ACTOR-CRITIC PROVABLY FINDS NASH EQUILIBRIA OF LINEAR-QUADRATIC MEAN-FIELD GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study discrete-time mean-field Markov games with infinite numbers of agents where each agent aims to minimize its ergodic cost. We consider the setting where the agents have identical linear state transitions and quadratic cost functions, while the aggregated effect of the agents is captured by the population mean of their states, namely, the mean-field state. For such a game, based on the Nash certainty equivalence principle, we provide sufficient conditions for the existence and uniqueness of its Nash equilibrium. Moreover, to find the Nash equilibrium, we propose a mean-field actor-critic algorithm with linear function approximation, which does not require knowing the model of dynamics. Specifically, at each iteration of our algorithm, we use the single-agent actor-critic algorithm to approximately obtain the optimal policy of the each agent given the current mean-field state, and then update the mean-field state. In particular, we prove that our algorithm converges to the Nash equilibrium at a linear rate. To the best of our knowledge, this is the first success of applying model-free reinforcement learning with function approximation to discrete-time mean-field Markov games with provable non-asymptotic global convergence guarantees.

1 INTRODUCTION

In reinforcement learning (RL) (Sutton and Barto, 2018), an agent learns to make decisions that minimize its expected total cost through sequential interactions with the environment. Multi-agent reinforcement learning (MARL) (Shoham et al., 2003; 2007; Busoniu et al., 2008) aims to extend RL to sequential decision-making problems involving multiple agents. In a non-cooperative game, we are interested in the Nash equilibrium (Nash, 1951), which is a joint policy of all the agents such that each agent cannot decrease its expected total cost by unilaterally deviating from its Nash policy. The Nash equilibrium plays a critical role in understanding the social dynamics of self-interested agents (Ash, 2000; Axtell, 2002) and constructing the optimal policy of a particular agent via fictitious self-play (Bowling and Veloso, 2000; Ganzfried and Sandholm, 2009). With the recent development in deep learning (LeCun et al., 2015), MARL with function approximation achieves tremendous empirical successes in applications, including Go (Silver et al., 2016; 2017), Poker (Heinrich and Silver, 2016; Moravčík et al., 2017), Star Craft (Vinyals et al., 2019), Dota (OpenAI, 2018), autonomous driving (Shalev-Shwartz et al., 2016), multi-robotic systems (Yang and Gu, 2004), and solving social dilemmas (de Cote et al., 2006; Leibo et al., 2017; Hughes et al., 2018). However, since the capacity of the joint state and action spaces grows exponentially in the number of agents, such MARL approaches become computationally intractable when the number of agents is large, which is common in real-world applications (Sandholm, 2010; Calderone, 2017; Wang et al., 2017a).

Mean-field game is proposed by Huang et al. (2003; 2006); Lasry and Lions (2006a;b; 2007) with the idea of utilizing mean-field approximation to model the strategic interactions within a large population. In a mean-field game, each agent has the same cost function and state transition, which depend on the other agents only through their aggregated effect. As a result, the optimal policy of each agent depends solely on its own state and the aggregated effect of the population, and such an optimal policy is symmetric across all the agents. Moreover, if the aggregated effect of the population corresponds to the Nash equilibrium, then the optimal policy of each agent jointly constitutes a Nash equilibrium. Although such a Nash equilibrium corresponds to an infinite number of agents, it well approximates the Nash equilibrium for a sufficiently large number of agents (Bensoussan et al.,

2016). Also, as the aggregated effect of the population abstracts away the strategic interactions between individual agents, it circumvents the computational intractability of the MARL approaches that do not exploit symmetry.

However, most existing work on mean-field games focuses on characterizing the existence and uniqueness of the Nash equilibrium rather than designing provably efficient algorithms. In particular, most existing work considers the continuous-time setting, which requires solving a pair of Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck (FP) equations, whereas the discrete-time setting is more common in practice, e.g., in the aforementioned applications. Moreover, most existing approaches, including the ones based on solving the HJB and FP equations, require knowing the model of dynamics (Bardi and Priuli, 2014), or having the access to a simulator, which generates the next state given any state-action pair and aggregated effect of the population (Guo et al., 2019), which is often unavailable in practice.

To address these challenges, we develop an efficient model-free RL approach to mean-field game, which provably attains the Nash equilibrium. In particular, we focus on discrete-time mean-field games with linear state transitions and quadratic cost functions, where the aggregated effect of the population is quantified by the mean-field state. Such games capture the fundamental difficulties of general mean-field games and well approximates a variety of real-world systems such as power grids (Minciardi and Sacile, 2011), swarm robots (Fang, 2014; Araki et al., 2017; Doerr et al., 2018), and financial systems (Zhou and Li, 2000; Huang and Li, 2018). In detail, based on the Nash certainty equivalence (NCE) principle (Huang et al., 2006; 2007), we propose a mean-field actor-critic algorithm which, at each iteration, given the mean-field state μ, approximately attains the optimal policy π_μ^* of each agent, and then updates the mean-field state μ assuming that all the agents follow π_μ^*. We parametrize the actor and critic by linear and quadratic functions, respectively, and prove that such a parameterization encompasses the optimal policy of each agent. Specifically, we update the actor parameter using policy gradient (Sutton et al., 2000) and natural policy gradient (Kakade, 2002; Peters and Schaal, 2008; Bhatnagar et al., 2009) and update the critic parameter using primal-dual gradient temporal difference (Sutton et al., 2009a;b). In particular, we prove that given the mean-field state μ, the sequence of policies generated by the actor converges linearly to the optimal policy π_μ^*. Moreover, when alternatingly update the policy and mean-field state, we prove that the sequence of policies and its corresponding sequence of mean-field states converge to the unique Nash equilibrium at a linear rate. Our approach can be interpreted from both "passive" and "active" perspectives: (i) Assuming that each self-interested agent employs the single-agent actor-critic algorithm, the policy of each agent converges to the unique Nash policy, which characterizes the social dynamics of a large population of model-free RL agents. (ii) For a particular agent, our approach serves as a fictitious self-play method for it to find its Nash policy, assuming the other agents give their best responses. To the best of our knowledge, our work establishes the first efficient model-free RL approach with function approximation that provably attains the Nash equilibrium of a discrete-time mean-field game. As a byproduct, we also show that the sequence of policies generated by the single-agent actor-critic algorithm converges at a linear rate to the optimal policy of a linear-quadratic regulator (LQR) problem in the presence of drift, which may be of independent interest.

Related Work. Mean-field game is first introduced in Huang et al. (2003; 2006); Lasry and Lions (2006a;b; 2007). In the last decade, there is growing interest in understanding continuous-time mean-field games. See, e.g., Guéant et al. (2011); Bensoussan et al. (2013); Gomes et al. (2014); Carmona and Delarue (2013; 2018) and the references therein. Due to their simple structures, continuous-time linear-quadratic mean-field games are extensively studied under various model assumptions. See Li and Zhang (2008); Bardi (2011); Wang and Zhang (2012); Bardi and Priuli (2014); Huang et al. (2016a;b); Bensoussan et al. (2016; 2017); Caines and Kizilkale (2017); Huang and Huang (2017); Moon and Başar (2018); Huang and Zhou (2019) for examples of this line of work. Meanwhile, the literature on discrete-time linear-quadratic mean-field games remains relatively scarce. Most of this line of work focuses on characterizing the existence of a Nash equilibrium and the behavior of such a Nash equilibrium when the number of agents goes to infinity (Gomes et al., 2010; Tembine and Huang, 2011; Moon and Başar, 2014; Biswas, 2015; Saldi et al., 2018a;b; 2019). See also Yang et al. (2018a), which applies maximum entropy inverse RL (Ziebart et al., 2008) to infer the cost function and social dynamics of discrete-time mean-field games with finite state and action spaces. Our work is most related to Guo et al. (2019), where they propose a mean-field Q-learning algorithm (Watkins and Dayan, 1992) for discrete-time mean-field games

with finite state and action spaces. Such an algorithm requires the access to a simulator, which, given any state-action pair and mean-field state, outputs the next state. In contrast, both our state and action spaces are infinite, and we do not require such a simulator but only observations of trajectories. Correspondingly, we study the mean-field actor-critic algorithm with linear function approximation, whereas their algorithm is tailored to the tabular setting. Also, our work is closely related to Mguni et al. (2018), which focuses on a more restrictive setting where the state transition does not involve the mean-field state. In such a setting, mean-field games are potential games, which is, however, not true in more general settings (Li et al., 2017; Briani and Cardaliaguet, 2018). In comparison, we allow the state transition to depend on the mean-field state. Meanwhile, they propose a fictitious self-play method based on the single-agent actor-critic algorithm and establishes its asymptotic convergence. However, their proof of convergence relies on the assumption that the single-agent actor-critic algorithm converges to the optimal policy, which is unverified therein. In addition, our work is related to Jayakumar and Aditya (2019), where the proposed algorithm is only shown to converge asymptotically to a stationary point of the mean-field game.

Our work also extends the line of work on finding the Nash equilibria of Markov games using MARL. Due to the computational intractability introduced by the large number of agents, such a line of work focuses on finite-agent Markov games (Littman, 1994; 2001; Hu and Wellman, 1998; Bowling, 2001; Lagoudakis and Parr, 2002; Hu and Wellman, 2003; Conitzer and Sandholm, 2007; Perolat et al., 2015; Pérolat et al., 2016b;a; 2018; Wei et al., 2017; Zhang et al., 2018; Zou et al., 2019; Casgrain et al., 2019). See also Shoham et al. (2003; 2007); Busoniu et al. (2008); Li (2018) for detailed surveys. Our work is related to Yang et al. (2018b), where they combine the mean-field approximation of actions (rather than states) and Nash Q-learning (Hu and Wellman, 2003) to study general-sum Markov games with a large number of agents. However, the Nash Q-learning algorithm is only applicable to finite state and action spaces, and its convergence is established under rather strong assumptions. Also, when the number of agents goes to infinity, their approach yields a variant of tabular Q-learning, which is different from our mean-field actor-critic algorithm.

For policy optimization, based on the policy gradient theorem, Sutton et al. (2000); Konda and Tsitsiklis (2000) propose the actor-critic algorithm, which is later generalized to the natural actor-critic algorithm (Peters and Schaal, 2008; Bhatnagar et al., 2009). Most existing results on the convergence of actor-critic algorithms are based on stochastic approximation using ordinary differential equations (Bhatnagar et al., 2009; Castro and Meir, 2010; Konda and Tsitsiklis, 2000; Maei, 2018), which are asymptotic in nature. For policy evaluation, the convergence of primal-dual gradient temporal difference is studied in Liu et al. (2015); Du et al. (2017); Wang et al. (2017b); Yu (2017); Wai et al. (2018). However, this line of work assumes that the feature mapping is bounded, which is not the case in our setting. Thus, the existing convergence results are not applicable to analyzing the critic update in our setting. To handle the unbounded feature mapping, we utilize a truncation argument, which requires more delicate analysis.

Finally, our work extends the line of work that studies model-free RL for LQR. For example, Bradtke (1993); Bradtke et al. (1994) show that policy iteration converges to the optimal policy, Tu and Recht (2017); Dean et al. (2017) study the sample complexity of least-squares temporal-difference for policy evaluation. More recently, Fazel et al. (2018); Malik et al. (2018); Tu and Recht (2018) show that the policy gradient algorithm converges at a linear rate to the optimal policy. See as also Hardt et al. (2016); Dean et al. (2018) for more in this line of work. Our work is also closely related to Yang et al. (2019), where they show that the sequence of policies generated by the natural actor-critic algorithm enjoys a linear rate of convergence to the optimal policy. Compared with this work, when fixing the mean-field state, we use the actor-critic algorithm to study LQR in the presence of drift, which introduces significant difficulties in the analysis. As we show in §3, the drift causes the optimal policy to have an additional intercept, which makes the state- and action-value functions more complicated.

Notations. We denote by $\|M\|_*$ the spectral norm, $\rho(M)$ the spectral radius, $\sigma_{\min}(M)$ the minimum singular value, and $\sigma_{\max}(M)$ the maximum singular value of a matrix M. We use $\|\alpha\|_2$ to represent the ℓ_2-norm of a vector α, and $(\alpha)_i^j$ to denote the sub-vector $(\alpha_i, \alpha_{i+1}, \ldots, \alpha_j)^\top$, where α_k is the k-th entry of the vector α. For scalars $a_1, \ldots, a_n$, we denote by $\mathrm{poly}(a_1, \ldots, a_n)$ the polynomial of $a_1, \ldots, a_n$, and this polynomial may vary from line to line. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$.

2 LINEAR-QUADRATIC MEAN-FIELD GAME

A linear-quadratic mean-field game involves $N_{\mathrm{a}} \in \mathbb{N}$ agents. Their state transitions are given by

$$x_{t+1}^i = Ax_t^i + Bu_t^i + \overline{A} \cdot \frac{1}{N_{\mathrm{a}}} \sum_{j=1}^{N_{\mathrm{a}}} x_t^j + d^i + \omega_t^i, \qquad \forall t \geq 0,\ i \in [N_{\mathrm{a}}],$$

where $x_t^i \in \mathbb{R}^m$ and $u_t^i \in \mathbb{R}^k$ are the state and action vectors of agent i, respectively, the vector $d^i \in \mathbb{R}^m$ is a drift term, and $\omega_t^i \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$. The agents are coupled through the mean-field state $1/N_{\mathrm{a}} \cdot \sum_{j=1}^{N_{\mathrm{a}}} x_t^j$. In the linear-quadratic mean-field game, the cost of agent $i \in [N_{\mathrm{a}}]$ at time $t \geq 0$ is given by

$$c_t^i = (x_t^i)^\top Q x_t^i + (u_t^i)^\top R u_t^i + \left(\frac{1}{N_{\mathrm{a}}} \sum_{j=1}^{N_{\mathrm{a}}} x_t^j \right)^\top \overline{Q} \left(\frac{1}{N_{\mathrm{a}}} \sum_{j=1}^{N_{\mathrm{a}}} x_t^j \right),$$

where u_t^i follows π^i, i.e., the policy of agent i. To measure the performance of agent i following its policy π^i under the influence of the other agents, we define the expected total cost of agent i as

$$J^i(\pi^1, \pi^2, \ldots, \pi^{N_{\mathrm{a}}}) = \lim_{T \to \infty} \mathbb{E} \left(\frac{1}{T} \sum_{t=0}^{T} c_t^i \right).$$

We are interested in finding a Nash equilibrium $(\pi^1, \pi^2, \ldots, \pi^{N_{\mathrm{a}}})$, which is defined by

$$J^i(\pi^1, \ldots, \pi^{i-1}, \pi^i, \pi^{i+1}, \ldots, \pi^{N_{\mathrm{a}}}) \leq J^i(\pi^1, \ldots, \pi^{i-1}, \widetilde{\pi}^i, \pi^{i+1}, \ldots, \pi^{N_{\mathrm{a}}}), \qquad \forall \widetilde{\pi}^i,\ i \in [N_{\mathrm{a}}].$$

That is, agent i cannot further decrease its expected total cost by unilaterally deviating from its Nash policy.

For the simplicity of discussion, we assume that the drift term d^i is identical for each agent. By the symmetry of the agents in terms of their state transitions and cost functions, we focus on a fixed agent and drop the superscript i hereafter. Further taking the infinite-population limit $N_{\mathrm{a}} \to \infty$ leads to the following definition of linear-quadratic mean-field game (LQ-MFG).

Problem 2.1 (LQ-MFG)**.** We consider the following formulation

$$\begin{aligned}
&x_{t+1} = Ax_t + Bu_t + \overline{A}\mathbb{E}x_t^* + d + \omega_t, \\
&c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + (\mathbb{E}x_t^*)^\top \overline{Q} (\mathbb{E}x_t^*), \\
&J(\pi) = \lim_{T \to \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=0}^{T} c(x_t, u_t) \right],
\end{aligned}$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\{x_t^*\}_{t \geq 0}$ is the trajectory generated by a Nash policy π^* (assuming it exists), $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. Here the expectation $\mathbb{E}x_t^*$ is taken over the identical agents. We aim to find π^* such that $J(\pi^*) = \inf_{\pi \in \Pi} J(\pi)$.

The formulation in Problem 2.1 is studied by Lasry and Lions (2007); Bensoussan et al. (2016); Saldi et al. (2018a;b). We propose a more general formulation in Problem C.1 (see §C of the appendix for details), where an additional interaction term between the state vector x_t and the mean-field state $\mathbb{E}x_t^*$ is included in the cost function. According to our analysis in §C, up to minor modification, the results in the following sections also carry over to Problem C.1. Therefore, for the sake of simplicity, we focus on Problem 2.1 in the sequel.

Note that the mean-field state $\mathbb{E}x_t^*$ converges to a constant vector μ^* as $t \to \infty$, which serves as a fixed mean-field state, since the Markov chain of states generated by the Nash policy π^* admits a stationary distribution. As we consider the ergodic setting, it suffices to study Problem 2.1 with $t \to \infty$, which motivates the following drifted LQR (D-LQR) problem, where the mean-field state acts as another drift term.

Problem 2.2 (D-LQR)**.** Given a mean-field state $\mu \in \mathbb{R}^m$, we consider the following formulation

$$x_{t+1} = Ax_t + Bu_t + \overline{A}\mu + d + \omega_t,$$
$$c_\mu(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu,$$
$$J_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} c_\mu(x_t, u_t)\right],$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find an optimal policy π_μ^* such that $J_\mu(\pi_\mu^*) = \inf_{\pi\in\Pi} J_\mu(\pi)$.

For the mean-field state $\mu = \mu^*$, which corresponds to the Nash equilibrium, solving Problem 2.2 gives $\pi_{\mu^*}^*$, which coincides with the Nash policy π^* defined in Problem 2.1. Compared with the most studied LQR problem (Lewis et al., 2012), both the state transition and the cost function in Problem 2.2 have drift terms, which act as the mean-field "force" that drives the states away from zero. Such a mean-field "force" introduces additional challenges when solving Problem 2.2 in the model-free setting (see §3.3 for details). On the other hand, the unique optimal policy π_μ^* of Problem 2.2 admits a linear form $\pi_\mu^*(x_t) = -K_{\pi_\mu^*}x_t + b_{\pi_\mu^*}$ (Anderson and Moore, 2007), where the matrix $K_{\pi_\mu^*} \in \mathbb{R}^{k\times m}$ and the vector $b_{\pi_\mu^*} \in \mathbb{R}^k$ are the parameters of π_μ^*. Motivated by such a linear form of the optimal policy, we define the class of linear-Gaussian policies as

$$\Pi = \{\pi(x) = -Kx + b + \sigma\cdot\eta\colon K \in \mathbb{R}^{k\times m}, b \in \mathbb{R}^k\}, \tag{2.1}$$

where the standard Gaussian term $\eta \in \mathbb{R}^k$ is included to encourage exploration. To solve Problem 2.2, it then suffices to find the optimal policy within Π.

Now, we introduce the definition of the Nash equilibrium pair (Saldi et al., 2018a;b). The Nash equilibrium pair is characterized by the NCE principle, which states that it suffices to find a pair of π^* and μ^*, such that the policy π^* is optimal for each agent when the mean-field state is μ^*, and all the agents following the policy π^* generate the mean-field state 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 state μ, and $\Lambda_2(\mu, \pi)$ as the mean-field state generated by the policy π given the current mean-field state μ as $t \to \infty$.

Definition 2.3 (Nash Equilibrium Pair)**.** The pair $(\mu^*, \pi^*) \in \mathbb{R}^m \times \Pi$ constitutes a Nash equilibrium pair of Problem 2.1 if it satisfies $\pi^* = \Lambda_1(\mu^*)$ and $\mu^* = \Lambda_2(\mu^*, \pi^*)$. Here μ^* is the Nash mean-field state and π^* is the Nash policy.

3 Mean-Field Actor-Critic

We first characterize the existence and uniqueness of the Nash equilibrium pair of Problem 2.1 under mild regularity conditions, and then propose a mean-field actor-critic algorithm for such a Nash equilibrium. As a building block of the mean-field actor-critic algortihm, we propose a natural actor-critic algorithm to solve Problem 2.2.

3.1 Existence and Uniqueness of Nash Equilibrium Pair

We now establish the existence and uniqueness of the Nash equilibrium pair defined in Definition 2.3. We impose the following regularity conditions.

Assumption 3.1. We assume that the following statements hold:

(i) The algebraic Riccati equation $X = A^\top XA + Q - A^\top XB(B^\top XB + R)^{-1}B^\top XA$ admits a unique positive definite solution X^*;

(ii) It holds that $L_0 := L_1L_3 + L_2 < 1$, where

$$L_1 = \|[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top]^{-1}\overline{A}\|_* \cdot \|[K^*Q^{-1}(I-A)^\top - R^{-1}B^\top]\|_*,$$
$$L_2 = [1-\rho(A-BK^*)]^{-1}\|\overline{A}\|_*, \qquad L_3 = [1-\rho(A-BK^*)]^{-1}\|B\|_*.$$

Here $K^* = -(B^\top X^*B + R)^{-1}B^\top X^*A$.

The first assumption is implied by mild regularity conditions on the matrices A, B, Q, and R. See Theorem 3.2 in De Souza et al. (1986) for details. The second assumption is standard in the literature (Bensoussan et al., 2016; Saldi et al., 2018b), which ensures the stability of the mean-field game. In the following proposition, we show that Problem 2.1 admits a unique Nash equilibrium pair.

Proposition 3.2 (Existence and Uniqueness of Nash Equilibrium Pair)**.** Under Assumption 3.1, the operator $\Lambda(\cdot) = \Lambda_2(\cdot, \Lambda_1(\cdot))$ is L_0-Lipschitz, where L_0 is given in Assumption 3.1. Moreover, there exists a unique Nash equilibrium pair (μ^*, π^*) of Problem 2.1.

Proof. See §E.1 for a detailed proof. □

3.2 Mean-Field Actor-Critic for LQ-MFG

The NCE principle motivates a fixed-point approach to solve Problem 2.1, which generates a sequence of policies $\{\pi_s\}_{s\geq 0}$ and mean-field states $\{\mu_s\}_{s\geq 0}$ satisfying the following two conditions: (i) Given the mean-field state μ_s, the policy π_s is optimal. (ii) The mean-field state becomes μ_{s+1} if all the agents follow π_s under the current mean-field state μ_s. Here (i) requires solving Problem 2.2 given μ_s, while (ii) requires simulating the agents following π_s given the current mean-field state μ_s. Based on such properties, we propose the mean-field actor-critic algorithm in Algorithm 1.

Algorithm 1 Mean-Field Actor-Critic Algorithm for solving LQ-MFG.

1: **Input:**
- Initial mean-field state μ_0 and Initial policy π_0 with parameters K_0 and b_0.
- Numbers of iterations S, $\{N_s\}_{s\in[S]}$, $\{H_s\}_{s\in[S]}$, $\{\widetilde{T}_{s,n}, T_{s,n}\}_{s\in[S],n\in[N_s]}$, $\{\widetilde{T}^b_{s,h}, T^b_{s,h}\}_{s\in[S],h\in[H_s]}$.
- Stepsizes $\{\gamma_s\}_{s\in[S]}$, $\{\gamma^b_s\}_{s\in[S]}$, $\{\gamma_{s,n,t}\}_{s\in[S],n\in[N_s],t\in[T_{s,n}]}$, $\{\gamma^b_{s,h,t}\}_{s\in[S],h\in[H_s],t\in[T^b_{s,h}]}$.

2: **for** $s = 0, 1, 2, \ldots, S-1$ **do**

3: **Policy Update:** Solve for the optimal policy π_{s+1} with parameters K_{s+1} and b_{s+1} of Problem 2.2 via Algorithm 2 with μ_s, π_s, N_s, H_s, $\{\widetilde{T}_{s,n}, T_{s,n}\}_{n\in[N_s]}$, $\{\widetilde{T}^b_{s,h}, T^b_{s,h}\}_{h\in[H_s]}$, γ_s, γ^b_s, $\{\gamma_{s,n,t}\}_{n\in[N_s],t\in[T_{s,n}]}$, and $\{\gamma^b_{s,h,t}\}_{h\in[H_s],t\in[T^b_{s,h}]}$ as inputs, which gives the estimated mean-field state $\widehat{\mu}_{K_{s+1},b_{s+1}}$.

4: **Mean-Field State Update:** Update the mean-field state via $\mu_{s+1} \leftarrow \widehat{\mu}_{K_{s+1},b_{s+1}}$.

5: **end for**

6: **Output**: Pair (π_S, μ_S).

Algorithm 1 requires solving Problem 2.2 at each iteration to obtain $\pi_s \approx \Lambda_1(\mu_s)$ and $\mu_{s+1} \approx \Lambda_2(\mu_s, \pi_s)$. To this end, we introduce a mixed actor-critic algorithm in §3.3 that solves Problem 2.2.

3.3 Natural Actor-Critic for D-LQR

Now we focus on solving Problem 2.2 for a fixed mean-field state μ, we thus drop the subscript μ in this section. We write $\pi_{K,b}(x) = -Kx + b + \sigma\eta$ to emphasize the dependence on K and b, and $J(K, b) = J(\pi_{K,b})$ consequently. Now, we propose an actor-critic algorithm to solve Problem 2.2.

For any policy $\pi_{K,b} \in \Pi$, by the state transition in Problem 2.2, we have

$$x_{t+1} = (A - BK)x_t + (Bb + \overline{A}\mu + d) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Psi_\epsilon), \tag{3.1}$$

where $\Psi_\epsilon = \sigma BB^\top + \Psi_\omega$. It is known that if $\rho(A - BK) < 1$, then the Markov chain $\{x_t\}_{t\geq 0}$ induced by (3.1) has a unique stationary distribution $\rho_{K,b} = \mathcal{N}(\mu_{K,b}, \Phi_K)$ (Anderson and Moore, 2007), where the mean-field state $\mu_{K,b}$ and the covariance Φ_K satisfy that

$$\mu_{K,b} = (I - A + BK)^{-1}(Bb + \overline{A}\mu + d), \tag{3.2}$$

$$\Phi_K = (A - BK)\Phi_K(A - BK)^\top + \Psi_\epsilon. \tag{3.3}$$

Meanwhile, the Bellman equation for Problem 2.2 takes the following form

$$P_K = (Q + K^\top RK) + (A - BK)^\top P_K(A - BK). \tag{3.4}$$

Then by calculation (see Proposition B.2 in §B.1 of the appendix for details), it holds that the expected total cost $J(K, b)$ is decomposed as

$$J(K, b) = J_1(K) + J_2(K, b) + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q} \mu, \tag{3.5}$$

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

$$\begin{aligned} J_1(K) &= \mathrm{tr}\big[(Q + K^\top R K)\Phi_K\big] = \mathrm{tr}(P_K \Psi_\epsilon), \\ J_2(K, b) &= \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top R K & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}. \end{aligned} \tag{3.6}$$

Here $J_1(K)$ is the expected total cost in the most studied LQR problems (Yang et al., 2019; Fazel et al., 2018), where the state transition does not have drift terms. Meanwhile, $J_2(K, b)$ corresponds to the expected cost induced by the drift terms. The following two propositions characterize the properties of $J_2(K, b)$. First, we show that $J_2(K, b)$ is strongly convex in b.

