

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 $\hat{\Theta}$ of Θ_K .
 Actor step. Update the policy parameter by $K \leftarrow K - \gamma \cdot (\hat{\Theta}^{22}K - \hat{\Theta}^{21})$.
end while
Output: The final policy π_K , matrix $\hat{\Theta}$ that estimates Θ_K , and \hat{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 $\hat{\Theta}$ 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 $t = 1, 2, \dots, T$ **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 $\omega_t^1 = (1 - \alpha_t) \cdot \omega_{t-1}^1 + \alpha_t \cdot (\vartheta_{t-1}^1 - c_{t-1})$.
 Update ω^2 by $\omega_t^2 = (1 - \alpha_t) \cdot \omega_{t-1}^2 + \alpha_t \cdot \delta_t \cdot \phi(x_{t-1})$.
 Project ϑ_t and ω_t to \mathcal{X}_Θ and \mathcal{X}_Ω , respectively.
end for
Define $\hat{\vartheta} = (\hat{\vartheta}^1, \hat{\vartheta}^2) = (\sum_{t=1}^T \alpha_t \cdot \vartheta_t) / (\sum_{t=1}^T \alpha_t)$ and $\hat{\omega} = (\sum_{t=1}^T \alpha_t \cdot \omega_t) / (\sum_{t=1}^T \alpha_t)$.
Return $\hat{\vartheta}^1$ and $\hat{\Theta} = \text{smat}(\hat{\vartheta}^2)$ 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 $\hat{\Theta}_K$. 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

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

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

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

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

$$\Lambda_K \vartheta = \lambda_K, \quad \Lambda_K = \begin{pmatrix} 1 & \bar{h}_K^\top \\ \bar{g}_K & \bar{\Xi}_K \end{pmatrix}, \quad \vartheta = \begin{pmatrix} \vartheta^1 \\ \vartheta^2 \end{pmatrix}, \quad \lambda_K = \begin{pmatrix} \bar{a}_K \\ \bar{b}_K \end{pmatrix}. \quad (\text{B.2})$$

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

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

Here, similar to Assumption (4.1), we let \mathcal{X}_Θ and \mathcal{X}_Ω be Euclidean balls given by \mathcal{X}_Θ and \mathcal{X}_Ω in (3.18) be defined as $\mathcal{X}_\Theta = \{\vartheta: \|\vartheta^1\|_2 \leq \tilde{R}_\Theta\}$ and $\mathcal{X}_\Omega = \{\omega: \|\omega\|_2 \leq \tilde{R}_\Omega\}$, where \tilde{R}_Θ and \tilde{R}_Ω are chosen appropriately. Notice that both $\bar{F}(\vartheta, \omega)$ 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 \hat{J} and $\hat{\Theta}$ 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 $t = 1, 2, \dots, T$ **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}_Θ and \mathcal{X}_Ω , respectively.

end for

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

Return $\hat{\vartheta}^1$ and $\hat{\Theta} = \text{smat}(\hat{\vartheta}^2)$ 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 $\hat{\vartheta}^1$ and $\hat{\Theta}$ 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

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

where $\Upsilon[\tilde{R}_\Theta, \tilde{R}_\Omega, J(K_0), \|K\|_{\text{fro}}, \sigma_{\min}^{-1}(Q)]$ is a polynomial of \tilde{R}_Ω , \tilde{R}_Θ , $J(K_0)$, $\|K\|_{\text{fro}}$, 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

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

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

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

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

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

Recall that $c(x, u) = \langle \phi(x, u), \text{svec}[\text{diag}(Q, R)] \rangle$. Since both $\phi(x, u)$ and $\tau_K(x, u) \cdot \phi(x, u)$ are sub-exponential random vectors, when \tilde{R}_Θ and \tilde{R}_Ω 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 $\|\hat{\Theta} - \Theta_K\|_{\text{fro}}^2$ to the primal-dual gap

$$\text{Gap}(\hat{\vartheta}, \hat{\omega}) = \max_{\omega \in \mathcal{X}_\Omega} \bar{F}(\hat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \bar{F}(\vartheta, \hat{\omega}). \quad (\text{B.6})$$

