Entropy Regularization for Mean Field Games with Learning

Xin Guo [*] Renyuan Xu [†] Thaleia Zariphopoulou [‡]

June 13, 2021

Abstract

Entropy regularization has been extensively adopted to improve the efficiency, the stability, and the convergence of algorithms in reinforcement learning. This paper analyzes both quantitatively and qualitatively the impact of entropy regularization for Mean Field Games (MFGs) with learning in a finite time horizon. Our study provides a theoretical justification that entropy regularization yields time-dependent policies and, furthermore, helps stabilizing and accelerating convergence to the game equilibrium. In addition, this study leads to a policy-gradient algorithm for exploration in MFG. With this algorithm, agents are able to learn the optimal exploration scheduling, with stable and fast convergence to the game equilibrium.

1 Introduction

Reinforcement learning (RL) is one of the three fundamental machine learning paradigms, alongside supervised learning and unsupervised learning. RL is learning via trial and error, through interactions with an environment and possibly with other agents; in RL, an agent takes an action and receives a reinforcement signal in terms of a numerical reward, which encodes the outcome of her action. In order to maximize the accumulated reward over time, the agent learns to select her actions based on her past experiences (exploitation) and/or by making new choices (exploration).

Exploration and exploitation are the essence of RL. Exploration provides opportunities to improve from current sub-optimal solutions to the ultimate global optimal one, yet is time consuming and computationally expensive as over-exploration may impair the convergence to the optimal solution. Meanwhile, pure exploitation, i.e., myopically picking the current solution based solely on past experience, though easy to implement, tends to yield sub-optimal global solutions. Therefore, an appropriate trade-off between exploration-exploitation is crucial for RL algorithm design to improve the learning and the optimization procedure.

Entropy regularization. One common approach to balance the exploration-exploitation in RL is to introduce entropy regularization [1, 28, 30]. In RL settings with more than one agent, there are two major sources of uncertainty: the unknown environment and the actions of the other agents. Shannon entropy and cross-entropy are two natural choices for entropy regularization: the former quantifies the information gain of exploring the environment while the latter measures the benefit

[*]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. Email: xinguo@berkeley.edu.

[†]Mathematical Institute, University of Oxford, Oxford, UK. Email: xur@maths.ox.ac.uk.

[‡]Departments of Mathematics and IROM, The University of Texas at Austin, Austin, USA, and the Oxford-Man Institute, University of Oxford, Oxford, UK. Email: zariphop@math.utexas.edu.

from exploring the actions of other agents. This information-theoretic perspective of exploration has been well understood in single-agent RL; see for instance [14, 17, 28, 30, 33].

However, there is virtually no theoretical study on the role of entropy regularization in multi-agent RL (MARL), with the exception of [2]. Indeed, most existing studies are empirical, demonstrating convergence improvement and variance reduction when entropy regularization is added. For instance, [19] showed via empirical analysis that policy features can be learned directly from pure observations of other agents and that the non-stationarity of the environment can be reduced by adding cross-entropy; [18] applied the cross-entropy regularization to demonstrate the convergence of fictitious play in a discrete-time model with a finite number of agents while [32] used the cross-entropy loss to train the prediction of other agents' actions via observations of their behavior. The only theoretical work so far can be found in [2] in an infinite horizon setting in which a regularized Q-learning algorithm for stationary discrete-time mean field games was proposed along with its convergence analysis. Still, the problem remains open for the finite time horizon case, which arise often in many applications.

Optimal exploration scheduling. Another major challenge for both single-agent RL and MARL is exploration efficiency. In practice, there are various heuristic designs of explorations for MARL, including adding random noise in the parameter space [31], the approach of ε-greedy policy [37], and the method with softmax [22]. However, there is no theoretical validation of these approaches.

Recently, time-invariant Gaussian exploration was applied to single-agent RL ([20, 29, 34]) and time-dependent "optimal exploration scheduling" was derived for single-agent mean-variance portfolio selection problem in [35]. In these works, the degree of exploration was characterized by the variance of the Gaussian distribution and the term "optimal exploration scheduling" was coined for the time-dependent variance of the Gaussian distribution.

Exploration schemes are inherently time-dependent, as it is necessary to balance the free exploration at the initial phase and the greedy control policy towards terminal time. Yet, it seems that there is no existing work on analyzing such time-dependent learning policies for MARL, neither empirical or theoretical.

Model-free vs model-based approach for MFG with learning. There are two popular approaches in single-agent reinforcement-learning to handle unknown or partially known environments: the model-based approach and the model-free approach. In the model-based paradigm, the agent is assumed to know the model structure but has no access to the model parameters. In this case, the agent estimates the unknown model parameters and, then, constructs a control policy based on the knowledge of the model [3, 11]. In the model-free paradigm, the agent learns the optimal policy *directly* via interacting with the system, without inferring the model parameters. Examples of model-free approach include policy gradient method [6] and actor critic method [13]. In practice, due to the lack of information on the actual system, model-based approach tends to suffer from model mis-specification [6]. On the other hand, the execution of the model-free algorithm does not rely on the assumptions of the model, thus is more robust against model mis-specification [16].

Given the robustness of the model-free approach and the additional complexity from the game interactions, model-free approach appears more appropriate for MFG with learning where the representative agent faces uncertainties about both the unknown environment and the large population of strategic opponents.

Our work. In this paper, we propose to study entropy regularization for MARL with a large population, namely, within the framework of the mean field game (MFG). This transition from

MARL to MFG with learning is critical to avoid the curse of dimensionality in MARL.

We analyze both quantitatively and qualitatively the impact of entropy regularization in MFG with learning in a finite time horizon. We adopt two different entropies: first, the Shannon entropy and, then, a combination of Shannon entropy and the cross-entropy, which we call the *enhanced entropy*.

- We derive explicit Nash equilibrium (NE) solutions (Theorems 2 and 4) for a class of linear-quadratic (LQ) stochastic games. Our study provides a theoretical justification to the fact that entropy regularization yields time-dependent policies. Furthermore, it helps stabilizing and accelerating convergence to the game equilibrium.
- This theoretical study enables us to design a model-free policy-gradient algorithm for MFG with learning. Under this algorithm, agents are able to learn efficiently the optimal exploration scheduling in an unknown environment and with a large group of competing agents. The convergence to the game equilibrium is stable and fast when appropriate exploration rates are chosen.

Additional related works. Our algorithm is inspired by the recent success of policy-gradient method for single-agent LQ regulators [12]. In addition, there is a concurrent work on the global convergence of policy gradient for MFG [36], yet without exploration. We also mention recent works on two-agent zero-sum LQ games [38] and the LQ mean field control problem with common noise [10].

Organization. The rest of the paper is organized as follows. Section 2 provides the mathematical framework for MFG with learning, Section 3 focuses on analyzing the impact of Shannon entropy and the enhanced entropy in a class of LQ games, and Section 4 proposes a policy-gradient based algorithm with entropy regularization, and provides its numerical performance.

2 Mathematical Formulation

We start with the mathematical formulation of the MFG with learning.

Key ideas. There are several key components for the formulation.

The first component is the *aggregation idea* from the theory of MFG to address the curse of dimensionality in MARL. Specifically, it is to consider N agents, and assume that they are all identical, indistinguishable and interchangeable, and that interactions among them are based on the *macroscopic information*, which is the empirical state distribution and action distribution of all agents. This allows us to work instead with a representative agent i, her state X_t^i, her policy π_t^i at time $t \in [0, T]$, and her interaction with other agents through the macroscopic information. Since agent i depends on other agents only through the empirical measure, we may then consider both the population state distribution and action distribution if such limits exist when $N \to \infty$. Moreover, the subscript i can be dropped and one can focus on a representative agent in this MFG formulation since all agents are assumed to be identical and indistinguishable.

The second component is how to model learning and exploration via the notion of *randomized policies*, known in the control literature as *relaxed controls* and in the game theory as *mixed strategies*. These are policies, say π_t, of the representative agent with $\pi_t \in \mathcal{P}(U)$, where the action space U is a closed subset of a Euclidean space and $\mathcal{P}(U)$ is the set of density functions of probability measures on U that are absolutely continuous with respect to the Lebesgue measure. Namely, $\pi_t \in \mathcal{P}(U)$ if and only if

$$\int_U \pi_t(u)du = 1 \quad \text{and} \quad \pi_t(u) \geq 0 \ a.e. \ on \ U. \tag{2.1}$$

The third ingredient is the *entropy regularization*, which is adopted to encourage exploration. For this, we will use both the Shannon entropy and the cross-entropy, denoted by $\mathcal{H}_{SE}$ and $\mathcal{H}_{CE}$, respectively (see (2.5) and (2.6)).

Controlled state process with randomized policies. We incorporate the above components in a finite horizon setting $[0,T]$, $0 < T < \infty$. For this, we introduce $\mu := \{\mu_s, \ s \in [t,T]\}$ to be the flow of population state distribution with $\mu_s \in \mathcal{P}(\mathbb{R})$ and $\alpha := \{\alpha_s, \ s \in [t,T]\}$ to be the flow of population action distribution with $\alpha_s(\cdot; x) \in \mathcal{P}(U)$ for any $x \in \mathbb{R}$, starting from time $t \in [0,T]$. Occasionally, α and μ will be also called the *mean field information.*

Next we define the controlled state process of the representative agent. Given $t \in [0,T]$ and exogenous flows, say α and μ with $\mu_t = \nu$, the representative agent adopts a randomized policy $\pi = \{\pi_s \in \mathcal{P}(U), s \in [t,T]\}$ over an admissible policy set $\mathcal{A}$ (to be specified below). Then, following the paradigm recently proposed in [34], the controlled state process is assumed to follow

$$\begin{aligned} dX_s^\pi &= \left(\int_U b(s, X_s^\pi, \mu_s, \alpha_s, u)\pi_s(u)du\right) ds + \left(\sqrt{\int_U \sigma^2(s, X_s^\pi, \mu_s, \alpha_s, u)\pi_s(u)du}\right) dW_s, \quad (2.2) \\ X_t^\pi &= \xi \sim \nu, \quad \mu_t = \nu, s \in [t,T]. \end{aligned}$$

Here $W = \{W_t\}_{t\in[0,T]}$ is a standard Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$, with $\{\mathcal{F}_t\}_{t\in[0,T]}$ satisfying the usual conditions; $\nu \in \mathcal{P}(U)$ is the distribution of the initial state satisfying $\int x^2\nu(dx) < \infty$; ξ is a random variable independent of W and $\mathcal{F}_t$-measurable; and $b, \sigma : [0,T] \times \mathbb{R} \times \mathcal{P}(\mathbb{R}) \times \mathcal{P}(U) \times U \hookrightarrow \mathbb{R}$.

We note the particular form of the state process (2.2) is a consequence of the aggregation of $\widehat{X}_s^u$ over action $u_s \in U$ where

$$d\widehat{X}_s^u = b\left(s, \widehat{X}_s^u, \mu_s, \alpha_s, u_s\right) ds + \sigma\left(s, \widehat{X}_s^u, \mu_s, \alpha_s, u_s\right) dW_s.$$

Such policies $u_s \in U$ are also called *pure strategies* in game theory. Pure strategies and mixed strategies are closely related. Indeed, $u = \{u_s, s \in [0,T]\}$ can be regarded as a Dirac measure $\pi = \{\pi_s(u), s \in [0,T]\}$ where $\pi_s(\cdot) = \delta_{u_s}(\cdot) \in U$. In this case π_s does not have a density, and hence $\pi_s \notin \mathcal{P}(U)$. (We refer the readers to [34] for more details).

