Provably Global Convergence of Actor-Critic: A Case for Linear Quadratic Regulator with Ergodic Cost

Zhuoran Yang
Princeton University
zy6@princeton.edu

Yongxin Chen
Georgia Institute of Technology
yongchen@gatech.edu

Mingyi Hong
University of Minnesota
mhong@umn.edu

Zhaoran Wang
Northwestern University
zhaoran.wang@northwestern.edu

Abstract

Despite the empirical success of the actor-critic algorithm, its theoretical understanding lags behind. In a broader context, actor-critic can be viewed as an online alternating update algorithm for bilevel optimization, whose convergence is known to be fragile. To understand the instability of actor-critic, we focus on its application to linear quadratic regulators, a simple yet fundamental setting of reinforcement learning. We establish a nonasymptotic convergence analysis of actor-critic in this setting. In particular, we prove that actor-critic finds a globally optimal pair of actor (policy) and critic (action-value function) at a linear rate of convergence. Our analysis may serve as a preliminary step towards a complete theoretical understanding of bilevel optimization with nonconvex subproblems, which is NP-hard in the worst case and is often solved using heuristics.

1 Introduction

The actor-critic algorithm [36] is one of the most used algorithms in reinforcement learning [46]. Compared with the classical policy gradient algorithm [73], actor-critic tracks the action-value function (critic) in policy gradient in an online manner, and alternatively updates the policy (actor) and the critic. On the one hand, the online update of critic significantly reduces the variance of policy gradient and hence leads to faster convergence. On the other hand, it also introduces algorithmic instability, which is often observed in practice [33] and parallels the notoriously unstable training of generative adversarial networks [50]. Such instability of actor-critic originates from several intertwining challenges, including (i) function approximation of actor and critic, (ii) improper choice of stepsizes, (iii) the noise arising from stochastic approximation, (iv) the asynchrony between actor and critic, and (v) possibly off-policy data used in the update of critic. As a result, the convergence of actor-critic remains much less well understood than that of policy gradient, which itself is open. Consequently, the practical use of actor-critic often lacks theoretical guidance.

In this paper, we aim to theoretically understand the algorithmic instability of actor-critic. In particular, under a bilevel optimization framework, we establish the global rate of convergence and sample complexity of actor-critic for linear quadratic regulators (LQR) with ergodic cost, a simple yet fundamental setting of reinforcement learning [54], which captures all the above challenges. Compared with the classical two-timescale analysis of actor-critic [11], which is asymptotic in nature and requires finite action space, our analysis is fully nonasymptotic and allows for continuous action space. Moreover, beyond the convergence to a stable equilibrium obtained by the classical two-timescale stochastic approximation via ordinary differential equations, we for the first time establish the linear rate of convergence to a globally optimal pair of actor and critic. In addition, we

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

characterize the required sample complexity. As a technical ingredient and byproduct, we for the first time establish the sublinear rate of convergence for the gradient temporal difference algorithm [66, 67] for ergodic cost and dependent data, which is of independent interest. Furthermore, although we only focus on the setting of LQR, our theoretical analysis framework can be readily extended to general RL problems with other policy optimization methods for the actor (e.g. trust-region policy optimization (TRPO) [57] and proximal policy optimization (PPO) [58]) and other policy evaluation methods for the critic such as TD(0) [65].

Our work adds to two lines of works in machine learning, stochastic analysis, and optimization:

(i) Actor-critic falls into the more general paradigm of bilevel optimization [42, 24, 4]. Bilevel optimization is defined by two nested optimization problems, where the upper-level optimization problem relies on the output of the lower-level one. As a special case of bilevel optimization, minimax optimization is prevalent in machine learning. Recent instances include training generative adversarial neural networks [27], (distributionally) robust learning [63], and imitation learning [31, 15]. Such instances of minimax optimization remain challenging as they lack convexity-concavity in general [25, 56, 16, 53, 38, 17, 18, 19, 41]. The more general paradigm of bilevel optimization remains even more challenging, as there does not exist a unified objective function for simultaneous minimization and maximization. In particular, actor-critic couples the nonconvex optimization of actor (policy gradient) as its upper level and the convex-concave minimax optimization of critic (gradient temporal difference) as its lower level, each of which is challenging to analyze by itself. Most existing convergence analysis of bilevel optimization is based on two-timescale analysis [10]. However, as two-timescale analysis abstracts away most technicalities via the lens of ordinary differential equations, which is asymptotic in nature, it often lacks the resolution to capture the nonasymptotic rate of convergence and sample complexity, which are obtained via our analysis.

(ii) As a proxy for analyzing more general reinforcement learning settings, LQR is studied in a recent line of works [12, 54, 26, 69, 70, 22, 23, 61, 21, 30]. In particular, a part of our analysis is based on the breakthrough of [26], which gives the global convergence of the population-version policy gradient algorithm for LQR and its finite-sample version based on the zeroth-order estimation of policy gradient based on the cumulative reward or cost. However, such zeroth-order estimation of policy gradient often suffers from large variance, as it involves the randomness of an entire trajectory. In contrast, actor-critic updates critic in an online manner, which reduces such variance but also introduces instability and complicates the convergence analysis. In particular, as the update of critic interleaves with the update of actor, the policy gradient for the update of actor is biased due to the inexactness of critic. Meanwhile, the update of critic has a "moving target", as it attempts to evaluate an actor that evolves along the iterations. A key to our analysis is to handle such asynchrony between actor and critic, which is a ubiquitous challenge in bilevel optimization. We hope our analysis may serve as the first step towards analyzing actor-critic in more general reinforcement learning settings.

Notation. For any integer $n > 0$, we denote $\{1, \ldots, n\}$ to be $[n]$. For any symmetric matrix X, let $\mathrm{svec}(X)$ denote the vectorization of the upper triangular submatrix of X with off-diagonal entries weighted by $\sqrt{2}$. Hence, for any symmetric matrices X and Y, we have $\mathrm{tr}(XY) = \langle X, Y \rangle = \mathrm{svec}(X)^\top \mathrm{svec}(Y)$. Meanwhile, let $\mathrm{smat}(\cdot)$ be the inverse operation of $\mathrm{svec}(\cdot)$, which maps a vector to a symmetric matrix. Besides, we denote by $A \otimes_s B$ the symmetric Kronecker product of A and B. We use $\|v\|_2$ to denote the ℓ_2-norm of a vector v. Finally, for a matrix A, we use $\|A\|$, $\|A\|_{\mathrm{fro}}$, and $\rho(A)$ to denote its the operator norm, Frobenius norm, and spectral radius, respectively.

2 Background

In the following, we introduce the background of actor-critic and LQR. In particular, we show that actor-critic can be cast as a first-order online alternating update algorithm for a bilevel optimization problem [42, 24, 4].

2.1 Actor-Critic Algorithm

We consider a Markov decision process, which is defined by $(\mathcal{X}, \mathcal{U}, P, c, D_0)$. Here $\mathcal{X}$ and $\mathcal{U}$ are the state and action spaces, respectively, $P\colon \mathcal{X} \times \mathcal{U} \to \mathcal{P}(\mathcal{X})$ is the Markov transition kernel, $c\colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is the cost function, and $D_0 \in \mathcal{P}(\mathcal{X})$ is the distribution of the initial state x_0. For

any $t \geq 0$, at the t-th time step, the agent takes action $u_t \in \mathcal{U}$ at state $x_t \in \mathcal{X}$, which incurs a cost $c(x_t, u_t)$ and moves the environment into a new state $x_{t+1} \sim P(\cdot \,|\, x_t, u_t)$. A policy specifies how the action u_t is taken at a given state x_t. Specifically, in order to handle infinite state and action spaces, we focus on a parametrized policy class $\{\pi_\omega \colon \mathcal{X} \to \mathcal{P}(\mathcal{U}), \omega \in \Omega\}$, where ω is the parameter of policy π_ω, and the agent takes action $u \sim \pi_\omega(\cdot \,|\, x)$ at a given state $x \in \mathcal{X}$. The agent aims to find a policy that minimizes the infinite-horizon time-average cost, that is,

$$\underset{\omega \in \Omega}{\text{minimize}} \; J(\omega) = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T} c(x_t, u_t) \,\middle|\, x_0 \sim D_0, u_t \sim \pi_\omega(\cdot \,|\, x_t), \forall t \geq 0 \right]. \tag{2.1}$$

Moreover, for policy π_ω, we define the (advantage) action-value and state-value functions respectively as

$$\begin{aligned} Q_\omega(x, u) &= \sum_{t \geq 0} \mathbb{E}_\omega \big[c(x_t, u_t) \,|\, x_0 = x, u_0 = u \big] - J(\omega), \\ V_\omega(x) &= \mathbb{E}_{u \sim \pi_\omega(\cdot \,|\, x)} \big[Q_\omega(x, u) \big], \end{aligned} \tag{2.2}$$

where we use $\mathbb{E}_\omega$ to indicate that the state-action pairs $\{(x_t, u_t)\}_{t \geq 1}$ are obtained from policy π_ω.

Actor-critic is based on the idea of solving the minimization problem in (2.1) via first-order optimization, which uses an estimator of $\nabla_\omega J(\omega)$. In detail, by the policy gradient theorem [68, 6, 36], we have

$$\nabla_\omega J(\omega) = \mathbb{E}_{x \sim \rho_\omega, u \sim \pi_\omega(\cdot \,|\, x)} \big[\nabla_\omega \log \pi_\omega(u \,|\, x) \cdot Q_\omega(x, u) \big], \tag{2.3}$$

where $\rho_\omega \in \mathcal{P}(\mathcal{X})$ is the stationary distribution induced by π_ω. Based on (2.3), actor-critic [36] consists two steps: (i) a policy evaluation step that estimates the action-value function Q_ω (critic) via temporal difference learning [20], where Q_ω is estimated using a parametrized function class $\{Q^\theta \colon \theta \in \Theta\}$, and (ii) a policy improvement step that updates the parameter ω of policy π_ω (actor) using a stochastic version of the policy gradient in (2.3), where Q_ω is replaced by the corresponding estimator Q^θ.

As shown in [74], actor-critic can be cast as solving a bilevel optimization problem, which takes the form

$$\underset{\omega \in \Omega}{\text{minimize}} \quad \mathbb{E}_{x \sim \rho_\omega, u \sim \pi_\omega(\cdot \,|\, x)} \big[Q^\theta(x, u) \big], \tag{2.4}$$

$$\text{subject to} \quad (\theta, J) = \underset{\theta \in \Theta, J \in \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}_{x \sim \rho_\omega, u \sim \pi_\omega(\cdot \,|\, s)} \Big\{ \big[Q^\theta(x, u) + J - c(x, u) - (\mathcal{B}^\omega Q^\theta)(x, u) \big]^2 \Big\}, \tag{2.5}$$

where $\mathcal{B}^\omega$ is an operator that depends on π_ω. In this problem, the actor and critic correspond to the upper-level and lower-level variables, respectively. Under this framework, the policy update can be viewed as a stochastic gradient step for the upper-level problem in (2.4). The objective in (2.5) is usually the mean-squared Bellman error or mean-squared projected Bellman error [20]. Moreover, when $\mathcal{B}^\omega$ is the Bellman evaluation operator associated with π_ω and we solve the lower-level problem in (2.5) via stochastic semi-gradient descent, we obtain the TD(0) update for policy evaluation [65]. Similarly, when $\mathcal{B}^\omega$ is the projected Bellman evaluation operator associated with π_ω, solving the lower level problem naturally recovers the GTD2 and TDC algorithms for policy evaluation [8]. Therefore, the actor-critic algorithm is a first-order online algorithm for the bilevel optimization problem in (2.4) and (2.5). We remark that bilevel optimization contains a family of extremely challenging problems. Even when the objective functions are linear, bilevel programming is NP-hard [29]. In practice, various heuristic algorithms are applied to solve them approximately [62].

2.2 Linear Quadratic Regulator

As the simplest optimal control problem, linear quadratic regulator serves as a perfect baseline to examine the performance of reinforcement learning methods. Viewing LQR from the lens of an MDP, the state and action spaces are $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{U} = \mathbb{R}^k$, respectively. The state transition dynamics and cost function are specified by

$$x_{t+1} = Ax_t + Bu_t + \epsilon_t, \qquad c(x, u) = x^\top Q x + u^\top R u, \tag{2.6}$$

where $\epsilon_t \sim N(0, \Psi)$ is the random noise that is i.i.d. for each $t \geq 0$, and A, B, Q, R, Ψ are matrices of proper dimensions with $Q, R, \Psi \succ 0$. Moreover, we assume that the dimensions d and k are fixed throughout this paper. For the problem of minimizing the infinite-horizon time-average cost $\limsup_{T\to\infty} T^{-1} \sum_{t=0}^{T} \mathbb{E}[c(x_t, u_t)]$ with $x_0 \sim D_0$, it is known that the optimal actions are linear in the corresponding state [76, 3, 7]. Specifically, the optimal actions $\{u_t^*\}_{t\geq 0}$ satisfy $u_t^* = -K^* x_t$ for all $t \geq 0$, where $K^* \in \mathbb{R}^{k\times d}$ can be written as $K^* = (R + B^\top P^* B)^{-1} B^\top P^* A$, with P^* being the solution to the discrete algebraic Riccati equation

$$P^* = Q + A^\top P^* A + A^\top P^* B (R + B^\top P^* B)^{-1} B^\top P^* A. \tag{2.7}$$

In the optimal control literature, it is common to solve LQR by first estimating matrices A, B, Q, R and then solving the Riccati equation in (2.7) with these matrices replaced by their estimates. Such an approach is known as model-based as it requires estimating the model parameters and the performance of the planning step in (2.7) hinges on how well the true model is estimated. See, e.g, [21, 70] for theoretical guarantees of model-based methods.

In contrast, from a purely data-driven perspective, the framework of model-free reinforcement learning offers a general treatment for optimal control problems without the prior knowledge of the model. Thanks to its simple structure, LQR enables us to assess the performances of reinforcement learning algorithms from a theoretical perspective. Specifically, it is shown that policy iteration [12, 14, 45], adaptive dynamic programming [52], and policy gradient methods [26, 44, 70] are all able to obtain the optimal policy of LQR. Also see [54] for a thorough review of reinforcement learning methods in the setting of LQR.

3 Actor-Critic Algorithm for LQR

In this section, we establish the actor-critic algorithm for the LQR problem introduced in §2.2. Recall that the optimal policy of LQR is a linear function of the state. Throughout the rest of this paper, we focus on the family of linear-Gaussian policies

$$\big\{\pi_K(\cdot \,|\, x) = N(-Kx, \sigma^2 I_k), K \in \mathbb{R}^{k\times d}\big\}, \tag{3.1}$$

where $\sigma > 0$ is a fixed constant. That is, for any $t \geq 0$, at state x_t, we could write the action u_t by $u_t = -Kx_t + \sigma \cdot \eta_t$, where $\eta_t \sim N(0, I_k)$. We note that if $\sigma = 0$, then the optimal policy $u = -K^* x$ belongs to our policy class. Here, instead of focusing on deterministic policies, we adopt Gaussian policies to encourage exploration. For policy π_K, the corresponding time-average cost $J(K)$, state-value function V_K, and action-value function Q_K are specified as in (2.1) and (2.2), respectively.

In the following, we first establish the policy gradient and value functions for the ergodic LQR in §3.1. Then, in §3.2, we present the on-policy natural actor-critic algorithm, which is further extended to the off-policy setting in §B.

3.1 Policy Gradient Theorem for Ergodic LQR

For any policy π_K, by (2.6), the state dynamics are given by a linear dynamical system

$$x_{t+1} = (A - BK)x_t + \varepsilon_t, \qquad \text{where} \quad \varepsilon_t = \epsilon_t + \sigma \cdot B\eta_t \sim N(0, \Psi_\sigma). \tag{3.2}$$

Here we define $\Psi_\sigma := \Psi + \sigma^2 \cdot BB^\top$ in (3.2) to simplify the notation. It is known that, when $\rho(A - BK) < 1$, the Markov chain in (3.2) has stationary distribution $N(0, \Sigma_K)$, denoted by ν_K hereafter, where Σ_K is the unique positive definite solution to the Lyapunov equation

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

In the following proposition, we establish $J(K)$, the value functions, and the gradient $\nabla_K J(K)$.

Proposition 3.1. For any $K \in \mathbb{R}^{k\times d}$ such that $\rho(A - BK) < 1$, let P_K be the unique positive definite solution to the Bellman equation

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

In the setting of LQR, for policy π_K, both the state- and action-value functions are quadratic. Specifically, we have

$$V_K(x) = x^\top P_K x - \mathrm{tr}(P_K \Sigma_K), \tag{3.5}$$

$$Q_K(x,u) = x^\top \Theta_K^{11} x + x^\top \Theta_K^{12} u + u^\top \Theta_K^{21} x + u^\top \Theta_K^{22} u - \sigma^2 \cdot \mathrm{tr}(R + P_K B B^\top) - \mathrm{tr}(P_K \Sigma_K), \tag{3.6}$$

where Σ_K is specified in (3.3), and we define matrix Θ_K by

$$\Theta_K = \begin{pmatrix} \Theta_K^{11} & \Theta_K^{12} \\ \Theta_K^{21} & \Theta_K^{22} \end{pmatrix} = \begin{pmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{pmatrix}. \tag{3.7}$$

Moreover, the time-average cost $J(K)$ and its gradient are given by

$$J(K) = \mathrm{tr}\big[(Q + K^\top R K)\Sigma_K\big] + \sigma^2 \cdot \mathrm{tr}(R) = \mathrm{tr}(P_K \Psi_\sigma) + \sigma^2 \cdot \mathrm{tr}(R). \tag{3.8}$$

$$\nabla_K J(K) = 2\big[(R + B^\top P_K B)K - B^\top P_K A\big]\Sigma_K = 2E_K \Sigma_K, \tag{3.9}$$

where we define $E_K := (R + B^\top P_K B)K - B^\top P_K A$.

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

To see the connection between (3.9) and the policy gradient theorem in (2.3), note that by direct computation we have

$$\nabla_K \log \pi_K(u \,|\, x) = \nabla_K \big[-(2\sigma^2)^{-1} \cdot (u + Kx)^2\big] = -\sigma^{-2} \cdot (u + Kx)x^\top. \tag{3.10}$$

Thus, combining, (3.6), (3.10), and the fact that $u = -Kx + \sigma \cdot \eta$ under π_K, the right-hand side of (2.3) can be written as

$$\mathbb{E}_{x\sim\nu_K, u\sim\pi_K}\big[\nabla_K \log \pi_K(u \,|\, x) \cdot Q_K(x,u)\big] = -\sigma^{-2} \cdot \mathbb{E}_{x\sim\nu_K, \eta\sim N(0,I_k)}[\sigma \cdot \eta x^\top \cdot Q_K(x, -Kx + \sigma\cdot\eta)].$$

Recall that for $\eta \in N(0, I_k)$, Stein's identity [64], $\mathbb{E}[\eta \cdot f(\eta)] = \mathbb{E}[\nabla f(\eta)]$, holds for all differentiable function $f\colon \mathbb{R}^k \to \mathbb{R}$, which implies that

$$\begin{aligned}
\nabla_K J(K) &= -\mathbb{E}_{x\sim\nu_K, \eta\sim N(0,I_d)}[(\nabla_u Q_K)(x, -Kx + \sigma\cdot\eta)\cdot x^\top] \\
&= -2\mathbb{E}_{x\sim\nu_K, \eta\sim N(0,I_d)}\{[(R + B^\top P_K B)(-Kx + \sigma\cdot\eta) + B^\top P_K A x]x^\top\} \\
&= 2\big[(R + B^\top P_K B)K - B^\top P_K A\big]\Sigma_K = 2E_K\Sigma_K.
\end{aligned} \tag{3.11}$$

Thus, (3.9) is exactly the policy gradient theorem (2.3) in the setting of LQR. Moreover, it is worth noting that Proposition 3.1 also holds for $\sigma = 0$. Thus, setting $\sigma = 0$ in (3.11), we obtain

$$\nabla_K J(K) = -\mathbb{E}_{x\sim\nu_K}[(\nabla_u Q_K)(x, -Kx)\cdot x^\top] = \mathbb{E}_{x\sim\nu_K}\big[(\nabla_u Q_K)(x,u)\big|_{u=\pi_K(x)} \nabla_K \pi_K(x)\big],$$

where $\pi_K(x) = -Kx$. Thus we obtain the deterministic policy gradient theorem [60] for LQR. Although the optimal policy for LQR is deterministic, due to the lack of exploration, behaving according to a deterministic policy may lead to suboptimal solutions. Thus, we focus on the family of stochastic policies where a Gaussian noise $\sigma \cdot \eta$ is added to the action so as to promote exploration.

