A Pseudocode of NENAC for Zero-Sum Games in General

In this section, we present the detailed description of NENAC for RSRL and RARL in general, which is introduced in §3. Note that the computation of the natural policy gradients requires taking expectation with respect to $\widetilde{\rho}_{K,L}$,

Algorithm 2 NENAC for Zero-Sum Games in General

Input: Initial parameters K_0 and L_0. Stepsize ι of the outer loop. Stepsize γ of the inner loop. Number of iterations $\mathcal{T}$ of the inner loop. Number of iterations $\mathcal{N}$ of the outer loop. Feasible parameter set $\mathcal{L}$.

Initialization: $K \leftarrow K_0, L \leftarrow L_0$.
for $n = 0, 1, ..., \mathcal{N}$ **do**
  Initialization: $K \leftarrow K_0(L)$.
  for $t = 0, 1, ..., \mathcal{T}$ **do**
    Inner Critic Step: Estimate $Q_{K,L}$ by $\widehat{Q}^\lambda$ via policy evaluation (e.g. GTD).
    Inner Actor Step: $K \leftarrow K - \gamma \cdot \mathcal{I}(K; L)^{-1}\mathbb{E}_{(x.u.v)\sim\widetilde{\rho}_{K,L}}[\nabla_K \log \pi_{K,L}(u, v \,|\, x) \cdot \widehat{Q}^\lambda(x, u, v)]$.
  end for
  Outer Critic Step: Estimate $Q_{K,L}$ by $\widehat{Q}^\lambda$ via policy evaluation (e.g. GTD).
  Outer Actor Step: $L \leftarrow \Pi_{\mathcal{L}}\{L + \iota \cdot \mathcal{I}(L; K)^{-1}\mathbb{E}_{(x,u,v)\sim\widetilde{\rho}_{K,L}}\mathbb{E}[\nabla_L \pi_{K,L}(u, v \,|\, x) \cdot \widehat{Q}^\lambda(x, u, v)]\}$.
end for

Output: (K, L) that estimates (K^*, L^*), where K is the optimal policy for RSRL and RARL.

which is the stationary distribution induced by $\pi_{K,L}$. For RSRL and RARL in general, such an expectation can be obtained by sampling from $\widetilde{\rho}_{K,L}$. For RSRL and RARL in the LQ setting, we develop a more efficient method to obtain the natural policy gradient, which is introduced in §4.

B Proof of the Main Results

In this section, we present the proof of Theorem 5.3. To this end, we first restate Theorem 5.3 with detailed dependency as follows.

Theorem B.1 (Theorem 5.3 restated)**.** Suppose that Assumptions 4.1, 4.3 and 5.1 hold. Let $\{(K_t^n, L^n)\}_{t,n\geq 0}$ be generated by Algorithm 1. Then, the following properties hold.

(i) For a fixed n, let $K_t = K_t^n$ and $L = L^n$. We set the number of GTD iterations at the inner critic step $T_{\mathrm{TD}}^{\mathrm{in}} = \Omega(\epsilon^{-5})$, the stepsize of the inner loop

$$\gamma \leq \left[\|R\| + \sigma_{\min}(\Psi_\sigma) \cdot \|B\|^2 \cdot (J(K_0, L) + \chi)\right]^{-1},$$

and the number of inner loop iterations $\mathcal{T} = \Omega(\log(1/\epsilon))$. Then, with probability at least $1 - \epsilon^{10}$, it holds that

$$J(K_{\mathcal{T}}, L) - J(K(L), L) \leq \epsilon.$$

(ii) We set the number of GTD iterations at the outer critic step $T_{\mathrm{TD}}^{\mathrm{out}} = \Omega(\varepsilon^{-10})$, the stepsize $\iota \leq \bar{\iota}$ for some $\bar{\iota} > 0$, and the number of outer loop iteration $\mathcal{N} = \mathcal{O}(\varepsilon^{-2})$. Then, it holds with probability at least $1 - \varepsilon^{10}$ that

$$\mathcal{N}^{-1} \cdot \sum_{n=1}^{\mathcal{N}} \|\widetilde{F}_{L^n}\| \leq \varepsilon.$$

Proof. Our analysis consists of two parts. In the first part, we show the global convergence of the inner loop. Then, in the second part, we show the global convergence of the outer loop, which corresponds to the global convergence of Algorithm 1. In what follows, for notational simplicity, we denote $\zeta_{K(L),L}$ by ζ_L^* for any notation ζ.

Global Convergence of the Inner Loop. The proof of the global convergence of the inner loop is based on Yang et al. (2019); Fazel et al. (2018). For notational simplicity, let $K_t = K_t^n$ and $L = L^n$ for a fixed n. In the

inner loop, we solve the following optimization problem

$$\min_{K'} J(K', L),$$

for a fixed $L \in \mathcal{L}$, where $\mathcal{L}$ is defined in (4.10). For a fixed L, we have that $v_t = -Lx_t + \sigma_2 \eta_t^2$. Thus, the transition takes the following form of

$$x_{t+1} = Ax_t + Bu_t + C(-Lx_t + \sigma_2 \eta_t^2) + e_t = A_L x_t + Bu_t + e_{C,t}, \tag{B.1}$$

where we write

$$A_L = A - CL, \qquad e_{C,t} = e_t + \sigma_2 \cdot C\eta_t^2. \tag{B.2}$$

The cost function takes the form of

$$\begin{aligned} c(x,u) &= \mathbb{E}_{\eta^2}\big[x^\top Qx + u^\top Ru - (-Lx + \sigma_2\eta^2)^\top S(-Lx + \sigma_2\eta^2)\big] \\ &= x^\top Q_L x + u^\top Ru + \mathbb{E}_{\eta^2}\big[(\sigma_2\eta^2)^\top S(\sigma_2\eta^2)\big], \end{aligned} \tag{B.3}$$

where we write

$$Q_L = Q - L^\top SL. \tag{B.4}$$

Note that $\mathbb{E}_{\eta^2}[(\sigma_2\eta^2)^\top S(\sigma_2\eta^2)]$ is independent of K and L. Thus, the cost function in (B.3) is equivalent to the following cost function

$$c_L(x,u) = x^\top Q_L x + u^\top Ru. \tag{B.5}$$

Thus, the inner loop is equivalent to a single agent LQ control problem with transition dynamics(B.1) and cost function (B.5). For a fixed $L \in \mathcal{L}$, we update K via

$$\widehat{E}_t = \widehat{\Lambda}^{22}_{K_t,L} \cdot K_t + \widehat{\Lambda}^{23} L - \widehat{\Lambda}^{21}_{K_t,L}, \qquad K_{t+1} = K_t - \gamma \cdot \widehat{E}_t,$$

where $\widehat{\Lambda}_{K,L}$ is the output of Algorithm 3 that estimates $\Lambda_{K,L}$. Ideally, K_t converges to $K(L)$, which is the optimal solution of the inner minimization problem.