Game payoff with entropy regularization. The objective of the representative agent is to maximize her payoff function J and solve for

$$V(t \,|\, \mu, \alpha) = \sup_{\pi \in \mathcal{A}} J(t, \pi \,|\, \mu, \alpha), \tag{2.3}$$

where the entropy-regularized payoff is defined as

$$\begin{aligned} J(t, \pi \,|\, \mu, \alpha) = \mathbb{E}\Bigg[\int_t^T \left(\int_U \left(r(s, X_s^\pi, \mu_s, \alpha_s, u)\pi_s(u)du + \lambda_{SE}\mathcal{H}_{SE}(\pi_s) + \lambda_{CE}\mathcal{H}_{CE}(\pi_s, \alpha_s, \mu_s)\right)\right) ds \\ + g(X_T^\pi, \mu_T, \alpha_T)\Big|\mu, \alpha\Bigg]. \end{aligned} \tag{2.4}$$

The Shannon entropy $\mathcal{H}_{SE}$ and cross-entropy $\mathcal{H}_{CE}$ are defined as

$$\mathcal{H}_{SE}(\pi_s) = -\int_U \pi_s(u) \ln \pi_s(u) du, \;\; \pi_s \in \mathcal{P}(U), \tag{2.5}$$

$$\mathcal{H}_{CE}(\pi_s, \alpha_s, \mu_s) = -\int_U \pi_s(u) \int \ln \alpha_s(u;x)\, \mu_s(dx) du, \;\; \pi_s \in \mathcal{P}(U). \tag{2.6}$$

In addition, $r : [0,T] \times \mathbb{R} \times \mathcal{P}(\mathbb{R}) \times \mathcal{P}(U) \times U \hookrightarrow \mathbb{R}$ and $g : [0,T] \times \mathbb{R} \times \mathcal{P}(U) \times \mathcal{P}(U) \hookrightarrow \mathbb{R}$ are the running reward and terminal reward functions of the representative agent, while $\lambda_{SE} > 0$ is the (temperature) parameter to control the degree of self-exploration and $\lambda_{CE} \geq 0$ is the (temperature) parameter to control the degree of exploration on the actions of the other agents. From an information-theoretic perspective, $\lambda_{SE}\mathcal{H}_{SE}$ and $\lambda_{CE}\mathcal{H}_{CE}$ quantify the information gain from exploring the unknown environment and the policies chosen by the other agents.

Observable quantities. In a game with learning, the functions b, σ, r and g are *unknown*. The representative agent takes actions while interacting with (the continuum) of the other agents. This interaction takes several rounds.

In each round starting from time 0, the agent observes $\{\alpha_s\}_{s\in[0,t]}$, $\{\mu_s\}_{s\in[0,t]}$ and $\{X_s^\pi\}_{s\in[0,t]}$ at time $t \in [0,T]$; the reward will not be revealed until time T, the end of each round; at time T, she will observe the *realized* cumulative reward $\widehat{j}\,(0, \pi|\alpha, \mu)$ with

$$\begin{aligned}\widehat{j}\,(0,\pi|\alpha,\mu) \;:=\; & \int_0^T \Big(\int_U \Big(r(s, X_s^\pi, \mu_s, \alpha_s, u)\pi_s(u)du + \lambda_{SE}\mathcal{H}_{SE}(\pi_s) + \lambda_{CE}\mathcal{H}_{CE}(\pi_s, \alpha_s, \mu_s)\Big)\Big)\, ds \\ & + g(X_T^\pi, \mu_T, \alpha_T),\end{aligned}$$

which is associated with the corresponding *single* trajectory $\{X_s^\pi\}_{s\in[0,T]}$ under policy π and the population behavior $\{\alpha_s\}_{s\in[0,T]}$, $\{\mu_s\}_{s\in[0,T]}$ in this round. Note that $\widehat{j}\,(0,\pi|\alpha,\mu)$ is one realized sample reward, which is different from the expected reward in (2.4).

Admissible policies. A policy $\pi \in \mathcal{A}(t, \mu, \alpha)$ is admissible if

(i) for each $s \in [t,T]$, $\pi_s \in \mathcal{P}(U)$ a.s.;

(ii) for each $Z \in \mathcal{B}(U)$ with $\mathcal{B}(U)$ being the Borel algebra on U, $\{\int_Z \pi_s(u)du, s \in [t,T]\}$ is $\mathcal{F}_t$-progressively measurable;

(iii) the SDE (2.2) admits a unique strong solution $X^\pi := \{X_s^\pi, s \in [t,T]\}$, with π being used;

(iv) the expectation on the right hand side of (2.4) is finite;

(v) there exists a measurable function $\widetilde{\pi} : [t,T] \times \mathbb{R} \to \mathcal{P}(U)$ such that

$$\mathbb{P}\Big(\pi_s(du) = \widetilde{\pi}_s(du; X_s^\pi), \quad \forall s \in [t,T]\Big) = 1.$$

Condition (v) imposes that the admissible policy is Markovian, i.e., closed-loop policy in feedback form.

Alternative formulation of the MFG with learning. We note that problem (2.3) treats the initial state ξ as a genuine source of randomness, in addition to the stochasticity from the Brownian motion W. Frequently, the following alternative interpretation, with a *deterministic* initial state x is useful for solving analytically the MFG. Specifically, let

$$
\begin{aligned}
\widetilde{V}(t,x\,|\,\mu,\alpha) :=\ & \sup_{\pi\in\mathcal{A}} \widetilde{J}(t,x\,|\,\pi,\mu,\alpha) \\
:=\ & \mathbb{E}\left[\int_t^T \left(\int_U \left(r(X_s^\pi,\mu_s,\alpha_s,u)\pi_s(u)du + \lambda_{SE}\mathcal{H}_{SE}(\pi_s) + \lambda_{CE}\mathcal{H}_{CE}(\pi_s,\alpha_s,\mu_s)\right)\right) ds \right.\\
& \left. + g(X_T^\pi,\mu_T,\alpha_T)\Big| X_t^\pi = x,\mu,\alpha\right],
\end{aligned}
\tag{2.7}
$$

subject to

$$
\begin{aligned}
dX_s^\pi &= \left(\int_U b(s,X_s^\pi,\mu_s,\alpha_s,u)\pi_s(u)du\right) ds + \left(\sqrt{\int_U \sigma^2(s,X_s^\pi,\mu_s,\alpha_s,u)\pi_s(u)du}\right) dW_s, \qquad (2.8)\\
X_t^\pi &= x,\ \ \mu_t = \nu,\ \ s\in[t,T].
\end{aligned}
$$

Then, it easily follows that

$$
\mathbb{E}_{\xi\sim\nu}[\widetilde{V}(t,\xi\,|\,\mu,\alpha)] = V(t\,|\,\mu,\alpha).
$$

While conceptually this approach is less general, it is frequently used - as in [25] and herein - to solve the MFG explicitly.

Nash Equilibrium (NE) for MFG with learning. To analyze game (2.2)-(2.3), we adopt the well-known NE criterion.

Definition 1 (NE for MFG)**.** *For game* (2.3) *with an initial state distribution ν and state process* (2.2)*, an agent-population profile $(\pi^*,\mu^*,\alpha^*) := \{(\pi_s^*,\mu_s^*,\alpha_s^*), t\le s\le T\}$ is called NE if the following conditions hold:*

- **A.** *(Single-agent-side) For the fixed population state-action distribution (μ^*,α^*) and any policy $\pi\in\mathcal{A}$,*
$$J(t,\pi\,|\,\mu^*,\alpha^*) \le J(t,\pi^*\,|\,\mu^*,\alpha^*).$$
- **B.** *(Population-side) $\pi_s^*(u;x) = \alpha_s^*(u;x)$, for all $x\in\mathbb{R}$. In addition, $\mathbb{P}_{X_s^*} = \mu_s^*$ for any $s\in[t,T]$, where X^* solves* (2.2) *when policy π^* is adopted with the initial population state distribution $\mu_t^* = \nu$.*

Given a NE (π^*,μ^*,α^*),

$$
V(t\,|\,\mu^*,\alpha^*) := J(t,\pi^*\,|\,\mu^*,\alpha^*) = \max_{\pi\in\mathcal{A}} J(t,\pi\,|\,\mu,\alpha^*)
$$

is called a game value associated with this NE.

Given (μ^*,α^*), condition **A** captures the optimality of π^* while condition **B** ensures the consistency of the solution so that the state and action flows of the single agent match those of the population. Note that uniqueness of NE for MFG is, in general, rare when mixed strategies are allowed (see, for example, [23]).

Solvability of MFG. There are three classical approaches to show the existence of MFG solution with pure strategies (or strict controls): the PDE (i.e., three-step fixed point) approach [26, 21], the probabilistic approach [8, 9] and the master equation approach [7]. The uniqueness of the MFG solution with pure strategies can be verified with certain technical conditions such as the small parameter conditions [26] or the monotonicity condition [21].

In the framework of MFG with relaxed controls, we follow the three-step fixed point approach to solve (2.2)-(2.3):

- **Step 1:** Fix a population state-action distribution (μ, α) and an initial state x. Then, solving the MFG (2.2)-(2.3) is reduced to solving a stochastic control problem with randomized policies (relaxed controls).

- **Step 2:** Let $X_s^{\pi,x}$ be the controlled state process under the optimal policy π from the initial state x in Step 1. Update $\alpha'_s(\cdot, y) = \pi_s(\cdot, y)$ for all $y \in \mathbb{R}$ and $s \in [t, T]$. Denote $X_s^{\pi,\xi}$ the controlled state process under π from some random initial state $\xi \sim \nu$. Then, update $\mu'_s = \mathbb{P}_{X_s^{\xi,\pi}}$.

- **Step 3:** Repeat Steps 1 and 2 until (μ', α') converges.

Note that there is no guarantee that the above procedure will yield any MFG solution since **Step 1** may have multiple solutions under relaxed controls. Moreover, by the nature of relaxed controls, the candidate fixed point(s) would be the fixed point(s) of a set-valued map as described in [23]. Nevertheless, for a family of linear-quadratic MFG, which will be introduced in Section 3, one can build proper verification arguments to show that the explicit fixed-point solution is indeed a solution to the MFG problem (2.2)-(2.3).

In general, the uniqueness of the MFG solution with relaxed controls does not hold unless there are additional convexity properties of the value function (see, for example, [24]). Here, the convexity in the linear-quadratic framework fails to hold when entropy regularization is included.

3 Shannon Entropy and Enhanced Entropy for MFG with Learning

In the mathematical formulation for MFG with learning of Section 2, we analyze the information theoretic gain for two types of entropies: Shannon entropy $\mathcal{H}_{SE}$ and enhanced entropy, which is a linear combination of Shannon entropy and cross-entropy $\lambda_{SE}\mathcal{H}_{SE} + \lambda_{CE}\mathcal{H}_{CE}$, with temperature parameters λ_{SE} and λ_{CE}. We study the impact of this entropy regularization within a class of LQ games in a finite time horizon. LQ games are the building blocks of stochastic games and often bring critical insights from their closed-form solutions ([4, 5]). Among others, we will see that the LQ games we analyze yield time-dependent optimal policies, with time-dependent Gaussian efficient explorations.

3.1 Game with Shannon Entropy

We start with the case of using only Shannon entropy for exploration, namely

$$\begin{aligned} V_{SE}(t \,|\, \mu) &:= \sup_{\pi \in \mathcal{A}} J_{SE}(t, \pi \,|\, \mu) \\ &:= \sup_{\pi \in \mathcal{A}} \mathbb{E}\left[\int_t^T \left(\int_{\mathbb{R}} -\frac{Q}{2}(X_s^\pi - m_s)^2 \pi_s(u) du + \lambda_{SE}\mathcal{H}_{SE}(\pi_s) \right) ds \; -\frac{\bar{Q}}{2}(X_T^\pi - m_T)^2 \middle| \, \mu \right], \end{aligned}$$

subject to

$$dX_s^\pi = \left(\int_{\mathbb{R}} (A(m_s - X_s^\pi) + Bu)\pi_s(u)du\right) ds + D\left(\sqrt{\int_{\mathbb{R}} u^2 \pi_s(u)du}\right) dW_s, \quad X_t^\pi = \xi \sim \nu. \tag{MFG-SE}$$

Here $\mu_t = \nu$, and $m_s = \int x\mu_s(dx)$ $(s \in [t,T])$. We assume $A > 0$, $Q > 0$, $\bar{Q} > 0$, and $\lambda_{SE} > 0$. We take the action space to be $U = \mathbb{R}$, and without loss of generality, $B > 0$ and $D > 0$.