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

$$|\hat{\vartheta}^1 - J(K)|^2 + \|\hat{\Theta} - \Theta_K\|_{\text{fro}}^2 \leq \kappa_K^{*-2} \cdot \text{Gap}(\hat{\vartheta}, \hat{\omega}). \quad (\text{B.7})$$

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

$$\mathcal{E} = \bigcap_{0 \leq t \leq T} \left\{ \|\phi(x_t, u_t)\|_2^2 \leq C_K \cdot \log T, \|\tau_K(x_t, u_t) \cdot \phi(x_t, u_t)\|_2^2 \leq C_K \cdot \log T \right\}, \quad (\text{B.8})$$

where C_b is a constant specified by the stationary distribution ν_b . Since both $\phi(x, u)$ 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

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

and the truncated minimax optimization problem,

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

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

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

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

$$\left| \text{Gap}(\hat{\vartheta}, \hat{\omega}) - \left[\max_{\omega \in \mathcal{X}_\Omega} \tilde{F}(\hat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \tilde{F}(\vartheta, \hat{\omega}) \right] \right| \leq 2/T. \quad (\text{B.11})$$

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

$$\max_{\omega \in \mathcal{X}_\Omega} \tilde{F}(\hat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \tilde{F}(\vartheta, \hat{\omega}) \leq \frac{C_K \cdot \log^6 T}{(1 - \rho) \cdot \sqrt{T}} \quad (\text{B.12})$$

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

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}_Θ and \mathcal{X}_Ω 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 $\|\hat{\Theta} - \Theta_K\|_{\text{fro}}^2$, where $\hat{\Theta} = \text{smat}(\hat{\vartheta}^2)$ 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

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

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

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

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

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

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}_Ω , where \mathcal{X}_Θ and \mathcal{X}_Ω 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 $\|\Theta_K\|_{\text{fro}}$. 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

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

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

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

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

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

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

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

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

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

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

Combining (C.5) and the definition of \tilde{R}_Θ 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}_Ω 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

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

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

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

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

$$\|\Xi_K \vartheta^2\|_2 \leq \|\Xi_K\| \cdot \|\vartheta^2\|_2 \leq 4(1 + \|K\|_{\text{fro}}^2)^2 \cdot \|\Sigma_K\|^2 \cdot \tilde{R}_\Theta. \quad (\text{C.8})$$

Moreover, for any positive definite matrix Γ , we have

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

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

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

which implies that

$$\|b_K\|_2 \leq 3(\|Q\|_{\text{fro}} + \|R\|_{\text{fro}}) \cdot \|\tilde{\Sigma}_K\|^2. \quad (\text{C.10})$$

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

$$\|\vartheta^1 \cdot \mathbb{E}_{(x,u)} [\phi(x, u)]\|_2 \leq J(K_0) \cdot \|\tilde{\Sigma}_K\|_{\text{fro}}. \quad (\text{C.11})$$

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

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

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 $\|\hat{\Theta} - \Theta_K\|_{\text{fro}}^2$ to the performance of the optimization in (3.18). Specifically, we consider the primal-dual gap

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

which characterizes the closeness between $(\hat{\vartheta}, \hat{\omega})$ 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

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

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

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

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

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

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

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

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

which establishes the connection between $\|\hat{\Theta} - \Theta_K\|_{\text{fro}}^2$ 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

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

Proof. See [55] for a detailed proof. \square

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

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

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

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

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 $t \in \{0, 1, \dots, T\}$, we define

$$\mathcal{E}_t = \left\{ \left| \|x_t\|_2^2 + \|u_t\|_2^2 - \text{tr}(\tilde{\Sigma}_K) \right| \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_K\| \right\}, \quad (\text{C.20})$$

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

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

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

$$\tilde{\phi}(x, u) = \phi(x, u) \cdot \mathbb{1}_{\left\{ \left| \|\phi(x, u)\|_2^2 - \text{tr}(\tilde{\Sigma}_K) \right| \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_K\| \right\}}. \quad (\text{C.22})$$

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

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

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

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

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

We remark that, when \mathcal{E} is true, $(\hat{\vartheta}, \hat{\omega})$ 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 $\max_{\omega \in \mathcal{X}_\Omega} \tilde{F}(\hat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \tilde{F}(\vartheta, \hat{\omega})$.