3.2 Natural Actor-Critic Algorithm

Natural policy gradient updates the variable along the steepest direction with respect to Fisher metric. For the Gaussian policies defined in (3.1), by (3.10), the Fisher's information of policy π_K, denoted by $\mathcal{I}(K)$, is given by

$$\begin{aligned}
[\mathcal{I}(K)]_{(i,j),(i',j')} &= \mathbb{E}_{x\sim\nu_K, a\sim\pi_K}\big[\nabla_{K_{ij}} \log \pi_K(u \,|\, x)\cdot \nabla_{K_{i'j'}} \log \pi_K(u \,|\, x)\big] \\
&= \sigma^{-2}\cdot \mathbb{E}_{x\sim\nu_K, \eta\sim N(0,I_k)}[\eta_i x_j \cdot \eta_{i'} x_{j'}] = \sigma^{-2}\cdot \mathbb{1}\{i = i'\}\cdot [\Sigma_K]_{jj'},
\end{aligned} \tag{3.12}$$

where $i, i' \in [k]$, $j, j' \in [d]$, K_{ij} and $K_{i'j'}$ are the (i,j)- and (i',j')-th entries of K, respectively, and $[\Sigma_K]_{jj'}$ is the (j,j')-th entry of Σ_K. Thus, in view of (3.9) in Proposition 3.1 and (3.12), natural policy gradient algorithm updates the policy parameter in the direction of

$$[\mathcal{I}(K)]^{-1}\nabla_K J(K) = \nabla_K J(K)\Sigma_K^{-1} = 2E_K.$$

By (3.7), we can write E_K as $\Theta_K^{22}K - \Theta_K^{21}$, where Θ_K is the coefficient matrix of the quadratic component of Q_K. Such a connection lays the foundation of the natural actor-critic algorithm. Specifically, in each iteration of the algorithm, the actor updates the policy via $K - \gamma \cdot (\widehat{\Theta}^{22}K - \widehat{\Theta}^{21})$, where γ is the stepsize and $\widehat{\Theta}$ is an estimator of Θ_K returned by any policy evaluation algorithm. We present such a general natural actor-critic method in Algorithm 1 §A, under the assumption that we are given a stable policy K_0 for initialization. Such an assumption is standard in literatures on model-free methods for LQR [22, 26, 44].

To obtain an online actor-critic algorithm, in the sequel, we propose an online policy evaluation algorithm for ergodic LQR based on temporal difference learning. Let π_K be the policy of interest. For notational simplicity, for any state-action pair $(x, u) \in \mathbb{R}^{d+k}$, we define the feature function

$$\phi(x, u) = \mathrm{svec}\left[\begin{pmatrix} x \\ u \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}^\top\right], \tag{3.13}$$

and denote $\mathrm{svec}(\Theta_K)$ by θ_K^*. Using this notation, the quadratic component in Q_K can be written as $\phi(x, u)^\top \theta_K^*$, and the Bellman equation for Q_K becomes

$$\langle \phi(x, u), \theta_K^* \rangle = c(x, u) - J(K) + \big\langle \mathbb{E}[\phi(x', u') \,|\, x, u], \theta_K^* \big\rangle, \ \forall (x, u) \in \mathbb{R}^{d+k}. \tag{3.14}$$

In order to further simplify the notation, hereafter, we define $\vartheta_K^* = (J(K), {\theta_K^*}^\top)^\top$, denote by $\mathbb{E}_{(x,u)}$ the expectation with respect to $x \sim \nu_K$ and $u \sim \pi_K(\cdot \,|\, x)$, and let (x', u') be the state-action pair subsequent to (x, u).

Furthermore, to estimate $J(K)$ and θ_K^* in (3.14) simultaneously, we define

$$\Xi_K = \mathbb{E}_{(x,u)}\big\{\phi(x, u)\big[\phi(x, u) - \phi(x', u')\big]^\top\big\}, \qquad b_K = \mathbb{E}_{(x,u)}\big[c(x, u)\phi(x, u)\big], \tag{3.15}$$

Notice that $J(K) = \mathbb{E}_{(x,u)}[c(x, u)]$. By direct computation, it can be shown that ϑ_K^* satisfies the following linear equation

$$\begin{pmatrix} 1 & 0 \\ \mathbb{E}_{(x,u)}[\phi(x, u)] & \Xi_K \end{pmatrix} \begin{pmatrix} \vartheta^1 \\ \vartheta^2 \end{pmatrix} = \begin{pmatrix} J(K) \\ b_K \end{pmatrix}, \tag{3.16}$$

whose solution is unique if and only if Ξ_K in (3.15) is invertible. The following lemma shows that, when π_K is a stable policy, Ξ_K is indeed invertible.

Lemma 3.2. When π_K is stable in the sense that $\rho(A - BK) < 1$, Ξ_K defined in (3.15) is invertible and thus ϑ_K^* is the unique solution to the linear equation (3.16). Furthermore, the minimum singular value of the matrix in the left-hand side of (3.16) is lower bounded by a constant $\kappa_K^* > 0$, where κ_K^* only depends on $\rho(A - BK)$, σ, and $\sigma_{\min}(\Psi)$.

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

By this lemma, when $\rho(A - BK) < 1$, policy evaluation for π_K can be reduced to finding the unique solution to a linear equation. Instead of solving the equation directly, it is equivalent to minimize the least-squares loss:

$$\underset{\vartheta}{\mathrm{minimize}}\Big\{[\vartheta^1 - J(K)]^2 + \big\|\vartheta^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)] + \Xi_K\vartheta^2 - b_K\big\|_2^2\Big\}, \tag{3.17}$$

where $\vartheta^1 \in \mathbb{R}$ and ϑ^2 has the same shape as $\mathrm{svec}(\Theta_K)$, which are the two components of ϑ. It is clear that the global minimizer of (3.17) is ϑ_K^*. Note that we have Fenchel's duality $x^2 = \sup_y\{2x \cdot y - y^2\}$. By this relation, we further write (3.17) as a minimax optimization problem

$$\begin{aligned} \min_{\vartheta \in \mathcal{X}_\Theta} \max_{\omega \in \mathcal{X}_\Omega} F(\vartheta, \omega) = {} & [\vartheta^1 - J(K)] \cdot \omega^1 \\ & + \big\langle \vartheta^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)] + \Xi_K\vartheta^2 - b_K, \omega^2 \big\rangle - 1/2 \cdot \|\omega\|_2^2, \end{aligned} \tag{3.18}$$

where the dual variable $\omega = (\omega^1, \omega^2)$ has the same shape as ϑ. Here we restrict the primal and dual variables to compact sets $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ for algorithmic stability, which will be specified in the next section. Note that the objective in (3.18) can be estimated unbiasedly using two consecutive state-action pairs (x, u) and (x', u'). Solving the minimax optimization in (3.18) using stochastic

gradient method, we obtain the gradient-based temporal difference (GTD) algorithm for policy evaluation [66, 67]. See Algorithm 2 for details. More specifically, by direct computation, we have

$$\nabla_{\vartheta^1} F(\theta,\omega) = \omega^1 + \big\langle \mathbb{E}_{(x,u)}(\phi), \omega^2 \big\rangle, \qquad \nabla_{\vartheta^2} F(\theta,\omega) = \mathbb{E}_{(x,u)}\big[(\phi-\phi')\cdot\phi^\top\omega^2\big], \tag{3.19}$$

$$\nabla_{\omega^1} F(\theta,\omega) = \vartheta^1 - J(K) - \omega^1, \qquad \nabla_{\omega^2} F(\theta,\omega) = \theta^1\cdot\mathbb{E}_{(x,u)}(\phi) + \Xi_K\vartheta^2 - b_K - \omega^2, \tag{3.20}$$

where we denote $\phi(x,u)$ and $\phi(x',u')$ by ϕ and ϕ', respectively. In the GTD algorithm, we update ϑ and ω in gradient directions where the gradients in (3.19) and (3.20) are replaced by their sample estimates. After T iterations of the algorithm, we output the averaged update $\widehat{\vartheta}^2 = (\sum_{t=1}^T \alpha_t\cdot\vartheta_t^2)/(\sum_{t=1}^T\alpha_t)$ and use $\widehat{\Theta} = \mathrm{smat}(\widehat{\vartheta}^2)$ to estimate Θ_K in (3.7), which is further used in Algorithm 1 to update the current policy π_K. Therefore, we obtain the online natural actor-critic algorithm [9] for ergodic LQR.

Meanwhile, using the perspective of bilevel optimization, similar to (2.4) and (2.5), our actor-critic algorithm can be viewed as a first-order online algorithm for

$$\begin{aligned} &\underset{K\in\mathbb{R}^{k\times d}}{\text{minimize}} \quad \mathbb{E}_{x\sim\nu_K,u\sim\pi_K}\big[\langle\phi(x,u),\vartheta^2\rangle\big],\\ &\text{subject to} \quad (\vartheta,\omega) = \underset{\theta\in\mathcal{X}_\Theta}{\operatorname{argmin}}\ \underset{\omega\in\mathcal{X}_\Omega}{\operatorname{argmax}}\ F(\vartheta,\omega), \end{aligned} \tag{3.21}$$

where $F(\vartheta,\omega)$ is defined in (3.18) and depends on π_K. In our algorithm, we solve the upper-level problem via natural gradient descent and solve the lower-level saddle point optimization problem using stochastic gradient updates.

Furthermore, we emphasize that our method defined by Algorithms 1 and 2 is online in the sense that each update only requires a single transition. More specifically, let $\{(x_n,u_n,c_n)\}_{n\geq 0}$ be the sequence of transitions experienced by the agent. Combining Algorithms 1 and 2 and neglecting the projections, we can write the updating rule as

$$\begin{aligned} K_{n+1} &= K_n - \overline{\gamma}_n\cdot\big\{[\mathrm{smat}(\vartheta_n^2)]^{22}K_n - [\mathrm{smat}(\vartheta_n^2)]^{21}\big\},\\ \vartheta_{n+1} &= \vartheta_n - \overline{\alpha}_n\cdot g_\vartheta(x_n,u_n,c_n,x_{n+1},u_{n+1}),\\ \omega_{n+1} &= \omega_n + \overline{\alpha}_n\cdot g_\omega(x_n,u_n,c_n,x_{n+1},u_{n+1}), \end{aligned} \tag{3.22}$$

where g_ϑ and g_ω are the update directions of ϑ and ω whose definitions are clear from Algorithm 2, $\{\overline{\gamma}_n\}$ and $\{\overline{\alpha}_n\}$ are the stepsizes. Moreover, there exists a monotone increasing sequence $\{N_t\}_{t\geq 0}$ such that $\overline{\gamma}_n = \gamma$ if $n = N_t$ for some t and $\overline{\gamma}_n = 0$ otherwise. Such a choice of the stepsizes reflects the intuition that, although both the actor and the critic are updated simultaneously, the critic should be updated at a faster pace. From the same viewpoint, classical actor-critic algorithms [36, 9, 28] establish convergence results under the assumption that

$$\sum_{n\geq 0}\overline{\gamma}_n = \sum_{n\geq 0}\overline{\alpha}_n = \infty, \qquad \sum_{n\geq 0}(\overline{\gamma}_n^2 + \overline{\alpha}_n^2) < \infty, \qquad \lim_{n\to\infty}\overline{\gamma}_n/\overline{\alpha}_n = 0.$$

The condition that $\overline{\gamma}_n/\overline{\alpha}_n = 0$ ensures that the critic updates at a faster timescale, which enables the asymptotic analysis utilizing two-timescale stochastic approximation [10, 37]. However, such an approach uses two ordinary differential equations (ODE) to approximate the updates in (3.22) and thus only offers asymptotic convergence results. In contrast, as shown in §4, our choice of the stepsizes yields nonasymptotic convergence results which shows that natural actor gradient converges in linear rate to the global optimum.

In addition, we note that in Algorithm 2 we assume that the initial state x_0 is sampled from the stationary distribution ν_K. Such an assumption is made only to simplify theoretical analysis. In practice, we could start the algorithm after sampling a sufficient number of transitions so that the Markov chain induced by π_K approximately mixes. Moreover, as shown in [69], when π_K is a stable policy such that $\rho(A-BK)<1$, the Markov chain induced by π_K is geometrically β-mixing and thus mixes rapidly.

Finally, we remark that the minimax formulation of the policy evaluation problem is first proposed in [40], which studies the sample complexity of the GTD algorithm for discounted MDPs with i.i.d. data. Using the same formulation, [72] establishes finite sample bounds with data generated from a Markov process. Our optimization problem in (3.18) can be viewed as the extension of their minimax formulation to the ergodic setting. Besides, our GTD algorithm can be applied to ergodic MDPs in general with dependent data, which might be of independent interest.

4 Theoretical Results

In this section, we establish the global convergence of the natural actor-critic algorithm. To this end, we first focus on the problem of policy evaluation by assessing the finite sample performance of the on-policy GTD algorithm.

Note that only $\widehat{\Theta}$ returned by the GTD algorithm is utilized in the natural actor-critic algorithm for policy update. Thus, in the policy evaluation problem for a linear policy π_K, we only need to study the estimation error $\|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2$, which characterizes the closeness between the direction of policy update in Algorithm 1 and the true natural policy gradient.

Furthermore, recall that we restrict the primal and dual variables respectively to compact sets $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ for algorithmic stability. We make the following assumption on $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$.

Assumption 4.1. Let π_{K_0} be the initial policy in Algorithm 1. We assume that π_{K_0} is a stable policy such that $\rho(A - BK_0) < 1$. Consider the policy evaluation problem for π_K. We assume that $J(K) \le J(K_0)$. Moreover, let $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ in (3.18) be defined as

$$\mathcal{X}_\Theta = \big\{\vartheta\colon 0 \le \vartheta^1 \le J(K_0), \|\vartheta^2\|_2 \le \widetilde{R}_\Theta\big\}, \tag{4.1}$$

$$\mathcal{X}_\Omega = \big\{\omega\colon |\omega^1| \le J(K_0), \|\omega^2\|_2 \le (1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Omega\big\}. \tag{4.2}$$

Here, $\widetilde{R}_\Theta$ and $\widetilde{R}_\Omega$ are two parameters that do not depend on K. Specifically, we have

$$\widetilde{R}_\Theta = \|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}} + \sqrt{d}/\sigma_{\min}(\Psi) \cdot (\|A\|_{\mathrm{fro}}^2 + \|B\|_{\mathrm{fro}}^2) \cdot J(K_0), \tag{4.3}$$

$$\widetilde{R}_\Omega = C \cdot \widetilde{R}_\Theta \cdot \sigma_{\min}^{-2}(Q) \cdot [J(K_0)]^2, \tag{4.4}$$

where $C > 0$ is a constant.

The assumption that we have access to a stable policy K_0 for initialization is commonly made in literatures on model-free methods for LQR [22, 26, 44]. Besides, $\rho(A - BK_0) < 1$ implies that $J(K_0)$ is finite. Here we assume $J(K) \le J(K_0)$ for simplicity. Even if $J(K) > J(K_0)$, we can replace $J(K_0)$ in (4.1) – (4.4) by $J(K)$ and the theory of policy evaluation still holds. Moreover, as we will show in Theorem 4.3, the actor-critic algorithm creates a sequence policies whose objective values decreases monotonically. Thus, here we assume $J(K) \le J(K_0)$ without loss of generality.

Furthermore, as shown in the proof, the construction of $\widetilde{R}_\Theta$ and $\widetilde{R}_\Omega$ ensures that $(\vartheta_K^*, 0)$ is the saddle-point of the minimax optimization in (3.18). In other words, the solution to (3.18) is the same as the unconstrained problem $\min_\vartheta \max_\omega F(\vartheta, \omega)$. When replacing the population problem by a sample-based optimization problem, restrictions on the primal and dual variables ensures that the iterates of the GTD algorithm remains bounded. Thus, setting $\widetilde{R}_\Theta$ and $\widetilde{R}_\Omega$ essentially guarantees that restricting (ϑ, ω) to $\mathcal{X}_\Theta \times \mathcal{X}_\Omega$ incurs no "bias" in the optimization problem.

We present the theoretical result for the online GTD algorithm as follows.

Theorem 4.2 (Policy evaluation)**.** Let $\widehat{\vartheta}^1$ and $\widehat{\Theta}$ be the output of Algorithm 2 based on T iterations. We set the stepsize to be $\alpha_t = \alpha/\sqrt{t}$ with $\alpha > 0$ being a constant. Under Assumption 4.1, for any $\rho \in (\rho(A - BK), 1)$, when the number of iterations T is sufficiently large, with probability at least $1 - T^{-4}$, we have

$$\|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2 \le \frac{\Upsilon\big[\widetilde{R}_\Theta, \widetilde{R}_\Omega, J(K_0), \|K\|_{\mathrm{fro}}, \sigma_{\min}^{-1}(Q)\big]}{{\kappa_K^*}^2 \cdot (1 - \rho)} \cdot \frac{\log^6 T}{\sqrt{T}}, \tag{4.5}$$

where $\Upsilon\big[\widetilde{R}_\Theta, \widetilde{R}_\Omega, J(K_0), \|K\|_{\mathrm{fro}}, \sigma_{\min}^{-1}(Q)\big]$ is a polynomial of $\widetilde{R}_\Omega$, $\widetilde{R}_\Omega$, $J(K_0)$, $\|K\|_{\mathrm{fro}}$, and $1/\sigma_{\min}(Q)$.

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

This theorem establishes the statistical rate of convergence for the on-policy GTD algorithm. Specifically, if we regard $\Upsilon[\widetilde{R}_\Theta, \widetilde{R}_\Omega, J(K_0), \|K\|_{\mathrm{fro}}, \sigma_{\min}^{-1}(Q)]$, ρ, and κ_K^* as constant, (4.5) implies that the estimation error is of order $\log^6 T/\sqrt{T}$. Thus, ignoring the logarithmic term, we conclude that the GTD algorithm converges in the sublinear rate $\mathcal{O}(1/\sqrt{T})$, which is optimal for convex-concave

stochastic optimization [49] and is also identical to the rate of convergence of the GTD algorithm in the discounted setting with bounded data [40, 72]. Note that we focus on the ergodic case and the feature mapping $\phi(x, u)$ defined in (3.13) is unbounded. We believe this theorem might be of independent interest. Furthermore, $1/\kappa_K^*$ is approximately the condition number of the linear equation of (3.16), which reflects the fundamental difficulty of estimating Θ_K. Specifically, when κ_K^* is close to zero, the matrix on the left-hand side of (3.16) is close to a singular matrix. In this case, estimating Θ_K can be viewed as solving an ill-conditioned regression problem and thus huge sample size is required for consistent estimation. Finally, $1/[1-\rho(A-BK)]$ also reflects the intrinsic hardness of estimating Θ_K. Specifically, for any $\rho \in (\rho(A-BK), 1)$, the Markov chain induced by π_K is β-mixing where the k-th mixing coefficients is bounded by $C \cdot \rho^k$ for some constant $C > 0$ [69]. Thus, when ρ is close to one, this Markov chain becomes more dependent, which makes the estimation problem more difficult.

Equipped with the finite sample error of the policy evaluation algorithm, now we are ready to present the global convergence of the actor-critic algorithm. For ease of presentation, we assume that Q, R, A, B, Ψ are all constant matrices.

Theorem 4.3 (Global convergence of actor-critic)**.** Let the initial policy K_0 be stable. We set the stepsize $\gamma = [\|R\| + \sigma_{\min}^{-1}(\Psi) \cdot \|B\|^2 \cdot J(K_0)]$ in Algorithm 1 and perform N actor updates in the actor-critic algorithm. Let $\{K_t\}_{0 \le t \le N}$ be the sequence of policy parameters generated by the algorithm. For any sufficiently small $\epsilon > 0$, we set $N > C \cdot \|\Sigma_{K^*}\|/\gamma \cdot \log\{2[J(K_0) - J(K^*)]/\epsilon\}$ for some constant $C > 0$. Moreover, for any $t \in \{0, 1, \ldots, N\}$, in the t-th iteration, we set the number T_t of GTD updates in Algorithm 2 to be

$$T_t \ge \Upsilon\big[\|K_t\|, J(K_0)\big] \cdot {\kappa_{K_t}^*}^{-5} \cdot [1-\rho(A-BK_t)]^{-5/2} \cdot \epsilon^{-5},$$

where $\Upsilon[\|K_t\|, J(K_0)]$ is a polynomial of $\|K_t\|$ and $J(K_0)$. Then with probability at least $1-\epsilon^{10}$, we have $J(K_N) - J(K^*) \le \epsilon$.

Proof Sketch. The proof of this Theorem is based on combining the convergence of the natural policy gradient and the finite sample analysis of the GTD algorithm established in Theorem 4.2. Specifically, for each K_t, we define $K'_{t+1} = K_t - \eta \cdot E_{K_t}$, which is the one-step natural policy gradient update starting from K_t. Similar to [26], for ergodic LQR, it can be shown that

$$J(K'_{t+1}) - J(K^*) \le \big[1 - C_1 \cdot \gamma \cdot \|\Sigma_{K^*}\|^{-1}\big] \cdot \big[J(K_t) - J(K^*)\big] \tag{4.6}$$

for some constant $C_1 > 0$. In addition, for policy π_{K_t}, when the number of GTD iteration T_t is sufficiently large, K_{t+1} is close to K'_{t+1}, which further implies that $|J(K'_{t+1}) - J(K_{t+1})|$ is small. Thus, combining this and (4.6), we obtain the linear convergence of the actor-critic algorithm. See §C.2 for a detailed proof. □

This theorem shows that natural actor-critic algorithm combined with GTD converges linearly to the optimal policy of LQR. Furthermore, the number of policy updates in this theorem matches those obtained by natural policy gradient algorithm [26, 44]. To the best of our knowledge, this result seems to be the first nonasymptotic convergence result for actor-critic algorithms with function approximation, whose existing theory are mostly asymptotic and based on ODE approximation. Furthermore, from the viewpoint of bilevel optimization, Theorem 4.3 offers theoretical guarantees for the actor-critic algorithm as a first-order online method for the bilevel program defined in (3.21), which serves a first attempt of understanding bilevel optimization with possibly nonconvex subproblems.

