**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 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](#) [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 aggregate 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., 2007](#)).
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 $\mu$, approximately attains the optimal policy $\pi^*_\mu$ of each agent, and then updates the mean-field state $\mu$ assuming that all the agents follow $\pi^*_\mu$. 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 $\mu$, the sequence of policies generated by the actor converges linearly to the optimal policy $\pi^*_\mu$. 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., Gueant 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 Ba şar, 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; Embine 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 under given mean-field state. 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., 2016; 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 Mei, 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 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 $\|M\|_2$ 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 $\ell_2$-norm of a vector $\alpha$, and $\{\alpha_i\}_k^j$ to denote the sub-vector $(\alpha_i, \alpha_{i+1}, \ldots, \alpha_j)^\top$, where $\alpha_k$ is the $k$-th entry of the vector $\alpha$. For scalars $a_1, \ldots, a_n$, we denote by 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 $N_a$-player game involves $N_a \in \mathbb{N}$ agents, whose state transitions are given by

$$x_{t+1}^i = Ax_{t}^i + Bu_{t}^i + \overline{Ax}_{t} + d^i + \omega_t^i, \quad \forall t \geq 0, \ i \in [N_a].$$

Here $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{m \times k}$, and $\overline{A} \in \mathbb{R}^{m \times m}$ are matrices, $x_t^i \in \mathbb{R}^m$ 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, $\omega_t^i \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi)$, and $\overline{A} = 1/N_a \cdot \sum_{j=1}^{N_a} x_t^j$ is the mean-field state. The agents are coupled through the mean-field state $\overline{A}$. In the linear-quadratic mean-field $N_a$-player game, the cost of agent $i \in [N_a]$ at time $t \geq 0$ is given by

$$c_t^i = (x_{t}^i)^\top Q x_{t}^i + (u_{t}^i)^\top R u_{t}^i + \pi_t^\top \overline{Q} \overline{x}_t,$$

where $Q \in \mathbb{R}^{m \times m}$, $R \in \mathbb{R}^{k \times k}$, and $\overline{Q} \in \mathbb{R}^{m \times m}$ are matrices, and $u_t^i$ is generated by $\pi^i$, i.e., the policy of agent $i$. To measure the performance of agent $i$ following its policy $\pi^i$ under the influence of the other agents, we define the expected total cost of agent $i$ as

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

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

$$J^i(\pi^1, \ldots, \pi^{i-1}, \pi^i, \pi^{i+1}, \ldots, \pi^{N_a}) \leq J^i(\pi^1, \ldots, \pi^{i-1}, \pi^i, \pi^{i+1}, \ldots, \pi^{N_a}), \quad \forall \pi^i, \ i \in [N_a].$$

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. We consider taking the infinite-population limit $N_a \to \infty$, where each agent has an infinitesimal contribution to the dynamics of the system. Thus, the joint policy of all the agents except agent $i$ can be modeled as a mean-field policy $\pi^i$, and all the agents following such a mean-field policy $\pi^i$ generate the mean-field state $\mathbb{E} x_t^i$, where $\{x_t^i\}_{t \geq 0}$ is generated following the policy $\pi^i$. 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.

Before we formally present the formulation of linear-quadratic mean-field games, we first introduce the following mean-field LQR (MF-LQR) problem, which aims to find an optimal policy for the fixed agent given the mean-field policy $\pi^i$.

**Problem 2.1 (MF-LQR).** Given the mean-field policy $\pi^i$, we consider the following formulation,

$$x_{t+1} = Ax_t + Bu_t + \mathbb{E}x_t + d + \omega_t,$$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + (\mathbb{E}x_t)^\top \overline{Q} (\mathbb{E}x_t),$$

$$J(\pi, \pi^i) = \lim_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T} c(x_t, u_t) \right],$$

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 $\pi$, $\{x_t^i\}_{t \geq 0}$ is the trajectory generated by the policy $\pi^i$, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi)$, and $d \in \mathbb{R}^m$ is a drift term. Here the expectation $\mathbb{E} x_t^i$ is taken across all the agents. We aim to find $\pi^*$ such that $J(\pi^*, \pi^i) = \inf_{\pi \in \Pi} J(\pi, \pi^i)$.

Note that a controllable linear system using linear quadratic optimal control is always stable. Further combining the fact that our linear closed-loop dynamics in Problem 2.1 is driven by the Gaussian noise term $\omega_t$, we know that the Markov chain of states generated by the policy $\pi^i$ admits a stationary distribution and converges to this stationary distribution. This implies that the mean-field state $\mathbb{E} x_t^i$ converges to a constant vector $\mu^i_t$ as $t \to \infty$, which serves as a time-invariant mean-field state. As we consider the ergodic setting, it then suffices to study Problem 2.1 with $t$ sufficiently large. Therefore, the influence of the mean-field policy $\pi^i$ is captured by the mean-field state $\mu^i_t$. By re-formulating Problem 2.1 with slight abuse of notations, we obtain the following drifted-LQR (D-LQR).
Problem 2.2 (D-LQR). Given a mean-field state $\mu \in \mathbb{R}^m$, we consider the following formulation,

$$
x_{t+1} = Ax_t + Bu_t + \mathcal{A}_t \mu + d + \omega_t,
$$

$$
c_\mu(x_t, u_t) = x_t^T Q x_t + u_t^T R u_t + \mu^T Q \mu,
$$

$$
J_\mu(\pi) = \lim_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T} c_\mu(x_t, u_t) \right],
$$

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 $\pi$, $\omega_t \in \mathbb{R}^m$ 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 $\pi^*_\mu$ such that $J_\mu(\pi^*_\mu) = \inf_{\pi \in \Pi} J_\mu(\pi)$.

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 $\pi^*_\mu$ of Problem 2.2 admits a linear form $\pi^*_\mu(x_t) = -K_{\pi^*_\mu} x_t + b_{\pi^*_\mu}$ under certain regularity conditions [Anderson and Moore, 2007], where the matrix $K_{\pi^*_\mu} \in \mathbb{R}^{m \times m}$ and the vector $b_{\pi^*_\mu} \in \mathbb{R}^k$ are the parameters of $\pi^*_\mu$. Motivated by such a linear form of the optimal policy, we define the class of linear-Gaussian policies as

$$
\Pi = \{ \pi(x) = -K_x x + b_x + \sigma \cdot \eta; \ K_x \in \mathbb{R}^{k \times m}, b_x \in \mathbb{R}^k \},
$$

where $\sigma \in \mathbb{R}$ and the standard Gaussian noise term $\eta \in \mathbb{R}^k$ is included to encourage exploration. To solve Problem 2.2, it suffices to find the optimal policy $\pi^*_\mu$ within $\Pi$. We define $\Lambda_1(\mu) = \pi^*_\mu$ as the optimal policy under the mean-field state $\mu$.

Assume that all the agents follow the linear policy $\pi(x) = -K_x x + b_x$ under the mean-field state $\mu$. By plugging $u_t = \pi(x_t)$ into the state transition in Problem 2.2 as $t \to \infty$, we know that these agents generate a new mean-field state $\mu_{\text{new}}$ such that

$$
\mu_{\text{new}} = (I - A + BK_x)^{-1} (Bb_x + A\mu + d).
$$

We define $\Lambda_2(\mu, \pi) = \mu_{\text{new}}$ as such a new mean-field state.

Now, we are ready to present the following linear-quadratic mean-field game (LQ-MFG).

Problem 2.3 (LQ-MFG). We consider the following formulation,

$$
x_{t+1} = Ax_t + Bu_t + \mathcal{A}_t \mu + d + \omega_t,
$$

$$
c(x_t, u_t) = x_t^T Q x_t + u_t^T R u_t + \mu^T Q \mu,
$$

$$
J(\pi, \mu) = \lim_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T} c(x_t, u_t) \right],
$$

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 $\pi$, $\mu \in \mathbb{R}^m$ is the mean-field state, $\omega_t \in \mathbb{R}^m$ 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 a pair $(\mu^*, \pi^*)$ such that (i) $J(\pi^*, \mu^*) = \inf_{\pi \in \Pi} J(\pi, \mu^*)$; (ii) $x_t^* \to \mu^*$ as $t \to \infty$, where $\{x_t^*\}_{t \geq 0}$ is the Markov chain of states generated by the policy $\pi^*$.

The formulation in Problem 2.3 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.2 (see §C of the appendix for details), where an additional interaction term between the state vector $x_t$ and the mean-field state $\mu$ is incorporated into 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.2. Therefore, for the sake of simplicity, we focus on Problem 2.3 in the sequel.

In Problem 2.3, condition (i) is equivalent to the optimality of the policy $\pi^*$ under the mean-field state $\mu^*$, namely, $\Lambda_1(\mu^*) = \pi^*$. Meanwhile, condition (ii) is equivalent to the invariance of the mean-field state $\mu^*$ given the policy $\pi^*$, namely, $\Lambda_2(\mu^*, \pi^*) = \mu^*$. Such equivalence follows from the NCE principle [Huang et al., 2006, 2007], which also motivates the following definition of the Nash equilibrium pair $\{Saldi et al., 2018a,b\}$. 
Definition 2.4 (Nash Equilibrium Pair). The pair \((\mu^*, \pi^*) \in \mathbb{R}^m \times \Pi\) constitutes a Nash equilibrium pair of Problem 2.3 if it satisfies \(\pi^* = \Lambda_1(\mu^*)\) and \(\mu^* = \Lambda_2(\mu^*, \pi^*)\). Here \(\mu^*\) is called the Nash mean-field state and \(\pi^*\) is called the Nash policy.

By Definition 2.4, Problem 2.3 aims to find a Nash equilibrium pair \((\mu^*, \pi^*)\).

3 MEAN-FIELD ACTOR-CRITIC

We first characterize the existence and uniqueness of the Nash equilibrium pair of Problem 2.3 under mild regularity conditions, and then propose a mean-field actor-critic algorithm to obtain such a Nash equilibrium. As a building block of the mean-field actor-critic, we propose the natural actor-critic 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 2.4. We impose the following regularity conditions.

Assumption 3.1. We assume that the following statements hold:

(i) The algebraic Riccati equation \(X = A^T X A + Q - A^T X B (B^T X B + R)^{-1} B^T X A\) admits a unique symmetric positive definite solution \(X^*\);
(ii) It holds for \(L_0 = L_1 L_3 + L_2\) that \(L_0 < 1\), where

\[
L_1 = \|[I - A]Q^{-1}(I - A)^T + BR^{-1}B^T\]^{-1}A\|_2; \|\[K^*Q^{-1}(I - A)^T - R^{-1}B^T\]\|_2,
\[
L_2 = |1 - \rho(A - BK^*)|^{-1} \cdot \|A\|_2, \quad L_3 = |1 - \rho(A - BK^*)|^{-1} \cdot \|B\|_2.
\]

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

The first assumption is implied by mild regularity conditions on the matrices \(A, B, Q,\) and \(R\), which are (1) the positivity of \(R\); (2) the non-negativity of \(Q = C^T C\); (3) the observability of \((A, C)\); (4) the stability of \((A, B)\). See De Souza et al. (1986); Lewis et al. (2012) for more details. The second assumption is standard in the literature (Bensoussan et al., 2016; Saldi et al., 2018b), which ensures the stability of the LQ-MFG. In the following proposition, we show that Problem 2.3 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 \((\mu^*, \pi^*)\) of Problem 2.3.

Proof. See E.1 for a detailed proof. \(\square\)

3.2 MEAN-FIELD ACTOR-CRITIC FOR LQ-MFG

The NCE principle motivates a fixed-point approach to solve Problem 2.3, which generates a sequence of policies \(\{\pi_s\}_{s \geq 0}\) and mean-field states \(\{\mu_s\}_{s \geq 0}\) satisfying the following two properties: (i) Given the mean-field state \(\mu_s\), the policy \(\pi_s\) is optimal. (ii) The mean-field state becomes \(\mu_{s+1}\) as \(t \to \infty\), if all the agents follow \(\pi_s\) under the current mean-field state \(\mu_s\). Here (i) requires solving Problem 2.2 given the mean-field state \(\mu_s\), while (ii) requires simulating the agents following the policy \(\pi_s\) given the current mean-field \(\mu_s\). Based on such properties, we propose the mean-field actor-critic in Algorithm 1.

Algorithm 1 requires solving Problem 2.2 at each iteration to obtain \(\pi_s = \Lambda_1(\mu_s)\) and \(\mu_{s+1} = \Lambda_2(\mu_s, \pi_s)\). To this end, we introduce the natural actor-critic 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 \(\mu\), we thus drop the subscript \(\mu\) hereafter. With slight abuse of notations, we write \(\pi_{K,b}(x) = -Kx + b + \sigma \cdot \eta\) to emphasize
Second, we show that \( \min \) induced by (3.1) has a unique stationary distribution \( \pi \) for any policy \( \text{actor-critic} \) to solve Problem 2.2.

First, we show that \( J(K, b) = J(\pi_{K,b}) \) consequently. Now, we propose the natural actor-critic to solve Problem 2.2.

For any policy \( \pi_{K,b} \in \Pi \), by the state transition in Problem 2.2, we have
\[
x_{t+1} = (A - BK)x_t + (Bb + \bar{A}\mu + d) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Psi_\epsilon),
\]
(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 \( \mathcal{N}(\mu_{K,b}, \Phi_K) \) (Anderson and Moore, 2007), where the mean-field state \( \mu_{K,b} \) and the covariance \( \Phi_K \) satisfy that
\[
\mu_{K,b} = (I - A + BK)^{-1}(Bb + \bar{A}\mu + d),
\]
(3.2)
\[
\Phi_K = (A - BK)\Phi_K(A - BK)^\top + \Psi_\epsilon.
\]
(3.3)

Meanwhile, the Bellman equation for Problem 2.2 takes the following form
\[
J(K, b) = J_1(K) + J_2(K, b) + \sigma^2 \cdot \text{tr}(R) + \mu^\top Q\mu,
\]
(3.5)
where \( J_1(K) \) and \( J_2(K, b) \) are defined as
\[
J_1(K) = \text{tr}[(Q + K^\top RK)\Phi_K] = \text{tr}(P_K \Psi_\epsilon),
\]
\[
J_2(K, b) = (\frac{\mu_{K,b}}{b})^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}.
\]
(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 \( \nu_K \)-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 \( t_K \)-Lipschitz continuous gradient in \( b \), where \( t_K \) is upper bounded as \( t_K \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_2 \cdot \|Q\|_2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2) \).

**Proof.** See [4] for a detailed proof. \( \Box \)

Second, we show that \( \min_b J_2(K, b) \) is independent of \( K \).
Proposition 3.4. We define $b^K = \arg\min_b J_2(K, b)$, where $J_2(K, b)$ is defined in (3.6). It holds that
\[ b^K = [KQ^{-1}(I - A)\top - R^{-1}B\top] : [(I - A)Q^{-1}(I - A)\top + BR^{-1}B\top]^{-1} \cdot (\overline{A}\mu + d). \]
Moreover, $J_2(K, b^K)$ takes the form of
\[ J_2(K, b^K) = (\overline{A}\mu + d)\top [(I - A)Q^{-1}(I - A)\top + BR^{-1}B\top]^{-1} \cdot (\overline{A}\mu + d), \]
which is independent of $K$.

Proof. See [E.2] for a detailed proof.

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 $\arg\min_K J_1(K)$. 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 $\Upsilon_K$, $p_{K,b}$, and $q_{K,b}$ as
\[
\Upsilon_K = \begin{pmatrix} Q + A\top P_K A & A\top P_K B \\ B\top P_K A & R + B\top P_K B \end{pmatrix} = \begin{pmatrix} \Upsilon_{K11} & \Upsilon_{K12} \\ \Upsilon_{K21} & \Upsilon_{K22} \end{pmatrix},
\]
\[ p_{K,b} = A\top [P_K \cdot (\overline{A}\mu + d) + f_{K,b}], \quad q_{K,b} = B\top [P_K \cdot (\overline{A}\mu + d) + f_{K,b}], \]
where $f_{K,b} = (I - A + BK)\top [(A - BK)\top P_K (Bb + \overline{A}\mu + d) - K\top Rb]$. 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
\[ \nabla_K J_1(K) = 2(\Upsilon_{K12}^2 K - \Upsilon_{K12}^2) \cdot \Phi_K, \quad \nabla_b J_2(K, b) = \Upsilon_{K12}^2 (K\mu_{K,b} + b) + \Upsilon_{K12}^2 K\mu_{K,b} + q_{K,b}. \]

Our algorithm follows the natural actor-critic method ([Bhatnagar et al. 2009]) and actor-critic method ([Konda and Tsitsiklis 2000]). Specifically, (i) To obtain the optimal $K^*$, in the critic update step, we estimate the matrix $\hat{\Upsilon}_K$ by $\hat{\Upsilon}_K$ via a policy evaluation algorithm, e.g., Algorithm 3 or Algorithm 4 (see [B.2] and [B.3] of the appendix for details); in the actor update step, we update $K$ via $K \leftarrow K - \gamma \cdot (\hat{\Upsilon}_{K12}^2 K - \hat{\Upsilon}_{K12}^2)$, where the term $\hat{\Upsilon}_{K12}^2 K - \hat{\Upsilon}_{K12}^2$ is the estimated natural gradient. (ii) To obtain the optimal $b^*$ given $K^*$, in the critic update step, we estimate $\hat{\Upsilon}_{K^*}, q_{K^*,b}$, and $\mu_{K^*,b}$ by $\hat{\Upsilon}_{K^*}, \hat{q}_{K^*,b}$, and $\hat{\mu}_{K^*,b}$ via a policy evaluation algorithm; In the actor update step, we update $b$ via $b \leftarrow b - \gamma \cdot \nabla_b J_2(K^*, b)$, where $\nabla_b J_2(K^*, b) = \hat{\Upsilon}_{K12}^2 (-K^*\hat{\mu}_{K^*,b} + b) + \hat{\Upsilon}_{K12}^2 \hat{\mu}_{K^*,b} + \hat{\mu}_{K^*,b}$ is the estimated gradient. Combining the above procedure, we obtain the natural actor-critic for Problem 2.2 which is stated in Algorithm 2.

