that in the limiting case, as $N \to \infty$, the problem becomes a constrained stochastic optimal control problem. In particular, as described by (4), a generic agent aims to find a control law μ that tracks a given reference signal (the mean-field trajectory). This control law, hereafter referred to as the *cost-minimizing control*, is characterized in closed-form by the following lemma.

Lemma 1. *Given a mean-field trajectory, $\bar{z} = (\bar{z}_0, \bar{z}_1, \ldots) \in \mathcal{Z}$, the control law that minimizes* (4)*, termed the* cost-minimizing control*, denoted by $\Phi(\bar{z}) = (\mu_0(z_0; \bar{z}), \mu_1(z_1; \bar{z}), \ldots)$, is given for each t by,*[1]

$$u_t = \mu_t(z_t; \bar{z}) := g_p(apz_t + \lambda_{t+1}(\bar{z})), \tag{7}$$

where $g_p := -\gamma b/(c_u + \gamma b^2 p)$, p is the unique positive solution to the discrete-time algebraic Riccati equation (DARE),

$$p^2 + \big([(1-\gamma a^2)c_u/(\gamma b^2)] - c_z\big)p - c_z c_u/(\gamma b^2) = 0, \tag{8}$$

that is

$$p = (-\alpha + \sqrt{\alpha^2 + 4\beta})/2, \tag{9}$$

[1]The cost-minimizing control policy μ_t (from the cost-minimizing control μ) is denoted by $\mu_t(\cdot; \bar{z})$ to illustrate that it is parameterized by the mean-field trajectory $\bar{z}$.

where $\alpha := \frac{c_u(1-\gamma a^2)}{\gamma b^2} - c_z$, $\beta := \frac{c_z c_u}{\gamma b^2}$, and the sequence $\{\lambda_t\}$, referred to as the co-state*, is generated backward-in-time by,*

$$\lambda_t(\bar{z}) = \gamma h_p \lambda_{t+1}(\bar{z}) - c_z \bar{z}_t, \tag{10}$$

where $h_p := a(1 + bpg_p)$.

To ensure the well-posedness of the cost-minimizing controller for mean-field $\bar{z} \in \mathcal{Z}$, the optimal cost must be bounded [14]. This is true given the following assumption.

Assumption 1. *Given γ, a, b, c_z, c_u and g_p, h_p, where p is the positive solution of* (8)*, as given by* (9)*, the quantity $T_p := |h_p| + |c_z b g_p/(1 - \gamma h_p)|$ satisfies $T_p < 1$.*

This assumption is analogous to condition (H6.1) of [27] for continuous-time settings. Lemma 2 shows that under Assumption 1, both the co-state process and the optimal cost are bounded.

Lemma 2. *1) If $\lambda_0(\bar{z}) = -c_z \sum_{s=0}^{\infty} (\gamma h_p)^s \bar{z}_s$ then $\lambda = (\lambda_0, \lambda_1, \ldots) \in \ell^\infty$. Moreover, with this initial condition,*

$$\lambda_t(\bar{z}) = -c_z \sum_{s=0}^{\infty} (\gamma h_p)^s \bar{z}_{t+s}, \text{ for } t = 0, 1, \ldots. \tag{11}$$

2) Under Assumption 1, $J(\Phi(\bar{z}), \bar{z})$ for any $\bar{z} \in \mathcal{Z}$ is bounded.

Substituting the cost-minimizing control, (7), into the state equation, (3), the closed-loop dynamics are

$$\begin{aligned} z_{t+1} &= az_t + bg_p(apz_t + \lambda_{t+1}(\bar{z})) + \omega_t \\ &= h_p z_t + bg_p \lambda_{t+1}(\bar{z}) + \omega_t. \end{aligned}$$

Taking expectation, the above equation becomes $\bar{z}'_{t+1} = h_p \bar{z}_t + bg_p \lambda_{t+1}(\bar{z})$ for $t = 0, 1, \ldots$, where $\bar{z}'_0 = \nu_0$. Substitution of the co-state process, (11), yields the following as the mean-field dynamics,

$$\bar{z}'_{t+1} = h_p \bar{z}_t - c_z b g_p \sum_{s=0}^{\infty} (\gamma h_p)^s \bar{z}_{t+1+s}. \tag{12}$$

In the same vein as [8], the above can be compactly summarized as an update rule, termed the *mean-field update operator*, on the space of (bounded) mean-field trajectories. The update rule, denoted by $\mathcal{T} : \mathcal{Z} \to \mathcal{Z}$, is given by,

$$\bar{z}' = \mathcal{T}(\bar{z}) := \Lambda(\Phi(\bar{z})). \tag{13}$$

The operator outputs an updated mean-field trajectory $\bar{z}'$, using (5), resulting from the cost-minimizing control for a mean-field trajectory $\bar{z}$, given by (7). The operator is a contraction mapping, as shown below.