We then establish the following lemma that characterizes the difference between $P_{K,L}$ and $P_{K',L'}$, which serves as a fundamental lemma though the proof.

Lemma B.2 (Cost Difference)**.** Let (K, L) and (K', L') be stable policy pairs. Let $\{x'_t\}_{t\geq 0}$ be the sequence generated by $x'_0 = x, x'_{t+1} = (A - BK' - CL')x'_t$ for $t \geq 0$. Then, it holds that

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

where

$$\begin{aligned} A_{K,L,K',L'}(x) = {} & 2x^\top (K'-K)^\top E_{K,L} x + x^\top (K'-K)^\top (R + B^\top P_{K,L} B)(K'-K)x \\ & + 2x^\top (L'-L) F_{K,L} x + x^\top (L'-L)^\top (-S + C^\top P_{K,L} C)(L'-L)x \\ & + 2x^\top (L'-L)^\top C^\top P_{K,L} B(K'-K)x. \end{aligned}$$

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

Applying Lemma B.2, we establish the following lemma, which characterizes the gradient dominance of $J(K, L)$ with respect to K.

Lemma B.3 (Gradient Dominance of $J(K, L)$)**.** For any stable policy (K, L), under Assumption 4.1, we have that

$$\begin{aligned} & \sigma_{\min}(\Psi) \cdot \|R + B^\top P_{K,L} B\|^{-1} \cdot \mathrm{Tr}(E_{K,L}^\top E_{K,L}) \\ & \quad \leq J(K,L) - J(K(L),L) \leq 1/\sigma_{\min}(R) \cdot \|\Sigma_L^*\| \cdot \mathrm{Tr}(E_{K,L}^\top E_{K,L}). \end{aligned} \tag{B.6}$$

Proof. See §E.5 for details. □

By Polyak (1963), Lemma B.3 is sufficient to show the linear convergence in objective function of natural policy gradient if we have the exact $E_{K_t,L}$. Since we only have the estimation $\widehat{E}_{K_t,L}$ of $E_{K_t,L}$, we need to aggregate the error of estimating $E_{K_t,L}$ by $\widehat{E}_{K_t,L}$. Note that we assume (K_0, L) is a stable policy pair by Assumption 5.1, which implies that $J(K_0, L)$ is finite. We use mathematical induction to prove that $\{J(K_t, L)\}_{t\geq 0}$ is a monotone decreasing sequence. Suppose that $J(K_t, L) \leq J_L(K_0, L) < \infty$. We define K'_{t+1} as the parameter obtained by a single step of natural policy gradient with $E_{K_t,L}$ as follows

$$K'_{t+1} = K_t - \gamma E_{K_t,L}.$$

We use K'_{t+1} as the interpolating term of K_t and K_{t+1}. The following lemma that characterizes the decrease of $J(K'_{t+1}, L)$.

Lemma B.4. We set γ to satisfy that

$$\gamma \leq \Big(\|R\| + \sigma_{\min}(\Psi_\sigma) \cdot \|B\|^2 \cdot \big(J(K_0, L) + \chi\big)\Big)^{-1}.$$

Then, we have that

$$J(K'_{t+1}, L) - J(K_t, L) \leq -\gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \sigma_{\min}(R) \cdot \|\Sigma_L^*\|^{-1} \cdot \big(J(K_t, L) - J(K(L), L\big).$$

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

In the sequel, we set γ to satisfy

$$\gamma \leq \Big(\|R\| + \sigma_{\min}(\Psi_\sigma) \cdot \|B\|^2 \cdot \big(J(K_0, L) + \chi\big)\Big)^{-1}.$$

We establish the following lemma that characterizes the difference $|J(K_{t+1}, L) - J(K'_{t+1}, L)|$.

Lemma B.5. At the the t-th inner iteration, we set the number of GTD iterations $T_{\mathrm{TD}}^{\mathrm{in}}$ at the inner critic step in Algorithm 1) to satisfy that

$$T_{\mathrm{TD}}^{\mathrm{in}} \geq \left(\frac{\mathrm{Poly}\big(J(K_t, L), \|K_t\|_{\mathrm{F}}, \|L\|_{\mathrm{F}}, \kappa^{-1}\big)}{\sigma_{\min}(\mathfrak{T}_{K_t,L}) \cdot \big(1 - \rho(A - BK_t - CL)\big) \cdot \big(J(K_t, L) - J(K(L), L)\big)}\right)^5. \tag{B.7}$$

Then, we have that

$$\begin{aligned}&\big|J(K'_{t+1}, L) - J(K_{t+1}, L)\big| \\ &\quad \leq 1/2 \cdot \gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \sigma_{\min}(R) \cdot \|\Sigma_L^*\|^{-1} \cdot \big(J(K_t, L) - J(K(L), L)\big),\end{aligned}$$

with probability at least $1 - (T_{\mathrm{TD}}^{\mathrm{in}})^{-4}$.

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

We set $T_{\mathrm{TD}}^{\mathrm{in}}$ to satisfy (B.7) of Lemma B.5. By Lemma B.4 and B.5, we have with probability at least $1 - (T_{\mathrm{TD}}^{\mathrm{in}})^{-4}$ that

$$\begin{aligned}J(K_{t+1}, L) - J(K_t, L) &\leq |J(K'_{t+1}, L) - J(K_{t+1}, L)| + J(K'_{t+1}, L) - J(K_t, L) \\ &\leq -1/2 \cdot \gamma \cdot \sigma_{\min}(\Psi) \cdot \sigma_{\min}(R) \cdot \|\Sigma_L^*\|^{-1} \cdot \big(J(K_t, L) - J(K(L), L)\big) \leq 0,\end{aligned} \tag{B.8}$$

which shows that the sequence $\{J(K_t, L)\}_{t\geq 0}$ decreases monotonously. By rearranging the inequality (B.8), we have with probability at least $1 - (T_{\mathrm{TD}}^{\mathrm{in}})^{-4}$ that

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

which shows the linear convergence of the inner loop of Algorithm 1. Specifically, to ensure that $J(K_t, L) - J(K(L), L) \leq \epsilon$, we set the number of iterations $\mathcal{T}$ of the inner loop of Algorithm 1 to be

$$\mathcal{T} \geq \gamma^{-1} \cdot \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \sigma_{\min}(R)^{-1} \cdot \|\Sigma_L^*\| \cdot \log\left(\frac{2\big[J(K_0, L) - J(K(L), L)\big]}{\epsilon}\right), \tag{B.9}$$

where we use that fact that $-\log^{-1}(1-x) \leq x^{-1}$ for $0 \leq x < 1$. Recall that we set $T_{\mathrm{TD}} = \Omega(\epsilon^{-5})$ to ensure the condition (B.7), which implies that the failure probability of each inner critic step is $(T_{\mathrm{TD}}^{\mathrm{in}})^{-4} = \mathcal{O}(\epsilon^{20})$. Thus, by noting that the total number of iterations is $\mathcal{T} = \Omega(\log(1/\epsilon))$, we have that the failure probability of the inner loop in Algorithm 1 is bounded by $\mathcal{O}(\epsilon^{20} \cdot \log(1/\epsilon)) \leq \epsilon^{10}$. Thus, we show that when we set the number of inner loop iteration to be $\mathcal{T} = \Omega(\log(1/\epsilon))$ and the number of GTD iterations at the inner critic step to be $T_{\mathrm{TD}}^{\mathrm{in}} = \Omega(\epsilon^{-5})$, it hold that $J(K_{\mathcal{T}}, L) - J(K(L), L) \leq \epsilon$ with probability at least $1 - \epsilon^{-10}$. Thus, we complete the proof of part (i) of Theorem 5.3.

