Global Convergence of Policy Gradient for Linear-Quadratic Mean-Field Control/Game in Continuous Time

Weichen Wang [1] Jiequn Han [2] Zhuoran Yang [3] Zhaoran Wang [4]

Abstract

Recent years have witnessed the success of multi-agent reinforcement learning, which has motivated new research directions for mean-field control (MFC) and mean-field game (MFG), as the multi-agent system can be well approximated by a mean-field problem when the number of agents grows to be very large. In this paper, we study the policy gradient (PG) method for the linear-quadratic mean-field control and game, where we assume each agent has identical linear state transitions and quadratic cost functions. While most recent works on policy gradient for MFC and MFG are based on discrete-time models, we focus on a continuous-time model where some of our analyzing techniques could be valuable to the interested readers. For both the MFC and the MFG, we provide PG update and show that it converges to the optimal solution at a linear rate, which is verified by a synthetic simulation. For the MFG, we also provide sufficient conditions for the existence and uniqueness of the Nash equilibrium.

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 2018) has become a very powerful tool for learning the optimal policy of a complicated system, with many successful applications including playing games achieving potential superhuman performance, such as Atari (Mnih et al., 2013), GO (Silver et al., 2016; 2017), Poker (Heinrich & Silver, 2016; Moravčík et al., 2017), multiplayer online video games Dota (OpenAI, 2018) and StarCraft (Vinyals et al., 2019), and more realistic real-world problems, such as robotic control (Yang & Gu, 2004), autonomous driving (Shalev-Shwartz et al., 2016), and social dilemmas (de Cote et al., 2006; Leibo et al., 2017; Hughes et al., 2018). The above are just some illustrative examples. More generally, RL has been applied to design efficient algorithms for decision making to minimize the long-term expected overall cost through interacting with the environment sequentially.

On a separate line of research, the subject of the optimal control assumes knowledge of the system dynamics and the observed reward/cost function, and studies the existence and uniqueness of the optimal solution. Extensive literature extends this area from the most basic setting of the linear-quadratic regulator problem (Willems, 1971; Bertsekas, 1995; Anderson & Moore, 2007) to the zero-sum game (Engwerda, 2005; Zhang, 2005) and to the multi-agent control/game (Egerstedt & Hu, 2001; Parsons & Wooldridge, 2002; Shamma, 2008; Semsar-Kazerooni & Khorasani, 2009; Dimarogonas & Johansson, 2010). However, the multi-agent control/game is typically computationally intractable for a large real-world problem, as the joint state and action spaces grow exponentially in the number of agents. Mean-field control/game proposed by (Huang et al., 2003; 2006; Lasry & Lions, 2006a;b; 2007) can be viewed as an approximation to the multi-agent control/game when the number of agents grows to infinity. In mean-field control/game, each agent shares the same cost function and state transition, which depend on other agents only through their aggregated mean effect. Consequently, each agent's optimal policy only depends on its own state and the mean-field state of the population. This symmetry across all agents significantly simplifies the analysis. Mean-field control/game has already found a lot of meaningful applications such as power grids (Minciardi & Sacile, 2011), swarm robots (Fang, 2014; Araki et al., 2017) and financial systems (Zhou & Li, 2000; Huang & Li, 2018).

Although the traditional optimal control approach lays a solid foundation for theoretical analysis, it fails to adapt well to the modern situation where we may have a huge system or complicated environment to explore. Therefore, recent years have witnessed increased interest in applying the RL techniques to various optimal control settings. See (Fazel et al., 2018; Zhang et al., 2019; Bu et al., 2019; Elie

[1] Faculty of Business and Economics, The University of Hong Kong. [2] Dept. of Mathematics, Princeton University. [3] Dept. of Operations Research & Financial Engineering, Princeton University. [4] Dept. of Industrial Engineering & Management Sciences, Northwestern University. Correspondence to: Weichen Wang <nickweichwang@gmail.com>, Jiequn Han <jiequnh@princeton.edu>, Zhuoran Yang <zy6@princeton.edu>, Zhaoran Wang <zhaoran.wang@northwestern.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

et al., 2020) for some examples. Specifically, this paper focuses on the RL technique of policy gradient (Sutton et al., 2000; Kakade, 2002; Silver et al., 2014), where we update the policy following the gradient of the cost function, and the setting of the linear-quadratic mean-field control/game (MFC/MFG), where we assume each agent has identical linear state transition and quadratic cost function. The MFC differs from the MFG in that the former allows all the agents to directly control the mean-field state and collaborate in order to maximize the social welfare together, while the latter can only allow each agent to make an individual decision with a guess on the mean-field output, hoping to achieve the Nash equilibrium of the system. The paper aims to show that policy gradient methods can achieve a desired linear convergence for both MFC and MFG. We follow the traditional optimal control setting for simplicity and choose the model-based approach to better present the theoretical results and algorithm. The corresponding model-free algorithm to estimate the gradient can be derived similarly following the works such as (Fazel et al., 2018; Carmona et al., 2019; Fu et al., 2019).

Many of the recent stochastic mean-field control/game literature are based on the continuous-time models, e.g. (Bensoussan et al., 2013; Cardaliaguet & Hadikhanloo, 2017; Carmona et al., 2018), where the main focus is on characterizing the properties of the optimal solution through solving a pair of Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck (FP) equations, rather than designing provably efficient learning algorithms. However, new developments on policy gradient algorithms for MFC and MFG are mainly based on discrete-time models, e.g. (Elliott et al., 2013; Guo et al., 2019; Carmona et al., 2019; Fu et al., 2019). One reason is that discrete-time models can be more straightforward to analyze. For example, (Fazel et al., 2018) pioneered the techniques to show the theoretical global convergence of PG for the classical linear-quadratic regulator (LQR) in discrete time. One contribution of the current paper is to extend those techniques to the setting of continuous-time stochastic models.

We will organize the paper as below. In Section 2, we review the continuous-time classical LQR problem and show that the PG converges to the optimal solution at a linear rate, with techniques designed for analyzing continuous stochastic dynamics. In Section 3, we formulate the MFC problem and reveal that with some reparametrization, MFC can be readily transformed into a classical LQR problem. The MFG however is much more involved to study, so we present the drifted LQR problem first in Section 4 as an intermediate step towards analyzing PG for MFG. Last but not least, in Section 5 we provide an algorithm for solving MFG which provably also enjoys the linear convergence rate. The algorithm naturally contains two update steps: for a given mean-field state, each agent seeks the best response by solving a drifted LQR problem; then to find the Nash equilibrium, we update the mean-field state assuming each agent follows the best strategy. We will define the Nash equilibrium more concretely and provide sufficient conditions for its existence and uniqueness in Section 5 as well. Finally, we demonstrate our theoretical results with synthetic simulations in Section 6 and conclude the paper with some discussions in Section 7.