One may want to apply gradient method to $J(K, b)$ directly in the joint space of $K$ and $b$. However, the gradient dominance property of $J_1(K)$ in the most studied LQR problem ([Yang et al. 2019]) no longer holds for $J(K, b)$. Therefore, the convergence of the gradient method to $J(K, b)$ is not guaranteed in our problem.

### 4 Global Convergence Results

The following theorem establishes the rate of convergence of Algorithm 1 to the Nash equilibrium pair $(\mu^*, \pi^*)$ of Problem 2.3.

**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
\[
S > \frac{\log(\|\mu_0 - \mu^*\|_2 \cdot \varepsilon^{-1})}{\log(1/L_0)}.
\]

For any $s \in [S]$, we define
\[
\varepsilon_s = \min \left\{ \left[1 - \rho(A - BK^*)\right]^4 \cdot \|B\|_2 + \|\overline{A}\|_2^{-4} \cdot \|\mu_s\|_2^{-2} \cdot \|d\|_2^{-2} \cdot \sigma_{\min}(\Psi_s) \cdot \sigma_{\min}(R) \cdot \varepsilon^2, \right. \\
\left. \nu_{K^*} \cdot \left[1 - \rho(A - BK^*)\right]^4 \cdot \|B\|_2^{-2} \cdot M_b(\mu_s) \cdot \varepsilon^2, \varepsilon \right\} \cdot 2^{-s-10},
\]

Under review as a conference paper at ICLR 2020
Algorithm 2 Natural Actor-Critic Algorithm for D-LQR.

1: **Input:**
   - Mean-field state \( \mu \) and initial policy \( \pi_{0,b_0} \).
   - Numbers of iterations \( N, H, \{\bar{T}_n, T_n\}_{n \in [N]}, \{\bar{T}_h^b, T_h^b\}_{h \in [H]} \).
   - Stepsizes \( \gamma, \gamma^b \), \( \{\gamma_n\}_{n \in [N]}, \{\gamma_h^b\}_{h \in [H]} \).

2: for \( n = 0, 1, 2, \ldots, N - 1 \) do
   3: **Critic Update:** Compute \( \bar{T}_{K_n} \) via Algorithm 3 with \( \pi_{K_n,b_n}, \mu, \bar{T}_n, T_n, \{\gamma_n\}_{t \in [T_n]}, K_0 \), and \( b_0 \) as inputs.
   4: **Actor Update:** Update the parameter via
      \[
      K_{n+1} \leftarrow K_n - \gamma \cdot (\bar{T}_{K_n}^2 K_n - \bar{T}_{K_n}^{21}).
      \]

5: end for

6: for \( h = 0, 1, 2, \ldots, H - 1 \) do
   7: **Critic Update:** Compute \( \hat{\mu}_{K_n,b_n}, \bar{T}_{K_n}, \bar{q}_{K_n,b_n} \) via Algorithm 3 with \( \pi_{K_n,b_n}, \mu, \bar{T}_h^b, T_h^b, \{\gamma_h^b\}_{t \in [T_h^b]}, K_0 \), and \( b_0 \).
   8: **Actor Update:** Update the parameter via
      \[
      b_{h+1} \leftarrow b_h - \gamma^b \cdot [\bar{T}_{K_n}^{22} (-K_N \hat{\mu}_{K_n,b_n} + b_h) + \bar{T}_{K_n}^{21} \hat{\mu}_{K_n,b_n} + \bar{q}_{K_n,b_n}].
      \]

9: end for

10: **Output:** Policy \( \pi_{K,h} = \pi_{K,b,h} \), estimated mean-field state \( \hat{\mu}_{K,h} = \hat{\mu}_{K,b,h} \).

where \( \nu_{K,*} \) is defined in Proposition 3.3 and
\[
M_b(\mu_s) = 4 \left\| Q^{-1} (I - A)^\top \cdot \left[ (I - A) Q^{-1} (I - A)^\top + BR^{-1} B^\top \right]^{-1} \cdot (\bar{A} \mu_s + d) \right\|_2
\cdot \left[ \nu_{K,*}^{-1} + \sigma_{\min}(\Psi_s) \cdot \sigma_{\min}(R) \right]^{1/2}.
\]

In the \( s \)-th policy update step in Line 3 of Algorithm 2, we set the inputs via Theorem 3.4 such that \( J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi_{s+1}^*) < \varepsilon \), where the expected total cost \( J_{\mu_s}(\cdot) \) is defined in Problem 2.2, and \( \pi_{s+1}^* = \Lambda_1(\mu_s) \) is the optimal policy under the mean-field state \( \mu_s \). Then it holds with probability at least \( 1 - \varepsilon^6 \) that
\[
\| \mu_s - \mu^* \|_2 \leq \varepsilon, \quad \| K_s - K^* \|_F \leq \varepsilon, \quad \| b_s - b^* \|_2 \leq (1 + L_1) \cdot \varepsilon.
\]

Here \( \mu^* \) is the Nash mean-field state, \( K_s \) and \( b_s \) are parameters of the policy \( \pi_s \), and \( K^* \) and \( b^* \) are parameters of the Nash policy \( \pi^* \).

**Proof Sketch.** The proof of the theorem is based on the convergence of the natural actor-critic algorithm and a contraction argument. First, we prove in Theorem 3.4 that Algorithm 2 converges linearly to the optimal policy of Problem 2.2. By this, in each iteration of Algorithm 2, we control the error between \( \mu_{s+1} \) and \( \mu_{s+1}^* \) to be \( \varepsilon \), with high probability, where \( \mu_{s+1}^* \) is the mean-field state generated by the optimal policy \( \Lambda_1(\mu_s) \); in other words, \( \mu_{s+1}^* = \Lambda(\mu_s) \). Combining the fact from Proposition 3.2 that \( \Lambda(\cdot) \) is a contraction, we deduce that
\[
\| \mu_{s+1} - \mu^* \|_2 \leq \| \Lambda(\mu_s) - \Lambda(\mu^*) \|_2 + \varepsilon \leq L_0 \cdot \| \mu_s - \mu^* \|_2 + \varepsilon
\]

with high probability, where \( \varepsilon > 0 \) is some error term and is specified in the detailed proof. Moreover, by telescoping sum, and note that the sum \( \sum_{t=1}^n \varepsilon_s \) is upper bounded by the desired error \( \varepsilon \), we conclude the theorem. See (D.1) for a detailed proof.

We highlight that if the inputs of Algorithm 2 satisfy the conditions stated in Theorem 3.4, it holds that \( J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi_{s+1}^*) < \varepsilon \) for any \( s \in [S] \). See Theorem 3.4 in (B.1) of the appendix for details. By Theorem 3.4, Algorithm 2 converges linearly to the unique Nash equilibrium pair \( (\mu^*, \pi^*) \) of Problem 2.3. 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 guarantees, 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 shown converges 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. For future research, we aim to extend our algorithm to other variations of mean-field games including risk-sensitive mean-field games (Saldi et al., 2018a; Tembine et al., 2014), robust mean-field games (Bauso et al., 2016), and partially observed mean-field games (Saldi et al., 2019).

REFERENCES


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


15


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 $\|M\|_F$ the Frobenius norm. We also denote by $\text{svec}(M)$ 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 $\text{smat}(\cdot)$ 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 $\pi$.

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

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

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

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

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

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

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

$$Q_{K,b}(x, u) = \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2 \left( p_{K,b}^\top q_{K,b} - \text{tr}(P_K\Phi_K) - \sigma^2 \cdot \text{tr}(R + P_KBB^\top) - \nu^\top Rb + 2b^\top R \mu_{K,b} - (\mu_{K,b})^\top Q_K + \nu^\top R \mu_{K,b} + 2f_{K,b}^\top (A\mu + d) - \mu_{K,b} \right) + (A\mu + d)^\top P_K(A\mu + d), \tag{B.4}$$

where the matrix $\Upsilon_K$ 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

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

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

$$J_1(K) = \text{tr}\left( (Q + K^\top RK)\Phi_K \right) = \text{tr}(P_K\Psi_K), \quad J_2(K, b) = \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}. \tag{B.5}$$

Here $\mu_{K,b}$ is given in (3.7), $\Phi_K$ 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

\[
\nabla_K J_1(K) = 2(\Upsilon_{K}^{21}K - \Upsilon_{K}^{22}) \cdot \Phi_K, \quad \nabla_b J_2(K, b) = 2[\Upsilon_{K}^{22}(-K\mu_{K,b} + b) + \Upsilon_{K}^{21}Kb + q_{K,b}],
\]

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

**Proof.** See \[(B.5)\] 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 \(\pi_{K_0, b_0}\) be stable. In Algorithm 2 for a sufficiently small tolerance \(\varepsilon > 0\), we set

\[
\gamma \leq \left[\|R\|_* + \|B\|_*^2 \cdot J(K_0, b_0) \cdot \sigma_{\min}^{-1}(\Phi_\varepsilon)\right]^{-1},
\]

\[
N \geq C \cdot \|\Phi_K\|_* \cdot \gamma^{-1} \cdot \log \{4[J(K_0, b_0) - J(K^*, b^*)] \cdot \varepsilon^{-1}\},
\]

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

\[
T_n \geq \text{poly}(\|K_n\|_F, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot \lambda_{K_n}^{-4} \cdot [1 - \rho(A - BK_n)]^{-9} \cdot \varepsilon^{-5},
\]

\[
\tilde{T}_n \geq \text{poly}(\|K_n\|_F, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot \lambda_{K_n}^{-2} \cdot [1 - \rho(A - BK_n)]^{-12} \cdot \varepsilon^{-12},
\]

\[
\gamma_{n,t} = \gamma_0 \cdot t^{-1/2},
\]

where \(\gamma_0\) is some positive constant, and \(\lambda_{K_n}\) is specified in Proposition B.6. Also, after obtaining \(K_N\) from Algorithm 2 we set

\[
\gamma^b \leq \min(1 - \rho(A - BK_N), [1 - \rho(A - BK_N)]^{-2} \cdot (\|B\|_*^2 \cdot \|K_N\|_2^2 \cdot \|R\|_* + \|B\|_* \cdot \|Q\|_*)),
\]

\[
H \geq C_0 \cdot \nu_{K_N}^{-1} \cdot (\gamma^b)^{-1} \cdot \log \left\{4[J(K_0, b_0) - J(K_N, b_{KN})] \cdot \varepsilon^{-1}\right\},
\]

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

\[
T_h^b \geq \text{poly}(\|K_N\|_F, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot \lambda_{K_N}^{-4} \cdot \nu_{K_N}^{-4} \cdot [1 - \rho(A - BK_N)]^{-10} \cdot \varepsilon^{-5},
\]

\[
\tilde{T}_h^b \geq \text{poly}(\|K_N\|_F, \|b_0\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot \lambda_{K_N}^{-2} \cdot \nu_{K_N}^{-2} \cdot [1 - \rho(A - BK_N)]^{-16} \cdot \varepsilon^{-8},
\]

\[
\gamma_{h,t} = \gamma_0 \cdot t^{-1/2},
\]

where \(\nu_{K_N}\) is specified in Proposition 3.3. Then with probability at least \(1 - \varepsilon^{10}\), it holds that

\[
J(K_N, b_H) - J(K^*, b^*) < \varepsilon, \|K_N - K^*\|_F \leq [\sigma_{\min}(\Psi_\varepsilon) \cdot \sigma_{\min}(R) \cdot \varepsilon]^{1/2},
\]

and \(\|b_H - b^*\|_2 \leq (2\varepsilon/\nu_{K^*})^{1/2}\).

**Proof.** See \[(B.2)\] for a detailed proof.

By this theorem, given any mean-field state \(\mu\), we show that Algorithm 2 converges linearly to the optimal policy \(\pi_{K^*}\) 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 \(\Upsilon_K\) and the vector \(q_{K,b}\). We now derive a policy evaluation algorithm to estimate these quantities, based on gradient temporal difference algorithm \cite{sutton2009}. For notational convenience, we denote by the feature vector

\[
\psi(x, u) = \begin{pmatrix}
\varphi(x, u) \\
x - \mu_{K,b} \\
u - (-K\mu_{K,b} + b)
\end{pmatrix},
\]

\[(B.5)\]
where

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

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

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

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

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

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:

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

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. \( \square \)

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

\[ \tilde{\Theta}_{K,b} : \zeta = \begin{pmatrix} \mathbb{E}_{\pi_{K,b}} [\psi(x, u)] & 0 \\ \Theta_{K,b} \end{pmatrix} \begin{pmatrix} J(K, b) \\ \alpha_{K,b} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{\pi_{K,b}} [c(x, u) \psi(x, u)] \\ J(K, b) \end{pmatrix}, \tag{B.8} \]

where the matrix \( \tilde{\Theta}_{K,b} \) takes the form

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

Note that in the above linear system \( \text{(B.5)} \), 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 \( 4(1 + \|K\|^2) \|\Phi_K\|^2 \). Moreover, \( \sigma_{\min}(\tilde{\Theta}_{K,b}) \) is lower bounded by a positive constant \( \lambda_K \), where \( \lambda_K \) only depends on \( \sigma, \sigma_{\min}(\Psi_{\omega}) \) and \( \rho(A - BK) \).

**Proof.** See [E.8] for a detailed proof. \( \square \)

Hereafter, for notational convenience, we denote by \( \hat{\psi}_t \) the estimated feature vector \( \hat{\psi}(x_t, u_t) \). 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 \( \text{(B.8)} \) directly, we minimize the following loss function

\[ \left[ \zeta_1 - J(K, b) \right]^2 + \left\| \mathbb{E}_{\pi_{K,b}} [\psi(x, u)] \zeta_1 + \Theta_{K,b} \zeta_2 - \mathbb{E}_{\pi_{K,b}} [c(x, u) \psi(x, u)] \right\|^2. \]
Algorithm 3 Primal-Dual Gradient Temporal Difference Algorithm.

1: **Input:** Policy $\pi_{K,b}$, mean-field state $\mu$, numbers of iteration $\bar{T}$ and $T$, stepsizes $\{\gamma_t\}_{t \in [T]}$, parameters $K_0$ and $b_0$.
2: Define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ via Condition [B.7] with $K_0$ and $b_0$.
3: Initialize the parameters by $\zeta_0 \in \mathcal{V}_\zeta$ and $\xi_0 \in \mathcal{V}_\xi$.
4: Sample $\bar{\xi}_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
5: for $t = 0, \ldots, \bar{T} - 1$ do
6:   Given the mean-field state $\mu$, take action $\tilde{u}_t$ following $\pi_{K,b}$ and generate the next state $\tilde{x}_{t+1}$.
7:   end for
8: Set $\hat{\mu}_{K,b} \leftarrow 1/\bar{T} \cdot \sum_{t=0}^{\bar{T}} \tilde{x}_t$ and compute the estimated feature vector $\hat{\psi}$ via [B.10].
9: Sample $x_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
10: for $t = 0, \ldots, T - 1$ do
11:   Given the mean-field state $\mu$, take action $u_t$ following $\pi_{K,b}$, observe the cost $c_t$, and generate the next state $x_{t+1}$.
12:   Set $\delta_{t+1} \leftarrow \zeta_t^1 + (\hat{\psi}_t - \hat{\psi}_{t+1})^\top \zeta_t^2 - c_t$.
13:   Update parameters via
14:      $$\zeta_{t+1}^1 \leftarrow \zeta_t^1 - \gamma_{t+1} \cdot (\zeta_t^1 + \hat{\psi}_t^\top \zeta_t^2),$$
15:      $$\zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \hat{\psi}_t (\hat{\psi}_t - \hat{\psi}_{t+1})^\top \zeta_t^2,$$
16:      $$\xi_{t+1}^1 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^1 + \gamma_{t+1} \cdot (\xi_t^1 - c_t),$$
17:      $$\xi_{t+1}^2 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^2 + \gamma_{t+1} \cdot \delta_{t+1} \cdot \hat{\psi}_t.$$
18:   Project $\zeta_{t+1}$ and $\xi_{t+1}$ to $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$, respectively.
19: end for
20: Set $\hat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^T \gamma_t)^{-1} \cdot (\sum_{t=1}^T \gamma_t \cdot \xi_t^2)$, and
21:      $$\hat{\Upsilon}_K \leftarrow \text{smat}(\hat{\alpha}_{K,b,1}), \quad \begin{pmatrix} \hat{p}_{K,b} \\ \hat{q}_{K,b} \end{pmatrix} \leftarrow \hat{\alpha}_{K,b,2} - \hat{\Upsilon}_K \begin{pmatrix} \hat{\mu}_{K,b} \\ -K \hat{\mu}_{K,b} + b \end{pmatrix},$$
22: where $\hat{\alpha}_{K,b,1} = (\hat{\alpha}_{K,b})_{1}^{(k+d+1)(k+d)/2}$ and $\hat{\alpha}_{K,b,2} = (\hat{\alpha}_{K,b})_{1}^{(k+d+3)(k+d)/2+1}$.
23: **Output:** Estimators $\hat{\mu}_{K,b}$, $\hat{\Upsilon}_K$, and $\hat{q}_{K,b}$. 
Further, by Fenchel’s duality, we convert the above problem to the following primal-dual min-max problem

\[
\min_{\zeta} \max_{\xi} F(\zeta, \xi) = \left\{ \mathbb{E}_{\pi_{K,b}} [\psi(x, u)] \zeta^1 + \Theta_{K,b} \xi^2 - \mathbb{E}_{\pi_{K,b}} [c(x, u) \psi(x, u)] \right\}^\top \xi^2 \\
+ \left[ \zeta^1 - J(K, b) \right] : \xi^1 - 1/2 \cdot \|\xi\|^2_2, \quad (B.9)
\]

where we restrict the primal variable \(\zeta\) in some compact set \(V_\zeta\) and the dual variable \(\xi\) in some compact set \(V_\xi\), which are detailedly specified in Condition [B.7]. By taking the gradient of the objective in (B.9) with respect to \(\zeta\) and \(\xi\), we obtain the following gradients

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

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 \(\hat{\mu}_{K,b}\) by simulating the MDP under the policy \(\pi_{K,b}\) for \(\tilde{T}\) iterations, and compute the estimated feature vector \(\hat{\psi}\) via

\[
\hat{\psi}(x, u) = \left( \begin{array}{c} \hat{\varphi}(x, u) \\ x - \hat{\mu}_{K,b} \end{array} \right), \quad (B.10)
\]

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

\[
\hat{\varphi}(x, u) = \text{svec} \left( \begin{array}{c} x - \hat{\mu}_{K,b} \\ u - (-K \hat{\mu}_{K,b} + b) \end{array} \right).
\]

Before stating the convergence result of Algorithm [3], we specify the sets \(V_\zeta\) and \(V_\xi\).