Furthermore, although we only consider the problem of LQR and analyze the natural actor-critic with GTD for policy evaluation, our theoretical framework can be applied to general reinforcement learning problems with other policy optimization methods for the actor (e.g. vanilla policy gradient [68], trust-region policy optimization [57] and proximal policy optimization [58]) and other policy evaluation methods for the critic such as TD(0) [65], least-square temporal-difference (LSTD) [13], and Retrace [47]. In particular, suppose the critic adopts compatible features [68] for policy evaluation using nonconvex optimization techniques, it can be shown that vanilla policy gradient converges to a local minimum of the expected total return with a sublinear rate [75]. Moreover, by leveraging the geometry of the expected total return as a functional of the policy, recently, [1, 39, 71, 59] prove that natural policy gradient [34], TRPO and PPO are all able to find the globally optimal policy. Using similar approaches, we can establish the convergence and global optimality of actor-critic methods.

References

[1] Agarwal, A., Kakade, S. M., Lee, J. D. and Mahajan, G. (2019). Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*.

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

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

[4] Bard, J. F. (2013). *Practical bilevel optimization: algorithms and applications*, vol. 30. Springer Science & Business Media.

[5] Basar, T. and Olsder, G. J. (1999). *Dynamic noncooperative game theory*, vol. 23. SIAM.

[6] Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, **15** 319–350.

[7] Bertsekas, D. P. (2012). *Dynamic programming and optimal control, Vol. II, 4th Edition*. Athena scientific.

[8] Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*.

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

[10] Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, **29** 291–294.

[11] Borkar, V. S. and Konda, V. R. (1997). The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, **22** 525–543.

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

[13] Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, **22** 33–57.

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

[15] Cai, Q., Hong, M., Chen, Y. and Wang, Z. (2019). On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*.

[16] Chen, X., Wang, J. and Ge, H. (2018). Training generative adversarial networks via primal-dual subgradient methods: A lagrangian perspective on gan. *arXiv preprint arXiv:1802.01765*.

[17] Dai, B., He, N., Pan, Y., Boots, B. and Song, L. (2017). Learning from conditional distributions via dual embeddings. In *International Conference on Artificial Intelligence and Statistics*.

[18] Dai, B., Shaw, A., He, N., Li, L. and Song, L. (2018). Boosting the actor with dual critic. *International Conference on Learning Representations*.

[19] Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J. and Song, L. (2018). Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*.

[20] Dann, C., Neumann, G. and Peters, J. (2014). Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, **15** 809–883.

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

[22] Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*.

[23] Dean, S., Tu, S., Matni, N. and Recht, B. (2018). Safely learning to control the constrained linear quadratic regulator. *arXiv preprint arXiv:1809.10121*.

[24] Dempe, S. (2002). *Foundations of bilevel programming*. Springer Science & Business Media.

[25] Du, S. S. and Hu, W. (2018). Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *arXiv preprint arXiv:1802.01504*.

[26] Fazel, M., Ge, R., Kakade, S. and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*.

[27] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

[28] Grondman, I., Busoniu, L., Lopes, G. A. and Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42** 1291–1307.

[29] Hansen, P., Jaumard, B. and Savard, G. (1992). New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, **13** 1194–1217.

[30] Hardt, M., Ma, T. and Recht, B. (2018). Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, **19** 1025–1068.

[31] Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*.

[32] Horn, R. A., Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis*. Cambridge university press.

[33] Islam, R., Henderson, P., Gomrokchi, M. and Precup, D. (2017). Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv preprint arXiv:1708.04133*.

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

[35] Kirk, D. E. (1970). *Optimal control theory: an introduction*. Springer.

[36] Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

[37] Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer.

[38] Lin, Q., Liu, M., Rafique, H. and Yang, T. (2018). Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *arXiv:1810.10207*.

[39] Liu, B., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*.

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

[41] Lu, S., Singh, R., Chen, X., Chen, Y. and Hong, M. (2019). Understand the dynamics of GANs via primal-dual optimization.
https://openreview.net/forum?id=rylIy3R9K7

[42] Luo, Z.-Q., Pang, J.-S. and Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge University Press.

[43] Magnus, J. R. (1978). The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, **32** 201–210.

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

[45] Meyn, S. P. (1997). The policy iteration algorithm for average reward markov decision processes with general state space. *IEEE Transactions on Automatic Control*, **42** 1663–1680.

[46] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*.

[47] Munos, R., Stepleton, T., Harutyunyan, A. and Bellemare, M. (2016). Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*.

[48] Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica: Journal of the Econometric Society* 575–595.

[49] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, **19** 1574–1609.

[50] Pfau, D. and Vinyals, O. (2016). Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*.

[51] Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, **3** 643–653.

[52] Powell, W. B. and Ma, J. (2011). A review of stochastic algorithms with continuous value function approximation and some new approximate policy iteration algorithms for multidimensional continuous applications. *Journal of Control Theory and Applications*, **9** 336–352.

[53] Rafique, H., Liu, M., Lin, Q. and Yang, T. (2018). Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv:1810.02060*.

[54] Recht, B. (2018). A tour of reinforcement learning: The view from continuous control. *arXiv preprint arXiv:1806.09460*.

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

[56] Sanjabi, M., Razaviyayn, M. and Lee, J. D. (2018). Solving non-convex non-concave min-max games under polyak-lojasiewicz condition. *arXiv preprint arXiv:1812.02878*.

[57] Schulman, J., Levine, S., Abbeel, P., Jordan, M. and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*.

[58] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[59] Shani, L., Efroni, Y. and Mannor, S. (2019). Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *arXiv preprint arXiv:1909.02769*.

[60] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. and Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International Conference on Machine Learning*.

[61] Simchowitz, M., Mania, H., Tu, S., Jordan, M. I. and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*.

[62] Sinha, A., Malo, P. and Deb, K. (2018). A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, **22** 276–295.

[63] Sinha, A., Namkoong, H. and Duchi, J. (2017). Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

[64] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* 1135–1151.

[65] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, **3** 9–44.

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

[67] Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

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

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

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

[71] Wang, L., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

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

[73] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8** 229–256.

[74] Yang, Z., Fu, Z., Zhang, K. and Wang, Z. (2018). Convergent reinforcement learning with function approximation: A bilevel optimization perspective. *Manuscript*.

[75] Zhang, K., Koppel, A., Zhu, H. and Başar, T. (2019). Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*.

[76] Zhou, K., Doyle, J. C. and Glover, K. (1996). *Robust and optimal control*, vol. 40. Prentice Hall.

A Algorithms

In this section, we present the details of the actor-critic algorithm.

Algorithm 1 Natural Actor-Critic Algorithm for Linear Quadratic Regulator

Input: Initial policy π_{K_0} such that $\rho(A - BK_0) < 1$, stepsizes γ for policy update, and a policy evaluation algorithm.
Initialization: Set the current policy π_K by letting $K \leftarrow K_0$.
while updating current policy **do**
  Critic step. Estimate Θ_K in (3.7) via a policy evaluation algorithm, e.g., the on-policy GTD algorithm (Algorithm 2), which returns an estimator $\widehat{\Theta}$ of Θ_K.
  Actor step. Update the policy parameter by $K \leftarrow K - \gamma \cdot (\widehat{\Theta}^{22} K - \widehat{\Theta}^{21})$.
end while
Output: The final policy π_K, matrix $\widehat{\Theta}$ that estimates Θ_K, and $\widehat{J}$ that approximates $J(K)$.

Algorithm 2 On-Policy Gradient-Based Temporal-Difference Algorithm for Policy Evaluation

Input: Policy π_K, number of iterations T, and stepsizes $\{\alpha_t\}_{t\in[T]}$.
Output: Estimator $\widehat{\Theta}$ of Θ_K in (3.7).
Initialize the primal and dual variables by $\vartheta_0 \in \mathcal{X}_\Theta$ and $\omega_0 \in \mathcal{X}_\Omega$, respectively.
Sample the initial state $x_0 \in \mathbb{R}^d$ from the stationary distribution ρ_K. Take action $u_0 \sim \pi_K(\cdot \,|\, x_0)$ and obtain the reward c_0 and the next state x_1.
for $t = 1, 2, \ldots, T$ **do**
  Take action u_t according to policy π_K, observe the reward c_t and the next state x_{t+1}.
  Compute the TD-error $\delta_t = \vartheta_{t-1}^1 - c_{t-1} + [\phi(x_{t-1}, u_{t-1}) - \phi(x_t, u_t)]^\top \vartheta_{t-1}^2$.
  Update ϑ^1 by $\vartheta_t^1 = \vartheta_{t-1}^1 - \alpha_t \cdot [\omega_{t-1}^1 + \phi(x_{t-1}, u_{t-1})^\top \omega_{t-1}^2]$.
  Update ϑ^2 by $\vartheta_t^2 = \vartheta_{t-1}^2 - \alpha_t \cdot [\phi(x_{t-1}, u_{t-1}) - \phi(x_t, u_t)] \cdot \phi(x_{t-1}, u_{t-1})^\top \omega_{t-1}^2$.
  Update ω^1 by $\omega_t^1 = (1 - \alpha_t) \cdot \omega_t^1 + \alpha_t \cdot (\vartheta_{t-1}^1 - c_{t-1})$.
  Update ω^2 by $\omega_t^2 = (1 - \alpha_t) \cdot \omega_t^2 + \alpha_t \cdot \delta_t \cdot \phi(x_{t-1})$.
  Project ϑ_t and ω_t to $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$, respectively.
end for
Define $\widehat{\vartheta} = (\widehat{\vartheta}^1, \widehat{\vartheta}^2) = (\sum_{t=1}^T \alpha_t \cdot \vartheta_t)/(\sum_{t=1}^T \alpha_t)$ and $\widehat{\omega} = (\sum_{t=1}^T \alpha_t \cdot \omega_t)/(\sum_{t=1}^T \alpha_t)$.
Return $\widehat{\vartheta}^1$ and $\widehat{\Theta} = \mathrm{smat}(\widehat{\vartheta}^2)$ as the estimators of $J(K)$ and Θ_K, respectively.

B Extension to the Off-Policy Setting

Recall that in our natural actor-critic algorithm, the critic can apply any policy evaluation algorithm to estimate $\widehat{\Theta}_K$. When using an off-policy method, we obtain an off-policy actor-critic algorithm. In this section, we extend Algorithm 2 to the off-policy setting using importance sampling. Specifically, let π_b be the behavior policy and suppose it induces a stationary distribution ν_b over the state space $\mathbb{R}^d$. Moreover, let π_K be the policy of interest and let $\tau_K(x, u) = \pi_K(u \,|\, x)/\pi_b(u \,|\, x)$ be the importance sampling ratio. Then, the Bellman equation in (3.14) can be written as

$$\langle \phi(x, u), \theta_K^* \rangle = c(x, u) - J(K) + \big\langle \mathbb{E}[\phi(x', u') \cdot \tau_K(x', u') \,|\, x, u], \theta_K^* \big\rangle, \tag{B.1}$$

where $(x, u) \in \mathbb{R}^{d+k}$, x' is the next state given (x, u), and $u' \sim \pi_b(\cdot \,|\, x)$. In the following, we denote by $\mathbb{E}_{(x,u)}$ the expectation with respect to $x \sim \nu_b$ and $u \sim \pi_b(\cdot \,|\, x)$. Similar to Ξ_K and b_K defined in (3.15), for the off-policy setting, we define

$$\overline{\Xi}_K = \mathbb{E}_{(x,u)}\big\{\phi(x, u)\big[\phi(x, u) - \tau_K(x', u') \cdot \phi(x', u')\big]^\top\big\}, \qquad \overline{b}_K = \mathbb{E}_{(x,u)}\big[c(x, u)\phi(x, u)\big],$$
$$\overline{h}_K = \mathbb{E}_{(x,u)}[\phi(x, u) - \tau_K(x', u') \cdot \phi(x', u')], \qquad \overline{g}_K = \mathbb{E}_{(x,u)}[\phi(x, u)], \qquad \bar{a}_K = \mathbb{E}_{(x,u)}[c(x, u)].$$

Based on (B.1) and direct computation, it can be shown that $\vartheta_K^* = (J(K), \mathrm{svec}(\Theta_K)^\top)^\top$ is the solution to linear equation

$$\Lambda_K \vartheta = \lambda_K, \qquad \Lambda_K = \begin{pmatrix} 1 & \overline{h}_K^\top \\ \overline{g}_K & \overline{\Xi}_K \end{pmatrix}, \qquad \vartheta = \begin{pmatrix} \vartheta^1 \\ \vartheta^2 \end{pmatrix}, \qquad \lambda_K = \begin{pmatrix} \bar{a}_K \\ \overline{b}_K \end{pmatrix}. \tag{B.2}$$

Similar to the derivations in §3.2, we propose to estimate ϑ_K^* by solving a minimax optimization problem:

$$\min_{\vartheta\in\mathcal{X}_\Theta}\max_{\omega\in\mathcal{X}_\Omega}\overline{F}(\vartheta,\omega)=\omega^\top[\Lambda_K\vartheta-\lambda_K]-1/2\cdot\|\omega\|_2^2. \tag{B.3}$$

Here, similar to Assumption (4.1), we let $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ be Euclidean balls given by $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ in (3.18) be defined as $\mathcal{X}_\Theta=\{\vartheta\colon\|\vartheta^1\|_2\le\widetilde{R}_\Theta\}$ and $\mathcal{X}_\Omega=\{\omega\colon\|\omega\|_2\le\widetilde{R}_\Omega\}$, where $\widetilde{R}_\Theta$ and $\widetilde{R}_\Omega$ are chosen appropriately. Notice that both $\overline{F}(\vartheta,\omega)$ and its gradient can be estimated unbiasedly using transitions sampled from the behavior policy. Solving (B.3) using stochastic gradient method, we obtain the off-policy GTD algorithm for the ergodic setting. See Algorithm 3 for the details. Combining this policy evaluation method with Algorithm 1, we establish the off-policy on-line natural actor-critic algorithm.

Algorithm 3 Off-Policy Gradient-Based Temporal-Difference Algorithm for Policy Evaluation

Input: Policy π_K, number of iterations T, and stepsizes $\{\alpha_t\}_{t\in[T]}$, the behavior policy π_b and its stationary distribution ν_b.

Output: Estimators $\widehat{J}$ and $\widehat{\Theta}$ of $J(K)$ in (3.8) and Θ_K in (3.7), respectively.

Initialize the primal and dual variables by $\vartheta_0\in\mathcal{X}_\Theta$ and $\omega_0\in\mathcal{X}_\Omega$, respectively.

Sample the initial state $x_0\in\mathbb{R}^d$ from the stationary distribution ν_b. Take action $u_0\sim\pi_b(\cdot\,|\,x_0)$ and obtain the reward c_0 and the next state x_1.

for $t=1,2,\ldots,T$ **do**

Take action u_t according to policy π_K, observe the reward c_t and the next state x_{t+1}.

Compute the TD-error $\delta_t=\vartheta_{t-1}^1-c_{t-1}+[\phi(x_{t-1},u_{t-1})-\tau_K(x_t,u_t)\cdot\phi(x_t,u_t)]^\top\vartheta_{t-1}^2$.

Update the primal variable ϑ by

$$\vartheta_t^1=\vartheta_{t-1}^1-\alpha_t\cdot[\omega_{t-1}^1+\phi(x_{t-1},u_{t-1})^\top\omega_{t-1}^2],$$
$$\vartheta_t^2=\vartheta_{t-1}^2-\alpha_t\cdot[\phi(x_{t-1},u_{t-1})-\tau_K(x_t,u_t)\cdot\phi(x_t,u_t)]\cdot[\phi(x_{t-1},u_{t-1})^\top\omega_{t-1}^2+\omega_{t-1}^1].$$

Update the dual variable ω by

$$\omega_t^1=(1-\alpha_t)\cdot\omega_t^1+\alpha_t\cdot\{\vartheta_{t-1}^1+[\phi(x_{t-1},u_{t-1})-\tau_K(x_t,u_t)\cdot\phi(x_t,u_t)]^\top\vartheta_{t-1}^2-c_{t-1}\},$$
$$\omega_t^2=(1-\alpha_t)\cdot\omega_t^2+\alpha_t\cdot\delta_t\cdot\phi(x_{t-1}).$$

Project ϑ_t and ω_t to $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$, respectively.

end for

Define $\widehat{\vartheta}=(\widehat{\vartheta}^1,\widehat{\vartheta}^2)=(\sum_{t=1}^T\alpha_t\cdot\vartheta_t)/(\sum_{t=1}^T\alpha_t)$ and $\widehat{\omega}=(\sum_{t=1}^T\alpha_t\cdot\omega_t)/(\sum_{t=1}^T\alpha_t)$.

Return $\widehat{\vartheta}^1$ and $\widehat{\Theta}=\mathrm{smat}(\widehat{\vartheta}^2)$ as the estimators of $J(K)$ and Θ_K, respectively.

Similar to Theorem 4.2, we have the following theorem that shows that Algorithm 3 converges at a sublinear rate to the desired solution ϑ_K^*.

Theorem B.1 (off-policy GTD)**.** Let $\widehat{\vartheta}^1$ and $\widehat{\Theta}$ be the output of Algorithm 3 based on T iterations. We set the stepsize to be $\alpha_t=\alpha/\sqrt{t}$ with $\alpha>0$ being a constant. We assume that Λ_K in (B.2) is invertible and that its minimum singular value is lower bounded by a constant $\kappa_k^*>0$. Moreover, we assume that the Markov chain induced by the behavioral policy π_b is geometrically β-mixing with parameter $\rho\in(0,1)$. Let ν_b be the stationary distribution of this induced Markov chain. We assume that, for $(x,u)\sim\nu_b$, both $\phi(x,u)$ and $\tau_K(x,u)$ are sub-exponential random variables. Then, when the number of iterations T is sufficiently large, with probability at least $1-T^{-4}$, we have

$$\|\widehat{\Theta}-\Theta_K\|_{\mathrm{fro}}^2\le\frac{\Upsilon\big[\widetilde{R}_\Theta,\widetilde{R}_\Omega,J(K_0),\|K\|_{\mathrm{fro}},\sigma_{\min}^{-1}(Q)\big]}{{\kappa_K^*}^2\cdot(1-\rho)}\cdot\frac{\log^6T}{\sqrt{T}},$$

where $\Upsilon\big[\widetilde{R}_\Theta,\widetilde{R}_\Omega,J(K_0),\|K\|_{\mathrm{fro}},\sigma_{\min}^{-1}(Q)\big]$ is a polynomial of $\widetilde{R}_\Omega$, $\widetilde{R}_\Omega$, $J(K_0)$, $\|K\|_{\mathrm{fro}}$, and $1/\sigma_{\min}(Q)$.

Proof. The proof of this theorem is parallel to that of Theorem 4.2, thus here we only sketch the proof for brevity.

The proof can be completed in three steps. In the first step, we show that $(\vartheta, \omega) = (\vartheta_K^*, 0)$ is the saddle point of the optimization problem in (B.3). To simplify the notation, we define a vector-valued function $G(x, u, x', u'; \vartheta)$ by

$$\begin{aligned} G^1(x, u, x', u'; \vartheta) &= \vartheta^1 - c(x, u) + \langle \phi(x, u) - \tau_K(x', u') \cdot \phi(x', u'), \vartheta^2 \rangle, \\ G^2(x, u, x', u'; \vartheta) &= \vartheta^1 \cdot \phi(x, u) + \big\{ \big[\phi(x, u) - \phi(x', u') \cdot \tau_K(x', u')\big]^\top \vartheta^2 - c(x, u) \big\} \cdot \phi(x, u). \end{aligned} \tag{B.4}$$

By definition, for all (ϑ, ω), $\overline{F}(\vartheta, \omega)$ in (B.3) can be equivalently written as

$$\overline{F}(\vartheta, \omega) = \big\langle \mathbb{E}_{(x,u,x',u')}[G(x, u, x', u'; \vartheta)], \omega \big\rangle - 1/2 \cdot \|\omega\|_2^2. \tag{B.5}$$

Thus, for any ϑ, the solution to the unconstrained maximization problem $\max_\omega F(\theta, \omega)$ is

$$w(\vartheta) = \mathbb{E}_{(x,u,x',u')}[G(x, u, x', u'; \vartheta)].$$

Recall that $c(x, u) = \langle \phi(x, u), \mathrm{svec}[\mathrm{diag}(Q, R)] \rangle$. Since both $\phi(x, u)$ and $\tau_K(x, u) \cdot \phi(x, u)$ are sub-exponential random vectors, when $\widetilde{R}_\Theta$ and $\widetilde{R}_\Omega$ are chosen properly, we can show that $\vartheta_K^* \in \mathcal{X}_\Theta$ and that $w(\vartheta) \in \mathcal{X}_\Omega$ for all $\vartheta \in \mathcal{X}_\Theta$. Notice that $w(\vartheta_K^*) = 0$. Thus, $(\vartheta_K^*, 0)$ is the solution to the minimax optimization problem in (B.3).