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

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

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

$$G^1(x, u, x', u'; \vartheta) - \tilde{G}^1(x, u, x', u'; \vartheta) = c(x, u) \cdot \mathbf{1}_{\mathcal{A}^c} \quad (\text{C.26})$$

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

where we denote $\{ \|\phi(x, u)\|_2^2 - \text{tr}(\tilde{\Sigma}_K) \} \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_K\|$ and $\{ \|\phi(x, u)\|_2^2 - \text{tr}(\tilde{\Sigma}_K) \} \leq C_1 \cdot \log T \cdot \|\tilde{\Sigma}_K\|$ 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

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

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

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

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

$$\begin{aligned}&\|\mathbb{E}_{(x, u, x', u')} [G^2(x, u, x', u'; \vartheta) - \tilde{G}^2(x, u, x', u'; \vartheta)]\|_2 \\ &\leq \left\{ \|\mathbb{E}_{(x, u, x', u')} [G^2(x, u, x', u'; \vartheta) \cdot \mathbf{1}_{\mathcal{A}^c}]\|_2 + \|\mathbb{E}_{(x, u, x', u')} [\phi(x', u')^\top \vartheta^2 \cdot \phi(x, u) \cdot \mathbf{1}_{\mathcal{A}} \cdot \mathbf{1}_{\mathcal{B}^c}]\|_2 \right\} \\ &\leq \left\{ \sqrt{\mathbb{P}(\mathcal{A}^c)} \cdot \sqrt{\mathbb{E}[\|G^2(x, u, x', u'; \vartheta)\|_2^2]} + \sqrt{\mathbb{P}(\mathcal{B}^c)} \cdot \sqrt{\mathbb{E}[\|\phi(x, u) \cdot \phi(x', u')^\top \vartheta^2\|_2^2]} \right\}.\end{aligned}\quad (\text{C.28})$$

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

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

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

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

$$\mathbb{E}[\|\phi(x, u) \cdot \phi(x', u')^\top \vartheta^2\|_2^2] \leq \left(\mathbb{E}[\|\phi(x', u')^\top \vartheta^2\|^4] \cdot \mathbb{E}[\|\phi(x, u)\|_2^4] \right)^{1/2}. \quad (\text{C.31})$$

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

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

$$\begin{aligned} & \left| \text{Gap}(\hat{\vartheta}, \hat{\omega}) - \left[\max_{\omega \in \mathcal{X}_\Omega} \tilde{F}(\hat{\vartheta}, \omega) - \min_{\vartheta \in \mathcal{X}_\Theta} \tilde{F}(\vartheta, \hat{\omega}) \right] \right| \\ & \leq \max_{\omega \in \mathcal{X}_\Omega} |F(\hat{\vartheta}, \omega) - \tilde{F}(\hat{\vartheta}, \omega)| + \max_{\vartheta \in \mathcal{X}_\Theta} |F(\vartheta, \hat{\omega}) - \tilde{F}(\vartheta, \hat{\omega})| \leq \frac{2}{T}. \end{aligned} \quad (\text{C.32})$$

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

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

where the function $\tilde{\phi}(x, u)$ is defined in (C.22), and we denote $\tilde{\phi}(x, u)$ and $\tilde{\phi}(x', u')$ by ϕ and ϕ' , respectively. Using this definition, the objective function $\tilde{F}(\vartheta, \omega)$ in (C.23) can be written as $\tilde{F}(\vartheta, \omega) = \mathbb{E}_{(x, u, x', u')} [H(\vartheta, \omega; \phi, \phi')]$, 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