**Condition B.7.** Given \(K_0\) and \(b_0\) such that \(\rho(A - BK_0) < 1\). We define the sets \(V_\zeta\) and \(V_\xi\) in (B.9) as

\[
V_\zeta = \left\{ \zeta: 0 \leq \zeta^1 \leq J(K_0, b_0), \|\zeta\|_2 \leq M_{\zeta,1} + M_{\zeta,2} \cdot (1 + \|K\|_F) \cdot \left[ 1 - \rho(A - BK) \right]^{-1} \right\}, \\
V_\xi = \left\{ \xi: \|\xi\|_2 \leq M_\xi \cdot (1 + \|K\|_F^2) \cdot \left[ 1 - \rho(A - BK) \right]^{-1} \right\}.
\]

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

\[
M_{\zeta,1} = \left( (\|Q\| + \|R\|) + (\|A\| + \|B\|) \cdot \sqrt{\mathbb{E} F J(K_0, b_0) / \sigma_{\min}(\Psi_w)} \right) \\
+ \left( (\|Q\| + \|R\|) + (\|A\| + \|B\|) \cdot J(K_0, b_0) / \sigma_{\min}(\Psi_w) \right) \\
\cdot \left( J(K_0, b_0) / \sigma_{\min}(Q) + J(K_0, b_0) / \sigma_{\min}(R) \right) \\
+ (\|A\| + \|B\|) \cdot J(K_0, b_0)^2 \cdot \sigma_{\min}(\Psi_w)^{-1} \cdot \sigma_{\min}(Q)^{-1}, \\
M_{\zeta,2} = (\|A\| + \|B\|) \cdot (\kappa_Q + \kappa_R), \\
M_\xi = C \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot J(K_0, b_0)^2 / \sigma_{\min}^2(Q),
\]

where \(C\) is a positive absolute constant, and \(\kappa_Q\) and \(\kappa_R\) 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 \(\rho(A - BK) < 1\) and \(J(K, b) \leq J(K_0, b_0)\), we define the sets \(V_\zeta\) and \(V_\xi\) through Condition [B.7]. Given the stepsize \(\gamma_t = \gamma_0 t^{-1/2}\) for some constant \(\gamma_0 > 0\), then for any \(\rho\) such that \(\rho(A - BK) < \rho < 1\), when the number of iterations \(T\) are sufficiently large such that \(T \leq \text{poly}_0(\|K\|_F, \|b\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot (1 - \rho)^{-6}\), with probability at least \(1 - T^{-4} - \tilde{T}^{-6}\), the output \(\hat{\alpha}_{K,b}\) satisfies that

\[
\|\hat{\alpha}_{K,b} - \alpha_{K,b}\|^2_2 \leq \lambda_K^2 \cdot \text{poly}(\|K\|_F, \|b\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot \left[ \frac{\log^6 T}{T^{1/2} \cdot (1 - \rho)^6} + \frac{\log \tilde{T}}{T^{1/4} \cdot (1 - \rho)^2} \right].
\]
where $\lambda_K$ is specified in Proposition B.6. Same bounds for $\|\hat{T}_K - T_K\|_F$, $\|\hat{\rho}_{K,b} - \rho_{K,b}\|$ and $\|\hat{q}_{K,b} - q_{K,b}\|_2$ hold. Also, with probability at least $1 - T^{-6}$, it holds that

$$\|\hat{\mu}_{K,b} - \mu_{K,b}\|_2 \leq \frac{\log T}{T^{1/4}} \cdot (1 - \rho)^{-2} \cdot \text{poly}(\|\Phi_K\|_F, \|\Phi_K\|_F, \|b\|_2, \|\mu\|_2, J(K_0, b_0)).$$

**Proof.** See §D.3 for a detailed proof. \qed

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.

1: **Input:** Policy $\pi_{K,b}$, number of iteration $T$ and $T$, stepsizes $\{\gamma_t\}_{t \in [T]}$.
2: Sample $\bar{x}_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
3: for $t = 0, \ldots, T - 1$ do
4:   Take action $\bar{u}_t$ under the policy $\pi_{K,b}$ and generate the next state $\bar{x}_{t+1}$.
5: end for
6: Set $\hat{\mu}_{K,b} \leftarrow 1/T \cdot \sum_{t=1}^{T} \bar{x}_t$.
7: Sample $x_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
8: for $t = 0, \ldots, T$ do
9:   Given the mean-field state $\mu$, take action $\bar{u}_t$ following $\pi_{K,b}$ and generate the next state $\bar{x}_{t+1}$.
10: Set $\delta_{t+1} \leftarrow \zeta_t + (\bar{y}_t - \bar{y}_{t+1})^\top \zeta_t - c_t$.
11: Update parameters via $\zeta_{t+1} \leftarrow (1 - \gamma_{t+1}) \cdot \zeta_t + \gamma_{t+1} \cdot c_t$ and $\zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \delta_{t+1} \cdot \bar{y}_t$.
12: Project $\zeta_t$ to some compact set $\mathcal{Y}_\zeta$.
13: end for
14: Set $\hat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^{T} \gamma_t)^{-1} : (\sum_{t=1}^{T} \gamma_t \cdot \zeta_t^2)$, and
15: Output: Estimators $\hat{\mu}_{K,b}$, $\hat{T}_K$, and $\hat{q}_{K,b}$.

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 2.3, a more general formulation includes an additional term $x_t^\top P \mu$ in the cost function. For the completeness of this paper, we define this general formulation here. Following from a same argument as in [3], it suffices to study the setting where $t$ is sufficiently large. First, we propose the following general drifted LQR (general D-LQR) problem, which is parallel to Problem 2.2.
Problem C.1 (General D-LQR). For any given mean-field state \( \mu \in \mathbb{R}^m \), consider the following formulation

\[
x_{t+1} = Ax_t + Bu_t + \bar{\mu} + d + \omega_t,
\]

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 \( \pi, \omega_t \in \mathbb{R}^m \) is an independent random noise term following the Gaussian distribution \( \mathcal{N}(0, \Psi) \), and \( d \in \mathbb{R}^m \) is a drift term. We aim to find an optimal policy \( \pi^* \) such that \( J_\pi(\pi^*) = \inf_{\pi \in \Pi} J_\pi(\pi) \).

In Problem C.1, the unique optimal policy \( \pi^*_\mu(\cdot) \) still admits a linear form \( \pi^*_\mu(x_t) = -K^*_\mu x_t + b^* \) \( \quad \text{(Anderson and Moore 2007)} \), where the matrix \( K^*_\mu, b^*_\mu \in \mathbb{R}^{k \times m} \) and the vector \( b^*_\mu \in \mathbb{R}^k \) are the parameters of the policy \( \pi^* \). It then suffices to find the optimal policy in the class \( \Pi \) defined in (2.1).

Parallel to Problem 2.3 we define the general LQ-MFG problem as follows.

Problem C.2 (General LQ-MFG). We consider the following formulation

\[
x_{t+1} = Ax_t + Bu_t + \bar{\mu} + d + \omega_t,
\]

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 \( \pi, \mu \in \mathbb{R}^m \) is the mean-field state, \( \omega_t \in \mathbb{R}^m \) 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 a pair \( (\mu^*, \pi^*) \) such that (i) \( J(\pi^*, \mu^*) = \inf_{\pi \in \Pi} J(\pi, \mu^*) \); (ii) \( E\pi^*_\mu(x_t) \) converges to \( \mu^* \) as \( t \to \infty \), where \( \{x_t\}_{t \geq 0} \) is the Markov chain of states generated by the policy \( \pi^* \).

One can see that Problem C.2 aims to find a Nash equilibrium pair \( (\mu^*, \pi^*) \).

Similar to the discussions in [3.2] to solve Problem C.2 one can design an algorithm similar to Algorithm 1 which solves Problem C.1 and obtain the new mean-field state at each iteration. We omit the detailed algorithm here. Now, we focus on solving Problem C.1 in the sequel.

Similar to [3.3] we drop the subscript \( \mu \) when we focus on Problem C.1 for a fixed \( \mu \). We write \( \pi_{K,b}(x) = -Kx + b + \sigma \cdot \eta \) to emphasize the dependence on \( K \) and \( b \), and \( J(K, b) = J(\pi_{K,b}) \) consequently. We derive an explicit form of the expected total cost \( J(K, b) \) in the following proposition.

Proposition C.3. The expected total cost \( J(K, b) \) in Problem C.1 is decomposed as

\[
J(K, b) = \bar{J}_1(K) + \bar{J}_2(K, b) + \sigma^2 \cdot \text{tr}(R) + \mu^\top \Psi \mu,
\]

where \( \bar{J}_1(K) \) and \( \bar{J}_2(K, b) \) take the forms of

\[
\bar{J}_1(K) = \text{tr}[(Q + K^\top RK) \Phi_K] = \text{tr}(P_K \Psi),
\]

\[
\bar{J}_2(K, b) = \left( \mu_{K,b} \right)^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix} + 2\mu^\top P \mu_{K,b}.
\]

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

\textbf{Proof.} The proof is similar to the one of Proposition B.2. Thus we omit it here. \( \square \)

Compared with the form of \( J(K, b) \) given in (3.5), we see that the only difference is that \( \bar{J}(K, b) \) contains an extra term \( 2\mu^\top P \mu_{K,b} \) in \( \bar{J}_2(K, b) \), which is only a linear term in \( b \) (recall that \( \mu_{K,b} \) is linear in \( b \) by (3.2)). Thus, \( \bar{J}_2(K, b) \) is still strongly convex in \( b \), as shown in the proposition below.
Proposition C.4. Given any $K$, the function $\tilde{J}_2(K, b)$ is $\nu_K$-strongly convex in $b$, here $\nu_K = \sigma_{\min}(Y_{1,K}^{1/2}Y_{1,K} + Y_{2,K}^{1/2}Y_{2,K})$, where $Y_{1,K} = R^{1/2}K(I - A + BK)^{-1}B - R^{1/2}$ and $Y_{2,K} = Q^{1/2}(I - A + BK)^{-1}$. Also, $\tilde{J}_2(K, b)$ has $\nu_K$-Lipschitz continuous gradient in $b$, where $\nu_K$ is upper bounded such that $\nu_K \leq \frac{1 - \rho(A - BK)}{2} \cdot \|B\|_2^2 \cdot \|I\|_2^2 \cdot \|\nu\|_2^2 \cdot \|Q\|_2$.

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 $\tilde{b}^K = \arg\min_b \tilde{J}_2(K, b)$, then $\tilde{J}_2(K, \tilde{b}^K)$ takes the form

$$
\tilde{J}_2(K, \tilde{b}^K) = \left(\bar{A}_K + d\right)^T \left(S(I - A)Q^{-1}\right) \left(\bar{A}_K + d\right),
$$

which is independent of $K$. Here $S = [(I - A)Q^{-1}(I - A)^T + BR^{-1}B^T]^{-1}$.

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

Similar to Problem 2.3, we define the state- and action-value functions as

$$
\tilde{V}_{K,b}(x) = \sum_{t=0}^{\infty} \mathbb{E}\left[\bar{c}(x_t, u_t) \mid x_0 = x, u_t = -Kx_t + b + \sigma \epsilon_t\right] - \tilde{J}(K, b),
$$

$$
\tilde{Q}_{K,b}(x, u) = \bar{c}(x, u) - \tilde{J}(K, b) + \mathbb{E}\left[\tilde{V}_{K,b}(x') \mid x, u\right],
$$

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.1, the state-value function $\tilde{V}_{K,b}(x)$ takes the form

$$
\tilde{V}_{K,b}(x) = x^T P_K x - \text{tr}(P_K \Phi_K) + 2\tilde{f}_{K,b}(x - \mu_{K,b}) - (\mu_{K,b})^T P_K \mu_{K,b},
$$

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

$$
\tilde{Q}_{K,b}(x, u) = x^T \Upsilon_K (x) + 2 \begin{pmatrix} \bar{p}_{K,b} \bar{q}_{K,b} \end{pmatrix} + \text{tr}(P_K \Phi_K) - \frac{1}{\sigma} \cdot \text{tr}(R + P_K BB^T) - b^T R b + 2b^T R K \mu_{K,b} - (\mu_{K,b})^T (Q + K^T R K + P_K) \mu_{K,b} + 2\tilde{f}_{K,b}^T (\bar{A}_K + d) - 2\mu^T P \mu_{K,b}.
$$

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

$$
\begin{pmatrix} \bar{p}_{K,b} \\ \bar{q}_{K,b} \end{pmatrix} = \begin{pmatrix} A^T P_K \cdot (\bar{A}_K + d) + \tilde{f}_{K,b} \\ B^T [P_K \cdot (\bar{A}_K + d) + \tilde{f}_{K,b}] \end{pmatrix},
$$

where the vector $\tilde{f}_{K,b} = (1 - A + BK)^{-1} [(A - BK)^T P_K (Bb + \bar{A}_K + d) - K^T R b + P \mu]$.

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

Now we establish the gradients of $\tilde{J}(K, b)$ for Problem C.1.

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

$$
\nabla_b \tilde{J}_1(K) = 2(\Upsilon_{K,b}^2 K - \Upsilon_{K,b}^1) \cdot \Phi_K, \\
\nabla_b \tilde{J}_2(K, b) = 2[\Upsilon_{K,b}^2 (-K \mu_{K,b} + b) + \Upsilon_{K,b}^1 \mu_{K,b} + \bar{q}_{K,b}],
$$

where the matrix $\Upsilon_K$ is given in (3.7), and the vector $\bar{q}_{K,b}$ 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.2 and Problem C.1, 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 \( \pi_{K^*(\mu_s)} = \Lambda_1(\mu_s) \) under the current mean-field state \( \mu_s \) (note that for any mean-field state \( \mu \), the optimal \( K^*(\mu) \) is independent of \( \mu \) by Proposition 3.4 therefore, we write \( K^* \) here for convenience). By (3.2), we know that

\[
\mu_{s+1}^* = \left( I - A + BK^* \right)^{-1} \cdot \left[ Bb^*(\mu_s) + \overline{\Lambda}_\mu s + d \right].
\]

Also, we denote by

\[
\tilde{\mu}_{s+1} = \left( I - A + B K_s \right)^{-1} (B b_s + \overline{\Lambda}_\mu s + d),
\]

which is the mean-field state generated by the policy \( \pi_{K_s, b_s} \) under the current mean-field state \( \mu_s \).

According to Algorithm 1, it is obvious that the output \( \mu_{s+1}^* \) is an estimator of \( \tilde{\mu}_{s+1} \). By triangular inequality, we have

\[
\| \mu_{s+1} - \mu^* \|_2 \leq \| \mu_{s+1} - \tilde{\mu}_{s+1} \|_2 + \| \tilde{\mu}_{s+1} - \mu_{s+1}^* \|_2 + \| \mu_{s+1}^* - \mu^* \|_2. \tag{D.1}
\]

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

\[
E_1 = \| \mu_{s+1} - \tilde{\mu}_{s+1} \|_2 < \varepsilon_s \leq \varepsilon / 8 \cdot 2^{-s} \tag{D.2}
\]

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

**Bound on \( E_2 \):** Note that by definition, combining with triangular inequality, we have

\[
E_2 = \left\| \left( I - A + BK_s \right)^{-1} (B b_s + \overline{\Lambda}_\mu s + d) - (I - A + BK^* \cdot Bb^*(\mu_s) + \overline{\Lambda}_\mu s + d) \right\|_2
\leq \| B b^*(\mu_s) + \overline{\Lambda}_\mu s + d \|_2 \cdot \left\| \left( I - A + BK^* \right)^{-1} \cdot B(\mu_s - K^*) \right\|_2
+ \| (I - A + BK_s)^{-1} \|_s \cdot \| B \|_s \cdot \| b_s - b^*(\mu_s) \|_2
\leq 2 \| B b^*(\mu_s) + \overline{\Lambda}_\mu s + d \|_2 \cdot \left( \| I - A + BK^* \|^{-1}_{s} \cdot \| B \|_{s} \cdot \| K_s - K^* \|_{s} + [1 - \rho(A - BK_s)]^{-1} \cdot \| B \|_{s} \cdot \| b_s - b^*(\mu_s) \|_2 \right).
\]

For the term \( \| B b^*(\mu_s) + \overline{\Lambda}_\mu s + d \|_2 \), combining Proposition 3.4, it holds that

\[
\| B b^*(\mu_s) + \overline{\Lambda}_\mu s + d \|_2 \leq L_1 \cdot \| B \|_s \cdot \| \mu_s \|_2 + \| \overline{\Lambda} \|_s \cdot \| \mu_s \|_2 + \| d \|_2
\leq (L_1 \cdot \| B \|_s + \| \overline{\Lambda} \|_s) \cdot \| \mu_s \|_2 + \| d \|_2, \tag{D.4}
\]

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

\[
\| K_s - K^* \|_{s} \leq \sqrt{\sigma_{\min}(\Psi_s) \cdot \sigma_{\min}(R) \cdot \varepsilon_s}, \quad \| b_s - b^*(\mu_s) \|_2 \leq \sqrt{2 \mu_K \cdot \varepsilon_s}. \tag{D.5}
\]

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

\[
E_2 \leq \varepsilon / 8 \cdot 2^{-s} \tag{D.6}
\]

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

**Bound on \( E_3 \):** We have

\[
E_3 = \| \mu_{s+1}^* - \mu^* \|_2 = \| \Lambda(\mu_s) - \Lambda(\mu^*) \|_2 \leq L_0 \cdot \| \mu_s - \mu^* \|_2, \tag{D.7}
\]
where we use the fact that the operator $\Lambda(\cdot)$ has Lipschitz constant $L_0 = L_1L_3 + L_2$ according to Proposition 3.2.