Notations. For a matrix M, we denote by $\|M\|_2$ (or $\|M\|$), $\|M\|_F$ the spectral and Frobenius norm, $\sigma_{\min}(M), \sigma_{\max}(M)$ its minimum and maximum singular value, and $\operatorname{tr}(M)$ the trace of M when M is a square matrix. Let $\langle M, N\rangle = \operatorname{tr}(M^\top N)$. We use $\|\alpha\|_2$ (or $\|\alpha\|$) to represent the ℓ_2-norm of a vector α. For scalars $a_1, \ldots, a_n$, we denote by $\operatorname{poly}(a_1, \ldots, a_n)$ the polynomial of $a_1, \ldots, a_n$.

2. Linear-Quadratic Regulator

As the simplest optimal control problem, linear quadratic regulator serves as a perfect baseline to examine the performance of reinforcement learning methods. Viewing LQR from the lens of Markov decision process (MDP), the state and action spaces are $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{U} = \mathbb{R}^k$, respectively. The continuous-time state transition dynamics is specified as the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = (AX_t + Bu_t)\mathrm{d}t + D\mathrm{d}W_t, \tag{1}$$

where W_t is standard d-dimensional Brownian motion. We consider the infinite-horizon time-average cost that each agent aims to minimize

$$\limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T c(X_t, u_t)\mathrm{d}t\right],$$
$$X_0 \sim \mu_0, \quad c(x,u) = x^\top Qx + u^\top Ru,$$

where the initial state X_0 is assumed to be sampled from the initial distribution μ_0. The $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times k}$, $D \in \mathbb{R}^{d\times d}$, $Q \in \mathbb{R}^{d\times d}$, $R \in \mathbb{R}^{k\times k}$ are matrices of proper dimensions with $Q, R \succ 0$.

It is known that the optimal action are linear in the corresponding state (Anderson & Moore, 2007; Bertsekas, 1995). Specifically, the optimal actions satisfy $u_t^* = -K^*X_t$ for all $t \geq 0$, where $K^* \in \mathbb{R}^{k\times d}$ can be written as $K^* = R^{-1}B^\top P^*$, with P^* being the solution to the continuous time algebraic Riccati equation

$$A^\top P^* + P^*A^\top - P^*BR^{-1}B^\top P^* + Q = 0. \tag{2}$$

2.1. Ergodic Cost and Relative Value Function

Inspired by the form of the optimal policy, we consider the general linear policy $u_t = -KX_t$, where $K \in \mathbb{R}^{k\times d}$ is the parameter to be optimized. The state dynamics becomes

$$\mathrm{d}X_t = (A - BK)X_t\mathrm{d}t + D\mathrm{d}W_t. \tag{3}$$

We assume $A - BK$ is stable[1], that is the real parts of all the eigenvalues of $A - BK$ are negative. Denote the invariant distribution of (3) as ρ_K. It is a Gaussian distribution $N(0, \Sigma_K)$, where Σ_K satisfies the continuous Lyapunov equation

$$(A - BK)\Sigma_K + \Sigma_K(A - BK)^\top + DD^\top = 0. \quad (4)$$

Then the associated ergodic cost $J(K)$ and the relative value function $V_K(x)$ can be expressed as

$$\begin{aligned} J(K) &:= \mathbb{E}_{X_t \sim \rho_K}[c(X_t, u_t)] \\ &= \mathbb{E}_{X_t \sim \rho_K}[X_t^\top (Q + K^\top RK)X_t] \quad (5) \\ &= \langle Q + K^\top RK, \Sigma_K \rangle, \end{aligned}$$

$$V_K(x) := \mathbb{E}\left[\int_0^\infty [c(X_t, u_t) - J(K)]\mathrm{d}t \,|\, X_0 = x\right].$$

Using dynamic programming, we have the Hamilton-Jacobi-Bellman (HJB) equation for $V_K(x)$

$$\begin{aligned} c(x, -Kx) - J(K) + [(A - BK)x]^\top \nabla V_K(x) & \\ + \frac{1}{2}\langle \nabla^2 V_K(x), DD^\top \rangle = 0. & \quad (6) \end{aligned}$$

Assuming the ansatz $V_K(x) = x^\top P_K x + C_K$ with a symmetric P_K and plugging it into (6), we have

$$\begin{aligned} x^T(Q + K^\top RK)x - J(K) + 2x^\top (A - BK)^\top P_K x & \\ + \langle P_K, DD^\top \rangle = 0. & \end{aligned}$$

This implies the following two equations need to be valid at the same time

$$(A - BK)^\top P_K + P_K(A - BK) + Q + K^\top RK = 0, \quad (7)$$

$$J(K) = \langle P_K, DD^\top \rangle. \quad (8)$$

To see it is possible, notice that the continuous Lyapunov equation (7) has a well-defined solution P_K if $A - BK$ is stable[2], and it also satisfies (8) since combining (4)(5)(7) we find

$$\begin{aligned} J(K) &= \langle Q + K^\top RK, \Sigma_K \rangle \\ &= -\operatorname{tr}[((A - BK)^\top P_K + P_K(A - BK))\Sigma_K] \\ &= -\operatorname{tr}[P_K(\Sigma_K(A - BK)^\top + (A - BK)\Sigma_K)] \\ &= \langle P_K, DD^\top \rangle. \end{aligned}$$

Therefore if $A - BK$ is stable, there exists a well-defined P_K satisfying (7)(8) simultaneously. Note that by definition $\mathbb{E}_{x \sim \rho_K}[V_K(x)] = 0$, so the constant term in $V_K(x)$ can be determined as

$$C_K = \mathbb{E}_{x \sim \rho_K}[x^\top P_K x] = \langle P_K, \Sigma_K \rangle.$$

2.2. Policy Gradient and Convergence

To implement the gradient descent method on $J(K)$, with a fixed stepsize η, we follow $K \leftarrow K - \eta \nabla_K J(K)$. The following proposition gives out the explicit formula for $\nabla_K J(K)$.

Proposition 1. *(Expression of the gradient).*

$$\nabla_K J(K) = 2(RK - B^\top P_K)\Sigma_K = 2E_K \Sigma_K,$$

where we define $E_K := RK - B^\top P_K$.

The proof of Proposition 1 can be found in Appendix A of the supplementary material (Proposition A.3). With the above explicit formula for policy gradient, we present an upper bound for $J(K) - J(K^*)$ below, which shows the cost function is gradient dominated (Karimi et al., 2016).