We remark that $\alpha := \{\alpha_s\}_{s\in[t,T]}$ does not appear in the game formulation (**MFG-SE**). This is because when only Shannon entropy is incorporated, there is no interaction between the policy of the representative agent and the population action distribution α.

There are two types of rewards in this game: the running reward $-\frac{Q}{2}(X_s^\pi - m_s)^2$ that penalizes any deviation from the current average state of the population at time $s \in [t,T]$, and the terminal reward $-\frac{\bar{Q}}{2}(X_T^\pi - m_T)^2$ that penalizes deviation from the average state of the population at terminal time T. There are also two types of interaction: the real time interaction $A(m_s - X_s^\pi)$ and $\frac{Q}{2}(X_s^\pi - m_s)^2$ for $s \in [t,T]$, and the interaction at terminal time $\frac{\bar{Q}}{2}(X_T^\pi - m_T)^2$.

Next, we present one of the main results herein which provides an explicit NE solution for the MFG we consider. For notational convenience, We denote by $\mathcal{N}(\cdot|\nu,\sigma^2)$ the density function of a Gaussian random variable with mean ν and variance σ^2.

Theorem 2 (**MFG-SE**). *Let* $m^* = \mathbb{E}[\xi]$ *and*

$$\widetilde{V}_{SE}(t,x) = -\frac{\eta_t^{SE}}{2}(x - m^*)^2 + \gamma_t^{SE}, \tag{3.1}$$

with

$$\eta_t^{SE} = \bar{Q}\exp\left(-\left(2A + \frac{B^2}{D^2}\right)(T-t)\right) + \frac{Q}{2A + \frac{B^2}{D^2}}\left(1 - \exp\left(-\left(2A + \frac{B^2}{D^2}\right)(T-t)\right)\right) > 0, \tag{3.2}$$

and

$$\gamma_t^{SE} = \frac{\lambda_{SE}}{2}\ln\left(\frac{2\pi\lambda_{SE}}{D^2}\right)(T-t) - \int_t^T \frac{\lambda_{SE}}{2}\ln\left(\eta_z^{SE}\right)dz.$$

Then,

$$V_{SE}^*(t) := \mathbb{E}_{\xi\sim\nu}[\widetilde{V}_{SE}(t,\xi)]$$

is a game value of (**MFG-SE**) *associated with the NE policy*

$$\pi_s^{SE*}(u;x) = \mathcal{N}\left(u \,\middle|\, \frac{B(m^* - x)}{D^2}, \frac{\lambda_{SE}}{D^2\eta_s^{SE}}\right), \quad s \in [t,T]. \tag{3.3}$$

The corresponding controlled state process under (3.3) *is the unique solution of the SDE,*

$$\begin{aligned} dX_s^* &= \left(A + \frac{B^2}{D^2}\right)(m^* - X_s^*)ds + \sqrt{\frac{B^2(X_s^* - m^*)^2}{D^2} + \frac{\lambda_{SE}}{\eta_s^{SE}}}dW_s, \\ X_t^* &= \xi \sim \nu, s \in [t,T]. \end{aligned} \tag{3.4}$$

In addition, the mean state of the population under policy (3.3) *is time-independent, i.e.,*

$$m_s^* = \mathbb{E}[X_s^*] = m^*, \quad s \in [t,T]. \tag{3.5}$$

Remark 3. *Theorem 2 provides important guidance for exploration from an information-theoretic perspective. It suggests that, with Shannon entropy regularization, the associated optimal policy* $\pi_s^{SE*}(u;x)$ *from* (3.3) *is Gaussian, mean-reverting and with time-dependent variance. This is useful for MARL algorithmic design as the agent can now focus on a much smaller class of policies*

$$\widehat{\pi}_s(u;x) \sim \mathcal{N}\left(\widehat{M}(m_s - x), \widehat{\sigma}_s^2\right), \tag{3.6}$$

with $m_s = \int_{\mathbb{R}} x\mu_s(dx)$, $\widehat{M}$ *some scalar and* $\widehat{\sigma}^2 = \{\widehat{\sigma}_s^2\}_{s\in[t,T]}$ *a variance exploration process. Meanwhile, she can improve her estimate on* $\widehat{M}$ *and* $\widehat{\sigma}^2$ *of the above policy while interacting with the system and other agents, and observing the outcome at the end of each round of play. Indeed, notice that the controlled state process becomes*

$$dX_s^{\widehat{\pi}} = \left(A + B\widehat{M}\right)(m_s - X_s^{\widehat{\pi}})ds + \left(D\sqrt{\widehat{M}^2(X_s^{\widehat{\pi}} - m_s)^2 + \widehat{\sigma}_s^2}\right)dW_s, \tag{3.7}$$

with $X_t = \xi$. *Thus, the following simple corollary will be useful for MARL (see also more details in Section 4 where this result is used for algorithm design).*

Corollary 3.1. *If the representative agent follows policy* (3.6) *under a given mean field information* $\mu = \{\mu_s\}_{s\in[t,T]}$, *then the payoff is given by*

$$\begin{aligned} J_{SE}(t, \widehat{\pi} \,|\, \mu) &= -\frac{Q}{2}\int_t^T \left(\phi_s^2 - 2m_s\,\widehat{m}_s + m_s^2\right) ds \\ &\quad + \frac{\lambda_{SE}}{2}\int_t^T \ln\left(2\pi e \widehat{\sigma}_s^2\right) ds - \frac{\bar{Q}}{2}\left(\phi_T^2 - 2m_T\,\widehat{m}_T + m_T^2\right), \end{aligned} \tag{3.8}$$

where

$$\phi_s^2 = e^{\left(2\widehat{K} + D^2\widehat{M}^2\right)(s-t)}\left(\mathbb{E}[\xi^2] + \int_t^s e^{-D^2\widehat{M}^2(z-t)}d(z)dz\right),$$

with

$$d(s) = -2\mathbb{E}[\xi]\,e^{-\widehat{K}(s-t)}\widehat{K}m_s + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)e^{-\widehat{K}(s-t)}\widehat{K}m_s + e^{-2\widehat{K}(s-t)}D^2\left(\widehat{M}^2 m_s^2 - 2\widehat{M}^2 m_s\widehat{m}_s + \widehat{\sigma}_s^2\right),$$

$$m_s = \int_{\mathbb{R}} x\mu_s(dx), \quad \widehat{m}_s = e^{\widehat{K}(s-t)}\mathbb{E}[\xi] + \int_t^s e^{\widehat{K}(s-z)}\widehat{K}m_z\,dz, \quad \text{and} \;\; \widehat{K} = -\left(A + B\widehat{M}\right).$$

Next, we analyze the game with an additional cross-entropy regularization.

3.2 Game with Enhanced Entropy (Linear Combination of Shannon Entropy and Cross-entropy)

The objective of this game is to find

$$\begin{aligned} V_{EE}(t \,|\, \mu, \alpha) :=& \sup_{\pi\in\mathcal{A}} J_{EE}(t, \pi \,|\, \mu, \alpha) \\ :=& \sup_{\pi\in\mathcal{A}} \mathbb{E}\left[\int_t^T \left(-\frac{Q}{2}(X_s^\pi - m_s)^2 + \lambda_{SE}\mathcal{H}_{SE}(\pi_s) + \lambda_{CE}\mathcal{H}_{CE}(\pi_s, \alpha_s, \mu_s)\right) ds \right. \\ & \left. -\frac{\bar{Q}}{2}(X_T^\pi - m_T)^2 \right| \alpha, \mu\Bigg], \end{aligned}$$

subject to

$$dX_s = \left(\int_{\mathbb{R}} (A(m_s - X_s^\pi) + Bu)\pi_s(u)du\right) ds + D\sqrt{\int_{\mathbb{R}} u^2 \pi_s(u) du} dW_s, \quad X_t^\pi = \xi \sim \nu. \quad \textbf{(MFG-EE)}$$

Here $\mu_t = \nu$, $m_s = \int x\mu_s(dx)$, $s \in [t,T]$, and we assume $\lambda_{SE} > 0$, $\lambda_{CE} \geq 0$, $Q > 0$, $\bar{Q} > 0$, and $A > 0$. Without loss of generality, we also take $B > 0$ and $D > 0$.

Theorem 4 (**MFG-EE**). *Let* $m^* = \mathbb{E}[\xi]$ *and*

$$\widetilde{V}_{EE}(t,x) := -\frac{\eta_t^{EE}}{2}(x - m^*)^2 + \gamma_t^{EE}, \tag{3.9}$$

with

$$\begin{aligned}
\eta_t^{EE} \quad &= \quad \bar{Q} \exp\left(-\left(2A + \frac{B^2}{D^2}\frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\right)(T-t)\right) \\
&\quad + \frac{Q}{2A + \frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}}\left(1 - \exp\left(-\left(2A + \frac{B^2}{D^2}\frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\right)(T-t)\right)\right),
\end{aligned} \tag{3.10}$$

where $\eta_t^{EE} > 0$, *for* $t \in [0,T]$, *and*

$$\begin{aligned}
\gamma_t^{EE} \quad &= \quad \frac{\lambda_{SE} + \lambda_{CE}}{2} \ln\left(\frac{2\pi(\lambda_{SE} + \lambda_{CE})}{D^2}\right)(T-t) - \frac{\lambda_{SE} + \lambda_{CE}}{2}\int_t^T \ln(\eta_z^{EE})dz \\
&\quad + \frac{B^2}{2D^2}\frac{\lambda_{CE}(\lambda_{SE} + \lambda_{CE})}{\lambda_{SE}^2}\int_t^T \eta_z^{EE}\,\kappa_z^{EE} dz,
\end{aligned}$$

with

$$\kappa_s^{EE} = e^{(2K+M)(s-t)}\, Var[\xi] + \int_t^s e^{(M+2K)(s-z)}\frac{\lambda_{SE} + \lambda_{CE}}{\eta_z^{EE}} dz \tag{3.11}$$

and

$$K = -\left(A + \frac{B^2}{D^2}\right)\frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}, \quad M = \left(\frac{B}{D}\frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\right)^2.$$

Then,

$$V_{EE}^*(t) := \mathbb{E}_{\xi\sim\nu}\left[\widetilde{V}_{EE}(t, \xi \,|\, \mu^*, \alpha^*)\right]$$

is a game value of (**MFG-EE**), *with the associated NE policy*

$$\pi_s^{EE*}(u;x) = \mathcal{N}\left(\frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\frac{B(m^* - x)}{D^2}, \frac{\lambda_{SE} + \lambda_{CE}}{D^2 \eta_s^{EE}}\right). \tag{3.12}$$

Furthermore, the optimal controlled state process X_s^* *under policy* (3.12) *is the unique solution of the SDE,*

$$\begin{aligned}
dX_s^* \quad &= \left(A + \frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\frac{B^2}{D^2}\right)(m^* - X_s^*)ds \\
&\quad + D\sqrt{\left(A + \frac{\lambda_{SE} + \lambda_{CE}}{\lambda_{SE}}\frac{B^2}{D^2}\right)^2 (X_s^* - m)^2 + \frac{\lambda_{SE} + \lambda_{CE}}{D^2 \eta_s^{EE}}} dW_s, \qquad (3.13) \\
X_t^* \quad &= \xi \sim \nu, \;\; s \in [t,T].
\end{aligned}$$

In addition, $\mu_t^* = \mathbb{P}_{X_t^*}$, $\alpha_s^*(u;x) = \pi_s^{EE*}(u;x)$, *and the mean state of the population under policy* (3.12) *is time independent, i.e.,*

$$m_s^* = \mathbb{E}[X_s^*] = m^*, \quad s \in [t,T]. \tag{3.14}$$

Before providing the proof, a few remarks are in place.