Lemma 3. *Under Assumption 1, the mean-field update operator $\mathcal{T}$ is a contraction mapping on $\mathcal{Z} = \ell^\infty$.*

Furthermore, iterated application of $\mathcal{T}$ results in a fixed point which corresponds to an MFE, as expressed below.

Theorem 1. *A mean-field trajectory $\bar{z}^*$ is a fixed point of $\mathcal{T}$,*

$$\bar{z}^* = \mathcal{T}(\bar{z}^*), \tag{14}$$

if and only if $(\Phi(\bar{z}^*), \bar{z}^*)$ *is an MFE.*

As a corollary to the above results, there exists a unique MFE, by the Banach fixed-point theorem [28]. Moreover, a straightforward approach for computing the equilibrium, *i.e.*, the fixed-point of $\mathcal{T}$, is to iterate the operator $\mathcal{T}$ until convergence. Indeed, we note that this process is referred to as *policy iteration* in the continuous-time LQ-MFGs setting of [11]. However, the cost-minimizing control given by Lemma 1 needs to be calculated backward-in-time, which makes the update of $\mathcal{T}$ in (13) not computable. In fact, to develop model-free learning algorithms, forward-in-time computation is necessary.

In what follows, we investigate properties of the mean-field operator that permit the construction of a *computable policy iteration* algorithm that proceeds forward-in-time.

IV. Approximate Computation of the MFE

A. Properties of the Mean-Field Update Operator

A prerequisite for the development of any algorithm is that the representations of all quantities in the algorithm are finite. Satisfying this requirement in our case is complicated by the fact that both the equilibrium mean-field trajectory and the cost-minimizing control are infinite dimensional (see Def. 1). To address the challenge, we represent the infinite sequences by finite sets of parameters.

The parameterization of the mean-field trajectory is inspired by a property of the update operator. To show this property, consider the following class of sequences.

Definition 2. *A sequence* $x = (x_0, x_1, \ldots)$ *is said to be a* τ-latent LTI sequence *if* $x_{t+1} = r x_t$ *for some* $r \in \mathbb{R}$ *for all* $t = \tau, \tau+1, \ldots$.

Any τ-latent LTI sequence, for $\tau < \infty$, can be represented by $\tau+2$ parameters, summarized by the pair $(x_{0:\tau}, r)$, where $x_{0:\tau} = (x_0, \ldots, x_\tau)$. This is illustrated in the following example.

Example 1. *Consider the following sequence* $(x_0, x_1, \ldots)$ *where* ϕ_0, ϕ_1 *are arbitrary functions and* $s_0, r \in \mathbb{R}$*,*

$$\begin{aligned}(x_0, x_1, x_2, x_3, x_4, \ldots) &= (s_0, \phi_0(x_0), \phi_1(x_1), r x_2, r x_3, \ldots)\\ &=: (x_{0:2}, r).\end{aligned}$$

The sequence obeys linear dynamics starting at $t = 2$*. As such, the above sequence is referred to as a* 2*-latent LTI sequence and is denoted by* $(x_{0:2}, r)$*.*

Our algorithm is based on the observation that, given any stable[2] τ-latent LTI sequence with constant r, the mean-field update operator outputs a stable $(\tau+1)$-latent LTI sequence with the same constant r, as summarized by Lemma 4 below.

Lemma 4. *If* $(x_{0:\tau}, r)$ *is a* τ*-latent LTI sequence with constant* r *satisfying* $|r| \leq 1$*, then* $(x'_{0:\tau+1}, r)$*, where* $x' = \mathcal{T}(x)$*, is a* $(\tau+1)$*-latent LTI sequence with constant* r*.*

By Lemma 4, each application of operator $\mathcal{T}$ increases the dimension of the mean-field trajectory's parameterization. This allows us to construct an iterative algorithm in which, for any finite iteration, all quantities are computable.

[2]Namely, $|r| \leq 1$.

B. A Computable Policy Iteration Algorithm

This section presents a policy iteration algorithm for approximately computing the mean-field equilibrium. The algorithm operates over iterations $k = 1, 2, \ldots$, where variables at the k^{th} iteration are denoted by superscript (k).

As mentioned in the discussion following Theorem 1, iterating the mean-field update operator $\mathcal{T}$ yields a process that converges to the MFE, though not computable due to the backward-in-time calculation of the cost-minimizing control. To address this issue, we propose an iterative algorithm that operates on parameterized sequences. Motivated by the result of Lemma 4, by initializing the algorithm with a τ-latent sequence, we can ensure that, after any finite number of iterations, the computed sequence is also τ-latent. Importantly, this structure allows one to describe the mean-field trajectory at any iteration by a finite set of parameters. Furthermore, the τ-latent LTI structure allows for the cost-minimizing control to be calculated forward-in-time. As a consequence, the aforementioned procedure can be carried out in a computable way, provided that the iteration number k remains finite.