Proposition 3.3. Given any K, the function $J_2(K, b)$ is ν_K-strongly convex in b. Here $\nu_K = \sigma_{\min}(Y_{1,K}^\top Y_{1,K} + Y_{2,K}^\top Y_{2,K})$, where $Y_{1,K} = R^{1/2} K (I - A + BK)^{-1} B - R^{1/2}$ and $Y_{2,K} = Q^{1/2}(I - A + BK)^{-1} B$. Also, $J_2(K, b)$ has ι_K-Lipschitz continuous gradient in b, where ι_K is upper bounded such that $\iota_K \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_*^2 \cdot \|K\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*)$.

Proof. See §E.4 for a detailed proof. □

Second, we show that $\min_b J_2(K, b)$ is independent of K.

Proposition 3.4. We define $b^K = \mathrm{argmin}_b J_2(K, b)$, where $J_2(K, b)$ is defined in (3.6). It holds that

$$b^K = \big[K Q^{-1}(I - A)^\top - R^{-1} B^\top\big] \cdot \big[(I - A) Q^{-1} (I - A)^\top + B R^{-1} B^\top\big]^{-1} \cdot (\overline{A}\mu + d).$$

Moreover, $J_2(K, b^K)$ takes the form

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top \big[(I - A) Q^{-1} (I - A)^\top + B R^{-1} B^\top\big]^{-1} (\overline{A}\mu + d),$$

which is independent of K.

Proof. See §E.2 for a detailed proof. □

We highlight that following from Proposition 3.3, we know that for any fixed K, the function $J_2(K, b)$ admits a unique minimizer b^K. Since $\min_b J_2(K, b)$ is independent of K by Proposition 3.4, it holds that the optimal K^* is the same as $\mathrm{argmin}_K J_1(K)$. This motivates us to minimize $J(K, b)$ by first updating K following the gradient direction $\nabla_K J_1(K)$ to the optimal K^*, then updating b following the gradient direction $\nabla_b J_2(K^*, b)$. We now design our algorithm based on this idea.

We define Υ_K, $p_{K,b}$, and $q_{K,b}$ as

$$\begin{aligned} \Upsilon_K &= \begin{pmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{pmatrix} = \begin{pmatrix} \Upsilon_K^{11} & \Upsilon_K^{12} \\ \Upsilon_K^{21} & \Upsilon_K^{22} \end{pmatrix}, \\ p_{K,b} &= A^\top \big[P_K \cdot (\overline{A}\mu + d) + f_{K,b}\big], \qquad q_{K,b} = B^\top \big[P_K \cdot (\overline{A}\mu + d) + f_{K,b}\big], \end{aligned} \tag{3.7}$$

where $f_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K (Bb + \overline{A}\mu + d) - K^\top R b]$. By calculation (see Proposition B.3 in §B.1 of the appendix for details), the gradients of $J_1(K)$ and $J_2(K, b)$ take the forms of $\nabla_K J_1(K) = 2(\Upsilon_K^{22} K - \Upsilon_K^{21})\Phi_K$ and $\nabla_b J_2(K, b) = \Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + q_{K,b}$.

Our algorithm follows the natural actor-critic method (Bhatnagar et al., 2009) and actor-critic method (Konda and Tsitsiklis, 2000). Specifically, to obtain the optimal K^*, in the critic update step, we estimate the matrix Υ_K by $\widehat{\Upsilon}_K$ via a policy evaluation algorithm, e.g., Algorithm 3 or Algorithm 4 (see §B.2 and §B.3 for details). In the actor update step, we update K via $K \leftarrow K - \gamma \cdot (\widehat{\Upsilon}_K^{22} K - \widehat{\Upsilon}_K^{21})$, where the term $\widehat{\Upsilon}_K^{22} K - \widehat{\Upsilon}_K^{21}$ is the estimated natural gradient. To obtain the optimal b^* given K^*, in the critic update step, we estimate Υ_{K^*}, $q_{K^*,b}$, and $\mu_{K^*,b}$ by $\widehat{\Upsilon}_K$, $\widehat{q}_{K,b}$, and $\widehat{\mu}_{K,b}$ via a policy evaluation algorithm. In the actor update step, we update b via $b \leftarrow b - \gamma \cdot \widehat{\nabla}_b J_2(K^*, b)$, where $\widehat{\nabla}_b J_2(K^*, b)$ is the estimated gradient. Then, we obtain the mixed actor-critic algorithm for Problem 2.2, which is stated in Algorithm 2.

Algorithm 2 Mixed Actor-Critic Algorithm for D-LQR.

1: **Input:**
- Mean-field state μ and initial policy π_{K_0,b_0}.
- Numbers of iterations N, H, $\{\widetilde{T}_n, T_n\}_{n\in[N]}$, $\{\widetilde{T}_h^b, T_h^b\}_{h\in[H]}$.
- Stepsizes γ, γ^b, $\{\gamma_{n,t}\}_{n\in[N],t\in[T_n]}$, $\{\gamma_{h,t}^b\}_{h\in[H],t\in[T_h^b]}$.

2: **for** $n = 0, 1, 2, \ldots, N-1$ **do**

3: **Critic Update:** Compute $\widehat{\Upsilon}_{K_n}$ via Algorithm 3 with π_{K_n,b_0}, μ, $\widetilde{T}_n$,T_n, $\{\gamma_{n,t}\}_{t\in[T_n]}$, K_0, and b_0 as inputs.

4: **Actor Update:** Update the parameter via $K_{n+1} \leftarrow K_n - \gamma \cdot (\widehat{\Upsilon}_{K_n}^{22} K_n - \widehat{\Upsilon}_{K_n}^{21})$.

5: **end for**

6: **for** $h = 0, 1, 2, \ldots, H-1$ **do**

7: **Critic Update:** Compute $\widehat{\mu}_{K_N,b_h}$, $\widehat{\Upsilon}_{K_N}$, $\widehat{q}_{K_N,b_h}$ via Algorithm 3 with π_{K_N,b_h}, μ, $\widetilde{T}_h^b$, T_h^b, $\{\gamma_{h,t}^b\}_{t\in[T_h^b]}$, K_0, and b_0 as inputs.

8: **Actor Update:** Update the parameter via

$$b_{h+1} \leftarrow b_h - \gamma^b \cdot \big[\widehat{\Upsilon}_{K_N}^{22}(-K_N\widehat{\mu}_{K,b_h} + b_h) + \widehat{\Upsilon}_{K_N}^{21}\widehat{\mu}_{K_N,b_h} + \widehat{q}_{K_N,b_h}\big].$$

9: **end for**

10: **Output:** Policy $\pi_{K,b} = \pi_{K_N,b_H}$, estimated mean-field state $\widehat{\mu}_{K,b} = \widehat{\mu}_{K_N,b_H}$.

4 Global Convergence Results

The following theorem establishes the convergence of Algorithm 1.

Theorem 4.1 (Convergence of Algorithm 1)**.** For a sufficiently small tolerance $\varepsilon > 0$, we set the number of iterations S in Algorithm 1 such that

$$S > \log\big(\|\mu_0 - \mu^*\|_2 \cdot \varepsilon^{-1}\big)/\log(1/L_0). \tag{4.1}$$

We define

$$\varepsilon_s = \min\Big\{\big[1-\rho(A-BK^*)\big]^4\big(\|B\|_* + \|\overline{A}\|_*\big)^{-4}\big(\|\mu_s\|_2^{-2} + \|d\|_2^{-2}\big)\cdot\sigma_{\min}(\Psi_\epsilon)\cdot\sigma_{\min}(R)\cdot\varepsilon^2,$$
$$\nu_{K^*}\cdot\big[1-\rho(A-BK_0)\big]^2\cdot\|B\|_*^{-2}\cdot\varepsilon^2,\ \varepsilon\Big\}\cdot 2^{-s-10}, \qquad s\in[S], \tag{4.2}$$

where ν_{K^*} is defined in Proposition 3.3. In the s-th policy update step in Line 3 of Algorithm 1, we set the inputs via Theorem B.4 such that $J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi_{\mu_s}^*) < \varepsilon_s$, where the expected total cost $J_{\mu_s}(\cdot)$ is defined in Problem 2.2, and $\pi_{\mu_s}^* = \Lambda_1(\mu_s)$ is the optimal policy under the mean-field state μ_s. Then it holds that $\|\mu_S - \mu^*\|_2 \le \varepsilon$ with probability at least $1-\varepsilon^5$, where μ_S is the output of Algorithm 1 and μ^* is the Nash mean-field state of Problem 2.1.

Proof. See §D.1 for a detailed proof. □

Here, we highlight that if the inputs of Algorithm 1 satisfy certain conditions, it holds that $J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi_{\mu_s}^*) < \varepsilon_s$ at the s-th iteration for any $s \in [S]$. See Theorem B.4 in §B.1 of the appendix for details. By Theorem 4.1, Algorithm 1 converges linearly to the unique Nash equilibrium pair (μ^*, π^*) of Problem 2.1. To the best of our knowledge, this theorem is the first successful attempt to establish that reinforcement learning with function approximation finds the Nash equilibrium pairs in mean-field games with theoretical guarantee, which lays the theoretical foundations for applying modern reinforcement learning techniques to general mean-field games.

5 Conclusion

For the discrete-time linear-quadratic mean-field games, we provide sufficient conditions for the existence and uniqueness of the Nash equilibrium pair. Moreover, we propose the mean-field actor-critic algorithm with linear function approximation that is proved to converge to the Nash equilibrium pair with linear rate of convergence. Our algorithm can be modified to use other parametrized function classes, including deep neural networks, for solving mean-field games.

REFERENCES

Alizadeh, F., Haeberly, J.-P. A. and Overton, M. L. (1998). Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, **8** 746–768.

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

Araki, B., Strang, J., Pohorecky, S., Qiu, C., Naegeli, T. and Rus, D. (2017). 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)*. IEEE.

Ash, C. (2000). Social-self-interest. *Annals of public and cooperative economics*, **71** 261–284.

Axtell, R. L. (2002). Non-cooperative dynamics of multi-agent teams. In *Autonomous Agents and Multiagent Systems*.

Bardi, M. (2011). Explicit solutions of some linear-quadratic mean field games. *Networks and heterogeneous media*, **7** 243–261.

Bardi, M. and Priuli, F. S. (2014). Linear-quadratic n-person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, **52** 3022–3052.

Bauso, D., Tembine, H. and Başar, T. (2016). Robust mean field games. *Dynamic games and applications*, **6** 277–303.

Bensoussan, A., Chau, M., Lai, Y. and Yam, S. C. P. (2017). Linear-quadratic mean field Stackelberg games with state and control delays. *SIAM Journal on Control and Optimization*, **55** 2748–2781.

Bensoussan, A., Frehse, J. and Yam, P. (2013). *Mean field games and mean field type control theory*. Springer.

Bensoussan, A., Sung, K., Yam, S. C. P. and Yung, S.-P. (2016). Linear-quadratic mean field games. *Journal of Optimization Theory and Applications*, **169** 496–529.

Bhandari, J., Russo, D. and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M. and Lee, M. (2009). Natural actor–critic algorithms. *Automatica*, **45** 2471–2482.

Biswas, A. (2015). Mean field games with ergodic cost for discrete time markov processes. *arXiv preprint arXiv:1510.08968*.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, **38** 447–469.

Bowling, M. (2001). Rational and convergent learning in stochastic games. In *International Conference on Artificial Intelligence*.

Bowling, M. and Veloso, M. (2000). An analysis of stochastic game theory for multiagent reinforcement learning. Tech. rep., Carnegie Mellon University.

Bradtke, S. J. (1993). Reinforcement learning applied to linear quadratic regulation. In *Advances in Neural Information Processing Systems*.

Bradtke, S. J., Ydstie, B. E. and Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *American Control Conference*, vol. 3. IEEE.

Briani, A. and Cardaliaguet, P. (2018). Stable solutions in potential mean field game systems. *Nonlinear Differential Equations and Applications*, **25** 1.

Busoniu, L., Babuska, R. and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, **38** 156–172.

Caines, P. E. and Kizilkale, A. C. (2017). ϵ-nash equilibria for partially observed LQG mean field games with a major player. *IEEE Transactions on Automatic Control*, **62** 3225–3234.

Calderone, D. J. (2017). *Models of Competition for Intelligent Transportation Infrastructure: Parking, Ridesharing, and External Factors in Routing Decisions*. University of California, Berkeley.

Carmona, R. and Delarue, F. (2013). Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, **51** 2705–2734.

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

Casgrain, P., Ning, B. and Jaimungal, S. (2019). Deep Q-learning for Nash equilibria: Nash-DQN. *arXiv preprint arXiv:1904.10554*.

Castro, D. D. and Meir, R. (2010). A convergent online single time scale actor critic algorithm. *Journal of Machine Learning Research*, **11** 367–410.

Conitzer, V. and Sandholm, T. (2007). AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, **67** 23–43.

de Cote, E. M., Lazaric, A. and Restelli, M. (2006). Learning to cooperate in multi-agent social dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems*. ACM.

De Souza, C., Gevers, M. and Goodwin, G. (1986). Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Transactions on Automatic Control*, **31** 831–838.

Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2017). On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*.

Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*.

Doerr, B., Linares, R., Zhu, P. and Ferrari, S. (2018). Random finite set theory and optimal control for large spacecraft swarms. *arXiv preprint arXiv:1810.00696*.

Du, S. S., Chen, J., Li, L., Xiao, L. and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

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

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

Ganzfried, S. and Sandholm, T. (2009). Computing equilibria in multiplayer stochastic games of imperfect information. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Gomes, D. A., Mohr, J. and Souza, R. R. (2010). Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, **93** 308–328.

Gomes, D. A. et al. (2014). Mean field games modelsa brief survey. *Dynamic Games and Applications*, **4** 110–154.

Guéant, O., Lasry, J.-M. and Lions, P.-L. (2011). Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*. Springer, 205–266.

Guo, X., Hu, A., Xu, R. and Zhang, J. (2019). Learning mean-field games. *arXiv preprint arXiv:1901.09585*.

Hardt, M., Ma, T. and Recht, B. (2016). Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*.

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

Hu, J. and Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.

Hu, J. and Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, **4** 1039–1069.

Huang, J. and Huang, M. (2017). Robust mean field linear-quadratic-gaussian games with unknown L^2-disturbance. *SIAM Journal on Control and Optimization*, **55** 2811–2840.

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

Huang, J., Li, X. and Wang, T. (2016a). Mean-field linear-quadratic-Gaussian (LQG) games for stochastic integral systems. *IEEE Transactions on Automatic Control*, **61** 2670–2675.

Huang, J., Wang, S. and Wu, Z. (2016b). Backward mean-field linear-quadratic-Gaussian (LQG) games: Full and partial information. *IEEE Transactions on Automatic Control*, **61** 3784–3796.

Huang, M., Caines, P. E. and Malhamé, R. P. (2003). Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *Conference on Decision and Control*. IEEE.

Huang, M., Caines, P. E. and Malhamé, R. P. (2007). Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. *IEEE transactions on automatic control*, **52** 1560–1571.

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

Huang, M. and Zhou, M. (2019). Linear quadratic mean field games: Asymptotic solvability and relation to the fixed point approach. *arXiv preprint arXiv:1903.08776*.

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. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*.

Jayakumar, S. and Aditya, M. (2019). Reinforcement learning in stationary mean-field games. In *International Conference on Autonomous Agents and Multiagent Systems*.

Kakade, S. M. (2002). A natural policy gradient. In *Advances in neural information processing systems*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*.

Korda, N. and La, P. (2015). On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media.

Lagoudakis, M. G. and Parr, R. (2002). Value function approximation in zero-sum Markov games. In *Uncertainty in Artificial Intelligence*.

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

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

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

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521** 436–444.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J. and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

Lewis, F. L., Vrabie, D. and Syrmos, V. L. (2012). *Optimal control*. John Wiley & Sons.

Li, S., Zhang, W. and Zhao, L. (2017). Connections between mean-field game and social welfare optimization. *arXiv preprint arXiv:1703.10211*.

Li, T. and Zhang, J.-F. (2008). Asymptotically optimal decentralized control for large population stochastic multiagent systems. *IEEE Transactions on Automatic Control*, **53** 1643–1660.

Li, Y. (2018). Deep reinforcement learning. *arXiv preprint arXiv:1810.06339*.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*. Elsevier, 157–163.

Littman, M. L. (2001). Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S. and Petrik, M. (2015). Finite-sample analysis of proximal gradient TD algorithms. In *Conference on Uncertainty in Artificial Intelligence*.

Maei, H. R. (2018). Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*.

Magnus, J. R. (1979). The expectation of products of quadratic forms in normal variables: the practice. *Statistica Neerlandica*, **33** 131–136.

Magnus, J. R. et al. (1978). *The moments of products of quadratic forms in normal variables*. Univ., Instituut voor Actuariaat en Econometrie.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P. L. and Wainwright, M. J. (2018). Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*.

Mguni, D., Jennings, J. and de Cote, E. M. (2018). Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Moon, J. and Başar, T. (2014). Discrete-time LQG mean field games with unreliable communication. In *Conference on Decision and Control*. IEEE.

Moon, J. and Başar, T. (2018). Linear quadratic mean field stackelberg differential games. *Automatica*, **97** 200–213.

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

Nash, J. (1951). Non-cooperative games. *Annals of mathematics* 286–295.

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

Pérolat, J., Piot, B., Geist, M., Scherrer, B. and Pietquin, O. (2016a). Softened approximate policy iteration for markov games. In *International Conference on Machine Learning*.

Pérolat, J., Piot, B. and Pietquin, O. (2018). Actor-critic fictitious play in simultaneous move multi-stage games. In *International Conference on Artificial Intelligence and Statistics*.

Pérolat, J., Piot, B., Scherrer, B. and Pietquin, O. (2016b). On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*.

Perolat, J., Scherrer, B., Piot, B. and Pietquin, O. (2015). Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning (ICML 2015)*.

Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, **71** 1180–1190.

Rudelson, M., Vershynin, R. et al. (2013). Hanson-wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, **18**.

Saldi, N., Basar, T. and Raginsky, M. (2018a). Discrete-time risk-sensitive mean-field games. *arXiv preprint arXiv:1808.03929*.

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

Saldi, N., Basar, T. and Raginsky, M. (2019). Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*.

Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. MIT Press.

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

Shoham, Y., Powers, R. and Grenager, T. (2003). Multi-agent reinforcement learning: a critical survey.

Shoham, Y., Powers, R. and Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, **171** 365–377.

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. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** 484–489.

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

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

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C. and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009b). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*.

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

Sznitman, A.-S. (1991). Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX1989*. Springer, 165–251.

Tembine, H. and Huang, M. (2011). Mean field difference games: Mckean-Vlasov dynamics. In *Conference on Decision and Control and European Control Conference*. IEEE.

Tembine, H., Zhu, Q. and Başar, T. (2014). Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, **59** 835–850.

Tu, S. and Recht, B. (2017). Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*.

Tu, S. and Recht, B. (2018). The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*.

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

Wai, H.-T., Yang, Z., Wang, P. Z. and Hong, M. (2018). Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*.

Wang, B.-C. and Zhang, J.-F. (2012). Mean field games for large-population multiagent systems with Markov jump parameters. *SIAM Journal on Control and Optimization*, **50** 2308–2334.

Wang, J., Zhang, W., Yuan, S. et al. (2017a). Display advertising with real-time bidding (RTB) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, **11** 297–435.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M. and Liu, T.-Y. (2017b). Finite sample analysis of the GTD policy evaluation algorithms in Markov setting. In *Advances in Neural Information Processing Systems*.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, **8** 279–292.

Wei, C.-Y., Hong, Y.-T. and Lu, C.-J. (2017). Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*.

Yang, E. and Gu, D. (2004). Multiagent reinforcement learning for multi-robot systems: A survey. *Manuscript*.

Yang, J., Ye, X., Trivedi, R., Xu, H. and Zha, H. (2018a). Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W. and Wang, J. (2018b). Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*.

Yang, Z., Chen, Y., Hong, M. and Wang, Z. (2019). On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*.

Yu, H. (2017). On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*.

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

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

Ziebart, B. D., Maas, A. L., Bagnell, J. A. and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, vol. 3.

Zou, S., Xu, T. and Liang, Y. (2019). Finite-sample analysis for SARSA and Q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234*.

A NOTATIONS IN THE PROOFS

In the proof, for convenience, for any invertible matrix M, we denote by $M^{-\top} = (M^{-1})^\top = (M^\top)^{-1}$ and $\|M\|_{\mathrm{F}}$ the Frobenius norm. We also denote by $\mathrm{svec}(M)$ the symmetric vectorization of the symmetric matrix M, which is the vectorization of the upper triangular matrix of the symmetric matrix M, with off-diagonal entries scaled by $\sqrt{2}$. Also, we denote by $\mathrm{smat}(\cdot)$ the inverse operation.

For notational simplicity, we write $z = [x^\top, u^\top]^\top$ as the combination of the state vector x and the action vector u. Also, we write $\mathbb{E}_\pi[\cdot]$ to emphasize that the expectation is taken following the policy π.

B AUXILIARY ALGORITHMS AND ANALYSIS

B.1 RESULTS IN D-LQR

In this section, we provide auxiliary results in analyzing Problem 2.2. First, we introduce the value functions of the Markov decision process (MDP) induced by Problem 2.2. We define the state-value function $V_{K,b}(x)$ as follows

$$V_{K,b}(x) = \sum_{t=0}^{\infty} \Big\{ \mathbb{E}\big[c(x_t, u_t) \,|\, x_0 = x, u_t = -Kx_t + b + \sigma\eta_t\big] - J(K,b) \Big\}, \tag{B.1}$$

and the action-value function $Q_{K,b}(x,u)$ as follows

$$Q_{K,b}(x,u) = c(x,u) - J(K,b) + \mathbb{E}\big[V_{K,b}(x') \,|\, x, u\big], \tag{B.2}$$

where the x' is the state generated by the state transition after the state-action pair (x,u). We establish the close forms of these value functions in the following proposition.

Proposition B.1. The state-value function $V_{K,b}(x)$ takes the form

$$V_{K,b}(x) = x^\top P_K x - \mathrm{tr}(P_K \Phi_K) + 2 f_{K,b}^\top (x - \mu_{K,b}) - (\mu_{K,b})^\top P_K \mu_{K,b}, \tag{B.3}$$

and the action-value function $Q_{K,b}(x,u)$ takes the form

$$\begin{aligned} Q_{K,b}(x,u) = {} & \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2 \begin{pmatrix} p_{K,b} \\ q_{K,b} \end{pmatrix}^\top \begin{pmatrix} x \\ u \end{pmatrix} - \mathrm{tr}(P_K \Phi_K) - \sigma^2 \cdot \mathrm{tr}(R + P_K B B^\top) - b^\top R b \\ & + 2 b^\top R K \mu_{K,b} - (\mu_{K,b})^\top (Q + K^\top R K + P_K) \mu_{K,b} + 2 f_{K,b}^\top \big[(\overline{A}\mu + d) - \mu_{K,b}\big] \\ & + (\overline{A}\mu + d)^\top P_K (\overline{A}\mu + d), \end{aligned} \tag{B.4}$$

where the matrix Υ_K and the vectors $p_{K,b}, q_{K,b}$ are given in (3.7).

Proof. See §E.6 for a detailed proof. □

By Proposition B.1, we know that $V_{K,b}(x)$ is quadratic in x, while $Q_{K,b}(x,u)$ is quadratic in x and u. Now, we show that $J(K,b)$ is decomposed as (3.5).

Proposition B.2. The expected total cost $J(K,b)$ defined in Problem 2.2 is decomposed as

$$J(K,b) = J_1(K) + J_2(K,b) + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q} \mu,$$

where $J_1(K)$ and $J_2(K,b)$ take the following forms

$$\begin{aligned} J_1(K) &= \mathrm{tr}\big[(Q + K^\top R K)\Phi_K\big] = \mathrm{tr}(P_K \Psi_\epsilon), \\ J_2(K,b) &= \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top R K & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}. \end{aligned}$$

Here $\mu_{K,b}$ is given in (3.2), Φ_K is given in (3.3), and P_K is given in (3.4).

Proof. See §E.3 for a detailed proof. □

We establish the gradients of $J(K, b)$ in the following proposition.

Proposition B.3. The gradient of $J_1(K)$ and the gradient of $J_2(K, b)$ with respect to b take the forms

$$\nabla_K J_1(K) = 2(\Upsilon_K^{22} K - \Upsilon_K^{21}) \cdot \Phi_K, \qquad \nabla_b J_2(K, b) = 2\big[\Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + q_{K,b}\big],$$

where the matrix Υ_K and the vector $q_{K,b}$ are given in (3.7).

Proof. See §E.5 for a detailed proof. □

Equipped with above results, the following theorem establishes the convergence of Algorithm 2.