3.3 Discussion.

In both linear-quadratic MFGs, with either only the Shannon entropy (**MFG-SE**) or with the additional cross-entropy (**MFG-EE**), there are several similarities.

- The form of the optimal policies (3.3) and (3.12) suggests that Gaussian exploration is optimal when entropy regularization is introduced in the MFG with learning. This is consistent with recent works of [34, 35] for continuous-time single-agent RL and is also supported by the empirical studies of [27] and [31].

- Both the means of the optimal policy $\pi_{SE}^*(u;x)$ in (3.3) and the optimal policy $\pi_{EE}^*(u;x)$ in (3.12) are influenced by both the mean field interaction and the current state of the representative agent. On the other hand, both their variances are time-dependent.

  In addition, the strength of their mean reversion is quantified by the coefficient $\frac{B}{D^2}$, which indicates that a smaller variance signifies less uncertainty in the game, hence a faster mean reverting policy.

- Equation (3.2) for η_s^{SE} and equation (3.10) for η_s^{CE} suggest that when time s is sufficiently small, the term $\frac{Q}{2A+\frac{B^2}{D^2}}\left(1-\exp\left(-\left(2A+\frac{B^2}{D^2}\right)(T-s)\right)\right)$ dominates η_s^{SE}, whereas η_s^{EE} is dominated by $\frac{Q}{2A+\frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}}\left(1-\exp\left(-\left(2A+\frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)(T-s)\right)\right)$. Thus, when time s is small, the cost of exploration is low and the representative agent has more incentive to explore in upcoming times.

  Conversely, when time s is sufficiently large and especially when $s \sim T$, η_s^{SE} is dominated by the term $\bar{Q}\exp\left(-\left(2A+\frac{B^2}{D^2}\right)(T-s)\right)$, whereas $\bar{Q}\exp\left(-\left(2A+\frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)(T-s)\right)$ dominates η_s^{EE}. Thus, the cost of exploration increases as time s approaches T. This implies that the agent is more sensitive to the terminal reward and explores less when the game approaches termination.

- In the very special case $A \equiv Q \equiv 0$, there is no intermediate payoff. Then, the variance of π_{SE}^* and π_{EE}^* decreases when time s increases, implying more exploration at the very beginning and less towards the very end.

Despite the above similarities, there is an important difference:

- The Shannon entropy and the cross-entropy affect the optimal policy π_{SE}^* and π_{EE}^* differently. Indeed, the mean of the optimal policy $\pi_{EE}^*(u;x)$ depends on the ratio between λ_{CE} and λ_{SE}, while λ_{SE} and λ_{CE} impact the variance of $\pi_{EE}^*(u;x)$ through both the $\frac{\lambda_{CE}}{\lambda_{SE}}$ and $\lambda_{SE}+\lambda_{CE}$ terms. In particular, with the additional cross-entropy, one will explore more and, consequently, the learning procedure would converge faster.

3.4 Derivations and Proofs of Main Results

The solution approach consists of two steps. The first is to find a candidate solution based on the classical fixed-point approach introduced in Section 2. The second is to verify the candidate solution via a verification theorem.

NE Derivation of (MFG-SE). To ease the exposition, we drop the subscript SE.

Proof. Proof of Theorem 2. For a given admissible policy $\pi \in \mathcal{A}$, the forward equation for $p(s,x)$, the density of X_s, is given by,

$$\partial_s p(s,x) = -\partial_x \left(\left(A(m_s - x) + B \int_{\mathbb{R}} u\, \pi_s(u;x) du \right) p(s,x) \right) + \frac{1}{2} \partial_{xx} \left(p(s,x) \int_{\mathbb{R}} D^2 u^2\, \pi_s(u;x) du \right),$$

with initial density $p(t,x) = \nu(x)$. Here, $m_s = \int x p(s,x) dx$, $s \in [t,T]$.

We first proceed heuristically with the associated HJB equation, derive a solution, and then validate this solution through a verification argument.

Step 1 (solving the control problem): Given fixed mean-field information $\{m_s\}_{s\in[0,T]}$ which is deterministic, the HJB equation for the value function $\widetilde{V}(s,x)$ can be written as

$$\begin{aligned} -\partial_s \widetilde{V}(s,x) = \quad & \max_{\pi_s \in \mathcal{P}(\mathbb{R})} \Bigg(\left(A(m_s - x) + B \int_{\mathbb{R}} u\pi(u;s,x) du \right) \widetilde{V}_x(s,x) \\ & - \frac{Q}{2}(m_s - x)^2 - \lambda_{SE} \int_{\mathbb{R}} \pi_s(u;x) \ln \pi_s(u;x) du + \frac{1}{2} \left(\int_{\mathbb{R}} D^2 u^2 \pi_s(u;x) du \right) \partial_{xx} \widetilde{V}(s,x) \Bigg), \end{aligned} \tag{3.15}$$

with terminal condition $\widetilde{V}(T,x) = -\frac{\overline{Q}}{2}(x - m_T)^2$. Recall that $\pi_s(u;x) \in \mathcal{P}(U)$ if and only if (2.1) holds. Solving the constrained maximization problem on the right hand side of (3.15) yields

$$\pi_s^*(u;x) = \frac{\exp\left(\frac{1}{\lambda_{SE}} \left(-\frac{Q}{2}(x - m_s)^2 + \frac{1}{2} D^2 u^2 \partial_{xx} \widetilde{V} + (A(m_s - x) + Bu) \widetilde{V}_x \right) \right)}{\int_{\mathbb{R}} \exp\left(\frac{1}{\lambda_{SE}} \left(-\frac{Q}{2}(x - m_s)^2 + \frac{1}{2} D^2 u^2 \partial_{xx} \widetilde{V} + (A(m_s - x) + Bu) \widetilde{V}_x \right) \right) du}.$$

Thus, the optimal policy is expected to be *Gaussian* with mean $\frac{B\partial_x \widetilde{V}}{-D^2 \partial_{xx} \widetilde{V}}$ and variance $\frac{\lambda_{SE}}{-D^2 \partial_{xx} \widetilde{V}}$, where it is for now assumed (and will be later verified) that $\partial_{xx} \widetilde{V} < 0$. Namely,

$$\pi_s^*(u;x) = \mathcal{N} \left(\frac{B\partial_x \widetilde{V}}{-D^2 \partial_{xx} \widetilde{V}}, \frac{\lambda_{SE}}{-D^2 \partial_{xx} \widetilde{V}} \right).$$

Therefore,

$$\int_{\mathbb{R}} u\pi_s^*(u;x) du = \frac{B\partial_x \widetilde{V}}{-D^2 \partial_{xx} \widetilde{V}} \quad \text{and} \quad \int_{\mathbb{R}} u^2 \pi_s^*(u;x) du = \left(\frac{B\partial_x \widetilde{V}}{-D^2 \partial_{xx} \widetilde{V}} \right)^2 + \frac{\lambda_{SE}}{-D^2 \partial_{xx} \widetilde{V}}.$$

Next, we introduce the ansatz

$$\widetilde{V}(s,x) = -\frac{\eta_s}{2}(x - m_s)^2 + \gamma_s, \tag{3.16}$$

for some $\eta_s > 0$ and γ_s to be appropriately defined. Then, $\partial_x \widetilde{V} = -\eta_s(x - m_s)$ and $\partial_{xx} \widetilde{V} = -\eta_s$, and thus,

$$\int_{\mathbb{R}} u\pi_s^*(u;x) du = \frac{B(m_s - x)}{D^2}$$

and

$$\int_{\mathbb{R}} u^2\pi_s^*(u;x)du = \frac{B^2(x-m_s)^2}{D^4} + \frac{\lambda_{SE}}{D^2\eta_s}.$$

Step 2 (updating the mean-field information): Denoting $\kappa = \frac{B}{D^2}$ and plugging in the forward equation for $p(s,x)$ yield

$$\begin{aligned}
\partial_s p(s,x) = \quad & -\partial_x\Big(\left((A+B\kappa)(m_s-x)\right) p(s,x)\Big) \\
& +\frac{1}{2}D^2\partial_{xx}\left(\left(\kappa^2(x-m_s)^2+\frac{\lambda_{SE}}{D^2\eta_s}\right)p(s,x)\right), \\
= \quad & (A+B\kappa)\, p(s,x) - \left((A+B\kappa)(m_s-x)\right)\partial_x p(s,x) \\
& +\frac{1}{2}D^2\left(2\kappa^2 p(s,x) + 4\kappa^2(x-m_s)\partial_x p(s,x)\right) \\
& +\frac{1}{2}D^2\left(\left(\kappa^2(x-m_s)^2+\frac{\lambda_{SE}}{D^2\eta_s}\right)\partial_{xx}p(s,x)\right).
\end{aligned} \tag{3.17}$$

Step 3 (finding a fixed-point): Multiplying both sides of (3.17) by x and integrating with respect to x yields that $dm_s = \left(\int x\partial_s p(s,dx)\right) ds = 0$. Therefore, $m_s^* = \mathbb{E}_{\xi\sim\nu}[\xi] =: m^*$.

Furthermore, the HJB equation for $s\in[t,T)$ is reduced to

$$\begin{aligned}
-\partial_s\widetilde{V}(s,x) = \quad & \max_{\pi_s\in\mathcal{P}(\mathbb{R})}\Bigg(\left(A(m^*-x)+B\int_{\mathbb{R}} u\pi_s(u;x)du\right)\partial_x\widetilde{V}(s,x) \\
& -\frac{Q}{2}(x-m^*)^2 - \lambda_{SE}\int_{\mathbb{R}}\pi_s(u;x)\ln\pi_s(u;x)du + \frac{1}{2}D^2\left(\int_{\mathbb{R}} u^2\pi_s(u;x)du\right)\partial_{xx}\widetilde{V}(s,x)\Bigg),
\end{aligned} \tag{3.18}$$

with $\widetilde{V}(T,x) = -\frac{\overline{Q}}{2}(m^*-x)^2$. Plugging $\pi_s^*(u;x)$ and using ansatz (3.16) with $m_s = m^*$ into the above HJB give

$$\begin{aligned}
\frac{\dot\eta_s}{2}(x-m^*)^2 - \dot\gamma_s \quad = \quad & -\left(A(m^*-x)+\frac{B^2(m^*-x)}{D^2}\right)\eta_s(x-m^*) \\
- \quad & \frac{Q}{2}(m^*-x)^2 + \lambda_{SE}\ln\left(\sqrt{\frac{2\pi e\lambda_{SE}}{D^2\eta_s}}\right) \\
- \quad & \frac{1}{2}D^2\left(\frac{B^2(x-m^*)^2}{D^4}+\frac{\lambda_{SE}}{D^2\eta_s}\right)\eta_s.
\end{aligned}$$

Direct calculations imply

$$\dot\eta_s = \left(2A+\frac{B^2}{D^2}\right)\eta_s - Q, \tag{3.19}$$

with $\eta_T = \bar{Q}$, and

$$\dot\gamma_s \quad = \quad -\frac{\lambda_{SE}}{2}\ln\left(\frac{2\pi\lambda_{SE}}{D^2}\right) + \frac{\lambda_{SE}}{2}\ln\eta_s \tag{3.20}$$

with $\gamma_T = 0$. Then, (3.19) admits the unique solution

$$\eta_s = \bar{Q} \exp\left(-\left(2A + \frac{B^2}{D^2}\right)(T-s)\right) + \frac{Q}{2A + \frac{B^2}{D^2}} \left(1 - \exp\left(-\left(2A + \frac{B^2}{D^2}\right)(T-s)\right)\right),$$

from which it is easy to verify that $\eta_s > 0$, since $A > 0, Q > 0$ and $\bar{Q} > 0$, $s \in [t,T]$. Moreover, (3.20) admits the unique solution

$$\gamma_s = \frac{\lambda_{SE}}{2} \ln\left(\frac{2\pi\lambda_{SE}}{D^2}\right)(T-s) - \int_s^T \frac{\lambda_{SE}}{2} \ln(\eta_z) dz.$$

Consequently, one NE (optimal) policy takes the form

$$\pi_s^*(u;x) = \mathcal{N}\left(\frac{B(m^* - x)}{D^2}, \frac{\lambda_{SE}}{D^2 \eta_s}\right),$$

and the associated optimal state process is the unique solution of the SDE (3.4).