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

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

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

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

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

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

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

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

where the last equality holds since $\tilde{R}_\Theta \geq \|Q\|_{\text{fro}} + \|R\|_{\text{fro}}$. Moreover, we have $\nabla_{\vartheta\vartheta}^2 H(\vartheta, \omega; \phi, \phi') = 0$ and $-\nabla_{\omega\omega}^2 H(\vartheta, \omega; \phi, \phi')$ 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 $\{X_t\}_{t \geq 0} \subseteq \mathbb{R}^m$, $\varepsilon \sim N(0, \Psi)$ 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 $t \geq 0$. Besides, the stationary distribution of this Markov chain is denoted by $N(0, \Sigma_\infty)$. For any integer $k \geq 1$, we define the k -th mixing coefficient as

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

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

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

where $C_{\rho, L}$ is a constant that solely depends on ρ and A . That is, $\{X_t\}_{t \geq 0}$ is geometrically β -mixing.

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

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

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

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

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

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

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

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

$$\max_{y \in \mathcal{Y}} F(\hat{x}, y) - \min_{x \in \mathcal{X}} F(x, \hat{y}) \leq \frac{C \cdot (D^2 + L_1^2 + L_1 L_2 D)}{\log(1/\rho)} \cdot \frac{\log^2 T + \log(1/\delta)}{\sqrt{T}} + \frac{C \cdot C_\xi L_1 D}{T}.$$

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

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

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

$$A_0 = D^2 + 12D \cdot \alpha_1 \cdot \tau(\eta) \quad A_1 = 4L_1 D \quad A_2 = 10L_1^2 + (24L_1^2 + 8L_1 L_2 D) \cdot \tau(\eta).$$

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

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

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

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

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

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

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

which concludes the proof of Theorem C.4. \square

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

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

Moreover, by the definitions of \mathcal{X}_Θ and \mathcal{X}_Ω in (4.1) and (4.2), respectively, we can set D by

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

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

$$L_1 \leq 16C_1^2 \cdot \log^2 T \cdot [\sigma^2 + (1 + \|K\|_{\text{fro}}^2) \cdot \|\Sigma_K\|]^2 \cdot [(1 + \|K\|_{\text{fro}}^2)^2 \cdot \tilde{R}_\Omega + \tilde{R}_\Theta], \quad L_2 = 1. \quad (\text{C.46})$$

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

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

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

$$\begin{aligned} \text{Gap}(\hat{\vartheta}, \hat{\omega}) &\leq C \cdot \log^4 T \cdot (1 + \|K\|_{\text{fro}}^2)^4 \cdot \|\Sigma_K\|^4 \cdot [(1 + \|K\|_{\text{fro}}^2)^2 \cdot \tilde{R}_\Omega + \tilde{R}_\Theta]^2 \\ &\cdot \left(\frac{\log^2 T + \log(T^5)}{\log(1/\rho) \cdot \sqrt{T}} + \frac{\sqrt{d + k}}{(1 - \rho) \cdot T} \right) + \frac{2}{T} \\ &\leq C \cdot (1 + \|K\|_{\text{fro}}^2)^4 \cdot \|\Sigma_K\|^4 \cdot [(1 + \|K\|_{\text{fro}}^2)^2 \cdot \tilde{R}_\Omega + \tilde{R}_\Theta]^2 \cdot \frac{\log^6 T}{(1 - \rho) \cdot \sqrt{T}} \end{aligned} \quad (\text{C.48})$$

holds with probability at least $1 - 3T^{-5} \geq 1 - T^{-4}$, 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 $1/\log(1/\rho) \leq 1/(1 - \rho)$. This further implies that the first term on the right-hand side of the first inequality dominates the second term. The upper bound of $\text{Gap}(\hat{\vartheta}, \hat{\omega})$ 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. \square

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

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

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 $A_{K, K'}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ 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