Global Convergence of the Outer Loop. The proof of the global convergence of the outer loop is based on Zhang et al. (2019b), which is based on the population version of the policy gradient and the exact $K(L)$. In contrast to their model-based analysis, our analysis is based on the actor-critic scheme, which requires us to propagate the errors due to the estimation of $K(L)$ and the policy gradient. In the outer loop, we update L via

$$L^{n+1} = \Pi_{\mathcal{L}}\{L^n + \iota \cdot \widehat{F}_n\}, \qquad \widehat{F}_n = \widehat{\Lambda}^{32}_{K^n, L^n} K^n + \widehat{\Lambda}^{33}_{K^n, L^n} L^n - \widehat{\Lambda}^{31}_{K^n, L^n},$$

where $K^n = K^n_{\mathcal{T}}$. By part (i) of Theorem 5.3, we have that K^n approximates $K(L^n)$ in the sense that $J(K^n, L^n) - J(K(L^n), L) \leq \epsilon$. We define

$$\bar{L}^{n+1} = \Pi^n_{\mathcal{L}}\big\{L^n + \iota \cdot F^*_{L^n}\big\}, \tag{B.10}$$

where $\mathcal{L}$ is defined in (4.10) and $\Pi^n_{\mathcal{L}} = \Pi^{L^n}_{\mathcal{L}}$ is defined in (4.13). We use $\bar{L}^{n+1}$ to connect L^{n+1} and L^n. We obtain the following lemma from Zhang et al. (2019b) for the projection operator $\Pi^L_{\mathcal{L}}$.

Lemma B.6 (Lemma 6.3 in Zhang et al. (2019b))**.** The set $\mathcal{L}$ is convex and compact, and the projection operator $\Pi_{\mathcal{L}}$ is convex. For and $L_1, L_2 \in \mathbb{R}^{d \times m_2}$, it holds that

$$\mathrm{Tr}\big(\Sigma^*_L \Pi^L_{\mathcal{L}}\{L_1 - L_2\}^\top (L_1 - L_2)\big) \geq \mathrm{Tr}\big(\Sigma^*_L \Pi^L_{\mathcal{L}}\{L_1 - L_2\}^\top \Pi^L_{\mathcal{L}}\{L_1 - L_2\}\big).$$

By Lemma B.2, we establish the following lemma, which lower bounds the value difference $x^\top P^*_{L'} x - x^\top P^*_L x$.

Lemma B.7. Suppose that $\big(K', L'\big), \big(K(L), L\big)$ are two stable policy pairs. Further, let $x'_0 = x, x'_{t+1} = (A - BK' - CL')x'_t$. Then we have

$$x^\top P_{K', L'} x - x^\top P^*_L x \geq \sum_{t \geq 0} 2{x'_t}^\top (L' - L)^\top F^*_L x'_t - \sum_{t \geq 0} {x'_t}^\top (L' - L)^\top W_L (L' - L) x'_t$$

where W_L takes the form of

$$\begin{aligned} W_L &= S - C^\top P^*_L C + C^\top P^*_L B (R + B^\top P^*_L B)^{-1} B^\top P^*_L C \\ &= S - C^\top \big((P^*_L)^{-1} + BS^{-1}B^\top\big)^{-1} C. \end{aligned}$$

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

Recall that $J^*(L) = \min_K J(K, L)$ is the inner minimum value for a fixed L. We show that $\{J^*(L^n)\}_{n \geq 0}$ is monotonously decreasing by induction. We assume that $J^*(L^n) \leq J^*(L^0)$. We define

$$\widetilde{F}_{L^n} = \iota^{-1} \cdot \big(\Pi^{L^n}_{\mathcal{L}}\{L^n + \iota \cdot F^*_{L^n}\} - L^n\big). \tag{B.11}$$

It then holds that $\bar{L}^{n+1} = L^n + \iota \cdot \widetilde{F}_{L^n}$. We establish the following lemma to demonstrate the increase from $J^*(L^n)$ to $J^*(\bar{L}^{n+1})$.

Lemma B.8. There exist some positive constant $\bar{\iota} > 0$ such that for any stepsize $\iota \in (0, \bar{\iota}]$,

$$J^*(\bar{L}^{n+1}) - J^*(L^n) \geq \iota/2 \cdot \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \|\widetilde{F}_{L^n}\|^2.$$

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

We now bound the difference $J^*(L^{n+1}) - J^*(\bar{L}^{n+1})$. We obtain the following lemma from Zhang et al. (2019b) that demonstrates the Lipschitz continuous of $J^*(L)$.

Lemma B.9 (Proposition B.1 in Zhang et al. (2019b))**.** For any $L, L' \in \mathcal{L}$, there exists a constant $\mathcal{B}_{\mathcal{L}}^L > 0$ such that it holds when $\|L - L'\| \leq \mathcal{B}_{\mathcal{L}}^L$ that

$$\|P_L^* - P_{L'}^*\| \leq \mathcal{B}_{\mathcal{L}}^P \cdot \|L - L'\|,$$

for a constant $\mathcal{B}_{\mathcal{L}}^P$.