Theorem B.4 (Convergence of Algorithm 2)**.** Let the initial policy π_{K_0,b_0} be stable. In Algorithm 2, for a sufficiently small tolerance $\varepsilon > 0$, we set

$$\gamma \leq [\|R\|_* + \|B\|_*^2 \cdot J(K_0, b_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon)]^{-1}$$
$$N \geq C \cdot \|\Phi_{K^*}\|_* \cdot \gamma^{-1} \cdot \log\Big\{4\big[J(K_0, b_0) - J(K^*, b^*)\big] \cdot \varepsilon^{-1}\Big\},$$

where C is a positive absolute constant. Let $\{K_n\}_{n\in[N]}$ and $\{b_h\}_{h\in[H]}$ be the sequences of parameters generated by Algorithm 2. In the n-th critic update step in Line 3 of Algorithm 2, we set

$$T_n \geq \text{poly}\big(\|K_n\|_{\text{F}}, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)\big) \cdot \lambda_{K_n}^{-4} \cdot \big[1 - \rho(A - BK_n)\big]^{-9} \cdot \varepsilon^{-5},$$
$$\widetilde{T}_n \geq \text{poly}\big(\|K_n\|_{\text{F}}, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)\big) \cdot \lambda_{K_n}^{-2} \cdot \big[1 - \rho(A - BK_n)\big]^{-12} \cdot \varepsilon^{-12},$$
$$\gamma_{n,t} = \gamma_0 \cdot t^{-1/2},$$

where γ_0 is some positive constant, and λ_{K_n} is specified in Proposition B.6. Also, after obtaining K_N from Algorithm 2, we set

$$\gamma^b \leq \min\{1 - \rho(A - BK_N), [1 - \rho(A - BK_N)]^{-2} \cdot (\|B\|_*^2 \cdot \|K_N\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*)\},$$
$$H \geq C_0 \cdot \nu_{K_N}^{-1} \cdot (\gamma^b)^{-1} \cdot \log\Big\{4\big[J(K_N, b_0) - J(K_N, b^{K_N})\big] \cdot \varepsilon^{-1}\Big\},$$

where C_0 is some positive absolute constant and the number ν_{K_N} is specified in Proposition 3.3. In the h-th critic update step in Line 7 of Algorithm 2, we set

$$T_h^b \geq \text{poly}\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \lambda_{K_N}^{-4} \cdot \nu_{K_N}^{-4} \cdot \big[1 - \rho(A - BK_N)\big]^{-10} \cdot \varepsilon^{-5},$$
$$\widetilde{T}_h^b \geq \text{poly}\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J(K_n, b_0)\big) \cdot \lambda_{K_N}^{-4} \cdot \nu_{K_N}^{-2} \cdot \big[1 - \rho(A - BK_N)\big]^{-16} \cdot \varepsilon^{-8},$$
$$\gamma_{h,t}^b = \gamma_0 \cdot t^{-1/2},$$

where ν_{K_N} is specified in Proposition 3.3. Then with probability at least $1 - \varepsilon^{10}$, it holds that $J(K_N, b_H) - J(K^*, b^*) < \varepsilon$, $\|K_N - K^*\|_{\text{F}} \leq [\sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \varepsilon]^{1/2}$, and $\|b_H - b^*\|_2 \leq (2\varepsilon/\nu_{K^*})^{1/2}$.

Proof. See §D.2 for a detailed proof. □

By this theorem, given any mean-field state μ, we show that Algorithm 2 converges linearly to the optimal policy π_μ^* of Problem 2.2. In the proof of Theorem B.4, we use the convergence results of Algorithm 3, which is provided in §B.2 of the appendix.

B.2 Primal-Dual Policy Evaluation Algorithm

Note that critic update steps in Algorithm 2 need estimators of the matrix Υ_K and the vector $q_{K,b}$. We now derive a policy evaluation algorithm to estimate these quantities, based on gradient temporal difference algorithm (Sutton et al., 2009a).

For notational convenience, we denote by the feature vector

$$\psi(x, u) = \begin{pmatrix} \varphi(x, u) \\ x - \mu_{K,b} \\ u - (-K\mu_{K,b} + b) \end{pmatrix}, \tag{B.5}$$

where

$$\varphi(x,u) = \mathrm{svec}\left[\begin{pmatrix} x-\mu_{K,b} \\ u-(-K\mu_{K,b}+b)\end{pmatrix}\begin{pmatrix} x-\mu_{K,b} \\ u-(-K\mu_{K,b}+b)\end{pmatrix}^\top\right].$$

Recall $\mathrm{svec}(M)$ gives the symmetric vectorization of the symmetric matrix M. We also denote by the vector

$$\alpha_{K,b} = \begin{pmatrix} \mathrm{svec}(\Upsilon_K) \\ \Upsilon_K\begin{pmatrix}\mu_{K,b} \\ -K\mu_{K,b}+b\end{pmatrix} + \begin{pmatrix} p_{K,b} \\ q_{K,b}\end{pmatrix}\end{pmatrix}, \tag{B.6}$$

where the matrix Υ_K and vectors $p_{K,b}$ and $q_{K,b}$ are given in (3.7). Note that to estimate Υ_K and $q_{K,b}$, we only need to estimate $\alpha_{K,b}$. We also define the following matrix

$$\Theta_{K,b} = \mathbb{E}_{\pi_{K,b}}\Big\{\psi(x,u)\big[\psi(x,u)-\psi(x',u')\big]^\top\Big\}, \tag{B.7}$$

where (x',u') is the state-action pair after (x,u) following the policy $\pi_{K,b}$ and the state transition. The matrix $\Theta_{K,b}$ defined in (B.7) connects with the vector $\alpha_{K,b}$ by the following proposition.

Proposition B.5. The following equation holds:

$$\begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b}\end{pmatrix}\begin{pmatrix} J(K,b) \\ \alpha_{K,b}\end{pmatrix} = \begin{pmatrix} J(K,b) \\ \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\end{pmatrix},$$

where the vector $\alpha_{K,b}$ is given in (B.6), the matrix $\Theta_{K,b}$ is given in (B.7), and the feature vector $\psi(x,u)$ is defined in (B.5).

Proof. See §E.7 for a detailed proof. □

Motivated by Proposition B.5, to obtain the vector $\alpha_{K,b}$, we only need to solve the following linear system in $\zeta = (\zeta_1, \zeta_2^\top)^\top$

$$\widetilde{\Theta}_{K,b}\cdot\zeta = \begin{pmatrix} J(K,b) \\ \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\end{pmatrix}, \tag{B.8}$$

where the matrix $\widetilde{\Theta}_{K,b}$ takes the form

$$\widetilde{\Theta}_{K,b} = \begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b}\end{pmatrix}.$$

Note that in the above linear system (B.8), if the matrix $\Theta_{K,b}$ is invertible, then the whole linear system admits the unique solution $\zeta_{K,b} = (J(K,b), \alpha_{K,b}^\top)^\top$. Under mild condition, the following proposition verifies the invertibility of $\Theta_{K,b}$.

Proposition B.6. If $\rho(A-BK) < 1$, then the matrix $\Theta_{K,b}$ is invertible, and its spectral norm is upper bounded by $4(1+\|K\|_{\mathrm{F}}^2)^2\cdot\|\Phi_K\|_*^2$. Moreover, $\sigma_{\min}(\widetilde{\Theta}_{K,b})$ is lower bounded by a positive constant λ_K, where λ_K only depends on σ, $\sigma_{\min}(\Psi_\omega)$ and $\rho(A-BK)$.

Proof. See §E.8 for a detailed proof. □

Hereafter, for notational convenience, we denote by $\widehat{\psi}_t$ the estimated feature vector $\widehat{\psi}(x_t,u_t)$. Now, we present the primal-dual gradient temporal difference algorithm in Algorithm 3.

Primal-Dual Gradient Method. In Algorithm 3, instead of solving the linear system (B.8) directly, we minimize the following loss function

$$\big[\zeta^1 - J(K,b)\big]^2 + \Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2.$$

Algorithm 3 Primal-Dual Gradient Temporal Difference Algorithm.

Input: Policy $\pi_{K,b}$, mean-field state μ, numbers of iteration $\widetilde{T}$ and T, stepsizes $\{\gamma_t\}_{t\in[T]}$, parameters K_0 and b_0.
Define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ via Condition B.7 with K_0 and b_0.
Initialize the parameters by $\zeta_0 \in \mathcal{V}_\zeta$ and $\xi_0 \in \mathcal{V}_\xi$.
Sample $\widetilde{x}_0$ from the the stationary distribution $\widetilde{\mathcal{N}}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, \widetilde{T} - 1$ **do**
  Given the mean-field state μ, take action $\widetilde{u}_t$ following $\pi_{K,b}$ and generate the next state $\widetilde{x}_{t+1}$.
end for
Set $\widehat{\mu}_{K,b} \leftarrow 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{x}_t$ and compute the estimated feature vector $\widehat{\psi}$ via (B.10).
Sample x_0 from the the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, T - 1$ **do**
  Given the mean-field state μ, take action u_t following $\pi_{K,b}$, observe the cost c_t, and generate the next state x_{t+1}.
  Set $\delta_{t+1} \leftarrow \zeta_t^1 + (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \zeta_t^2 - c_t$.
  Update parameters via

$$\zeta_{t+1}^1 \leftarrow \zeta_t^1 - \gamma_{t+1} \cdot (\xi_t^1 + \widehat{\psi}_t^\top \xi_t^2), \qquad \zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \widehat{\psi}_t (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \xi_t^2,$$
$$\xi_{t+1}^1 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^1 + \gamma_{t+1} \cdot (\zeta_t^1 - c_t), \qquad \xi_{t+1}^2 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^2 + \gamma_{t+1} \cdot \delta_{t+1} \cdot \widehat{\psi}_t.$$

  Project ζ_{t+1} and ξ_{t+1} to $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$, respectively.
end for
Set $\widehat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^T \gamma_t)^{-1} \cdot (\sum_{t=1}^T \gamma_t \cdot \zeta_t^2)$, and

$$\widehat{\Upsilon}_K \leftarrow \mathrm{smat}(\widehat{\alpha}_{K,b,1}), \qquad \begin{pmatrix} \widehat{p}_{K,b} \\ \widehat{q}_{K,b} \end{pmatrix} \leftarrow \widehat{\alpha}_{K,b,2} - \widehat{\Upsilon}_K \begin{pmatrix} \widehat{\mu}_{K,b} \\ -K\widehat{\mu}_{K,b} + b \end{pmatrix},$$

where $\widehat{\alpha}_{K,b,1} = (\widehat{\alpha}_{K,b})_1^{(k+d+1)(k+d)/2}$ and $\widehat{\alpha}_{K,b,2} = (\widehat{\alpha}_{K,b})_{(k+d+1)(k+d)/2+1}^{(k+d+3)(k+d)/2}$.

Output: Estimators $\widehat{\mu}_{K,b}$, $\widehat{\Upsilon}_K$, and $\widehat{q}_{K,b}$.

Further, by Fenchel's duality, we convert the above problem to the following primal-dual min-max problem

$$\min_{\zeta\in\mathcal{V}_\zeta}\max_{\xi\in\mathcal{V}_\xi} F(\zeta,\xi) = \Big\{\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\}^\top \xi^2 + \big[\zeta^1 - J(K,b)\big]\cdot\xi^1 - 1/2\cdot\|\xi\|_2^2, \tag{B.9}$$

where we restrict the primal variable ζ in some compact set $\mathcal{V}_\zeta$ and the dual variable ξ in some compact set $\mathcal{V}_\xi$, which are detailedly specified in Condition B.7. By taking the gradient of the objective in (B.9) with respect to ζ and ξ, we obtain the following gradients

$$\nabla_{\zeta^1}F = \xi^1 + \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]^\top\xi^2, \qquad \nabla_{\zeta^2}F = \Theta_{K,b}^\top\xi^2,$$
$$\nabla_{\xi^1}F = \zeta^1 - J(K,b) - \xi^1, \qquad \nabla_{\xi^2}F = \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] - \xi^2.$$

This gives a primal-dual gradient method to solve the min-max problem (B.9).

Estimate of Mean State $\mu_{K,b}$. Note that in the definition of the feature vector in (B.5), we need the mean $\mu_{K,b}$ of the state to evaluate the feature vector $\psi(x,u)$. Therefore, prior to the main body of Algorithm 3, we obtain the estimator $\widehat{\mu}_{K,b}$ by simulating the MDP under the policy $\pi_{K,b}$ for $\widetilde{T}$ iterations, and compute the estimated feature vector $\widehat{\psi}$ via

$$\widehat{\psi}(x,u) = \begin{pmatrix}\widehat{\varphi}(x,u)\\ x-\widehat{\mu}_{K,b}\\ u-(-K\widehat{\mu}_{K,b}+b)\end{pmatrix}, \tag{B.10}$$

where $\widehat{\varphi}(x,u)$ takes the following form

$$\widehat{\varphi}(x,u) = \mathrm{svec}\left[\begin{pmatrix}x-\widehat{\mu}_{K,b}\\ u-(-K\widehat{\mu}_{K,b}+b)\end{pmatrix}\begin{pmatrix}x-\widehat{\mu}_{K,b}\\ u-(-K\widehat{\mu}_{K,b}+b)\end{pmatrix}^\top\right].$$

Before stating the convergence result of Algorithm 3, we specify the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$.

Condition B.7 . Given K_0 and b_0 such that $\rho(A-BK_0)<1$. We define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ in (B.9) as

$$\mathcal{V}_\zeta = \Big\{\zeta\colon 0\le\zeta^1\le J(K_0,b_0), \|\zeta^2\|_2\le M_{\zeta,1}+M_{\zeta,2}\cdot(1+\|K\|_{\mathrm{F}})\cdot\big[1-\rho(A-BK)\big]^{-1}\Big\},$$
$$\mathcal{V}_\xi = \Big\{\xi\colon |\xi^1|\le J(K_0,b_0), \|\xi^2\|_2\le M_\xi\cdot\big(1+\|K\|_{\mathrm{F}}^2\big)^3\cdot\big[1-\rho(A-BK)\big]^{-1}\Big\}.$$

Here $M_{\zeta,1}$, $M_{\zeta,2}$ and M_ξ are constants independent of K and b, which take the following forms

$$\begin{aligned}M_{\zeta,1} = &\Big[\big(\|Q\|_{\mathrm{F}}+\|R\|_{\mathrm{F}}\big)+\big(\|A\|_{\mathrm{F}}^2+\|B\|_{\mathrm{F}}^2\big)\cdot\sqrt{d}\cdot J(K_0,b_0)/\sigma_{\min}(\Psi_\omega)\Big]\\ &+\Big[\big(\|Q\|_*+\|R\|_*\big)+\big(\|A\|_*+\|B\|_*\big)^2\cdot J(K_0,b_0)/\sigma_{\min}(\Psi_\omega)\Big]\\ &\quad\cdot\big[J(K_0,b_0)/\sigma_{\min}(Q)+J(K_0,b_0)/\sigma_{\min}(R)\big]\\ &+\big(\|A\|_*+\|B\|_*\big)\cdot J(K_0,b_0)^2\cdot\sigma_{\min}^{-1}(\Psi_\omega)\cdot\sigma_{\min}^{-1}(Q),\\ M_{\zeta,2} = &\big(\|A\|_*+\|B\|_*\big)\cdot(\kappa_Q+\kappa_R), \qquad M_\xi = C\cdot(M_{\zeta,1}+M_{\zeta,2})\cdot J(K_0,b_0)^2/\sigma_{\min}^2(Q),\end{aligned}$$

where C is a positive absolute constant, and κ_Q and κ_R are condition numbers of the matrices Q and R, respectively.

We now characterize the convergence of Algorithm 3.

Theorem B.8 (Convergence of Algorithm 3)**.** Given K_0, b_0, K and b such that $\rho(A-BK_0)<1$ and $J(K,b)\le J(K_0,b_0)$, we define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ through Condition B.7. Given the stepsize $\gamma_t = \gamma_0\cdot t^{-1/2}$ for some constant $\gamma_0>0$, then for any ρ such that $\rho(A-BK)<\rho<1$, when the number of iterations $\widetilde{T}$ and T are sufficiently large such that $\widetilde{T}\le\mathrm{poly}_0(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0))\cdot(1-\rho)^{-6}$, with probability at least $1-T^{-4}-\widetilde{T}^{-6}$, the output $\widehat{\alpha}_{K,b}$ satisfies that

$$\|\widehat{\alpha}_{K,b}-\alpha_{K,b}\|_2^2\le\lambda_K^{-2}\cdot\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big)\cdot\left[\frac{\log^6 T}{T^{1/2}\cdot(1-\rho)^4}+\frac{\log\widetilde{T}}{\widetilde{T}^{1/4}\cdot(1-\rho)^2}\right],$$

where λ_K is specified in Proposition B.6. Same bounds for $\|\widehat{\Upsilon}_K - \Upsilon_K\|_{\mathrm{F}}$, $\|\widehat{p}_{K,b} - p_{K,b}\|_2$ and $\|\widehat{q}_{K,b} - q_{K,b}\|_2$ hold. Also, with probability at least $1 - \widetilde{T}^{-6}$, it holds that

$$\|\widehat{\mu}_{K,b} - \mu_{K,b}\|_2 \leq \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\big(\|\Phi_K\|_*, \|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\big).$$

Proof. See §D.3 for a detailed proof. □

The policy evaluation algorithm is used in the critic update steps in Algorithm 2. Therefore, the above theorem is crucial in the proof of the corresponding convergence result stated in Theorem B.4.

B.3 Temporal Difference Policy Evaluation Algorithm

We also apply TD(0) method (Sutton and Barto, 2018) to evaluate the policy, which is presented in Algorithm 4.

Algorithm 4 Temporal Difference Policy Evaluation Algorithm.

Input: Policy $\pi_{K,b}$, number of iteration $\widetilde{T}$ and T, stepsizes $\{\gamma_t\}_{t\in[T]}$.
Sample $\widetilde{x}_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, \widetilde{T} - 1$ **do**
  Take action $\widetilde{u}_t$ under the policy $\pi_{K,b}$ and generate the next state $\widetilde{x}_{t+1}$.
end for
Set $\widehat{\mu}_{K,b} \leftarrow 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{x}_t$.
Sample x_0 from the the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, T$ **do**
  Given the mean-field state μ, take action $\widetilde{u}_t$ following $\pi_{K,b}$ and generate the next state $\widetilde{x}_{t+1}$.
  Set $\delta_{t+1} \leftarrow \zeta_t^1 + (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \zeta_t^2 - c_t$.
  Update parameters via $\zeta_{t+1}^1 \leftarrow (1-\gamma_{t+1}) \cdot \zeta_t^1 + \gamma_{t+1} \cdot c_t$ and $\zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \delta_{t+1} \cdot \widehat{\psi}_t$.
  Project ζ_t to some compact set $\mathcal{V}_\zeta'$.
end for
Set $\widehat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^T \gamma_t)^{-1} \cdot (\sum_{t=1}^T \gamma_t \cdot \zeta_t^2)$, and

$$\widehat{\Upsilon}_K \leftarrow \mathrm{smat}(\widehat{\alpha}_{K,b,1}), \qquad \begin{pmatrix} \widehat{p}_{K,b} \\ \widehat{q}_{K,b} \end{pmatrix} \leftarrow \widehat{\alpha}_{K,b,2} - \widehat{\Upsilon}_K \begin{pmatrix} \widehat{\mu}_{K,b} \\ -K\widehat{\mu}_{K,b} + b \end{pmatrix},$$

where $\widehat{\alpha}_{K,b,1} = (\widehat{\alpha}_{K,b})_1^{(k+d+1)(k+d)/2}$ and $\widehat{\alpha}_{K,b,2} = (\widehat{\alpha}_{K,b})_{(k+d+1)(k+d)/2+1}^{(k+d+3)(k+d)/2}$.
Output: Estimators $\widehat{\mu}_{K,b}$, $\widehat{\Upsilon}_K$, and $\widehat{q}_{K,b}$.

Note that in related literature (Bhandari et al., 2018; Korda and La, 2015), non-asymptotic convergence analysis of TD(0) method with linear function approximation is only applied to discounted MDP; as for our ergodic setting, the convergence of TD(0) method is only shown asymptotically (Borkar and Meyn, 2000; Kushner and Yin, 2003) using ordinary differential equation method. Therefore, in the convergence theorem proposed in §3, we only focus on the primal-dual gradient temporal difference method (Algorithm 3) to establish non-asymptotic convergence result. However, in practice, we still use TD(0) algorithm.

C General Formulation

Compared with Problem 2.1, a more general formulation includes an additional term $x_t^\top P \mathbb{E} x_t^*$ in the cost function. For the completeness of this paper, we define this general formulation here.

Problem C.1 (General LQ-MFG)**.** We consider the following formulation

$$\begin{aligned}
&x_{t+1} = Ax_t + Bu_t + \overline{A}\mathbb{E}x_t^* + d + \omega_t,\\
&\widetilde{c}(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + (\mathbb{E}x_t^*)^\top \overline{Q}(\mathbb{E}x_t^*) + 2x_t^\top P(\mathbb{E}x_t^*),\\
&\widetilde{J}(\pi) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} \widetilde{c}(x_t, u_t)\right],
\end{aligned}$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\{x_t^*\}_{t\geq 0}$ is the trajectory generated by a Nash policy π^* (assuming it exists), $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. Here the expectation in $\mathbb{E}x_t^*$ is taken over the identical agents. We aim to find π^* such that $\widetilde{J}(\pi^*) = \inf_{\pi\in\Pi} \widetilde{J}(\pi)$.

Following similar discussion as in §2, it suffices to study Problem C.1 with $t \to \infty$, which motivates the following general drifted LQR (general D-LQR) problem.

Problem C.2 (General D-LQR)**.** For any given mean-field state $\mu \in \mathbb{R}^m$, consider the following formulation

$$\begin{aligned}
&x_{t+1} = Ax_t + Bu_t + \overline{A}\mu + d + \omega_t,\\
&\widetilde{c}_\mu(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu + 2x_t^\top P\mu,\\
&\widetilde{J}_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} \widetilde{c}_\mu(x_t, u_t)\right],
\end{aligned}$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find an optimal policy π_μ^* such that $\widetilde{J}_\mu(\pi_\mu^*) = \inf_{\pi\in\Pi} \widetilde{J}_\mu(\pi)$.

In Problem C.2, the unique optimal policy $\pi_\mu^*(\cdot)$ still admits a linear form $\pi_\mu^*(x_t) = -K_{\pi_\mu^*}x_t + b_{\pi_\mu^*}$ (Anderson and Moore, 2007), where the matrix $K_{\pi_\mu^*} \in \mathbb{R}^{k\times m}$ and the vector $b_{\pi_\mu^*} \in \mathbb{R}^k$ are the parameters of the policy π. It then suffices to find the optimal policy in the class Π introduced in (2.1). Similar to §3.3, we drop the subscript μ when we focus on Problem C.2 for a fixed μ. We write $\pi_{K,b}(x) = -Kx + b + \sigma\eta$ to emphasize the dependence on K and b, and $\widetilde{J}(K, b) = \widetilde{J}(\pi_{K,b})$ consequently. We derive an explicit form of the expected total cost $\widetilde{J}(K, b)$ in the following proposition.

Proposition C.3. The expected total cost $\widetilde{J}(K, b)$ in Problem C.2 is decomposed as

$$\widetilde{J}(K, b) = \widetilde{J}_1(K) + \widetilde{J}_2(K, b) + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q}\mu,$$

where $\widetilde{J}_1(K)$ and $\widetilde{J}_2(K, b)$ take the following forms

$$\begin{aligned}
&\widetilde{J}_1(K) = \mathrm{tr}\big[(Q + K^\top RK)\Phi_K\big] = \mathrm{tr}(P_K\Psi_\epsilon),\\
&\widetilde{J}_2(K, b) = \begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}^\top \begin{pmatrix} Q + K^\top RK & -K^\top R\\ -RK & R\end{pmatrix}\begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix} + 2\mu^\top P\mu_{K,b}.
\end{aligned}$$

Here $\mu_{K,b}$ is given in (3.2), Φ_K is given in (3.3), and P_K is given in (3.4).

Proof. The proof is similar to the one of Proposition B.2. Thus we omit it here. □

Compared with the form of $J(K, b)$ given in (3.5), we see that the only difference is that $\widetilde{J}(K, b)$ contains an extra term $2\mu^\top P\mu_{K,b}$ in $\widetilde{J}_2(K, b)$, which is only a linear term in b (recall that $\mu_{K,b}$ is linear in b by (3.2)). Thus, $\widetilde{J}_2(K, b)$ is still strongly convex in b, as shown in the proposition below.

Proposition C.4. Given any K, the function $\widetilde{J}_2(K, b)$ is ν_K-strongly convex in b, here $\nu_K = \sigma_{\min}(Y_{1,K}^\top Y_{1,K} + Y_{2,K}^\top Y_{2,K})$, where $Y_{1,K} = R^{1/2}K(I - A + BK)^{-1}B - R^{1/2}$ and $Y_{2,K} = Q^{1/2}(I - A + BK)^{-1}B$. Also, $\widetilde{J}_2(K, b)$ has ι_K-Lipschitz continuous gradient in b, where ι_K is upper bounded such that $\iota_K \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_*^2 \cdot \|K\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*)$.

Proof. The proof is similar to the one of Proposition 3.3. Thus we omit it here. □

Parallel to Proposition 3.4, we derive a similar proposition in the sequel.

Proposition C.5. Denote by $\widetilde{b}^K = \operatorname{argmin}_b \widetilde{J}_2(K, b)$, then $\widetilde{J}_2(K, \widetilde{b}^K)$ takes the form

$$\widetilde{J}_2(K, \widetilde{b}^K) = \begin{pmatrix} \overline{A}\mu + d \\ P^\top \mu \end{pmatrix}^\top \begin{pmatrix} S & S(I-A)Q^{-1} \\ Q^{-1}(I-A)^\top S & 3Q^{-1}(I-A)^\top S(I-A)Q^{-1} - Q^{-1} \end{pmatrix} \begin{pmatrix} \overline{A}\mu + d \\ P^\top \mu \end{pmatrix},$$

which is independent of K. Here $S = [(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top]^{-1}$. And $\widetilde{b}^K$ takes the form

$$\widetilde{b}^K = \big[KQ^{-1}(I-A)^\top - R^{-1}B^\top\big] \cdot S \cdot \big[(\overline{A}\mu + d) + (I-A)Q^{-1}P^\top \mu\big] - KQ^{-1}P^\top \mu.$$

Proof. The proof is similar to the one of Proposition 3.4. Thus we omit it here. □

Similar to Problem 2.2, we define the state- and action-value functions as

$$\widetilde{V}_{K,b}(x) = \sum_{t=0}^{\infty} \Big\{ \mathbb{E}\big[\widetilde{c}(x_t, u_t) \,|\, x_0 = x, u_t = -Kx_t + b + \sigma\eta_t\big] - \widetilde{J}(K, b) \Big\},$$
$$\widetilde{Q}_{K,b}(x, u) = \widetilde{c}(x, u) - \widetilde{J}(K, b) + \mathbb{E}\big[\widetilde{V}_{K,b}(x') \,|\, x, u\big],$$

where the x' is the state generated by the state transition after the state-action pair (x, u). A slight modification of Proposition B.1 gives the proposition below.