Then, in the second step, we relate the estimation error $\|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2$ to the primal-dual gap

$$\mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}) = \max_{\omega \in \mathcal{X}_\Omega} \overline{F}(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \overline{F}(\vartheta, \widehat{\omega}). \tag{B.6}$$

Similar to the derivations in (C.14)–(C.17), since the minimum singular value of Λ_K is lower bounded by κ_K^*, it holds that

$$|\widehat{\vartheta}^1 - J(K)|^2 + \|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2 \le {\kappa_K^*}^{-2} \cdot \mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}). \tag{B.7}$$

Thus, it suffices to bound the primal-dual gap in (B.6), which is achieved in the last step.

Specifically, we would like to utilize Theorem C.4 obtained from [72]. Since this theorem requires bounded iterates and Lipschitz gradient, similar to the third step in §C.1, we truncate the feature vector $\phi(x, u)$. In particular, we define

$$\mathcal{E} = \bigcap_{0 \le t \le T} \Big\{ \|\phi(x_t, u_t)\|_2^2 \le C_K \cdot \log T, \|\tau_K(x_t, u_t) \cdot \phi(x_t, u_t)\|_2^2 \le C_K \cdot \log T \Big\}, \tag{B.8}$$

where C_b is a constant specified by the stationary distribution ν_b. Since both is $\phi(x, u)$ and $\phi(x, u) \cdot \tau_K(x, u)$ are sub-exponential random vector when $(x, u) \sim \nu_b$, it can be shown that $\mathcal{E}$ holds with probability at least $1 - 2T^{-5}$. Then we define truncated random vectors

$$\widetilde{\phi}(x, u) = \phi(x, u) \cdot \mathbb{1}_{\mathcal{E}}, \qquad \widetilde{\varphi}_K(x, u) = \phi(x, u) \cdot \tau_K(x, u) \cdot \mathbb{1}_{\mathcal{E}}$$

and the truncated minimax optimization problem,

$$\min_{\vartheta \in \mathcal{X}_\Theta} \max_{\omega \in \mathcal{X}_\Omega} \widetilde{F}(\vartheta, \omega) = \big\langle \mathbb{E}_{(x,u,x',u')} \big[\widetilde{G}(x, u, x', u'; \vartheta)\big], \omega \big\rangle - 1/2 \cdot \|\omega\|_2^2, \tag{B.9}$$

where we define $\widetilde{G}(x, u, x', u'; \vartheta)$ by

$$\begin{aligned} \widetilde{G}^1(x, u, x', u'; \vartheta) &= \vartheta^1 - \widetilde{c}(x, u) + \langle \widetilde{\phi}(x, u) - \widetilde{\varphi}_K(x', u'), \vartheta^2 \rangle, \\ G^2(x, u, x', u'; \vartheta) &= \vartheta^1 \cdot \widetilde{\phi}(x, u) + \big\{ \big[\widetilde{\phi}(x, u) - \widetilde{\varphi}_K(x', u')\big]^\top \vartheta^2 - \widetilde{c}(x, u) \big\} \cdot \widetilde{\phi}(x, u). \end{aligned} \tag{B.10}$$

Here we let $\widetilde{c}(x, u) = \langle \widetilde{\phi}(x, u), \mathrm{svec}[\mathrm{diag}(Q, R)] \rangle$ in (B.10). Similar to the derivations from (C.25) to (C.31), we can show that $\sup_{\vartheta, \omega} |\overline{F}(\vartheta, \omega) - \widetilde{F}(\vartheta, \omega)| \le 1/T$, which implies that

$$\Big| \mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}) - \Big[\max_{\omega \in \mathcal{X}_\Omega} \widetilde{F}(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \widetilde{F}(\vartheta, \widehat{\omega}) \Big] \Big| \le 2/T. \tag{B.11}$$

Since $\widetilde{F}$ defined in (B.9) have Lipschitz gradients, by Theorem C.4, we have

$$\max_{\omega \in \mathcal{X}_\Omega} \widetilde{F}(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \widetilde{F}(\vartheta, \widehat{\omega}) \le \frac{C_K \cdot \log^6 T}{(1 - \rho) \cdot \sqrt{T}} \tag{B.12}$$

with probability at least $1 - T^{-5}$, where C_K is a constant that depends polynomially on $\widetilde{R}_\Theta$, $\widetilde{R}_\Omega$, $J(K_0)$, $\|K\|_{\mathrm{fro}}$, and $1/\sigma_{\min}(Q)$. Finally, combining (B.7), (B.11), and (B.12), we conclude the proof of this theorem. □

C Proofs of the Main Results

In this section, we provide the proofs of the main results, namely, Theorems 4.2 and 4.3, which are proved in §C.1 and §C.2, respectively. The proofs of the supporting results are deferred to the appendix.

C.1 Proof of Theorem 4.2

Proof. Our proof can be decomposed into three steps. In the first step, we show that, with $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ given in (4.1) and (4.2), $(\vartheta, \omega) = (\vartheta_K^*, 0)$ is the solution to the minimax optimization problem in (3.18). Then, in the second step, we show that the primal-dual gap of this optimization problem yields an upper bound for the estimation error $\|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2$, where $\widehat{\Theta} = \mathrm{smat}(\widehat{\vartheta}^2)$ is the estimator of Θ_K returned by the GTD algorithm. Finally, in the last step, we study the performance of such a minimax optimization problem, which enables us to establish the error of policy evaluation.

Step 1. In the first step, we show that $(\vartheta, \omega) = (\vartheta_K^*, 0)$ is the saddle point of the optimization problem in (3.18). To simplify the notation, we define a vector-valued function $G(x, u, x', u'; \vartheta)$ by

$$\begin{aligned} G^1(x, u, x', u'; \vartheta) &= \vartheta^1 - c(x, u), \\ G^2(x, u, x', u'; \vartheta) &= \vartheta^1 \cdot \phi(x, u) + \big\{ \big[\phi(x, u) - \phi(x', u')\big]^\top \vartheta^2 - c(x, u) \big\} \cdot \phi(x, u). \end{aligned} \tag{C.1}$$

By definition, $G(x, u, x', u'; \vartheta)$ is of the same shape as ϑ and ω. Moreover, for all (ϑ, ω), $F(\vartheta, \omega)$ in (3.18) can be equivalently written as

$$F(\vartheta, \omega) = \big\langle \mathbb{E}_{(x,u,x',u')}[G(x, u, x', u'; \vartheta)], \omega \big\rangle - 1/2 \cdot \|\omega\|_2^2. \tag{C.2}$$

Thus, for any ϑ, the solution to the unconstrained maximization problem $\max_\omega F(\theta, \omega)$ is

$$w(\vartheta) = \mathbb{E}_{(x,u,x',u')}[G(x, u, x', u'; \vartheta)]. \tag{C.3}$$

In the following, we show that $\vartheta_K^* \in \mathcal{X}_\Theta$. Moreover, we also prove that, for any $\vartheta \in \mathcal{X}_\Theta$, $w(\vartheta)$ in (C.3) belongs to $\mathcal{X}_\Omega$, where $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ are defined in (4.1) and (4.2), respectively. Since $w(\vartheta_K^*) = 0$, it holds that $(\vartheta_K^*, 0)$ is the solution to the minimax optimization problem in (3.18).

Recall that we assume $J(K) \leq J(K_0)$, where K_0 is the initial policy that is stable. Thus, $J(K_0)$ is finite. By the definition of ϑ_K^*, to show $\vartheta_K^* \in \mathcal{X}_\Theta$, it suffices to bound $\|\Theta_K\|_{\mathrm{fro}}$. By the definition of Θ_K in (3.7), we have

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

which implies that

$$\|\Theta_K\|_{\mathrm{fro}} \leq (\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}}) + (\|A\|_{\mathrm{fro}}^2 + \|B\|_{\mathrm{fro}}^2) \cdot \|P_K\|_{\mathrm{fro}}. \tag{C.4}$$

Now we apply the following lemma to obtain an upper bound on $\|P_K\|_{\mathrm{fro}}$.

Lemma C.1. When π_K is a stable policy, we have

$$\|\Sigma_K\| \leq J(K)/\sigma_{\min}(Q), \qquad \|P_K\| \leq J(K)/\sigma_{\min}(\Psi),$$

where $\sigma_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix.

Proof. By (3.8) in Proposition 3.1, we have

$$\begin{aligned} J(K) &\geq \mathrm{tr}[(Q + K^\top R K)\Sigma_K] \geq \sigma_{\min}(Q) \cdot \mathrm{tr}(\Sigma_K) \geq \sigma_{\min}(Q) \cdot \|\Sigma_K\|, \\ J(K) &\geq \mathrm{tr}(P_K \Psi_\sigma) \geq \sigma_{\min}(\Psi_\sigma) \cdot \mathrm{tr}(P_K) \geq \|P_K\| \geq J(K)/\sigma_{\min}(\Psi), \end{aligned}$$

where we use the fact that $\Psi_\sigma \succeq \Psi$. Therefore, we conclude the proof. □

Applying Lemma C.1 to (C.4), we have

$$\|\Theta_K\|_{\mathrm{fro}} \leq (\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}}) + (\|A\|_{\mathrm{fro}}^2 + \|B\|_{\mathrm{fro}}^2) \cdot \sqrt{d} \cdot J(K)/\sigma_{\min}(\Psi). \tag{C.5}$$

Combining (C.5) and the definition of $\widetilde{R}_\Theta$ in (4.3) we conclude that $\vartheta_K^* \in \mathcal{X}_\Theta$.

Furthermore, it remains to show that the vector in (C.3) belongs to $\mathcal{X}_\Omega$ for all $\vartheta \in \mathcal{X}_\Theta$. We consider the two components of $G(x, u, x', u'; \vartheta)$ separately. By (C.1), we have

$$\big|\mathbb{E}_{(x,u,x',u')}[G^1(x, u, x', u'; \vartheta)]\big| = |\vartheta^1 - J(K)| \le J(K_0), \tag{C.6}$$

where the second inequality follows from the fact that $0 \le \vartheta^1 \le J(K_0)$. Moreover, by (C.1), for the second component of $G(x, u, x', u'; \vartheta)$, we have

$$\mathbb{E}_{(x,u,x',u')}[G^2(x, u, x', u'; \vartheta)] = \vartheta^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)] + \Xi_K \vartheta^2 - b_K, \tag{C.7}$$

where Ξ_K and b_K are defined in (3.15). By Lemma D.2, we have

$$\|\Xi_K \vartheta^2\|_2 \le \|\Xi_K\| \cdot \|\vartheta^2\|_2 \le 4(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \|\Sigma_K\|^2 \cdot \widetilde{R}_\Theta. \tag{C.8}$$

Moreover, for any positive definite matrix Γ, we have

$$b_K^\top \operatorname{svec}(\Gamma) = \mathbb{E}_{(x,u)}\big\{\big\langle \phi(x, u), \operatorname{svec}[\operatorname{diag}(Q, R)]\big\rangle \cdot \big\langle \phi(x, u), \operatorname{smat}(\Gamma)\big\rangle\big\}, \tag{C.9}$$

where $\operatorname{diag}(Q, R)$ is the block diagonal matrix constructed by Q and R. Note that the joint distribution of (x, u) is the Gaussian distribution $N(0, \widetilde{\Sigma}_K)$, where $\widetilde{\Sigma}_K$ is defined in (D.16). Thus, $b_K^\top \operatorname{svec}(\Gamma)$ can be written as the product of two quadratic forms of Gaussian random variables. Applying Lemma D.3 to (C.9), we obtain that

$$b_K^\top \operatorname{svec}(\Gamma) = 2\big\langle \widetilde{\Sigma}_K \operatorname{diag}(Q, R) \widetilde{\Sigma}_K, \Gamma\big\rangle \cdot + \big\langle \widetilde{\Sigma}_K, \operatorname{diag}(Q, R)\big\rangle \cdot \big\langle \widetilde{\Sigma}_K, \Gamma\big\rangle,$$

which implies that

$$\|b_K\|_2 \le 3(\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}}) \cdot \|\widetilde{\Sigma}_K\|^2. \tag{C.10}$$

In addition, the first term on the right-hand side of (C.7) is bounded by

$$\big\|\vartheta^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)]\big\|_2 \le J(K_0) \cdot \big\|\widetilde{\Sigma}_K\big\|_{\mathrm{fro}}. \tag{C.11}$$

Finally, combining (C.8), (C.10), (C.11), and the upper bounds in (D.17), we have

$$\begin{aligned}
&\big\|\mathbb{E}_{(x,u,x',u')}[G^2(x, u, x', u'; \vartheta)]\big\|_2 \\
&\quad \le 2(d + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\| + 4(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \|\Sigma_K\|^2 \cdot \widetilde{R}_\Theta \\
&\qquad\quad + 12(\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}}) \cdot (d + \|K\|_{\mathrm{fro}}^2)^2 \cdot \|\Sigma_K\|^2 \\
&\quad \le C \cdot (1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Theta \cdot \sigma_{\min}^{-2}(Q) \cdot [J(K_0)]^2,
\end{aligned} \tag{C.12}$$

where $C > 0$ is an absolute constant.

Hence, combining (4.4), (C.6) and (C.12), we conclude that $w(\vartheta) \in \mathcal{X}_\Omega$ for all $\vartheta \in \mathcal{X}_\Theta$. Therefore, we have shown that $(\vartheta_K^*, 0)$ is the saddle point of the optimization problem in (3.18), which concludes the first step of the proof.

Step 2. In the following, we relate the estimation error $\|\widehat{\Theta} - \Theta_K\|_{\mathrm{fro}}^2$ to the performance of the optimization in (3.18). Specifically, we consider the primal-dual gap

$$\mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}) = \max_{\omega \in \mathcal{X}_\Omega} F(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} F(\vartheta, \widehat{\omega}), \tag{C.13}$$

which characterizes the closeness between $(\widehat{\vartheta}, \widehat{\phi})$ and the optimal solution $(\vartheta_K^*, 0)$, quantified by the objective value.

Recall that $w(\vartheta)$ defined in (C.3) is the optimal dual variable for each $\theta \in \mathcal{X}_\Theta$. Hence, for any $\omega \in \mathcal{X}_\Omega$, it holds that

$$\begin{aligned}
\min_{\vartheta \in \mathcal{X}_\Theta} F(\vartheta, \omega) &\le \min_{\theta \in \mathcal{X}_\Theta} \max_{\omega \in \mathcal{X}_\Omega} F(\theta, \omega) \\
&\le \min_{\vartheta \in \mathcal{X}_\Omega} \big\{[\vartheta^1 - J(K)]^2 + \|\vartheta^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)] + \Xi_K \vartheta^2 - b_K\|_2^2\big\} = 0.
\end{aligned} \tag{C.14}$$

Thus, for $\widehat{\vartheta}$ returned by the GTD algorithm, we have

$$\begin{aligned}
&\big\{[\widehat{\vartheta}^1 - J(K)]^2 + \|\widehat{\vartheta}^1 \cdot \mathbb{E}_{(x,u)}[\phi(x, u)] + \Xi_K \widehat{\vartheta}^2 - b_K\|_2^2\big\} = \max_{\omega \in \mathcal{X}_\Omega} F(\widehat{\vartheta}, \omega) \\
&\quad = \max_{\omega \in \mathcal{X}_\Omega} F(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} F(\vartheta, \widehat{\omega}) + \min_{\vartheta \in \mathcal{X}_\Theta} F(\vartheta, \widehat{\omega}) \le \mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}),
\end{aligned} \tag{C.15}$$

where the last inequality follows from (C.14).

Furthermore, by direct computation, we can bound the left-hand side of (C.15) via

$$\begin{aligned}&\left\|\begin{pmatrix}1 & 0\\ \mathbb{E}_{(x,u)}[\phi(x,u)] & \Xi_K\end{pmatrix}(\widehat{\vartheta}-\vartheta_K^*)\right\|_2^2\\ &\quad\geq {\kappa_K^*}^2\cdot\|\widehat{\vartheta}-\vartheta_K^*\|_2^2=\kappa_K^*\cdot\big[\|\widehat{\Theta}-\Theta_K\|_{\mathrm{fro}}^2+|\widehat{\vartheta}^1-J(K)|^2\big],\end{aligned}\tag{C.16}$$

where we utilize the fact that ϑ_K^* is the solution to the linear equation in (3.16) and κ_K^* is specified in Lemma 3.2. Therefore, combining (C.15) and (C.16), we have

$$|\widehat{\vartheta}^1-J(K)|^2+\|\widehat{\Theta}-\Theta_K\|_{\mathrm{fro}}^2\leq {\kappa_K^*}^{-2}\cdot\mathtt{Gap}(\widehat{\vartheta},\widehat{\omega}),\tag{C.17}$$

which establishes the connection between $\|\widehat{\Theta}-\Theta_K\|_{\mathrm{fro}}^2$ and the primal-dual gap in (C.13).

Step 3. In the last step, we construct an upper bound for the primal-dual gap. By (C.17), this yields an upper bound for the error of parameter estimation.

Note that the distribution of the state-action pair (x,u) have unbounded support. We first construct an event such that $\{\phi(x_t,u_t)\}_{t=0}^T$ are bounded conditioning on this event. To this end, we establish an upper bound for tail probability of the $\|\phi(x,u)\|_2$ using the Hansen-Wright inequality stated as follows.

Lemma C.2 (Hansen-Wright inequality)**.** For any integer $m>0$, let A be a matrix in $\mathbb{R}^{m\times m}$ and let $\eta\sim N(0,I_m)$ be the standard Gaussian random variable in $\mathbb{R}^m$. Then, there exists an absolute constant $C>0$ such that, for any $t\geq 0$, we have

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

Proof. See [55] for a detailed proof. □

Applying Lemma C.2 to $(x,u)\sim N(0,\widetilde{\Sigma}_K)$ with $\widetilde{\Sigma}_K$ defined in (D.16), we obtain

$$\mathbb{P}\big[\big|\|x\|_2^2+\|u\|_2^2-\mathrm{tr}\big(\widetilde{\Sigma}_K\big)\big|>t\big]\leq 2\cdot\exp\big[-C\cdot\min\big(t^2\cdot\big\|\widetilde{\Sigma}_K\big\|_{\mathrm{fro}}^{-2},\,t\cdot\big\|\widetilde{\Sigma}_K\big\|^{-1}\big)\big].\tag{C.18}$$

Setting $t=C_1\cdot\log T\cdot\big\|\widetilde{\Sigma}_K\big\|$ in (C.18) with constant C_1 sufficiently large, it holds that

$$t^2\cdot\big\|\widetilde{\Sigma}_K\big\|_{\mathrm{fro}}^{-2}=\big\|\widetilde{\Sigma}_K\big\|_{\mathrm{fro}}^{-2}\cdot C_1^2\cdot\log^2 T\cdot\big\|\widetilde{\Sigma}_K\big\|^2\geq C_1^2\cdot(d+k)^{-1}\cdot\log^2 T\geq t\cdot\big\|\widetilde{\Sigma}_K\big\|^{-1},\tag{C.19}$$

where the first inequality follows from the relation between the operator and Frobenius norms, and the second inequality holds when $\log T\geq C_1^{-1}\cdot(d+k)$. For ease of presentation, for any $t\in\{0,1,\ldots,T\}$, we define

$$\mathcal{E}_t=\Big\{\big|\|x_t\|_2^2+\|u_t\|_2^2-\mathrm{tr}\big(\widetilde{\Sigma}_K\big)\big|\leq C_1\cdot\log T\cdot\big\|\widetilde{\Sigma}_K\big\|\Big\},\tag{C.20}$$

and write $\mathcal{E}=\bigcap_{0\leq t\leq T}\mathcal{E}_t$. Combining (C.18) and (C.19), we obtain that $\mathcal{E}_t$ holds with probability at least $1-T^{-6}$. Thus, by taking a union bound for $\{(x_t,u_t)\}_{t=0}^T$, we have $\mathbb{P}(\mathcal{E})\geq 1-2T^{-5}$. Moreover, combining (C.20) and (D.17) further implies that, on event $\mathcal{E}$, we have

$$\begin{aligned}\max_{0\leq t\leq T}\big\{\|x_t\|_2^2+\|u_t\|_2^2\big\}&\leq C_1\cdot\log T\cdot\big\|\widetilde{\Sigma}_K\big\|+\mathrm{tr}\big(\widetilde{\Sigma}_K\big)\leq\big(C_1\cdot\log T+d+k\big)\cdot\big\|\widetilde{\Sigma}_K\big\|\\ &\leq 2C_1\cdot\log T\cdot\big\|\widetilde{\Sigma}_K\big\|\leq 2C_1\cdot\log T\cdot\big[\sigma^2+(1+\|K\|_{\mathrm{fro}}^2)\cdot\|\Sigma_K\|\big].\end{aligned}\tag{C.21}$$

In the sequel, we study the stochastic optimization problem in (3.18) with the restriction that $\mathcal{E}$ holds. Specifically, for any state-action pair (x,u), we define the truncated feature function as

$$\widetilde{\phi}(x,u)=\phi(x,u)\cdot\mathbb{1}\Big\{\big|\|\phi(x,u)\|_2^2-\mathrm{tr}(\widetilde{\Sigma}_K)\big|\leq C_1\cdot\log T\cdot\big\|\widetilde{\Sigma}_K\big\|\Big\}.\tag{C.22}$$

By this definition, for any $t\in\{0,\ldots,t\}$, we have $\widetilde{\phi}(x_t,u_t)=\phi(x_t,u_t)\cdot\mathbb{1}_{\mathcal{E}_t}$. Now we replace $\phi(x,u)$ by $\widetilde{\phi}(x,u)$ in (3.18) and consider the following minimiax optimization problem:

$$\min_{\vartheta\in\mathcal{X}_\Theta}\max_{\omega\in\mathcal{X}_\Omega}\widetilde{F}(\vartheta,\omega)=\big\langle\mathbb{E}_{(x,u,x',u')}\big[\widetilde{G}(x,u,x',u';\vartheta)\big],\omega\big\rangle-1/2\cdot\|\omega\|_2^2,\tag{C.23}$$

where, similar to $G(x,u,x',u';\vartheta)$ in (C.1), we define $\widetilde{G}(x,u,x',u';\vartheta)$ by

$$\begin{aligned}\widetilde{G}^1(x,u,x',u';\vartheta) &= \vartheta^1 - \widetilde{c}(x,u),\\ \widetilde{G}^2(x,u,x',u';\vartheta) &= \vartheta^1 \cdot \widetilde{\phi}(x,u) + \big\{\big[\widetilde{\phi}(x,u) - \widetilde{\phi}(x',u')\big]^\top \vartheta^2 - \widetilde{c}(x,u)\big\} \cdot \widetilde{\phi}(x,u).\end{aligned} \tag{C.24}$$

Here we denote $\widetilde{c}(x,u) = \langle \widetilde{\phi}(x,u), \mathrm{svec}[\mathrm{diag}(Q,R)]\rangle$ in (C.24) to simplify the notation.