Lemma 2. *(Gradient domination).*

$$J(K) - J(K^*) \leq \frac{\|\Sigma_{K^*}\| \operatorname{tr}(\nabla_K J(K)^\top \nabla_K J(K))}{\sigma_{\min}(R)\sigma_{\min}^2(DD^\top)}.$$

Lemma 2 indicates that $J(\cdot)$ is regular in the sense that, despite nonconvex, all stationary points of $J(\cdot)$ are global minima. This property is essential in establishing the linear convergence of policy gradient. Although the current paper only focuses on the LQ setting, as a first step towards understanding policy gradient for continuous-time control, the analysis could be generalized beyond the LQ setting for a regular cost function.

The following theorem is the main result for this section, revealing that the policy gradient method for the continuous-time LQR achieves linear convergence rate. Its proof, together with those for the above proposition and lemma can be found in Appendix A of the supplementary material.

Theorem 3. *(Global convergence of model-based gradient descent). Assume the policy gradient starts from an initial K_0 satisfying that $A - BK_0$ is stable. With an appropriate constant setting of the stepsize η in the form of $\eta = poly\left(\frac{\sigma_{\min}(Q)}{C(K_0)}, \sigma_{\min}(DD^\top), \|B\|^{-1}, \|R\|^{-1}\right)$, and number of iterations*

$$N \geq \frac{\|\Sigma_{K^*}\|}{\eta \sigma_{\min}^2(DD^\top)\sigma_{\min}(R)} \log \frac{J(K_0) - J(K^*)}{\varepsilon},$$

the iterates of gradient descent enjoys $J(K_N) - J(K^*) \leq \varepsilon$. Comparing to Theorem 7 of (Fazel et al., 2018) for the linear convergence of policy gradient for the discrete-time LQR, the results for the continuous case is simpler in that η does not depend on $\|A\|$ and $\sigma_{\min}(R)$.

[1] Actually, we only require a stable initial K_0 in the policy gradient method, as long as the step size η is small, we can guarantee that all K's following policy gradient descent will be stable. See Section 2.2 and Lemma A.5 in the supplementary material.

[2] Please refer to Lemma A.1 in the supplementary material and the reference therein.

3. Linear-Quadratic Mean-Field Control

Now we consider a linear-quadratic regulator with mean-field interactions

$$\mathrm{d}X_t = (AX_t+\bar{A}\mathbb{E}_0[X_t]+Bu_t+\bar{B}\mathbb{E}_0[u_t])\mathrm{d}t+D\mathrm{d}W_t+\bar{D}\mathrm{d}W_t^0, \tag{9}$$

in which W_t, W_t^0 are the idiosyncratic and common noise modeled by two independent d-dimensional Brownian motions and $\mathbb{E}_0$ denotes the conditional expectation given W_t^0. We call $\mathbb{E}_0[X_t]$ and $\mathbb{E}_0[u_t]$ in (9) the mean-field state and mean-field action respectively. In MFC problem, the agent seeks for policy in terms of $u_t = u(X_t, \mathbb{E}_0[X_t])$ to minimize the following infinite-horizon time-average cost

$$\limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T c(X_t, \mathbb{E}_0[X_t], u_t, \mathbb{E}_0[u_t])\mathrm{d}t\right],$$
$$X_0 \sim \mu_0, \quad Q, \bar{Q}, R, \bar{R} \succ 0\,,$$
$$c(x, \bar{x}, u, \bar{u}) = x^\top Qx + \bar{x}^\top \bar{Q}\bar{x} + u^\top Ru + \bar{u}^\top \bar{R}\bar{u}\,. \tag{10}$$

The discrete version of this model for MFC has been studied in (Carmona et al., 2019). Here we consider the continuous-time model for both MFC and MFG in this and the next sections. Another comment is that assuming identical transition and cost may seem to be unrealistic in practice, but this allows us to focus more on how individual agents interact with the mean-field and to study the asymptotic limit when the number of agents grows to infinity. This simplification has been adopted by some recent theoretical works such as (Elie et al., 2020; Guo et al., 2019; Carmona et al., 2019; Fu et al., 2019)

3.1. Reparametrization

For this problem, under some suitable conditions, one can prove the optimal control is a linear combination of X_t and $\mathbb{E}_0[X_t]$, see e.g. (Carmona et al., 2018). We can actually recast the original MFC into a LQR problem with a larger state space. Specifically, motivated by the form of the optimal policy, we consider the general linear policy

$$u_t = -K(X_t - \mathbb{E}_0[X_t]) - L\mathbb{E}_0[X_t], \tag{11}$$

where $\theta = (K, L)$ are the two parameter matrices to be optimized. Denote by $Y_t^1 = X_t - \mathbb{E}_0[X_t]$ and $Y_t^2 = \mathbb{E}_0[X_t]$. An important observation is that, under the policy (11), the dynamics of these two processes are decoupled

$$\mathrm{d}Y_t^1 = (A - BK)Y_t^1\mathrm{d}t + D\mathrm{d}W_t,$$
$$\mathrm{d}Y_t^2 = (A + \bar{A} - (B + \bar{B})L)Y_t^2\mathrm{d}t + \bar{D}\mathrm{d}W_t^0.$$

Moreover, the running cost can also be written as a quadratic function of (Y_t^1, Y_t^2). Therefore one can essentially optimize K and L similar to the LQR, and all the theoretical results should follow.

4. Drifted Linear-Quadratic Regulator

In this section, we extend the simplest linear SDE dynamics to include an intercept in the drift. This extension is going to be useful for MFG. The state transition dynamics considered in this section is

$$\mathrm{d}X_t = (a + AX_t + Bu_t)\mathrm{d}t + D\mathrm{d}W_t\,. \tag{12}$$

Each agent still aims to minimize the same quadratic cost $c(x, u) = x^\top Qx + u^\top Ru$.

4.1. Ergodic Cost

We again consider the general linear policy, but with an extra intercept, $u_t = -KX_t + b$, where $K \in \mathbb{R}^{k\times d}$ and $b \in \mathbb{R}^k$ are the parameters to be optimized. The state dynamics becomes

$$\mathrm{d}X_t = ((A - BK)X_t + a + Bb)\mathrm{d}t + D\mathrm{d}W_t. \tag{13}$$

The invariant distribution $\rho_{K,b}$ of (13) is a Gaussian distribution $N(\mu_{K,b}, \Sigma_K)$, where $\mu_{K,b}$ satisfies $\mu_{K,b} = -(A - BK)^{-1}(a + Bb)$ and Σ_K does not depend on b and still satisfies the continuous Lyapunov equation $(A - BK)\Sigma_K + \Sigma_K(A - BK)^\top + DD^\top = 0$. The associated ergodic cost can be expressed as

$$J(K, b) := \mathbb{E}_{X_t\sim\rho_{K,b}}[c(X_t, u_t)] = J_1(K) + J_2(K, b)\,, \tag{14}$$

where $J_1(K)$ and $J_2(K, b)$ are defined as

$$J_1(K) = \left\langle Q + K^\top RK, \Sigma_K\right\rangle = \left\langle P_K, DD^\top\right\rangle,$$

$$J_2(K, b) = \begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}^\top \begin{pmatrix}Q + K^\top RK & -K^\top R\\ -RK & R\end{pmatrix}\begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}$$

Here $J_1(K)$ is the expected total cost in the regular LQR problem without intercept and P_K is the solution of the continuous Lyapunov equation (7). Meanwhile, $J_2(K, b)$ corresponds to the expected cost induced by the drift a of the transition dynamics and the policy intercept b.