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

$$\|\mu_{t+1} - \mu^*\|_2 \leq L_0 \cdot \|\mu_t - \mu^*\|_2 + \varepsilon \cdot 2^{-s-2}. \quad (D.8)$$

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

$$J_1(K) = \text{tr}(P_K \Psi_\omega) = \mathbb{E}_{\psi \sim \mathcal{N}(0, \psi_\omega)}(y^\top P_K y). \quad (D.9)$$

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

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

where the matrix $\Upsilon_K$ 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 $J_1(K) - J_1(K^*)$

$$J_1(K) - J_1(K^*) \geq \sigma_{\min}(\Psi_\omega) \cdot \|\Upsilon_{K^*}\|_*^{-1} \cdot \text{tr}[(\Upsilon_{K^*}^2 K - \Upsilon_{K}^2) (\Upsilon_{K^*}^2 K - \Upsilon_{K}^2)].$$

and the following upper bound

$$J_1(K) - J_1(K^*) \leq \sigma_{\min}^{-1}(R) \cdot \|\Phi_K\|_* \cdot \text{tr}[(\Upsilon_{K^*}^2 K - \Upsilon_{K}^2) (\Upsilon_{K^*}^2 K - \Upsilon_{K}^2)].$$

**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

$$K_{n+1} = K_n - \gamma \cdot (\hat{Y}_{K_n}^{22} K_n - \hat{Y}_{K_n}^{21}), \quad (D.10)$$

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

$$\tilde{K}_{n+1} = K_n - \gamma \cdot (Y_{K_n}^{22} K_n - Y_{K_n}^{21}), \quad (D.11)$$

where $Y_{K_n}$ is given in (D.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 $\tilde{K}_{n+1}$ in (D.10) and the definition of $J_1(K)$ given in (3.6), we have

$$J_1(\tilde{K}_{n+1}) - J_1(K_n) = \mathbb{E}_{y \sim N(0, \Psi)}[y^\top (P_{\tilde{K}_{n+1}} - P_{K_n}) y]$$

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

$$\|Y_{K_n}^{22}\|_* \leq \|R\|_* + \|B\|_*^2 \cdot \|P_{K_n}\|_* \leq \|R\|_* + \|B\|_*^2 \cdot J_1(K_n) \cdot \sigma_{\text{min}}^{-1}(\Psi) \sigma_{\text{min}}^{-1}(\Psi)$$

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

$$J_1(\tilde{K}_{n+1}) - J_1(K_n) \leq -\gamma \cdot \text{tr} \left[ \Phi_{\tilde{K}_{n+1}} \cdot (Y_{K_n}^{22} K_n - Y_{K_n}^{21})^\top (Y_{K_n}^{22} K_n - Y_{K_n}^{21}) \right]$$

where the last line comes from Lemma D.2. This also implies that $J_1(\tilde{K}_{n+1}) \leq J_1(K_n)$. Now, we use the following lemma to establish an upper bound of $|J_1(\tilde{K}_{n+1}) - J_1(K_1)|$.

**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 $1 - \frac{\epsilon}{4}$, it holds that

$$\left| J_1(\tilde{K}_{n+1}) - J_1(K_{n+1}) \right| \leq \gamma \cdot \sigma_{\text{min}}(\Psi) \cdot \sigma_{\text{min}}(R) \cdot \|\Phi_K\|_*^{-1} \cdot \epsilon/4, \quad (D.14)$$

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

**Proof.** See (F.10) for a detailed proof.

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

$$J_1(K_{n+1}) - J_1(K_n) \leq -\gamma \cdot \sigma_{\text{min}}(\Psi) \cdot \sigma_{\text{min}}(R) \cdot \|\Phi_K\|_*^{-1} \cdot \epsilon/4 < 0. \quad (D.15)$$

This shows that with probability at least $1 - \frac{\epsilon}{15}$, 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 $\log(1/\epsilon)$, we know that $J_1(K_N) \leq J_1(K_{N-1}) \leq \cdots \leq J_1(K_0)$ with probability at least $1 - \frac{\epsilon}{15}$, as long as $J_1(K_n) - J_1(K^*) \geq \epsilon/2$ 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 \( \text{(D.13)} \) and \( \text{(D.15)} \), with probability at least \( 1 - \varepsilon^{13} \), it holds that

\[
J_1(K_{n+1}) - J_1(K^*) \leq \left[ 1 - \gamma \cdot \sigma_{\min}(\Psi_e) \cdot \sigma_{\min}(R) \cdot \|\Phi K^*\|_F^{-1} \right] \cdot \left[ J_1(K_n) - J_1(K^*) \right].
\]

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 \( \|K_n - K^*\|_F \). We upper bound \( \|K - K^*\|_F \) using \( J_1(K) - J_1(K^*) \) in the following lemma.

**Lemma D.4.** For any \( K \), we have

\[
\|K - K^*\|_F^2 \leq \sigma_{\min}^{-1}(\Psi_e) \cdot \sigma_{\min}^{-1}(R) \cdot \left[ J_1(K) - J_1(K^*) \right].
\]

**Proof.** See \( \text{[F.4]} \) for a detailed proof.

Combining Lemma D.4, we show the error bound \( \|K_N - K^*\|_F \).

**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 \( J_2(K_N, b_H) - J_2(K_N, b_{H-1}) \leq \cdots \leq J_2(K_N, b_1) \leq J_2(K_N, b_0) \), 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

\[
b_{h+1} = b_h - \gamma^b \cdot \nabla_b J_2(K_N, b_h),
\]

where \( \nabla_b J_2(K_N, b_h) = \partial_2^{K_N}(-K_N \hat{\mu}_{K_N, b} + b_h) + \partial_2^{K_N} \mu_{K_N, b} + \hat{q}_{K_N, b} \) is the estimated gradient at \( b_h \), and the matrix \( \hat{\mu}_{K_N} \) and the vector \( \hat{q}_{K_N, b} \) are the outputs of Algorithm 3. We define \( b_{h+1} \) as an exact update

\[
\tilde{b}_{h+1} = b_h - \gamma^b \cdot \nabla_b J_2(K_N, b_h),
\]

where \( \nabla_b J_2(K_N, b_h) = \partial_2^{K_N}(-K_N \mu_{K_N, b} + b_h) + \partial_2^{K_N} \mu_{K_N, b} + q_{K_N, b} \) is the exact gradient at \( b_h \), and \( \mu_{K_N} \) and \( q_{K_N, b} \) 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 \( b^{K_N} = \argmin_{b} J_2(K_N, b) \).

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

\[
J_2(K_N, \tilde{b}_{h+1}) - J_2(K_N, b_h) \leq -\gamma^b \cdot \|\nabla_b J_2(K_N, b_h)\|_2^2 \leq -\nu_{K_N} \cdot \gamma^b \cdot \left[ J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \right] \leq -\nu_{K_N} \cdot \gamma^b \cdot \varepsilon < 0,
\]

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

**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

\[
|J_2(K_N, b_{h+1}) - J_2(K_N, \tilde{b}_{h+1})| \leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2,
\]

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

**Proof.** See \( \text{[F.4]} \) 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 $1 - \varepsilon^{11}$. 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

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

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 $\zeta_{K,b} \in V_{\zeta}$ and $\xi(\zeta) \in V_{\xi}$ for any $\zeta \in V_{\zeta}$, where $\xi(\zeta)$ is defined as $\xi(\zeta) = \arg\max_{\xi} F(\zeta, \xi)$, by the following lemma.

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

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

$$\text{gap}(\hat{\zeta}, \hat{\xi}) = \max_{\zeta \in V_{\zeta}} F(\hat{\zeta}, \xi) - \min_{\zeta \in V_{\zeta}} F(\zeta, \hat{\xi}),$$

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

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

$$\left\| \mathbb{E}_{\pi_{K,b}} [\psi(x, u)] \hat{\zeta}_1 + \Theta_{K,b} \hat{\zeta}^2 - \mathbb{E}_{\pi_{K,b}} [c(x, u) \psi(x, u)] \right\|_2^2 + |\zeta^1 - J(K, b)|^2 = F(\hat{\zeta}, \xi(\hat{\zeta})) = \max_{\zeta \in V_{\zeta}} F(\hat{\zeta}, \xi(\hat{\zeta})) = \text{gap}(\hat{\zeta}, \hat{\xi}) + \min_{\zeta \in V_{\zeta}} F(\zeta, \hat{\xi}),$$

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

$$\min_{\zeta \in V_{\zeta}} F(\zeta, \xi) \leq \min_{\zeta \in V_{\zeta}} \max_{\xi \in V_{\xi}} F(\zeta, \xi) = \min_{\zeta \in V_{\zeta}} F(\zeta, \xi(\zeta))$$

$$= \frac{1}{2} \min_{\zeta \in V_{\zeta}} \left\{ \left\| \mathbb{E}_{\pi_{K,b}} [\psi(x, u)] \zeta_1 + \Theta_{K,b} \zeta^2 - \mathbb{E}_{\pi_{K,b}} [c(x, u) \psi(x, u)] \right\|_2^2 + |\zeta^1 - J(K, b)|^2 \right\} = 0,$$

(D.22)
Lemma D.7. The average \( E_{\pi_{K, b}} [\psi(x, u)] \mathcal{C}_1 + \Theta_{K, b} \mathcal{C}_2 - E_{\pi_{K, b}} [c(x, u) \psi(x, u)] \) is given by (D.21) as follows:
\[
\|E_{\pi_{K, b}} [\psi(x, u)] \mathcal{C}_1 + \Theta_{K, b} \mathcal{C}_2 - E_{\pi_{K, b}} [c(x, u) \psi(x, u)]\|_2^2 + \|\mathcal{C}_1 - J(K, b)\|_2^2
= \|\Theta_{K, b}(\mathcal{C}_1 - \mathcal{C}_2)\|_2^2 \geq \lambda_K^2 \cdot \|\mathcal{C}_1 - \mathcal{C}_2\|_2 \geq \lambda_K^2 \cdot \|\alpha_{K, b} - \alpha_{K, b}\|_2^2,
\]
where the matrix \( \Theta_{K, b} \) is given in (B.8) and the scalar \( \lambda_K \) is specified in Proposition B.6. Combining (D.21), (D.22) and (D.23), we obtain the relation between gap \( (\mathcal{C}_1, \mathcal{C}_2) \) and \( \|\alpha_{K, b} - \alpha_{K, b}\|_2^2 \):
\[
\|\alpha_{K, b} - \alpha_{K, b}\|_2^2 \leq \lambda_K^2 \cdot \text{gap}(\mathcal{C}_1, \mathcal{C}_2).
\]

Part 3. We now proceed to upper bound the primal-dual gap \( (\hat{\mathcal{C}}, \hat{\mathcal{C}}) \), then by (D.24) in Part 2, we establish the upper bound of the estimation error \( \|\hat{\alpha}_{K, b} - \alpha_{K, b}\|_2 \). 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 \( \mu_z = 1/\bar{T} \cdot \sum_{t=1}^{\bar{T}} \tilde{z}_t \), where \( \tilde{z}_t = [\tilde{x}_t^T, \tilde{u}_t^T]^T \) is the concatenation of the state \( \tilde{x}_t \) and the action \( \tilde{u}_t \). Then following transition holds
\[
\tilde{z}_{t+1} = L \tilde{z}_t + \nu + \delta_t,
\]
where
\[
\nu = \left( \begin{array}{c} \frac{\bar{T}}{2} \mu_d + d \\ -K(\bar{T} \mu_d + d) + b \end{array} \right), \quad \delta_t = \left( \begin{array}{c} \omega_t \\ -K \omega_t + \sigma \eta \end{array} \right), \quad L = \left( \begin{array}{cc} A & B \\ -KA & -KB \end{array} \right).
\]

Note that we have
\[
L = \left( \begin{array}{cc} A & B \\ -KA & -KB \end{array} \right) = \left( \begin{array}{c} I \\ -K \end{array} \right) \left( \begin{array}{cc} A & B \end{array} \right).
\]

Then by the property of spectral radius, it holds that
\[
\rho(L) = \rho \left( \begin{array}{cc} A & B \\ -KA & -KB \end{array} \right) = \rho(A - BK) < 1.
\]

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

Lemma D.7. The average \( \mu_z = 1/\bar{T} \cdot \sum_{t=1}^{\bar{T}} \tilde{z}_t \), where \( \tilde{z}_t = [\tilde{x}_t, \tilde{u}_t]^T \), satisfies that
\[
\mu_z \sim \mathcal{N} \left( \mu_z + \frac{1}{\bar{T}} \mu_{\bar{T}}, \frac{1}{\bar{T}} \Sigma_{\bar{T}} \right),
\]
where \( \mu_{\bar{T}} \) and \( \Sigma_{\bar{T}} \) are bounded such that there exist positive absolute constants \( M_\mu \) and \( M_\Sigma \) such that \( \|\mu_{\bar{T}}\|_2 \leq M_\mu \cdot (1 - \rho)^{-2} \cdot \|\mu_z\|_2 \) and \( \|\Sigma_{\bar{T}}\|_F \leq M_\Sigma \cdot (1 - \rho)^{-1} \cdot \|\Sigma_z\|_F \). This gives that with probability at least \( 1 - \bar{T}^{-6} \), it holds that
\[
\|\bar{z} - \mu_z\|_2 \leq \frac{\log \bar{T}}{\bar{T}^{1/4}} \cdot (1 - \rho)^{-2} \cdot \text{poly}(\|\Phi_K\|_\infty, \|K\|_F, \|b\|_2, \|\mu\|_2).
\]

Proof. See [F.6] for a detailed proof. \( \square \)

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

Now, note that for any \( z \sim \mathcal{N}(\mu_z, \Sigma_z) \), the random variable \( z - \mu_z + 1/\bar{T} \cdot \mu_{\bar{T}} \) follows \( \mathcal{N}(0, \Sigma_z + 1/\bar{T} \cdot \Sigma_{\bar{T}}) \). By Lemma G.3 there exists an absolute constant \( C_0 > 0 \) such that
\[
P \left[ \|z - \mu_z + 1/\bar{T} \cdot \mu_{\bar{T}}\|_2^2 - \text{tr}(\Sigma_z) > \tau \right] \leq 2 \cdot \exp \left[ -C_0 \cdot \min(\tau^2 \|\Sigma_z\|_F^{-2}, \tau \|\Sigma_z\|_F^{-1}) \right],
\]
where, for notational convenience, we write the matrix $\tilde{\Sigma}_z = \Sigma_z + 1/\bar{T} \cdot \hat{\Sigma}_z$. By taking $r = C_1 \cdot \log T \cdot 1/\tilde{\Sigma}_z$ for some sufficiently large positive absolute constant $C_1$, the above inequality takes the form

$$P\left[ \left| \|z - \bar{\mu}_z + 1/\bar{T} \cdot \mu_T \|_2^2 - \text{tr}(\tilde{\Sigma}_z) \right| > C_1 \cdot \log T \cdot \|\tilde{\Sigma}_z\|_F \right] \leq T^{-6}. \quad (D.25)$$

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

$$\mathcal{E}_{t,1} = \left\{ \|z_t - \bar{\mu}_z + 1/\bar{T} \cdot \mu_T \|_2^2 - \text{tr}(\tilde{\Sigma}_z) \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_z\|_F \right\}.$$

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

$$\max_{t \in [T]} \|z_t - \bar{\mu}_z\|_2^2 \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_z\|_F + \text{tr}(\tilde{\Sigma}_z) + \|1/\bar{T} \cdot \mu_T\|_2^2$$

$$\leq 2C_1 \cdot \left(1 + M_\Sigma(1-\rho)^{-1}/\bar{T}^2\right) \cdot \log T \cdot \|\tilde{\Sigma}_z\|_F + M_\mu(1-\rho)^{-2}/\bar{T}^2 \cdot \|\mu_T\|_2^2$$

$$\leq C_2 \cdot \log T \cdot (1 + \|\Phi_K\|_F) \cdot \|\Phi_K\|_F \cdot (1-\rho)^{-1} + C_3 \cdot (\|b\|_2^2 + \|\mu\|_2^2) \cdot (1-\rho)^{-4} \cdot \bar{T}^{-2}$$

$$\leq C_2 \cdot \log T \cdot (1 + \|\Phi_K\|_F) \cdot \|\Phi_K\|_F \cdot (1-\rho)^{-1}$$

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

$$\mathcal{E}_2 = \{\|\bar{\mu}_z - \mu_T\|_2^2 \leq C_1\}.$$

Then by Lemma [D.7], we know that $P(\mathcal{E}_2) \geq 1 - T^{-6}$ for sufficiently large $\bar{T}$. 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 $P(\mathcal{E}) \geq 1 - T^{-5} - T^{-6}$.

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

$$\bar{F}(\zeta, \xi) = \left\{ \mathbb{E}(\tilde{\psi}) \xi^1 + \mathbb{E} \left[ (\tilde{\psi} - \tilde{\psi})\tilde{\psi}^T \right] \xi^2 - \mathbb{E}(\tilde{\psi}) \right\}^\top \xi^2 + \left[ \xi^1 - \mathbb{E}(\tilde{\psi}) \right] \cdot \xi^1 - 1/2 \cdot \|\xi\|_2^2$$

(D.27)

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

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

$$|F(\zeta, \xi) - \bar{F}(\zeta, \xi)| \leq \left[ \frac{1}{2T} + \frac{\log \bar{T}}{T^{1/4}} \right] \cdot (1-\rho)^{-2} \cdot \text{poly}(\|K\|_F, \|b\|_2, \|\mu\|_2, J(K_0, b_0)),$$

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

Proof. See [E.7] for a detailed proof.