We remark that, when $\mathcal{E}$ is true, $(\widehat{\vartheta},\widehat{\omega})$ is also the solution returned by the gradient-based algorithm for the minimax optimization problem in (C.23). As a result, when $\mathcal{E}$ holds, the primal-dual gap of (C.23) is equal to $\max_{\omega\in\mathcal{X}_\Omega} \widetilde{F}(\widehat{\vartheta},\omega) - \min_{\vartheta\in\mathcal{X}_\Theta} \widetilde{F}(\vartheta,\widehat{\omega})$.

In the following, we characterize the difference between the objective functions in (3.18) and (C.23). For any $(\vartheta,\omega)\in\mathcal{X}_\Theta\times\mathcal{X}_\Omega$, by (C.2) and (C.23) we have

$$\begin{aligned}\big|F(\vartheta,\omega) - \widetilde{F}(\vartheta,\omega)\big| &= \big|\big\langle \mathbb{E}_{(x,u,x',u')}\big[G(x,u,x',u';\vartheta) - \widetilde{G}(x,u,x',u';\vartheta)\big], \omega\big\rangle\big|\\ &\le \big|\mathbb{E}_{(x,u,x',u')}\big[G^1(x,u,x',u';\vartheta) - \widetilde{G}^1(x,u,x',u';\vartheta)\big]\big| \cdot J(K_0)\\ &\quad + \big\|\mathbb{E}_{(x,u,x',u')}\big[G^2(x,u,x',u';\vartheta) - \widetilde{G}^2(x,u,x',u';\vartheta)\big]\big\|_2 \cdot \widetilde{R}_\Omega.\end{aligned} \tag{C.25}$$

By the definitions of $G(x,u,x',u';\vartheta)$ and $\widetilde{G}(x,u,x',u';\vartheta)$ in (C.1) and (C.24), we have

$$G^1(x,u,x',u';\vartheta) - \widetilde{G}^1(x,u,x',u';\vartheta) = c(x,u)\cdot \mathbb{1}_{\mathcal{A}^c} \tag{C.26}$$

$$G^2(x,u,x',u';\vartheta) - \widetilde{G}^2(x,u,x',u';\vartheta) = G^2(x,u,x',u';\vartheta)\cdot \mathbb{1}_{\mathcal{A}^c} + \phi(x',u')^\top\vartheta^2\cdot\phi(x,u)\cdot\mathbb{1}_{\mathcal{A}}\cdot\mathbb{1}_{\mathcal{B}^c},$$

where we denote $\{|\|\phi(x,u)\|_2^2 - \mathrm{tr}(\widetilde{\Sigma}_K)| \le C_1\cdot\log T\cdot\|\widetilde{\Sigma}_K\|\}$ and $\{|\|\phi(x,u)\|_2^2 - \mathrm{tr}(\widetilde{\Sigma}_K)| \le C_1\cdot\log T\cdot\|\widetilde{\Sigma}_K\|\}$ by $\mathcal{A}$ and $\mathcal{B}$, respectively, and $\mathcal{A}^c$, $\mathcal{B}^c$ are the complement sets of $\mathcal{A}$ and $\mathcal{B}$.

For the first term on the right-hand side of (C.25), Cauchy-Schwarz inequality implies that

$$\big|\mathbb{E}_{(x,u,x',u')}\big[G^1(x,u,x',u';\vartheta) - \widetilde{G}^1(x,u,x',u';\vartheta)\big]\big| \le \sqrt{\mathbb{P}(\mathcal{A}^c)}\cdot\sqrt{\mathbb{E}[c^2(x,u)]}. \tag{C.27}$$

Since $c(x,u)$ is a quadratic form of a Gaussian random variable, by Lemma D.3, we have

$$\begin{aligned}\mathbb{E}[c^2(x,u)] &= 2\,\mathrm{tr}\big[\widetilde{\Sigma}_K\mathrm{diag}(Q,R)\widetilde{\Sigma}_K\mathrm{diag}(Q,R)\big] + \big\{\mathrm{tr}\big[\widetilde{\Sigma}_K\mathrm{diag}(Q,R)\big]\big\}^2\\ &\le 3(\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}})^2\cdot\|\widetilde{\Sigma}_K\|_{\mathrm{fro}}^2 \le 3(\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}})^2\cdot\big[\sigma^2\cdot k + (d+\|K\|_{\mathrm{fro}}^2)^2\cdot\|\Sigma_K\|^2\big],\end{aligned}$$

where the last inequality follows from (D.17). Besides, for the second term on the right-hand side of (C.25), combining (C.25), (C.26), triangle inequality, and Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\big\|\mathbb{E}_{(x,u,x',u')}\big[G^2(x,u,x',u';\vartheta) - \widetilde{G}^2(x,u,x',u';\vartheta)\big]\big\|_2\\ &\quad\le \Big\{\big\|\mathbb{E}_{(x,u,x',u')}[G^2(x,u,x',u';\vartheta)\cdot\mathbb{1}_{\mathcal{A}^c}]\big\|_2 + \big\|\mathbb{E}_{(x,u,x',u')}[\phi(x',u')^\top\vartheta^2\cdot\phi(x,u)\,\mathbb{1}_{\mathcal{B}^c}]\big\|_2\Big\}\\ &\quad\le \Big\{\sqrt{\mathbb{P}(\mathcal{A}^c)}\cdot\sqrt{\mathbb{E}\big[\big\|G^2(x,u,x',u';\vartheta)\big\|_2^2\big]} + \sqrt{\mathbb{P}(\mathcal{B}^c)}\cdot\sqrt{\mathbb{E}\big[\big\|\phi(x,u)\cdot\phi(x',u')^\top\vartheta^2\big\|_2^2\big]}\Big\}.\end{aligned} \tag{C.28}$$

For the expectations on the right-hand side of (C.28), using the inequality $(a+b)^2\le 2a^2+2b^2$, we have

$$\begin{aligned}&\mathbb{E}\big[\big\|G^2(x,u,x',u';\vartheta)\big\|_2^2\big]\\ &\quad\le 2\cdot\mathbb{E}\Big\{\big[\vartheta^1 - c(x,u) + \phi(x,u)^\top\vartheta^2\big]^2\cdot\|\phi(x,u)\|_2^2\Big\} + 2\cdot\mathbb{E}\big[\big\|\phi(x,u)\cdot\phi(x',u')^\top\vartheta^2\big\|_2^2\big].\end{aligned} \tag{C.29}$$

Further applying Cauchy-Schwarz inequality to (C.29), we have

$$\begin{aligned}&\mathbb{E}\Big\{\big[\vartheta^1 - c(x,u) + \phi(x,u)^\top\vartheta^2\big]^2\cdot\|\phi(x,u)\|_2^2\Big\}\\ &\quad\le \Big(\mathbb{E}\big\{\big[\vartheta^1 - c(x,u) + \phi(x,u)^\top\vartheta^2\big]^4\big\}\cdot\mathbb{E}\big[\|\phi(x,u)\|_2^4\big]\Big)^{1/2},\end{aligned} \tag{C.30}$$

$$\mathbb{E}\big[\big\|\phi(x,u)\cdot\phi(x',u')^\top\vartheta\big\|_2^2\big] \le \Big(\mathbb{E}\big[|\phi(x',u')^\top\vartheta|^4\big]\cdot\mathbb{E}\big[\|\phi(x,u)\|_2^4\big]\Big)^{1/2}. \tag{C.31}$$

Since the marginal distributions of (x,u) and (x',u') are both $N(0,\widetilde{\Sigma}_K)$, in (C.30) and (C.31) we bound the two terms in (C.29) using the fourth moments of $N(0,\widetilde{\Sigma}_K)$, which can be written as a polynomial of $J(K_0)$, $\|K\|_{\mathrm{fro}}$, $\|Q\|$, $\|R\|$, $\widetilde{R}_\Theta$, and $\widetilde{R}_\Omega$.

Meanwhile, recall that we have shown that $\mathbb{P}(\mathcal{A}^c) \le T^{-6}$ and $\mathbb{P}(\mathcal{B}^c) \le T^{-6}$. Thus, when T is sufficiently large, by combining (C.25), (C.27), (C.28), and (C.29), we have $|F(\vartheta,\omega) - \widetilde{F}(\vartheta,\omega)| \le 1/T$, which implies that

$$\begin{aligned}&\Big|\mathtt{Gap}(\widehat{\vartheta},\widehat{\omega}) - \Big[\max_{\omega\in\mathcal{X}_\Omega}\widetilde{F}(\widehat{\vartheta},\omega) - \min_{\vartheta\in\mathcal{X}_\Theta}\widetilde{F}(\vartheta,\widehat{\omega})\Big]\Big| \\ &\quad\le \max_{\omega\in\mathcal{X}_\Omega}\big|F(\widehat{\vartheta},\omega) - \widetilde{F}(\widehat{\vartheta},\omega)\big| + \max_{\vartheta\in\mathcal{X}_\Theta}\big|F(\vartheta,\widehat{\omega}) - \widetilde{F}(\vartheta,\widehat{\omega})\big| \le \frac{2}{T}. \end{aligned} \tag{C.32}$$

Hereafter, we study the primal-dual gap in (C.13) conditioning on event $\mathcal{E}$. To simplify the notation, we define function $H(\vartheta,\omega;\phi,\phi')$ on $\mathcal{X}_\Theta\times\mathcal{X}_\Omega$ by

$$H(\vartheta,\omega;\phi,\phi') = \big\langle \widetilde{G}(x,u,x',u';\vartheta),\omega\big\rangle - 1/2\cdot\|\omega\|_2^2,$$

where the function $\widetilde{\phi}(x,u)$ is defined in (C.22), and we denote $\widetilde{\phi}(x,u)$ and $\widetilde{\phi}(x',u')$ by ϕ and ϕ', respectively. Using this definition, the objective function $\widetilde{F}(\vartheta,\omega)$ in (C.23) can be written as $\widetilde{F}(\vartheta,\omega) = \mathbb{E}_{(x,u,x',u')}[H(\vartheta,\omega;\phi,\phi')]$, where (x,u) and (x',u') are two consecutive state-action pairs. Note that $H(\vartheta,\omega;\phi,\phi')$ is a quadratic function of (ϑ,ω) for all ϕ and ϕ'. The partial gradients of $H(\vartheta,\omega;\phi,\phi')$ are given by

$$\nabla_{\vartheta^1}H(\vartheta,\omega;\phi,\phi') = \omega^1 + \widetilde{\phi}(x,u)^\top\omega^2, \tag{C.33}$$

$$\nabla_{\vartheta^2}H(\vartheta,\omega;\phi,\phi') = [\widetilde{\phi}(x,u)^\top\omega^2]\cdot[\widetilde{\phi}(x,u) - \widetilde{\phi}(x',u')], \tag{C.34}$$

$$\nabla_{\omega^1}H(\vartheta,\omega;\phi,\phi') = \vartheta^1 - \widetilde{c}(x,u) - \omega^1, \tag{C.35}$$

$$\nabla_{\omega^2}H(\vartheta,\omega;\phi,\phi') = \widetilde{G}^2(x,u,x',u';\vartheta) - \omega^2. \tag{C.36}$$

By combining (C.22), (C.33), and (C.34), we can bound the norm of $\nabla_\vartheta H(\vartheta,\omega;\phi,\phi')$ by

$$\begin{aligned}\big\|\nabla_\vartheta H(\vartheta,\omega;\phi,\phi')\big\|_2 &\le |\omega^1 + \widetilde{\phi}(x,u)^\top\omega^2| + \big\|[\widetilde{\phi}(x,u)^\top\omega^2]\cdot[\widetilde{\phi}(x,u) - \widetilde{\phi}(x',u')]\big\|_2 \\ &\le |\omega^1| + 2\|\widetilde{\phi}(x,u)\|_2\cdot\|\omega^2\|_2\cdot\big[\|\widetilde{\phi}(x,u)\|_2 + \|\widetilde{\phi}(x',u')\|_2\big] \\ &\le J(K_0) + 16C_1^2\cdot(1+\|K\|_{\mathrm{fro}}^2)^2\cdot\log^2 T\cdot\big[\sigma^2 + (1+\|K\|_{\mathrm{fro}}^2)\cdot\|\Sigma_K\|\big]^2\cdot\widetilde{R}_\Omega.\end{aligned} \tag{C.37}$$

Here the second inequality holds when $\|\widetilde{\phi}(x,u)\|_2 \ge 1$ and the last inequality follows from (C.21). Similarly, combining triangle inequality, (C.35), and (C.36), we have

$$\begin{aligned}\big\|\nabla_\omega H(\vartheta,\omega;\phi,\phi')\big\|_2 &\le |\vartheta^1 - \widetilde{c}(x,u) - \omega^1| + + \big[(\|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}})\cdot\|\widetilde{\phi}(x,u)\|_2 \\ &\qquad + (\|\widetilde{\phi}(x',u')\|_2 + \|\widetilde{\phi}(x,u)\|_2)\cdot\widetilde{R}_\Theta\big]\cdot\|\widetilde{\phi}(x,u)\|_2 \\ &\le 2J(K_0) + 16C_1^2\cdot\log^2 T\cdot\big[\sigma^2 + (1+\|K\|_{\mathrm{fro}}^2)\cdot\|\Sigma_K\|\big]^2\cdot\widetilde{R}_\Theta.\end{aligned} \tag{C.38}$$

where the last equality holds since $\widetilde{R}_\Theta \ge \|Q\|_{\mathrm{fro}} + \|R\|_{\mathrm{fro}}$. Moreover, we have $\nabla^2_{\vartheta\vartheta}H(\vartheta,\omega;\phi,\phi') = 0$ and $-\nabla^2_{\omega\omega}H(\vartheta,\omega;\phi,\phi')$ is the identity matrix.

We utilize the following lemma, obtained from [69], to handle the dependence along the trajectory.

Lemma C.3 (Geometrically β-mixing)**.** Consider a linear dynamical system $X_{t+1} = LX_t + \varepsilon$, where $\{X_t\}_{t\ge0}\subseteq\mathbb{R}^m$, $\varepsilon\sim N(0,\Psi)$ is the random noise, and $L\in\mathbb{R}^{m\times m}$ has spectral radius smaller than one. We denote by ν_t the marginal distribution of X_t for all $t\ge0$. Besides, the stationary distribution of this Markov chain is denoted by $N(0,\Sigma_\infty)$. For any integer $k\ge1$, we define the k-th mixing coefficient as

$$\beta(k) = \sup_{t\ge0}\mathbb{E}_{x\sim\nu_t}\big[\big\|\mathbb{P}_{X_k}(\cdot\,|\,X_0 = x) - \mathbb{P}_{N(0,\Sigma_\infty)}(\cdot)\big\|_{\mathrm{TV}}\big].$$

Furthermore, for any $\rho\in(\rho(L),1)$ and any $k\ge1$, we have

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

where $C_{\rho,L}$ is a constant that solely depends on ρ and A. That is, $\{X_t\}_{t\ge0}$ is geometrically β-mixing.

Proof. See Proposition 3.1 in [69] for a detailed proof. □

Recall that under policy π_K, $\{(x_t, u_t)\}_{t\geq 0}$ form a linear dynamic system characterized by (D.13) and (D.14). Since $\rho(L) = \rho(A - BK) < 1$, Lemma C.3 implies that, for all $\rho \in (\rho(A - BK), 1)$, $(x_t, u_t)_{t\geq 0}$ is a geometrically β-mixing stochastic process with parameter ρ. The following theorem, adapted from Theorem 1 in [72], establishes the primal-dual gap for a convex-concave minimax optimization problem involving a geometrically β-mixing stochastic process.

Theorem C.4 (Primal-dual gap for minimax optimization)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ are bounded and closed convex sets such that $\|x - x'\|_2 \leq D$ for all $x, x' \in \mathcal{X}$ and $\|y - y'\|_2 \leq D$ for all $y, y' \in \mathcal{Y}$. Consider the gradient algorithm for stochastic minimax optimization problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x, y) = \mathbb{E}_{\xi\sim\pi_\xi}[\Phi(x, y; \xi)], \tag{C.39}$$

where ξ is a random variable with distribution π_ξ and $F(x, y)$ is convex in x and concave in y. In addition, we assume that π_ξ is the stationary distribution of a Markov chain $\{\xi_t\}_{t\geq 0}$ which is geometrically β-mixing with parameter $\rho \in (0, 1)$. Specifically, we assume that there exists a constant $C_\xi > 0$ such that, for all $k \geq 1$, the k-th mixing coefficient satisfy $\beta(k) \leq C_\xi \cdot \rho^k$. Furthermore, we consider the case where, almost surely for every $\xi \sim \pi_\xi$, $\Phi(x, y; \xi)$ is L_1-Lipschitz in both x and y, $\nabla_x\Phi(x, y; \xi)$ is L_2-Lipschitz in x for all $y \in \mathcal{Y}$, and $\nabla_y\Phi(x, y; \xi)$ is L_2-Lipschitz in y for all $x \in \mathcal{X}$. Here, without loss of generality, we assume that $D, L_1, L_2 > 1$. Consider solving the optimization problem in (C.39) via T iterations of the gradient-based updates

$$x_t = \Pi_{\mathcal{X}}\big[x_{t-1} - \alpha_t\nabla_x\Phi(x_{t-1}, y_{t-1}; \xi_{t-1})\big], \qquad y_t = \Pi_{\mathcal{Y}}\big[y_{t-1} + \alpha_t \cdot \nabla_y\Phi(x_{t-1}, y_{t-1}; \xi_{t-1})\big],$$

where $t \in [T]$, $\Pi_{\mathcal{X}}$ and $\Pi_{\mathcal{Y}}$ are projection operators, and $\{\alpha_t = \alpha/\sqrt{t}\}_{t\in[T]}$ are the stepsizes, where $\alpha > 0$ is a constant. Let

$$\widehat{x} = \frac{\sum_{t\in[T]} \alpha_t \cdot x_t}{\sum_{t\in[T]} \alpha_t}, \qquad \widehat{y} = \frac{\sum_{t\in[T]} \alpha_t \cdot y_t}{\sum_{t\in[T]} \alpha_t}$$

be the final output of the algorithm. Then, there exists an absolute constant $C > 0$ such that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the primal-dual gap satisfies

$$\max_{y\in\mathcal{Y}} F(\widehat{x}, y) - \min_{x\in\mathcal{X}} F(x, \widehat{y}) \leq \frac{C \cdot (D^2 + L_1^2 + L_1L_2D)}{\log(1/\rho)} \cdot \frac{\log^2 T + \log(1/\delta)}{\sqrt{T}} + \frac{C \cdot C_\xi L_1 D}{T}.$$