Verification argument. The final step is to verify that m^* is the mean state under policy (3.3) and $V^*(t) := \mathbb{E}_{\xi\sim\nu}[\tilde{V}(\xi,t)] = \mathbb{E}_{\xi\sim\nu}[-\frac{\eta_t}{2}(\xi - m^*)^2 + \gamma_t]$ is the corresponding game value.

First, let us fix the mean field information as $m_s = m^*$, $s \in [t,T]$, and also fix the initial state $x \in \mathbb{R}$ and initial time $t \in [0,T]$. Let $\pi \in \mathcal{A}(x)$ and X^π be the associated state process under π solving

$$dX_s^\pi = \left(\int_{\mathbb{R}} (A(m^* - X_s^\pi) + Bu)\pi_s(u) du\right) ds + D\left(\sqrt{\int_{\mathbb{R}} u^2 \pi_s(u) du}\right) dW_s.$$

Denote $\tilde{r}(x,\pi) = -\frac{Q}{2}(x - m^*)^2$, $\tilde{b}(x,\pi) = \int_{\mathbb{R}}(A(m^* - x) + Bu)\pi_s(u)du$, and $\tilde{\sigma}(x,\pi) = D\left(\sqrt{\int_{\mathbb{R}} u^2 \pi(u) du}\right)$. Further, define the stopping time $\tau_n^\pi := \left\{ s \geq t : \int_t^T \partial_x \widetilde{V}(t, X_s^\pi)\tilde{\sigma}^2(X_s^\pi, \pi_s)^2 ds \geq n \right\}$, for $n \geq 1$. Then, Itô's formula yields

$$\begin{aligned} \widetilde{V}(T \wedge \tau_n^\pi, X_{T\wedge\tau_n^\pi}^\pi) - \widetilde{V}(t,x) &= \int_t^{T\wedge\tau_n^\pi} \left(\frac{1}{2}\partial_{xx}\widetilde{V}(s, X_s^\pi)\tilde{\sigma}^2(X_s^\pi,\pi_s) + \partial_x \widetilde{V}(s, X_s^\pi)\tilde{b}(X_s^\pi, \pi_s)\right) ds \\ &\quad + \int_t^{T\wedge\tau_n^\pi} \partial_x \widetilde{V}(s, X_s^\pi) dW_s. \end{aligned}$$

Taking expectations, using that $\widetilde{V}$ solves the HJB equation (3.18), and that π is in general sub-optimal, we deduce that

$$\begin{aligned} &\mathbb{E}\left[\widetilde{V}(T, X_{T\wedge\tau_n^\pi}^\pi)\right] \\ =\ & \widetilde{V}(t,x) + \mathbb{E}\left[\int_t^{T\wedge\tau_n^\pi} \left(\frac{1}{2}\partial_{xx}\widetilde{V}(s, X_s^\pi)\tilde{\sigma}^2(X_s^\pi,\pi_s) + \partial_x\widetilde{V}(s, X_s^\pi)\tilde{b}(X_s^\pi,\pi_s)\right) ds + \int_t^{T\wedge\tau_n^\pi} \partial_x \widetilde{V}(s, X_s^\pi) dW_s\right] \\ \leq\ & \widetilde{V}(t,x) - \mathbb{E}\left[\int_t^{T\wedge\tau_n^\pi} \left(\tilde{r}(X_s^\pi, \pi_s) - \lambda \int_{\mathbb{R}} \pi_s(u) \ln \pi_s(u) du\right) ds\right]. \end{aligned}$$

Standard calculations yield that $\mathbb{E}[\sup_{t\leq s\leq T} |X_s^\pi|^2] \leq N(1 + x^2)e^{NT}$ for some constant $N > 0$, which is independent of n. Sending $n \to \infty$ yields

$$\widetilde{V}(t,x) \geq \mathbb{E}\left[\int_t^T \left(\tilde{r}(X_s^\pi, \pi_s) - \lambda \int_{\mathbb{R}} \pi_s(u)\ln \pi_s(u) du\right) ds - \frac{\bar{Q}}{2}(X_T^\pi - m^*)^2\right],$$

for each $x \in \mathbb{R}$ and $\pi \in \mathcal{A}$. Hence, $\widetilde{V}(t,x) \geq V^*(t,x)$, for all $x \in \mathbb{R}$.

On the other hand, the right-hand of (3.15) is maximized for

$$\pi_s^*(u;x) = \mathcal{N}\left(\frac{B(m^*-x)}{D^2}, \frac{\lambda_{SE}}{D^2\eta_s}\right). \tag{3.21}$$

Thus,

$$\widetilde{V}(t,x) = \mathbb{E}\left[\int_0^T \left(\tilde{r}(X_s^*,\pi_s) - \lambda \int_{\mathbb{R}} \pi_s^*(u;X_s^*) \ln \pi_s^*(u;X_s^*)du\right) ds - \frac{\bar{Q}}{2}(X_T^* - m^*)^2\right],$$

where X_s^* is the controlled state process under policy (3.21).

Next, let us show that for $s \in [t,T]$,

$$m^* = \mathbb{E}[X_s^*].$$

To this end, let $K = -\left(A + \frac{B^2}{D^2}\right)$. Then,

$$dX_s^* = (KX_s^* - Km^*)ds + f(s, X_s^*, m^*)dW_s,$$

with

$$f(s,x,m) = \left(\sqrt{\left(\frac{B}{D}(x-m)\right)^2 + \frac{\lambda_{SE}}{\eta_s}}\right).$$

Therefore,

$$e^{-K(s-t)}X_s^* \quad = \quad \xi + \int_t^s e^{-K(z-t)}\left(-Km^*dz + f(z,X_z^*,m^*)dW_z\right),$$

and $e^{-K(s-t)}\mathbb{E}[X_s^*] = \mathbb{E}[\xi] + \left(e^{-K(s-t)} - 1\right)m^*$. Hence, $\mathbb{E}[X_s^*] = m^*$, $s \in [t,T]$.

□

NE Derivation of Game (MFG-EE). To ease the exposition, we drop the subscript EE.

Proof. Proof of Theorem 4 For a given Markovian policy $\pi_s(u;x)$, the forward equation for $p(s,x)$, the density of X_s, $s \in [t,T]$ satisfies

$$\partial_s p(s,x) = -\partial_x\left(\left(A(m_s - x) + B\int_{\mathbb{R}} u\pi_s(u;x)du\right)p(s,x)\right) + \frac{1}{2}\partial_{xx}\left(p(s,x)\int_{\mathbb{R}} D^2u^2\pi_s(u;x)du\right),$$

with initial density $p(t,x) = \nu(x)$ and $m_s = \int xp(s,x)dx$.

Step 1 (solving the control problem): Given fixed mean-field information $\{m_s\}_{s\in[0,T]}$, the HJB equation for the value function $\widetilde{V}(s,x,m)$ can be written as

$$\begin{aligned}
-\partial_s\widetilde{V}(s,x) = \quad & \max_{\pi_s \in \mathcal{P}(\mathbb{R})}\Bigg(\left(A(m_s - x) + B\int_{\mathbb{R}} ud\pi_s(u;x)\right)\partial_x\widetilde{V}(s,x) - \frac{Q}{2}(m_s - x)^2 \\
& -\lambda_{CE}\int_{\mathbb{R}} \pi_s(u;x)\int \ln \alpha_s(u;x)\mu_s(dx)du - \lambda_{SE}\int_{\mathbb{R}} \pi_s(u;x)\ln \pi_s(u;x)du \\
& +\frac{1}{2}\left(\int_{\mathbb{R}} D^2u^2\pi_s(u;x)du\right)\partial_{xx}\widetilde{V}(s,x)\Bigg),
\end{aligned}$$

with $\widetilde{V}(T,x) = -\frac{\bar{Q}}{2}(x-m_T)^2$. Recall that $\pi_s \in \mathcal{P}(U)$ if and only if (2.1) holds. The constrained maximization problem on the right hand side of (3.15) yields

$$\pi_s^*(u;x) = \frac{\exp\left(\frac{1}{\lambda_{SE}}\left(-\frac{Q}{2}(x-m_s)^2 + \frac{1}{2}D^2u^2\partial_{xx}\widetilde{V} + (A(m_s-x)+Bu)\partial_x\widetilde{V} - \lambda_{CE}\int\Big(\ln\alpha_s(u;x)\Big)\mu_s(dx)\right)\right)}{\int_{\mathbb{R}}\exp\left(\frac{1}{\lambda_{SE}}\left(-\frac{Q}{2}(x-m_s)^2 + \frac{1}{2}D^2u^2\partial_{xx}\widetilde{V} + (A(m_s-x)+Bu)\partial_x\widetilde{V} - \lambda_{CE}\int\Big(\ln\alpha_s(u;x)\Big)\mu_s(dx)\right)\right)du}.$$

Next, we introduce the ansatz for the population action distribution for the agent in state y,

$$\alpha_s(u;y) = \mathcal{N}(u\,|\,H_s(y-m_s), L_s), \tag{3.22}$$

with some (to be defined) deterministic processes H_s and $L_s > 0$, $s \in [t,T]$. Then, $\alpha_s(u;y)$ is Gaussian with mean $H_s(y-m_s)$ and variance L_s. We stress that the Gaussian property of $\alpha_s(u;y)$ does not imply the Gaussian property of the aggregated population action distribution $\widetilde{\alpha}(s) = \int \alpha_s(u;y)\mu_s(dy)$.