More formally, our (computable) policy iteration algorithm proceeds as follows. Without loss of generality, we start with a 0-latent LTI mean-field trajectory $\bar{z}^{(0)}$ with $\bar{z}_0^{(0)} = \nu_0$ at iteration 0. Thus, at any iteration k, by Lemma 4, the mean-field trajectory $\bar{z}^{(k)}$ is a k-latent LTI sequence. Hence, the cost-minimizing control under $\bar{z}^{(k)}$ can be written in parameterized form[3] as:

$$u_t^{(k)} = \mu_t(z_t; (\bar{z}_{0:k}^{(k)}, r)) := g_p\big(apz_t - c_z l_k(t, \bar{z}^{(k)}, r)\big) \quad (15)$$

where

$$l_k(t, \bar{z}, r) := \begin{cases} \frac{(\gamma h_p)^{k-t} r \bar{z}_k}{1-\gamma h_p r} + q_k(t, \bar{z}) & \text{if } t < k \\ \frac{r^{t-k+1}\bar{z}_k}{1-\gamma h_p r^{(k)}} & \text{if } t \geq k \end{cases}$$

and $q_k(t, \bar{z}) := \sum_{s=0}^{k-t-1} (\gamma h_p)^s \bar{z}_{t+1+s}$.

Note that the control expressed in (15) has a closed-form (without infinite sums) and is indeed calculated forward-in-time. The mean-field trajectory is then updated by the operator $\mathcal{T}$, which first executes the control in (15), then aggregates the generated mean-field trajectory by averaging the states over all agents,

$$\bar{z}_{t+1}^{(k+1)} = a\bar{z}_t^{(k)} + bu_t^{(k)}, \quad (16)$$

where $\bar{z}_0^{(k+1)} = \nu_0$, $0 \leq t \leq k$. This closes the loop and leads to a computable version of iterating the operator $\mathcal{T}$. The details of the algorithm are summarized in Algorithm 1.

Algorithm 1 generates iterates $\bar{z}^{(k)}$ that approach the equilibrium mean-field trajectory $\bar{z}^*$. Furthermore, the minimum number of iterations required to reach a given accuracy can be represented in terms of the desired accuracy, the

[3]With some abuse of notation, we have replaced the (infinite) mean-field trajectory with its parameterized form.

Algorithm 1: Policy iteration for LQ-MFGs

Data: $a, b, c_z, c_u, \gamma, \nu_0$, $|r| \leq 1$, and $\varepsilon_s > 0$

Initialize: Set $\bar{z}^0$ as a 0-latent LTI mean-field with $\bar{z}_0^{(0)} = \nu_0$, $k = 0$;

$p \leftarrow (-\alpha + \sqrt{\alpha^2 + 4\beta})/2$, where $\alpha = \frac{c_u(1-\gamma a^2)}{\gamma b^2} - c_z$ and $\beta = \frac{c_z c_u}{\gamma b^2}$

$g \leftarrow -\gamma b/(c_u + \gamma b^2 p)$

$h \leftarrow a(1 + bpg)$

$T \leftarrow |h| + |bgc_z/(1 - \gamma h)|$

while $\max_{0 \leq t \leq k} \left| \bar{z}_t^{(k)} - \bar{z}_t^{(k-1)} \right| > \varepsilon_s(1 - T)/T$ **do**

  $\bar{z}_0^{(k+1)} \leftarrow \nu_0$

  for $m \in \{0, 1, \ldots, k\}$ **do**

    $\bar{z}_{m+1}^{(k+1)} \leftarrow a\bar{z}_m^{(k)} + b\mu_m(\bar{z}_m^{(k)}; (\bar{z}_{0:k}^{(k)}, r))$

  $k \leftarrow k + 1$

return Parameter tuple $(\bar{z}_{0:k}^{(k)}, r)$ that yields the control $\mu(\cdot; (\bar{z}_{0:k}^{(k)}, r))$ (see (15))

initial approximation error, the contraction coefficient, and the constant of the linear dynamics. The convergence is summarized by the following theorem.