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

$$\left| \text{gap}(\zeta, \xi) - \left[ \max_{\xi \in \mathcal{V}_\zeta} F(\zeta, \xi) - \min_{\xi \in \mathcal{V}_\zeta} \bar{F}(\zeta, \xi) \right] \right| \leq \left[ \frac{1}{2T} + \frac{\log \bar{T}}{T^{1/4}} \right] \cdot (1-\rho)^{-2} \cdot \text{poly}(\|K\|_F, \|b\|_2, \|\mu\|_2, J(K_0, b_0)).$$

(D.28)

Therefore, to obtain the bound of $\text{gap}(\zeta, \xi)$, we only need to bound the term $\max_{\xi \in \mathcal{V}_\zeta} \bar{F}(\zeta, \xi) - \min_{\xi \in \mathcal{V}_\zeta} \bar{F}(\zeta, \xi)$ in (D.25). 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.
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 \( \pi_t \) 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 \( \Sigma_{\infty} \) is the covariance matrix.

We define the \( \beta \)-mixing coefficients for any \( n \geq 1 \) as follows:

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

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

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

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

Proof. See Proposition 3.1 in [Tu and Recht, 2017] for a detailed proof. \qed

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 \( \{z_t = (x_t, u_t')\}_{t \geq 0} \) is \( \beta \)-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

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

where the objective function \( F(x, y) \) is convex in \( x \) and concave in \( y \). In addition, we assume that the distribution \( \pi_x \) is the stationary distribution induced by a Markov chain \( \{\epsilon_t\}_{t \geq 0} \), which is \( \beta \)-mixing with \( \beta(n) \leq C_{\epsilon} \cdot \rho^n \), where \( C_{\epsilon} \) is a constant. Moreover, we assume that it holds almost surely that \( G(x, y; \epsilon) \) is \( \tilde{L}_0 \)-Lipschitz in both \( x \) and \( y \), the gradient \( \nabla_x G(x, y; \epsilon) \) is \( \tilde{L}_1 \)-Lipschitz in \( x \) for any \( y \in \mathcal{Y} \), the gradient \( \nabla_y G(x, y; \epsilon) \) is \( \tilde{L}_2 \)-Lipschitz in \( y \) for any \( x \in \mathcal{X} \), where for simplicity we assume that \( C_{\epsilon}, \tilde{L}_0, \tilde{L}_1 > 1 \). Each step of our gradient-based method takes the following form:

\[
x_{t+1} = \Gamma_x [x_t - \gamma_t \cdot \nabla_x G(x_t, y_t; \epsilon_t)], \quad y_{t+1} = \Gamma_y [y_t - \gamma_t \cdot \nabla_y G(x_t, y_t; \epsilon_t)],
\]

where the operators \( \Gamma_x \) and \( \Gamma_y \) 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 \( \bar{x} = (\sum_{t=1}^{T} \gamma_t)^{-1} (\sum_{t=1}^{T} \gamma_t x_t) \) and \( \bar{y} = (\sum_{t=1}^{T} \gamma_t)^{-1} (\sum_{t=1}^{T} \gamma_t y_t) \) be the output of the primal-dual method after \( T \) iterations, then there exists an absolute constant \( C_0 > 0 \), such that for any \( \delta \in (0, 1) \), the primal-dual gap to the minimax optimization problem satisfies that

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

with probability at least \( 1 - \delta \).

Proof. See Theorem 5.4 in [Yang et al., 2019] for a detailed proof. \qed

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

\[
G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') = \left[ \tilde{\psi} \zeta^1 + (\tilde{\psi} - \tilde{\psi}') \tilde{\psi}^T \xi^2 - \tilde{c} \tilde{\psi} \right]^T \xi^2 + (\zeta^1 - \tilde{c}) \cdot \xi^1 - 1/2 \cdot ||\xi||_2^2,
\]

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

\[
\nabla_{\zeta} G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') = \left( \frac{\tilde{\psi}^T \xi^2 + \xi^1}{\tilde{\psi}(\tilde{\psi} - \tilde{\psi}')^T \xi^2} \right), \quad \nabla_{\xi} G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') = \left( \frac{\zeta^1 - \tilde{c} - \xi^1}{\tilde{\psi}(\tilde{\psi} - \tilde{\psi}')^T \xi^2 - \tilde{c} \tilde{\psi} - \xi^2} \right).
\]
By Condition [B.7] and Lemma [D.6], we know that
\[ \| \nabla G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') \|_2 \leq \text{poly}(\| K \|_F, J(K_0, b_0)) \cdot \log^2 T \cdot (1 - \rho)^{-2}, \]
\[ \| \nabla G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') \|_2 \leq \text{poly}(\| K \|_F, \| \mu \|_2, J(K_0, b_0)) \cdot \log^2 T \cdot (1 - \rho)^{-2}. \]  
(D.29)
This gives the Lipschitz constant $\tilde{L}_0$ in Lemma [D.10] for $G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}')$. Also, the Hessians take the forms
\[ \nabla^2 G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') = 0, \quad \nabla^2 G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') = -I, \]
which follows that
\[ \| \nabla^2 G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') \|_2 = 0, \quad \| \nabla^2 G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}') \|_2 = 1. \]  
(D.30)
This gives the Lipschitz constant $\tilde{L}_1$ in Lemma [D.10] for $\nabla G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}')$ and $\nabla G(\zeta, \xi; \tilde{\psi}, \tilde{\psi}')$. Moreover, note that (D.26) provides an upper bound of $M$, combining (D.29), (D.30) and Lemma [D.10] we know that
\[ \max_{\zeta \in \mathcal{V}_1} \tilde{F}(\zeta, \xi) - \min_{\zeta \in \mathcal{V}_1} \tilde{F}(\zeta, \xi) \leq \text{poly}(\| K \|_F, \| \mu \|_2, J(K_0, b_0)) \cdot \log^2 T \cdot (1 - \rho)^{-4} \cdot \sqrt{T} \]
holds with probability at least $1 - T^{-5}$. Combining (D.24) and (D.28), we conclude the theorem. \qed

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 $\Pi$ 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 $\Pi$ as
\[ \| \pi_{K_1,b_1} - \pi_{K_2,b_2} \|_2 = c_1 \cdot \| K_1 - K_2 \|_2 + c_2 \cdot \| b_1 - b_2 \|_2, \]
where $c_1$ and $c_2$ are positive constants.

One can easily verify that the above definition satisfies the requirement of being a metric.

We compute the Lipschitz constants for both operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot, \cdot)$ in the sequel.

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
\[ \| \Lambda_1(\mu_1) - \Lambda_1(\mu_2) \|_2 \leq c_2 \cdot \left( \| (I - A)Q^{-1}(I - A)^T + BR^{-1}B^T \|^{-1} \right)_\ast \cdot \| K^*Q^{-1}(I - A)^T - R^{-1}B^T \|_\ast \cdot \| \mu_1 - \mu_2 \|_2 \\
= c_2 \cdot L_1 \cdot \| \mu_1 - \mu_2 \|_2. \]  
(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 $\pi$ is an optimal policy under some $\mu \in \mathbb{R}^m$, it holds that
\[ \| \Lambda_2(\mu_1, \pi) - \Lambda_2(\mu_2, \pi) \|_2 = \| (I - A + BK^*)^{-1} \cdot \overline{A} \cdot (\mu_1 - \mu_2) \|_2 \\
\leq \left( 1 - \rho(A - BK^*) \right)^{-1} \| \overline{A} \|_\ast \cdot \| \mu_1 - \mu_2 \|_2 \\
= L_2 \cdot \| \mu_1 - \mu_2 \|_2. \]  
(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 $\pi_1, \pi_2$ are optimal under $\mu_1, \mu_2$, and for any population mean $\mu \in \mathbb{R}^m$, we have
\[
\| \Lambda_2(\mu, \pi_1) - \Lambda_2(\mu, \pi_2) \|_2 = \| (I + A + BK^*)^{-1}B \cdot (b_{\mu_1} - b_{\mu_2}) \|_2 \\
\leq 1 - \rho(A + BK^*)^{-1} \| B \|_* \cdot \| b_{\mu_1} - b_{\mu_2} \|_2 \\
= c_2^{-1} \cdot L_3 \cdot \| \pi_1 - \pi_2 \|_2. \quad \text{(E.3)}
\]

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:
\[
\| \Lambda(\mu_1) - \Lambda(\mu_2) \|_2 = \| \Lambda_2(\mu_1, \Lambda_1(\mu_1)) - \Lambda_2(\mu_2, \Lambda_1(\mu_2)) \|_2 \\
\leq \| \Lambda_2(\mu_1, \Lambda_1(\mu_1)) - \Lambda_2(\mu_1, \Lambda_1(\mu_2)) \|_2 + \| \Lambda_2(\mu_1, \Lambda_1(\mu_2)) - \Lambda_2(\mu_2, \Lambda_1(\mu_2)) \|_2 \\
\leq c_2^{-1} \cdot L_3 \cdot \| \Lambda_1(\mu_1) - \Lambda_1(\mu_2) \|_2 + L_2 \cdot \| \mu_1 - \mu_2 \|_2 \\
\leq c_2^{-1} \cdot L_3 \cdot c_2 \cdot L_1 \cdot \| \mu_1 - \mu_2 \|_2 + L_2 \cdot \| \mu_1 - \mu_2 \|_2 = (L_1L_3 + L_2) \cdot \| \mu_1 - \mu_2 \|_2,
\]
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 $L_0 = L_1L_3 + L_2 < 1$, 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 2.3 \qed

\subsection*{E.2 Proof of Proposition 3.4}

\textbf{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
\[
\min_{\mu_{K,b}} \left( \frac{\mu_{K,b}}{b} \right)^\top \left( \begin{array}{cc} Q + K^\top RK & -K^\top R \\ -RK & R \end{array} \right) \left( \begin{array}{c} \mu_{K,b} \\ b \end{array} \right) \\
\text{s.t. } (I - A + BK)\mu_{K,b} - (Bb + \overline{A}\mu + d) = 0. \quad \text{(E.4)}
\]
Consider the KKT conditions of (E.4), the minimizer satisfies that
\[
2MK\left( \frac{\mu_{K,b}}{b} \right) + NK\lambda = 0, \quad NK\left( \frac{\mu_{K,b}}{b} \right) + \overline{A}\mu + d = 0, \quad \text{(E.5)}
\]
where we that
\[
MK = \left( \begin{array}{cc} Q + K^\top RK & -K^\top R \\ -RK & R \end{array} \right), \quad NK = \left( \begin{array}{c} -(I - A + BK)^\top \\ B^\top \end{array} \right).
\]
By solving (E.5), we obtain the minimizer to the program as follows
\[
\left( \frac{\mu_{K,b}}{b} \right) = -M_K^{-1}NK\left( N_K^\top M_K^{-1}NK \right)^{-1}(\overline{A}\mu + d). \quad \text{(E.6)}
\]
By substituting (E.6) into the definition of $J_2(K, b)$ in (3.6), we have
\[
J_2(K, b_K) = (\overline{A}\mu + d)^\top \left( N_K^\top M_K^{-1}NK \right)^{-1}(\overline{A}\mu + d). \quad \text{(E.7)}
\]
Moreover, by algebra, we have
\[
M_K^{-1} = \left( \begin{array}{cc} Q^{-1} & Q^{-1}K^\top \\ KQ^{-1} & KQ^{-1}K^\top + R^{-1} \end{array} \right).
\]
Therefore, the term $N_K^\top M_K^{-1}NK$ in (E.7) takes the following form
\[
N_K^\top M_K^{-1}NK = (I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top.
\]
Combining (E.7), we have
\[
J_2(K, b_K) = (\overline{A}\mu + d)^\top \left( (I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top \right)^{-1}(\overline{A}\mu + d).
\]
Also, combining (E.6), we have
\[
\left( \frac{\mu_{K,b}}{b} \right) = \left( \begin{array}{c} Q^{-1}(I - A)^\top \\ KQ^{-1}(I - A)^\top + R^{-1}B^\top \end{array} \right) \left( (I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top \right)^{-1}(\overline{A}\mu + d).
\]
Then we finish the proof of the proposition. \qed
E.3 Proof of Proposition B.3

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

$$E_c = E(x_t^T Q x_t + u_t^T R u_t + \mu^T Q \mu)$$
$$= E(x_t^T Q x_t + x_t^T K^T R K x_t - 2b^T R K x_t + b^T R B + \sigma^2 \eta_t^T R \eta_t + \mu^T Q \mu)$$
$$= E(x_t^T (Q + K^T R K)x_t - 2b^T R K x_t) + b^T R B + \sigma^2 \cdot \text{tr}(R) + \mu^T Q \mu,$$

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

$$J(K, b) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \left\{ E[x_t^T (Q + K^T R K)x_t - 2b^T R K x_t] + b^T R B + \mu^T Q \mu \right\}$$
$$= E_{x \sim N(\mu_{K,b}, \Phi_K)}[x^T (Q + K^T R K)x - 2b^T R K x] + b^T R B + \sigma^2 \cdot \text{tr}(R) + \mu^T Q \mu$$
$$= \text{tr}[(Q + K^T R K)\Phi_K] + (\mu_{K,b})^T (Q + K^T R K)\mu_{K,b} - 2b^T R K \mu_{K,b} + b^T R B + \sigma^2 \cdot \text{tr}(R) + \mu^T Q \mu.$$

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

$$\text{tr}[(Q + K^T R K)\Phi_K] = \text{tr}(P_K \Psi_e),$$

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

$$\nabla^2_{b,b} J_2(K, b) = B^T(I - A + BK)^{-T}(Q + K^T R K)(I - A + BK)^{-1}B$$
$$- [RK(I - A + BK)^{-1}B + B^T(I - A + BK)^{-T}K^T R] + R$$
$$= [\sqrt{R}K(I - A + BK)^{-1}B - \sqrt{R}]^T [\sqrt{R}K(I - A + BK)^{-1}B - \sqrt{R}]$$
$$+ B^T(I - A + BK)^{-T}Q(I - A + BK)^{-1}B,$$

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

$$\|\nabla^2_{b,b} J_2(K, b)\|_* \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_2^2 \cdot \|K\|_2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2).$$

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

$$\lambda_K \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_2^2 \cdot \|K\|_2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2).$$

This finishes the proof.

E.5 Proof of Proposition B.3

Proof. It holds that

$$\nabla_K [\text{tr}(P_K \Psi_e)] = 2[(R + B^T P_K B)K - B^T P_K A] \Phi_K,$$

also, we have

$$\nabla_K [(\mu_{K,b})^T (Q + K^T R K)\mu_{K,b}]$$
$$= 2[RK - B^T(I - A + BK)^{-T}(Q + K^T R K)]\mu_{K,b}(\mu_{K,b})^T,$$
$$\nabla_K (b^T R K \mu_{K,b}) = [R - B^T(I - A + BK)^{-T}K^T R]b(\mu_{K,b})^T.$$

Combining the above equations, we obtain the form of $\nabla_K J_1(K)$. Moreover, the form of $\nabla_{b} J_2(K, \hat{b})$ is obtained simply by taking gradient w.r.t. $b$, combining the definition of $\mu_{K,b}$ in (B.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

$$V_{K,b}(x) = \sum_{t=0}^{\infty} \mathbb{E} \left[ x_t^T (Q + K^T RK)x_t - 2b^T RKx_t \right. \left. + b^T Rb + \sigma^2 \eta_t^T R R \eta_t + \mu^T Q \mu | x_0 = x \right] - J(K,b).$$

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

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

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

$$x^T Gx + r^T x + h = x^T (Q + K^T RK)x - 2b^T RKx + b^T Rb + \mu^T Q \mu$$

$$- \left[ \text{tr}(P_K \Psi) + (\mu_K)^T (Q + K^T RK) \mu_K - 2b^T RK \mu_K + \mu^T Q \mu + b^T Rb \right]$$

$$+ \left[ (A - BK)x + (Bb + \bar{A} \mu + d) \right]^T G \left[ (A - BK)x + (Bb + \bar{A} \mu + d) \right]$$

$$+ \text{tr}(G \Psi) + r^T \left[ (A - BK)x + (Bb + \bar{A} \mu + d) \right] + h - \sigma^2 \cdot \text{tr}(R).$$

(E.9)

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

$$G = P_K, \quad 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

$$h = -2f_{K,b} \mu_K - (\mu_K)^T P_K \mu_K - \text{tr}(P_K \Phi_K).$$

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. \qed

E.7 Proof of Proposition B.5

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

$$Q_{K,b}(x,u) = \psi(x,u)^T \alpha_{K,b} + \beta_{K,b},$$

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

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

(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

$$\psi(x,u)^T \alpha_{K,b} = c(x,u) - J(K,b) + \mathbb{E}_{\pi_{K,b}} [\psi(x',u') | x,u]^T \alpha_{K,b}.$$

(E.12)

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

$$\mathbb{E}_{\pi_{K,b}} \left\{ \psi(x,u) [\psi(x,u) - \psi(x',u')]^T \right\} \cdot \alpha_{K,b} + \mathbb{E}_{\pi_{K,b}} [\psi(x,u)] \cdot J(K,b) = \mathbb{E}_{\pi_{K,b}} [c(x,u) \psi(x,u)].$$

Combining the definition of the matrix $\Theta_{K,b}$ in (B.7), we conclude the proposition. \qed
E.8 Proof of Proposition B.6

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

\[
z' = Lz + \nu + \delta,
\]

where the matrix \( L \) and vectors \( \nu, \delta \) are given as

\[
\nu = \left( \frac{A\mu + d}{-K(\mu + d) + b} \right), \quad \delta = \left( -K\omega + \sigma \eta \right), \quad L = \left( \begin{array}{cc} A & B \\ -KA & -KB \end{array} \right).
\]

Note that the matrix \( L \) also takes the form

\[
L = \left( \begin{array}{cc} I & 0 \\ 0 & K \end{array} \right) \left( \begin{array}{cc} A & B \\ -KA & -KB \end{array} \right).
\]

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 \( \mu_z \) and the covariance \( \Sigma_z \) satisfy the following equations

\[
\mu_z = L\mu_z + \nu, \quad \Sigma_z = L\Sigma_z L^T + \Psi_\delta,
\]

where the matrix \( \Psi_\delta \) is the covariance of the vector \( \delta \), which takes the form

\[
\Psi_\delta = \left( \begin{array}{cc} \Psi_\omega & -\Psi_\omega K^T \\ -K\Psi_\omega & K\Psi_\omega K^T + \sigma^2 I \end{array} \right).
\]

Also, the covariance matrix \( \Sigma_z \) takes the following form

\[
\Sigma_z = \left( \begin{array}{cc} \Phi_K & -\Phi_K K^T \\ -K\Phi_K & K\Phi_K K^T + \sigma^2 I \end{array} \right) = \left( \begin{array}{cc} 0 & 0 \\ 0 & \sigma^2 I \end{array} \right) + \left( \begin{array}{cc} I & 0 \\ 0 & -K \end{array} \right) \Phi_K \left( \begin{array}{cc} I \\ -K \end{array} \right)^T.
\]

(E.15)

Now, we establish the following lemma.