By Lemma B.9 and (C.7) of Proposition C.1, we have that

$$|J^*(\bar{L}^{n+1}) - J^*(L^{n+1})| \leq \mathcal{B}_{\mathcal{L}}^P \cdot \|\Psi_\sigma\| \cdot \|L^{n+1} - \bar{L}^{n+1}\|, \tag{B.12}$$

for $\|L^{n+1} - \bar{L}^{n+1}\| \leq \mathcal{B}_{\mathcal{L}}^L$. By the definition of $\bar{L}^n$ in (B.10) and the convexity of $\Pi_{\mathcal{L}}$ in Lemma B.6, we have that

$$\|L^{n+1} - \bar{L}^{n+1}\| \leq \iota \cdot \|\widehat{F}_n - F_{L^n}^*\|. \tag{B.13}$$

Thus, we proceed to bound the difference between $\widehat{F}_n$ and $F_{L^n}^*$. By the triangle inequality, we have

$$\|\widehat{F}_n - F_{L^n}^*\| \leq \|\widehat{F}_n - F_{K^n,L^n}\| + \|F_{K^n,L^n} - F_{L^n}^*\|. \tag{B.14}$$

By (B.12), (B.13), and (B.14), we have that

$$\big|J^*(\bar{L}^{n+1}) - J^*(L^{n+1})\big| \leq \iota \cdot \mathcal{B}_{\mathcal{L}}^P \cdot \|\Psi_\sigma\| \cdot \big(\|\widehat{F}_n - F_{K^n,L^n}\| + \|F_{K^n,L^n} - F_{L^n}^*\|\big). \tag{B.15}$$

For the first term on the right-hand side of (B.15), we have that

$$\|\widehat{F}_n - F_{K^n,L^n}\| \leq \|\widehat{\Lambda}_{K^n,L^n} - \Lambda_{K^n,L^n}\| \cdot (\|K^n\| + \|L^n\| + 1). \tag{B.16}$$

Thus, by Theorem 5.2, the difference $\|\widehat{F}_n - F_{K^n,L^n}\|$ is sufficiently small when we set the number of GTD iterations $T_{\text{TD}}^{\text{out}}$ at the outer critic step to be sufficiently large. For the second term on the right-hand side of (B.15), we establish the following lemma. The following lemma bounds the difference $\|F_{K^n,L^n} - F_{L^n}^*\|$ on the right of (B.15).

Lemma B.10. For stable policy pair (K, L), it holds that that

$$\|F_{K,L} - F_L^*\| \leq C \cdot \Big(J(K, L) - J\big(K(L), L\big)\Big),$$

where $C > 0$ is a constant.

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

Thus, by plugging (B.16) and Lemma B.10 into (B.15), we have that

$$\begin{aligned}
&|J^*(\bar{L}^{n+1}) - J^*(L^{n+1})| \\
&\quad \leq \iota \cdot \mathcal{B}_{\mathcal{L}}^P \cdot \|\Psi_\sigma\| \cdot \Big(\|\widehat{\Lambda}_{K^n,L^n} - \Lambda_{K^n,L^n}\| \cdot \big(\|K^n\| + \|L^n\| + 1\big) + C \cdot \big(J(K^n, L^n) - J^*(L^n)\big)\Big).
\end{aligned} \tag{B.17}$$

By Lemma B.8, to establish the increase from $J^*(L^n)$ to $J^*(L^{n+1})$, it suffices to show that

$$|J^*(\bar{L}^{n+1}) - J^*(L^{n+1})| \leq \iota/4 \cdot \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \|\widetilde{F}_{L^n}\|^2. \tag{B.18}$$

By comparing (B.17) and (B.18), we note that (B.18) holds if the following two inequalities holds simultaneously

$$\|\widehat{\Lambda}_{K^n,L^n} - \Lambda_{K^n,L^n}\| \leq \frac{1/8 \cdot \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \|\widetilde{F}_{L^n}\|^2}{\mathcal{B}_{\mathcal{L}}^P \cdot \|\Psi_\sigma\| \cdot (\|K^n\| + \|L^n\| + 1)}, \tag{B.19}$$

$$J(K^n, L^n) - J^*(L^n) \leq \frac{1/8 \cdot \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \|\widetilde{F}_{L^n}\|^2}{C \cdot \mathcal{B}_{\mathcal{L}}^P \cdot \|\Psi_\sigma\|}. \tag{B.20}$$

We now fix a $\varepsilon > 0$ and consider $\|\widetilde{F}_{L^n}\| \geq \epsilon$. By Theorem 5.2, if we set the number of GTD iterations at the outer critic step $T_{\mathrm{TD}}^{\mathrm{out}} = \Omega(\varepsilon^{-10})$, we have that the condition (B.19) holds with failure probability $\mathcal{O}(\varepsilon^{40})$. Meanwhile, following from (B.9) by setting the number of inner loop iterations $\mathcal{T} = \Omega(\log(1/\varepsilon))$ and the number of GTD iterations at the inner critic step $T_{\mathrm{TD}}^{\mathrm{in}} = \Omega(\varepsilon^{-10})$, we have that the condition (B.20) holds with failure probability $\mathcal{O}(\varepsilon^{20})$. Note that when (B.19) and (B.20) hold, by (B.18) and Lemma B.8, we have that

$$\|\widetilde{F}_{L^n}\|^2 \leq \frac{4\big(J^*(L^{n+1}) - J^*(L^n)\big)}{\sigma_{\min}(\Psi_\sigma)}. \tag{B.21}$$

Upon telescoping (B.21), we have that

$$\frac{1}{\mathcal{N}} \sum_{n=0}^{\mathcal{N}-1} \|\widetilde{F}_{L^n}\|_{\mathrm{F}} \leq \sqrt{\frac{1}{\mathcal{N}} \sum_{n=0}^{\mathcal{N}-1} \|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2} \leq 2\sqrt{\frac{\sigma_{\min}(\Psi_\sigma)\big[J^* - J^*(L^0)\big]}{\iota \cdot \mathcal{N}}}.$$

By setting $\mathcal{N} = \Omega(\varepsilon^{-2})$, we have that $\mathcal{N}^{-1} \cdot \sum_{n=0}^{\mathcal{N}-1} \|\widetilde{F}_{L^n}\| \leq \varepsilon$. By the union bound, the failure probability is $\mathcal{N} \cdot (\mathcal{O}(\varepsilon^{20}) + \mathcal{O}(\varepsilon^{40})) \leq \varepsilon^{10}$. Thus, we complete the proof of Theorem 5.3. □

C Policy Gradient Theorem in the LQ setting

We first give a detailed version of Proposition 4.2. The proof is a modification of that of Proposition 3.1 in Yang et al. (2019), which establishes the policy gradient theorem for the single-agent linear quadratic regulator under the ergodic setting. Compared with Yang et al. (2019), we need to consider the system determined by the two competing players. Note that Proposition C.1 recovers Proposition 3.1 in Yang et al. (2019) when $L = 0$.