4.2. Policy Gradient and Convergence

Proposition 4. *The optimal intercept b^K to minimize $J_2(K, b)$ for any given K is that*

$$b^K = -(KQ^{-1}A^\top + R^{-1}B^\top)(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a \tag{15}$$

Furthermore, $J_2(K, b^K)$ takes the form of

$$J_2(K, b^K) = a^\top(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a \tag{16}$$

which is independent of K.

Since $\min_b J_2(K, b)$ does not depend on K, it holds that the optimal K^* can be obtained by minimizing $J_1(K)$ similar to the case of no intercept, that is, updating K following the gradient direction $\nabla_K J_1(K)$. So the optimal K^* does not depend on the intercept a at all. Once we have the optimal K^*, the optimal $b^* = b^{K^*}$ is obtained by plugging in K^* in (15). From Proposition 1, we know $\nabla_K J(K) = \nabla_K J_1(K) = 2(RK - B^\top P_K)\Sigma_K$.

Define μ^K to be the mean of the invariant density corresponding to $u_t = -KX_t + b^K$. Then $\mu^K = -(A - BK)^{-1}(a + Bb^K) = -Q^{-1}A^\top(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a$, which does not depend on K. The state dynamics can be written as

$$\mathrm{d}(X_t - \mu^K) = (A - BK)(X_t - \mu^K)\mathrm{d}t + D\mathrm{d}W_t.$$

And the cost function $J(K) = J(K, b^K) = J_1(K) + a^\top(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a$. These derivations reveal that we can directly apply convergence theorem of the policy gradient for the classical LQR to $X_t - \mu^K$. We relegate all the proofs to Appendix B of the supplementary material.

Theorem 5. *(Global convergence for drifted LQR). With the initial $A - BK_0$ stable and the stepsize η in the same form as Theorem 3 and the number of iterations*

$$N \geq \frac{\|\Sigma_{K^*}\|}{\eta\sigma_{\min}^2(DD^\top)\sigma_{\min}(R)} \log \frac{J_1(K_0) - J_1(K^*)}{\varepsilon},$$

if we follow $b^K = -(KQ^{-1}A^\top + R^{-1}B^\top)(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a$, we have $J(K_N, b^{K_N}) - J(K^*, b^*) \leq \varepsilon$. Furthermore,

$$\|K_N - K^*\|_F \leq \sigma_{\min}^{-1/2}(R)\sigma_{\min}^{-1/2}(DD^\top)\sqrt{\varepsilon},$$

$$\|b^{K_N} - b^*\|_2 \leq C_b(a)\sigma_{\min}^{-1/2}(R)\sigma_{\min}^{-1/2}(DD^\top)\sqrt{\varepsilon}\,,$$

where $C_b(a) = \|Q^{-1}A^\top(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}a\|_2$ is a constant depending on the intercept a.

5. Linear-Quadratic Mean-Field Game

The linear-quadratic MFG has the same dynamics (9) and cost function (10) as the MFC problem. But the key difference is that MFC allows all the agents to conduct the control together, whereas in MFG each agent has to optimize its own objective assuming a guess of the mean-field state/action. Therefore, the ultimate goal of studying MFG is to see if multiple agents can reach a Nash equilibrium, where given the mean-field state/action, the policy of each agent is optimal and given all the agents carry out the optimal policy, we recover exactly the same mean-field state/action.

So the idea of policy gradient for MFG is straightforward: for any given mean-field state/action, we update policy by following the gradient and then with the updated policy we update the mean-field state/action. We will provide sufficient conditions for the existence and uniqueness of the Nash equilibrium and show that PG can converge to the Nash equilibrium at a linear rate.

To that end, we need to study the linear-quadratic control problem for any given mean-field state μ_x and mean-field action μ_u, that is,

$$\begin{aligned}
&\mathrm{d}X_t = (AX_t + \bar{A}\mu_x + Bu_t + \bar{B}\mu_u)\mathrm{d}t + D\mathrm{d}W_t + \bar{D}\mathrm{d}W_t^0,\\
&c(X_t, u_t) = X_t^\top QX_t + \mu_x^\top\bar{Q}\mu_x + u_t^\top Ru_t + \mu_u^\top\bar{R}\mu_u,\\
&J_{\mu_x,\mu_u}(\pi) = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T c(X_t, u_t)\mathrm{d}t\right], \quad X_0 \sim \mu_0,
\end{aligned} \tag{17}$$

where u_t is the action vector generated by playing policy π. Define $\mu = (\mu_x^\top, \mu_u^\top)^\top \in \mathbb{R}^{d+k}$. We hope to find the optimal policy $\pi_\mu^* = \inf_{\pi\in\Pi} J_\mu(\pi)$. This is clearly a drifted LQR problem with an intercept $\bar{A}\mu_x + \bar{B}\mu_u$ in the drift. As in the drifted LQR, we consider the class of linear policies with an intercept, that is,

$$\Pi = \{\pi(x) = -Kx + b : K \in \mathbb{R}^{k\times d}, b \in \mathbb{R}^k\}.$$

Hence it suffices to find the optimal policy π_μ^* within Π.

Now, we introduce the definition of Nash equilibrium (Saldi et al., 2018). The Nash equilibrium is obtained if we can find a pair (π^*, μ^*), such that the policy π^* is optimal for each agent when the mean-field state/action is μ^*, while all the agents following the policy π^* generate the mean-field state/action μ^* as $t \to \infty$. To present its formal definition, we define $\Lambda_1(\mu)$ as the optimal policy in Π given the mean-field μ, and define $\Lambda_2(\mu, \pi)$ as the mean-field state/action generated by the policy π given the current mean-field μ as $t \to \infty$.