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

\[
\Theta_{K,b} = \left( \begin{array}{cc} 2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L) & 0 \\ 0 & \Sigma_z (I - L)^T \end{array} \right).
\]

Proof. See (E.8) for a detailed proof. \( \square \)

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

\[
\| \Theta_{K,b} \|_* \leq 2 \max \left\{ \| \Sigma_z \|_2, \| \Sigma_z \|_* \right\} \leq 4 \| \Sigma_z \|_2 \leq 4(1 + \| K \|_F^2) \| \Phi_K \|_*^2.
\]

This proves the upper bound of the spectral norm.

Minimum singular value. To lower bound \( \sigma_{\min}(\overline{\Theta}_{K,b}) \), we only need to upper bound \( \sigma_{\max}(\overline{\Theta}_{K,b}) \).

We first proceed to calculate \( \overline{\Theta}_{K,b}^{-1} \). Recall that the matrix \( \overline{\Theta}_{K,b} \) in the linear system (B.8) takes the following form

\[
\overline{\Theta}_{K,b} = \left( \begin{array}{cc} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}[\psi(x,u)] & \Theta_{K,b} \end{array} \right).
\]

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

\[
\overline{\sigma}_z = \mathbb{E}_{\pi_{K,b}}[\psi(x,u)] = \left( \begin{array}{c} \text{svec}(\Sigma_z) \\ 0_{k+m} \end{array} \right).
\]

(E.16)
where $0_{k+m}$ denotes the all-zero column vector with dimension $k + m$. Also, since $\Theta_{K,b}$ is invertible, the matrix $\Theta_{K,b}$ is also invertible, whose inverse is given as

$$
\tilde{\Theta}_{K,b}^{-1} = \begin{pmatrix}
\frac{1}{\sigma_z} & 0 \\
-\Theta_{K,b}^{-1} & \Theta_{K,b}^{-1}
\end{pmatrix}.
$$

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

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

**Proof.** See [F.9] for a detailed proof. \qed

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

## 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

$$
y^\top P_{K_2} y = \sum_{t \geq 0} y^\top [(A - BK_2)^t] (Q + K_2^\top RK_2)(A - BK_2)^t y.
$$

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

$$
y^\top P_{K_2} y = \sum_{t \geq 0} y_t^\top (Q + K_2^\top RK_2) y_t = \sum_{t \geq 0} (y_t^\top Q y_t + y_t^\top K_2^\top RK_2 y_t).
$$

Then by telescoping, we have

$$
y^\top P_{K_2} y - y^\top P_{K_1} y = \sum_{t \geq 0} (y_t^\top Q y_t + y_t^\top K_2^\top RK_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t).
$$

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

$$
y_t^\top Q y_t + y_t^\top K_2^\top RK_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t
$$

$$
= y_t^\top [Q + (K_2 - K_1 + K_1)\top R(K_2 - K_1 + K_1)] y_t
$$

$$
+ y_t^\top [A - BK_1 - B(K_2 - K_1)]^\top P_{K_1} [A - BK_1 - B(K_2 - K_1)] y_t - y_t^\top P_{K_1} y_t
$$

$$
= 2y_t^\top (K_2 - K_1)^\top [(R + A^\top P_{K_1} B)K_1 - B^\top P_{K_1} A] y_t + y_t^\top (K_2 - K_1)^\top (R + A^\top P_{K_1} B)(K_2 - K_1) y_t
$$

$$
= 2y_t^\top (K_2 - K_1)^\top (T_{21}^2, K_1 - T_{21}^1) y_t + y_t^\top (K_2 - K_1)^\top T_{22}^2, K_1 - T_{21}^1) y_t.
$$

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

### F.2 PROOF OF LEMMA D.2

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

$$
J_1(K) - J_1(K^*) = \text{tr}(P_e \Psi_\epsilon - P_{K^*} \Psi_\epsilon) = \mathbb{E}_{y \sim \mathcal{N}(0, \Psi_\epsilon)} (y^\top P_{K} y - y^\top P_{K^*} y)
$$

$$
= -\mathbb{E}_{y \sim \mathcal{N}(0, \Psi_\epsilon)} \left[ \sum_{t \geq 0} D_{K, K^*}(y_t) \right],
$$

(E.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

\[
D_{K,K^*}(y) = 2y^T(K^* - K)(\bar{Y}_{22}^K K - \bar{Y}_{21}^K) + y^T(K^* - K)^T \bar{Y}_{22}^K(K^* - K)x
\]


where the second line comes from direct computation, while the last line comes from the fact that

\[
\bar{Y}_{22}^K K - \bar{Y}_{21}^K = 0
\]

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

\[
D_{K,K^*}(y) \geq -y^T(\bar{Y}_{22}^K K - \bar{Y}_{21}^K) (\bar{Y}_{22}^K K - \bar{Y}_{21}^K) y.
\]

(F.5)

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

\[
J_1(K) - J_1(K^*) \leq \left| \mathbb{E}_{y_t \sim N(0, \sigma^*)} \left[ \sum_{t \geq 0} y_t y_t^T \right] \right| \cdot \text{tr}[\bar{Y}_{22}^K K - \bar{Y}_{21}^K] (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)
\]

\[
= \left| \Phi \right| \cdot \text{tr}[\bar{Y}_{22}^K K - \bar{Y}_{21}^K] (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)
\]

\[
\leq \|\bar{Y}_{22}^K K - \bar{Y}_{21}^K\|_\ast \cdot \|\Phi\| \cdot \text{tr}[\bar{Y}_{22}^K K - \bar{Y}_{21}^K] (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)
\]

\[
\leq \sigma_{\min}(R) \cdot \|\Phi\| \cdot \text{tr}[\bar{Y}_{22}^K K - \bar{Y}_{21}^K] (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)
\]

where the second line comes from direct computation, while the last line comes from the fact that

Under review as a conference paper at ICLR 2020

where

\[
\Phi
\]

and

\[
\{y_t\}_{t \geq 0} \]

follows the transition \( y_{t+1} = (A - BK^*)y_t \). Therefore, by choosing \( K = (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)^{-1} \), by a similar calculation in (F.4), the function \( D_{K,K^*}(y) \) takes the form

\[
D_{K,K^*}(y) = -y^T(\bar{Y}_{22}^K K - \bar{Y}_{21}^K) (\bar{Y}_{22}^K K - \bar{Y}_{21}^K) y.
\]

(F.7)

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

\[
J(K) - J(K^*) \geq \text{tr}[\Phi (\bar{Y}_{22}^K K - \bar{Y}_{21}^K) (\bar{Y}_{22}^K K - \bar{Y}_{21}^K)]
\]

\[
\geq \sigma_{\min}(\Phi) \cdot \|\bar{Y}_{22}^K K - \bar{Y}_{21}^K\|_\ast \cdot \text{tr}[\bar{Y}_{22}^K K - \bar{Y}_{21}^K] (\bar{Y}_{22}^K K - \bar{Y}_{21}^K).
\]

Here in the last line, we use the fact that \( \Phi \geq \Psi \). This finishes the proof of the lower bound.

**F.3 PROOF OF LEMMA D.4**

**Proof.** Note that from Lemma D.4 it holds that

\[
J_1(K) - J_1(K^*) = \mathbb{E} \left\{ \sum_{t \geq 0} \left[ 2y_t^T (K^* - K)(\bar{Y}_{22}^K K - \bar{Y}_{21}^K) y_t + y_t^T (K^* - K)^T \bar{Y}_{22}^K (K - K^*) y_t \right] \right\}
\]

\[
= \text{tr}[\Phi (K^* - K)^T \bar{Y}_{22}^K (K - K^*)]
\]

\[
\geq \|\Phi\| \cdot \|\bar{Y}_{22}^K\|_\ast \cdot \text{tr}[(K^* - K)^T (K - K^*)]
\]

\[
\geq \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \|K - K^*\|_2^2,
\]

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 \( \bar{Y}_{22}^K K^* - \bar{Y}_{21}^K \) is the natural gradient of \( J_1(\cdot) \) evaluated at \( K^* \), which implies that \( \bar{Y}_{22}^K K^* - \bar{Y}_{21}^K = 0 \); while in the last line, we use the fact that \( \|\Phi\| \geq \sigma_{\min}(\Psi) \) and \( \|\bar{Y}_{22}^K\|_\ast \geq \sigma_{\min}(R) \). This concludes the lemma.
Also, we upper bound \( \| \nabla_b J_2(K_N, \hat{b}_{h+1}) \|_2 \) as
\[
\| \nabla_b J_2(K_N, \hat{b}_{h+1}) \|_2 \leq \text{poly}_1(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)) \cdot [1 - \rho(A - BK_N)]^{-1}. \tag{F.9}
\]
Combining (F.8), (F.9) and the fact that \( \nu_{K_N} \leq \nu_{K_N} \leq [1 - \rho(A - BK_N)]^{-2} \cdot \text{poly}_2(\|K_N\|_*) \), we know that
\[
|J_2(K_N, b_{h+1}) - J_2(K_N, \hat{b}_{h+1})| \leq \gamma^b \cdot \text{poly}_1(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)) \cdot \| \nabla_b J_2(K_N, b_h) - \tilde{\nabla}_b J_2(K_N, b_h) \|_2 \cdot [1 - \rho(A - BK_N)]^{-1}
\]
\[
+ (\gamma^b)^2 \cdot \text{poly}_2(\|K_N\|_*) \cdot \| \nabla_b J_2(K_N, b_h) - \tilde{\nabla}_b J_2(K_N, b_h) \|_2 \cdot [1 - \rho(A - BK_N)]^{-2}.
\]

Note that from the definition of \( \nabla_b J_2(K_N, b_h) \) and \( \tilde{\nabla}_b J_2(K_N, b_h) \), we have
\[
\| \nabla_b J_2(K_N, b_h) - \tilde{\nabla}_b J_2(K_N, b_h) \|_2 \
\leq \| \tilde{\nabla}^{22}_{K_N} - \nabla^{22}_{K_N} \|_2 \cdot \| K_N \|_* \cdot \| \hat{\mu}_{K_N, b_h} \|_2 + \| \nabla_{K_N} \|_* \cdot \| \hat{\mu}_{K_N, b_h} - \mu_{K_N, b_h} \|_2 + \| \nabla_{K_N} \|_* \cdot \| \hat{\mu}_{K_N, b_h} - \mu_{K_N, b_h} \|_2 + \| \nabla_{K_N} \|_* \cdot \| \hat{\mu}_{K_N, b_h} - \mu_{K_N, b_h} \|_2.
\]

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 \( \| \mu_{K_N, b_h} \|_2 \leq J(K_N, b_0)/\sigma_{\min}(Q) \), we know that with probability at least \( 1 - (T^n)^{-4} - (T^b)^{-6} \), it holds that
\[
\| \nabla_b J_2(K_N, b_h) - \tilde{\nabla}_b J_2(K_N, b_h) \|_2 \
\leq \lambda_{K_N}^{-1} \cdot \text{poly}_2(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)) \cdot \left[\frac{\log^3 T_n}{(T^n)^{1/4}(1 - \rho)^2} + \frac{\log^{1/2} T^b}{(T^n)^{1/8} \cdot (1 - \rho)}\right].
\]

To show that (D.19) holds, we only need to pick \( \gamma^b, T^n \) and \( \hat{T}^b_n \) such that
\[
\gamma^b \cdot \text{poly}_1(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)) \cdot \lambda_{K_N}^{-1} \cdot \text{poly}_2(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)) \cdot \left[\frac{\log^3 T^n}{(T^n)^{1/4}(1 - \rho)^2} + \frac{\log^{1/2} \hat{T}^b}{(T^n)^{1/8} \cdot (1 - \rho)}\right] \cdot [1 - \rho(A - BK_N)]^{-1} \cdot (\gamma^b)^2 \cdot \text{poly}_2(\|K_N\|_*) \cdot \lambda_{K_N}^{-1}
\]
\[
\leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2,
\]
in other words, we pick
\[
\gamma^b \leq \min\left\{1 - \rho(A - BK_N), [1 - \rho(A - BK_N)]^{-2} \cdot (\|B\|_*^2 \cdot \|K_N\|_*^2 \cdot \|R\|_* + \|B\|_* \cdot \|Q\|_*),\right\},
\]
\[
T^n_b = \text{poly}_4(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)) \cdot \lambda_{K_N}^{-4} \cdot \nu_{K_N}^{-4} \cdot (1 - \rho)^{-10} \cdot \varepsilon^{-5},
\]
\[
\hat{T}^b_n = \text{poly}_5(\|K_N\|_F, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)) \cdot \lambda_{K_N}^{-2} \cdot \nu_{K_N}^{-2} \cdot (1 - \rho)^{-16} \cdot \varepsilon^{-8},
\]
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. \(\square\)
F.5 Proof of Lemma \[D.6\]

\textit{Proof.} \textbf{Part 1.} First we proceed to prove that \( \zeta_{K,b} \in \mathcal{V}_c \). Note that from Condition \[B.7\] we know that \( \zeta_{K,b} = J(K,b) \) satisfies that \( 0 \leq \zeta_{K,b} \leq J(K_0,b_0) \). It remains to show that \( \zeta_{K,b} = \alpha_{K,b} \) satisfies that \( \| \alpha_{K,b} \|_2 \leq M_\zeta \). By the definition of \( \alpha_{K,b} \) in \[B.6\], we know that

\[
\| \alpha_{K,b} \|_2^2 \leq \| \Upsilon_K \|_F^2 + \| \Upsilon_K \|_F^2 \cdot \left( \left\| \mu_{K,b} \right\|_F^2 + \left\| \mu_{K,b} \right\|_F^2 \right) + \left( \| A \|_F + \| B \|_F \right)^2 \cdot \left( \| P_{K,b} \|_F + \| A \|_F + \| B \|_F \right) \cdot \left( \| f_{K,b} \|_2 \right)^2
\]

(\ref{F.10})

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

\[
J(K,b) = \text{tr}(P_K \Psi_e) + (\mu_{K,b})^\top Q \mu_{K,b} + (\mu_{K,b})^\top R \mu_{K,b} + \sigma^2 \cdot \text{tr}(R) + \mu^\top Q \mu.
\]

Thus, we have

\[
J(K_0,b_0) \geq J(K_0,b) \geq \sigma_{\min}(\Psi_w) \cdot \text{tr}(P_K) \geq \sigma_{\min}(\Psi_w) \cdot \| P_K \|_F,
\]

\[
J(K_0,b_0) \geq J(K_0,b) \geq (\mu_{K,b})^\top Q \mu_{K,b} \geq \sigma_{\min}(Q) \cdot \| \mu_{K,b} \|_2,
\]

\[
J(K_0,b_0) \geq J(K_0,b) \geq (\mu_{K,b}^u)^\top R \mu_{K,b}^u \geq \sigma_{\min}(R) \cdot \| \mu_{K,b}^u \|_2,
\]

which imply that

\[
\| P_K \|_F = J(K_0,b_0)/\sigma_{\min}(\Psi_w), \quad \| \mu_{K,b} \|_2 = J(K_0,b_0)/\sigma_{\min}(Q), \quad \| \mu_{K,b}^u \|_2 = J(K_0,b_0)/\sigma_{\min}(R).
\]

\textbf{Part 2.} For \( \Upsilon_K \), note that we decompose the matrix in the following way:

\[
\Upsilon_K = \begin{pmatrix} Q & R \\ A^\top & \mu \end{pmatrix} P_K \left( \begin{array}{c} 0 \\ B \end{array} \right),
\]

which gives

\[
\| \Upsilon_K \|_F \leq (\| Q \|_F + \| R \|_F) + (\| A \|_F + \| B \|_F) \cdot \| P_K \|_F,
\]

\[
\| \Upsilon_K \|_F \leq (\| Q \|_F + \| R \|_F) + \sqrt{m} \cdot \| J(K_0,b_0)/\sigma_{\min}(\Psi_w),
\]

\[
\| \Upsilon_K \|_F \leq (\| Q \|_F + \| R \|_F) + \sqrt{m} \cdot \| J(K_0,b_0)/\sigma_{\min}(\Psi_w),
\]

\textbf{F.11}

Combining \textbf{F.11} and the fact that \( \| U \|_F \leq \sqrt{m} \cdot \| U \|_F \) for any \( U \in \mathbb{R}^{m \times m} \), we know that

\[
\| \Upsilon_K \|_F \leq (\| Q \|_F + \| R \|_F) + \sqrt{m} \cdot \| J(K_0,b_0)/\sigma_{\min}(\Psi_w),
\]

\[
\| \Upsilon_K \|_F \leq (\| Q \|_F + \| R \|_F) + \sqrt{m} \cdot \| J(K_0,b_0)/\sigma_{\min}(\Psi_w),
\]