Theorem 2. *Under Assumption 1, given $\varepsilon_s > 0$ there exists a $k^* > K(\varepsilon_s) := \lceil (\log \varepsilon_s - \log \|\bar{z}^{(0)} - \bar{z}^*\|_\infty) / \log T_p \rceil$ such that $\|\bar{z}^{(k^*)} - \bar{z}^*\|_\infty < \varepsilon_s$, where T_p was introduced in Assumption 1.*

V. Numerical Results

In this section we present simulations to demonstrate the performance of Algorithm 1 that approximates the equilibrium mean-field of the LQ-MFG. We use a normal distribution with mean and variance $\nu_0 = 20.0$ and $\sigma_0^2 = 1.0$, respectively, to generate the initial condition of the generic agent z_0. The dynamics of the generic agent are defined as in (3) and the parameters are $a = 1.1315$ and $b = 0.7752$. The standard deviation of the noise process is $\sigma_\omega = 0.03$. The cost function has the form shown in (4) with values $c_z = 0.0392$ and $c_u = 1.6864$. The positive solution of the resulting Riccati equation, given by (9). is $p = 0.8787$ with $g_p = -0.3227, h_p = 0.8828$ and $T_p = 0.9305 < 1$. The algorithm starts with initial mean field $\bar{z}^{(0)}$, which is a 0-latent LTI mean-field with parameters $\bar{z}_0 = \nu_0, \quad r = 0.6$.

Figure 1 shows approximations of the mean-field for different values of ε_s. As shown, for decreasing values of ε_s the approximations approach the equilibrium mean-field. Interestingly, the algorithm reaches a good approximation ($\varepsilon_s = 0.005$) in a small number of iterations ($k = 40$). Figure 2 depicts the average cost per agent for different numbers of agents, N and for different values of ε_s. Each plot in the figure corresponds to a different number of agents N. As N increases, the average cost is seen to decrease. This provides evidence that our conjecture, regarding policies obtained from the infinite population case when applied to the finite population case, is correct. The figure also shows

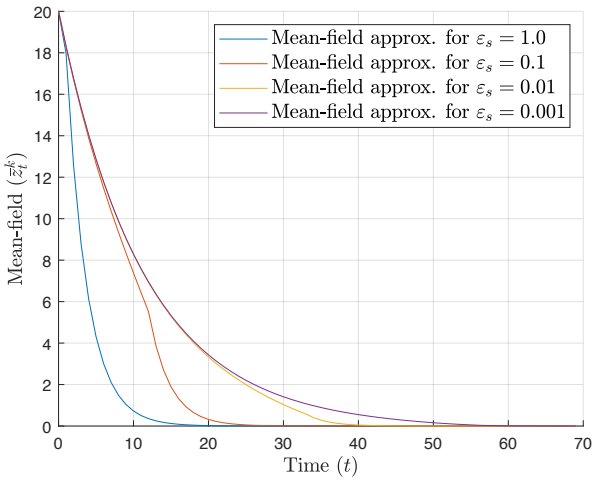

Fig. 1. Mean-field approximation for different values of ε_s. Notice the convergence of the mean-field trajectory as ε_s decreases.

that as the approximations become better, there is a decrease in the average cost per agent.

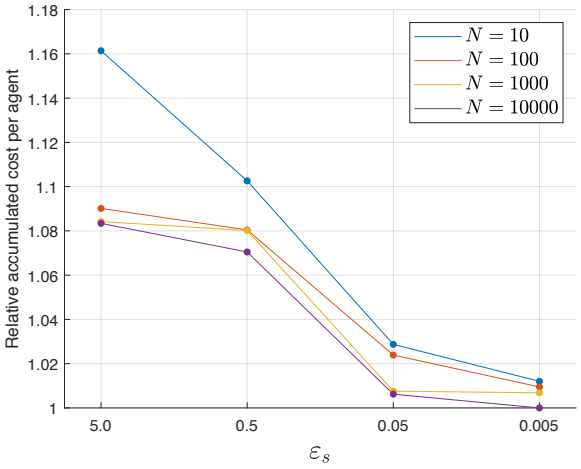

Fig. 2. Relative accumulated cost per agent w.r.t. ε_s. Values are normalized to the lowest cost obtained ($N = 10000$, $\varepsilon_s = 0.005$).

VI. Concluding Remarks and Future Directions

We have developed a policy iteration algorithm for approximating equilibria in infinite-horizon LQ-MFGs with discounted cost. The main challenge in the algorithm development arises from the fact that the optimal control is computed backward in time. By investigating properties of the mean-field update operator (which we term the τ-latent property), we can represent the mean-field trajectory at any given iteration by a finite set of parameters, resulting in a forward-in-time construction of the optimal control. The algorithm is provably convergent, with numerical results demonstrating the nature of convergence. The optimality of the computed equilibrium has been empirically studied; naturally, the optimality of the approximate equilibrium improves as the iteration index increases and the stopping

threshold decreases. The results derived in this paper provide an algorithmic viewpoint of the nature of mean-field equilibria for LQ-MFGs. We believe that such insights will be useful for developing model-free RL algorithms. Future work includes an extension to the multivariate case as well as consideration of a nonlinear/non-quadratic model (see [26]).