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

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

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

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

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

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

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

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. \square

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

$$\begin{aligned} \sigma_{\min}(\Psi) \cdot \|R + B^\top P_K B\|^{-1} \cdot \text{tr}(E_K^\top E_K) &\leq J(K) - J(K^*) \\ &\leq 1/\sigma_{\min}(R) \cdot \|\Sigma_{K^*}\| \cdot \text{tr}(E_K^\top E_K). \end{aligned} \quad (\text{C.53})$$

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

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

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

$$J(K) - J(K^*) = -\mathbb{E}_{x_0^* \sim N(0, \Psi_\sigma)} \left[\sum_{t \geq 0} A_{K, K^*}(x_t^*) \right], \quad (\text{C.55})$$

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

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

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

$$\begin{aligned} A_{K, K'}(x) &= 2x^\top (K' - K)^\top E_K x + x^\top (K' - K)^\top (R + B^\top P_K B)(K' - K)x \\ &= \text{tr} \left\{ x x^\top [K' - K + (R + B^\top P_K B)^{-1} E_K]^\top (R + B^\top P_K B) [K' - K + (R + B^\top P_K B)^{-1} E_K] \right\} \\ &\quad - \text{tr} [x x^\top E_K^\top (R + B^\top P_K B)^{-1} E_K] \\ &\geq -\text{tr} [x x^\top E_K^\top (R + B^\top P_K B)^{-1} E_K], \end{aligned} \quad (\text{C.57})$$

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

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

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 $J(K) - J(K^*) \leq 1/\sigma_{\min}(R) \cdot \|\Sigma_{K^*}\| \cdot \text{tr}(E_K^\top E_K)$, 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

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

where in the first equality we define $x_t' = (A - B K')^t x_0^*$ for all $t \geq 0$, 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. \square

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

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

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

$$K_{t+1} = K_t - \gamma \cdot \hat{E}_{K_t}, \quad \hat{E}_{K_t} = \hat{\Theta}_t^{22} K_t - \hat{\Theta}_t^{21}, \quad (\text{C.59})$$

where $\hat{\Theta}_t$ 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

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

When γ is sufficiently small such that

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

by triangle inequality, we have

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

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

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

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

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

Now we utilize the following Lemma, obtained from [26], to construct an 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

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

then we have

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

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

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

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

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

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

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

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

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

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

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

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

In addition, by triangle inequality, we have

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

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

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

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

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

In the sequel, we set

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Meanwhile, (C.63) implies that,

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

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

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

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

$$N \geq 2\|\Sigma_{K^*}\|/\gamma \cdot \sigma_{\min}^{-1}(\Psi) \cdot \sigma_{\min}^{-1}(R) \cdot \log\{2[J(K_0) - J(K^*)]/\epsilon\} \quad (\text{C.80})$$

policy updates, we have $J(K_N) - J(K^*) \leq \epsilon$ 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

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

for some polynomial function $\Upsilon_4(\cdot, \cdot)$. With such a T_t , the fail probability $T_t^{-4} \leq \epsilon^{-20}$. 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. \square

D Proofs of the Auxiliary Results

In this section, we provide 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 $t \geq 0$, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\alpha_K = -\text{tr}(P_K \Sigma_K)$. Hence, we establish (3.5).

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

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

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

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

where the second term denotes that we first take compute the gradient $\nabla_K \text{tr}[Q_0 \Sigma_K]$ 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

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

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

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

which implies that

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

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

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

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

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

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

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

Therefore, we conclude the proof of the lemma. \square

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

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

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. \square

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 $N(0, \tilde{\Sigma}_K)$ 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

$$\Xi_K = (\tilde{\Sigma}_K \otimes_s \tilde{\Sigma}_K) - (\tilde{\Sigma}_K L^\top) \otimes_s (\tilde{\Sigma}_K L^\top) = (\tilde{\Sigma}_K \otimes_s \tilde{\Sigma}_K)(I - L^\top \otimes_s L^\top). \quad (\text{D.12})$$