Proof. This theorem follows from Theorem 1 in [72], where we set $\alpha_t = \alpha/\sqrt{t}$ for all $t \geq 1$, and focus on the case where $\{\xi_t\}_{t\geq 0}$ is geometrically β-mixing. Under the mixing assumption, for any $k \geq 1$, the k-th mixing coefficient of $\{\xi_t\}_{t\geq 0}$ satisfies $\beta(k) \leq C_\xi \cdot \rho^k$. Then, for any $\delta, \eta \in (0, 1)$, Theorem 1 in [72] implies

$$\begin{aligned}\max_{y\in\mathcal{Y}} F(\widehat{x}, y) - \min_{x\in\mathcal{X}} F(x, \widehat{y}) \leq \bigg(\sum_{t=1}^T \alpha_t\bigg)^{-1} \bigg(A_0 + A_1 \cdot \eta \cdot \sum_{t=1}^T \alpha_t + A_2 \sum_{t=1}^T \alpha_t^2 + \\ 16DL_1 \cdot \bigg\{2\tau(\eta) \cdot \log[\tau(\eta)/\delta] \cdot \bigg[\sum_{t=1}^T \alpha_t^2 + \tau(\eta) \cdot \alpha_1\bigg]\bigg\}^{1/2}\bigg),\end{aligned} \tag{C.40}$$

where we define $\tau(\eta) = \log(\eta/C_\xi)/\log(\rho)$ and denote

$$A_0 = D^2 + 12D \cdot \alpha_1 \cdot \tau(\eta) \qquad A_1 = 4L_1D \qquad A_2 = 10L_1^2 + (24L_1^2 + 8L_1L_2D) \cdot \tau(\eta).$$

Now we set $\alpha_t = \alpha/\sqrt{t}$ and $\eta = C_\xi/T$ in (C.40), which implies that $\tau(\eta) = \log T/\log(1/\rho)$. Moreover, note that for all $T \geq 1$, we have $2\sqrt{T+1} - 2 \leq \sum_{t=1}^T 1/\sqrt{t} \leq 2\sqrt{T} - 1$ and $\sum_{t=1}^T 1/t \leq \log T + 1$. The last term on the right-hand side of (C.40) can be upper bounded by

$$\begin{aligned}&16DL_1 \cdot \big\{2\log T/\log(1/\rho) \cdot \log[\tau(\eta)/\delta] \cdot \big[\log T + 1 + \alpha \cdot \log T/\log(1/\rho)\big]\big\}^{1/2} \\ &\quad\leq 16DL_1 \cdot \big\{2\log T/\log(1/\rho) \cdot [\log\log T + \log(1/\delta)] \cdot \big[\log T + 1 + \alpha \cdot \log T/\log(1/\rho)\big]\big\}^{1/2} \\ &\quad\leq C \cdot DL_1 \cdot \log T/\log(1/\rho) \cdot \sqrt{\log\log T + \log(1/\delta)},\end{aligned} \tag{C.41}$$

where C is an absolute constant. Moreover, for the first three terms, we have

$$A_0 = D^2 + 12D \cdot \alpha \cdot \log T / \log(1/\rho) \le C \cdot D^2 \log T / \log(1/\rho), \quad A_1 \cdot \eta \le C \cdot C_\xi L_1 D / T, \tag{C.42}$$

$$\begin{aligned} A_2 \cdot \sum_{t=1}^{T} \alpha_t^2 &\le \big[10L_1^2 + (24L_1^2 + 8L_1 L_2 D) \cdot \log T / \log(1/\rho)\big] \cdot (\log T + 1) \\ &\le C \cdot [L_1^2 + L_1 L_2 D] \cdot \log^2 T / \log(1/\rho). \end{aligned} \tag{C.43}$$

Thus, combining (C.40), (C.41), (C.42), and (C.43), we obtain that

$$\max_{y \in \mathcal{Y}} F(\widehat{x}, y) - \min_{x \in \mathcal{X}} F(x, \widehat{y}) \le C \cdot \big[(D^2 + L_1^2 + L_1 L_2 D)/\log(1/\rho) \cdot \log T \cdot \log(T/\delta)/\sqrt{T} + C_\xi L_1 D / T\big],$$

which concludes the proof of Theorem C.4. □

In order to apply Theorem C.4 to the minimax optimization in (C.23), we only need to specify parameters C_ξ, D, L_1, and L_2. First, for any $\rho \in (\rho(A - BK), 1)$, by Lemma C.3, we can set

$$\begin{aligned} C_\xi &= C_{\rho,L} \cdot \big[\mathrm{tr}(\widetilde{\Sigma}_K) + (d+k) \cdot (1-\rho)^2\big]^{1/2} \\ &\le 2C_{\rho,L} \cdot \sqrt{d+k} \cdot \big\{\big[\sigma^2 + (1 + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\|\big]^{1/2} + (1-\rho)^{-1}\big\}. \end{aligned} \tag{C.44}$$

Moreover, by the definitions of $\mathcal{X}_\Theta$ and $\mathcal{X}_\Omega$ in (4.1) and (4.2), respectively, we can set D by

$$D^2 = 2[J(K_0)]^2 + \widetilde{R}_\Theta^2 + (1 + \|K\|_{\mathrm{fro}}^2)^4 \cdot \widetilde{R}_\Omega^2. \tag{C.45}$$

Moreover, by (C.37), (C.38), and the form of $\nabla^2 G(\theta, \omega; \phi, \phi')$, we have

$$L_1 \le 16 C_1^2 \cdot \log^2 T \cdot \big[\sigma^2 + (1 + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\|\big]^2 \cdot \big[(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Omega + \widetilde{R}_\Theta\big], \qquad L_2 = 1. \tag{C.46}$$

Combining Theorem C.4, (C.44), (C.45), and (C.46), we to obtain an upper bound for the primal-dual gap in (C.13). Specifically, for any $\rho \in (\rho(A - BK), 1)$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the primal-dual gap of the optimization problem in (C.23) is bounded by

$$\begin{aligned} &C \cdot \log^4 T \cdot \big[\sigma^2 + (1 + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\|\big]^4 \cdot \big[(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Omega + \widetilde{R}_\Theta\big]^2 \\ &\quad \cdot \bigg(\frac{\log^2 T + \log(1/\delta)}{\log(1/\rho) \cdot \sqrt{T}} + \frac{\sqrt{d+k}}{(1-\rho) \cdot T}\bigg). \end{aligned} \tag{C.47}$$

where $C > 0$ is an absolute constant. Besides, we note that σ is a constant and that $\|\Sigma_K\| \ge \sigma_{\min}(\Psi) > 0$. Finally, recall that, when event $\mathcal{E}$ holds, the primal-dual gap is equal to $\max_{\omega \in \mathcal{X}_\Omega} \widetilde{F}(\widehat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \widetilde{F}(\vartheta, \widehat{\omega})$. Combining (C.32), (C.47) with $\delta = T^{-5}$, and the fact that $\mathbb{P}(\mathcal{E}) \ge 1 - 2T^{-5}$, we conclude that

$$\begin{aligned} \mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega}) &\le C \cdot \log^4 T \cdot (1 + \|K\|_{\mathrm{fro}}^2)^4 \cdot \|\Sigma_K\|^4 \cdot \big[(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Omega + \widetilde{R}_\Theta\big]^2 \\ &\quad \cdot \bigg(\frac{\log^2 T + \log(T^5)}{\log(1/\rho) \cdot \sqrt{T}} + \frac{\sqrt{d+k}}{(1-\rho) \cdot T}\bigg) + \frac{2}{T} \\ &\le C \cdot (1 + \|K\|_{\mathrm{fro}}^2)^4 \cdot \|\Sigma_K\|^4 \cdot \big[(1 + \|K\|_{\mathrm{fro}}^2)^2 \cdot \widetilde{R}_\Omega + \widetilde{R}_\Theta\big]^2 \cdot \frac{\log^6 T}{(1-\rho) \cdot \sqrt{T}} \end{aligned} \tag{C.48}$$

holds with probability at least $1 - 3T^{-5} \ge 1 - T^{-4}$, where in the second inequality we use the fact that $1 - 1/x < \log x < x + 1$ holds for all $x > 0$, which implies that $1/\log(1/\rho) \le 1/(1-\rho)$. This further implies that the first term on the right-hand side of the first inequality dominates the second term. The upper bound of $\mathtt{Gap}(\widehat{\vartheta}, \widehat{\omega})$ in (C.48) concludes the last step of our proof. Finally, combining (C.17) and (C.48), we complete the proof of Theorem 4.2. □

C.2 Proof of Theorem 4.3

Proof. Our proof of the global convergence can be decomposed into two steps. In the first step, similar to the analysis in [26], we study the geometry of the average return $J(K)$, as a function of K. Specifically, we show that $J(K)$ is gradient dominated [51]. Note that we study the ergodic setting with system noise and stochastic policies. In contrast, [26] study the case where both the transition and the policy are deterministic. Thus, their analysis of the geometry of $J(K)$ cannot be directly applied to our problem. Motivated by their analysis, we follow the similar approach to with modifications for our setting. In addition, in the second step, we utilize the geometry of $J(K)$ to show the global convergence of the actor-critic algorithm. Specifically, combining Theorem 4.2, we show that, with high probability, Algorithm 1 constructs a sequence of policies that converges linearly to the optimal policy π_{K^*}.

Step 1. As shown in (3.8) in Proposition 3.1, we can write $J(K)$ as

$$J(K) = \mathrm{tr}(P_K \Psi_\sigma) + \sigma^2 \cdot \mathrm{tr}(R) = \mathbb{E}_{x \in N(0,\Psi_\sigma)}\big(x^\top P_K x\big) + \sigma^2 \cdot \mathrm{tr}(R).$$

In the following lemma, for two policies π_K and $\pi_{K'}$, we bound the difference between $x^\top P_K x$ and $x^\top P_{K'} x$. Then, taking expectation with respect to $x \in N(0, \Psi_\sigma)$ yields the difference between $J(K)$ and $J(K')$.

Lemma C.5. Let K and K' be two stable policies such that both $\rho(A - BK)$ and $\rho(A - BK')$ are smaller than one. For any $x \in \mathbb{R}^d$, let $\{x'_t\}_{t\geq 0} \subseteq \mathbb{R}^d$ be the sequence of states satisfying $x'_0 = x$ and $x'_{t+1} = (A - BK')x'_t$ for all $t \geq 0$. Then it holds that

$$x^\top P_{K'} x - x^\top P_K x = \sum_{t\geq 0} A_{K,K'}(x'_t),$$

where the function $A_{K,K'} \colon \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$A_{K,K'}(x) = 2x^\top (K' - K)^\top E_K x + x^\top (K' - K)^\top (R + B^\top P_K B)(K' - K)x.$$

Proof. Note that both P_K and $P_{K'}$ satisfy the Bellman equation specified in (3.4). Moreover, using the operator $\mathcal{T}_K^\top$ defined in (D.3), we have $P_{K'} = \mathcal{T}_{K'}^\top(Q + K'^\top R K')$, which is equivalent to

$$x^\top P_{K'} x = \sum_{t\geq 0} x^\top [(A - BK')^t]^\top \big(Q + K'^\top R K'\big)[(A - BK')^t]x. \tag{C.49}$$

By the construction in Lemma (C.5), for all $t \geq 0$, we have $(A - BK')^t x = x'_t$. Thus, by (C.49) we have

$$x^\top P_{K'} x = \sum_{t\geq 0} x_t'^\top \big(Q + K'^\top R K'\big) x'_t = \sum_{t\geq 0} \big(x_t'^\top Q x'_t + u_t'^\top R u'_t\big), \tag{C.50}$$

where we define $u'_t = -K'x'_t$ for all $t \geq 0$. Thus, by (C.50), we have the following telescoping sum:

$$\begin{aligned} x^\top P_{K'} x - x^\top P_K x &= \sum_{t\geq 0} \big[\big(x_t'^\top Q x'_t + u_t'^\top R u'_t\big) + x_t'^\top P_K x'_t - x_t'^\top P_K x'_t\big] - x_0'^\top P_K x'_0 \\ &= \sum_{t\geq 0} \big[\big(x_t'^\top Q x'_t + u_t'^\top R u'_t\big) + x_{t+1}'^\top P_K x'_{t+1} - x_t'^\top P_K x'_t\big]. \end{aligned} \tag{C.51}$$

Thus, in (C.51) we write $x^\top P_{K'} x - x^\top P_K x$ as a summation where each term can be written as a quadratic function of x_t. To further simplify (C.51), for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned} &x^\top Q x + (-K'x)^\top R(-K'x) + [(A - BK')x]^\top P_K[(A - BK')x] - x^\top P_K x \\ &\quad = x^\top \big[Q + (K' - K + K)^\top R(K' - K + K)\big]x + \\ &\qquad x^\top \big[A - BK - B(K' - K)\big]^\top P_K \big[A - BK - B(K' - K)\big]x - x^\top P_K x \\ &\quad = 2x^\top (K' - K)^\top \big[(R + B^\top P_K B)K - B^\top P_K A\big]x + x^\top (K' - K)^\top (R + B^\top P_K B)(K' - K)x. \\ &\quad = 2x^\top (K' - K)^\top E_K x + x^\top (K' - K)^\top (R + B^\top P_K B)(K' - K)x, \end{aligned} \tag{C.52}$$

where $E_K = (R + B^\top P_K B)K - B^\top P_K A$. Finally, combining (C.51) and (C.52), we complete the proof of this lemma. □

In the following lemma, we utilize Lemma C.5 to show that $J(K)$ is gradient dominated.

Lemma C.6 (Gradient domination of $J(K)$)**.** Let K^* be an optimal policy. Suppose K has finite cost. Then, it holds that

$$\sigma_{\min}(\Psi) \cdot \|R + B^\top P_K B\|^{-1} \cdot \mathrm{tr}(E_K^\top E_K) \le J(K) - J(K^*) \le 1/\sigma_{\min}(R) \cdot \|\Sigma_{K^*}\| \cdot \mathrm{tr}(E_K^\top E_K). \tag{C.53}$$

Proof. For the upper bound in (C.53), bu (3.8) we obtain that

$$J(K) - J(K^*) = \mathrm{tr}[(P_K - P_K^*)\Psi_\sigma] = \mathbb{E}_{x\sim N(0,\Psi_\sigma)}\big[x^\top (P_K - P_K^*)x\big], \tag{C.54}$$

where $\Psi_\sigma = \Psi + \sigma^2 BB^\top$ does not involve K or K^*. Applying Lemma C.5 to (C.54) with $K' = K^*$, we have

$$J(K) - J(K^*) = -\mathbb{E}_{x_0^*\sim N(0,\Psi_\sigma)}\bigg[\sum_{t\ge 0} A_{K,K^*}(x_t^*)\bigg], \tag{C.55}$$

where we define $x_t^* = (A - BK^*)^t x_0^*$ for all $t \ge 0$. Besides, by direct computation, we have

$$\begin{aligned}&\mathbb{E}_{x_0^*\sim N(0,\Psi_\sigma)}\bigg[\sum_{t\ge 0} x_t^*(x_t^*)^\top\bigg] \\ &\quad= \mathbb{E}_{x\sim N(0,\Psi_\sigma)}\bigg\{\sum_{t\ge 0}(A - BK^*)^t xx^\top[(A - BK^*)^t]^\top\bigg\} = \mathcal{T}_{K^*}(\Psi_\sigma) = \Sigma_{K^*},\end{aligned} \tag{C.56}$$

where the operator $\mathcal{T}_K$ is defined in (D.3).

Meanwhile, by the definition of $A_{K,K'}$, for any $x \in \mathbb{R}^d$, by completing the squares we have

$$\begin{aligned}A_{K,K'}(x) &= 2x^\top (K' - K)^\top E_K x + x^\top (K' - K)^\top (R + B^\top P_K B)(K' - K)x \\ &= \mathrm{tr}\Big\{xx^\top \big[K' - K + (R + B^\top P_K B)^{-1} E_K\big]^\top (R + B^\top P_K B)\big[K' - K + (R + B^\top P_K B)^{-1} E_K\big]\Big\} \\ &\qquad - \mathrm{tr}\big[xx^\top E_K^\top (R + B^\top P_K B)^{-1} E_K\big] \\ &\ge -\mathrm{tr}\big[xx^\top E_K^\top (R + B^\top P_K B)^{-1} E_K\big],\end{aligned} \tag{C.57}$$

where the equality is attained by $K' = K - (R + B^\top P_K B)^{-1} E_K$.

Thus, combining (C.55), (C.56), and (C.57), we obtain that

$$\begin{aligned}J(K) - J(K^*) &\le \mathrm{tr}\big[\Sigma_{K^*} E_K^\top (R + B^\top P_K B)^{-1} E_K\big] \le \|\Sigma_{K^*}\| \cdot \mathrm{tr}\big[\Sigma_{K^*} E_K^\top (R + B^\top P_K B)^{-1} E_K\big] \\ &\le \|\Sigma_{K^*}\| \cdot \|(R + B^\top P_K B)^{-1}\| \cdot \mathrm{tr}(E_K^\top E_K).\end{aligned} \tag{C.58}$$

Notice that $R + B^\top P_K B \succeq R$ implies $(R + B^\top P_K B)^{-1} \preceq R^{-1}$. Therefore, by (C.58) we obtain that $J(K) - J(K^*) \le 1/\sigma_{\min}(R) \cdot \|\Sigma_{K^*}\| \cdot \mathrm{tr}(E_K^\top E_K)$, which establishes the upper bound in (C.53).

Furthermore, for the lower bound, since $K' = K - (R + B^\top P_K B)^{-1} E_K$ attains the lower bound in (C.57) and K^* is the optimal policy, similar to (C.55) and (C.56), we have

$$\begin{aligned}J(K) - J(K^*) \ge J(K) - J(K') &= -\mathbb{E}_{x_0^*\sim N(0,\Psi_\sigma)}\bigg[\sum_{t\ge 0} A_{K,K'}(x_t')\bigg] \\ &= \mathrm{tr}\big[\Sigma_{K'} E_K^\top (R + B^\top P_K B)^{-1} E_K\big] \ge \sigma_{\min}(\Psi) \cdot \|R + B^\top P_K B\|^{-1} \cdot \mathrm{tr}(E_K^\top E_K),\end{aligned}$$

where in the first equality we define $x_t' = (A - BK')^t$ for all $t \ge 0$, and the last inequality follows from the fact that $\Sigma_{K'} \succeq \Psi \succeq \sigma_{\min}(\Psi) \cdot I_d$. Therefore, we conclude the proof of Lemma C.6. □

Notice that $K = K^*$ achieves the minimum of $J(K)$. Lemma C.6 implies that

$$J(K) - J(K^*) \le \lambda \cdot \langle E_K, E_K\rangle,$$

where $\lambda = 1/\sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|$. That is, the difference of the objective can be bounded by the norm of the natural gradient. Therefore, updating the policy parameter K in the direction of natural gradient E_K yields decreases the objective value. Therefore, we conclude the first step.

Step 2. In the second part of the proof, equipped with Lemma C.6, we establish the global convergence of the natural actor-critic algorithm. Recall that we assume that the initial policy π_{K_0} is stable, which implies that $J(K_0)$ is finite. Moreover, according to Algorithm 1, the policy parameters are updated via

$$K_{t+1} = K_t - \gamma \cdot \widehat{E}_{K_t}, \qquad \widehat{E}_{K_t} = \widehat{\Theta}_t^{22} K_t - \widehat{\Theta}_t^{21}, \tag{C.59}$$

where $\widehat{\Theta}_t$ is the estimator of Θ_{K_t} returned by Algorithm 2.

We use mathematical induction to show that $\{J(K_t)\}_{t\geq 0}$ is a monotone decreasing sequence. Suppose $J(K_t) \leq J(K_0)$. We define $K'_{t+1} = K_t - \gamma \cdot E_{K_t}$, i.e., K'_{t+1} is obtained by a single step of natural policy gradient, starting from K_t. In the sequel, we use $J(K'_{t+1})$ to connect $J(K_t)$ and $J(K_{t+1})$. By Lemma C.5, we have

$$\begin{aligned} J(K'_{t+1}) - J(K_t) &= \mathbb{E}_{x\sim N(0,\Psi_\sigma)}[x^\top (P_{K'_{t+1}} - P_{K_t})x] \\ &= -2\gamma \cdot \mathrm{tr}\big(\Sigma_{K'_{t+1}} \cdot E_{K_t}^\top E_{K_t}\big) + \gamma^2 \cdot \mathrm{tr}\big[\Sigma_{K'_{t+1}} \cdot E_{K_t}^\top (R + B^\top P_{K_t} B) E_{K_t}\big] \\ &= -2\gamma \cdot \mathrm{tr}\big(\Sigma_{K'_{t+1}} \cdot E_{K_t}^\top E_{K_t}\big) + \gamma^2 \cdot \|R + B^\top P_{K_t} B\| \cdot \mathrm{tr}\big(\Sigma_{K'_{t+1}} \cdot E_{K_t}^\top E_{K_t}\big). \end{aligned} \tag{C.60}$$

When γ is sufficiently small such that

$$\gamma \cdot \big[\|R\| + \sigma_{\min}^{-1}(\Psi) \cdot \|B\|^2 \cdot J(K_0)\big] \leq 1, \tag{C.61}$$

by triangle inequality, we have

$$\gamma \cdot \|R + B^\top P_{K_t} B\| \leq \gamma \cdot \big[\|R\| + \|B\|^2 \cdot \|P_{K_t}\|\big] \leq \gamma \cdot \big[\|R\| + \sigma_{\min}^{-1}(\Psi) \cdot \|B\|^2 \cdot J(K_0)\big] < 1, \tag{C.62}$$

where the second inequality follows from Lemma C.1 and the induction assumption that $J(K_t) \leq J(K_0)$, and the last inequality follows from (C.61). Thus, combining (C.60) and (C.62), we have

$$\begin{aligned} J(K'_{t+1}) - J(K_t) &\leq -\gamma \cdot \mathrm{tr}\big(\Sigma_{K'_{t+1}} \cdot E_{K_t}^\top E_{K_t}\big) \leq -\gamma \cdot \sigma_{\min}(\Psi) \cdot \mathrm{tr}\big(E_{K_t}^\top E_{K_t}\big), \\ &\leq -\gamma \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1} \cdot \big[J(K_t) - J(K^*)\big]. \end{aligned} \tag{C.63}$$

where the third inequality follows from the fact that $\Sigma_{K'_{t+1}} \succeq \Psi$, and the last inequality follows from Lemma C.6. Note that (C.63) implies that $J(K'_{t+1}) \leq J(K_t) \leq J(K_0)$.