Proposition C.6. For Problem C.2, the state-value function $\widetilde{V}_{K,b}(x)$ takes the form

$$\widetilde{V}_{K,b}(x) = x^\top P_K x - \operatorname{tr}(P_K \Phi_K) + 2\widetilde{f}_{K,b}^\top (x - \mu_{K,b}) - (\mu_{K,b})^\top P_K \mu_{K,b},$$

and the action-value function $\widetilde{Q}_{K,b}(x, u)$ takes the form

$$\begin{aligned} \widetilde{Q}_{K,b}(x, u) = {} & \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2 \begin{pmatrix} \widetilde{p}_{K,b} \\ \widetilde{q}_{K,b} \end{pmatrix}^\top \begin{pmatrix} x \\ u \end{pmatrix} - \operatorname{tr}(P_K \Phi_K) - \sigma^2 \cdot \operatorname{tr}(R + P_K BB^\top) - b^\top Rb \\ & + 2b^\top RK\mu_{K,b} - (\mu_{K,b})^\top (Q + K^\top RK + P_K)\mu_{K,b} + 2\widetilde{f}_{K,b}^\top \big[(\overline{A}\mu + d) - \mu_{K,b}\big] \\ & + (\overline{A}\mu + d)^\top P_K (\overline{A}\mu + d) - 2\mu^\top P \mu_{K,b}. \end{aligned}$$

Here the matrix Υ_K is given in (3.7), and the vectors $\widetilde{p}_{K,b}, \widetilde{q}_{K,b}$ are given as

$$\begin{pmatrix} \widetilde{p}_{K,b} \\ \widetilde{q}_{K,b} \end{pmatrix} = \begin{pmatrix} A^\top \big[P_K \cdot (\overline{A}\mu + d) + \widetilde{f}_{K,b}\big] + P\mu \\ B^\top \big[P_K \cdot (\overline{A}\mu + d) + \widetilde{f}_{K,b}\big] \end{pmatrix}, \tag{C.1}$$

where the vector $\widetilde{f}_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K (Bb + \overline{A}\mu + d) - K^\top Rb + P\mu]$.

Proof. The proof is similar to the one of Proposition B.1. Thus we omit it here. □

Now we establish the gradients of $\widetilde{J}(K, b)$ for Problem C.2.

Proposition C.7. The gradient of $\widetilde{J}_1(K)$ and the gradient of $\widetilde{J}_2(K, b)$ w.r.t. b takes the form

$$\nabla_K \widetilde{J}_1(K) = 2(\Upsilon_K^{22} K - \Upsilon_K^{21}) \cdot \Phi_K, \qquad \nabla_b \widetilde{J}_2(K, b) = 2\big[\Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + \widetilde{q}_{K,b}\big],$$

where the matrix Υ_K is given in (3.7), and the vector $\widetilde{q}_{K,b}$ is given in (C.1).

Proof. The proof is similar to the one of Proposition B.3. Thus we omit it here. □

Equipped with above results, parallel to the analysis in §3, it is clear that by slight modification of Algorithms 1, 2, and 3, we derive similar actor-critic algorithms to solve both Problem C.1 and Problem C.2, where all the non-asymptotic convergence results hold. We omit the algorithms and the convergence results here.

D PROOFS OF THEOREMS

D.1 PROOF OF THEOREM 4.1

We denote by $\mu_{s+1}^* = \Lambda(\mu_s)$, which is the mean-field state generated by the optimal policy $\pi_{K^*,b^*(\mu_s)} = \Lambda_1(\mu_s)$ under the current mean-field state μ_s (note that for any mean-field state μ, the optimal $K^*(\mu)$ is independent of μ by Proposition 3.4; therefore, we write K^* here for convenience). By (3.2), we know that

$$\mu_{s+1}^* = (I - A + BK^*)^{-1} \cdot \big[Bb^*(\mu_s) + \overline{A}\mu_s + d\big].$$

Also, we denote by

$$\widetilde{\mu}_{s+1} = (I - A + BK_s)^{-1}(Bb_s + \overline{A}\mu_s + d),$$

which is the mean-field state generated by the policy π_{K_s,b_s} under the current mean-field state μ_s. According to Algorithm 1, it is obvious that the output μ_{s+1} is an estimator of $\widetilde{\mu}_{s+1}$. By triangular inequality, we have

$$\|\mu_{s+1} - \mu^*\|_2 \le \underbrace{\|\mu_{s+1} - \widetilde{\mu}_{s+1}\|_2}_{E_1} + \underbrace{\|\widetilde{\mu}_{s+1} - \mu_{s+1}^*\|_2}_{E_2} + \underbrace{\|\mu_{s+1}^* - \mu^*\|_2}_{E_3}. \tag{D.1}$$

We bound E_1, E_2 and E_3 in the sequel.

Bound on E_1: Note that from Theorem B.8 and the choice of the inputs in Algorithm 3 stated in Theorem B.4, it holds that

$$E_1 = \|\mu_{s+1} - \widetilde{\mu}_{s+1}\|_2 < \varepsilon_s \le \varepsilon/8 \cdot 2^{-s} \tag{D.2}$$

with probability at least $1 - \varepsilon^{10}$, where ε_s is given in (4.2).

Bound on E_2: Note that by definition, combining with triangular inequality, we have

$$\begin{aligned}
E_2 &= \Big\|(I - A + BK_s)^{-1}(Bb_s + \overline{A}\mu_s + d) - (I - A + BK^*)^{-1} \cdot \big[Bb^*(\mu_s) + \overline{A}\mu_s + d\big]\Big\|_2 \\
&\le \big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 \cdot \Big\|\big[I - A + BK^* + B(K_s - K^*)\big]^{-1} - (I - A + BK^*)^{-1}\Big\|_* \\
&\quad + \big\|(I - A + BK_s)^{-1}\big\|_* \cdot \|B\|_* \cdot \big\|b_s - b^*(\mu_s)\big\|_2 \\
&\le 2\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 \cdot \big\|(I - A + BK^*)^{-1}B(K_s - K^*)(I - A + BK^*)^{-1}\big\|_* \\
&\quad + \big\|(I - A + BK_s)^{-1}\big\|_* \cdot \|B\|_* \cdot \big\|b_s - b^*(\mu_s)\big\|_2 \\
&\le 2\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 \cdot \big[1 - \rho(A - BK_*)\big]^{-2} \cdot \|B\|_* \cdot \|K_s - K^*\|_* \\
&\quad + \big[1 - \rho(A - BK_0)\big]^{-1} \cdot \|B\|_* \cdot \big\|b_s - b^*(\mu_s)\big\|_2.
\end{aligned} \tag{D.3}$$

For the term $\|Bb^*(\mu_s) + \overline{A}\mu_s + d\|_2$, combining Proposition 3.4, it holds that

$$\begin{aligned}
\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 &\le L_1 \cdot \|B\|_* \cdot \|\mu_s\|_2 + \|\overline{A}\|_* \cdot \|\mu_s\|_2 + \|d\|_2 \\
&\le \big(L_1 \cdot \|B\|_* + \|\overline{A}\|_*\big) \cdot \|\mu_s\|_2 + \|d\|_2,
\end{aligned} \tag{D.4}$$

where the scalar L_1 is given in Assumption 3.1. Moreover, from Theorem B.4, by the choice of N_s, it holds with probability at least $1 - \varepsilon^{10}$ that

$$\|K_s - K^*\|_{\mathrm{F}} \le \sqrt{\sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \varepsilon_s}, \qquad \big\|b_s - b^*(\mu_s)\big\|_2 \le \sqrt{2\nu_{K^*}^{-1} \cdot \varepsilon_s}. \tag{D.5}$$

Combining (D.3), (D.4), (D.5) and the choice of ε_s in (4.2), we deduce that

$$E_2 \le \varepsilon/8 \cdot 2^{-s} \tag{D.6}$$

holds with probability at least $1 - \varepsilon^{10}$.

Bound on E_3: We have

$$E_3 = \|\mu_{s+1}^* - \mu^*\|_2 = \big\|\Lambda(\mu_s) - \Lambda(\mu^*)\big\|_2 \le L_0 \cdot \|\mu_s - \mu^*\|_2, \tag{D.7}$$

where we use the fact that the operator $\Lambda(\cdot)$ has Lipschitz constant $L_0 = L_1 L_3 + L_2$ according to Proposition 3.2.

From (D.2), (D.6) and (D.7), combining (D.1), we know that

$$\|\mu_{s+1} - \mu^*\|_2 \leq L_0 \cdot \|\mu_s - \mu^*\|_2 + \varepsilon \cdot 2^{-s-2}. \tag{D.8}$$

By telescoping (D.8), we obtain that

$$\|\mu_S - \mu^*\|_2 \leq L_0^S \cdot \|\mu_0 - \mu^*\|_2 + \varepsilon/2.$$

Moreover, by the choice of S in (4.1) and the definition of L_0 in Assumption 3.1, we know that $\|\mu_S - \mu^*\| < \varepsilon$. This concludes the theorem.

D.2 Proof of Theorem B.4

Proof. We first show that $J_1(K_N) - J_1(K^*) < \varepsilon/2$ with a high probability, then show that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$ with a high probability. Then we have

$$J(K_N, b_N) - J(K^*, b^*) = J_1(K_N) + J_2(K_N, b_H) - J_1(K^*) - J_2(K^*, b^*) < \varepsilon$$

with a high probability, which proves the theorem.

Part 1. We show that $J_1(K_N) - J_1(K^*) < \varepsilon/2$ with a high probability. We proceed our proof by first showing that $J_1(K)$ is gradient dominant, and then by gradient dominance, we prove the convergence of Algorithm 2 combining the convergence result of Algorithm 3 in Theorem B.8.

To show that $J_1(K)$ is gradient dominant, we first bound $J_1(K_1) - J_1(K_2)$ for any K_1 and K_2. By Proposition B.2, $J_1(K)$ takes the form

$$J_1(K) = \mathrm{tr}(P_K \Psi_\epsilon) = \mathbb{E}_{y \sim \mathcal{N}(0, \Psi_\epsilon)}(y^\top P_K y). \tag{D.9}$$

We use the following cost difference lemma to bound $y^\top P_{K_1} y - y^\top P_{K_2} y$.

Lemma D.1 (Cost Difference)**.** Denote by K_1 and K_2 two parameters, which satisfy that $\rho(A - BK_1) < 1$ and $\rho(A - BK_2) < 1$. Moreover, for any state vector y, we denote by $\{y_t\}_{t \geq 0}$ the sequence generated by the transition $y_{t+1} = (A - BK_2) y_t$ with initial state $y_0 = y$. Then it satisfies that

$$y^\top P_{K_2} y - y^\top P_{K_1} y = \sum_{t \geq 0} D_{K_1, K_2}(y_t).$$

Here we denote by the function

$$D_{K_1, K_2}(y) = 2y^\top (K_2 - K_1)(\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y + y^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y,$$

where the matrix Υ_K is given in (3.7).

Proof. See §F.1 for a detailed proof. □

Now, based on Lemma D.1, the following lemma shows that $J_1(K)$ is gradient dominant.

Lemma D.2 (Gradient Dominance)**.** Let K^* be the optimal parameter and K be a parameter such that $J_1(K)$ is finite, then we have the following lower bound for $J_1(K) - J_1(K^*)$

$$J_1(K) - J_1(K^*) \geq \sigma_{\min}(\Psi_\omega) \cdot \|\Upsilon_K^{22}\|_*^{-1} \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\big],$$

and the following upper bound

$$J_1(K) - J_1(K^*) \leq \sigma_{\min}^{-1}(R) \cdot \|\Phi_{K^*}\|_* \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\big].$$

Proof. See §F.2 for a detailed proof. □

We now use the above results to show the convergence of Algorithm 2. Recall that according to Algorithm 2, the parameter K is updated via

$$K_{n+1} = K_n - \gamma \cdot (\widehat{\Upsilon}^{22}_{K_n} K_n - \widehat{\Upsilon}^{21}_{K_n}), \tag{D.10}$$

where $\widehat{\Upsilon}_{K_n}$ is the output of Algorithm 3. We also define $\widetilde{K}_{n+1}$ as the exact update

$$\widetilde{K}_{n+1} = K_n - \gamma \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n}), \tag{D.11}$$

where Υ_{K_n} is given in (3.7).

We proceed to bound $|J_1(K_{n+1}) - J_1(K^*)|$ in the sequel. First, we claim that $J_1(K_N) \leq J_1(K_{N-1}) \leq \cdots \leq J_1(K_0)$ with high probability. We prove the claim by mathematical induction. Suppose that $J_1(K_n) \leq J_1(K_{n-1}) \leq \cdots \leq J_1(K_0)$ (this holds for $n = 0$). Recall the definition of $\widetilde{K}_{n+1}$ in (D.11) and the definition of $J_1(K)$ given in (3.6), we have

$$\begin{aligned}
J_1(\widetilde{K}_{n+1}) - J_1(K_n) &= \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}\big[y^\top (P_{\widetilde{K}_{n+1}} - P_{K_n}) y\big] \\
&= -2\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\
&\quad + \gamma^2 \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top \Upsilon^{22}_{K_n} (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\
&\leq -2\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\
&\quad + \gamma^2 \cdot \|\Upsilon^{22}_{K_n}\|_* \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big],
\end{aligned} \tag{D.12}$$

where the first line comes from (D.9) and the second line comes from Lemma D.1. Note that by the definition of Υ_K in (3.7), the term $\|\Upsilon^{22}_{K_n}\|_*$ is upper bounded as

$$\begin{aligned}
\|\Upsilon^{22}_{K_n}\|_* &\leq \|R\|_* + \|B\|_*^2 \cdot \|P_{K_n}\|_* \leq \|R\|_* + \|B\|_*^2 \cdot J_1(K_n) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon) \\
&\leq \|R\|_* + \|B\|_*^2 \cdot J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon).
\end{aligned}$$

Combining (D.12) and the choice of stepsize $\gamma \leq [\|R\|_* + \|B\|_*^2 \cdot J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon)]^{-1}$, it holds that

$$\begin{aligned}
J_1(\widetilde{K}_{n+1}) - J_1(K_n) &\leq -\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\
&\leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \mathrm{tr}\big[(\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\
&\leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1} \cdot \big[J_1(K_n) - J_1(K^*)\big],
\end{aligned} \tag{D.13}$$

where the last line comes from Lemma D.2. This also implies that $J_1(\widetilde{K}_{n+1}) \leq J_1(K_n)$.

Now, we use the following lemma to establish an upper bound of $|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})|$.

Lemma D.3. Suppose that $J_1(K_n) \leq J_1(K_0)$. Under the conditions stated in Theorem B.4, with probability at least $1 - \varepsilon^{15}$, it holds that

$$\big|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\big| \leq \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1} \cdot \varepsilon/4, \tag{D.14}$$

where K_{n+1} is given in (D.10) and $\widetilde{K}_{n+1}$ is given in (D.11).

Proof. See §F.10 for a detailed proof. □

Now, by applying Lemma D.3, we know that if $J_1(K_n) - J_1(K^*) \geq \varepsilon/2$, combining (D.13) and (D.14), it holds with probability at least $1 - \varepsilon^{15}$ that

$$J_1(K_{n+1}) - J_1(K_n) \leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1} \cdot \varepsilon/4 < 0. \tag{D.15}$$

This shows that with probability at least $1-\varepsilon^{15}$, it holds that $J_1(K_{n+1}) \leq J_1(K_n) \leq \cdots \leq J_1(K_0)$. Finally, note that the total number of iterations N goes only with $\log(1/\varepsilon)$, we know that $J_1(K_N) \leq J_1(K_{N-1}) \leq \cdots \leq J_1(K_0)$ with probability at least $1 - \varepsilon^{13}$, as long as $J_1(K_n) - J_1(K^*) \geq \varepsilon/2$ for any $n \leq N$. This finishes the proof of the claim.

Also, when $J_1(K_n) - J_1(K^*) \geq \varepsilon/2$, by (D.13) and (D.15), with probability at least $1 - \varepsilon^{13}$, it holds that

$$J_1(K_{n+1}) - J_1(K^*) \leq \big[1 - \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1}\big] \cdot \big[J_1(K_n) - J_1(K^*)\big].$$

Again, combining the fact that N goes only with $\log(1/\varepsilon)$, we know that

$$J_1(K_N) - J_1(K^*) \leq \varepsilon/2$$

with probability at least $1 - \varepsilon^{11}$.

Now we only need to give an error bound of $\|K_n - K^*\|_{\mathrm{F}}$. We upper bound $\|K - K^*\|_{\mathrm{F}}$ using $J_1(K) - J_1(K^*)$ in the following lemma.

Lemma D.4. For any K, we have

$$\|K - K^*\|_{\mathrm{F}}^2 \leq \sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \big[J_1(K) - J_1(K^*)\big].$$

Proof. See §F.3 for a detailed proof. □

Combining Lemma D.4, we show the error bound $\|K_N - K^*\|_{\mathrm{F}}$.

Part 2. We show that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$ with a high probability. Following from Proposition 3.4, it holds that $J_2(K^*, b^*) = J_2(K_N, b^{K_N})$. Therefore, to show that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$, we only need to prove that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$.

First, we use mathematical induction to show that $J_2(K_N, b_H) \leq J_2(K_N, b_{H-1}) \leq \cdots \leq J_2(K_N, b_1) \leq J_2(K_N, b_0)$, as long as $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon/2$ for any $h \leq H$. Suppose that $J_2(K_N, b_h) \leq J_2(K_N, b_{h-1}) \leq \cdots \leq J_2(K_N, b_0)$ (this holds for $h = 0$). Recall that according to Algorithm 2, the parameter b is updated by

$$b_{h+1} = b_h - \gamma^b \cdot \widehat{\nabla}_b J_2(K_N, b_h), \tag{D.16}$$

where $\widehat{\nabla}_b J_2(K_N, b_h) = \widehat{\Upsilon}_{K_N}^{22}(-K_N \widehat{\mu}_{K_N, b_h} + b_h) + \widehat{\Upsilon}_{K_N}^{21} \widehat{\mu}_{K_N, b_h} + \widehat{q}_{K_N, b_h}$ is the estimated gradient at b_h, and the matrix $\widehat{\Upsilon}_{K_N}$ and the vector $\widehat{q}_{K_N, b_h}$ are the outputs of Algorithm 3. We define $\widetilde{b}_{h+1}$ as an exact update

$$\widetilde{b}_{h+1} = b_h - \gamma^b \cdot \nabla_b J_2(K_N, b_h), \tag{D.17}$$

where $\nabla_b J_2(K_N, b_h) = \Upsilon_{K_N}^{22}(-K_N \mu_{K_N, b_h} + b_h) + \Upsilon_{K_N}^{21} \mu_{K_N, b_h} + q_{K_N, b_h}$ is the exact gradient at b_h, and Υ_{K_N} and q_{K_N, b_h} are given in (3.7). We proceed to bound $J_2(K_N, b_{h+1}) - J_2(K_N, b^{K_N})$ in the sequel, where $b^{K_N} = \operatorname{argmin}_b J_2(K_N, b)$.

Note that by the convexity and smoothness of $J_2(K_N, b)$ in Proposition 3.3, we have

$$\begin{aligned} J_2(K_N, \widetilde{b}_{h+1}) - J_2(K_N, b_h) &\leq -\gamma^b/2 \cdot \big\|\nabla_b J_2(K_N, b_h)\big\|_2^2 \leq -\nu_{K_N} \cdot \gamma^b \cdot \big[J_2(K_N, b_h) - J_2(K_N, b^{K_N})\big] \\ &\leq -\nu_{K_N} \cdot \gamma^b \cdot \varepsilon < 0, \end{aligned} \tag{D.18}$$

where in the inequalities, we use the fact that $J_2(K_N, b)$ has ι_{K_N}-Lipschitz continuous gradient and is ι_{K_N}-strongly convex by Proposition 3.3. Note that (D.18) also implies that $J_2(K_N, \widetilde{b}_{h+1}) < J_2(K_N, b_h)$. Now, we only need to establish the error bound $|J_2(\widetilde{K}_N, b_{h+1}) - J_2(K_N, b_{h+1})|$. The following lemma quantifies $\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big|$.

Lemma D.5. Assume that $J_2(K_N, b_h) \leq J_2(K_N, b_0)$. Under the conditions stated in Theorem B.4, with probability at least $1 - \varepsilon^{15}$, we have

$$\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big| \leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2, \tag{D.19}$$

where $\widehat{\nabla}_b J_2(K_N, b_h)$ is the approximate gradient given in (D.16), while $\nabla_b J_2(K_N, b_h)$ is the exact gradient given in (D.17).

Proof. See §F.4 for a detailed proof. □

Now, by applying Lemma D.5, we know that if $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon$, combining (D.18), it holds with probability at least $1 - \varepsilon^{15}$ that

$$J_2(K_N, b_{h+1}) - J_2(K_N, b_h) \leq -\nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2 < 0.$$

This shows that with probability at least $1 - \varepsilon^{15}$, it holds that $J_2(K_N, b_{h+1}) \leq J_2(K_N, b_h) \leq \cdots \leq J_2(K_N, b_0)$. Note that the total number of iterations H only goes by $\log(1/\varepsilon)$. Therefore, we know that as long as $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon$ for any $h \leq H$, it holds that $J_2(K_N, b_H) \leq J_2(K_N, b_{H-1}) \leq \cdots \leq J_2(K_N, b_0)$ with probability at least $1 - \varepsilon^{13}$. By this, we finish the claim.

Then, if $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon/2$, by (D.18) and (D.19), with probability at least $1 - \varepsilon^{13}$, it holds that

$$J_2(K_N, b_{h+1}) - J_2(K_N, b^{K_N}) \leq (1 - \nu_{K_N} \cdot \gamma^b) \cdot \big[J_2(K_N, b_h) - J_2(K_N, b^{K_N})\big].$$

Again, combining the fact that H goes only with $\log(1/\varepsilon)$, we know that

$$J_2(K_N, b_H) - J_2(K_N, b^{K_N}) \leq \varepsilon/2$$

with probability at least $1 - \varepsilon^{11}$. Moreover, combining the fact in Proposition 3.3 that $J_2(K_N, b)$ is strongly convex in b, we concludes the theorem. □

D.3 Proof of Theorem B.8

Proof. We follow the proof of Theorem 4.2 in Yang et al. (2019), where the authors only consider LQR without drift terms. Since our proof requires much more delicate analysis, we present it here.

Part 1. We proceed to show that $(\zeta_{K,b}, 0)$ is a saddle point of the problem (B.9). We first show that $\zeta_{K,b} \in \mathcal{V}_\zeta$ and $\xi(\zeta) \in \mathcal{V}_\xi$ for any $\zeta \in \mathcal{V}_\zeta$, where $\xi(\zeta)$ is defined as $\xi(\zeta) = \operatorname{argmax}_\xi F(\zeta, \xi)$, by the following lemma.

Lemma D.6. The vector $\zeta_{K,b} = (J(K,b), \alpha_{K,b}^\top)^\top \in \mathcal{V}_\zeta$. Moreover, for any $\zeta \in \mathcal{V}_\zeta$, the vector $\xi(\zeta)$ defined above satisfies that $\xi(\zeta) \in \mathcal{V}_\xi$.

Proof. See §F.5 for detailed proof. □

Note that $\nabla_\zeta F(\zeta_{K,b}, 0) = 0$ and $\nabla_\xi F(\zeta_{K,b}, 0) = 0$, combining the above Lemma D.6, we know that $(\zeta_{K,b}, 0)$ is a saddle point of the function $F(\zeta, \xi)$ defined in the problem (B.9). This finishes the proof of our **Part 1**.

Part 2. We define the primal-dual gap of (B.9) as

$$\text{gap}(\widehat{\zeta}, \widehat{\xi}) = \max_{\xi \in \mathcal{V}_\xi} F(\widehat{\zeta}, \xi) - \min_{\zeta \in \mathcal{V}_\zeta} F(\zeta, \widehat{\xi}), \tag{D.20}$$

which indeed captures the performance of $(\widehat{\zeta}, \widehat{\xi})$. In the sequel, we proceed to relate (D.20) with the estimation error $\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2$.