APPENDIX

Proof of Lemma 1. Substituting $K_t = \gamma^t p_t$ and $g_t = \gamma^t \lambda_t$ into (21)–(26) of [29] yields (similar derivations in [30] and on p. 234 of [31]),

$$\begin{aligned}
u_t &= -\gamma b(c_u + b^2\gamma p_{t+1})^{-1}(p_{t+1}az_t + \lambda_{t+1}),\\
\lambda_t &= \gamma a\left(1 - \gamma p_{t+1}b^2(c_u + \gamma b^2 p_{t+1})^{-1}\right)\lambda_{t+1} - c_z\bar{z}_t,\\
p_t &= \gamma a^2 p_{t+1} + c_z - \gamma^2 a^2 b^2 p_{t+1}^2 (c_u + \gamma b^2 p_{t+1})^{-1}.
\end{aligned}$$

Since it is an infinite horizon problem, the Riccati equation will have a steady state solution. This can be written as,

$$\begin{aligned}
u_t &= -\gamma b(apz_t + \lambda_{t+1})/(c_u + \gamma b^2 p),\\
p &= \gamma a^2 p + c_z - \gamma^2 a^2 b^2 p^2/(c_u + \gamma b^2 p),\\
\lambda_t &= \gamma a\left(1 - \gamma p b^2/(c_u + \gamma b^2 p)\right)\lambda_{t+1} - c_z\bar{z}_t.
\end{aligned}$$

Defining $g_p := -\gamma b/(c_u + \gamma b^2 p)$ and $h_p := a(1 + bpg_p)$, the above expressions for u_t and λ_t correspond to (7) and (10), respectively. Rearranging and grouping p terms in the above expression yields (8), with unique positive solution (9). □

Proof of Lemma 2. 1) First, we show that $\gamma|h_p| < 1$. It is well known [30] that the DARE for variables (a_c, b_c, s_c, r_c) and average cost is $\hat{k} = a_c^2\hat{k} + s_c - a_c^2\hat{k}^2 b_c^2/(r_c + b_c^2\hat{k})$. If $b_c \neq 0$ and $s_c > 0$, then this equation will have a positive solution. Moreover, the optimal feedback gain is $l_c := -a_c b_c\hat{k}/(r_c + b_c^2\hat{k})$ and the closed-loop gain $|a_c + b_c l_c| < 1$. By using a change of variables with $a_c = \sqrt{\gamma}a, b_c = b, s_c = c_z, r_c = c_u/\gamma$, the equation (8) is recovered with $\hat{k} = p$. Hence there exists a unique positive solution for (8), given by (9). Moreover $|a_c + b_c l_c| = \left|\sqrt{\gamma}a - \frac{\gamma\sqrt{\gamma}ab^2p}{c_u + \gamma b^2 p}\right| = \sqrt{\gamma}\,|a(1 + bpg)| = \sqrt{\gamma}|h_p| < 1$ and thus $\gamma|h_p| < \sqrt{\gamma}|h_p| < 1$. From (10), recursing backwards yields $\lambda_0 = (\gamma h_p)^t \lambda_t - c_z\sum_{s=0}^{t-1}(\gamma h_p)^s \bar{z}_s$. Under the assumption $\lambda_0 = -c_z\sum_{s=0}^{\infty}(\gamma h_p)^s\bar{z}_s$, it follows that $(\gamma h_p)^t\lambda_t = -c_z\sum_{s=t}^{\infty}(\gamma h_p)^s\bar{z}_s \Rightarrow \lambda_t = -c_z\sum_{s=0}^{\infty}(\gamma h_p)^s\bar{z}_{t+s}$. As $\bar{z} \in \ell^\infty$ there exists some $0 \leq \bar{z}_\infty < \infty$ s.t. $|\bar{z}_t| \leq \bar{z}_\infty$. This translates to $|\lambda_t| \leq c_z\bar{z}_\infty/(1 - \gamma h_p)$ for all t. Hence $\lambda \in \ell^\infty$.