Definition 6. *(Nash Equilibrium Pair). The pair $(\mu^*, \pi^*) \in \mathbb{R}^{d+k} \times \Pi$ constitutes a Nash equilibrium pair of (17) if it satisfies $\pi^* = \Lambda_1(\mu^*)$ and $\mu^* = \Lambda_2(\mu^*, \pi^*)$. Here μ^* is called the Nash mean-field state/action and π^* is called the Nash policy.*

5.1. Existence and Uniqueness of Nash Equilibrium

Let us first rewrite (17) as follows:

$$\begin{aligned}
&\mathrm{d}X_t = (\tilde{a}_\mu + AX_t + Bu_t)\mathrm{d}t + \tilde{D}\mathrm{d}\tilde{W}_t,\\
&c(X_t, u_t) = X_t^\top QX_t + u_t^\top Ru_t + \tilde{C}_\mu,
\end{aligned} \tag{18}$$

where $\tilde{a}_\mu = \bar{A}\mu_x + \bar{B}\mu_u$ is the intercept in the drift, $\tilde{D} = (D, \bar{D}) \in \mathbb{R}^{d\times 2d}$ is an expanded matrix, $\tilde{W}_t = (W_t^\top, {W_t^0}^\top)^\top \in \mathbb{R}^{2d}$ is $2d$-dimensional Brownian motion, $\tilde{C}_\mu = \mu_x^\top\bar{Q}\mu_x + \mu_u^\top\bar{R}\mu_u$ is a constant. So this is exactly the drifted LQR problem we considered in (12) with the same quadratic cost function ignoring the constant term.

Therefore, for the mapping $\pi_\mu^* = \Lambda_1(\mu)$, from (15) in Proposition 4, we know $\pi_\mu^*(x) = -K^*x + b_\mu^*$ where

$$b_\mu^* = -(K^*Q^{-1}A^\top + R^{-1}B^\top)(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}\tilde{a}_\mu\,.$$

Note that K^* is fixed for all μ. For the mapping $\mu_{\text{new}} = \Lambda_2(\mu, \pi) = (\mu_{\text{new},x}^\top, \mu_{\text{new},u}^\top)^\top$ where $\pi(x) = -K_\pi x + b_\pi$, it is not hard to see the new mean of the mean-field state/action should be

$$\mu_{\text{new},x} = -(A - BK_\pi)^{-1}(Bb_\pi + \tilde{\alpha}_\mu)\,, \tag{19}$$

$$\mu_{\text{new},u} = b_\pi + K_\pi(A - BK_\pi)^{-1}(Bb_\pi + \tilde{\alpha}_\mu)\,. \tag{20}$$

With the more detailed formulas for the mapping Λ_1 and Λ_2, we then establish the existence and uniqueness of the Nash equilibrium. The following conditions are required.

Assumption 7. *We assume the following conditions hold.*

(i) The continuous-time Riccati equation $A^\top P^* + P^* A^\top - P^* B R^{-1} B^\top P^* + Q = 0$ admits a unique symmetric positive definite solution P^*.

(ii) The optimal $K^* = R^{-1}B^\top P^*$. It holds that $L_0 = L_1 L_3 + L_2 < 1$, where

$$L_1 = \|K^* Q^{-1} A^\top + R^{-1} B^\top\| \cdot \max\left\{\left\|\Gamma^{-1}\bar{A}\right\|, \left\|\Gamma^{-1}\bar{B}\right\|\right\},$$

$$L_2 = \max\left\{\|\Delta_A\| + \|K^*\Delta_A\|, \|\Delta_B\| + \|K^*\Delta_B\|\right\},$$

$$L_3 = \|(A - BK^*)^{-1}B\| + \|I + K^*(A - BK^*)^{-1}B\|\,,$$

and $\Gamma = AQ^{-1}A^\top + BR^{-1}B^\top$, $\Delta_A = (A - BK^*)^{-1}\bar{A}$ and $\Delta_B = (A - BK^*)^{-1}\bar{B}$.

Proposition 8. *(Existence and Uniqueness of Nash Equilibrium). Under Assumption 7, the operator $\Lambda(\cdot) = \Lambda_2(\cdot, \Lambda_1(\cdot))$ is L_0-Lipschitz, where L_0 is given in Assumption 7. Moreover, there exists a unique Nash equilibrium pair (μ^*, π^*) of the MFG.*

Note that Assumption 7 (ii) essentially assumes a contractive mapping of $\Lambda(\cdot)$. This condition is solely for technical purpose, without which it is not clear to us whether Nash equilibrium for the MFG problem even exists. That being said, we hope to argue that the assumption is not very restrictive. For example, if we do not have the mean-field effect, i.e. $\bar{A} = \bar{B} = 0$, then $L_1 = L_2 = 0$ and the Lipschitz constant $L_0 = 0$. This reduces to the regular LQR problem for each agent and the Nash equilibrium exists trivially since no agents interact at all. This condition in some sense requires that the mean-field effect cannot be too strong to adversely affect the existence of the Nash equilibrium.

5.2. Policy Gradient Algorithm and Convergence

To achieve the Nash equilibrium, the natural algorithm is that (i) for any given mean-field state/action μ_s, we solve the drifted LQR problem in (17) by policy gradient update until sufficient accuracy is achieved, say $J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi^*_{\mu_s}) \le \varepsilon_s$ where $\pi^*_{\mu_s} = \Lambda_1(\mu_s)$ and ε_s will be determined later; (ii) with the given π_{s+1}, we update the mean-field state/action μ_{s+1} by $\mu_{s+1} = \Lambda_2(\mu_s, \pi_{s+1})$ where the detailed formulas for $\Lambda_2(\cdot, \cdot)$ are provided in (19) (20). We summarize the above procedure in Algorithm 1.

Algorithm 1 Policy Gradient for Mean-Field Game

Input: Total number of iterations S, stepsize η, number of iterations N_s for each policy update;
Initial mean-field state/action $\mu_0 = (\mu_{0,x}^\top, \mu_{0,u}^\top)^\top$, initial policy π_0 with parameters K_{π_0} and b_{π_0}.
Output: Pair (π_S, μ_S).

```
for $s = 0, 1, \ldots, S-1$ do
    Policy Update:
    $K^0 = K_{\pi_s}$; $\tilde{a}_{\mu_s} \leftarrow \bar{A}\mu_{s,x} + \bar{B}\mu_{s,u}$;
    for $n = 0, 1, \ldots, N_s - 1$ do
        $K^{n+1} \leftarrow K^n - 2\eta(RK^n - B^\top P_{K^n})\Sigma_{K^n}$;
    end for
    $K_{\pi_{s+1}} \leftarrow K^{N_s}$;
    $b_{\pi_{s+1}} = -(K_{\pi_{s+1}} Q^{-1} A^\top + R^{-1} B^\top)(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}\tilde{a}_{\mu_s}$;
    $\pi_{s+1}(x) = -K_{\pi_{s+1}} x + b_{\pi_{s+1}}$;
    Mean-Field State/Action Update:
    $\mu_{s+1,x} \leftarrow -(A - BK_{\pi_{s+1}})^{-1}(Bb_{\pi_{s+1}} + \tilde{\alpha}_{\mu_s})$;
    $\mu_{s+1,u} \leftarrow b_{\pi_{s+1}} + K_{\pi_{s+1}}(A - BK_{\pi_{s+1}})^{-1}(Bb_{\pi_{s+1}} + \tilde{\alpha}_{\mu_s})$;
end for
```

The following theorem shows the linear convergence of Algorithm 1 to the MFG Nash equilibrium. The proof is deferred to Appendix C in the supplementary material.