\textbf{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

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

we therefore upper bound \( f_{K,b} \) as

\[
\| f_{K,b} \|_2 \leq \left[ \| K_0,b_0 \|_F \cdot \sigma_{\min}(\Psi_w)^{-1} \right] \cdot \left[ (\| A - BK \|_F)^{-1} \cdot (\| A - BK \|_F) \right] \cdot \left[ \| f_{K,b} \|_2 \right]
\]

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

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

\[
\xi^1(c) = \zeta^1 - J(K,b^K), \quad \xi^2(c) = \mathbb{E}_{\pi_{K,b,K}} \left[ \psi(x,u) \right] \xi^1 + \Theta_{K,b,K} \xi^2 - \mathbb{E}_{\pi_{K,b,K}} \left[ c(x,u) \psi(x,u) \right].
\]

\textbf{F.14}

Then we have

\[
\| \xi^1 \|_2 = \| \zeta^1 - J(K,b^K) \|_2 \leq J(K_0,b_0)
\]

\textbf{F.15}
and also
\[ \| \xi^2(\zeta) \|_2 \leq \left\| \mathbb{E}_{\pi_{\zeta, K}} [\psi(x, u)] \xi^2 \right\|_2 + \left\| \Theta_{K, b} \|_* \cdot \| \xi^2 \|_2 + \left\| \mathbb{E}_{\pi_{\zeta, K}} [c(x, u)\psi(x, u)] \right\|_2. \] (F.16)

Note that we upper bound \( B_1 \) as
\[ B_1 \leq J(K_0, b_0) \cdot \left\| \mathbb{E}_{\pi_{\zeta, K}} [\psi(x, u)] \right\|_2 \leq J(K_0, b_0) \cdot \| \Sigma_z \|_F, \] (F.17)
where \( \Sigma_z \) is given in (E.15). Also, by Proposition B.6, we bound \( B_2 \) as
\[ B_2 \leq 4(1 + \| K \|_F^3) \cdot \| \Phi_K \|_* \cdot (M_{\zeta, 1} + M_{\zeta, 2}) \cdot \left[ 1 - \rho(A - BK) \right]^{-1}. \] (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 \( \mathbb{E}_{\pi_{\zeta, K}} [c(x, u)\psi(x, u)] \) has the following form
\[
\mathbb{E}_{\pi_{\zeta, K}} [c(x, u)\psi(x, u)] = \left(2 \text{vec} [\Sigma_z \text{diag}(Q, R) \Sigma_z + (\Sigma_z, \text{diag}(Q, R)) \Sigma_z] \right) \Sigma_z \left(2Q\mu_{K,b} \cdot 2R\mu_{K,b} \cdot \| \text{vec}(\Sigma_z) \|_0 \right) \cdot \left(0_n \cdot 0_k \right).
\]

Here the matrix \( \Sigma_z \) 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. \( \square \)

From Lemma F.1 and (F.11), we obtain the upper bound for \( B_3 \)
\[ B_3 \leq 3[\| Q \|_F + \| R \|_F + J(K_0, b_0)\| Q \|_* / \sigma_{\text{min}}(Q) + J(K_0, b_0)\| R \|_* / \sigma_{\text{min}}(R)] \cdot \| \Sigma_z \|_F^2. \] (F.19)

Moreover, by the definition of \( \Sigma_z \) in (E.15), combining the triangle inequality, we have the following bounds for the Frobenius norm and spectral norm of \( \Sigma_z \), respectively:
\[ \| \Sigma_z \|_F \leq 2(\| K \|_F^2) \cdot \| \Phi_K \|_*, \quad \| \Sigma_z \|_* \leq 2(1 + \| K \|_F^2) \cdot \| \Phi_K \|_* . \] (F.20)

Also, by similar techniques we used in deriving (F.11), we have
\[ J(K_0, b_0) \geq J(K, b^K) \geq \text{tr}[(Q + K^\top RR)\Phi_K] \geq \| \Phi_K \|_* \cdot \sigma_{\text{min}}(Q), \]
which gives the upper bound for \( \Phi_K \) as follows
\[ \| \Phi_K \|_* \leq J(K_0, b_0) / \sigma_{\text{min}}(Q). \] (F.21)

Therefore, combining (F.16), (F.17), (F.18), (F.19), (F.20) and (F.21), we know that
\[ \| \xi^2(\zeta) \|_2 \leq C \cdot (M_{\zeta, 1} + M_{\zeta, 2}) \cdot J(K_0, b_0)^2 / \sigma_{\text{min}}^2(Q) \cdot (1 + \| K \|_F^2)^3 \cdot \left[ 1 - \rho(A - BK) \right]^{-1}. \] (F.22)

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

**F.6 PROOF OF LEMMA D.7**

**Proof.** Assume that \( \tilde{z}_0 \sim \mathcal{N}(\mu, \Sigma) \). Recall that the following transition holds
\[ \tilde{z}_{t+1} = L\tilde{z}_t + \nu + \delta_t. \] (F.23)

We know that
\[ \tilde{z}_t \sim \mathcal{N} \left( L^t \mu_1 + \sum_{i=0}^{t-1} L^i \cdot \nu_i, (L^\top)^t \Sigma_i L^t + \sum_{i=0}^{t-1} (L^\top)^i \Psi_i L^i \right), \] (F.24)
where $\Psi_\delta$ is the covariance matrix of $\delta_t$, 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 $\mu_z$ takes the form

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

Therefore, combining (F.24), we have

$$
\mathbb{E}(\tilde{\mu}_z) = \mu_z + \frac{1}{T} \sum_{t=1}^{\tilde{T}} L^t \mu_t - \frac{1}{T} \sum_{t=1}^{\tilde{T}} \sum_{i=t}^{\infty} L^i \nu.
$$

It holds that

$$
\left\| \frac{1}{T} \sum_{t=1}^{\tilde{T}} L^t \mu_t - \frac{1}{T} \sum_{t=1}^{\tilde{T}} \sum_{i=t}^{\infty} L^i \nu \right\|_2 \leq \left[ \frac{1 - \rho(L)}{T} \right] \cdot \| \mu_t \|_2 + \left[ \frac{1 - \rho(L)}{T} \right] \cdot \| \nu \|_2 \leq M_\mu \cdot (1 - \rho)^{-2} \cdot \| \mu_z \|_2 / \tilde{T},
$$

where $M_\mu$ 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 $\| \Sigma \|_F \leq 2 \| \Sigma_1 \|_F + 2 \| \Sigma_2 \|_F$. Combining (F.24), we know that $\tilde{\mu}_z \sim \mathcal{N}(\mathbb{E} \tilde{\mu}_z, \tilde{\Sigma}_\tilde{T}/\tilde{T})$, where $\tilde{\Sigma}_\tilde{T}$ satisfies that

$$
\tilde{T} / 2 \cdot \| \tilde{\Sigma}_\tilde{T} \|_F \leq \sum_{t=1}^{\tilde{T}} \rho(L)^{2t} \cdot \| \Sigma_t \|_F + \sum_{t=1}^{\tilde{T}} \sum_{i=0}^{t-1} \rho(L)^{2t} \cdot \| \Psi_\delta \|_F \leq \left[ \frac{1 - \rho(L)}{2} \right] \cdot \| \Sigma \|_F + \tilde{T} \cdot \left[ \frac{1 - \rho(L)}{2} \right] \cdot \| \Psi_\delta \|_F,
$$

which implies that

$$
\| \tilde{\Sigma}_\tilde{T} \|_F \leq M_\Sigma \cdot (1 - \rho)^{-1} \cdot \| \Sigma \|_F,
$$

where $M_\Sigma$ is a positive absolute constant. Combining (F.25), (F.26) and (F.27), we conclude that $\tilde{\mu}_z$ 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

$$
\hat{F}(\zeta, \xi) = \left\{ \mathbb{E}(\hat{\psi}) \xi^1 + \mathbb{E}[(\hat{\psi} - \hat{\psi}) \hat{\psi}^\top] \xi^2 - \mathbb{E}(\hat{\psi} \hat{\psi}) \right\}^\top \xi^2 + [\xi^1 - \mathbb{E}(\xi^1)] \cdot \xi^1 - 1/2 \cdot \| \xi \|_2^2,
$$

where $\hat{\psi} = \hat{\psi}(x, u)$ 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 $\hat{F}(\zeta, \xi)$ is still random, where the randomness comes from the estimated feature vectors. Note that $|F(\zeta, \xi) - \hat{F}(\zeta, \xi)| \leq |F(\zeta, \xi) - \tilde{F}(\zeta, \xi)| + |\tilde{F}(\zeta, \xi) - \hat{F}(\zeta, \xi)|$. Thus, we only need to characterize $|F(\zeta, \xi) - \tilde{F}(\zeta, \xi)|$ and $|\tilde{F}(\zeta, \xi) - \hat{F}(\zeta, \xi)|$. 

43
Part 1. First we characterize \(|F(\zeta, \xi) - \hat{F}(\zeta, \xi)|\). Note that by algebra, we have

\[
|F(\zeta, \xi) - \hat{F}(\zeta, \xi)| = \left| \left\{ E(\psi - \hat{\psi}) \xi^1 + E[(\psi - \hat{\psi})^T (\hat{\psi} - \psi)] \xi^2 - E[c(\psi - \hat{\psi})] \right\} \right|_D^2 \\
\leq E(||\psi - \hat{\psi}||_2) \left[ ||\xi^1|| + E(||\psi - \hat{\psi}||_2 + 2||\hat{\psi}||_2) ||\xi^2||_2 + E(c) \right] \cdot ||\xi^2||_2,
\]

where the expectation is only taken over the trajectory generated by the state transition and the policy \(\pi_{K,b}\). From \(P(||\bar{\mu}_z - \mu_z + 1/T \cdot \mu_{\bar{\psi}}||_2 \leq C_1 \geq 1 - \bar{T}^{-6}\), we know that with probability at least \(1 - \bar{T}^{-6}\), the term \(E(||\psi - \hat{\psi}||_2 + 2||\hat{\psi}||_2)\) is upper bounded by a polynomial \(\text{poly}(||\Phi_K||_*, ||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0))\). Also, the term \(E(c)\) is also upper bounded by a polynomial \(\text{poly}(||\Phi_K||_*, ||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0))\) by the definition of the cost function \(c(x,u)\). Therefore, combining Condition B.7 and (F.28), with probability at least \(1 - \bar{T}^{-6}\), we obtain that

\[
|F(\zeta, \xi) - \hat{F}(\zeta, \xi)| \leq E(||\psi - \hat{\psi}||_2) \cdot \text{poly}(||\Phi_K||_*, ||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0)).
\]

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

\[
||\psi(x,u) - \hat{\psi}(x,u)||_2^2 = ||\bar{\mu}_z - \mu_z||_2^2 + ||z(\bar{\mu}_z - \mu_z)^T (\bar{\mu}_z - \mu_z)||_F^2 + ||\mu_z \bar{\mu}_z^T - \bar{\mu}_z \mu_z^T||_F^2 \\
\leq \text{poly}(||\Phi_K||_*, ||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0)) \cdot ||\mu_z - \bar{\mu}_z||_2^2,
\]

where \(z = [x^T, u^T]^T\). Also, by Lemma D.7, we know that

\[
||\mu_z - \bar{\mu}_z||_2 \leq \frac{\log \bar{T}}{T^{1/4}} \cdot (1 - \rho)^{-2} \cdot \text{poly}(||\Phi_K||_*, ||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0)),
\]

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

\[
|F(\zeta, \xi) - \hat{F}(\zeta, \xi)| \leq \frac{\log \bar{T}}{T^{1/4}} \cdot (1 - \rho)^{-2} \cdot \text{poly}(||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0))
\]

with probability at least \(1 - \bar{T}^{-6}\).

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

\[
|\hat{F}(\zeta, \xi) - \hat{F}(\zeta, \xi)| = \left| \left\{ E(\psi - \hat{\psi}) \xi^1 + E[(\psi - \hat{\psi})^T (\hat{\psi} - \psi)] \xi^2 - E(c(\psi - \hat{\psi}) \right\} \right|_D^2 \\
\leq \left| E(\psi) \xi^1 + E(\hat{\psi}) \xi^2 - E(\hat{\psi}) \xi^1 \right|_D^2 + \text{E}(\hat{\psi}) \xi^1 \cdot \mathbb{1}_{E^*} + \left| \text{E}(\hat{\psi} \hat{\psi}^T) \xi^2 \right|_D^2 \cdot \mathbb{1}_{(E' \cap \mathcal{E}'^c)},
\]

where \(E'\) is an event defined as

\[
E' = \left( \bigcap_{t \in [T]} \left\{ ||z^t_q - \mu_z + 1/T \cdot \mu_{\bar{\psi}}||_2^2 - \text{tr}(\hat{\Sigma}_z) \right\} \right) \bigcap \mathcal{E}_2,
\]

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

\[
|\hat{F}(\zeta, \xi) - \hat{F}(\zeta, \xi)| \leq \frac{1}{T} + \frac{1}{T^{1/4}} \cdot \text{poly}(||K||_F, ||b||_2, ||\mu||_2, J(K_0, b_0))
\]

for sufficiently large \(T\) and \(\bar{T}\). Therefore, combining (F.32) and (F.34), by triangular inequality, we finish the proof of the lemma. \(\square\)
F.8 Proof of Lemma E.2

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

$$\psi(x, u) = \left( \text{svec} \left[ (z - \mu_z)(z - \mu_z)^T \right] \right).$$

We then have

$$\psi(x, u) - \psi(x', u') = \left( \text{svec} \left[ yy^T - (Ly + \delta)(Ly + \delta)^T \right] \right),$$

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

$$\left( \text{svec}(M) \right)^T \Theta_{K,b} \left( \text{svec}(N) \right) = \mathbb{E}_{y, \delta} \left\{ \left( \text{svec}(M) \right)^T \left( \text{svec}(y y^T) \right) \left( \text{svec}[yy^T - (Ly + \delta)(Ly + \delta)^T] \right)^T \left( \text{svec}(N) \right) \right\}$$

$$= \mathbb{E}_{y, \delta} \left\{ (M, yy^T) + m^T y \cdot [(N, yy^T - (Ly + \delta)(Ly + \delta)^T) + n^T (y - Ly - \delta)] \right\}$$

$$= \mathbb{E}_y \left[ (yy^T, M) \cdot (yy^T - Lyy^T L^T - \Psi_\delta, N) \right] + \mathbb{E}_y \left[ (yy^T, M) \cdot n^T (y - Ly) \right]$$

$$+ \mathbb{E}_y \left[ m^T y \cdot (yy^T - Lyy^T L^T - \Psi_\delta, N) \right] + \mathbb{E}_y \left[ m^T y \cdot n^T (y - Ly) \right].$$

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

$$A_4 = \mathbb{E}_y \left[ m^T y \cdot (y - Ly)^T n \right] = \mathbb{E}_y \left[ m^T y y^T (I - L)^T n \right] = m^T \Sigma_z (I - L)^T n. \quad (F.36)$$

For $A_1$, by algebra, we have

$$A_1 = \mathbb{E}_y \left[ (yy^T, M) \cdot (yy^T - Lyy^T L^T - \Psi_\delta, N) \right]$$

$$= \mathbb{E}_y \left[ (yy^T, M) \cdot (yy^T - Lyy^T L^T, N) \right] - \mathbb{E}_y \left[ (yy^T, M) \cdot (\Psi_\delta, N) \right]$$

$$= \mathbb{E}_y \left[ y^T M y \cdot y^T (N - L^T NL)y \right] - \left( \Sigma_z, M \right) \cdot (\Psi_\delta, N)$$

$$= \mathbb{E}_{u \sim \mathcal{N}(0, I)} \left[ u^T \Sigma_z^{1/2} M \Sigma_z^{1/2} u \cdot u^T \Sigma_z^{1/2} (N - L^T NL) \Sigma_z^{1/2} u \right] - \left( \Sigma_z, M \right) \cdot (\Psi_\delta, N). \quad (F.37)$$

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

$$A_1 = 2 \text{tr} \left[ \Sigma_z^{1/2} M \Sigma_z^{1/2} \cdot \Sigma_z^{1/2} (N - L^T NL) \Sigma_z^{1/2} \right]$$

$$+ \text{tr} \left[ \Sigma_z^{1/2} M \Sigma_z^{1/2} \cdot \text{tr} \left[ \Sigma_z^{1/2} (N - L^T NL) \Sigma_z^{1/2} \right] - \left( \Sigma_z, M \right) \cdot (\Psi_\delta, N) \right]$$

$$= 2 \left( \Sigma_z (N - L^T NL) \Sigma_z \right)$$

$$+ \left( \Sigma_z, M \right) \cdot (\Psi_\delta, N)$$

$$= 2 \left( \Sigma_z (N - L^T NL) \Sigma_z \right) + \left( \Sigma_z, M \right) \cdot (\Psi_\delta, N). \quad (F.38)$$

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

$$A_1 = 2 \text{svec}(M)^T \text{svec} \left[ \Sigma_z (N - L^T NL) \Sigma_z \right]$$

$$= 2 \text{svec}(M)^T \left[ \Sigma_z \otimes_s \Sigma_z - (\Sigma_z L^T) \otimes_s (\Sigma_z L^T) \right] \text{svec}(N)$$

$$= 2 \text{svec}(M)^T \left[ (\Sigma_z \otimes_s \Sigma_z) (I - L \otimes_s L)^T \right] \text{svec}(N). \quad (F.38)$$
Combining (F.35), (F.36) and (F.38), we obtain that
\[
\begin{pmatrix}
\text{svec}(M) \\
\text{svec}(N)
\end{pmatrix}_m^T \Theta_{K,b} \begin{pmatrix}
\text{svec}(M) \\
\text{svec}(N)
\end{pmatrix}_n
= \text{svec}(M)^T [2(\Sigma_z \otimes \Sigma_z)(I - L \otimes_s L)^\top] \text{svec}(N) + m^\top \Sigma_z (I - L)^\top n
= \begin{pmatrix}
\text{svec}(M) \\
\text{svec}(N)
\end{pmatrix}_m^T \begin{pmatrix}
2(\Sigma_z \otimes \Sigma_z)(I - L \otimes_s L)^\top & 0 \\
0 & \Sigma_z (I - L)^\top
\end{pmatrix} \begin{pmatrix}
\text{svec}(M) \\
\text{svec}(N)
\end{pmatrix}_n.
\]
Thus, the matrix \( \Theta_{K,b} \) takes the following form
\[
\Theta_{K,b} = \begin{pmatrix}
2(\Sigma_z \otimes \Sigma_z)(I - L \otimes_s L)^\top & 0 \\
0 & \Sigma_z (I - L)^\top
\end{pmatrix}.
\]
This concludes the lemma. \( \square \)