In turn, $\ln\alpha_s(u;y) = -\frac{1}{2}\ln(2\pi L_s) - \frac{1}{2L_s}(u-H_s(y-m_s))^2$ and

$$\begin{aligned}\int \mu_s(dy)\ln(\alpha_s(u;y)) &= -\frac{1}{2}\ln(2\pi L_s) - \frac{1}{2L_s}\int (u-H_s(y-m_s))^2\mu_t(dy)\\ &= -\frac{1}{2}\ln(2\pi L_s) - \frac{1}{2L_s}(u^2 + H_s^2\mathrm{Var}(\mu_s)),\end{aligned}$$

with $\mathrm{Var}(\mu_s) = \int x^2\mu_s(dx) - (\int x\mu_s(dx))^2$. Therefore, the optimal policy is *Gaussian* with mean $\frac{B\partial_x\widetilde{V}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}$ and variance $\frac{\lambda_{SE}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}$, namely,

$$\pi_s^*(u;x) = \mathcal{N}\left(u\;\middle|\;\frac{B\partial_x\widetilde{V}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}, \frac{\lambda_{SE}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}\right).$$

Let us for now assume (and will verify later) that $-D^2\partial_{xx}\widetilde{V} - \frac{\lambda_{CE}}{L_s} > 0$. In turn,

$$\int_{\mathbb{R}} u\pi_s^*(u;x)du = \frac{B\partial_x\widetilde{V}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}$$

and

$$\int_{\mathbb{R}} u^2\pi_s^*(u;x)du = \left(\frac{B\partial_x\widetilde{V}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}\right)^2 + \frac{\lambda_{SE}}{-D^2\partial_{xx}\widetilde{V}-\frac{\lambda_{CE}}{L_s}}.$$

Next, consider the ansatz

$$\widetilde{V}(s,x) = -\frac{\eta_s}{2}(x-m_s)^2 + \gamma_s. \tag{3.23}$$

In turn, $\partial_x\widetilde{V} = -\eta_s(x-m_s)$ and $\partial_{xx}\widetilde{V} = -\eta_s$, together with

$$\int_{\mathbb{R}} u\pi_s^*(u;x)du = \frac{B\eta_s(m_s-x)}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}},$$

and

$$\int_{\mathbb{R}} u^2\pi_s^*(u;x)du = \left(\frac{B\eta_s(m_s-x)}{D^2\eta_s-\frac{\lambda_{CE}}{L_s}}\right)^2 + \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}.$$

Step 2 (updating the mean-field information): Denoting $\kappa_s := \frac{B\eta_s}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}$, $s \in [t,T]$, and plugging in the forward equation for $p(s,x)$, we deduce that

$$
\begin{aligned}
\partial_s p(s,x) = \quad & -\partial_x \left((A+B\kappa_s)(m_s - x)p(s,x)\right) \\
& + \frac{1}{2}D^2 \partial_{xx}\left(\left(\kappa_t^2 (x-m_s)^2 + \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)p(s,x)\right) \\
= \quad & (A+B\kappa_s)\,p(s,x) - \left((A+B\kappa_s)(m_s-x)\right)\partial_x p(s,x) \\
& + \frac{1}{2}D^2\left(2\kappa_s^2 p(s,x) + 4\kappa_s^2(x-m_s)\partial_x p(s,x)\right) \\
& + \frac{1}{2}D^2\left(\left(\kappa_s^2(x-m_s)^2 + \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)\partial_{xx}p(s,x)\right). \qquad (3.24)
\end{aligned}
$$

Multiplying both sides of (3.24) by x and integrating with respect to x give $\partial_s m_s = \int x\partial_s p(s,dx) = 0$ and, thus, $m_s^* = m^* = \mathbb{E}_{\xi\sim\nu}[\xi]$, for $s \in [t,T]$. Hence, the HJB equation reduces to

$$
\begin{aligned}
-\partial_s \widetilde{V}(s,x) = \quad & \max_{\pi_s \in \mathcal{P}(\mathbb{R})}\Big(\left(A(m^*-x) + B\textstyle\int_{\mathbb{R}} u d\pi_s(u)\right)\partial_x \widetilde{V}(s,x) - \frac{Q}{2}(m^*-x)^2 \\
& -\lambda_{SE}\textstyle\int_{\mathbb{R}}\pi_s(u)\ln\pi_s(u)du - \lambda_{CE}\int_{\mathbb{R}}\pi_s(u)\int\mu_s(dy)\ln\alpha_s(u;y)dydu \\
& + \tfrac{1}{2}D^2\textstyle\int_{\mathbb{R}} u^2\pi_s(u)du\partial_{xx}\widetilde{V}(s,x)\Big).
\end{aligned}
$$

Plugging $\pi_s^*(u;x)$ and using ansatz (3.23) with $m_s = m^*$ for the above HJB, we obtain

$$
\begin{aligned}
\frac{\dot{\eta_s}}{2}(x-m^*)^2 - \dot{\gamma}_s = \quad & - \quad \left(A(m^*-x) + B\frac{B\eta_s(m^*-x)}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)\eta_s(x-m^*) + \lambda_{SE}\ln\left(\sqrt{\frac{2\pi e\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}}\right) \\
& - \quad \frac{1}{2}D^2\left(\left(\frac{(B\eta_s)(x-m^*)}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)^2 + \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)\eta_s - \frac{Q}{2}(x-m^*)^2 \\
& + \quad \frac{\lambda_{CE}}{2}\ln(2\pi L_s) + \frac{\lambda_{CE}}{2L_s}\left(\left(\frac{B\eta_s(x-m^*)}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right)^2 + \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}\right) \\
& + \quad \frac{\lambda_{CE}}{2L_s}H_s^2\mathrm{Var}(\mu_s).
\end{aligned}
$$

Direct calculations yield

$$
\dot{\eta}_s = 2A\eta_s + \frac{(B\eta_s)^2}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}} - Q, \qquad (3.25)
$$

$$
\dot{\gamma}_s = \frac{\lambda_{SE}}{2} - \lambda_{SE}\ln\left(\sqrt{\frac{2\pi e\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}}}\right) - \frac{\lambda_{CE}}{2}\ln(2\pi L_s) - \frac{\lambda_{CE}}{2L_s}H_s^2\mathrm{Var}(\mu_s). \qquad (3.26)
$$

Step 3 (finding a fixed-point): Setting

$$L_s = \frac{\lambda_{SE}}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}} \quad \textit{and} \quad H_s = \frac{-B\eta_s}{D^2\eta_s - \frac{\lambda_{CE}}{L_s}},$$

we deduce that

$$H_s = -\frac{B}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}} \quad \textit{and} \quad L_s = \frac{\lambda_{SE}+\lambda_{CE}}{D^2\eta_s},$$

with

$$\dot{\eta}_s \quad = \quad 2A\eta_s + \frac{B^2\eta_s}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}} - Q, \tag{3.27}$$

$$\dot{\gamma}_s \quad = \quad -\frac{\lambda_{SE}+\lambda_{CE}}{2}\ln\left(\frac{2\pi(\lambda_{SE}+\lambda_{CE})}{D^2\eta_s}\right) - \frac{\lambda_{CE}}{2}\frac{B^2\eta_s(\lambda_{SE}+\lambda_{CE})}{D^2\lambda_{SE}^2}\mathrm{Var}(\mu_s). \tag{3.28}$$

Consequently,

$$\pi_s^*(u;x) = \mathcal{N}\left(u \;\middle|\; \frac{B}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}(m^*-x), \frac{\lambda_{SE}+\lambda_{CE}}{D^2\eta_s}\right).$$

Denote $-K = \left(A + B\frac{B}{D^2}\right)\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}$ and

$$f(s,x,m) = \left(\sqrt{\left(\frac{B}{D}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}(x-m)\right)^2 + \frac{\lambda_{SE}+\lambda_{CE}}{\eta_s}}\right).$$

In turn,

$$dX_s^* = K(X_s^* - m^*)ds + f(s, X_s^*, m^*)dW_s,$$

and

$$\begin{aligned} d(e^{-K(s-t)}X_s^*) &= -Ke^{-K(s-t)}X_s^*ds + e^{-K(s-t)}dX_s^* \\ &= -Ke^{-K(s-t)}X_s^*ds + e^{-K(s-t)}((KX_s^* - Km^*)ds + f(s,X_s^*,m^*)dW_s) \\ &= e^{-K(s-t)}(-Km^*ds + f(s,X_s^*,m^*)dW_s). \end{aligned}$$

Therefore,

$$e^{-K(s-t)}X_s^* \quad = \quad \xi + \int_t^s e^{-K(z-t)}(-Km^*dz + f(z,X_z^*,m^*)dW_z), \tag{3.29}$$

and

$$e^{-2K(s-t)}\mathrm{Var}[X_s^*] = \mathrm{Var}[\xi] + \mathbb{E}\left[\left(\int_t^s e^{-K(z-t)}f(z,X_z^*,m^*)dW_z\right)^2\right].$$

By Itô's isometry,

$$\begin{aligned} &\mathbb{E}\left[\left(\int_t^s e^{-K(z-t)}f(z,X_z^*,m^*)dW_z\right)^2\right] = \mathbb{E}\left[\int_t^s e^{-2K(z-t)}f^2(z,X_z^*,m^*)dz\right] \\ &= \mathbb{E}\left[\int_t^s e^{-2K(z-t)}\left(\left(\frac{B}{D}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}(X_z^*-m^*)\right)^2 + \frac{\lambda_{SE}+\lambda_{CE}}{\eta_z}\right)dz\right] \\ &= \int_t^s e^{-2K(z-t)}\left(\left(\frac{B}{D}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)^2\mathrm{Var}[X_z^*] + \frac{\lambda_{SE}+\lambda_{CE}}{\eta_z}\right)dz. \end{aligned}$$

Let

$$y(s) = e^{-2K(s-t)}\text{Var}[X_s^*],$$

$$M = \left(\frac{B}{D}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)^2 \quad \text{and} \quad b(s) = e^{-2K(s-t)}\frac{\lambda_{SE}+\lambda_{CE}}{\eta_s}.$$

Thus,

$$y(s) = e^{M(s-t)}\left(y(t) + \int_t^s e^{-M(z-t)}b(z)dz\right),$$

and

$$e^{-2K(s-t)}\text{Var}(X_s^*) = e^{M(s-t)}\left(\text{Var}(\xi) + \int_t^s e^{-(M+2K)(z-t)}\frac{\lambda_{SE}+\lambda_{CE}}{\eta_z}dz\right).$$

Therefore,

$$\begin{aligned}\text{Var}[X_s^*] &= e^{(2K+M)(s-t)}\text{Var}[\xi] + e^{2K(s-t)}\int_t^s e^{M(s-z)}b(z)dz \\ &= e^{(2K+M)(s-t)}\text{Var}[\xi] + \int_t^s e^{(M+2K)(s-z)}\frac{\lambda_{SE}+\lambda_{CE}}{\eta_z}dz. \end{aligned} \tag{3.30}$$

Assume for the moment that $\eta_s > 0$ for $s \in [t,T]$. Then, $\text{Var}[X_s^*]$ is well-defined. Hence (3.28) reduces to

$$\begin{aligned}\dot{\gamma}_s &= -\frac{\lambda_{SE}+\lambda_{CE}}{2}\ln\frac{2\pi(\lambda_{SE}+\lambda_{CE})}{D^2} \\ &\quad +\frac{\lambda_{SE}+\lambda_{CE}}{2}\ln\eta_s - \frac{\lambda_{CE}}{2}\frac{B^2\eta_s(\lambda_{SE}+\lambda_{CE})}{D^2\lambda_{SE}^2}\kappa_s,\end{aligned} \tag{3.31}$$

with $\gamma_T = 0$, where κ_s, $s \in [t,T]$,

$$\kappa_s := e^{(M+2K)(s-t)}\text{Var}[\xi] + \int_t^s e^{(M+2K)(s-z)}\frac{\lambda_{SE}+\lambda_{CE}}{\eta_z}dz \tag{3.32}$$

with

$$K = -\left(A + B\frac{B}{D^2}\right)\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}} \quad \text{and} \quad M = D^2\left(\frac{B}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)^2.$$

Therefore, equation (3.27) admits the unique solution

$$\begin{aligned}\eta_s &= \bar{Q}\exp\left(-\left(2A + \frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)(T-s)\right) \\ &\quad + \frac{Q}{2A + \frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}}\left(1 - \exp\left(-\left(2A + \frac{B^2}{D^2}\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\right)(T-s)\right)\right).\end{aligned}$$

We easily deduce that $\eta_s > 0$, $s \in [t,T]$, since $\bar{Q} > 0, Q > 0$ and $A > 0$.