Proposition C.1. We assume that the policy $\pi_{K,L}$is stable in the sense that $\rho(A - BK - CL) < 1$. Let $P_{K,L}, \Sigma_{K,L}$ be the unique positive definite solutions to the following Lyapunov equations

$$P_{K,L} = (Q + K^\top RK - L^\top SL) + (A - BK - CL)^\top P_{K,L}(A - BK - CL), \tag{C.1}$$

$$\Sigma_{K,L} = \Psi_\sigma + (A - BK - CL)\Sigma_{K,L}(A - BK - CL)^\top. \tag{C.2}$$

Then, the Markov chain $x_{t+1} = (A - BK - CL)x_t + \epsilon_t$ has a stationary distribution $\rho_{K,L} = N(0, \Sigma_{K,L})$. We denote the second tensor power of state-action pair (x, u, v) by $\Phi(x, u, v) = (x, u, v)^{\otimes 2}$. We define the parameter matrix $\Lambda_{K,L}$ for the function $Q_{K,L}$ as follows,

$$\Lambda_{K,L} = \begin{pmatrix} \Lambda_{K,L}^{11} & \Lambda_{K,L}^{12} & \Lambda_{K,L}^{13} \\ \Lambda_{K,L}^{21} & \Lambda_{K,L}^{22} & \Lambda_{K,L}^{23} \\ \Lambda_{K,L}^{31} & \Lambda_{K,L}^{32} & \Lambda_{K,L}^{33} \end{pmatrix} = \begin{pmatrix} Q + A^\top P_{K,L}A & A^\top P_{K,L}B & A^\top P_{K,L}C \\ B^\top P_{K,L}A & R + B^\top P_{K,L}B & B^\top P_{K,L}C \\ C^\top P_{K,L}A & C^\top P_{K,L}B & -S + C^\top P_{K,L}C \end{pmatrix}. \tag{C.3}$$

Then, the state and state-action value functions are quadratic functions, taking the form

$$V_{K,L}(x) = x^\top P_{K,L}x - \mathrm{Tr}(P_{K,L}\Sigma_{K,L}), \tag{C.4}$$

$$Q_{K,L}(x, u, v) = \mathrm{Tr}\big(\Lambda_{K,L}\Phi(x, u, v)\big) + q_{K,L}, \tag{C.5}$$

where $q_{K,L} = -\sigma_1^2 \cdot \mathrm{Tr}(R + P_{K,L}BB^T) - \sigma_2^2 \cdot \mathrm{Tr}(-S + P_{K,L}CC^\top)$ is independent of (x, u, v). We have the following Bellman equation,

$$\mathrm{Tr}\big(\Lambda_{K,L}\Phi(x, u, v)\big) = c(x, u, v) - J(K, L) + \mathbb{E}^{\pi_{K,L}}\Big[\mathrm{Tr}\big(\Lambda_{K,L}\Phi(x', u', v')\big) \,\Big|\, x, u, v\Big], \tag{C.6}$$

where (x', u', v') denotes the subsequent state-action pair following policy $\pi_{K,L}$. Further, for the ergodic cost $J(K, L)$, it holds that

$$\begin{aligned} J(K,L) &= \mathrm{Tr}\big((Q + K^\top RK - L^\top SL)\Sigma_{K,L}\big) + \sigma_1^2 \mathrm{Tr}(R) - \sigma_2^2 \mathrm{Tr}(S), \\ &= \mathrm{Tr}(P_{K,L}\Psi_\sigma) + \sigma_1^2 \mathrm{Tr}(R) - \sigma_2^2 \mathrm{Tr}(S), \end{aligned} \tag{C.7}$$

$$\nabla_K J(K,L) = 2\big((R + B^\top P_{K,L} B)K - B^\top P_{K,L}(A - CL)\big)\Sigma_{K,L} = 2E_{K,L}\Sigma_{K,L}, \tag{C.8}$$

$$\nabla_L J(K,L) = 2\big((-S + C^\top P_{K,L} C)L - C^\top P_{K,L}(A - BK)\big)\Sigma_{K,L} = 2F_{K,L}\Sigma_{K,L}. \tag{C.9}$$

Here $E_{K,L}$ and $F_{K,L}$ are given by

$$E_{K,L} = (R + B^\top P_{K,L} B)K - B^\top P_{K,L}(A - CL) = \Lambda_{K,L}^{22} K + \Lambda_{K,L}^{23} L - \Lambda_{K,L}^{21}, \tag{C.10}$$

$$F_{K,L} = (-S + C^\top P_{K,L} C)L - C^\top P_{K,L}(A - BK) = \Lambda_{K,L}^{33} L + \Lambda_{K,L}^{32} K - \Lambda_{K,L}^{31}. \tag{C.11}$$

where $\Lambda_{K,L}$ is defined in (C.3). Further, $2E_{K,L}$ and $2F_{K,L}$ corresponds to the natural gradient of $J(K, L)$.

Proof. By direct computation, the Markov chain defined in (4.2) has a stationary distribution $N(0, \Sigma_{K,L})$, where $\Sigma_{K,L}$ is the positive definite solution to the Lyapunov equation (C.2). By the definition of $J(K, L)$ in (3.1), the ergodic cost $J(K, L)$ is the the cost evaluated on the stationary distribution $\rho_{K,L}$, which implies that

$$\begin{aligned} J(K,L) &= \mathbb{E}_{x\sim\rho_{K,L}, u\sim\pi_{K,L}(\cdot\,|\,x)}\big[c(x,u,v)\big] \\ &= \mathrm{Tr}\big((Q + K^\top RK - L^\top SL)\Sigma_{K,L}\big) + \sigma_1^2 \mathrm{Tr}(R) - \sigma_2^2 \mathrm{Tr}(S). \end{aligned} \tag{C.12}$$

This establishes the first equality of (C.7).

For notational simplicity, we define the following operators for the stable policy pair (K, L)

$$\mathcal{T}_{K,L}(\Omega) = \sum_{t\geq 0} (A - BK - CL)^t \Omega \big((A - BK - CL)^t\big)^\top, \tag{C.13}$$

$$\mathcal{T}_{K,L}^\top(\Omega) = \sum_{t\geq 0} \big((A - BK - CL)^t\big)^\top \Omega (A - BK - CL)^t, \tag{C.14}$$

which satisfy the following Lyapunov equations

$$\mathcal{T}_{K,L}(\Omega) = \Omega + (A - BK - CL)\mathcal{T}_{K,L}(\Omega)(A - BK - CL)^\top, \tag{C.15}$$

$$\mathcal{T}_{K,L}^\top(\Omega) = \Omega + (A - BK - CL)^\top \mathcal{T}_{K,L}(\Omega)(A - BK - CL). \tag{C.16}$$

For any positive definite matrices Ω_1 and Ω_2, since $\rho(A - BK - CL) < 1$, it holds that

$$\begin{aligned} \mathrm{Tr}\big(\Omega_1 \mathcal{T}_{K,L}(\Omega_2)\big) &= \sum_{t\geq 0} \mathrm{Tr}\Big(\Omega_1 (A - BK - CL)^t \Omega_2 \big((A - BK - CL)^t\big)^\top\Big) \\ &= \sum_{t\geq 0} \mathrm{Tr}\Big(\big((A - BK - CL)^t\big)^\top \Omega_1 (A - BK - CL)^t \Omega_2\Big) \\ &= \mathrm{Tr}\big(\mathcal{T}_{K,L}^\top(\Omega_1)\Omega_2\big). \end{aligned} \tag{C.17}$$