Furthermore, by the difference between $J(K_{t+1})$ and $J(K'_{t+1})$ can be bounded by

$$\begin{aligned} \big|J(K_{t+1}) - J(K'_{t+1})\big| &= \big|\mathrm{tr}\big[(P_{K_{t+1}} - P_{K'_{t+1}}) \cdot \Psi_\sigma\big]\big| \leq \|\Psi_\sigma\|_{\mathrm{fro}} \cdot \big\|P_{K_{t+1}} - P_{K'_{t+1}}\big\| \\ &\leq \big[\|\Psi\|_{\mathrm{fro}} \cdot + \sigma^2 \cdot \|B\|_{\mathrm{fro}}^2\big] \cdot \big\|P_{K_{t+1}} - P_{K'_{t+1}}\big\|. \end{aligned} \tag{C.64}$$

Now we utilize the following Lemma, obtained from [26], to construct and upper bound for $\|P_{K_{t+1}} - P_{K'_{t+1}}\|$.

Lemma C.7 (Perturbation of P_K)**.** Suppose $\pi_{K'}$ is a small perturbation of π_K in the sense that

$$\|K' - K\| \leq \sigma_{\min}(\Psi)/4 \cdot \|\Sigma_K\|^{-1} \|B\|^{-1} \cdot (\|A - BK\| + 1)^{-1}, \tag{C.65}$$

then we have

$$\begin{aligned} \|P_{K'} - P_K\| \leq\ & 6\sigma_{\min}^{-1}(\Psi) \cdot \|\Sigma_K\| \cdot \|K\| \cdot \|R\| \\ & \cdot \big(\|K\| \cdot \|B\| \cdot \|A - BK\| + \|K\| \cdot \|B\| + 1\big) \cdot \|K - K'\|. \end{aligned} \tag{C.66}$$

Proof. This lemma is a slight modification of Lemma 24 in [26]. Here we sketch the proof. See [26, Lemmas 17 and 24] for a detailed proof.

Recall that we define operator $\mathcal{T}_K$ in (D.3). The operator norm of $\mathcal{T}_K$ is defined as $\|\mathcal{T}_K\| \leq \sup_\Omega \|\mathcal{T}_K(\Omega)\|/\|\Omega\|$, where the supremum is taken over all symmetric matrices. As shown in Lemma 17 in [26], we have $\|\mathcal{T}_K\| \leq \sigma_{\min}^{-1}(\Psi) \cdot \|\Sigma_K\|$. Moreover, under the condition in (C.65), in the proof of Lemma 24 in [26], it is shown that

$$\|P_{K'} - P_K\| \leq 6\|\mathcal{T}_K\| \cdot \|K\| \cdot \|R\| \cdot \big(\|K\| \cdot \|B\| \cdot \|A - BK\| + \|K\| \cdot \|B\| + 1\big) \cdot \|K - K'\|.$$

Combining this with the upper bound on $\|\mathcal{T}_K\|$, we conclude the proof. □

To use this lemma, we need to verify (C.65). That is,

$$4\|K_{t+1} - K'_{t+1}\| \cdot (1 + \|A - BK'_{t+1}\|) \cdot \|B\| \cdot \|\Sigma_{K'_{t+1}}\| \leq \sigma_{\min}(\Psi). \tag{C.67}$$

By the definition of K_{t+1} and K'_{t+1}, we have

$$\|K_{t+1} - K'_{t+1}\| = \gamma \cdot \|\widehat{E}_{K_t} - E_{K_t}\| \leq \gamma \cdot \|\widehat{\Theta}_t - \Theta_{K_t}\|_{\text{fro}} \cdot (1 + \|K_t\|), \tag{C.68}$$

where $\widehat{E}_{K_t}$ is defined in (C.59). Plugging (C.68) into the left-hand side of (C.67), we obtain that

$$\begin{aligned} &4\|K_{t+1} - K'_{t+1}\| \cdot (1 + \|A - BK'_{t+1}\|) \cdot \|B\| \cdot \|\Sigma_{K'_{t+1}}\| \\ &\quad \leq 4\gamma \cdot \|\widehat{\Theta}_t - \Theta_{K_t}\|_{\text{fro}} \cdot (1 + \|K_t\|) \cdot (1 + \|A - BK'_{t+1}\|) \cdot \|B\| \cdot \|\Sigma_{K'_{t+1}}\|. \end{aligned} \tag{C.69}$$

Utilizing Lemma (C.1) and the fact that $J(K'_{t+1}) \leq J(K_0)$, we have

$$\|\Sigma_{K'_{t+1}}\| \leq J(K'_{t+1})/\sigma_{\min}(Q) \leq J(K_0)/\sigma_{\min}(Q). \tag{C.70}$$

In addition, by triangle inequality, we have

$$\begin{aligned} \|A - BK'_{t+1}\| &\leq \|A - BK_t\| + \gamma \cdot \|B\| \cdot \|E_{K_t}\| \\ &\leq \|A - BK_t\| + \gamma \cdot \|B\| \cdot \|\Theta_{K_t}\| \cdot (1 + \|K_t\|). \end{aligned} \tag{C.71}$$

By the definition of Θ_K in (3.7), we have

$$\begin{aligned} \|\Theta_{K_t}\| &\leq \|Q\| + \|R\| + (\|A\|_{\text{fro}} + \|B\|_{\text{fro}})^2 \cdot \|P_{K_t}\| \\ &\leq \|Q\| + \|R\| + (\|A\|_{\text{fro}} + \|B\|_{\text{fro}})^2 \cdot J(K_0)/\sigma_{\min}(\Psi), \end{aligned} \tag{C.72}$$

where the last inequality follows from Lemma (C.1) and the induction assumption. Furthermore, by triangle inequality, it holds that

$$\begin{aligned} \|K_{t+1}\| &\leq \|K_t\| + \gamma \cdot \|E_{K_t}\| \leq \|K_t\| + \gamma \cdot \|\Theta_{K_t}\| \cdot (1 + \|K_t\|) \\ &\leq \|K_t\| + \gamma \cdot \big[\|Q\| + \|R\| + (\|A\|_{\text{fro}} + \|B\|_{\text{fro}})^2 \cdot J(K_0)/\sigma_{\min}(\Psi)\big] \cdot (1 + \|K_t\|). \end{aligned} \tag{C.73}$$

In the sequel, we set

$$\gamma = \big[\|R\| + \sigma_{\min}^{-1}(\Psi) \cdot \|B\|^2 \cdot J(K_0)\big]^{-1}. \tag{C.74}$$

Note that we assume that $\|Q\|$, $\|R\|$, $\|A\|$, $\|B\|$, $\sigma_{\min}(Q)$, $\sigma_{\min}(R)$ are all constants. Combining (C.69), (C.70), (C.71), and (C.72), we conclude that there exists a polynomial $\Upsilon_1(\cdot, \cdot)$ such that

$$4\|K_{t+1} - K'_{t+1}\| \cdot (1 + \|A - BK'_{t+1}\|) \cdot \|B\| \cdot \|\Sigma_{K'_{t+1}}\| \leq \Upsilon_1\big[\|K_t\|, J(K_0)\big] \cdot \|\widehat{\Theta}_t - \Theta_{K_t}\|_{\text{fro}}. \tag{C.75}$$

Furthermore, for the right-hand side of (C.66), combining (C.68), (C.69), (C.70), (C.71), (C.72), and (C.73). we conclude that there exists a polynomial $\Upsilon_2(\cdot, \cdot)$ such that

$$\begin{aligned} &\big[\|\Psi\|_{\text{fro}} \cdot + \sigma^2 \cdot \|B\|_{\text{fro}}^2\big] \cdot 6\sigma_{\min}^{-1}(\Psi) \cdot \|\Sigma_{K'_{t+1}}\| \cdot \|K_{t+1'}\| \cdot \|R\| \\ &\qquad \cdot \big(\|K'_{t+1}\| \cdot \|B\| \cdot \|A - BK'_{t+1}\| + \|K'_{t+1}\| \cdot \|B\| + 1\big) \cdot \|K_{t+1} - K'_{t+1}\| \\ &\leq \Upsilon_2\big[\|K_t\|, J(K_0)\big] \cdot \|\widehat{\Theta}_t - \Theta_{K_t}\|_{\text{fro}}. \end{aligned} \tag{C.76}$$

Meanwhile, in Theorem 4.2 we have shown that, there exists a polynomial $\Upsilon_3(\cdot, \cdot)$ such that, for T sufficiently large, Algorithm 2 with T iterations returns an estimator $\widehat{\Theta}_t$ for Θ_{K_t} such that

$$\|\widehat{\Theta}_t - \Theta_{K_t}\|_{\text{fro}} \leq \frac{\Upsilon_3\big[\|K_t\|, J(K_0)\big]}{\kappa^*_{K_t} \cdot \sqrt{(1-\rho)}} \cdot \frac{\log^3 T}{T^{1/4}} \tag{C.77}$$

holds with probability at least $1 - T^{-4}$, where $\rho \in (\rho(A - BK_t), 1)$ and $\kappa^*_{K_t}$ is specified in Lemma 3.2, which depends only on ρ, σ, and $\sigma_{\min}(\Psi)$. Notice that $\log^3 T \cdot T^{-1/4} \leq T^{-1/5}$ for T sufficiently

large. Therefore, in the GTD algorithm for estimating Θ_{K_t}, we set the number of iterations T_t sufficiently large such that

$$\begin{aligned}
&\Upsilon_1\big[\|K_t\|, J(K_0)\big] \cdot \Upsilon_3\big[\|K_t\|, J(K_0)\big] \cdot {\kappa^*_{K_t}}^{-1} \cdot (1-\rho)^{-1/2} \cdot T_t^{-1/5} \le \sigma_{\min}(\Psi), \\
&\Upsilon_2\big[\|K_t\|, J(K_0)\big] \cdot \Upsilon_3\big[\|K_t\|, J(K_0)\big] \cdot {\kappa^*_{K_t}}^{-1} \cdot (1-\rho)^{-1/2} \cdot T_t^{-1/5} \\
&\qquad \le \epsilon/2 \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1}
\end{aligned} \tag{C.78}$$

hold simultaneously. For such a T_t, combining (C.75) and (C.77), we conclude that (C.67) holds. Lemma C.7 implies that (C.66) is true. Combining (C.64), (C.66), (C.76), and (C.77), we conclude that

$$\big|J(K_{t+1}) - J(K'_{t+1})\big| \le \epsilon/2 \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1} \tag{C.79}$$

holds with probability at least $1 - T_t^{-4}$. Thus, when $J(K_t) - J(K^*) > \epsilon$, combining (C.63) and (C.79) we have

$$J(K_{t+1}) - J(K_t) \le -\epsilon/2 \cdot \gamma\sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1} < 0.$$

Therefore, we have shown that, as long as $J(K_t) - J(K^*) \ge \epsilon$, $J(K_{t+1}) < J(K_t)$ holds with probability at least $1 - T_t^{-1/4}$.

Meanwhile, (C.63) implies that,

$$J(K'_{t+1}) - J(K^*) \le \big[1 - \gamma \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1}\big] \cdot \big[J(K_t) - J(K^*)\big]$$

By (C.79), when $J(K_t) - J(K^*) \ge \epsilon$, with probability $1 - T_t^{-4}$, we have

$$J(K_{t+1}) - J(K^*) \le \big[1 - \gamma/2 \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K^*}\|^{-1}\big] \cdot \big[J(K_t) - J(K^*)\big],$$

which shows that, in terms of the policy parameter, natural actor-critic algorithm converges linearly. Specifically, with

$$N \ge 2\|\Sigma_{K^*}\|/\gamma \cdot \sigma^{-1}_{\min}(\Psi) \cdot \sigma^{-1}_{\min}(R) \cdot \log\big\{2[J(K_0) - J(K^*)]/\epsilon\big\} \tag{C.80}$$

policy updates, we have $J(K_N) - J(K^*) \le \epsilon$ with high-probability, where γ is specified in (C.74).

Finally, it remains to determine T_t for all $t \in [N]$. Notice that T_t satisfies the two inequalities in (C.78). Thus, we set

$$T_t \ge \Upsilon_4[\|K_t\|, J(K_0)] \cdot {\kappa^*_{K_t}}^{-5} \cdot (\Xi_{K_t}) \cdot \big[1 - \rho(A - BK_t)\big]^{-5/2} \cdot \epsilon^{-5}$$

for some polynomial function $\Upsilon_4(\cdot,\cdot)$. With such a T_t, the fail probability $T_t^{-4} \le \epsilon^{-20}$. Notice that the total number of iterations depends on ϵ only through $\log(1/\epsilon)$. Thus, the total fail probability can be bounded by ϵ^{10}. Therefore, we conclude the proof. □

D Proofs of the Auxiliary Results

In this section, we provides the proofs for Proposition 3.1 and Lemma 3.2.

D.1 Proof of Proposition 3.1

Proof. We first establish (3.8). Note that under π_K, we can write u_t as $-Kx_t + \sigma \cdot \eta_t$, where $\eta_t \sim N(0, I_d)$. This implies that, for all ≥ 0, we have

$$\begin{aligned}
\mathbb{E}[c(x_t, u_t) \,|\, x_t] &= x_t^\top Q x_t + \mathbb{E}_{\eta_t \sim N(0,I_d)}[(-Kx_t + \sigma \cdot \eta_t)^\top R(-Kx_t + \sigma \cdot \eta_t)] \\
&= x_t^\top (Q + K^\top R K) x_t + \sigma^2 \cdot \mathrm{tr}(R).
\end{aligned} \tag{D.1}$$

Thus, combining (D.1) and the definition of $J(K)$ in (2.1), we have

$$\begin{aligned}
J(K) &= \lim_{T\to\infty} \mathbb{E}\bigg\{\frac{1}{T}\sum_{t\ge 0}^{T} \mathbb{E}[c(x_t, u_t) \,|\, x_t]\bigg\} = \lim_{T\to\infty} \mathbb{E}\bigg\{\frac{1}{T}\sum_{t\ge 0}^{T} [x_t^\top (Q + K^\top R K) x_t + \sigma^2 \cdot \mathrm{tr}(R)]\bigg\} \\
&= \mathbb{E}_{x\sim\nu_K}[x^\top (Q + K^\top R K) x] + \sigma^2 \cdot \mathrm{tr}(R) = \mathrm{tr}\big[(Q + K^\top R K)\Sigma_K\big] + \sigma^2 \cdot \mathrm{tr}(R),
\end{aligned} \tag{D.2}$$

where the third inequality in (D.2) holds because the limiting distribution of $\{x_t\}_{t\geq 0}$ is ν_K.

It remains to establish the second equality in (3.8). To this end, for $K \in \mathbb{R}^{k\times d}$ such that $\rho(A-BK) < 1$, we define operators we define $\mathcal{T}_K$ and $\mathcal{T}_K^\top$ by

$$\mathcal{T}_K(\Omega) = \sum_{t\geq 0}(A-BK)^t\Omega\big[(A-BK)^t\big]^\top, \qquad \mathcal{T}_K^\top(\Omega) = \sum_{t\geq 0}\big[(A-BK)^t\big]^\top\Omega(A-BK)^t, \tag{D.3}$$

where $\Omega \in \mathbb{R}^{d\times d}$ is positive definite. By definition, $\mathcal{T}_K(\Omega)$ and $\mathcal{T}_K^\top(\Omega)$ satisfy Lyapunov equations

$$\mathcal{T}_K(\Omega) = \Omega + (A-BK)\mathcal{T}_K(\Omega)(A-BK)^\top, \tag{D.4}$$

$$\mathcal{T}_K^\top(\Omega) = \Omega + (A-BK)^\top\mathcal{T}_K^\top(\Omega)(A-BK), \tag{D.5}$$

respectively. Moreover, for any positive definite matrices Ω_1, Ω_2, since $\rho(A-BK) < 1$, we have

$$\begin{aligned}\mathrm{tr}[\Omega_1\cdot\mathcal{T}_K(\Omega_2)] &= \sum_{t\geq 0}\mathrm{tr}\big\{\Omega_1(A-BK)^t\Omega_2[(A-BK)^t]^\top\big\}\\ &= \sum_{t\geq 0}\mathrm{tr}\big\{[(A-BK)^t]^\top\Omega_1(A-BK)^t\Omega_2\big\} = \mathrm{tr}[\mathcal{T}_K^\top(\Omega_1)\cdot\Omega_2].\end{aligned} \tag{D.6}$$

Meanwhile, by combining (3.3), (3.4), (D.4), and (D.5), we have $\Sigma_K = \mathcal{T}_K(\Psi_\sigma)$ and$P_K = \mathcal{T}_K^\top(Q + K^\top RK)$. Thus, (D.6) implies that

$$\mathrm{tr}\big[(Q+K^\top RK)\cdot\Sigma_K\big] = \mathrm{tr}\big[(Q+K^\top RK)\cdot\mathcal{T}_K(\Psi_\sigma)\big] = \mathrm{tr}\big[\mathcal{T}_K^\top(Q+K^\top RK)\cdot\Psi_\sigma\big] = \mathrm{tr}(P_K\Psi_\sigma).$$

Combining this equation with (D.2), we establish the second equation of (3.8).

In the following, we establish the value functions. In the setting of LQR, the state-value function V_K is given by

$$\begin{aligned}V_K(x) &= \sum_{t=0}^{\infty}\big\{\mathbb{E}[c(x_t,u_t)\,|\,x_0 = x, u_t = -Kx_t + \sigma\cdot\eta_t] - J(K)\big\}\\ &= \sum_{t=0}^{\infty}\big\{\mathbb{E}[x_t^\top(Q+K^\top RK)x_t] + \sigma^2\cdot\mathrm{tr}(R) - J(K)\big\}.\end{aligned} \tag{D.7}$$

Combining the linear dynamics in (3.2) and (D.7), we see that V_K is a quadratic function, which is denoted by $V_k(x) = x^\top P_K x + \alpha_K$, where both P_K and α_K depends on K. Note that V_K satisfies the Bellman equation

$$V_K(x) = \mathbb{E}_{u\sim\pi_K}[c(x,u)] - J(K) + \mathbb{E}[V_K(x')\,|\,x],$$

where x' is the next state given (x,u). Thus, for any $x\in\mathbb{R}^d$, we have

$$x^\top P_K x = x(Q+K^\top RK)x + x^\top(A-BK)^\top P_K(A-BK)x.$$

Thus, P_K is the unique positive definite solution to the Bellman equation in (3.4). Meanwhile, since $\mathbb{E}_{x\sim\nu_K}[V_K(x)] = 0$, we have $\alpha_K = -\mathrm{tr}(P_K\Sigma_K)$. Hence, we establish (3.5).

Furthermore, for any state-action pair (x,u), we have

$$\begin{aligned}Q_K(x,u) &= c(x,u) - J(K) + \mathbb{E}[V_K(x')\,|\,x,u]\\ &= c(x,u) - J(K) + (Ax+Bu)^\top P_K(Ax+Bu) + \mathrm{tr}(P_K\Psi) - \mathrm{tr}(P_K\Sigma_K)\\ &= x^\top Qx + u^\top Ru + (Ax+Bu)^\top P_K(Ax+Bu) - \sigma^2\cdot\mathrm{tr}(R + P_KBB^\top) - \mathrm{tr}(P_K\Sigma_K),\end{aligned}$$

where x' in the first equality is the next state following (x,u), and the last equality follows from (3.8) and the fact that $\Psi_\sigma = \Psi + \sigma^2\cdot BB^\top$. Thus, we prove (3.6).

It remains to derive the policy gradient $\nabla_K J(K)$. By (3.8), we have

$$\nabla_K J(K) = 2RK\Sigma_K + \nabla_K\,\mathrm{tr}(Q_0\cdot\Sigma_K)\big|_{Q_0 = Q+K^\top RK}, \tag{D.8}$$

where the second term denotes that we first take compute the gradient $\nabla_K\,\mathrm{tr}[Q_0\Sigma_K]$ with respect to K and then set $Q_0 = Q + K^\top RK$. Recall that we can write $\Sigma_K = \mathcal{T}_K(\Psi_\sigma)$. The following lemma enables us to compute the gradient involving $\mathcal{T}_K$.