2) The closed-loop dynamics of z_t under the cost-minimizing control are, $z_{t+1} = az_t + bg_p(paz_t + \lambda_{t+1}) + \omega_t$. Using this equation recursively, the expression for z_t in terms of z_0 is, $z_t = h_p^{t-1}z_0 + bg_p\sum_{s=0}^{t-1}h_p^s\lambda_{t-s} + \sum_{s=0}^{t-1}h_p^s\omega_{t-1-s}$. The expression for $\mathbb{E}\left[\left(z_t - \bar{z}_t\right)^2\right]$ is thus,

$$\begin{aligned}
\mathbb{E}\left[\,(z_t - \bar{z}_t)^2\,\right] &= \left(h_p^{t-1}\right)^2(\sigma_0^2 + \nu_0^2) + \sigma_w^2\sum_{s=0}^{t-1}h_p^s +\\
&\left(bg_p\sum_{s=0}^{t-1}h_p^s\lambda_{t-s} - \bar{z}_t\right)^2 + 2h_p^{t-1}\nu_0\left(bg_p\sum_{s=0}^{t-1}h_p^s\lambda_{t-s} - \bar{z}_t\right).
\end{aligned}$$

Assumption 1 implies that $|h_p| < 1$. Furthermore, since $\bar{z} \in \ell^\infty$ and $\lambda \in \ell^\infty$, there exist constants $\bar{z}_\infty, \lambda_\infty < \infty$ such that $\bar{z}_t \leq \bar{z}_\infty$ and $\lambda_t \leq \lambda_\infty$ for all t. Thus,

$$\begin{aligned}
\mathbb{E}\left[(z_t - \bar{z}_t)^2\right] &\leq \left(h_p^{t-1}\right)^2(\sigma_0^2 + \nu_0^2) + \sigma_w^2 t\\
&+ (bg_p t\lambda_\infty - \bar{z}_\infty)^2 + 2h_p^{t-1}\nu_0\,(bg_p t\lambda_\infty - \bar{z}_\infty)\,.
\end{aligned}$$

Similarly, $\mathbb{E}\left[u_t^2\right]$ is bounded above as,

$$\begin{aligned}
\mathbb{E}\big[u_t^2\big] &= (ag_p p)^2\big[\left(h_p^{t-1}\right)^2\left(\sigma_0^2 + \nu_0^2\right) + \sum_{s=0}^{t-1}\sigma_w^2 h_p^{2s}\\
&+ bg_p h_p^s\lambda_{t-s} + 2h_p^{t-1}\nu_0 bg_p h_p^s\lambda_{t-s}\big] + g_p^2\lambda_{t+1}^2\\
&+ 2ag_p^2 p\lambda_{t+1}\left(h_p^{t-1}\nu_0 + bg_p\sum_{s=0}^{t-1}h_p^s\lambda_{t-1}\right)\\
&\leq (ag_p p)^2\left[\left(h_p^{t-1}\right)^2\left(\sigma_0^2 + \nu_0^2\right) + \sigma_w^2 t\right.\\
&\left.+ (bg_p t\lambda_\infty) + 2h_p^{t-1}\nu_0 bg_p t\lambda_\infty\right] + g_p^2\left(\lambda_\infty\right)^2\\
&+ 2ag_p^2 p\lambda_\infty\left(h_p^{t-1}\nu_0 + bg_p t\lambda_\infty\right).
\end{aligned}$$

Since the optimal cost is, $\sum_{t=0}^{\infty}c_z\mathbb{E}\big[\gamma^t\left(z_t - \bar{z}_t\right)^2\big] + c_u\mathbb{E}\big[\gamma^t u_t^2\big]$, and $\sum_{t=0}^{\infty}t\gamma^t = \frac{\gamma}{(1-\gamma)^2}$, $\sum_{t=0}^{\infty}\gamma^t\left(h_p^{t-1}\right)^2 = \frac{1}{h_p^2\left(1-\gamma h_p^2\right)}$, $\sum_{t=0}^{\infty}t\gamma^t h_p^{t-1} = \frac{\gamma}{(1-\gamma h_p)^2}$ it can be concluded that the optimal cost is bounded. □

Proof of Lemma 3. Let us define two mean-fields $\bar{z}, \hat{z} \in \ell^\infty$ and their next iterates $\bar{z}' = \mathcal{T}(\bar{z}), \hat{z}' = \mathcal{T}(\hat{z})$. Let us define the difference sequences $\delta_t = \bar{z}_t - \hat{z}_t$ and $\delta_t' = \bar{z}_t' - \hat{z}_t'$. Using (12), the equation expressing the connection between δ_t and δ_t' is $\delta_{t+1}' = h_p\delta_t - c_z bg_p\sum_{s=0}^{\infty}(\gamma h_p)^s\,\delta_{t+1+s}$. Hence,

$$\begin{aligned}
\|\delta'\|_\infty &\leq \|\delta\|_\infty\left(\left|h_p\right| + \left|c_z bg_p\sum_{s=0}^{\infty}\left(\gamma h_p\right)^s\right|\right)\\
&\leq \|\delta\|_\infty\left(\left|h_p\right| + \left|c_z bg_p/(1 - \gamma h_p)\right|\right) = \|\delta\|_\infty T_p
\end{aligned}$$

where the last inequality follows from $\gamma|h_p| < 1$ (see Lemma 2). By Assumption 1, $\mathcal{T}$ is a contraction. □