First, note that for the estimator $\widehat{\zeta}$ returned by Algorithm 3, we have

$$\begin{aligned} &\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\widehat{\zeta}^1 + \Theta_{K,b}\widehat{\zeta}^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\widehat{\zeta}^1 - J(K,b)\big|^2 \\ &\quad = F\big[\widehat{\zeta}, \xi(\widehat{\zeta})\big] = \max_{\xi \in \mathcal{V}_\xi} F(\widehat{\zeta}, \xi) = \text{gap}(\widehat{\zeta}, \widehat{\xi}) + \min_{\zeta \in \mathcal{V}_\zeta} F(\zeta, \widehat{\xi}), \end{aligned} \tag{D.21}$$

where the second line comes from (F.14) and the definition of $\xi(\zeta)$ in **Part 1**. Moreover, note that for any $\xi \in \mathcal{V}_\xi$, the follows hold:

$$\begin{aligned} &\min_{\zeta \in \mathcal{V}_\zeta} F(\zeta, \xi) \leq \min_{\zeta \in \mathcal{V}_\zeta} \max_{\xi \in \mathcal{V}_\xi} F(\zeta, \xi) = \min_{\zeta \in \mathcal{V}_\zeta} F\big[\zeta, \xi(\zeta)\big] \\ &\quad = \frac{1}{2} \cdot \min_{\zeta \in \mathcal{V}_\zeta} \Big\{ \Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\zeta^1 - J(K,b)\big|^2 \Big\} \\ &\quad = 0, \end{aligned} \tag{D.22}$$

where the first line comes from the definition of $\xi(\zeta)$ in **Part 1**, the second line comes from (F.14), while the last equality holds by taking $\zeta = \zeta_{K,b} \in \mathcal{V}_\zeta$. Further, we establish a lower for the LHS of (D.21) as follows

$$\begin{aligned}&\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\widehat{\zeta}^1 + \Theta_{K,b}\widehat{\zeta}^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\widehat{\zeta}^1 - J(K,b)\big|^2 \\ &\quad = \big\|\widetilde{\Theta}_{K,b}(\widehat{\zeta} - \zeta_{K,b})\big\|_2^2 \geq \lambda_K^2 \cdot \|\widehat{\zeta} - \zeta_{K,b}\|_2^2 \geq \lambda_K^2 \cdot \|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2^2, \end{aligned} \tag{D.23}$$

where the matrix $\widetilde{\Theta}_{K,b}$ is given in (B.8) and the scalar λ_K is specified in Proposition B.6. Combining (D.21), (D.22) and (D.23), we obtain the relation between $\text{gap}(\widehat{\zeta}, \widehat{\xi})$ and $\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2$:

$$\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2^2 \leq \lambda_K^{-2} \cdot \text{gap}(\widehat{\zeta}, \widehat{\xi}). \tag{D.24}$$

Part 3. We now proceed to upper bound the primal-dual gap $\text{gap}(\widehat{\zeta}, \widehat{\xi})$, then by (D.24) in **Part 2**, we establish the upper bound of the estimation error $\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2$. Note that since the state x and action u follow Gaussian distributions, therefore, they are unbounded. First we utilize Hansen-Wright inequality (Lemma G.3) to pick up an event where these two random variables are bounded. Before pick up such an event, the following Lemma D.7 characterize the distribution of $\widehat{\mu}_z = 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{z}_t$, where $\widetilde{z}_t = [\widetilde{x}_t^\top, \widetilde{u}_t^\top]^\top$ is the concatenation of the state $\widetilde{x}_t$ and the action $\widetilde{u}_t$. Then following transition holds

$$\widetilde{z}_{t+1} = L\widetilde{z}_t + \nu + \delta_t,$$

where

$$\nu = \begin{pmatrix} \overline{A}\mu + d \\ -K(\overline{A}\mu + d) + b \end{pmatrix}, \qquad \delta_t = \begin{pmatrix} \omega_t \\ -K\omega_t + \sigma\eta \end{pmatrix}, \qquad L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix}.$$

Note that we have

$$L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix} = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix}.$$

Then by the property of spectral radius, it holds that

$$\rho(L) = \rho\left(\begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} I \\ -K \end{pmatrix} \right) = \rho(A - BK) < 1.$$

Thus, the Markov chain $\{\widetilde{z}_t\}_{t \geq 0}$ generated by the transition admits a unique stationary distribution $\mathcal{N}(\mu_z, \Sigma_z)$. The following lemma characterizes the average $\widehat{\mu}_z = 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{z}_t$.

Lemma D.7. The average $\widehat{\mu}_z = 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{z}_t$, where $\widetilde{z}_t = [\widetilde{x}_t^\top, \widetilde{u}_t^\top]^\top$, satisfies that

$$\widehat{\mu}_z \sim \mathcal{N}\left(\mu_z + \frac{1}{\widetilde{T}} \mu_{\widetilde{T}},\ \frac{1}{\widetilde{T}} \widetilde{\Sigma}_{\widetilde{T}} \right),$$

where $\mu_{\widetilde{T}}$ and $\widetilde{\Sigma}_{\widetilde{T}}$ are bounded such that there exist positive absolute constants M_μ and M_Σ such that $\|\mu_{\widetilde{T}}\|_2 \leq M_\mu \cdot (1-\rho)^{-2} \cdot \|\mu_z\|_2$ and $\|\widetilde{\Sigma}_{\widetilde{T}}\|_{\text{F}} \leq M_\Sigma \cdot (1-\rho)^{-1} \cdot \|\Sigma_z\|_{\text{F}}$. This gives that with probability at least $1 - \widetilde{T}^{-6}$, it holds that

$$\|\widehat{\mu}_z - \mu_z\|_2 \leq \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \text{poly}\big(\|\Phi_K\|_*, \|K\|_{\text{F}}, \|b\|_2, \|\mu\|_2\big).$$

Proof. See §F.6 for a detailed proof. □

The above lemma also gives the error bound $\|\widehat{\mu}_{K,b} - \mu_{K,b}\|_2$.

Now, note that for any $z \sim \mathcal{N}(\mu_z, \Sigma_z)$, the random variable $z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}$ follows $\mathcal{N}(0, \Sigma_z + 1/\widetilde{T} \cdot \widetilde{\Sigma}_{\widetilde{T}})$. By Lemma G.3, there exists an absolute constant $C_0 > 0$ such that

$$\mathbb{P}\Big[\big|\|z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \text{tr}(\widetilde{\Sigma}_z)\big| > \tau\Big] \leq 2 \cdot \exp\Big[-C_0 \cdot \min\big(\tau^2 \|\widetilde{\Sigma}_z\|_{\text{F}}^{-2},\ \tau \|\widetilde{\Sigma}_z\|_*^{-1}\big)\Big],$$

where, for notational convenience, we write the matrix $\widetilde{\Sigma}_z = \Sigma_z + 1/\widetilde{T} \cdot \widetilde{\Sigma}_{\widetilde{T}}$. By taking $\tau = C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}$ for some sufficiently large positive absolute constant C_1, the above inequality takes the form

$$\mathbb{P}\Big[\big|\|z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \mathrm{tr}(\widetilde{\Sigma}_z)\big| > C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}\Big] \leq T^{-6}. \tag{D.25}$$

We define the following event for any $t \in [T]$:

$$\mathcal{E}_{t,1} = \Big\{\big|\|z_t - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \mathrm{tr}(\widetilde{\Sigma}_z)\big| \leq C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}\Big\}.$$

Then by (D.25), we know that $\mathbb{P}(\mathcal{E}_{t,1}) \geq 1 - T^{-6}$ for any $t \in [T]$. Also, we define $\mathcal{E}_1 = \cap_{t\in[T]}\mathcal{E}_{t,1}$, then we know that $\mathbb{P}(\mathcal{E}_1) \geq 1 - T^{-5}$ by union bound. Also, by the definition of $\mathcal{E}_1$, conditioning on the event $\mathcal{E}_1$, it holds that

$$\begin{aligned}
\max_{t\in[T]} \|z_t - \widehat{\mu}_z\|_2^2 &\leq C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}} + \mathrm{tr}(\widetilde{\Sigma}_z) + \|1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 \\
&\leq 2\widetilde{C}_1 \cdot \big[1 + M_\Sigma(1-\rho)^{-1}/\widetilde{T}^2\big] \cdot \log T \cdot \|\Sigma_z\|_* + M_\mu(1-\rho)^{-2}/\widetilde{T}^2 \cdot \|\mu_z\|_2^2 \\
&\leq C_2 \cdot \log T \cdot \big(1 + \|K\|_{\mathrm{F}}^2\big) \cdot \|\Phi_K\|_* \cdot (1-\rho)^{-1} + C_3 \cdot \big(\|b\|_2^2 + \|\mu\|_2^2\big) \cdot (1-\rho)^{-4} \cdot \widetilde{T}^{-2} \\
&\leq 2C_2 \cdot \log T \cdot \big(1 + \|K\|_{\mathrm{F}}^2\big) \cdot \|\Phi_K\|_* \cdot (1-\rho)^{-1}
\end{aligned} \tag{D.26}$$

for sufficiently large $\widetilde{T}$. Here C_2 and C_3 are positive absolute constants. Moreover, we define the following event

$$\mathcal{E}_2 = \big\{\|\widehat{\mu}_z - \mu_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2 \leq C_1\big\}.$$

Then by Lemma D.7, we know that $\mathbb{P}(\mathcal{E}_2) \geq 1 - \widetilde{T}^{-6}$ for sufficiently large $\widetilde{T}$. We define the event $\mathcal{E}$ as the intersection of the event $\mathcal{E}_1$ and $\mathcal{E}_2$, then by union bound, we know that $\mathbb{P}(\mathcal{E}) \geq 1 - T^{-5} - \widetilde{T}^{-6}$.

Now, we define the truncated feature vector $\widetilde{\psi}(x,u)$ as $\widetilde{\psi}(x,u) = \widehat{\psi}(x,u)\,\mathbb{1}_{\mathcal{E}}$, and also the truncated objective function as

$$\widetilde{F}(\zeta,\xi) = \Big\{\mathbb{E}(\widetilde{\psi})\zeta^1 + \mathbb{E}\big[(\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top\big]\zeta^2 - \mathbb{E}(\widetilde{c}\widetilde{\psi})\Big\}^\top \xi^2 + \big[\zeta^1 - \mathbb{E}(\widetilde{c})\big] \cdot \xi^1 - 1/2 \cdot \|\xi\|_2^2, \tag{D.27}$$

where we write $\widetilde{\psi} = \widetilde{\psi}(x,u)$ to simplify the notations, the function $\widetilde{c} = \widetilde{c}(x,u)$ is the truncated cost, which is defined as $\widetilde{c}(x,u) = c(x,u)\,\mathbb{1}_{\mathcal{E}}$, and the expectation is taken over the trajectory generated by the policy $\pi_{K,b}$. The following lemma establishes the upper bound of $|F(\zeta,\xi) - \widetilde{F}(\zeta,\xi)|$, where $F(\zeta,\xi)$ and $\widetilde{F}(\zeta,\xi)$ are given in (B.9) and (D.27) respectively.

Lemma D.8. Given $F(\zeta,\xi)$ and $\widetilde{F}(\zeta,\xi)$ in (B.9) and (D.27), it holds that

$$|F(\zeta,\xi) - \widetilde{F}(\zeta,\xi)| \leq \bigg[\frac{1}{2T} + \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}}\bigg] \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\big(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\big),$$

with probability at least $1 - \widetilde{T}^{-6}$.

Proof. See §F.7 for a detailed proof. □

By Lemma D.8, we know that with probability at least $1 - \widetilde{T}^{-6}$, it holds that

$$\begin{aligned}
&\bigg|\mathrm{gap}(\widehat{\zeta},\widehat{\xi}) - \bigg[\max_{\xi\in\mathcal{V}_\xi} \widetilde{F}(\widehat{\zeta},\xi) - \min_{\zeta\in\mathcal{V}_\zeta} \widetilde{F}(\zeta,\widehat{\xi})\bigg]\bigg| \\
&\qquad \leq \bigg[\frac{1}{2T} + \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}}\bigg] \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\big(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\big).
\end{aligned} \tag{D.28}$$

Therefore, to obtain the bound of $\mathrm{gap}(\zeta,\xi)$, we only need to bound the term $\max_{\xi\in\mathcal{V}_\xi} \widetilde{F}(\widehat{\zeta},\xi) - \min_{\zeta\in\mathcal{V}_\zeta} \widetilde{F}(\zeta,\widehat{\xi})$ in (D.28). We first use the following lemma to characterize the dependency of the trajectories generated by the policy $\pi_{K,b}$ and the state transition in Problem 2.2.

Lemma D.9. Consider a linear system $x_{t+1} = Dx_t + d + \epsilon_t$, where $\{x_t\}_{t\geq 0} \subset \mathbb{R}^m$, the matrix $D \in \mathbb{R}^{m\times m}$ satisfying $\rho(D) < 1$, the vector $d \in \mathbb{R}^m$, and $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ is the Gaussians. We denote by ϖ_t the marginal distribution of x_t for each $t \geq 0$. Besides, it is easy to verify that the stationary distribution of this Markov chain is a Gaussian distribution $\mathcal{N}((I-D)^{-1}d, \Sigma_\infty)$, where Σ_∞ is the covariance matrix. We define the β-mixing coefficients for any $n \geq 1$ as follows

$$\beta(n) = \sup_{t\geq 0} \mathbb{E}_{x\sim\varpi_t}\Big[\big\|\mathbb{P}_{x_n}(\cdot \,|\, x_0 = x) - \mathbb{P}_{\mathcal{N}((I-D)^{-1}d,\Sigma_\infty)}(\cdot)\big\|_{\mathrm{TV}}\Big].$$

Then, for any $\rho \in (\rho(D), 1)$, the mixing coefficients satisfy that

$$\beta(n) \leq C_{\rho,D} \cdot \big[\mathrm{tr}(\Sigma_\infty) + m \cdot (1-\rho)^{-2}\big]^{1/2} \cdot \rho^n,$$

where the scalar $C_{\rho,D,d}$ is a constant which only depends on ρ, D and d. We say that the sequence $\{x_t\}_{t\geq 0}$ is β-mixing with parameter ρ.

Proof. See Proposition 3.1 in Tu and Recht (2017) for a detailed proof. □

Note that under the state transition in Problem 2.2, the sequence $\{x_t\}_{t\geq 0}$ follows (3.1), where the matrix $A - BK$ satisfies that $\rho(A - BK) < 1$. Therefore, according to Lemma D.9, we know that the sequence $\{z_t = (x_t^\top, u_t^\top)^\top\}_{t\geq 0}$ is β-mixing with parameter $\rho \in (\rho(A - BK), 1)$. Now the following lemma helps us to establish the primal-dual gap for a convex-concave problem.

Lemma D.10. Let $\mathcal{X}$ and $\mathcal{Y}$ be two compact and convex sets such that $\|x - x'\|_2 \leq M$ and $\|y - y'\|_2 \leq M$ for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. We consider solving the following problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x, y) = \mathbb{E}_{\epsilon\sim\varpi_\epsilon}\big[G(x, y; \epsilon)\big],$$

where the objective function $F(x, y)$ is convex in x and concave in y. In addition, we assume that the distribution ϖ_ϵ is the stationary distribution induced by a Markov chain $\{\epsilon_t\}_{t\geq 0}$, which is β-mixing with $\beta(n) \leq C_\epsilon \cdot \rho^n$, where C_ϵ is a constant. Moreover, we assume that it holds almost surely that $G(x, y; \epsilon)$ is $\widetilde{L}_0$-Lipschitz in both x and y, the gradient $\nabla_x G(x, y; \epsilon)$ is $\widetilde{L}_1$-Lipschitz in x for any $y \in \mathcal{Y}$, the gradient $\nabla_y G(x, y; \epsilon)$ is $\widetilde{L}_1$-Lipschitz in y for any $x \in \mathcal{X}$, where for simplicity we assume that $C_\epsilon, \widetilde{L}_0, \widetilde{L}_1 > 1$. Each step of our gradient-based method takes the following form:

$$x_{t+1} = \Gamma_{\mathcal{X}}\big[x_t - \gamma_{t+1} \cdot \nabla_x G(x_t, y_t; \epsilon_t)\big], \qquad y_{t+1} = \Gamma_{\mathcal{Y}}\big[y_t - \gamma_{t+1} \cdot \nabla_y G(x_t, y_t; \epsilon_t)\big],$$

where the operators $\Gamma_{\mathcal{X}}$ and $\Gamma_{\mathcal{Y}}$ projects the variables back to $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the stepsizes take the form $\gamma_t = \gamma_0 \cdot t^{-1/2}$ for some constant $\gamma_0 > 0$. Moreover, let $\widehat{x} = (\sum_{t=1}^T \gamma_t)^{-1}(\sum_{t=1}^T \gamma_t x_t)$ and $\widehat{y} = (\sum_{t=1}^T \gamma_t)^{-1}(\sum_{t=1}^T \gamma_t y_t)$ be the final output of the gradient method after T iterations, then there exists an absolute constant $C > 0$, such that for any $\delta \in (0, 1)$, the primal-dual gap to the minimax optimization problem satisfies that

$$\max_{x\in\mathcal{X}} F(\widehat{x}, y) - \min_{y\in\mathcal{Y}} F(x, \widehat{y}) \leq \frac{C \cdot (M^2 + \widetilde{L}_0^2 + \widetilde{L}_0\widetilde{L}_1 M)}{\log(1/\rho)} \cdot \frac{\log^2 T + \log(1/\delta)}{\sqrt{T}} + \frac{C \cdot C_\epsilon \widetilde{L}_0 M}{T}$$

with probability at least $1 - \delta$.

Proof. See Theorem 5.4 in Yang et al. (2019) for a detailed proof. □

To use Lemma D.10, we define the function $G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}')$ as

$$G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \Big[\widetilde{\psi}\zeta^1 + (\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top\zeta^2 - \widetilde{c}\widetilde{\psi}\Big]^\top \xi^2 + (\zeta^1 - \widetilde{c}) \cdot \xi^1 - 1/2 \cdot \|\xi\|_2^2,$$

where recall that $\widetilde{\psi} = \widetilde{\psi}(x, u)$ and $\widetilde{\psi}' = \widetilde{\psi}(x', u')$. Note that its gradients take the form

$$\nabla_\zeta G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \begin{pmatrix} \widetilde{\psi}^\top\xi^2 + \xi^1 \\ \widetilde{\psi}(\widetilde{\psi} - \widetilde{\psi}')^\top\xi^2 \end{pmatrix}, \qquad \nabla_\xi G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \begin{pmatrix} \zeta^1 - \widetilde{c} - \xi^1 \\ \widetilde{\psi}\zeta^1 + (\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top\zeta^2 - \widetilde{c}\widetilde{\psi} - \xi^2 \end{pmatrix}.$$

By Condition B.7 and Lemma D.6, we know that

$$\begin{aligned}\big\|\nabla_\zeta G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 &\le \mathrm{poly}\big(\|K\|_{\mathrm{F}}, J(K_0,b_0)\big)\cdot\log^2 T\cdot(1-\rho)^{-2},\\ \big\|\nabla_\xi G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 &\le \mathrm{poly}\big(\|K\|_{\mathrm{F}}, \|\mu\|_2, J(K_0,b_0)\big)\cdot\log^2 T\cdot(1-\rho)^{-2}.\end{aligned} \tag{D.29}$$

This gives the Lipschitz constant $\widetilde{L}_0$ in Lemma D.10 for $G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$. Also, the Hessians take the forms

$$\nabla^2_{\zeta\zeta}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')=0,\qquad \nabla^2_{\xi\xi}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')=-I,$$

which follows that

$$\big\|\nabla^2_{\zeta\zeta}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2=0,\qquad \big\|\nabla^2_{\xi\xi}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2=1. \tag{D.30}$$

This gives the Lipschitz constant $\widetilde{L}_1$ in Lemma D.10 for $\nabla_\zeta G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$ and $\nabla_\xi G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$. Moreover, note that (D.26) provides an upper bound of M, combining (D.29), (D.30) and Lemma D.10, we know that

$$\max_{\xi\in\mathcal{V}_\xi}\widetilde{F}(\widehat{\zeta},\xi)-\min_{\zeta\in\mathcal{V}_\zeta}\widetilde{F}(\zeta,\widehat{\xi})\le\frac{\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|\mu\|_2,J(K_0,b_0)\big)\cdot\log^6 T}{(1-\rho)^4\cdot\sqrt{T}}$$

holds with probability at least $1-T^{-5}$. Combining (D.24) and (D.28), we conclude the theorem. □

E Proofs of Propositions

E.1 Proof of Proposition 3.2

Proof. We follow a similar argument as shown in the proof of Theorem 1.1 in Sznitman (1991) and Theorem 3.2 in Bensoussan et al. (2016). Note that for any policy $\pi_{K,b}\in\Pi$, the matrix K and the vector b uniquely determine the policy. This motivates us to define the metric on Π as follows.

Definition E.1. For any $\pi_{K_1,b_1},\pi_{K_2,b_2}\in\Pi$, we define the metric on the set of policies Π as

$$\|\pi_{K_1,b_1}-\pi_{K_2,b_2}\|_2=c_1\cdot\|K_1-K_2\|_*+c_2\cdot\|b_1-b_2\|_2,$$

where c_1 and c_2 are positive constants.

One can easily verify that the above definition satisfies the requirement of a metric. We first evaluate the forms of the operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot,\cdot)$.

Forms of the operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot,\cdot)$. It is obvious that

$$\Lambda_1(\mu)=\pi^*_\mu,$$

where π^*_μ solves Problem 2.2. We now turn to $\Lambda_2(\mu,\pi)$, which gives the new mean-field state generated by the policy π under the old mean-field state μ. Note that in Problem 2.2, the sequence of states $\{x_t\}_{t\ge0}$ constitutes a Markov chain, which has a stationary distribution. Thus, by taking the expectation of the state transition in Problem 2.2 and using the linear-Gaussian policy $\pi(x)=-K_\pi x+b_\pi+\sigma\eta$, we have $\mu_{\text{new}}=(A-BK_\pi)\mu_{\text{new}}+(Bb_\pi+\overline{A}\mu+d)$. By solving the equation, it holds that

$$\Lambda_2(\mu,\pi)=\mu_{\text{new}}=(I-A+BK_\pi)^{-1}(Bb_\pi+\overline{A}\mu+d).$$

This gives the form of $\Lambda_2(\cdot,\cdot)$.

Next, we compute the Lipschitz constants for both operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot,\cdot)$.

Lipschitz constant for $\Lambda_1(\cdot)$. Note that by Proposition 3.4, for any $\mu_1,\mu_2\in\mathbb{R}^m$, the optimal parameter policy K^* is fixed for Problem 2.2. Therefore, by the form of the optimal b^K given in Proposition 3.4, it holds that

$$\begin{aligned}\big\|\Lambda_1(\mu_1)-\Lambda_1(\mu_2)\big\|_2 &\le c_2\cdot\Big\|\big[(I-A)Q^{-1}(I-A)^\top+BR^{-1}B^\top\big]^{-1}\overline{A}\Big\|_*\\ &\qquad\cdot\Big\|\big[K^*Q^{-1}(I-A)^\top-R^{-1}B^\top\big]\Big\|_*\cdot\|\mu_1-\mu_2\|_2\\ &=c_2\cdot L_1\cdot\|\mu_1-\mu_2\|_2.\end{aligned} \tag{E.1}$$

Lipschitz constants for $\Lambda_2(\cdot,\cdot)$. Note that by Proposition 3.4, for any $\mu_1, \mu_2 \in \mathbb{R}^m$, the induced optimal parameter policy K^* is fixed for Problem 2.2. We thus have for any $\pi \in \Pi$ such that π is an optimal policy under some $\mu \in \mathbb{R}^m$, it holds that

$$\begin{aligned}\big\|\Lambda_2(\mu_1,\pi) - \Lambda_2(\mu_2,\pi)\big\|_2 &= \big\|(I - A + BK_\pi)^{-1} \cdot \overline{A} \cdot (\mu_1 - \mu_2)\big\|_2 \\ &\le \big[1 - \rho(A - BK^*)\big]^{-1} \|\overline{A}\|_* \cdot \|\mu_1 - \mu_2\|_2 \\ &= L_2 \cdot \|\mu_1 - \mu_2\|_2. \end{aligned} \tag{E.2}$$

Moreover, for any $\pi_1, \pi_2 \in \Pi$ such that there exists $\mu_1, \mu_2 \in \mathbb{R}^m$ such that π_1, π_2 are optimal under μ_1, μ_2, and for any population mean $\mu \in \mathbb{R}^m$, we have

$$\begin{aligned}\big\|\Lambda_2(\mu,\pi_1) - \Lambda_2(\mu,\pi_2)\big\|_2 &= \big\|(I - A + BK^*)^{-1} B \cdot (b_{\mu_1} - b_{\mu_2})\big\|_2 \\ &\le \big[1 - \rho(A - BK^*)\big]^{-1} \|B\|_* \cdot \|b_{\mu_1} - b_{\mu_2}\|_2 \\ &= c_2^{-1} \cdot L_3 \cdot \|\pi_1 - \pi_2\|_2. \end{aligned} \tag{E.3}$$

Now we proceed to show that the operator is a contraction. For any $\mu_1, \mu_2 \in \mathbb{R}^m$, the following inequality holds:

$$\begin{aligned}\big\|\Lambda(\mu_1) - \Lambda(\mu_2)\big\|_2 =& \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_1)\big) - \Lambda_2\big(\mu_2, \Lambda_1(\mu_2)\big)\Big\|_2 \\ \le& \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_1)\big) - \Lambda_2\big(\mu_1, \Lambda_1(\mu_2)\big)\Big\|_2 + \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_2)\big) - \Lambda_2\big(\mu_2, \Lambda_1(\mu_2)\big)\Big\|_2 \\ \le& c_2^{-1} \cdot L_3 \cdot \big\|\Lambda_1(\mu_1) - \Lambda_1(\mu_2)\big\|_2 + L_2 \cdot \|\mu_1 - \mu_2\|_2 \\ \le& c_2^{-1} \cdot L_3 \cdot c_2 \cdot L_1 \cdot \|\mu_1 - \mu_2\|_2 + L_2 \cdot \|\mu_1 - \mu_2\|_2 = (L_1 L_3 + L_2) \cdot \|\mu_1 - \mu_2\|_2,\end{aligned}$$

where in the second line, we use triangular inequality; in the third line, we use (E.2) and (E.3); in the last line, we use (E.1). By Assumption 3.1, we know that $L_0 = L_1 L_3 + L_2 < 1$, which shows that the operator $\Lambda(\cdot)$ is a contraction. Therefore, by Banach fixed-point theorem, we conclude that $\Lambda(\cdot)$ has a unique fixed point, which gives the equilibrium pair of Problem 2.1. □

E.2 Proof of Proposition 3.4

Proof. Note that by the definition of $J_2(K,b)$ in (3.6) and the definition of $\mu_{K,b}$ in (3.2), the problem $\min_b J_2(K,b)$ is a constrained optimization program

$$\begin{aligned}&\min_{\mu_{K,b},b} \begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}^\top \begin{pmatrix}Q + K^\top RK & -K^\top R\\ -RK & R\end{pmatrix} \begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix} \\ &\text{s.t. } (I - A + BK)\mu_{K,b} - (Bb + \overline{A}\mu + d) = 0.\end{aligned} \tag{E.4}$$

Consider the KKT conditions of (E.4), the minimizer satisfies that

$$2M_K \begin{pmatrix}\mu_{K,b^K}\\ b^K\end{pmatrix} + N_K \lambda = 0, \qquad N_K^\top \begin{pmatrix}\mu_{K,b^K}\\ b^K\end{pmatrix} + \overline{A}\mu + d = 0, \tag{E.5}$$

where we that

$$M_K = \begin{pmatrix}Q + K^\top RK & -K^\top R\\ -RK & R\end{pmatrix}, \qquad N_K = \begin{pmatrix}-(I - A + BK)^\top\\ B^\top\end{pmatrix}.$$