Combining (C.1), (C.2), (C.15), and (C.16), we have $P_{K,L} = \mathcal{T}_{K,L}^\top(Q + K^\top RK - L^\top SL)$ and $\Sigma_{K,L} = \mathcal{T}_{K,L}(\Psi_\sigma)$. Thus, by (C.17), it holds that

$$\begin{aligned} \mathrm{Tr}\big((Q + K^\top RK - L^\top SL)\Sigma_{K,L}\big) &= \mathrm{Tr}\big((Q + K^\top RK - L^\top SL)\mathcal{T}_{K,L}(\Psi_\sigma)\big) \\ &= \mathrm{Tr}\big(\mathcal{T}_{K,L}^\top(Q + K^\top RK - L^\top SL)\Psi_\sigma\big) = \mathrm{Tr}(P_{K,L}\Psi_\sigma). \end{aligned}$$

Combining with (C.12), we establish the second equality of (C.7). Second, we proceed to establish the state value function $V_{K,L}(x)$ in (C.4) and state-action value function $Q_{K,L}$ in (C.5). By the definition of $V_{K,L}(x)$ in (3.2), it holds that

$$\begin{aligned} V_{K,L}(x) &= \sum_{t\geq 0} \mathbb{E}^{\pi_{K,L}}\big[c(x_t, u_t, v_t)\,|\,x_0 = x, u_t = -Kx_t + \sigma_1\eta_t^1, v_t = -Lx_t + \sigma_2\eta_t^2\big] \\ &= \sum_{t\geq 0} \Big(\mathbb{E}^{\pi_{K,L}}\big[x_t^\top (Q + K^\top RK - L^\top SL)x_t \,\big|\, x_0 = x\big] - J(K) + \sigma_1^2 \mathrm{Tr}(R) - \sigma_2^2 \mathrm{Tr}(S)\Big). \end{aligned}$$

Combining with the linear transition (4.2), $V_{K,L}(x)$ is quadratic in x. We write $V_{K,L}(x) = x^\top \mathbf{V}_{K,L} x + v_{K,L}$ for matrix $\mathbf{V}_{K,L}$ and constant $v_{K,L}$, which are determined in the sequel. Note that $V_{K,L}(x)$ satisfies the following Bellman equation,

$$V_{K,L}(x) = \mathbb{E}_{(u,v)\sim\pi_{K,L}}\big[c(x,u,v)\big] - J(K,L) + \mathbb{E}_{K,L}\big[V(x') \,\big|\, x\big],$$

where x' is the subsequent state of x following the policy $\pi_{K,L}$. Therefore, the matrix $\mathbf{V}_{K,L}$ satisfies the following equation.

$$x^\top \mathbf{V}_{K,L} x = x^\top (Q + K^\top RK - L^\top SL)x + x^\top (A - BK - CL)^\top \mathbf{V}_{K,L}(A - BK - CL)x,$$

which implies that $\mathbf{V}_{K,L}$ satisfies the same Lyapunov equation as $P_{K,L}$. By the uniqueness of $P_{K,L}$, we deduce that $\mathbf{V}_{K,L} = P_{K,L}$. Combining with the fact that $\mathbb{E}_{x\sim\rho_{K,L}}[V_{K,L}(x)] = 0$, we complete the proof of (C.4).

Furthermore, by the definition of $Q_{K,L}(x,u,v)$ in (3.3), we have the following equation.

$$Q_{K,L}(x,u,v) = c(x,u,v) - J(K) + \mathbb{E}_{K,L}\big[V_{K,L}(x') \,|\, x,u,v\big]. \tag{C.18}$$

Combining (C.4) and (C.18), by direct computation, we obtain (C.5), and then the Bellman equation (C.6) of $Q_{K,L}$ follows.

In the following, we proceed to characterize the expressions of policy gradient (C.8) and (C.9). By (C.7), it holds that

$$\nabla_K J(K,L) = 2RK\Sigma_{K,L} + \nabla_K \mathrm{Tr}(Q_0 \Sigma_{K,L}), \tag{C.19}$$
$$\nabla_L J(K,L) = -2SL\Sigma_{K,L} + \nabla_L \mathrm{Tr}(Q_0 \Sigma_{K,L}), \tag{C.20}$$

where $Q_0 = Q + K^\top RK - L^\top L$. We first compute the expression (C.8). Note that $\mathcal{T}_{K,L}(\Psi_\sigma) = \Sigma_{K,L}$ and (C.15), for the second term on the right of (C.19), we have that

$$\begin{aligned}\nabla_K \mathrm{Tr}(Q_0\Sigma_{K,L}) &= \nabla_K \mathrm{Tr}\big[(A - BK - CL)^\top Q_0 (A - BK - CL)\Sigma_{K,L}\big] \\ &= -2B^\top Q_0 (A - BK - CL)\Sigma_{K,L} + \nabla_K \mathrm{Tr}(Q_1 \Sigma_{K,L}),\end{aligned}$$

where $Q_1 = (A - BK - CL)^\top Q_0 (A - BK - CL)$. Moreover, we denote by $Q_t := (A - BK - CL)^{t\top} Q_0 (A - BK - CL)^t$. Note that $\sum_{t=0}^\infty Q_t = P_{K,L}$ and that $Q_t \to 0$ when t goes to infinity, we have that

$$\begin{aligned}\nabla_K \mathrm{Tr}(Q_0\Sigma_{K,L}) &= -2B^\top \big[\sum_{t=0}^n Q_t\big](A - BK - CL)\Sigma_{K,L} + \nabla_K \mathrm{Tr}\big(Q_t \Sigma_{K,L}\big) \\ &= -2B^\top P_{K,L}(A - BK - CL)\Sigma_{K,L}.\end{aligned}$$

Combining with (C.19), we establish (C.8). Similarly, we establish (C.9).

Finally, we establish the fact that $E_{K,L}$ and $F_{K,L}$ relate with the natural policy gradients. For the Gaussian policy $\pi_{K,L}$ defined in (4.1), the corresponding Fisher information, denoted by $\mathcal{I}(K,L)$, has the following structure

$$\begin{aligned}\big[\mathcal{I}(K;L)\big]_{K_{ij}K_{i'j'}} &= \mathbb{E}_{x\sim\rho_{K,L},(u,v)\sim\pi_{K,L}}\big[\nabla_{K_{ij}} \log \pi_{K,L}(u,v\,|\,x) \nabla_{K_{i'j'}} \log \pi_{K,L}(u,v\,|\,x)\big] \\ &= \sigma_1^{-2} \cdot \mathbb{E}_{x\sim\rho_{K,L},(u,v)\sim\pi_{K,L}}\big[\eta_i^1 x_j \cdot \eta_{i'}^1 x_{j'}\big] = \sigma_1^{-2} \cdot \mathbb{1}_{i=i'} \cdot (\Sigma_{K,L})_{jj'}.\end{aligned}$$