Moreover, Ξ_K is a invertible matrix whose operator norm is bounded by $2[\sigma^2 + (1 + \|K\|_{\text{fro}}^2) \cdot \|\Sigma_K\|]$. 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 $\text{svec}(\Theta_K)$, 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

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

where $\epsilon \sim N(0, \Psi)$ and $\eta \sim N(0, I_k)$. 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

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

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

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

which is denoted by $\tilde{\Psi}_\sigma$ 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 $\tilde{\nu}_K$, which is a centered Gaussian distribution in \mathbb{R}^{d+k} . Since $x \sim N(0, \Sigma_K)$ and $u = -Kx + \sigma \cdot I_k$, we can write $\tilde{\nu}_K$ as $N(0, \tilde{\Sigma}_K)$, where $\tilde{\Sigma}_K \in \mathbb{R}^{(d+k) \times (d+k)}$ can be written as

$$\tilde{\Sigma}_K = \begin{pmatrix} \Sigma_K & -\Sigma_K K^\top \\ -K\Sigma_K & K\Sigma_K K^\top + \sigma^2 \cdot I_k \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 \cdot I_k \end{pmatrix} + \begin{pmatrix} I_d \\ -K \end{pmatrix} \Sigma_K \begin{pmatrix} I_d \\ -K \end{pmatrix}^\top. \quad (\text{D.16})$$

Thus, by triangle inequality we have

$$\|\tilde{\Sigma}_K\|_{\text{fro}} \leq \sigma^2 \cdot k + \|\Sigma_K\| \cdot (d + \|K\|_{\text{fro}}^2), \quad \|\tilde{\Sigma}_K\| \leq \sigma^2 + (1 + \|K\|_{\text{fro}}^2) \cdot \|\Sigma_K\|, \quad (\text{D.17})$$

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

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

$$\tilde{\Sigma}_K = L\tilde{\Sigma}_K L^\top + \tilde{\Psi}_\sigma, \quad (\text{D.18})$$

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

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

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

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

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

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

where $\tilde{\Sigma}_K^{1/2}$ is the square root of $\tilde{\Sigma}_K$ 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

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

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

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

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

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

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

Thus, by (D.20) we have

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

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,

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

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

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

whose operator norm is bounded via

$$\|\tilde{\Xi}_K^{-1}\|^2 \leq 1 + \|\Xi_K^{-1} \text{svec}(\tilde{\Sigma}_K)\|_2^2 + \|\Xi_K^{-1}\|^2. \quad (\text{D.22})$$

By (D.12), we have

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

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 $\lambda_1, \dots, \lambda_m$ and μ_1, \dots, μ_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

$$\|\Xi_K^{-1} \text{svec}(\tilde{\Sigma}_K)\|_2 \leq [1 - \rho^2(L)]^{-1} \cdot \|\tilde{\Sigma}_K^{-1}\|_F \leq \sqrt{d+k} \cdot [1 - \rho^2(L)]^{-1} \cdot \|\tilde{\Sigma}_K^{-1}\|. \quad (\text{D.24})$$

Besides, by (D.12) we have

$$\|\tilde{\Xi}_K^{-1}\| \leq \|(I - L^\top \otimes_s L^\top)^{-1}\| \cdot \|\tilde{\Sigma}_K^{-1} \otimes_s \tilde{\Sigma}_K^{-1}\| \leq [1 - \rho^2(L)]^{-1} \cdot \|\tilde{\Sigma}_K^{-1}\|^2. \quad (\text{D.25})$$

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

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

which implies that

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

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

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

Thus, suppose σ^2 is sufficiently large such that $\sigma^2 - \sigma_{\min}(\Psi) \cdot \|K\|^2 > 0$, $\sigma_{\min}(\tilde{\Sigma}_K)$ 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 $\sigma_{\min}(\tilde{\Xi}_K) \geq \kappa_K^*$.

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

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

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

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

which concludes the proof. \square