By solving (E.5), we obtain the minimizer to the program as follows

$$\begin{pmatrix}\mu_{K,b^K}\\ b^K\end{pmatrix} = -M_K^{-1} N_K (N_K^\top M_K^{-1} N_K)^{-1} (\overline{A}\mu + d). \tag{E.6}$$

By substituting (E.6) into the definition of $J_2(K,b)$ in (3.6), we have

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top (N_K^\top M_K^{-1} N_K)^{-1} (\overline{A}\mu + d). \tag{E.7}$$

Moreover, by algebra, we have

$$M_K^{-1} = \begin{pmatrix}Q^{-1} & Q^{-1}K^\top\\ KQ^{-1} & KQ^{-1}K^\top + R^{-1}\end{pmatrix}.$$

Therefore, the term $N_K^\top M_K^{-1} N_K$ in (E.7) takes the following form

$$N_K^\top M_K^{-1} N_K = (I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top.$$

Combining (E.7), we have

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top \big[(I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top\big]^{-1}(\overline{A}\mu + d).$$

Also, combining (E.6), we have

$$\begin{pmatrix} \mu_{K,b^K} \\ b^K \end{pmatrix} = \begin{pmatrix} Q^{-1}(I - A)^\top \\ KQ^{-1}(I - A)^\top - R^{-1}B^\top \end{pmatrix} \big[(I - A)Q^{-1}(I - A)^\top + BR^{-1}B^\top\big]^{-1}(\overline{A}\mu + d).$$

Then we finish the proof of the proposition. □

E.3 Proof of Proposition B.2

Proof. By the definition of the cost function $c(x, u)$ in Problem 2.2, we have

$$\begin{aligned} \mathbb{E}c_t &= \mathbb{E}(x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu) \\ &= \mathbb{E}(x_t^\top Q x_t + x_t^\top K^\top R K x_t - 2b^\top R K x_t + b^\top R b + \sigma^2 \eta_t^\top R \eta_t + \mu^\top \overline{Q}\mu) \\ &= \mathbb{E}\big[x_t^\top (Q + K^\top R K) x_t - 2b^\top R K x_t\big] + b^\top R b + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q}\mu, \end{aligned} \tag{E.8}$$

where in the second line we use the form of the linear policy. Therefore, combining (E.8) and the definition of $J(K, b)$ in Problem 2.2, we have

$$\begin{aligned} J(K, b) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \Big\{ \mathbb{E}\big[x_t^\top (Q + K^\top R K) x_t - 2b^\top R K x_t\big] + b^\top R b + \mu^\top \overline{Q}\mu \Big\} \\ &= \mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K)} \big[x^\top (Q + K^\top R K) x - 2b^\top R K x\big] + b^\top R b + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q}\mu \\ &= \mathrm{tr}\big[(Q + K^\top R K)\Phi_K\big] + (\mu_{K,b})^\top (Q + K^\top R K)\mu_{K,b} - 2b^\top R K \mu_{K,b} + b^\top R b + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q}\mu. \end{aligned}$$

Now, by applying (3.3) and (3.4) repeatedly, we have

$$\mathrm{tr}\big[(Q + K^\top R K)\Phi_K\big] = \mathrm{tr}(P_K \Psi_\epsilon),$$

where P_K is given in (3.4). This finishes the proof of the proposition. □

E.4 Proof of Proposition 3.3

Proof. By calculating the Hessian matrix of $J_2(K, b)$ directly, we have

$$\begin{aligned} \nabla_{bb}^2 J_2(K, b) =& B^\top (I - A + BK)^{-\top} (Q + K^\top R K)(I - A + BK)^{-1} B \\ &- \big[R K (I - A + BK)^{-1} B + B^\top (I - A + BK)^{-\top} K^\top R\big] + R \\ =& \big[\sqrt{R} K (I - A + BK)^{-1} B - \sqrt{R}\big]^\top \big[\sqrt{R} K (I - A + BK)^{-1} B - \sqrt{R}\big] \\ &+ B^\top (I - A + BK)^{-\top} Q (I - A + BK)^{-1} B, \end{aligned}$$

which is a positive constant matrix, whose minimum singular value is denoted by $\nu_K > 0$. Moreover, note that the spectral norm of $\nabla_{bb}^2 J_2(K, b)$ is upper bounded as

$$\big\|\nabla_{bb}^2 J_2(K, b)\big\|_* \le \big[1 - \rho(A - BK)\big]^{-2} \cdot \big(\|B\|_*^2 \cdot \|K\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*\big).$$

Therefore, we know that the maximum singular value ι_K of $\nabla_{bb}^2 J_2(K, b)$ is upper bounded

$$\iota_K \le \big[1 - \rho(A - BK)\big]^{-2} \cdot \big(\|B\|_*^2 \cdot \|K\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*\big).$$

This finishes the proof. □

E.5 PROOF OF PROPOSITION B.3

Proof. It holds that

$$\nabla_K\big[\mathrm{tr}(P_K\Psi_\epsilon)\big] = 2\big[(R+B^\top P_K B)K - B^\top P_K A\big]\Phi_K,$$

also, we have

$$\begin{aligned}
&\nabla_K\big[(\mu_{K,b})^\top(Q+K^\top RK)\mu_{K,b}\big]\\
&\quad= 2\big[RK - B^\top(I-A+BK)^{-\top}(Q+K^\top RK)\big]\mu_{K,b}(\mu_{K,b})^\top,\\
&\nabla_K(b^\top RK\mu_{K,b}) = \big[R - B^\top(I-A+BK)^{-\top}K^\top R\big]b(\mu_{K,b})^\top.
\end{aligned}$$

Combining the above equations, we obtain the form of $\nabla_K J_1(K)$. Moreover, the form of $\nabla_b J_2(K,b)$ is obtained simply by taking gradient w.r.t. b, combining the definition of $\mu_{K,b}$ in (3.2) and the Lyapunov equation (3.3). This concludes the theorem. □

E.6 PROOF OF PROPOSITION B.1

Proof. From the definition of $V_{K,b}(x)$ in (B.1) and the definition of the cost function $c(x,u)$ in Problem 2.2, it holds that

$$\begin{aligned}
V_{K,b}(x) = \sum_{t=0}^{\infty}\Big\{&\mathbb{E}\big[x_t^\top(Q+K^\top RK)x_t - 2b^\top RKx_t\\
&+ b^\top Rb + \sigma^2\eta_t^\top R\eta_t + \mu^\top\overline{Q}\mu \,|\, x_0 = x\big] - J(K,b)\Big\}.
\end{aligned}$$

Combining (3.1), we see that $V_{K,b}(x)$ is indeed a quadratic function taking the form $V_{K,b}(x) = x^\top Gx + r^\top x + h$, where G, r and h are functions of K and b. Note that $V_{K,b}(x)$ satisfies the Bellman equation, i.e.,

$$V_{K,b}(x) = c(x,-Kx+b) - J(K,b) + \mathbb{E}\big[V_{K,b}(x')\,|\,x\big],$$

then by substituting the form of $c(x,-Kx+b)$ in Problem 2.2 and $J(K,b)$ in (3.5), we obtain that

$$\begin{aligned}
x^\top Gx + r^\top x + h =& x^\top(Q+K^\top RK)x - 2b^\top RKx + b^\top Rb + \mu^\top\overline{Q}\mu\\
&- \big[\mathrm{tr}(P_K\Psi_\epsilon) + (\mu_{K,b})^\top(Q+K^\top RK)\mu_{K,b} - 2b^\top RK\mu_{K,b} + \mu^\top\overline{Q}\mu + b^\top Rb\big]\\
&+ \big[(A-BK)x + (Bb+\overline{A}\mu+d)\big]^\top G\big[(A-BK)x + (Bb+\overline{A}\mu+d)\big]\\
&+ \mathrm{tr}(G\Psi_\epsilon) + r^\top\big[(A-BK)x + (Bb+\overline{A}\mu+d)\big] + h - \sigma^2\cdot\mathrm{tr}(R).
\end{aligned} \tag{E.9}$$

By comparing the quadratic terms and linear terms in (E.9), we obtain that

$$G = P_K, \qquad r = 2f_{K,b}.$$

Also, by the definition of $V_{K,b}(x)$ in (B.1), we know that $\mathbb{E}[V_{K,b}(x)] = 0$. Therefore, we have

$$h = -2f_{K,b}\mu_{K,b} - (\mu_{K,b})^\top P_K\mu_{K,b} - \mathrm{tr}(P_K\Phi_K).$$

This finishes the proof of (B.3).

For the action-value function $Q_{K,b}(x,u)$, by the definition in (B.2) and (B.3), we obtain (B.4). This finishes the proof of the proposition. □

E.7 PROOF OF PROPOSITION B.5

Proof. By Proposition B.1, it holds that $Q_{K,b}$ takes the following linear form

$$Q_{K,b}(x,u) = \psi(x,u)^\top\alpha_{K,b} + \beta_{K,b}, \tag{E.10}$$

where $\beta_{K,b}$ is a scalar independent of x and u. Given the ergodic cost, recall that the Bellman equation takes the form

$$Q_{K,b}(x,u) = c(x,u) - J(K,b) + \mathbb{E}_{\pi_{K,b}}\big[Q_{K,b}(x',u')\,|\,x,u\big], \tag{E.11}$$

where (x', u') is the state-action pair after (x, u) following the policy $\pi_{K,b}$ and the state transition. Combining (E.10) and (E.11), we obtain that

$$\psi(x,u)^\top \alpha_{K,b} = c(x,u) - J(K,b) + \mathbb{E}_{\pi_{K,b}}\big[\psi(x',u') \,|\, x,u\big]^\top \alpha_{K,b}. \tag{E.12}$$

By left multiplying $\psi(x, u)$ to both sides of (E.12), and taking the total expectation, we have

$$\mathbb{E}_{\pi_{K,b}}\Big\{\psi(x,u)\big[\psi(x,u) - \psi(x',u')\big]^\top\Big\} \cdot \alpha_{K,b} + \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] \cdot J(K,b) = \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big].$$

Combining the definition of the matrix $\Theta_{K,b}$ in (B.7), we conclude the proposition. □

E.8 Proof of Proposition B.6

Proof. **Invertibility and Upper Bound.** We first introduce some notations. We define the vector $z = (x^\top, u^\top)^\top$. Then following the state transition and the policy $\pi_{K,b}$, the transition of z takes the form

$$z' = Lz + \nu + \delta, \tag{E.13}$$

where the matrix L and vectors ν, δ are given as

$$\nu = \begin{pmatrix} \overline{A}\mu + d \\ -K(\overline{A}\mu + d) + b \end{pmatrix}, \qquad \delta = \begin{pmatrix} \omega \\ -K\omega + \sigma\eta \end{pmatrix}, \qquad L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix}.$$

Note that the matrix L also takes the form

$$L = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix}.$$

Combining the fact that $\rho(UV) = \rho(VU)$ for any matrices U and V, we know that $\rho(L) = \rho(A - BK) < 1$, which gives the stability of (E.13).

Note that by (E.13), we see that the mean μ_z and the covariance Σ_z satisfy the following equations

$$\mu_z = L\mu_z + \nu, \qquad \Sigma_z = L\Sigma_z L^\top + \Psi_\delta, \tag{E.14}$$

where the matrix Ψ_δ is the covariance of the vector δ, which takes the form

$$\Psi_\delta = \begin{pmatrix} \Psi_\omega & -\Psi_\omega K^\top \\ -K\Psi_\omega & K\Psi_\omega K^\top + \sigma^2 I \end{pmatrix}.$$

Also, the covariance matrix Σ_z takes the following form

$$\Sigma_z = \begin{pmatrix} \Phi_K & -\Phi_K K^\top \\ -K\Phi_K & K\Phi_K K^\top + \sigma^2 \cdot I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 \cdot I \end{pmatrix} + \begin{pmatrix} I \\ -K \end{pmatrix} \Phi_K \begin{pmatrix} I \\ -K \end{pmatrix}^\top. \tag{E.15}$$

Now, we establish the following lemma.

Lemma E.2. The matrix $\Theta_{K,b}$ in (B.7) takes the following form

$$\Theta_{K,b} = \begin{pmatrix} 2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top & 0 \\ 0 & \Sigma_z(I - L)^\top \end{pmatrix}.$$

Proof. See §F.8 for a detailed proof. □

Note that since $\rho(L) < 1$, both the matrices $I - L \otimes_s L$ and $I - L$ are positive definite. Therefore, by Lemma E.2, the matrix $\Theta_{K,b}$ is invertible. This finishes the proof of the invertibility of the matrix $\Theta_{K,b}$. Moreover, from (E.15) and Lemma E.2, we upper bound the spectral norm of $\Theta_{K,b}$ as

$$\begin{aligned} \|\Theta_{K,b}\|_* &\le 2\max\Big\{\|\Sigma_z\|_*^2 \cdot \big(1 + \|L\|_*^2\big),\ \|\Sigma_z\|_* \cdot \big(1 + \|L\|_*\big)\Big\} \\ &\le 4\|\Sigma_z\|_*^2 \le 4\big(1 + \|K\|_{\mathrm{F}}^2\big)^2 \cdot \|\Phi_K\|_*^2. \end{aligned}$$

This proves the upper bound of the spectral norm.

Minimum singular value. To lower bound $\sigma_{\min}(\widetilde{\Theta}_{K,b})$, we only need to upper bound $\sigma_{\max}(\widetilde{\Theta}_{K,b}^{-1})$. We first proceed to calculate $\widetilde{\Theta}_{K,b}^{-1}$. Recall that the matrix $\widetilde{\Theta}_{K,b}$ in the linear system (B.8) takes the following form

$$\widetilde{\Theta}_{K,b} = \begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b} \end{pmatrix}.$$

By the definition of the feature vector $\psi(x,u)$, we know that the vector $\widetilde{\sigma}_z = \mathbb{E}_{\pi_{K,b}}[\psi(x,u)]$ takes the form

$$\widetilde{\sigma}_z = \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] = \begin{pmatrix} \mathrm{svec}(\Sigma_z) \\ \mathbf{0}_{k+m} \end{pmatrix}, \tag{E.16}$$

where $\mathbf{0}_{k+m}$ denotes the all-zero column vector with dimension $k+m$. Also, since $\Theta_{K,b}$ is invertible, the matrix $\widetilde{\Theta}_{K,b}$ is also invertible, whose inverse is given as

$$\widetilde{\Theta}_{K,b}^{-1} = \begin{pmatrix} 1 & 0 \\ -\Theta_{K,b}^{-1}\cdot\widetilde{\sigma}_z & \Theta_{K,b}^{-1} \end{pmatrix}.$$

The following lemma characterizes the spectral norm of the matrix $\widetilde{\Theta}_{K,b}^{-1}$.

Lemma E.3. The spectral norm of the matrix $\widetilde{\Theta}_{K,b}^{-1}$ is upper bounded by some positive constant $\widetilde{\lambda}_K$, where $\widetilde{\lambda}_K$ only depends on $\rho(A-BK)$, σ and $\sigma_{\min}(\Psi_\omega)$.

Proof. See §F.9 for a detailed proof. □

By Lemma E.3, we know that minimum singular value of the matrix $\widetilde{\Theta}_{K,b}$ is lower bounded by a positive constant $\lambda_K = 1/\widetilde{\lambda}_K$, which only depends on $\rho(A-BK)$, σ and $\sigma_{\min}(\Psi_\omega)$. This concludes the proposition. □

F Proofs of Lemmas

F.1 Proof of Lemma D.1

Proof. By the fact that P_{K_2} satisfies the Bellman equation in (3.4), we have

$$y^\top P_{K_2} y = \sum_{t\geq 0} y^\top\big[(A-BK_2)^t\big]^\top (Q + K_2^\top R K_2)(A-BK_2)^t y.$$

By the transition $y_{t+1} = (A-BK_2)y_t$, we know that $y_t = (A-BK_2)^t y$. Therefore, it holds that

$$y^\top P_{K_2} y = \sum_{t\geq 0} y_t^\top (Q + K_2^\top R K_2) y_t = \sum_{t\geq 0}(y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t).$$

Then by telescoping, we have

$$y^\top P_{K_2} y - y^\top P_{K_1} y = \sum_{t\geq 0}(y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t). \tag{F.1}$$

Also, for any $t \geq 0$, we have

$$\begin{aligned}
&y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t \\
&\quad = y_t^\top\big[Q + (K_2 - K_1 + K_1)^\top R (K_2 - K_1 + K_1)\big] y_t \\
&\qquad + y_t^\top\big[A - BK_1 - B(K_2 - K_1)\big]^\top P_{K_1}\big[A - BK_1 - B(K_2 - K_1)\big] y_t - y_t^\top P_{K_1} y_t \\
&\quad = 2 y_t^\top (K_2 - K_1)^\top\big[(R + B^\top P_{K_1} B)K_1 - B^\top P_{K_1} A\big] y_t + y_t^\top (K_2 - K_1)^\top (R + B^\top P_{K_1} B)(K_2 - K_1) y_t \\
&\quad = 2 y_t^\top (K_2 - K_1)^\top (\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y_t + y_t^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y_t.
\end{aligned} \tag{F.2}$$

Combining (F.1) and (F.2), we finish the proof. □

F.2 PROOF OF LEMMA D.2

Proof. **Upper Bound.** From the definition of $J_1(K)$ in (3.6), we have

$$J_1(K) - J_1(K^*) = \mathrm{tr}(P_K\Psi_\epsilon - P_{K^*}\Psi_\epsilon) = \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}(y^\top P_K y - y^\top P_{K^*} y)$$
$$= -\mathbb{E}_{y_0\sim\mathcal{N}(0,\Psi_\epsilon)}\left[\sum_{t\geq 0} D_{K,K^*}(y_t)\right], \tag{F.3}$$

where in the last equality, we apply Lemma D.1 and $\{y_t\}_{t\geq 0}$ follows the transition $y_{t+1} = (A - BK^*)y_t$. Also, we write $D_{K,K^*}(y)$ as

$$\begin{aligned} D_{K,K^*}(y) &= 2y^\top(K^* - K)(\Upsilon_K^{22}K - \Upsilon_K^{21})y + y^\top(K^* - K)^\top\Upsilon_K^{22}(K^* - K)x \\ &= y^\top\big[K^* - K + (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\big]^\top\Upsilon_K^{22}\big[K^* - K + (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\big]y \\ &\quad - y^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})y. \end{aligned} \tag{F.4}$$

Note that the first term on the RHS of (F.4) is positive, due to the fact that it is a quadratic form of a positive definite matrix, we therefore lower bound $D_{K,K^*}(y)$ as

$$D_{K,K^*}(y) \geq -y^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})y. \tag{F.5}$$

Therefore, combining (F.3) and (F.5), it holds that

$$\begin{aligned} J_1(K) - J_1(K^*) &\leq \left\|\mathbb{E}_{y_0\sim\mathcal{N}(0,\Psi_\epsilon)}\left[\sum_{t\geq 0} y_t y_t^\top\right]\right\|_* \cdot \mathrm{tr}\big[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\big] \\ &= \|\Phi_{K^*}\|_* \cdot \mathrm{tr}\big[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\big] \\ &\leq \big\|(\Upsilon_K^{22})^{-1}\big\|_* \cdot \|\Phi_{K^*}\|_* \cdot \mathrm{tr}\big[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})\big] \\ &\leq \sigma_{\min}^{-1}(R) \cdot \|\Phi_{K^*}\|_* \cdot \mathrm{tr}\big[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})\big], \end{aligned}$$

where the second line comes from direct computation, while the last line comes from the fact that $\Upsilon_K^{22} = R + B^\top KP_B \preceq R$. This complete the proof of the upper bound.

Lower Bound. Note that for any $\widetilde{K}$, it holds that

$$J(K) - J(K^*) \geq J(K) - J(\widetilde{K}) = -\mathbb{E}_{y_0\sim\mathcal{N}(0,\Psi_\epsilon)}\left[\sum_{t\geq 0} D_{K,\widetilde{K}}(y_t)\right], \tag{F.6}$$

where $\{y_t\}_{t\geq 0}$ follows the transition $y_{t+1} = (A - B\widetilde{K})y_t$. Therefore, by choosing $\widetilde{K} = K - (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})$, by a similar calculation in (F.4), the function $D_{K,\widetilde{K}}(y)$ takes the form

$$D_{K,\widetilde{K}}(y) = -y^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})y. \tag{F.7}$$

Combining (F.6) and (F.7), we obtain a lower bound as

$$\begin{aligned} J(K) - J(K^*) &\geq \mathrm{tr}\big[\Phi_{\widetilde{K}}(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\big] \\ &\geq \sigma_{\min}(\Psi_\epsilon) \cdot \|\Upsilon_K^{22}\|_*^{-1} \cdot \mathrm{tr}\big[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})\big]. \end{aligned}$$

Here in the last line, we use the fact that $\Phi_{\widetilde{K}} \succeq \Psi_\epsilon$. This finishes the proof of the lower bound. □

F.3 PROOF OF LEMMA D.4

Proof. Note that from Lemma D.1, it holds that

$$\begin{aligned} J_1(K) - J_1(K^*) &= \mathbb{E}\left\{\sum_{t\geq 0}\Big[2y_t^\top(K - K^*)(\Upsilon_{K^*}^{22}K^* - \Upsilon_{K^*}^{21})y_t + y_t^\top(K - K^*)^\top\Upsilon_{K^*}^{21}(K - K^*)y_t\Big]\right\} \\ &= \mathrm{tr}\big[\Phi_K(K - K^*)^\top\Upsilon_{K^*}^{21}(K - K^*)\big] \\ &\geq \|\Phi_K\|_* \cdot \|\Upsilon_{K^*}^{21}\|_* \cdot \mathrm{tr}\big[(K - K^*)^\top(K - K^*)\big] \\ &\geq \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|K - K^*\|_{\mathrm{F}}^2, \end{aligned}$$

where $\{y_t\}_{t\geq 0}$ follows the transition $y_{t+1} = (A - BK)y_t$. Here in the second line, we use the fact that $\Upsilon_{K^*}^{22}K^* - \Upsilon_{K^*}^{21}$ is the natural gradient of $J_1(\cdot)$ evaluated at K^*, which implies that $\Upsilon_{K^*}^{22}K^* - \Upsilon_{K^*}^{21} = 0$; while in the last line, we use the fact that $\|\Phi_K\|_* \geq \sigma_{\min}(\Psi_\epsilon)$ and $\|\Upsilon_{K^*}^{21}\|_* \geq \sigma_{\min}(R)$. This concludes the lemma. □

F.4 PROOF OF LEMMA D.5

Proof. Note that by Proposition 3.3, we have

$$\begin{aligned} J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1}) &\leq \gamma^b \cdot \nabla_b J_2(K_N, \widetilde{b}_{h+1})^\top \big[\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big] \\ &\quad + (\gamma^b)^2 \cdot \nu_{K_N}/2 \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2^2, \\ J_2(K_N, \widetilde{b}_{h+1}) - J_2(K_N, b_{h+1}) &\leq -\gamma^b \cdot \nabla_b J_2(K_N, \widetilde{b}_{h+1})^\top \big[\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big] \\ &\quad - (\gamma^b)^2 \cdot \iota_{K_N}/2 \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2^2. \end{aligned} \tag{F.8}$$

Also, we upper bound $\|\nabla_b J_2(K_N, \widetilde{b}_{h+1})\|_2$ as

$$\big\|\nabla_b J_2(K_N, \widetilde{b}_{h+1})\big\|_2 \leq \mathrm{poly}_1\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \big[1 - \rho(A - BK_N)\big]^{-1}. \tag{F.9}$$

Combining (F.8), (F.9) and the fact that $\nu_{K_N} \leq \iota_{K_N} \leq [1 - \rho(A - BK_N)]^{-2} \cdot \mathrm{poly}_2(\|K_N\|_*)$, we know that

$$\begin{aligned} &\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big| \\ &\quad \leq \gamma^b \cdot \mathrm{poly}_1\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2 \cdot \big[1 - \rho(A - BK_N)\big]^{-1} \\ &\qquad + (\gamma^b)^2 \cdot \mathrm{poly}_2\big(\|K_N\|_*\big) \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2^2 \cdot \big[1 - \rho(A - BK_N)\big]^{-2}. \end{aligned}$$

Note that from the definition of $\nabla_b J_2(K_N, b_h)$ and $\widehat{\nabla}_b J_2(K_N, b_h)$, we have

$$\begin{aligned} &\big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2 \\ &\quad \leq \|\widehat{\Upsilon}_{K_N}^{22} - \Upsilon_{K_N}^{22}\|_* \cdot \|K_N\|_* \cdot \|\widehat{\mu}_{K_N, b_h}\|_2 + \|\Upsilon_{K_N}^{22}\|_* \cdot \|K_N\|_* \cdot \|\widehat{\mu}_{K_N, b_h} - \mu_{K_N, b_h}\|_2 + \|\widehat{\Upsilon}_{K_N}^{22} - \Upsilon_{K_N}^{22}\|_* \cdot \|b_h\|_2 \\ &\qquad + \|\widehat{\Upsilon}_{K_N}^{21} - \Upsilon_{K_N}^{21}\|_* \cdot \|\widehat{\mu}_{K_N, b_h}\|_2 + \|\Upsilon_{K_N}^{21}\|_* \cdot \|\widehat{\mu}_{K_N, b_h} - \mu_{K_N, b_h}\|_2 + \|\widehat{q}_{K_N, b_h} - q_{K_N, b_h}\|_2. \end{aligned}$$

From Theorem B.8, combining the fact that $J_2(K_N, b_h) \leq J_2(K_N, b_0)$ and the fact that $\|\mu_{K_N, b}\|_2 \leq J(K_N, b_0)/\sigma_{\min}(Q)$, we know that with probability at least $1 - (T^b)^{-4} - (\widetilde{T}^b)^{-6}$, it holds that

$$\begin{aligned} &\big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2 \\ &\quad \leq \lambda_{K_N}^{-1} \cdot \mathrm{poly}_3\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big) \cdot \bigg[\frac{\log^3 T_n^b}{(T_n^b)^{1/4}(1-\rho)^2} + \frac{\log^{1/2} \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/8} \cdot (1-\rho)}\bigg]. \end{aligned}$$