### F.9 Proof of Lemma F.3

**Proof.** By the property of the spectral norm, we upper bound \( \|\Theta_{K,b}^{-1}\|_* \) as
\[
\|\Theta_{K,b}^{-1}\|_*^2 \leq 1 + \|\Theta_{K,b}^{-1}\|_2^2 + \|\Theta_{K,b}^{-1}\|_*^2.
\]
(\( F.39 \))

We proceed to bound the RHS of (\( F.39 \)). For the term \( \Theta_{K,b}^{-1} \tilde{\sigma}_z \), combining Lemma F.2 and (E.16), we have
\[
\Theta_{K,b}^{-1} \tilde{\sigma}_z = \begin{pmatrix}
1/2 \cdot (I - L \otimes_s L)^{-\top} (\Sigma_z \otimes \Sigma_z)^{-1} \cdot \text{svec}(\Sigma_z) \\
0_k \end{pmatrix}
= \begin{pmatrix}
1/2 \cdot (I - L \otimes_s L)^{-\top} (\Sigma_z^{-1} \otimes \Sigma_z^{-1}) \cdot \text{svec}(\Sigma_z) \\
0_k \end{pmatrix}
= \begin{pmatrix}
1/2 \cdot (I - L \otimes_s L)^{-\top} \cdot \text{svec}(\Sigma_z^{-1}) \\
0_k \end{pmatrix},
\]
where we use the property of the symmetric Kronecker product in the second and last line. Therefore, we have
\[
\|\Theta_{K,b}^{-1} \tilde{\sigma}_z\|_2 = 1/2 \cdot \| (I - L \otimes_s L)^{-\top} \cdot \text{svec}(\Sigma_z^{-1})\|_2
\leq 1/2 \cdot \| (I - L \otimes_s L)^{-\top}\|_* \cdot \| \text{svec}(\Sigma_z^{-1})\|_2
\leq 1/2 \cdot [1 - \rho^2(L)]^{-1} \cdot \| \Sigma_z^{-1}\|_*
\leq 1/2 \cdot \sqrt{k + m} \cdot [1 - \rho^2(L)]^{-1} \cdot \| \Sigma_z^{-1}\|_*
= 1/2 \cdot \sqrt{k + m} \cdot [1 - \rho^2(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-1},
\]
where in the third line we use Lemma G.2 on the matrix \( L \otimes_s L \). Similarly, we upper bound \( \|\Theta_{K,b}^{-1}\|_* \) in the sequel
\[
\|\Theta_{K,b}^{-1}\|_* \leq \min \left\{ 1/2 \cdot [1 - \rho^2(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-2}, [1 - \rho(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-1} \right\}.
\]
(\( F.41 \))

Thus, combining (\( F.39 \)), (\( F.40 \)) and (\( F.41 \)), we obtain that
\[
\|\Theta_{K,b}^{-1}\|_*^2 \leq 1 + 1/2 \cdot \sqrt{k + m} \cdot [1 - \rho^2(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-1}
+ \min \left\{ 1/2 \cdot [1 - \rho^2(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-2}, [1 - \rho(L)]^{-1} \cdot [\sigma_{\min}(\Sigma_z)]^{-1} \right\}.
\]
(\( F.42 \))

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
\[
\text{svec}(z) = \mathbb{E}_{\pi_{K,b},x \sim (\mu_{K,b}, \Phi_K)} \left\{ [s^\top (x - \mu_{K,b}) + r^\top (u + K(u - K,b) - b)]^2 \right\}
= \mathbb{E}_{x \sim (\mu_{K,b}, \Phi_K), \eta \sim \mathcal{N}(0, I)} \left\{ (s - K^\top r)^\top (x - \mu_{K,b}) + \sigma r^\top \eta \right\}
= \mathbb{E}_{x \sim (\mu_{K,b}, \Phi_K)} \left\{ (s - K^\top r)^\top (x - \mu_{K,b}) \right\} + \mathbb{E}_{\eta \sim \mathcal{N}(0, I)} [(\sigma r^\top \eta)^2] \quad \text{(F.43)}
The first term on the RHS of (F.43) is lower bounded as
\[
E_{x \sim (\mu_K, \Phi_K)} \left\{ \left( s - K^T r \right)^\top (x - \mu_K) \right\} = (s - K^T r)^\top \Phi_K (s - K^T r)
\]
\[
\geq \| s - K^T r \|^2_2 \cdot \sigma_{\min}(\Phi_K) \geq \| s - K^T r \|^2_2 \cdot \sigma_{\min}(\Psi_\omega),
\]
where the last inequality comes from the fact that \( \sigma_{\min}(\Phi_K) \geq \sigma_{\min}(\Psi_\omega) \) by (5.3). The second term on the RHS of (F.43) takes the form \( E_{\eta \sim N(0, I)} \left[ (\sigma r^\top \eta)^2 \right] = \sigma^2 \| r \|^2_2 \). Therefore, combining (F.43) and (F.44), we have
\[
\left( s \right)^\top \Sigma z \left( s \right) \geq \| s - K^T r \|^2_2 \cdot \sigma_{\min}(\Psi_\omega) + \sigma^2 \| r \|^2_2
\]
\[
\geq \sigma_{\min}(\Psi_\omega) \cdot \| s \|^2_2 + \left[ \sigma^2 - \| K \|^2_2 \cdot \sigma_{\min}(\Psi_\omega) \right] \cdot \| r \|^2_2.
\]
From this, we see that \( \sigma_{\min}(\Sigma_z) \) is lower bounded by \( \min\{ \sigma_{\min}(\Psi_\omega), \sigma^2 - \| K \|^2_2 \cdot \sigma_{\min}(\Psi_\omega) \} \), which only depends on \( \sigma_{\min}(\Psi_\omega) \) and \( \sigma \). Thus, combining (F.42), we know that \( \| \Theta_{K,b} \|_\ast \) is upper bounded by some constant \( \lambda_K \), where \( \lambda_K \) only depends on \( \rho(A - BK) \), \( \sigma \) 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 D.2, we have
\[
| J_1(\tilde{K}_{n+1}) - J_1(K_{n+1}) | = \max \left\{ \| P_{K_{n+1}} - P_{\tilde{K}_{n+1}} \|_\ast , \| \Psi_\omega \|_F \right\} \leq \| P_{K_{n+1}} - P_{\tilde{K}_{n+1}} \|_\ast \cdot \| \Psi_\omega \|_F.
\]
(F.45)
The following lemma helps establish the upper bound of the term \( \| P_{K_{n+1}} - P_{\tilde{K}_{n+1}} \|_\ast \).

**Lemma F.2.** Suppose that the parameters \( K \) and \( \tilde{K} \) satisfy
\[
\| \tilde{K} - K \|_\ast \cdot \left( \| A - BK \|_\ast + 1 \right) \cdot \| \Phi_K \|_\ast \leq \sigma_{\min}(\Psi_\omega) / 4 \cdot \| B \|_\ast^{-1},
\]
then it holds that
\[
\| P_{\tilde{K}} - P_K \|_\ast \leq 6 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \| \Phi_K \|_\ast \cdot \| K \|_\ast \cdot \| R \|_\ast \cdot \| \tilde{K} - K \|_\ast
\]
\[
\cdot \left( \| B \|_\ast \cdot \| K \|_\ast \right) \cdot \| A - BK \|_\ast + \| B \|_\ast \cdot \| K \|_\ast + 1).
\]
(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 \( \tilde{K}_{n+1} \) and \( K_{n+1} \) satisfy (F.46). Note that from (D.10) and (D.11), we have
\[
\| \tilde{K}_{n+1} - K_{n+1} \|_\ast \cdot \left( \| A - BK_{n+1} \|_\ast + 1 \right) \cdot \| \Phi_{\tilde{K}_{n+1}} \|_\ast \leq \gamma \cdot \| \tilde{K}_n - K_n \|_F \cdot \left( 1 + \| K_n \|_\ast \right) \cdot \left( \| A - B\tilde{K}_{n+1} \|_\ast + 1 \right) \cdot \| \Phi_{\tilde{K}_{n+1}} \|_\ast.
\]
(F.48)
Now we proceed to upper bound the RHS of (F.48). For the term \( \| A - B\tilde{K}_{n+1} \|_\ast \), we have
\[
\| A - B\tilde{K}_{n+1} \|_\ast \leq \| A - BK_{n+1} \|_\ast + \gamma \cdot \| B \|_\ast \cdot \| \tilde{K}_n - K_n \|_\ast \cdot \| K_n \|_\ast \cdot \left( \| A - B\tilde{K}_{n+1} \|_\ast + 1 \right) \cdot \| \Phi_{\tilde{K}_{n+1}} \|_\ast.
\]
(F.49)
And by the definition of \( \tilde{Y}_K \) in (3.7), we upper bound \( \| \tilde{Y}_K \|_\ast \) as
\[
\| \tilde{Y}_K \|_\ast \leq \| Q \|_\ast + \| R \|_\ast + \left( \| A \|_F + \| B \|_F \right)^2 \cdot \| P_{\tilde{K}} \|_\ast
\]
\[
\leq \| Q \|_\ast + \| R \|_\ast + \left( \| A \|_F + \| B \|_F \right)^2 \cdot J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon),
\]
(F.50)
where the last line comes from the fact that \( \| P_{\tilde{K}} \|_\ast \leq J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon) \leq J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon). \) As for the term \( \| \Phi_{\tilde{K}_{n+1}} \|_\ast \) in (F.48), we upper bound it as
\[
\| \Phi_{\tilde{K}_{n+1}} \|_\ast \leq J_1(\tilde{K}_{n+1}) \cdot \sigma_{\min}^{-1}(Q) \leq J_1(K_0) \cdot \sigma_{\min}^{-1}(Q).
\]
(F.51)
Therefore, combining (F.48), (F.49), (F.50) and (F.51), we know that

\[ \| \tilde{K}_{n+1} - K_{n+1} \|_s \cdot (\| A - B \tilde{K}_{n+1} \|_s + 1) \cdot \| \Phi_{K,n+1} \|_s \leq \text{poly}_1 (\| K_n \|_s, J_1(K_0)) \cdot \| \tilde{\Upsilon}_{K_n} - \Upsilon_{K_n} \|_F. \]

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

\[ 6 \cdot \sigma_{\text{min}}^{-1} (\Psi_{\omega}) \cdot \| \Phi_{K,n+1} \|_s \cdot \| \tilde{K}_{n+1} \|_s \cdot \| R \|_s \cdot \| \tilde{K}_{n+1} - K_{n+1} \|_s \]

\[ \cdot (\| B \|_s \cdot \| \tilde{K}_{n+1} \|_s \cdot \| A - B \tilde{K}_{n+1} \|_s + \| B \|_s \cdot \| \tilde{K}_{n+1} \|_s + 1) \]

\[ \leq \text{poly}_2 (\| K_n \|_s, J_1(K_0)) \cdot \| \tilde{\Upsilon}_{K_n} - \Upsilon_{K_n} \|_F. \]

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

\[ \| \tilde{\Upsilon}_{K_n} - \Upsilon_{K_n} \|_F \leq \frac{\text{poly}_3 (\| K_n \|_F, \| \mu \|_2)}{\lambda_{K_n} \cdot (1 - \rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} \]

\[ + \frac{\text{poly}_4 (\| K_n \|_F, \| b^{K_n} \|_2, \| \mu \|_2)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \tilde{T}_n}{\tilde{T}_n^{1/8} \cdot (1 - \rho)}. \]

By choosing \( T_n \) and \( \tilde{T}_n \) such that

\[ \frac{\text{poly}_3 (\| K_n \|_F, \| \mu \|_2)}{\lambda_{K_n} \cdot (1 - \rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} + \frac{\text{poly}_4 (\| K_n \|_F, \| b^{K_n} \|_2, \| \mu \|_2)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \tilde{T}_n}{\tilde{T}_n^{1/8} \cdot (1 - \rho)} \]

\[ \leq \max \left\{ \left[ \text{poly}_1 (\| K_n \|_s, J_1(K_0)) \right]^{-1} \cdot \sigma_{\text{min}} (\Psi_{\omega}) / 4 \cdot \| B \|_F^{-1}, \right\} \]

in other words, we pick

\[ T_n \geq \text{poly}_5 (\| K_n \|_F, \| b^{K_n} \|_2, \| \mu \|_2) \cdot \lambda_{K_n}^{-4} \cdot [1 - \rho (A - BK_n)]^{-9} \cdot \varepsilon^{-5}, \]

\[ \tilde{T}_n \geq \text{poly}_6 (\| K_n \|_F, \| b^{K_n} \|_2, \| \mu \|_2) \cdot \lambda_{K_n}^{-2} \cdot [1 - \rho (A - BK_n)]^{-12} \cdot \varepsilon^{-12}, \]

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

\[ \| J_1(\tilde{K}_{n+1}) - J_1(K_{n+1}) \| \leq \gamma \cdot \sigma_{\text{min}} (\Psi_{\omega}) \cdot \sigma_{\text{min}} (R) \cdot \| \Phi_{K,n+1} \|_s^{-1} \cdot \varepsilon / 4 \]

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

\[ \square \]

F.11 Proof of Lemma F.1

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

\[ c(x, u) = \psi(x, u)^\top \left( \text{svec} \left[ \text{diag}(Q, R) \right] \right) + ((\mu_{K,b})^\top Q_{K,b} + (\mu_{K,b}^u)^\top R_{K,b} + \mu^\top \overline{\Omega} \mu). \]
For any test matrix $V$ and vectors $v_x, v_u$, we proceed by the following calculations

$$
\mathbb{E}_{\pi_{K,b}} \left[ c(x, u) \psi(x, u) \right]^T \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix}
= \mathbb{E}_{\pi_{K,b}} \left\{ \psi(x, u)^T \begin{pmatrix} \text{svec} \left[ \text{diag}(Q, R) \right] \\ \frac{2Q\mu_{K,b}}{2R\mu_{K,b}} \end{pmatrix} \psi(x, u)^T \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix} \right\}
+ \mathbb{E}_{\pi_{K,b}} \left\{ \psi(x, u)^T \left( (\mu_{K,b})^T Q\mu_{K,b} + (\mu_{K,b}^u)^T R\mu_{K,b}^u + \mu^T \mathcal{Q}\mu \right) \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix} \right\} =: D_1 + D_2.
$$

(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

$$
D_1 = \mathbb{E}_{\pi_{K,b}} \left\{ \left[ (z - \mu_z)^T \text{diag}(Q, R)(z - \mu_z) + (z - \mu_z)^T \frac{2Q\mu_{K,b}}{2R\mu_{K,b}} \right] \cdot \left[ (z - \mu_z)^T V(z - \mu_z) + (z - \mu_z)^T \begin{pmatrix} v_x \\ v_u \end{pmatrix} \right] \right\}
= \mathbb{E}_{\pi_{K,b}} \left\{ (z - \mu_z)^T \text{diag}(Q, R)(z - \mu_z) \cdot (z - \mu_z)^T V(z - \mu_z) \right\}
= \mathbb{E}_{\pi_{K,b}} \left\{ \left[ \frac{2Q\mu_{K,b}}{2R\mu_{K,b}} \right]^T (z - \mu_z)(z - \mu_z)^T \begin{pmatrix} v_x \\ v_u \end{pmatrix} \right\}.
$$

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

$$
\mathbb{E}_{\pi_{K,b}} \left\{ (z - \mu_z)^T \text{diag}(Q, R)(z - \mu_z) \cdot (z - \mu_z)^T V(z - \mu_z) \right\}
= 2 \langle \Sigma_z \text{diag}(Q, R) \Sigma_z, V \rangle + \langle \Sigma_z, \text{diag}(Q, R) \rangle \cdot \langle \Sigma_z, V \rangle
= \text{svec} \left\{ 2\Sigma_z \text{diag}(Q, R) \Sigma_z + \langle \Sigma_z, \text{diag}(Q, R) \rangle \cdot \Sigma_z \right\}^T \text{svec}(V).
$$

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

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

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

$$
D_1 = \left( \text{svec} \left[ \Sigma_z \text{diag}(Q, R) \Sigma_z + \langle \Sigma_z, \text{diag}(Q, R) \rangle \Sigma_z \right] \right)^T \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix}.
$$

(F.54)

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

$$
D_2 = ((\mu_{K,b})^T Q\mu_{K,b} + (\mu_{K,b}^u)^T R\mu_{K,b}^u + \mu^T \mathcal{Q}\mu)^T \begin{pmatrix} \text{svec}(\Sigma_z) \\ 0_m \\ 0_k \end{pmatrix} \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix}.
$$

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

Proof. See [Magnus et al. (1978); Magnus (1979)] for a detailed proof. \hfill \Box

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 $1 \leq i \leq j \leq n$.

Proof. See Lemma 2 in [Alizadeh et al. (1998)] for a detailed proof. \hfill \Box

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 $t \geq 0$, we have
\[
\mathbb{P}
\left[
\left|\eta^\top A \eta - \mathbb{E}(\eta^\top A \eta)\right| > t
\right]
\leq 2 \cdot \exp\left[-C \cdot \min\left(t^2 \|A\|_{F}^{-2}, t \|A\|_{*}^{-1}\right)\right].
\]

Proof. See [Rudelson et al. (2013)] for a detailed proof. \hfill \Box