Similarly, we have that

$$\big[\mathcal{I}(L;K)\big]_{L_{ij}L_{i'j'}} = \sigma_2^{-2} \cdot \mathbb{1}_{i=i'} \cdot (\Sigma_{K,L})_{jj'}.$$

Thus, the natural policy gradients take the form of

$$\begin{aligned}\big[\mathcal{I}(K;L)\big]^{-1} \nabla_K J(K,L) &= \nabla_K J(K,L)\Sigma_{K,L}^{-1} = 2E_{K,L}, \\ \big[\mathcal{I}(L;K)\big]^{-1} \nabla_L J(K,L) &= \nabla_L J(K,L)\Sigma_{K,L}^{-1} = 2F_{K,L}.\end{aligned}$$

Therefore, we conclude the proof of Proposition (C). □

D Gradient Temporal-Difference

In this section, we present the GTD algorithm that estimate $\Lambda_{K,L}$ defined in (C.3) at the critic steps. Our GTD is based on Yang et al. (2019), which develop GTD for the single-agent LQR problem. We extend this approach to the two-agent zero-sum LQ game. For notational simplicity, we denote by $z = (x, u, v)$ the state-action pair, and by $\mathcal{Z} = \mathcal{X} \times \mathcal{U} \times \mathcal{V}$ the state-action space.

To solve the RSRL and RARL, we need to sample from the MDP $\mathcal{M}$ for the two-player zero-sum game introduced in §2.3, while we only have access to the original MDP $\bar{\mathcal{M}} = (\mathcal{X}, \mathcal{U}, \bar{P}, \bar{c}, \mathcal{D}_0)$ in §2.1 for the single-agent RSRL and RARL. To sample from $\mathcal{M}$ in the sequel, we first sample the initial state $x_0 \sim \mathcal{D}_0$. At state x_t $(t \geq 0)$, the agent takes action u_t and receives a cost $\bar{c}_t$. Then, the system transits to the next state $\bar{x}_{t+1}$. Letting $c_t = \bar{c}_t - v_t^\top S v_t$ and $x_{t+1} = \bar{x}_{t+1} + C v_t$, we have that $\{(x_t, c_t)\}_{t \geq 0}$ is a trajectory from the MDP $\mathcal{M}$ by following $\{u_t\}$ and $\{v_t\}$.

We define the following feature vector

$$\phi(z) = \mathrm{svec}\big[(z)^{\otimes 2}\big], \tag{D.1}$$

for any $z \in \mathcal{Z}$. We denote the vectorization of $\Lambda_{K,L}$ by $\lambda^* = \mathrm{svec}(\Lambda_{K,L})$. Note that we omit the dependence of λ^* on (K, L) for notational simplicity. Then, we write the Bellman equation in (4.7) of state-action function as follows,

$$\langle \phi(z), \lambda^* \rangle = c(z) - J(K, L) + \big\langle \mathbb{E}_{z' \sim \widetilde{P}(\cdot \,|\, z; \pi_{K,L})}\big[\phi(z')\big], \lambda^* \big\rangle.$$

Here $z' = (x', u', v')$, and the expectation $\mathbb{E}_{z' \sim \widetilde{P}(\cdot \,|\, z)}$ is with respect to the conditional distribution $\widetilde{P}(\cdot \,|\, z; \pi_{K,L})$ that takes the form of

$$\widetilde{P}(z' \,|\, z; \pi_{K,L}) = P(x' \,|\, z) \cdot \pi_{K,L}(u', v' \,|\, x').$$

We define

$$\beta^* = \begin{pmatrix} (\beta^*)^1 \\ (\beta^*)^2 \end{pmatrix} = \begin{pmatrix} J(K, L) \\ \lambda^* \end{pmatrix}. \tag{D.2}$$

Along the formulation of Liu et al. (2015); Yang et al. (2019), we define the following notations,

$$\begin{aligned} \mathfrak{T}_{K,L} &= \mathbb{E}_{z \sim \widetilde{\rho}_{K,L}, z' \sim \widetilde{P}(\cdot \,|\, z; \pi_{K,L})}\Big[\phi(z)\big[\phi(z) - \phi(z')\big]^\top\Big], \\ b_{K,L} &= \mathbb{E}_{z \sim \widetilde{\rho}_{K,L}, z' \sim \widetilde{P}(\cdot \,|\, z; \pi_{K,L})}\big[c(z)\phi(z)\big]. \end{aligned} \tag{D.3}$$

Here $\widetilde{\rho}_{K,L}$ is the stationary distribution induced by $\pi_{K,L}$ over $\mathcal{Z}$, which takes the form of

$$\widetilde{\rho}_{K,L}(z) = \rho_{K,L}(x) \cdot \pi_{K,L}(u, v \,|\, x), \qquad \forall z = (x, u, v) \in \mathcal{X} \times \mathcal{U} \times \mathcal{V}.$$

Then, by direct computation, we have the following linear equation

$$\begin{pmatrix} 1 & 0 \\ \mathbb{E}_{(z) \sim \widetilde{\rho}_{K,L}}\big[\phi(z)\big] & \mathfrak{T}_{K,L} \end{pmatrix} \beta^* = \begin{pmatrix} J(K, L) \\ b_{K,L} \end{pmatrix}. \tag{D.4}$$

The solution of (D.4) is unique if the matrix $\mathfrak{T}_{K,L}$ is invertible, which is verified by the following lemma.

Lemma D.1. For any policy $\pi_{K,L}$ that is stable in the sense that $\rho(A - BK - CL) < 1$, the matrix $\mathfrak{T}_{K,L}$ defined in (D.3) is invertible, and it holds that

$$\|\mathfrak{T}_{K,L}\| \leq 2\big(\sigma_1^2 + \sigma_2^2 + (1 + \|K\|^2 + \|L\|^2) \cdot \|\Sigma_{K,L}\|\big)^2.$$

It then holds that $\beta^*_{K,L}$ is the unique solution of (D.4).