To show that (D.19) holds, we only need to pick γ^b, T_n^b and $\widetilde{T}_n^b$ such that

$$\begin{aligned} &\gamma^b \cdot \mathrm{poly}_1\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \lambda_{K_N}^{-1} \cdot \mathrm{poly}_3\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big) \\ &\quad \cdot \bigg[\frac{\log^3 T_n^b}{(T_n^b)^{1/4}(1-\rho)^2} + \frac{\log^{1/2} \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/8} \cdot (1-\rho)}\bigg] \cdot \big[1 - \rho(A - BK_N)\big]^{-1} + (\gamma^b)^2 \cdot \mathrm{poly}_2\big(\|K_N\|_*\big) \cdot \lambda_{K_N}^{-1} \\ &\quad \cdot \mathrm{poly}_3\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big) \cdot \bigg[\frac{\log^6 T_n^b}{(T_n^b)^{1/2}(1-\rho)^4} + \frac{\log \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/4} \cdot (1-\rho)^2}\bigg] \cdot \big[1 - \rho(A - BK_N)\big]^{-2} \\ &\quad \leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2, \end{aligned}$$

in other words, we pick

$$\begin{aligned} &\gamma^b \leq \min\Big\{1 - \rho(A - BK_N),\ \big[1 - \rho(A - BK_N)\big]^{-2} \cdot \big(\|B\|_*^2 \cdot \|K_N\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*\big)\Big\}, \\ &T_n^b = \mathrm{poly}_4\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \lambda_{K_N}^{-4} \cdot \nu_{K_N}^{-4} \cdot (1-\rho)^{-10} \cdot \varepsilon^{-5}, \\ &\widetilde{T}_n^b = \mathrm{poly}_5\big(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \lambda_{K_N}^{-2} \cdot \nu_{K_N}^{-2} \cdot (1-\rho)^{-16} \cdot \varepsilon^{-8}, \end{aligned}$$

then (D.19) holds with probability at least $1 - \varepsilon^{15}$. This corresponds to the choices of parameters in the statement of Theorem B.4. We then finish the proof. □

F.5 PROOF OF LEMMA D.6

Proof. **Part 1.** First we proceed to prove that $\zeta_{K,b} \in \mathcal{V}_\zeta$. Note that from Condition B.7, we know that $\zeta^1_{K,b} = J(K,b)$ satisfies that $0 \le \zeta^1_{K,b} \le J(K_0,b_0)$. It remains to show that $\zeta^2_{K,b} = \alpha_{K,b}$ satisfies that $\|\zeta^2_{K,b}\|_2 \le M_\zeta$. By the definition of $\alpha_{K,b}$ in (B.6), we know that

$$\begin{aligned}\|\alpha_{K,b}\|_2^2 \le \|\Upsilon_K\|_{\mathrm{F}}^2 + \|\Upsilon_K\|_*^2 \cdot \big(\|\mu_{K,b}\|_2^2 + \|\mu^u_{K,b}\|_2^2\big) \\ + \big(\|A\|_* + \|B\|_*\big)^2 \cdot \big(\|P_K\|_* \cdot \|\overline{A}\mu + d\|_2 + \|f_{K,b}\|_2\big)^2\end{aligned} \tag{F.10}$$

where for notational simplicity, we denote by $\mu^u_{K,b}$ the mean-field action, i.e., $\mu^u_{K,b} = -K\mu_{K,b} + b$. We only need to bound Υ_K, $\mu_{K,b}$, $\mu^u_{K,b}$, P_K, and $f_{K,b}$. Note that the expected total cost $J(K,b)$ takes the form

$$J(K,b) = \mathrm{tr}(P_K\Psi_\epsilon) + (\mu_{K,b})^\top Q\mu_{K,b} + (\mu^u_{K,b})^\top R\mu^u_{K,b} + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q}\mu.$$

Thus, we have

$$\begin{aligned}&J(K_0,b_0) \ge J(K,b) \ge \sigma_{\min}(\Psi_\omega)\cdot \mathrm{tr}(P_K) \ge \sigma_{\min}(\Psi_\omega)\cdot\|P_K\|_*,\\ &J(K_0,b_0) \ge J(K,b) \ge (\mu_{K,b})^\top Q\mu_{K,b} \ge \sigma_{\min}(Q)\cdot\|\mu_{K,b}\|_2,\\ &J(K_0,b_0) \ge J(K,b) \ge (\mu^u_{K,b})^\top R\mu^u_{K,b} \ge \sigma_{\min}(R)\cdot\|\mu^u_{K,b}\|_2,\end{aligned}$$

which imply that

$$\|P_K\|_* \le J(K_0,b_0)/\sigma_{\min}(\Psi_\omega), \qquad \|\mu_{K,b}\|_2 \le J(K_0,b_0)/\sigma_{\min}(Q), \qquad \|\mu^u_{K,b}\|_2 \le J(K_0,b_0)/\sigma_{\min}(R). \tag{F.11}$$

For Υ_K, note that we decompose the matrix in the following way:

$$\Upsilon_K = \begin{pmatrix} Q & \\ & R\end{pmatrix} + \begin{pmatrix} A^\top \\ B^\top\end{pmatrix} P_K \begin{pmatrix} A & B\end{pmatrix},$$

which gives

$$\begin{aligned}&\|\Upsilon_K\|_{\mathrm{F}} \le (\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}}) + \big(\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2\big)\cdot\|P_K\|_{\mathrm{F}},\\ &\|\Upsilon_K\|_* \le (\|Q\|_* + \|R\|_*) + \big(\|A\|_* + \|B\|_*\big)^2\cdot\|P_K\|_*.\end{aligned}$$

Combining (F.11) and the fact that $\|U\|_{\mathrm{F}} \le \sqrt{m}\cdot\|U\|_*$ for any $U \in \mathbb{R}^{m\times m}$, we know that

$$\begin{aligned}&\|\Upsilon_K\|_{\mathrm{F}} \le (\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}}) + \big(\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2\big)\cdot\sqrt{m}\cdot J(K_0,b_0)/\sigma_{\min}(\Psi_\omega),\\ &\|\Upsilon_K\|_* \le (\|Q\|_* + \|R\|_*) + \big(\|A\|_* + \|B\|_*\big)^2\cdot J(K_0,b_0)/\sigma_{\min}(\Psi_\omega).\end{aligned} \tag{F.12}$$

Now we focus on the bound of the vector $f_{K,b}$. Note that we write $f_{K,b}$ in the following way

$$f_{K,b} = -P_K\mu_{K,b} + (I - A + BK)^{-T}\big[Q\mu_{K,b} - K^\top R\mu^u_{K,b}\big],$$

we therefore upper bound $f_{K,b}$ as

$$\|f_{K,b}\|_2 \le J(K_0,b_0)^2\cdot\sigma_{\min}^{-1}(\Psi_\omega)\cdot\sigma_{\min}^{-1}(Q) + \big[1-\rho(A-BK)\big]^{-1}\cdot(\kappa_Q + \kappa_R\cdot\|K\|_{\mathrm{F}}) \tag{F.13}$$

Combining (F.10), (F.11), (F.12) and (F.13), we know that $\|\zeta^2_{K,b^K}\|_2 = \|\alpha_{K,b^K}\|_2 \le M_{\zeta,1} + M_{\zeta,2}\cdot(1+\|K\|_{\mathrm{F}})\cdot[1-\rho(A-BK)]^{-1}$. Therefore, we know that $\zeta_{K,b^K} \in \mathcal{V}_\zeta$.

Part 2. Now we show that for any $\zeta \in \mathcal{V}_\zeta$, we have $\xi(\zeta) \in \mathcal{V}_\xi$. Note that $\xi(\zeta)$ is given by

$$\begin{aligned}&\xi^1(\zeta) = \zeta^1 - J(K,b^K),\\ &\xi^2(\zeta) = \mathbb{E}_{\pi_{K,b^K}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b^K}\zeta^2 - \mathbb{E}_{\pi_{K,b^K}}\big[c(x,u)\psi(x,u)\big].\end{aligned} \tag{F.14}$$

Then we have

$$\big|\xi^1(\zeta)\big| = \big|\zeta^1 - J(K,b^K)\big| \le J(K_0,b_0) \tag{F.15}$$

and also

$$\left\|\xi^2(\zeta)\right\|_2 \leq \underbrace{\left\|\mathbb{E}_{\pi_{K,b^K}}\left[\psi(x,u)\right]\zeta^1\right\|_2}_{B_1} + \underbrace{\|\Theta_{K,b^K}\|_* \cdot \|\zeta^2\|_2}_{B_2} + \underbrace{\left\|\mathbb{E}_{\pi_{K,b^K}}\left[c(x,u)\psi(x,u)\right]\right\|_2}_{B_3}. \tag{F.16}$$

Note that we upper bound B_1 as

$$B_1 \leq J(K_0, b_0) \cdot \left\|\mathbb{E}_{\pi_{K,b^K}}\left[\psi(x,u)\right]\right\|_2 \leq J(K_0, b_0) \cdot \|\Sigma_z\|_{\mathrm{F}}, \tag{F.17}$$

where Σ_z is given in (E.15). Also, by Proposition B.6, we bound B_2 as

$$B_2 \leq 4(1 + \|K\|_{\mathrm{F}}^2)^3 \cdot \|\Phi_K\|_*^2 \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot \left[1 - \rho(A - BK)\right]^{-1}. \tag{F.18}$$

As for the term B_3 in (F.16), we utilize the following lemma to provide an upper bound.

Lemma F.1. The vector $\mathbb{E}_{\pi_{K,b^K}}\left[c(x,u)\psi(x,u)\right]$ has the following form

$$\mathbb{E}_{\pi_{K,b^K}}\left[c(x,u)\psi(x,u)\right] = \begin{pmatrix} 2\mathrm{svec}\left[\Sigma_z \mathrm{diag}(Q,R)\Sigma_z + \langle \Sigma_z, \mathrm{diag}(Q,R)\rangle \Sigma_z\right] \\ \Sigma_z \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} \end{pmatrix}$$
$$+ \left((\mu_{K,b})^\top Q \mu_{K,b} + (\mu_{K,b}^u)^\top R \mu_{K,b}^u + \mu^\top \overline{Q}\mu\right) \begin{pmatrix} \mathrm{svec}(\Sigma_z) \\ \mathbf{0}_m \\ \mathbf{0}_k \end{pmatrix}.$$

Here the matrix Σ_z takes the form

$$\Sigma_z = \begin{pmatrix} \Phi_K & -\Phi_K K^\top \\ -K\Phi_K & K\Phi_K K^\top + \sigma^2 \cdot I \end{pmatrix}.$$

Proof. See §F.11 for detailed proof. □

From Lemma F.1 and (F.11), we obtain the upper bound for B_3

$$B_3 \leq 3\left[\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}} + J(K_0,b_0)\|Q\|_*/\sigma_{\min}(Q) + J(K_0,b_0)\|R\|_*/\sigma_{\min}(R)\right] \cdot \|\Sigma_z\|_*^2. \tag{F.19}$$

Moreover, by the definition of Σ_z in (E.15), combining the triangular inequality, we have the following bounds for the Frobenius norm and spectral norm of Σ_z, respectively:

$$\|\Sigma_z\|_{\mathrm{F}} \leq 2(d + \|K\|_{\mathrm{F}}^2) \cdot \|\Phi_K\|_*, \qquad \|\Sigma_z\|_* \leq 2(1 + \|K\|_{\mathrm{F}}^2) \cdot \|\Phi_K\|_*. \tag{F.20}$$

Also, by similar techniques we used in deriving (F.11), we have

$$J(K_0, b_0) \geq J(K, b^K) \geq \mathrm{tr}\left[(Q + K^\top R K)\Phi_K\right] \geq \|\Phi_K\|_* \cdot \sigma_{\min}(Q),$$

which gives the upper bound for Φ_K as follows

$$\|\Phi_K\|_* \leq J(K_0, b_0)/\sigma_{\min}(Q). \tag{F.21}$$

Therefore, combining (F.16), (F.17), (F.18), (F.19), (F.20) and (F.21), we know that

$$\left\|\xi^2(\zeta)\right\|_2 \leq C \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot J(K_0,b_0)^2/\sigma_{\min}^2(Q) \cdot \left(1 + \|K\|_{\mathrm{F}}^2\right)^3 \cdot \left[1 - \rho(A - BK)\right]^{-1}. \tag{F.22}$$

By (F.15) and (F.22), we know that $\xi(\zeta) \in \mathcal{V}_\xi$ for any $\zeta \in \mathcal{V}_\zeta$. From this, we conclude the lemma. □

F.6 Proof of Lemma D.7

Proof. Assume that $\widetilde{z}_0 \sim \mathcal{N}(\mu_\dagger, \Sigma_\dagger)$. Recall that the following transition holds

$$\widetilde{z}_{t+1} = L\widetilde{z}_t + \nu + \delta_t. \tag{F.23}$$

We know that

$$\widetilde{z}_t \sim \mathcal{N}\left(L^t\mu_\dagger + \sum_{i=0}^{t-1} L^i \cdot \nu,\ (L^\top)^t \Sigma_\dagger L^t + \sum_{i=0}^{t-1} (L^\top)^i \Psi_\delta L^i\right), \tag{F.24}$$

where Ψ_δ is the covariance matrix of δ_t, which takes the form

$$\Psi_\delta = \begin{pmatrix} \Psi_\omega & K\Psi_\omega \\ K\Psi_\omega & K\Psi_\omega K^\top + \sigma^2 I \end{pmatrix}.$$

From (F.23), we know that μ_z takes the form

$$\mu_z = (I - L)^{-1}\nu = \sum_{j=0}^{\infty} L^j \nu.$$

Therefore, combining (F.24), we have

$$\mathbb{E}(\widehat{\mu}_z) = \mu_z + \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} L^t \mu_\dagger - \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} L^i \nu. \tag{F.25}$$

It holds that

$$\begin{aligned} &\left\| \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} L^t \mu_\dagger - \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} L^i \nu \right\|_2 \\ &\quad \le \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} \rho(L)^t \cdot \|\mu_\dagger\|_2 + \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} \rho(L)^i \cdot \|\nu\|_2 \\ &\quad \le \frac{\big[1 - \rho(L)\big]^{-1}}{\widetilde{T}} \cdot \|\mu_\dagger\|_2 + \frac{\big[1 - \rho(L)\big]^{-2}}{\widetilde{T}} \cdot \|\nu\|_2 \\ &\quad \le M_\mu \cdot (1 - \rho)^{-2} \cdot \|\mu_z\|_2 / \widetilde{T}, \end{aligned} \tag{F.26}$$

where M_μ is a positive absolute constant. For the covariance, note that for any random variables $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$, we know that $Z = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma)$, where $\|\Sigma\|_{\mathrm{F}} \le 2\|\Sigma_1\|_{\mathrm{F}} + 2\|\Sigma_2\|_{\mathrm{F}}$. Combining (F.24), we know that $\widehat{\mu}_z \sim \mathcal{N}(\mathbb{E}\widehat{\mu}_z, \widetilde{\Sigma}_{\widetilde{T}}/\widetilde{T})$, where $\widetilde{\Sigma}_{\widetilde{T}}$ satisfies that

$$\begin{aligned} \widetilde{T}/2 \cdot \|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} &\le \sum_{t=1}^{\widetilde{T}} \rho(L)^{2t} \cdot \|\Sigma_\dagger\|_{\mathrm{F}} + \sum_{t=1}^{\widetilde{T}} \sum_{i=0}^{t-1} \rho(L)^{2i} \cdot \|\Psi_\delta\|_{\mathrm{F}} \\ &\le \big[1 - \rho(L)^2\big]^{-1} \cdot \|\Sigma_\dagger\|_{\mathrm{F}} + \widetilde{T} \cdot \big[1 - \rho(L)^2\big]^{-1} \cdot \|\Psi_\delta\|_{\mathrm{F}}, \end{aligned}$$

which implies that

$$\|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} \le M_\Sigma \cdot (1 - \rho)^{-1} \cdot \|\Sigma_z\|_{\mathrm{F}}, \tag{F.27}$$

where M_Σ is a positive absolute constant. Combining (F.25), (F.26) and (F.27), we conclude that $\widehat{\mu}_z$ follows the distribution. The inequality follows by a Gaussian tail bounds. Then we finish the proof of the lemma. □

F.7 Proof of Lemma D.8

Proof. We continue using the notations given in §D.3. We define

$$\widehat{F}(\zeta, \xi) = \Big\{ \mathbb{E}(\widehat{\psi})\zeta^1 + \mathbb{E}\big[(\widehat{\psi} - \widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2 - \mathbb{E}(c\widehat{\psi}) \Big\}^\top \xi^2 + \big[\zeta^1 - \mathbb{E}(c)\big] \cdot \xi^1 - 1/2 \cdot \|\xi\|_2^2,$$

where $\widehat{\psi} = \widehat{\psi}(x, u)$ is the estimated feature vector. Here the expectation is only taken over the trajectory generated by the state transition and the policy $\pi_{K,b}$, instead of the randomness when calculating the estimated feature vectors. Thus, the function $\widehat{F}(\zeta, \xi)$ is still random, where the randomness comes from the estimated feature vectors. Note that $|F(\zeta, \xi) - \widetilde{F}(\zeta, \xi)| \le |F(\zeta, \xi) - \widehat{F}(\zeta, \xi)| + |\widehat{F}(\zeta, \xi) - \widetilde{F}(\zeta, \xi)|$. Thus, we only need to characterize $|F(\zeta, \xi) - \widehat{F}(\zeta, \xi)|$ and $|\widehat{F}(\zeta, \xi) - \widetilde{F}(\zeta, \xi)|$.

Part 1. First we characterize $|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)|$. Note that by algebra, we have

$$\begin{aligned}\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big| &= \bigg|\Big\{\mathbb{E}(\psi-\widehat{\psi})\zeta^1+\mathbb{E}\big[(\psi-\psi')\psi^\top-(\widehat{\psi}-\widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2-\mathbb{E}\big[c(\psi-\widehat{\psi})\big]\Big\}^\top\xi^2\bigg| \\ &\leq \mathbb{E}\big(\|\psi-\widehat{\psi}\|_2\big)\cdot\Big[|\zeta^1|+\mathbb{E}\big(\|\psi-\psi'\|_2+2\|\widehat{\psi}\|_2\big)\|\zeta^2\|_2+\mathbb{E}(c)\Big]\cdot\|\xi^2\|_2,\end{aligned} \tag{F.28}$$

where the expectation is only taken over the trajectory generated by the state transition and the policy $\pi_{K,b}$. From $\mathbb{P}(\|\widehat{\mu}_z-\mu_z+1/\widetilde{T}\cdot\mu_{\widetilde{T}}\|_2\leq C_1)\geq 1-\widetilde{T}^{-6}$, we know that with probability at least $1-\widetilde{T}^{-6}$, the term $\mathbb{E}(\|\psi-\psi'\|_2+2\|\widehat{\psi}\|_2)$ is upper bounded by a polynomial $\mathrm{poly}(\|\Phi_K\|_*,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0))$. Also, the term $\mathbb{E}(c)$ is also upper bounded by a polynomial $\mathrm{poly}(\|\Phi_K\|_*,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0))$ by the definition of the cost function $c(x,u)$. Therefore, combining Condition B.7 and (F.28), with probability at least $1-\widetilde{T}^{-6}$, we obtain that

$$\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big|\leq\mathbb{E}\big(\|\psi-\widehat{\psi}\|_2\big)\cdot\mathrm{poly}\big(\|\Phi_K\|_*,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big). \tag{F.29}$$

We upper bound the term $\|\psi(x,u)-\widehat{\psi}(x,u)\|_2$ for any x and u as

$$\begin{aligned}\|\psi(x,u)-\widehat{\psi}(x,u)\|_2^2 &= \|\widehat{\mu}_z-\mu_z\|_2^2+\big\|z(\widehat{\mu}_z-\mu_z)^\top+(\widehat{\mu}_z-\mu_z)z^\top\big\|_{\mathrm{F}}^2+\|\mu_z\mu_z^\top-\widehat{\mu}_z\widehat{\mu}_z^\top\|_{\mathrm{F}}^2 \\ &\leq \mathrm{poly}\big(\|\Phi_K\|_*,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big)\cdot\|\widehat{\mu}_z-\mu_z\|_2^2,\end{aligned} \tag{F.30}$$

where $z=[x^\top,u^\top]^\top$. Also, by Lemma D.7, we know that

$$\|\widehat{\mu}_z-\mu_z\|_2\leq\frac{\log\widetilde{T}}{\widetilde{T}^{1/4}}\cdot(1-\rho)^{-2}\cdot\mathrm{poly}\big(\|\Phi_K\|_*,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big), \tag{F.31}$$

holds with probability at least $1-\widetilde{T}^{-6}$. Combining (F.29), (F.30) and (F.31), it holds that

$$\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big|\leq\frac{\log\widetilde{T}}{\widetilde{T}^{1/4}}\cdot(1-\rho)^{-2}\cdot\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big) \tag{F.32}$$

with probability at least $1-\widetilde{T}^{-6}$.

Part 2. We now characterize $|\widehat{F}(\zeta,\xi)-\widetilde{F}(\zeta,\xi)|$ in the sequel. By definitions, we have

$$\begin{aligned}&\big|\widetilde{F}(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big| \\ &\quad= \bigg|\Big\{\mathbb{E}(\widetilde{\psi}-\widehat{\psi})\zeta^1+\mathbb{E}\big[(\widetilde{\psi}-\widetilde{\psi}')\widetilde{\psi}^\top-(\widehat{\psi}-\widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2-\mathbb{E}(\widetilde{c}\widetilde{\psi}-\widehat{c}\widehat{\psi})\Big\}^\top\xi^2+\mathbb{E}(\widehat{c}-\widetilde{c})\xi^1\bigg| \\ &\quad\leq \bigg|\Big\{\mathbb{E}(\widehat{\psi})\zeta^1+\mathbb{E}(\widehat{\psi}\widehat{\psi}^\top)\zeta^2-\mathbb{E}(\widehat{c}\widehat{\psi})\Big\}^\top\xi^2+\mathbb{E}(\widehat{c})\xi^1\bigg|\cdot\mathbb{1}_{\mathcal{E}^c}+\bigg|\big[\mathbb{E}(\widehat{\psi}'\widehat{\psi}^\top)\zeta^2\big]^\top\xi^2\bigg|\cdot\mathbb{1}_{(\mathcal{E}'\cap\mathcal{E})^c},\end{aligned} \tag{F.33}$$

where $\mathcal{E}'$ is an event defined as

$$\mathcal{E}'=\bigg(\bigcap_{t\in[T]}\Big\{\big|\|z_t'-\mu_z+1/\widetilde{T}\cdot\mu_{\widetilde{T}}\|_2^2-\mathrm{tr}(\widetilde{\Sigma}_z)\big|\leq C_1\cdot\log T\cdot\|\widetilde{\Sigma}_z\|_*\Big\}\bigg)\bigcap\mathcal{E}_2,$$

which, by Lemma G.3, satisfies that $\mathbb{P}(\mathcal{E}')\geq 1-T^{-5}-\widetilde{T}^{-6}$. By a similar argument as in **Part 1**, we upper bound the two absolute values on the RHS of (F.33) as

$$\big|\widetilde{F}(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big|\leq\bigg[\frac{1}{T}+\frac{1}{\widetilde{T}^{1/4}}\bigg]\cdot\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big) \tag{F.34}$$

for sufficiently large T and $\widetilde{T}$. Therefore, combining (F.32) and (F.34), by triangular inequality, we finish the proof of the lemma. □

F.8 PROOF OF LEMMA E.2

Proof. Recall that the feature vector $\psi(x,u)$ takes the following form

$$\psi(x,u) = \begin{pmatrix} \text{svec}\big[(z-\mu_z)(z-\mu_z)^\top\big] \\ z-\mu_z \end{pmatrix}.$$

We then have

$$\psi(x,u) - \psi(x',u') = \begin{pmatrix} \text{svec}\big[yy^\top - (Ly+\delta)(Ly+\delta)^\top\big] \\ y-(Ly+\delta) \end{pmatrix},$$

where we denote by $y = z - \mu_z$, and x' and u' are the state and action following the state transition and the policy $\pi_{K,b}$. Therefore, for any symmetric matrices M, N and any vectors m, n, the following equation holds

$$\begin{aligned}
&\begin{pmatrix} \text{svec}(M) \\ m \end{pmatrix}^\top \Theta_{K,b} \begin{pmatrix} \text{svec}(N) \\ n \end{pmatrix} \\
&\quad = \mathbb{E}_{y,\delta}\bigg\{ \begin{pmatrix} \text{svec}(M) \\ m \end{pmatrix}^\top \begin{pmatrix} \text{svec}(yy^\top) \\ y \end{pmatrix} \begin{pmatrix} \text{svec}\big[yy^\top - (Ly+\delta)(Ly+\delta)^\top\big] \\ y-(Ly+\delta) \end{pmatrix}^\top \begin{pmatrix} \text{svec}(N) \\ n \end{pmatrix} \bigg\} \\
&\quad = \mathbb{E}_{y,\delta}\Big\{ \big(\langle M, yy^\top\rangle + m^\top y\big) \cdot \big[\langle N, yy^\top - (Ly+\delta)(Ly+\delta)^\top\rangle + n^\top(y - Ly - \delta)\big] \Big\} \\
&\quad = \underbrace{\mathbb{E}_y\big[\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\big]}_{A_1} + \underbrace{\mathbb{E}_y\big[\langle yy^\top, M\rangle \cdot n^\top(y - Ly)\big]}_{A_2} \\
&\qquad + \underbrace{\mathbb{E}_y\big[m^\top y \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\big]}_{A_3} + \underbrace{\mathbb{E}_y\big[m^\top y \cdot n^\top(y - Ly)\big]}_{A_4},
\end{aligned} \tag{F.35}$$

where the expectations are taken over the distribution $y \sim \mathcal{N}(0, \Sigma_z)$ and $\delta \sim \mathcal{N}(0, \Psi_\delta)$.