Moreover, equation (3.31) admits the (unique) solution

$$\begin{aligned}\gamma_s &= \frac{\lambda_{SE}+\lambda_{CE}}{2}(T-s)\ln\frac{2\pi(\lambda_{SE}+\lambda_{CE})}{D^2} - \frac{\lambda_{SE}+\lambda_{CE}}{2}\int_s^T \ln\,\eta_z\,dz \\ &\quad + \int_s^T \frac{\lambda_{CE}}{2}\frac{B^2\eta_z(\lambda_{SE}+\lambda_{CE})}{D^2\lambda_{SE}^2}\kappa_z dz.\end{aligned}$$

We then obtain that the associated optimal feedback policy is given by

$$\pi_s^*(u;x) = \mathcal{N}\left(\frac{\lambda_{SE}+\lambda_{CE}}{\lambda_{SE}}\frac{B(m^*-x)}{D^2}, \frac{\lambda_{SE}+\lambda_{CE}}{D^2\eta_s}\right), \tag{3.33}$$

and the optimal controlled state process is the unique solution of the SDE (3.13). The verification is similar to the verification of Theorem 2 and is therefore omitted. □

Proof of Corollary 3.1

Proof. Let $\widehat{K} := -\left(A + B\widehat{M}\right)$ and $f\left(m_s, X_s, \widehat{M}, \widehat{\sigma}_s^2\right) := D\sqrt{\widehat{M}^2(m_s - X_s)^2 + \widehat{\sigma}_s^2}$, $s \in [t,T]$. Then, under policy $\widehat{\pi}$ given in (3.6),

$$\begin{aligned} dX_s^{\widehat{\pi}} &= \left(\left(A + B\widehat{M}\right)m_s - \left(A + B\widehat{M}\right)X_s^{\widehat{\pi}}\right)ds + f\left(m_s, X_s^{\widehat{\pi}}, \widehat{M}, \widehat{\sigma}_s^2\right)dW_s \\ &= -\widehat{K}m_s ds + KX_s^{\widehat{\pi}}ds + fdW_s. \end{aligned}$$

Using that $d(e^{-\widehat{K}s}X_s^{\widehat{\pi}}) = e^{-\widehat{K}s}\left(-\widehat{K}m_s ds + f\,dW_s\right)$, we have

$$e^{-\widehat{K}(s-t)}X_s^{\widehat{\pi}} = \xi + \int_t^s e^{-\widehat{K}(z-t)}(-\widehat{K}m_z dz + f\,dW_z). \tag{3.34}$$

Hence,

$$\mathbb{E}[X_s^{\widehat{\pi}}] = e^{\widehat{K}(s-t)}\mathbb{E}[\xi] - \int_t^s e^{\widehat{K}(s-z)}\widehat{K}m_z\,dz.$$

Let $\widehat{m}_s := \mathbb{E}[X_s^{\widehat{\pi}}]$. From (3.34) and routine calculations, we deduce that

$$\begin{aligned} e^{-2\widehat{K}(s-t)}\left(X_s^{\widehat{\pi}}\right)^2 &= \xi^2 + 2\xi\int_t^s e^{-\widehat{K}(z-t)}(-\widehat{K}m_z dz + f\,dW_z) + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)^2 \\ &\quad + \left(\int_t^s e^{-\widehat{K}(z-t)}f\,dW_z\right)^2 - 2\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\int_t^s e^{-\widehat{K}(z-t)}f\,dW_z. \end{aligned}$$

Hence,

$$\begin{aligned} &e^{-2\widehat{K}(s-t)}\mathbb{E}\left[\left(X_s^{\widehat{\pi}}\right)^2\right] \\ &= \mathbb{E}[\xi^2] - 2\mathbb{E}[\xi]\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)^2 + \mathbb{E}\left(\int_t^s e^{-\widehat{K}(z-t)}f\,dW_z\right)^2 \end{aligned} \tag{3.35}$$

By Itô's isometry,

$$\begin{aligned} &\mathbb{E}\left(\int_t^s e^{-\widehat{K}(z-t)}f\,dW_z\right)^2 = \mathbb{E}\left[\int_t^s e^{-2\widehat{K}(z-t)}f^2\,dz\right] \\ &= \mathbb{E}\left[\int_t^s e^{-2\widehat{K}(z-t)}D^2\left(\widehat{M}^2\left(X_z^{\widehat{\pi}} - m_z\right)^2 + \widehat{\sigma}_z^2\right)dz\right] \\ &= \mathbb{E}\left[\int_t^s e^{-2\widehat{K}(z-t)}D^2\widehat{M}^2\left(m_z^2 - 2m_zX_z^{\widehat{\pi}} + \left(X_z^{\widehat{\pi}}\right)^2\right)dz + \int_t^s e^{-2K(z-t)}D^2\widehat{\sigma}_z^2 dz\right] \\ &= \mathbb{E}\left[\int_t^s e^{-2\widehat{K}(z-t)}D^2\widehat{M}^2\left(X_z^{\widehat{\pi}}\right)^2 dz\right] + \int_t^s e^{-2\widehat{K}(z-t)}D^2\left(\widehat{M}^2m_z^2 - 2\widehat{M}^2m_z\widehat{m}_z + \widehat{\sigma}_z^2\right)dz \\ &= D^2\widehat{M}^2\int_t^s e^{-2\widehat{K}(z-t)}\mathbb{E}\left[\left(X_z^{\widehat{\pi}}\right)^2\right]dz + \int_t^s e^{-2\widehat{K}(z-t)}D^2\left(\widehat{M}^2m_z^2 - 2\widehat{M}^2m_z\widehat{m}_z + \widehat{\sigma}_z^2\right)dz \end{aligned} \tag{3.36}$$

Combining (3.35) and (3.36) yields

$$
\begin{aligned}
e^{-2K(s-t)}\mathbb{E}\left[\left(X_s^{\widehat{\pi}}\right)^2\right] &= \mathbb{E}[\xi^2] - 2\mathbb{E}[\xi]\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)^2 \\
&\quad + D^2\widehat{M}^2\int_t^s e^{-2\widehat{K}(z-t)}\mathbb{E}\left[\left(X_z^{\widehat{\pi}}\right)^2\right]dz + \int_t^s e^{-2\widehat{K}(z-t)}D^2\left(\widehat{M}^2 m_z^2 - 2\widehat{M}^2 m_z\widehat{m}_z + \widehat{\sigma}_z^2\right)dz.
\end{aligned}
$$

Letting $y_s := e^{-2\widehat{K}(s-t)}\mathbb{E}\left[\left(X_s^{\widehat{\pi}}\right)^2\right]$ for $s\in[t,T]$, we have

$$
y_s - y_t = b_s + D^2\widehat{M}^2\int_t^s y_z dz,
$$

where,

$$
b_s := -2\mathbb{E}[\xi]\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)^2 + \int_t^s e^{-2\widehat{K}(z-t)}D^2\left(\widehat{M}^2 m_z^2 - 2\widehat{M}^2 m_z\widehat{m}_z + \widehat{\sigma}_z^2\right)dz,
$$

with $b_t = 0$. Therefore,

$$
\int_t^s \dot{y}_z dz = \int_t^s \dot{b}_z dz + D^2\widehat{M}^2\int_t^s y_z dz
$$

and, thus,

$$
y_s = e^{D^2\widehat{M}^2(s-t)}\left(y_t + \int_t^s e^{-D^2\widehat{M}^2(z-t)}\dot{b}(z)dz\right).
$$

Finally, for $s\in[t,T]$,

$$
\mathbb{E}\left[\left(X_s^{\widehat{\pi}}\right)^2\right] = e^{\left(2\widehat{K}+D^2\widehat{M}^2\right)(s-t)}\left(\mathbb{E}[\xi^2] + \int_t^s e^{-D^2\widehat{M}^2(z-t)}\dot{b}(z)dz\right)
$$

with

$$
\begin{aligned}
\dot{b}(s) &= -2\mathbb{E}[\xi]\,e^{-\widehat{K}(s-t)}\widehat{K}m_s + \left(\int_t^s e^{-\widehat{K}(z-t)}\widehat{K}m_z dz\right)e^{-\widehat{K}(s-t)}\widehat{K}m_s \\
&\quad + e^{-2\widehat{K}(s-t)}D^2\left(\widehat{M}^2 m_s^2 - 2\widehat{M}^2 m_s\widehat{m}_s + \widehat{\sigma}_s^2\right).
\end{aligned}
$$

The rest of the proof follows easily. □

4 Experiment

We now demonstrate how the theoretical results of Theorems 2 and 4 can be used to design algorithms for MFG with learning. The experiment aims to highlight

- how entropy regularization helps to "explore optimally" in a game with learning, and especially in improving the speed of convergence to the NE, and
- how the agent manages to eventually learn the optimal scheduling of the exploration and, in particular, the time-dependent variances (as in (3.3) and (3.12)) over a finite time horizon.

Throughout this section, the experiment is with the inclusion of Shannon entropy only, as the case with the additional cross-entropy may be studied in a similar fashion.

4.1 Set-up

The algorithm design is with discrete time steps $s = 0, 1, 2, \cdots, N$, where $\delta = \frac{T}{N}$ is the step-size. According to Theorem 2 and Corollary 3.1, it suffices to focus on a considerably smaller class of policies of form

$$\widehat{\pi}_s \sim \mathcal{N}\left(\widehat{M}(x_s - m_s), \widehat{\sigma}_s^2\right),$$

which can be fully characterized by the mean state process $m := \{m_s\}_{s=0}^N$ and $\widehat{R} := (\widehat{M}, \widehat{\sigma}^2)$, with $\widehat{\sigma}^2 := \{\widehat{\sigma}_s^2\}_{s=0}^N$. We, then, consider the discrete-time LQ-MFG problem

$$J\left(\widehat{R}, m\right) := \mathbb{E}\left[\sum_{s=0}^{N-1}\left(-\frac{Q}{2}(x_s - m_s)^2 + \lambda_{SE}\mathcal{H}_{SE}(\pi_s)\right)\delta - \frac{\bar{Q}}{2}(X_N - m_N)^2\right], \tag{4.1}$$

where, for $s = 0, 1, \cdots, N-1$,

$$x_{s+1} = x_s + \left(\int_{\mathbb{R}}(A(m_s - X_s) + Bu)\pi_s(u)du\right)\delta + \left(D\sqrt{\int_{\mathbb{R}} u^2\pi_s(u)du}\right)\Delta W_s,\ x_0 = \xi \sim \nu. \tag{4.2}$$

Here, ΔW_s are i.i.d $\mathcal{N}(0, \delta)$ random variables and ν is the distribution of the initial state ξ.

4.2 Mean Field Policy Gradient with Exploration

Recall that in the learning setting, the model parameters A, B, D, Q, and $\bar{Q}$ are assumed to be unknown to the agent. She only has access to the *simulated* reward function

$$\widehat{J}\left(\widehat{R}, m\right) := \sum_{s=0}^{N-1}\left((x_s - m_s)^2 + \lambda_{SE}\mathcal{H}_{SE}(\pi_s)\right)\delta - \frac{Q}{2}(X_N - m_N)^2,$$

which is associated with a *single* trajectory $\{x_s\}_{s=0}^N$ under the policy characterized by $\widehat{R}$ and the mean state process $m = \{m_s\}_{s=0}^N$. Note, however, that this assumption is weaker than being able to observe $J\left(\widehat{R}, m\right)$ defined in (4.1), as $J\left(\widehat{R}, m\right)$ involves calculating an expectation and requiring observing infinite number of samples.

The algorithm has the following key elements:

- *Information adaptiveness.* In each outer iteration k, $k = 1, 2, \cdots, K$, the representative agent can improve her decision (lines 4-11) based on the mean field information m^{k-1} from the previous outer iteration $k-1$. This implies that she has access to a simulator with which she may exercise different policies when other agents keep applying the same policy from the previous outer iteration. This is a standard assumption, see, for example, [13, 15]. Once the agent stops improving her policy, the mean field information is updated assuming that all agents follow the same improved policy (line 12). In the RL literature, this procedure is sometimes called *self play.*

- *Agent update.* Within each outer iteration k under a fixed mean field information m^{k-1}, the agent will update her estimation of the optimal policy $\widehat{R}$ for I rounds (lines 4-11). Each round corresponds to one gradient descent step (line 10) and requires n samples of the simulated reward function (line 7) associated with the perturbed version of $\widehat{R}^i$ (line 6).