Lemma D.1. Let W and Ψ be two positive definite matrices. Then it holds that

$$\nabla_K \operatorname{tr}\big[W \cdot \mathcal{T}_K(\Psi)\big] = -2B^\top \mathcal{T}_K^\top(W)(A - BK)\mathcal{T}_K(\Psi).$$

Proof. To simplify the notation, we define operator $\mathcal{F}_K$ by

$$\mathcal{F}_K^\top(\Omega) = (A - BK)^\top \Omega (A - BK)$$

and let $\mathcal{F}_K^{\top,t}$ be the t-th composition of $\mathcal{F}_K$. Thus, by the definition of $\mathcal{T}_K^\top$ and $\mathcal{F}_K^\top$, we have

$$\mathcal{T}_K^\top(\Omega) = \sum_{t\geq 0} \mathcal{F}_K^{\top,t}(\Omega).$$

Moreover, by (D.4) we have

$$\operatorname{tr}\big[W \cdot \mathcal{T}_K(\Psi)\big] = \operatorname{tr}(W\Psi) + \operatorname{tr}\big[(A - BK)^\top W (A - BK) \cdot \mathcal{T}_K(\Psi)\big],$$

which implies that

$$\nabla_K \operatorname{tr}\big[W \cdot \mathcal{T}_K(\Psi)\big] = -2B^\top W(A - BK)\mathcal{T}_K(\Psi) + \nabla_K \operatorname{tr}[W_1 \mathcal{T}_K(\Psi)]\Big|_{W_1 = \mathcal{F}_K(\Omega)}. \tag{D.9}$$

For any $k \geq 1$, by recursively applying (D.9) for k times, we have

$$\begin{aligned}
&\nabla_K \operatorname{tr}\big[W \cdot \mathcal{T}_K(\Psi)\big] \\
&\quad = -2B^\top \bigg[\sum_{t=0}^{k} \mathcal{F}_K^{\top,t}(W)\bigg](A - BK)\mathcal{T}_K(\Psi) + \nabla_K \operatorname{tr}[W_1 \mathcal{T}_K(\Psi)]\Big|_{W_1 = \mathcal{F}_K^{(k+1)}(\Omega)}.
\end{aligned} \tag{D.10}$$

Meanwhile, since $\rho(A - BK) < 1$, we have

$$\lim_{k\to\infty} \operatorname{tr}\big[\mathcal{F}_K^{\top,k}(W)\mathcal{T}_K(\Psi)\big] \leq \lim_{k\to\infty} \|W\| \cdot \operatorname{tr}[\mathcal{T}_K(\Psi)] \cdot \rho(A - BK)^{2k} = 0.$$

Thus, by letting k on the right-hand side of (D.10) go to infinity, we obtain

$$\nabla_K \operatorname{tr}\big[W \cdot \mathcal{T}_K(\Psi)\big] = -2B^\top \bigg[\sum_{t=0}^{\infty} \mathcal{F}_K^{\top,t}(W)\bigg](A - BK)\mathcal{T}_K(\Psi) = -2B^\top \mathcal{T}_K^\top(W)(A - BK)\mathcal{T}_K(\Psi).$$

Therefore, we conclude the proof of the lemma. □

By the above lemma, since $\Sigma_K = \mathcal{T}_K(\Psi_\sigma)$, we have

$$\begin{aligned}
\nabla_K \operatorname{tr}(Q_0 \cdot \Sigma_K)\big|_{Q_0 = Q + K^\top RK} &= \nabla_K \operatorname{tr}\big[Q_0 \cdot \mathcal{T}_K(\Psi_\sigma)\big]\Big|_{Q_0 = Q + K^\top RK} \\
&= -2B^\top \mathcal{T}_K^\top(Q + K^\top RK)(A - BK)\mathcal{T}_K(\Psi_\sigma) = -2B^\top P_K(A - BK)\Sigma_K,
\end{aligned} \tag{D.11}$$

where we use the fact that $P_K = \mathcal{T}_K^\top(Q + K^\top RK)$. Therefore, combining (D.8) and (D.11), we establish (3.9), which completes the proof of Proposition 3.1. □

D.2 Proof of Lemma 3.2

We present a stronger lemma than Lemma 3.2, whose proof automatically validates Lemma 3.2.

Lemma D.2. Suppose $\rho(A - BK) < 1$. Let $N(0, \widetilde{\Sigma}_K)$ be the stationary distribution of the state-action pair (x, u) when following policy π_K. Then for Ξ_K defined in (3.15), we have

$$\Xi_K = \big(\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K\big) - \big(\widetilde{\Sigma}_K L^\top\big) \otimes_s \big(\widetilde{\Sigma}_K L^\top\big) = \big(\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K\big)\big(I - L^\top \otimes_s L^\top\big). \tag{D.12}$$

Moreover, Ξ_K is a invertible matrix whose operator norm is bounded by $2[\sigma^2 + (1 + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\|]$. There exists a positive number κ_K^* such that the minimum singular value of the matrix in the left-hand side of (3.16) is lower bounded by a constant $\kappa_K^* > 0$, where κ_K^* only depends on $\rho(A - BK)$, σ, and $\sigma_{\min}(\Psi)$. Furthermore, since Ξ_K is invertible, the linear equation in (3.16) has unique solution ϑ_K^*, whose first and second components are $J(K)$ and $\operatorname{svec}(\Theta_K)$, respectively.

Proof. Throughout the proof of Lemma D.2, for any state-action pair $(x, u) \in \mathbb{R}^{d+k}$, we denote the next state-action pair following policy π_K by (x', u'). Then we can write

$$x' = Ax + Bu + \epsilon, \qquad u' = -Kx' + \sigma \cdot \eta = -KAx - KBu - K\epsilon + \sigma \cdot \eta, \tag{D.13}$$

where $\epsilon \sim N(0, \Psi)$ and $\eta \in N(0, I_k)$. For notational simplicity, we denote (x, u) and (x', u') by z and z', respectively. Thus, we can write $z' = Lz + \varepsilon$, where we define

$$L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix} = \begin{pmatrix} I_d \\ -K \end{pmatrix} (A \quad B), \qquad \varepsilon = \begin{pmatrix} \epsilon \\ -K\epsilon + \sigma \cdot \eta \end{pmatrix}. \tag{D.14}$$

Since it holds that $\rho(MN) = \rho(NM)$ for any two matrices M and N [32, Theorem 1.3.22], we have $\rho(L) = \rho(A - BK) < 1$. Meanwhile, by definition, $\varepsilon \in \mathbb{R}^{d+k}$ is a centered Gaussian random variable with covariance

$$\begin{pmatrix} \Psi & -\Psi K^\top \\ -K\Psi & K\Psi K^\top + \sigma^2 \cdot I_k \end{pmatrix}, \tag{D.15}$$

which is denoted by $\widetilde{\Psi}_\sigma$ for notational simplicity. In addition, for $x \sim \nu_K$ and $u \sim \pi_K(\cdot \,|\, x)$, we denote the joint distribution of $z = (x, u)$ by $\widetilde{\nu}_K$, which is a centered Gaussian distribution in $\mathbb{R}^{d\times k}$. Since $x \sim N(0, \Sigma_K)$ and $u = -Kx + \sigma \cdot I_k$, we can write $\widetilde{\nu}_K$ as $N(0, \widetilde{\Sigma}_K)$, where $\widetilde{\Sigma}_K \in \mathbb{R}^{(d+k)\times(d+k)}$ can be written as

$$\widetilde{\Sigma}_K = \begin{pmatrix} \Sigma_K & -\Sigma_K K^\top \\ -K\Sigma_K & K\Sigma_K K^\top + \sigma^2 \cdot I_k \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 \cdot I_k \end{pmatrix} + \begin{pmatrix} I_d \\ -K \end{pmatrix} \Sigma_K \begin{pmatrix} I_d \\ -K \end{pmatrix}^\top. \tag{D.16}$$

Thus, by triangle inequality we have

$$\big\|\widetilde{\Sigma}_K\big\|_{\mathrm{fro}} \le \sigma^2 \cdot k + \|\Sigma_K\| \cdot (d + \|K\|_{\mathrm{fro}}^2), \qquad \big\|\widetilde{\Sigma}_K\big\| \le \sigma^2 + (1 + \|K\|_{\mathrm{fro}}^2) \cdot \|\Sigma_K\|, \tag{D.17}$$

where in (D.17) we use the fact that $\|AB\|_{\mathrm{fro}} \le \|A\|_{\mathrm{fro}} \cdot \|B\|$.

Furthermore, since L defined in (D.14) satisfy $\rho(L) < 1$, $\widetilde{\Sigma}_K$ is the unique positive definite solution to the Lyapunov equation

$$\widetilde{\Sigma}_K = L\widetilde{\Sigma}_K L^\top + \widetilde{\Psi}_K, \tag{D.18}$$

where $\widetilde{\Psi}_K$ is defined in (D.15). Moreover, the feature mapping can be written as $\phi(x, u) = \phi(z) = \mathrm{svec}(zz^\top)$, which implies that

$$\begin{aligned} \phi(x, u) - \phi(x', u') &= \mathrm{svec}\big[zz^\top - (Lz + \varepsilon)(Lz + \varepsilon)^\top\big] \\ &= \mathrm{svec}\big(zz^\top - Lzz^\top L^\top - Lz\varepsilon^\top - \varepsilon z^\top L^\top - \varepsilon\varepsilon^\top\big). \end{aligned}$$

Hence, since ε is independent of z, by the definition of Ξ_K in (3.15), we have

$$\Xi_K = \mathbb{E}_{z\sim\widetilde{\nu}_K}[\phi(z)\,\mathrm{svec}(xx^\top - Lxx^\top L^\top - \widetilde{\Psi}_\sigma)^\top].$$

Now let M and N by any two matrices, by direct computation, we have

$$\begin{aligned} &\mathrm{svec}(M)^\top \Xi_K\, \mathrm{svec}(N) = \mathbb{E}_{z\sim\widetilde{\nu}_K}\big[\langle zz^\top, M\rangle \cdot \langle zz^\top - Lzz^\top L^\top - \widetilde{\Psi}_\sigma, N\rangle\big] \\ &\quad = \mathbb{E}_{z\sim\widetilde{\nu}_K}\big[z^\top Mzz^\top (N - L^\top NL)z\big] - \mathbb{E}_{z\sim\widetilde{\nu}_K}[z^\top Mz] \cdot \langle \widetilde{\Psi}_\sigma, N\rangle \\ &\quad = \mathbb{E}_{g\sim N(0, I_{d+k})}\big[g^\top \widetilde{\Sigma}_K^{1/2} M \widetilde{\Sigma}_K^{1/2} gg^\top \widetilde{\Sigma}_K^{1/2} (N - L^\top NL)\widetilde{\Sigma}_K^{1/2} g\big] - \big\langle \widetilde{\Sigma}_K, M\big\rangle \cdot \big\langle \widetilde{\Psi}_\sigma, N\big\rangle, \end{aligned} \tag{D.19}$$

where $\widetilde{\Sigma}_K^{1/2}$ is the square root of $\widetilde{\Sigma}_K$ defined in (D.18). We utilize the following Lemma to compute the expectation of the product of quadratic forms of Gaussian random variables.

Lemma D.3. Let $g \sim N(0, I_d)$ be the standard Gaussian random variable in $\mathbb{R}^d$ and let A_1, A_2 be two symmetric matrices. Then we have

$$\mathbb{E}[g^\top A_1 g \cdot g^\top A_2 g] = 2\,\mathrm{tr}(A_1 A_2) + \mathrm{tr}(A_1) \cdot \mathrm{tr}(A_2).$$

Proof. See, e.g., [48, 43] for a detailed proof. □

Applying this lemma to (D.19), we have

$$\begin{aligned}
&\operatorname{svec}(M)^\top \Xi_K \operatorname{svec}(N) \\
&\quad = 2\operatorname{tr}\big[\widetilde{\Sigma}_K^{1/2} M \widetilde{\Sigma}_K^{1/2} \cdot \widetilde{\Sigma}_K^{1/2}(N - L^\top N L)\widetilde{\Sigma}_K^{1/2}\big] \\
&\qquad + \operatorname{tr}\big(\widetilde{\Sigma}_K^{1/2} M \widetilde{\Sigma}_K^{1/2}\big) \cdot \operatorname{tr}\big[\widetilde{\Sigma}_K^{1/2}(N - L^\top N L)\widetilde{\Sigma}_K^{1/2}\big] - \langle \widetilde{\Sigma}_K, M\rangle \cdot \langle \widetilde{\Psi}_\sigma, N\rangle \\
&\quad = 2\big\langle M, \widetilde{\Sigma}_K(N - L^\top N L)\widetilde{\Sigma}_K\big\rangle + \langle M, \widetilde{\Sigma}_K\rangle \cdot \big[\langle N - L^\top N L, \widetilde{\Sigma}_K\rangle - \langle \widetilde{\Psi}_\sigma, N\rangle\big]. \qquad \text{(D.20)}
\end{aligned}$$

Note that $\widetilde{\Sigma}_K$ satisfy the Lyapunov equation in (D.18), which implies that

$$\langle N - L^\top N L, \widetilde{\Sigma}_K\rangle = \langle N, \widetilde{\Sigma}_K\rangle - \langle N, L\widetilde{\Sigma}_K L^\top\rangle = \langle N, \widetilde{\Psi}_\sigma\rangle.$$

Thus, by (D.20) we have

$$\begin{aligned}
\operatorname{svec}(M)^\top \Xi_K \operatorname{svec}(N) &= 2\big\langle M, \widetilde{\Sigma}_K(N - L^\top N L)\widetilde{\Sigma}_K\big\rangle = 2\operatorname{svec}(M)^\top \operatorname{svec}\big[\widetilde{\Sigma}_K(N - L^\top N L)\widetilde{\Sigma}_K\big] \\
&= 2\operatorname{svec}(M)^\top \big(\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K - \widetilde{\Sigma}_K L^\top \otimes_s \widetilde{\Sigma}_K L^\top\big) \operatorname{svec}(N)^\top \\
&= 2\operatorname{svec}(M)^\top \big[\big(\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K\big)(I - L^\top \otimes L^\top)\big] \operatorname{svec}(N),
\end{aligned}$$

where the last equality follows from the fact that

$$(A \otimes_s B)(C \otimes_s D) = 1/2 \cdot (AC \otimes_s BD + AD \otimes_s BC)$$

holds for any matrices A, B, C, D. Thus, we have established (D.12). Since $\rho(L) = \rho(A - BK) < 1$, $I - L^\top \otimes L^\top$ is positive definite, which implies that Ξ_K is invertible.

Now we consider the linear equation in (3.16). Since Ξ_K is invertible,

$$\widetilde{\Xi}_K = \begin{pmatrix} 1 & 0 \\ \mathbb{E}_{(x,u)}[\phi(x,u)] & \Xi_K \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \operatorname{svec}(\widetilde{\Sigma}_K) & \Xi_K \end{pmatrix} \qquad \text{(D.21)}$$

is also invertible. Thus, (3.16) has unique solution ϑ_K^*. Moreover, to bound the smallest singular value of $\widetilde{\Xi}_K$, we note that the inverse of $\widetilde{\Xi}_K$ can be written as

$$\widetilde{\Xi}_K^{-1} = \begin{pmatrix} 1 & 0 \\ -\Xi_K^{-1} \operatorname{svec}(\widetilde{\Sigma}_K) & \Xi_K^{-1} \end{pmatrix},$$

whose operator norm is bounded via

$$\big\|\widetilde{\Xi}_K^{-1}\big\|^2 \le 1 + \big\|\Xi_K^{-1} \operatorname{svec}(\widetilde{\Sigma}_K)\big\|_2^2 + \|\Xi_K^{-1}\|^2. \qquad \text{(D.22)}$$

By (D.12), we have

$$\begin{aligned}
\Xi_K^{-1} \operatorname{svec}(\widetilde{\Sigma}_K) &= (I - L^\top \otimes_s L^\top)^{-1}(\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K)^{-1} \operatorname{svec}(\widetilde{\Sigma}_K) \\
&= (I - L^\top \otimes_s L^\top)^{-1}(\widetilde{\Sigma}_K^{-1} \otimes_s \widetilde{\Sigma}_K^{-1}) \operatorname{svec}(\widetilde{\Sigma}_K) = (I - L^\top \otimes_s L^\top)^{-1} \operatorname{svec}(\widetilde{\Sigma}_K^{-1}). \qquad \text{(D.23)}
\end{aligned}$$

The following lemma characterizes the eigenvalues of symmetric Kronecker matrices.

Lemma D.4 (Lemma 7.2 in [2])**.** Let A and B be two matrices in $\mathbb{R}^{m\times m}$ that can be diagonalized simultaneously. Moreover, let $\lambda_1, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_m$ be the eigenvalues of A and B, respectively. Then, the eigenvalues of $A \otimes_s B$ are given by $\{1/2 \cdot (\lambda_i \mu_j + \lambda_j \mu_i), i, j \in [m]\}$.

By Lemma D.4, the spectral radius of $L^\top \otimes_s L^\top$ is bounded by $\rho^2(L) = \rho^2(A - BK) < 1$. By (D.23) we have

$$\big\|\Xi_K^{-1} \operatorname{svec}(\widetilde{\Sigma}_K)\big\|_2 \le \big[1 - \rho^2(L)\big]^{-1} \cdot \|\widetilde{\Sigma}_K^{-1}\|_F \le \sqrt{d+k} \cdot \big[1 - \rho^2(L)\big]^{-1} \cdot \|\widetilde{\Sigma}_K^{-1}\|. \qquad \text{(D.24)}$$

Besides, by (D.12) we have

$$\|\Xi_K^{-1}\| \le \big\|(I - L^\top \otimes_s L^\top)^{-1}\big\| \cdot \big\|\widetilde{\Sigma}_K^{-1} \otimes_s \widetilde{\Sigma}_K^{-1}\big\| \le \big[1 - \rho^2(L)\big]^{-1} \cdot \big\|\widetilde{\Sigma}_K^{-1}\big\|^2. \qquad \text{(D.25)}$$

Notice that $\|\widetilde{\Sigma}_K^{-1}\| = 1/\sigma_{\min}(\widetilde{\Sigma}_K)$. Hence, combining (D.22), (D.24), and (D.25) we conclude that

$$\big\|\widetilde{\Xi}_K^{-1}\big\|^2 \le 1 + (d+k) \cdot \big[1 - \rho(L)^2\big]^{-2} \cdot [\sigma_{\min}(\widetilde{\Sigma}_K)]^{-2} + \big[1 - \rho(L)^2\big]^{-2} \cdot [\sigma_{\min}(\widetilde{\Sigma}_K)]^{-4},$$

which implies that

$$\sigma_{\min}(\widetilde{\Xi}_K) \geq \frac{[1-\rho^2(A-BK)]\cdot[\sigma_{\min}(\widetilde{\Sigma}_K)]^2}{\Big(1+[1-\rho^2(A-BK)]^2\cdot[\sigma_{\min}(\widetilde{\Sigma}_K)]^4+(d+k)\cdot[\sigma_{\min}(\widetilde{\Sigma}_K)]^2\Big)^{1/2}} > 0.$$

Moreover, to see that $\sigma_{\min}(\widetilde{\Sigma}_K)$ only depends on σ and $\sigma_{\min}(\Psi)$, for any $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^k$, we have

$$\begin{aligned}\begin{pmatrix} a \\ b \end{pmatrix}^\top \widetilde{\Sigma}_K \begin{pmatrix} a \\ b \end{pmatrix} &= \mathbb{E}_{(x,u)\sim\widetilde{\nu}_K}[(a^\top x + b^\top u)^2] = \mathbb{E}_{x\sim\nu_K,\eta\sim N(0,I_k)}\big\{[(a-K^\top b)x+\sigma\cdot\eta]^2\big\} \\ &\geq \sigma^2\cdot\|b\|_2^2 + \sigma_{\min}(\Psi)\cdot\|a-K^\top b\|_2^2 \geq (\sigma^2-\sigma_{\min}(\Psi)\cdot\|K\|^2)\cdot\|b\|_2^2+\sigma_{\min}(\Psi)\cdot\|a\|_2^2.\end{aligned}$$

Thus, suppose σ^2 is sufficiently large such that $\sigma^2 - \sigma_{\min}(\Psi)\cdot\|K\|^2 > 0$, $\sigma_{\min}(\widetilde{\Sigma}_K)$ is lower bounded by $\min\{\sigma^2-\sigma_{\min}(\Psi)\cdot\|K\|^2, \sigma_{\min}(\Psi)\}$. Therefore, we can find a constant κ_K^* depending only on $\rho(A-KB)$, σ, and $\sigma_{\min}(\Psi)$ such that $\sigma_{\min}(\widetilde{\Xi}_K) \geq \kappa_K^*$.

Finally, to obtain an upper bound on $\|\Xi_K\|$, by triangle inequality and Lemma D.4 we have

$$\|\Xi_K\| \leq \big\|\widetilde{\Sigma}_K \otimes_s \widetilde{\Sigma}_K\big\|\cdot\big(1+\|L^\top \otimes_s L^\top\|\big) \leq \big\|\widetilde{\Sigma}_K\big\|^2\cdot\big(1+\|L\|^2\big) \leq 2\big\|\widetilde{\Sigma}_K\big\|^2,$$

where we use the fact that $\rho(L) < 1$. Applying (D.17) to the inequality above, we obtain that

$$\|\Xi_K\| \leq 2\big[\sigma^2 + (1+\|K\|_{\mathrm{fro}}^2)\cdot\|\Sigma_K\|\big],$$

which concludes the proof. □