For the terms A_2 and A_3 in (F.35), by the fact that $y = z - \mu_z \sim \mathcal{N}(0, \Sigma_z)$, we know that these two terms vanish. For A_4, it holds that

$$A_4 = \mathbb{E}_y\big[m^\top y \cdot (y - Ly)^\top n\big] = \mathbb{E}_y\big[m^\top yy^\top (I-L)^\top n\big] = m^\top \Sigma_z (I-L)^\top n. \tag{F.36}$$

For A_1, by algebra, we have

$$\begin{aligned}
A_1 &= \mathbb{E}_y\big[\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\big] \\
&= \mathbb{E}_y\big[\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top, N\rangle\big] - \mathbb{E}_y\big[\langle yy^\top, M\rangle \cdot \langle \Psi_\delta, N\rangle\big] \\
&= \mathbb{E}_y\big[y^\top M y \cdot y^\top (N - L^\top N L) y\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle \\
&= \mathbb{E}_{u\sim\mathcal{N}(0,I)}\big[u^\top \Sigma_z^{1/2} M \Sigma_z^{1/2} u \cdot u^\top \Sigma_z^{1/2}(N - L^\top N L)\Sigma_z^{1/2} u\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle.
\end{aligned} \tag{F.37}$$

Now, by applying Lemma G.1 to the first term on the RHS of (F.37), we know that

$$\begin{aligned}
A_1 &= 2\,\text{tr}\big[\Sigma_z^{1/2} M \Sigma_z^{1/2} \cdot \Sigma_z^{1/2}(N - L^\top N L)\Sigma_z^{1/2}\big] \\
&\qquad + \text{tr}(\Sigma_z^{1/2} M \Sigma_z^{1/2}) \cdot \text{tr}\big[\Sigma_z^{1/2}(N - L^\top N L)\Sigma_z^{1/2}\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle \\
&= 2\langle M, \Sigma_z(N - L^\top N L)\Sigma_z\rangle + \langle \Sigma_z, M\rangle \cdot \langle \Sigma_z - L\Sigma_z L^\top - \Psi_\delta, N\rangle \\
&= 2\langle M, \Sigma_z(N - L^\top N L)\Sigma_z\rangle,
\end{aligned}$$

where we use (E.14) in the last equality. By using the property of the operator $\text{svec}(\cdot)$ and the definition of the symmetric Kronecker product, we obtain that

$$\begin{aligned}
A_1 &= 2\text{svec}(M)^\top \text{svec}\big[\Sigma_z(N - L^\top N L)\Sigma_z\big] \\
&= 2\text{svec}(M)^\top \big[\Sigma_z \otimes_s \Sigma_z - (\Sigma_z L^\top) \otimes_s (\Sigma_z L^\top)\big]\text{svec}(N) \\
&= 2\text{svec}(M)^\top \big[(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top\big]\text{svec}(N).
\end{aligned} \tag{F.38}$$

Combining (F.35), (F.36) and (F.38), we obtain that

$$\begin{aligned}
&\begin{pmatrix}\mathrm{svec}(M)\\ m\end{pmatrix}^\top \Theta_{K,b}\begin{pmatrix}\mathrm{svec}(N)\\ n\end{pmatrix}\\
&\quad= \mathrm{svec}(M)^\top\big[2(\Sigma_z\otimes_s\Sigma_z)(I-L\otimes_s L)^\top\big]\mathrm{svec}(N)+m^\top\Sigma_z(I-L)^\top n\\
&\quad= \begin{pmatrix}\mathrm{svec}(M)\\ m\end{pmatrix}^\top\begin{pmatrix}2(\Sigma_z\otimes_s\Sigma_z)(I-L\otimes_s L)^\top & 0\\ 0 & \Sigma_z(I-L)^\top\end{pmatrix}\begin{pmatrix}\mathrm{svec}(N)\\ n\end{pmatrix}.
\end{aligned}$$

Thus, the matrix $\Theta_{K,b}$ takes the following form

$$\Theta_{K,b}=\begin{pmatrix}2(\Sigma_z\otimes_s\Sigma_z)(I-L\otimes_s L)^\top & 0\\ 0 & \Sigma_z(I-L)^\top\end{pmatrix}.$$

This concludes the lemma. □

F.9 Proof of Lemma E.3

Proof. By the property of the spectral norm, we upper bound $\|\widetilde{\Theta}_{K,b}^{-1}\|_*$ as

$$\|\widetilde{\Theta}_{K,b}^{-1}\|_*^2\le 1+\|\Theta_{K,b}^{-1}\|_*^2+\|\Theta_{K,b}^{-1}\widetilde{\sigma}_z\|_2^2. \tag{F.39}$$

We proceed to bound the RHS of (F.39). For the term $\Theta_{K,b}^{-1}\widetilde{\sigma}_z$, combining Lemma E.2 and (E.16), we have

$$\begin{aligned}
\Theta_{K,b}^{-1}\widetilde{\sigma}_z &= \begin{pmatrix}1/2\cdot(I-L\otimes_s L)^{-\top}(\Sigma_z\otimes_s\Sigma_z)^{-1}\cdot\mathrm{svec}(\Sigma_z)\\ \mathbf{0}_{k+m}\end{pmatrix}\\
&= \begin{pmatrix}1/2\cdot(I-L\otimes_s L)^{-\top}(\Sigma_z^{-1}\otimes_s\Sigma_z^{-1})\cdot\mathrm{svec}(\Sigma_z)\\ \mathbf{0}_{k+m}\end{pmatrix}\\
&= \begin{pmatrix}1/2\cdot(I-L\otimes_s L)^{-\top}\cdot\mathrm{svec}(\Sigma_z^{-1})\\ \mathbf{0}_{k+m}\end{pmatrix},
\end{aligned}$$

where we use the property of the symmetric Kronecker product in the second and last line. Therefore, we have

$$\begin{aligned}
\|\Theta_{K,b}^{-1}\widetilde{\sigma}_z\|_2 &= 1/2\cdot\big\|(I-L\otimes_s L)^{-\top}\cdot\mathrm{svec}(\Sigma_z^{-1})\big\|_2\\
&\le 1/2\cdot\big\|(I-L\otimes_s L)^{-\top}\big\|_*\cdot\big\|\mathrm{svec}(\Sigma_z^{-1})\big\|_2\\
&\le 1/2\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\|\Sigma_z^{-1}\|_{\mathrm{F}}\\
&\le 1/2\cdot\sqrt{k+m}\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\|\Sigma_z^{-1}\|_*\\
&= 1/2\cdot\sqrt{k+m}\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\big[\sigma_{\min}(\Sigma_z)\big]^{-1},
\end{aligned}\tag{F.40}$$

where in the third line we use Lemma G.2 to the matrix $L\otimes_s L$. Similarly, we upper bound $\|\Theta_{K,b}^{-1}\|_*$ in the sequel

$$\|\Theta_{K,b}^{-1}\|_*\le\min\Big\{1/2\cdot\big[1-\rho^2(L)\big]^{-1}\big[\sigma_{\min}(\Sigma_z)\big]^{-2},\big[1-\rho(L)\big]^{-1}\big[\sigma_{\min}(\Sigma_z)\big]^{-1}\Big\}. \tag{F.41}$$

Thus, combining (F.39), (F.40) and (F.41), we obtain that

$$\begin{aligned}
\|\widetilde{\Theta}_{K,b}^{-1}\|_*^2 \le{}& 1+1/2\cdot\sqrt{k+m}\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\big[\sigma_{\min}(\Sigma_z)\big]^{-1}\\
&+\min\Big\{1/2\cdot\big[1-\rho^2(L)\big]^{-1}\big[\sigma_{\min}(\Sigma_z)\big]^{-2},\big[1-\rho(L)\big]^{-1}\big[\sigma_{\min}(\Sigma_z)\big]^{-1}\Big\}.
\end{aligned}\tag{F.42}$$

Now it remains to characterize $\sigma_{\min}(\Sigma_z)$. For any vectors $s\in\mathbb{R}^m$ and $r\in\mathbb{R}^k$, we have

$$\begin{aligned}
\begin{pmatrix}s\\ r\end{pmatrix}^\top\Sigma_z\begin{pmatrix}s\\ r\end{pmatrix} &= \mathbb{E}_{\pi_{K,b},x\sim(\mu_{K,b},\Phi_K)}\Big\{\big[s^\top(x-\mu_{K,b})+r^\top(u+K\mu_{K,b}-b)\big]^2\Big\}\\
&= \mathbb{E}_{x\sim(\mu_{K,b},\Phi_K),\eta\sim\mathcal{N}(0,I)}\Big\{\big[(s-K^\top r)^\top(x-\mu_{K,b})+\sigma r^\top\eta\big]^2\Big\}\\
&= \mathbb{E}_{x\sim(\mu_{K,b},\Phi_K)}\Big\{\big[(s-K^\top r)^\top(x-\mu_{K,b})\big]^2\Big\}+\mathbb{E}_{\eta\sim\mathcal{N}(0,I)}\big[(\sigma r^\top\eta)^2\big].
\end{aligned}\tag{F.43}$$

The first term on the RHS of (F.43) is lower bounded as

$$\mathbb{E}_{x\sim(\mu_{K,b},\Phi_K)}\Big\{\big[(s-K^\top r)^\top(x-\mu_{K,b})\big]^2\Big\} = (s-K^\top r)^\top\Phi_K(s-K^\top r)$$
$$\geq \|s-K^\top r\|_2^2\cdot\sigma_{\min}(\Phi_K)\geq\|s-K^\top r\|_2^2\cdot\sigma_{\min}(\Psi_\omega), \tag{F.44}$$

where the last inequality comes from the fact that $\sigma_{\min}(\Phi_K)\geq\sigma_{\min}(\Psi_\omega)$ by (3.3). The second term on the RHS of (F.43) takes the form $\mathbb{E}_{\eta\sim\mathcal{N}(0,I)}[(\sigma r^\top\eta)^2]=\sigma^2\|r\|_2^2$. Therefore, combining (F.43) and (F.44), we have

$$\begin{pmatrix}s\\r\end{pmatrix}^\top\Sigma_z\begin{pmatrix}s\\r\end{pmatrix}\geq\|s-K^\top r\|_2^2\cdot\sigma_{\min}(\Psi_\omega)+\sigma^2\|r\|_2^2$$
$$\geq\sigma_{\min}(\Psi_\omega)\cdot\|s\|_2^2+\big[\sigma^2-\|K\|_*^2\cdot\sigma_{\min}(\Psi_\omega)\big]\cdot\|r\|_2^2.$$

From this, we see that $\sigma_{\min}(\Sigma_z)$ is lower bounded by $\min\{\sigma_{\min}(\Psi_\omega),\sigma^2-\|K\|_*^2\cdot\sigma_{\min}(\Psi_\omega)\}$, which only depends on $\sigma_{\min}(\Psi_\omega)$ and σ. Thus, combining (F.42), we know that $\|\widetilde{\Theta}_{K,b}^{-1}\|_*$ is upper bounded by some constant $\widetilde{\lambda}_K$, where $\widetilde{\lambda}_K$ only depends on $\rho(A-BK)$, σ and $\sigma_{\min}(\Psi_\omega)$. This finishes the proof of the lemma. □

F.10 Proof of Lemma D.3

Proof. By direct calculation and Proposition B.2, we have

$$\big|J_1(\widetilde{K}_{n+1})-J_1(K_{n+1})\big|=\mathrm{tr}\big[(P_{\widetilde{K}_{n+1}}-P_{K_{n+1}})\Psi_\epsilon\big]\leq\|P_{\widetilde{K}_{n+1}}-P_{K_{n+1}}\|_*\cdot\|\Psi_\epsilon\|_{\mathrm{F}}. \tag{F.45}$$

The following lemma helps establish the upper bound of the term $\|P_{\widetilde{K}_{n+1}}-P_{K_{n+1}}\|_*$.

Lemma F.2. Suppose that the parameters K and $\widetilde{K}$ satisfy that

$$\|\widetilde{K}-K\|_*\cdot\big(\|A-BK\|_*+1\big)\cdot\|\Phi_K\|_*\leq\sigma_{\min}(\Psi_\omega)/4\cdot\|B\|_*^{-1}, \tag{F.46}$$

then it holds that

$$\|P_{\widetilde{K}}-P_K\|_*\leq 6\cdot\sigma_{\min}^{-1}(\Psi_\omega)\cdot\|\Phi_K\|_*\cdot\|K\|_*\cdot\|R\|_*\cdot\|\widetilde{K}-K\|_*$$
$$\cdot\big(\|B\|_*\cdot\|K\|_*\big)\cdot\|A-BK\|_*+\|B\|_*\cdot\|K\|_*+1\big). \tag{F.47}$$

Proof. See Lemma 5.7 in Yang et al. (2019) for a detailed proof. □

To use Lemma F.2, we only need to verify that $\widetilde{K}_{n+1}$ and K_{n+1} satisfy (F.46). Note that from (D.10) and (D.11), we have

$$\|\widetilde{K}_{n+1}-K_{n+1}\|_*\cdot\big(\|A-B\widetilde{K}_{n+1}\|_*+1\big)\cdot\|\Phi_{\widetilde{K}_{n+1}}\|_*$$
$$\leq\gamma\cdot\|\widehat{\Upsilon}_{K_n}-\Upsilon_{K_n}\|_{\mathrm{F}}\cdot\big(1+\|K_n\|_*\big)\cdot\big(\|A-B\widetilde{K}_{n+1}\|_*+1\big)\cdot\|\Phi_{\widetilde{K}_{n+1}}\|_*. \tag{F.48}$$

Now we proceed to upper bound the RHS of (F.48). For the term $\|A-B\widetilde{K}_{n+1}\|_*$, we have

$$\|A-B\widetilde{K}_{n+1}\|_*\leq\|A-BK_n\|_*+\gamma\cdot\|B\|_*\cdot\|\Upsilon_{K_n}^{22}K_n-\Upsilon_{K_n}^{21}\|_*$$
$$\leq\|A-BK_n\|_*+\gamma\cdot\|B\|_*\cdot\|\Upsilon_{K_n}\|_*\cdot\big(1+\|K_n\|_*\big). \tag{F.49}$$

And by the definition of Υ_{K_n} in (3.7), we upper bound $\|\Upsilon_{K_n}\|_*$ as

$$\|\Upsilon_{K_n}\|_*\leq\|Q\|_*+\|R\|_*+\big(\|A\|_{\mathrm{F}}+\|B\|_{\mathrm{F}}\big)^2\cdot\|P_{K_n}\|_*$$
$$\leq\|Q\|_*+\|R\|_*+\big(\|A\|_{\mathrm{F}}+\|B\|_{\mathrm{F}}\big)^2\cdot J_1(K_0)\cdot\sigma_{\min}^{-1}(\Psi_\epsilon), \tag{F.50}$$

where the last line comes from the fact that $\|P_{K_n}\|_*\leq J_1(K_n)\cdot\sigma_{\min}^{-1}(\Psi_\epsilon)\leq J_1(K_0)\cdot\sigma_{\min}^{-1}(\Psi_\epsilon)$. As for the term $\|\Phi_{\widetilde{K}_{n+1}}\|_*$ in (F.48), we upper bound it as

$$\|\Phi_{\widetilde{K}_{n+1}}\|_*\leq J_1(\widetilde{K}_{n+1})\cdot\sigma_{\min}^{-1}(Q)\leq J_1(K_0)\cdot\sigma_{\min}^{-1}(Q). \tag{F.51}$$

Therefore, combining (F.48), (F.49), (F.50) and (F.51), we know that

$$\begin{aligned}&\|\widetilde{K}_{n+1} - K_{n+1}\|_* \cdot \big(\|A - B\widetilde{K}_{n+1}\|_* + 1\big) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_* \\ &\quad \leq \mathrm{poly}_1\big(\|K_n\|_*, J_1(K_0)\big) \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}}.\end{aligned}$$

Also, by the above results, the RHS of (F.47) is upper bounded as

$$\begin{aligned}&6 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_* \cdot \|\widetilde{K}_{n+1}\|_* \cdot \|R\|_* \cdot \|\widetilde{K}_{n+1} - K_{n+1}\|_* \\ &\qquad \cdot \big(\|B\|_* \cdot \|\widetilde{K}_{n+1}\|_*) \cdot \|A - B\widetilde{K}_{n+1}\|_* + \|B\|_* \cdot \|\widetilde{K}_{n+1}\|_* + 1\big) \\ &\quad \leq \mathrm{poly}_2\big(\|K_n\|_*, J_1(K_0)\big) \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}}.\end{aligned}$$

Note that from the policy evaluation theorem B.8, we know that with probability at least $1 - T^{-4} - \widetilde{T}^{-6}$, it holds that

$$\begin{aligned}\|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}} \leq\ & \frac{\mathrm{poly}_3\big(\|K_n\|_{\mathrm{F}}, \|\mu\|_2\big)}{\lambda_{K_n} \cdot (1-\rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} \\ &+ \frac{\mathrm{poly}_4\big(\|K_n\|_{\mathrm{F}}, \|b^{K_n}\|_2, \|\mu\|_2\big)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \widetilde{T}_n}{\widetilde{T}_n^{1/8} \cdot (1-\rho)}.\end{aligned}$$

By choosing T_n and $\widetilde{T}_n$ such that

$$\begin{aligned}&\frac{\mathrm{poly}_3\big(\|K_n\|_{\mathrm{F}}, \|\mu\|_2\big)}{\lambda_{K_n} \cdot (1-\rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} + \frac{\mathrm{poly}_4\big(\|K_n\|_{\mathrm{F}}, \|b^{K_n}\|_2, \|\mu\|_2\big)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \widetilde{T}_n}{\widetilde{T}_n^{1/8} \cdot (1-\rho)} \\ &\quad \leq \max\bigg\{\Big[\mathrm{poly}_1\big(\|K_n\|_*, J_1(K_0)\big)\Big]^{-1} \cdot \sigma_{\min}(\Psi_\omega)/4 \cdot \|B\|_*^{-1}, \\ &\qquad \Big[\mathrm{poly}_2\big(\|K_n\|_*, J_1(K_0)\big)\Big]^{-1} \cdot \varepsilon/2 \cdot \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1} \cdot \|\Psi_\epsilon\|_{\mathrm{F}}^{-1}\bigg\},\end{aligned}$$

in other words, we pick

$$\begin{aligned}T_n &\geq \mathrm{poly}_5\big(\|K_n\|_{\mathrm{F}}, \|b^{K_n}\|_2, \|\mu\|_2\big) \cdot \lambda_{K_n}^{-4} \cdot \big[1 - \rho(A - BK_n)\big]^{-9} \cdot \varepsilon^{-5}, \\ \widetilde{T}_n &\geq \mathrm{poly}_6\big(\|K_n\|_{\mathrm{F}}, \|b^{K_n}\|_2, \|\mu\|_2\big) \cdot \lambda_{K_n}^{-2} \cdot \big[1 - \rho(A - BK_n)\big]^{-12} \cdot \varepsilon^{-12},\end{aligned}$$

then we know that (F.46) holds with probability at least $1 - \varepsilon^{10}$ for sufficiently small $\varepsilon > 0$. This corresponds to the parameters that we choose in the statement of Theorem B.4. By applying Lemma F.2, combining (F.45), we derive that

$$\big|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\big| \leq \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_*^{-1} \cdot \varepsilon/4$$

holds with probability at least $1 - \varepsilon^{15}$. By this, we finish the proof of the lemma. □

F.11 Proof of Lemma F.1

Proof. First, note that the cost function $c(x, u)$ is written in the following way

$$c(x,u) = \psi(x,u)^\top \begin{pmatrix} \mathrm{svec}\big[\mathrm{diag}(Q,R)\big] \\ 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} + ((\mu_{K,b})^\top Q \mu_{K,b} + (\mu_{K,b}^u)^\top R \mu_{K,b}^u + \mu^\top \overline{Q} \mu).$$

For any test matrix V and vectors v_x, v_u, we proceed by the following calculations

$$\begin{aligned}
&\mathbb{E}_{\pi_{K,b^K}}\big[c(x,u)\psi(x,u)\big]^\top \begin{pmatrix}\mathrm{svec}(V)\\ v_x\\ v_u\end{pmatrix}\\
&\quad= \mathbb{E}_{\pi_{K,b^K}}\left\{\psi(x,u)^\top \begin{pmatrix}\mathrm{svec}\big[\mathrm{diag}(Q,R)\big]\\ 2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}\psi(x,u)^\top\begin{pmatrix}\mathrm{svec}(V)\\ v_x\\ v_u\end{pmatrix}\right\}\\
&\qquad+\mathbb{E}_{\pi_{K,b^K}}\left\{\psi(x,u)^\top\big((\mu_{K,b})^\top Q\mu_{K,b}+(\mu^u_{K,b})^\top R\mu^u_{K,b}+\mu^\top\overline{Q}\mu\big)\begin{pmatrix}\mathrm{svec}(V)\\ v_x\\ v_u\end{pmatrix}\right\} =: D_1+D_2.
\end{aligned}\tag{F.52}$$

In the sequel, we calculate D_1 and D_2 respectively.

Calculation of D_1. Note that from the definition of $\psi(x,u)$ in (B.5), we calculate D_1 as

$$\begin{aligned}
D_1 &= \mathbb{E}_{\pi_{K,b^K}}\left\{\left[(z-\mu_z)^\top\mathrm{diag}(Q,R)(z-\mu_z)+(z-\mu_z)^\top\begin{pmatrix}2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}\right]\right.\\
&\qquad\left.\cdot\left[(z-\mu_z)^\top V(z-\mu_z)+(z-\mu_z)^\top\begin{pmatrix}v_x\\ v_u\end{pmatrix}\right]\right\}\\
&= \mathbb{E}_{\pi_{K,b^K}}\big[(z-\mu_z)^\top\mathrm{diag}(Q,R)(z-\mu_z)\cdot(z-\mu_z)^\top V(z-\mu_z)\big]\\
&\qquad+\mathbb{E}_{\pi_{K,b^K}}\left[\begin{pmatrix}2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}^\top(z-\mu_z)(z-\mu_z)^\top\begin{pmatrix}v_x\\ v_u\end{pmatrix}\right].
\end{aligned}\tag{F.53}$$

Here $z=(x^\top,u^\top)^\top$ and $\mu_z=\mathbb{E}_{\pi_{K,b^K}}(z)$. For the first term on the RHS of (F.53), note that $z-\mu_z$ is a centralized Gaussian, whose covariance matrix is Σ_z; therefore, by Lemma G.1, we obtain that

$$\begin{aligned}
&\mathbb{E}_{\pi_{K,b^K}}\big[(z-\mu_z)^\top\mathrm{diag}(Q,R)(z-\mu_z)\cdot(z-\mu_z)^\top V(z-\mu_z)\big]\\
&\quad= 2\big\langle\Sigma_z\mathrm{diag}(Q,R)\Sigma_z,V\big\rangle+\big\langle\Sigma_z,\mathrm{diag}(Q,R)\big\rangle\cdot\langle\Sigma_z,V\rangle\\
&\quad= \mathrm{svec}\Big[2\Sigma_z\mathrm{diag}(Q,R)\Sigma_z+\big\langle\Sigma_z,\mathrm{diag}(Q,R)\big\rangle\cdot\Sigma_z\Big]^\top\mathrm{svec}(V).
\end{aligned}$$

Moreover, the second term on the RHS of (F.53) is calculated as

$$\mathbb{E}_{\pi_{K,b^K}}\left[\begin{pmatrix}2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}^\top(z-\mu_z)(z-\mu_z)^\top\begin{pmatrix}v_x\\ v_u\end{pmatrix}\right]=\left[\Sigma_z\begin{pmatrix}2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}\right]^\top\begin{pmatrix}v_x\\ v_u\end{pmatrix}.$$

Combining the above two equations and (F.53), we obtain that

$$D_1=\begin{pmatrix}2\mathrm{svec}\big[\Sigma_z\mathrm{diag}(Q,R)\Sigma_z+\langle\Sigma_z,\mathrm{diag}(Q,R)\rangle\Sigma_z\big]\\ \Sigma_z\begin{pmatrix}2Q\mu_{K,b}\\ 2R\mu^u_{K,b}\end{pmatrix}\end{pmatrix}^\top\begin{pmatrix}\mathrm{svec}(V)\\ v_x\\ v_u\end{pmatrix}.\tag{F.54}$$

Calculation of D_2: By the definition of the feature vector $\psi(x,u)$ in (B.5), we know that

$$D_2=\big((\mu_{K,b})^\top Q\mu_{K,b}+(\mu^u_{K,b})^\top R\mu^u_{K,b}+\mu^\top\overline{Q}\mu\big)\begin{pmatrix}\mathrm{svec}(\Sigma_z)\\ \mathbf{0}_m\\ \mathbf{0}_k\end{pmatrix}^\top\begin{pmatrix}\mathrm{svec}(V)\\ v_x\\ v_u\end{pmatrix}.\tag{F.55}$$

Now, combining (F.52), (F.54) and (F.55), we conclude the lemma. □

G AUXILIARY RESULTS

Lemma G.1. Assume that the random variable $w \sim \mathcal{N}(0, I)$, and let U and V be two symmetric matrices, then it holds that

$$\mathbb{E}[w^\top U w \cdot w^\top V w] = 2\,\mathrm{tr}(UV) + \mathrm{tr}(U) \cdot \mathrm{tr}(V).$$

Proof. See Magnus et al. (1978); Magnus (1979) for a detailed proof. □

Lemma G.2. Let M, N be commuting symmetric matrices, and let $\alpha_1, \ldots, \alpha_n$, $\beta_1, \ldots, \beta_n$ denote their eigenvalues with $v_1, \ldots, v_n$ a common basis of orthogonal eigenvectors. Then the $n(n+1)/2$ eigenvalues of $M \otimes_s N$ are given by $(\alpha_i \beta_j + \alpha_j \beta_i)/2$, where $1 \le i \le j \le n$.

Proof. See Lemma 2 in Alizadeh et al. (1998) for a detailed proof. □

Lemma G.3. For any integer $m > 0$, let $A \in \mathbb{R}^{m \times m}$ and $\eta \sim \mathcal{N}(0, I_m)$. Then, there exists some absolute constant $C > 0$ such that for any $t \ge 0$, we have

$$\mathbb{P}\Big[\big|\eta^\top A \eta - \mathbb{E}(\eta^\top A \eta)\big| > t\Big] \le 2 \cdot \exp\Big[-C \cdot \min\big(t^2 \|A\|_{\mathrm{F}}^{-2},\ t\|A\|_*^{-1}\big)\Big].$$

Proof. See Rudelson et al. (2013) for a detailed proof. □