Theorem 9. *(Convergence of Algorithm 1). For a sufficiently small tolerance $0 < \varepsilon < 1$, we choose the number of iterations S in Algorithm 1 such that*

$$S \ge \frac{\log(2\|\mu_0 - \mu^*\|_2 \cdot \varepsilon^{-1})}{\log(1/L_0)}\,.$$

For any $s = 0, 1, \ldots, S-1$, define

$$\varepsilon_s = \min\Big\{2^{-2}\|B\|_2^{-2}\|(A - BK^*)^{-1}\|_2^{-2}, C_b(\mu_s)^{-2}\varepsilon^2,$$
$$2^{-2s-4}(L_3 C_b(\mu_s) + 2C_K(\mu_2))^{-2}\varepsilon^2, \varepsilon^2\Big\}$$
$$\times\, \sigma_{\min}(R)\sigma_{\min}(DD^\top)\,,$$

where

$$C_b(\mu_s) = \|Q^{-1}A^\top(AQ^{-1}A^\top + BR^{-1}B^\top)^{-1}\tilde{a}_{\mu_s}\|_2\,,$$

$$C_K(\mu_s) = \Big(\|\tilde{\alpha}_{\mu_s}\|_2 + (1 + L_1\|\mu_s\|_2)\|B\|_2\Big)$$
$$\times \Big((1 + \|K^*\|_2)\|(A - BK^*)^{-1}\|_2^2\|B\|_2$$
$$+ \|(A - BK^*)^{-1}\|_2\Big)\,.$$

Assume $A - BK_{\pi_0}$ is stable. In the s-th policy update, we choose η as in Theorem 3 and number of iterations

$$N_s \ge \frac{\|\Sigma_{K^*}\|}{\eta\sigma_{\min}^2(DD^\top)\sigma_{\min}(R)} \log\frac{J_{\mu_s,1}(K_{\pi_s}) - J_{\mu_s,1}(K^*)}{\varepsilon_s},$$

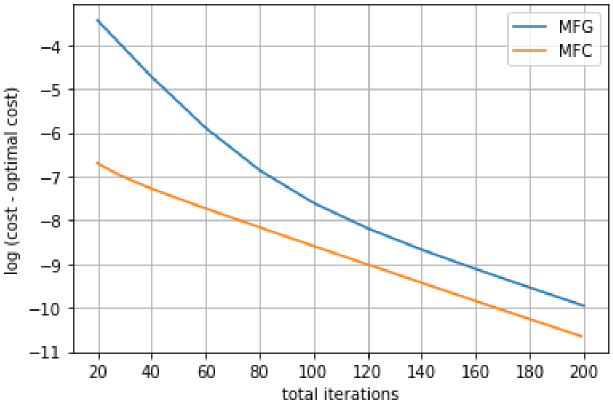

Figure 1. Linear Convergence of Policy Gradient for the MFC and the MFG. The orange curve for the MFC uses the initial values $K = 0, L = 0$, the learning rate $\eta = 0.01$ and plots $\log(J(K,L) - J(K^*,L^*))$ against the iterations $n = 1, 2, \ldots, 200$. The blue curve for the MFG runs Algorithm 1 with the initial values $K = 0, b = 0, \mu_x = 0.5(1,1,1)^\top, \mu_u = 0.5$, the learning rate $\eta = 0.005$, the total number of iterations $S = 10$ for the outer loop, and for each $s = 1, \ldots, 10$, the number of iterations $N_s = 20$ for the inner policy gradient updates. It plots $\log(J_{\mu_s}(K_{\pi_s}, b_{\pi_s}) - J_{\mu^*}(K^*, b^*))$ against $s = 1, 2, \ldots, 10$.

such that $J_{\mu_s}(K_{\pi_{s+1}}, b_{\pi_{s+1}}) - J_{\mu_s}(K^*, b^*_{\mu_s}) \le \varepsilon_s$ where $K^*, b^*_{\mu_s}$ are parameters of the optimal policy $\pi^*_{\mu_s} = \Lambda_1(\mu_s)$ generated from the mean-field state/action μ_s, $J_{\mu_s}(K_\pi, b_\pi) = J_{\mu_s}(\pi)$ is defined in the drifted MFG problem (17)*, and $J_{\mu_s,1}(K_\pi)$ is defined in* (14) *corresponding to $J_{\mu_s}(K_\pi, b_\pi)$. Then it holds that*

$$\|\mu_S - \mu^*\|_2 \le \varepsilon, \qquad \|K_{\pi_S} - K^*\|_F \le \varepsilon,$$

$$\|b_{\pi_S} - b^*\|_2 \le (1 + L_1)\varepsilon.$$

Here μ^* is the Nash mean-field state/action, K_{π_S}, b_{π_S} are parameters of the final output policy π_S, and K^*, b^* are the parameteris of the Nash policy $\pi^* = \Lambda_1(\mu^*)$.

Theorem 9 shows the linear convergence of the proposed Algorithm 1. This confirms that for the continuous-time MFG, policy gradient can achieve the ideal linear convergence performance in finding the Nash equilibrium. This lays an important theoretical foundation for applying modern reinforcement learning techniques to the general continuous mean-field game.

6. Simulation for Mean-Field Control / Game

In this section, we provide numerical results to demonstrate the linear convergence of the policy gradient algorithms for the mean-field control and game, and make an empirical comparison of them. We consider the following setting:

$$A = \begin{pmatrix} -1 & 0.1 & -0.05 \\ 0.05 & -1 & -0.05 \\ 0 & 0 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} -0.5 \\ -0.5 \\ 0.8 \end{pmatrix},$$

and $\bar{A} = -0.5A, \bar{B} = -0.5B, D = \bar{D} = I_3, Q = 0.1I_3, \bar{Q} = 0.05I_3, R = 1, \bar{R} = 2$. We can manually check that the conditions in Assumption 7 hold. Firstly, the continuous-time Riccati equation indeed has the following unique solution

$$P^* = \begin{pmatrix} 0.049798 & 0.003367 & -0.000802 \\ 0.003367 & 0.049963 & -0.000824 \\ -0.000802 & -0.000824 & 0.049272 \end{pmatrix}.$$

The second condition also holds with $L_1 = 0.0458, L_2 = 0.8752, L_3 = 2.0201$ and $L_0 = 0.9678 < 1$.