Algorithm 1 Mean Field Policy Gradient with Exploration

1: **Input**: Initial beliefs $m^0 := \{m_s^0\}_{s=0}^N$, distribution of initial policy $\mathcal{D}$, number of trajectories n, smoothing parameter r, learning rate η.
2: **for** $k \in \{1, \dots, K\}$ **do**
3: Sample initial policy $\widehat{R}^0 \sim \mathcal{D}$.
4: **for** $i \in \{0, \dots, I\}$ **do**
5: **for** $j \in \{1, \dots, n\}$ **do**
6: Sample policy $\widehat{R}^{i,j} = \widehat{R}^i + U^{i,j}$ where $U^{i,j} \in \mathbb{R}^{N+2}$ is drawn uniformly at random over matrices such that $\|U^{i,j}\|_F = r$, where $\|\cdot\|_F$ is the Frobenius norm.
7: Denote $\widehat{j}\left(\widehat{R}^{i,j}, m^{k-1}\right)$ as the single trajectory cost with policy $\widehat{R}^{i,j}$ starting from $x_0^{i,j} \sim \nu$ under fixed mean state m^{k-1}.
8: **end for**
9: Obtain the estimate of $\nabla J(\widehat{R}^i, m^{k-1})$:

$$\widehat{\nabla J(\widehat{R}^i, m^{k-1})} = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{r^2} \widehat{j}\left(\widehat{R}^{i,j}, m^{k-1}\right) U^{i,j}. \tag{4.3}$$

10: Perform policy gradient descent step:

$$\widehat{R}^{i+1} = \widehat{R}^i - \eta \widehat{\nabla J(\widehat{R}^i, m^{k-1})}. \tag{4.4}$$

11: **end for**
12: Update mean field information $m^k = \{m_s^k\}_{s=0}^N$ assuming all agents follow policy $\widehat{R}^I$.
13: **end for**

- The gradient term $\nabla J(\widehat{R}^i, m^{k-1})$ in (4.3) is estimated using a *zeroth-order optimization approach* (line 9). That is, the agent only has query access to a sample of the reward function $\widehat{j}(\cdot)$ at input points (R, m), without querying the gradients and higher order derivatives of $\widehat{j}(\cdot)$. Moreover, to avoid the issue of ill-definedness of $\mathbb{E}_{U \sim \mathcal{N}(0,\sigma^2 I)}[J(\widehat{R} + U, m)]$ with a Gaussian smoothing, we choose $\mathbb{S}_r$ by smoothing over the sphere of a ball; hence, step (4.3) in Algorithm 1 is to find, for a given m, a bounded and biased estimate $\widehat{\nabla J(\widehat{R}, m)}$ of $\nabla J(\widehat{R}, m)$.

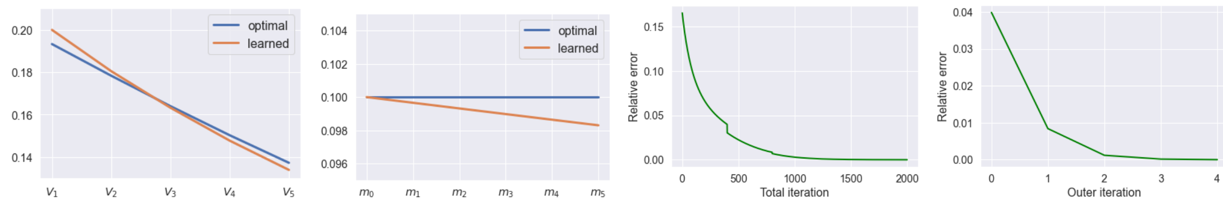
4.3 Results

Model set-up. We take $T = 0.1$, $\delta = 0.02$ (hence $N = \frac{T}{\delta} = 5$), $A = 2.0$, $B = 3.0$, $D = 2.0$, $Q = 3.0$, $\bar{Q} = 2.0$, $\mathbb{E}[\xi] = 0.1$, and $\mathbb{E}[\xi^2] = 1$.

Experiment set-up. We set $r = 0.01$ and $\eta = 0.05$, $m_s^0 = 0.0$ ($s = 0, 1, \cdots, N$), $\widehat{M}^0 \sim \mathcal{N}(0.5, 1)$, and $\widehat{\sigma}_s^0 \sim \mathcal{N}(0.5, 0.1)$ ($s = 0, 1, \cdots, N-1$), with $K = 10$, $n = 50$, and $I = 400$.

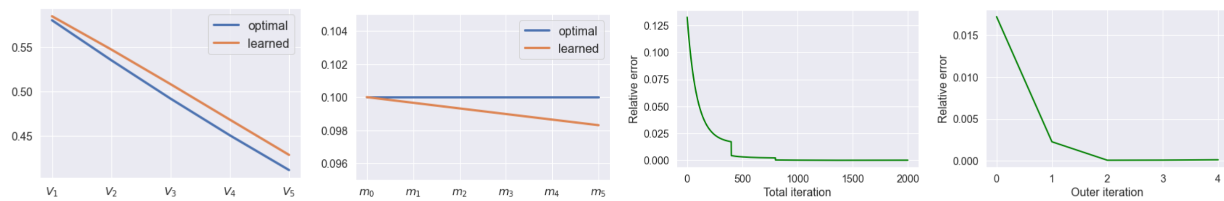
Performance evaluation. Given policy $\widehat{R}$ and mean field information m, define the *relative error* between $(\widehat{R}, m)$ and the mean field solution (R^*, m^*) of problem (4.1)-(4.2) as

$$Err(\widehat{R}, m) := \frac{|J(\widehat{R}, m) - J(R^*, m^*)|}{|J(R^*, m^*)|}. \tag{4.5}$$



(a) $\{V_s\}_{s=0}^N$ (b) $\{m_s\}_{s=0}^N$ (c) Relative error in total iterations. (d) Relative error in outer iterations

Figure 1: Performance of the algorithm when $\lambda_{SE} = 1.0$. (True $M = 0.75$ and learned $\widehat{M} = 0.732$.)



(a) $\{V_s\}_{s=0}^N$ (b) $\{m_s\}_{s=0}^N$ (c) Relative error in total iterations. (d) Relative error in outer iterations

Figure 2: Performance of the algorithm when $\lambda_{SE} = 3.0$, with true $M = 0.75$ and learned $\widehat{M} = 0.736$.

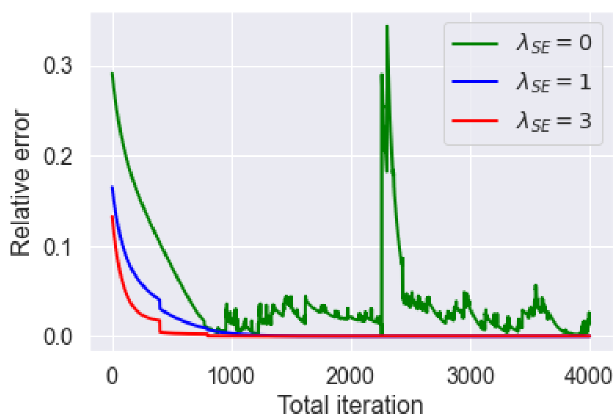


Figure 3: Comparison of relative errors with different λ_{SE}.

Results.

1. *Stability.* As seen from Figure 3, when $\lambda_{SE} = 0$, i.e., when there is no exploration, the algorithm is unstable. Within each outer iteration, the error level fluctuates when the representative agent updates her policy under a fixed mean field information. At the end of each outer iteration, there is a sudden jump in the error when the population updates its mean field policy. In contrast, the algorithm is stable when exploration is included, i.e., when $\lambda_{SE} > 0$.

2. *Speed of convergence.* As Figures 1b and 2b show, Shannon entropy ($\lambda_{SE} > 0$) improves the speed of convergence to the mean field equilibrium. In fact, the algorithm does not

converge without entropy regularization, i.e., when $\lambda_{SE} = 0$; On the other hand, the algorithm converges to the equilibrium solution when $\lambda_{SE} = 1$ and $\lambda_{SE} = 3$. Moreover, the convergence speed is faster with $\lambda_{SE} = 3$ than with $\lambda_{SE} = 1$, with the former converging to the mean field equilibrium within three outer iterations and the latter in five outer iterations.

3. *Accuracy of learned mean field equilibrium.* Figures 1b and 2b show consistency with Theorems 2 and 4. The algorithm is able to learn the mean field information with small errors ($< 5\%$) for both cases $\lambda_{SE} = 1$ and $\lambda_{SE} = 3$.

4. *Learning optimal scheduling of the exploration policy.* With given parameters, the variance of the Gaussian mean field policy (a.k.a., the optimal exploration scheduling) is a decreasing function of time t for both $\lambda_{SE} = 1$ and $\lambda_{SE} = 3$. Figures 1a and 2a suggest that the agent can learn this decreasing function $\{\widehat{\sigma}_s^2\}_{s=0}^T$ with small error ($< 5\%$).

References

[1] Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160, 2019.

[2] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*, 2020.

[3] Karl J Aström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[4] Martino Bardi. Explicit solutions of some linear-quadratic mean field games. *Networks & Heterogeneous Media*, 7(2):243, 2012.

[5] Alain Bensoussan, KCJ Sung, Sheung Chi Phillip Yam, and Siu-Pang Yung. Linear-quadratic mean field games. *Journal of Optimization Theory and Applications*, 169(2):496–529, 2016.

[6] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[7] Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. *The Master Equation and the Convergence Problem in Mean Field Games:(AMS-201)*, volume 201. Princeton University Press, 2019.

[8] René Carmona, François Delarue, et al. Mean field forward-backward stochastic differential equations. *Electronic Communications in Probability*, 18, 2013.

[9] René Carmona, Daniel Lacker, et al. A probabilistic weak formulation of mean field games and applications. *Annals of Applied probability*, 25(3):1189–1231, 2015.

[10] René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: Convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[11] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

[12] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

[13] Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*, 2019.

[14] Jordi Grau-Moya, Felix Leibfried, and Peter Vrancx. Soft Q-learning with mutual-information regularization. In *International Conference on Learning Representations*, 2018.

[15] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, pages 4967–4977, 2019.

[16] Ben M Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *Available at SSRN*, 2020.

[17] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

[18] Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

[19] Zhang-Wei Hong, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. A deep policy inference Q-network for multi-agent systems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1388–1396. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[20] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

[21] Minyi Huang, Peter E Caines, and Roland P Malhamé. Large-population cost-coupled lqg problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria. *IEEE transactions on automatic control*, 52(9):1560–1571, 2007.

[22] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 2961–2970, 2019.

[23] Daniel Lacker. Mean field games via controlled martingale problems: existence of markovian equilibria. *Stochastic Processes and their Applications*, 125(7):2856–2894, 2015.

[24] Daniel Lacker. Mean field games via controlled martingale problems: existence of markovian equilibria. *Stochastic Processes and their Applications*, 125(7):2856–2894, 2015.

[25] Daniel Lacker and Thaleia Zariphopoulou. Mean field and n-agent games for optimal investment under relative performance criteria. *Mathematical Finance*, 2017.

[26] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[27] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[28] Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2461–2470, 2019.

[29] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

[30] Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

[31] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

[32] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. *arXiv preprint arXiv:1802.09640*, 2018.

[33] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[34] Haoran Wang, Thaleia Zariphopoulou, and Xunyu Zhou. Exploration versus exploitation in reinforcement learning: a stochastic control approach. *Journal of Machine Learning Research, To Appear*, 2020.

[35] Haoran Wang and Xun Yu Zhou. Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Available at SSRN 3382932*, 2019.

[36] Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. *arXiv preprint arXiv:2008.06845*, 2020.

[37] Michael Wunder, Michael L Littman, and Monica Babes. Classes of multiagent Q-learning dynamics with epsilon-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1167–1174. Citeseer, 2010.

[38] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.