Proof of Theorem 1. Consider an MFE $(\mu^*, \bar{z}^*)$ that satisfies Definition 1. Then, by definition, $\mu^* = \Phi(\bar{z}^*)$. The second part of Definition 1 states that $\bar{z}^* = \Lambda(\mu^*)$. Thus $\bar{z}^* = \Lambda(\Phi(\bar{z}^*)) = \mathcal{T}(\bar{z}^*)$. Now let us prove the converse. Consider a mean-field $\bar{z}^*$ which is the fixed point of $\mathcal{T}$ i.e. $\bar{z}^* = \mathcal{T}(\bar{z}^*)$. Then if μ^* is the cost-minimizing control for $\bar{z}^*$ i.e. $\mu^* = \Phi(\bar{z}^*)$, $(\mu^*, \bar{z}^*)$ is an MFE since (1) $\mu^* = \Phi(\bar{z}^*)$, and (2) $\Lambda(\mu^*) = \Lambda(\Phi(\bar{z}^*)) = \mathcal{T}(\bar{z}^*) = \bar{z}^*$. □

Proof of Lemma 4. For $t \in \{\tau, \tau+1, \ldots\}$ using (12) and the fact that $|r| \leq 1$, we can write $x_{t+1}' = h_p x_t - c_z bg_p\sum_{s=0}^{\infty}(\gamma h_p r)^s\,r x_t = \hat{r}_p x_t$ where $\hat{r}_p := h_p - \frac{c_z bg_p r}{1 - \gamma h_p r}$. Similarly, x_{t+2}' is generated as $x_{t+2}' = \hat{r}_p x_{t+1} = \hat{r}_p r x_t$ for all $t \in \{\tau, \tau+1, \ldots\}$. Grouping terms, we obtain $x_{t+2}' = r x_{t+1}'$ for all $t \in \{\tau, \tau+1, \ldots\}$. □

Proof of Theorem 2. We first state and prove in Lemma 5 below that the expression in the stopping condition of the algorithm $\max_{0\leq t\leq k}\left|\bar{z}_t^{(k)} - \bar{z}_t^{(k-1)}\right|$ is equal to $\left\|\bar{z}^{(k)} -\right.$

$\bar{z}^{(k-1)}\big\|_\infty$. This is due to the fact that $\bar{z}^{(k)}$ and $\bar{z}^{(k-1)}$ both follow stable linear dynamics for $t \geq k$.

Lemma 5. $\max_{0\leq t\leq k} \big|\bar{z}_t^{(k)} - \bar{z}_t^{(k-1)}\big| = \big\|\bar{z}^{(k)} - \bar{z}^{(k-1)}\big\|_\infty$.

Proof. By definition $\big\|\bar{z}^{(k)} - \bar{z}^{(k-1)}\big\|_\infty = \sup_{t\geq 0} \big|\bar{z}_t^{(k)} - \bar{z}_t^{(k-1)}\big|$. Hence for all $t \geq k$, $\big|\bar{z}_t^{(k)} - \bar{z}_t^{(k-1)}\big| = \big|r^{t-k}\big(\bar{z}_k^{(k)} - \bar{z}_k^{(k-1)}\big)\big| \leq \big|\bar{z}_k^{(k)} - \bar{z}_k^{(k-1)}\big|$. Using this property, $\big\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\big\|_\infty = \sup_{t\geq 0} \big|\bar{z}_t^{(k+1)} - \bar{z}_t^{(k)}\big| = \max_{0\leq t\leq k+1} \big|\bar{z}_t^{(k+1)} - \bar{z}_t^{(k)}\big|$. □

Since $\mathcal{T}$ is contractive with a fixed point of $\bar{z}^*$,

$$\big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty \leq T_p \big\|\bar{z}^{(k)} - \bar{z}^*\big\|_\infty \tag{17}$$

for any $k = 1, 2, \ldots$. The algorithm terminates at iteration k when $\big\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\big\|_\infty < \varepsilon_s(1 - T_p)/T_p$. Thus,

$$\begin{aligned}
\varepsilon_s(1 - T_p)/T_p &> \big\|\bar{z}^{(k+1)} - \bar{z}^{(k)}\big\|_\infty \\
&\geq \big\|\bar{z}^{(k)} - \bar{z}^*\big\|_\infty - \big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty \\
&\geq \frac{1}{T_p}\big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty - \big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty \\
&= \big(1 - T_p\big)\big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty / T_p.
\end{aligned}$$