For the MFC, we start iterations from $K = 0, L = 0$, which are indeed stabilizing. We chose $\eta = 0.01$ and let the PG run for $N = 200$ updates. The linear convergence can be clearly seen from the orange curve of Figure 1, where we plot $\log(J(K,L) - J(K^*,L^*))$ against $n = 1, 2, \ldots, N$. For the MFG, we start iterations from $K = 0, b = 0, \mu_x = 0.5(1,1,1)^\top, \mu_u = 0.5$, and set $\eta = 0.005$, the total number of iterations $S = 10$ for the outer loop, and for each $s = 1, \ldots, 10$ the number of iterations $N_s = 20$ for the inner policy gradient updates. The blue curve of Figure 1 shows $\log(J_{\mu_s}(K_{\pi_s}, b_{\pi_s}) - J_{\mu^*}(K^*, b^*))$ against total iterations sN_s for $s = 1, 2, \ldots, 10$. The linear convergence of the algorithm matches well with our theoretical results. Note that here $J(K, L)$ is the cost of the MFC problem (10), while $J_{\mu_s}(K_{\pi_s}, b_{\pi_s})$ is the cost of the drifted LQR problem (17) corresponding to the MFG. It is not hard to calculate that $J(K^*, L^*) = 0.5986$ and $J_{\mu^*}(K^*, b^*) = 0.2981$, where J_{μ^*} is smaller as it ignores the dynamics of the conditional mean $\mathbb{E}_0[X_t], \mathbb{E}_0[u_t]$.

Since the mean-field control and game share the same model dynamics and cost function, we compare the total cost they achieve in Figure 2. As the target of the MFC is indeed minimizing the total cost, the effective control of policy gradient guarantees that the cost of the MFC (orange curve) converges to the optimal level at a linear rate. However, each agent of the MFG only cares about minimizing the cost with a given estimate of the mean-field state/action, i.e. solving the drifted LQR problem. Even when the estimate μ_s gets very close to the optimal μ^* and the Nash equilibrium is approximately obtained, the total cost of the MFG (blue curve) is still much larger than the optimal level. This is expected in MFG; obviously, agents have no control over the mean-field state and do not have access to $\bar{Q}, \bar{R}$ at all.

Last but not least, we empirically study the impact of L_0 on the convergence rate. Let us now take $\bar{A} = \theta A, \bar{B} = \theta B$. From Algorithm 1, it is not hard to calculate $\tilde{a}_{\mu_{s+1}} =$

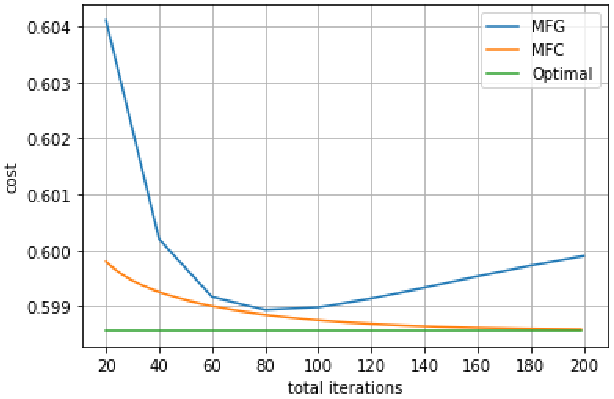

Figure 2. Total Cost of the MFC and the MFG. The cost of the MFC (orange curve) converges to the optimal level (green line) at a linear rate, while the cost of the MFG (blue curve) fails to converge to the optimal level, although Nash equilibrium has been reached (Figure 1 shows the convergence). For the MFG, we get the cost every $N_s = 20$ inner policy gradient iterations.

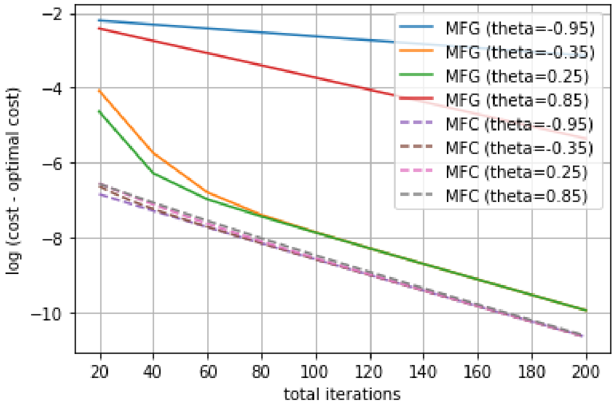

Figure 3. Impact of L_0 on Linear Convergence of Policy Gradient for MFC and MFG. We consider $\bar{A} = \theta A, \bar{B} = \theta B$ where $\theta \in \{-0.95, -0.35, 0.25, 0.85\}$, correspondingly $L_0 \in \{1.84, 0.68, 0.48, 1.65\}$. For each θ, the same cost curves as Figure 1 are plotted.

$\bar{A}\mu_{s+1,x} + \bar{B}\mu_{s+1,u} = \theta(A\mu_{s+1,x} + B\mu_{s+1,u}) = -\theta\tilde{a}_{\mu_s}$. So in order for the algorithm to converge, we require $\theta \in (-1, 1)$, where L_0 can be as large as $1.9 > 1$. This tells us $L_0 < 1$ in Assumption 7 is only a sufficient condition to guarantee linear convergence of PG for MFG and may be further relaxed; however, we cannot expect arbitrarily large L_0 for the convergence theory to hold. In Figure 3, we plot the same cost curves of MFC/MFG as Figure 1 for $\theta \in \{-0.95, -0.35, 0.25, 0.85\}$, correspondingly $L_0 \in \{1.84, 0.68, 0.48, 1.65\}$. As expected, the convergence of MFC almost has no change with different θ's or L_0's. For MFG, smaller L_0 (and L_1) leads to faster convergence (both smaller intercept and steeper slope) as Theorem 9 reveals.

7. Discussions and Conclusions

The paper aims to study the policy gradient method for the continuous-time MFC and MFG problems under the same framework. Specifically, we provided the linear convergence of PG algorithm for each problem setting. Although the paper is theory-oriented, we demonstrated the theory through a simple simulation, and made the comparison between the mean-field control and game. The key observation is that the MFG accumulates a larger total cost compared to the MFC, although the Nash equilibrium has been reached.

We hope to comment on the limitations of the current work.

- Firstly, our paper focuses on model-based PG for clearness of theoretical analysis. Extension to model-free is rather standard by zeroth-order optimization. Similar to Algorithm 1 of (Fazel et al., 2018) and Algorithm 2 of (Carmona et al., 2019), we can estimate $\nabla_K J(K)$ by $m^{-1}\sum_{i=1}^m \widehat{J}(K + U_i) \cdot U_i$, where U_i is a random matrix and $\widehat{J}(K + U_i)$ is an estimator of $J(K + U_i)$. Since the trajectory of the perturbed policy $K + U_i$ mixes exponentially fast, $J(K + U_i)$ can be well estimated using the time-averaged cost. See (Fazel et al., 2018; Fu et al., 2019) for more details.

- Secondly, We focus on the LQ setting as an initial step towards understanding PG for continuous-time control. The advantage of LQ setting is that we can derive PG in closed form. However, the analysis can be generalized beyond LQ setting. See discussions under Lemma 2 for gradient dominated cost function. In addition, other variations of the MFC and the MFG can be considered for future research, including the risk-sensitive mean-field setting (Tembine et al., 2013), the robust mean-field games (Bauso et al., 2012) and the mean-field models with partially observed information (Saldi et al., 2019).

- Thirdly, in mean-field approximation, we do need to assume permutation invariance among the agents. When permutation invariance fails, we have to deal with a multi-agent problem, whose complexity grows exponentially as the number of agents grows. Mean-field approximation is a common tool for alleviating such a curse of dimensionality, and is widely applied to problems in traffic control and finance. Such an assumption may be relaxed when the agents are divided into multiple groups and within each group the agents are identical, see e.g. (Bensoussan et al., 2018).

Finally, we emphasize again that directly studying continuous control is indeed a practical need, especially for applications in robotics and physics, where naive discretization could lead to exponentially large discretization error. Our contribution is to show for the first time PG works well with the continuous-time control, which fills the theoretical gap, at least for the most basic setting of LQR, MFC and MFG. Our analysis of PG convergence could be generalized to other continuous-time control problems whose objective functions satisfy benign properties.

Acknowledgements

Zhaoran Wang's work is supported by National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma Investments. Zhuoran Yang would like to acknowledge Simons Institute (Theory of Reinforcement Learning) for their research support. We would also like to thank all the anonymous reviewers for their helpful suggestions and supportive comments.

References

Anderson, B. D. O. and Moore, J. B. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

Araki, B., Strang, J., Pohorecky, S., Qiu, C., Naegeli, T., and Rus, D. Multi-robot path planning for a swarm of robots that can both fly and drive. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5575–5582. IEEE, 2017.

Bauso, D., Tembine, H., and Başar, T. Robust mean field games with application to production of an exhaustible resource. *IFAC Proceedings Volumes*, 45(13):454–459, 2012.

Bensoussan, A., Frehse, J., Yam, P., et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Bensoussan, A., Huang, T., and Laurière, M. Mean field control and mean field game models with several populations, 2018.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Bu, J., Ratliff, L. J., and Mesbahi, M. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

Cardaliaguet, P. and Hadikhanloo, S. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

Carmona, R., Delarue, F., et al. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Carmona, R., Laurière, M., and Tan, Z. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

de Cote, E. M., Lazaric, A., and Restelli, M. Learning to cooperate in multi-agent social dilemmas. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pp. 783–785, 2006.

Dimarogonas, D. V. and Johansson, K. H. Stability analysis for multi-agent systems using the incidence matrix: Quantized communication and formation control. *Automatica*, 46(4):695–700, 2010.

Egerstedt, M. and Hu, X. Formation constrained multi-agent control. *IEEE transactions on robotics and automation*, 17(6):947–951, 2001.

Elie, R., Perolat, J., Laurière, M., Geist, M., and Pietquin, O. On the convergence of model free learning in mean field games. In *AAAI Conference one Artificial Intelligence (AAAI 2020)*, 2020.

Elliott, R., Li, X., and Ni, Y.-H. Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, 49(11):3222–3233, 2013.

Engwerda, J. *LQ dynamic optimization and differential games*. John Wiley & Sons, 2005.

Fang, J. The LQR controller design of two-wheeled self-balancing robot based on the particle swarm optimization algorithm. *Mathematical Problems in Engineering*, 2014, 2014.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Fu, Z., Yang, Z., Chen, Y., and Wang, Z. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*, 2019.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. In *Advances in Neural Information Processing Systems*, pp. 4967–4977, 2019.

Heinrich, J. and Silver, D. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Huang, J. and Li, N. Linear–quadratic mean-field game for stochastic delayed systems. *IEEE Transactions on Automatic Control*, 63(8):2722–2729, 2018.

Huang, M., Caines, P. E., and Malhamé, R. P. Individual and mass behaviour in large population stochastic wireless power control problems: centralized and Nash equilibrium solutions. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, volume 1, pp. 98–103. IEEE, 2003.

Huang, M., Malhamé, R. P., Caines, P. E., et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3): 221–252, 2006.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pp. 3326–3336, 2018.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9): 619–625, 2006a.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. ii–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, 343(10):679–684, 2006b.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Minciardi, R. and Sacile, R. Optimal control in a cooperative network of smart power grids. *IEEE Systems Journal*, 6 (1):126–133, 2011.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

Parsons, S. and Wooldridge, M. Game theory and decision theory in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 5(3):243–254, 2002.

Saldi, N., Basar, T., and Raginsky, M. Markov-Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Saldi, N., Başar, T., and Raginsky, M. Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

Semsar-Kazerooni, E. and Khorasani, K. Multi-agent team cooperation: A game theory approach. *Automatica*, 45 (10):2205–2213, 2009.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Shamma, J. *Cooperative control of distributed multi-agent systems*. John Wiley & Sons, 2008.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 387–395, Bejing, China, 22–24 Jun 2014. PMLR.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Tembine, H., Zhu, Q., and Başar, T. Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, 59(4):835–850, 2013.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, pp. 2, 2019.

Willems, J. Least squares stationary optimal control and the algebraic riccati equation. *IEEE Transactions on Automatic Control*, 16(6):621–634, 1971.

Yang, E. and Gu, D. Multiagent reinforcement learning for multi-robot systems: A survey. Technical report, tech. rep, 2004.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pp. 11598–11610, 2019.

Zhang, P. Some results on two-person zero-sum linear quadratic differential games. *SIAM journal on control and optimization*, 43(6):2157–2165, 2005.

Zhou, X. Y. and Li, D. Continuous-time mean-variance portfolio selection: A stochastic lq framework. *Applied Mathematics and Optimization*, 42(1):19–33, 2000.