Hence, $\big\|\bar{z}^{(k+1)} - \bar{z}^*\big\|_\infty < \varepsilon_s$ for any $\varepsilon_s > 0$. Now we prove the bound on the number of iterations. If the number of iterations is $k > K(\varepsilon_s)$, then,

$$\begin{aligned}
&k > \big(\log \varepsilon_s - \log \big\|\bar{z}^{(0)} - \bar{z}^*\big\|_\infty\big)/\log T_p \\
&\Leftrightarrow k \log T_p < \log \varepsilon_s - \log \big\|\bar{z}^{(0)} - \bar{z}^*\big\|_\infty \\
&\Leftrightarrow T_p^k < \frac{\varepsilon_s}{\big\|\bar{z}^{(0)} - \bar{z}^*\big\|_\infty} \Leftrightarrow T_p^k \big\|\bar{z}^{(0)} - \bar{z}^*\big\|_\infty < \varepsilon_s.
\end{aligned} \tag{18}$$

The inequality flip in the second step is due to the fact that $T_p < 1$ (Assumption 1) and $\log T_p < 0$. From (17) $\big\|\bar{z}^{(k)} - \bar{z}^*\big\|_\infty \leq T_p^k \big\|\bar{z}^{(0)} - \bar{z}^*\big\|_\infty$ and using the inequality (18), $\big\|\bar{z}^{(k)} - \bar{z}^*\big\|_\infty < \varepsilon_s$. □

REFERENCES

[1] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[2] M. Wooldridge, *An Introduction to Multiagent Systems*. John Wiley & Sons, 2009.

[3] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, "Distributed event-triggered control for multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1291–1297, 2011.

[4] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*, 2018, pp. 5867–5876.

[5] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Finite-sample analyses for fully decentralized multi-agent reinforcement learning," *arXiv preprint arXiv:1812.02783*, 2018.

[6] R. Breban, R. Vardavas, and S. Blower, "Mean-field analysis of an inductive reasoning game: Application to influenza vaccination," *Physical Review E*, vol. 76, no. 3, p. 031127, 2007.

[7] R. Couillet, S. M. Perlaza, H. Tembine, and M. Debbah, "Electrical vehicles in the smart grid: A mean field game analysis," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1086–1096, 2012.

[8] M. Huang, R. P. Malhamé, P. E. Caines *et al.*, "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.

[9] M. Huang, P. E. Caines, and R. P. Malhamé, "Individual and mass behaviour in large population stochastic wireless power control problems: Centralized and Nash equilibrium solutions," in *IEEE International Conference on Decision and Control*, vol. 1. IEEE, 2003, pp. 98–103.

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

[11] 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.

[12] A. Bensoussan, K. Sung, S. C. P. Yam, and S.-P. Yung, "Linear-quadratic mean field games," *Journal of Optimization Theory and Applications*, vol. 169, no. 2, pp. 496–529, 2016.

[13] M. Huang and M. Zhou, "Linear quadratic mean field games–part I: The asymptotic solvability problem," *arXiv preprint arXiv:1811.00522*, 2018.

[14] J. Moon and T. Başar, "Discrete-time LQG mean field games with unreliable communication," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 2697–2702.

[15] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 251–259.

[16] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," *arXiv preprint arXiv:1901.09585*, 2019.

[17] R. Elie, J. Pérolat, M. Laurière, M. Geist, and O. Pietquin, "Approximate fictitious play for mean field games," *arXiv preprint arXiv:1907.02633*, 2019.

[18] Z. Fu, Z. Yang, Y. Chen, and Z. Wang, "Actor-critic provably finds Nash equilibria of linear quadratic mean field games," in *International Conference on Learning Representations*, 2020.

[19] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv preprint arXiv:1911.10635*, 2019.

[20] S. J. Bradtke, "Reinforcement learning applied to linear quadratic regulation," in *Advances in Neural Information Processing Systems*, 1993, pp. 295–302.

[21] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*, 2018, pp. 1467–1476.

[22] K. Zhang, Z. Yang, and T. Başar, "Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games," in *Advances in Neural Information Processing Systems*, 2019.

[23] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, 2014.

[24] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic control*, vol. 59, no. 11, pp. 3051–3056, 2014.

[25] P. Cardaliaguet and C.-A. Lehalle, "Mean field game of controls and an application to trade crowding," *Mathematics and Financial Economics*, vol. 12, no. 3, pp. 335–363, 2018.

[26] N. Saldi, T. Başar, 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] M. Huang, "Stochastic control for distributed systems with applications to wireless communications," Ph.D. dissertation, McGill University, 2003.

[28] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

[29] K. Yazdani and M. Hale, "Technical report: Infinite horizon discrete-time linear quadratic Gaussian tracking control derivation," *arXiv preprint arXiv:1807.04700*, 2018.

[30] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.

[31] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Siam, 1999, vol. 23.