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

The linear equation defined in (D.4) is equivalent to the following minimization problem,

$$\min_{\beta} |\beta^1 - J(K,L)|^2 + \big\|\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big]\beta^1 + \mathfrak{T}_{K,L}\beta^2 - b_{K,L}\big\|^2,$$

where the expectation is with respect to By Fenchel duality, solving the above linear equation is equivalent to the following minimax problem,

$$\min_{\beta\in\mathcal{X}_\beta}\max_{w\in\mathcal{X}_w} G(\beta,w) = \mathbb{E}_{z\sim\widetilde{\rho}_{K,L},z'\sim\widetilde{P}(\cdot\,|\,z;\pi_{K,L})}\big[g(\beta,w;z,z')\big], \tag{D.5}$$

where g is defined as follows,

$$g(\beta,w;z,z') = \big[\beta^1 - c(z)\big]\cdot w^1 - \frac{1}{2}\cdot\|w\|_2^2 + \big\langle\phi(z)\cdot\big\{\beta^1 + \big[\phi(z)-\phi(z')\big]^\top\beta^2 - c(z)\big\}, w^2\big\rangle. \tag{D.6}$$

By the property of Fenchel duality and (D.4), the optimal solution of (D.5) is $(\beta^*, 0)$. We note that $g(\beta,w;z,z')$ is convex with respect to β and concave with respect to w. To solve minimax problem defined in (D.5), we apply stochastic primal-dual update to (β,w) as follows,

$$\beta_{t+1} = \Pi_{\mathcal{X}_B}\Big\{\beta_t - \alpha_t\cdot\nabla_\beta g(\beta_t,w_t;z_t,z_{t+1})\Big\},$$
$$w_{t+1} = \Pi_{\mathcal{X}_W}\Big\{w_t - \alpha_t\cdot\nabla_w g(\beta_t,w_t;z_t,z_{t+1})\Big\},$$

where $\big\{z_t\big\}_{t\geq 0}$ is drawn from $z_0\sim\mathcal{D}, z_{t+1}\sim\widetilde{P}(\cdot\,|\,z_t;\pi_{K,L})$. Here for the stability of the algorithm, we project β and w to some compact sets $\mathcal{X}_B$ and $\mathcal{X}_W$, which are defined as follows,

$$\mathcal{X}_B = \Big\{\beta : 0\leq\beta^1\leq J(K,L), \|\beta^2\|\leq\bar{R}_B\Big\}, \tag{D.7}$$
$$\mathcal{X}_W = \Big\{w : |w^1|\leq J(K,L), \|w^2\|_2\leq\bar{R}_W\Big\}. \tag{D.8}$$

Here $\bar{R}_B$ and $\bar{R}_W$ are defined by

$$\bar{R}_B = C\cdot\big(J(K,L)/\sigma_{\min}(\Psi_\sigma)\big) + C_0, \tag{D.9}$$
$$\bar{R}_W = C\cdot\bar{R}_B\cdot\sigma_{\min}(Q - L^\top SL)^{-2}\cdot\big(d + \|K\|_{\mathrm{F}}^2 + \|L\|_{\mathrm{F}}^2\big)^2\cdot J(K,L)^2 + C_0, \tag{D.10}$$

for absolute constants $C>0$ and C_0. We conclude the above discussion in Algorithm 3.

We establish Theorem 5.2 that characterizes the sublinear convergence of Algorithm 3. For the completeness, we restate the theorem as follows.

Theorem D.2 (Convergence of GTD (Algorithm 3))**.** Let $\widehat{\Lambda}$ be the output of Algorithm 3 after T iterations. We set the stepsize $\alpha_t = 1/\sqrt{t}$. Then, For the sufficiently large T, it holds with probability at least $1-T^{-4}$ that

$$\|\widehat{\Lambda} - \Lambda_{K,L}\|_{\mathrm{F}}^2 \leq \frac{\mathrm{Poly}\Big(J(K,L), \|K\|_{\mathrm{F}}, \|L\|_{\mathrm{F}}, \sigma_{\min}(Q - L^\top SL)^{-1}\Big)}{\sigma_{\min}(\mathfrak{T}_{K,L})\cdot\big(1-\rho(A-BK-CL)\big)}\cdot\frac{\log^4 T}{\sqrt{T}},$$

where Poly is a polynomial.

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

Algorithm 3 GTD for Estimating $\Lambda_{K,L}$.

Input: Policy $\pi_{K,L}$, number of GTD iterations T_{TD}, and step sizes $\{\alpha_t\}_{t=0}^T$.
Initialize the primal and dual variables $\beta_0 \in \mathcal{X}_\beta$ and $w_0 \in \mathcal{X}_w$.
Sample $x_0 \sim \mathcal{D}$. Take action $(u_0, v_0) \sim \pi(\cdot \,|\, x_0)$, and observe the cost c_0 and next state x_1. Let $z_0 = (x_0, u_0, v_0)$
for $t = 1, 2, ..., T$ **do**
Take action $(u_t, v_t) \sim \pi_{K,L}(\cdot \,|\, x_t)$, and observe the cost c_t and next state x_{t+1}. Let $z_t = (x_t, u_t, v_t)$ and $\phi(z) = \mathrm{svec}(zz^\top)$.
$\beta_t^1 = \beta_{t-1}^1 - \alpha_t \cdot \Big[w_{t-1}^1 + \langle \phi(z_{t-1}), w_{t-1}^2 \rangle\Big]$,
$\beta_t^2 = \beta_{t-1}^2 - \alpha_t \cdot \Big[\big(\phi(z_{t-1}) - \phi(z_t)\big)\phi(z_{t-1})^\top w_{t-1}^2\Big]$,
$w_t^1 = (1-\alpha_t) \cdot w_{t-1}^1 + \alpha_t \cdot [\beta_{t-1}^1 - c_{t-1}]$,
$w_t^2 = (1-\alpha_t) \cdot w_{t-1}^2 + \alpha_t \cdot \phi(z_{t-1})\Big[\beta_{t-1}^1 + \big[\phi(z_{t-1}) - \phi(z_t)\big]^\top \beta_{t-1}^2 - c_{t-1}\Big]$.
Project β_t and w_t to the space $\mathcal{X}_B$ and $\mathcal{X}_W$ specified in (D.7) and (D.8).
end for
$\widehat{\beta} = \frac{\sum_{t=1}^T \alpha_t \beta_t}{\sum_{t=1}^T \alpha_t}, \widehat{w} = \frac{\sum_{t=1}^T \alpha_t w_t}{\sum_{t=1}^T \alpha_t}$.
Output: $\widehat{\Lambda} = \mathrm{smat}(\widehat{\beta}^2)$ that estimates $\Lambda_{K,L}$.

D.1 Proof of Theorem 5.2

Proof. The proof of Theorem 5.2 is based on Yang et al. (2019). In comparison, we establish the global convergence of GTD for the two-agent zero sum LQ game, whereas they do it for the single-agent LQR. The proof of Theorem 5.2 takes three steps. First, we prove that the optimal solution of (D.5) is contained in the parameter domain $\mathcal{X}_B \times \mathcal{X}_W$. Second, we define the optimality gap of the minimax problem in (D.5), and relate the gap to the difference between current iterates and the optimal solution. At last, we prove that the optimality gap converges to zero sublinearly.

We first establish the following lemma that shows the saddle point of the objective function defined in (D.5) is contained in the parameter domain $\mathcal{X}_B \times \mathcal{X}_W$.

Lemma D.3. For any $\beta \in \mathcal{X}_B$, we define

$$w(\beta) = \operatorname*{argmax}_w G(\beta, w).$$

Then, it holds that $w(\beta) \in \mathcal{X}_W$. Moreover, the optimal solution $(\beta^*, 0)$ of (D.5) is contained in $\mathcal{X}_W \times \mathcal{X}_B$.

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

To analyze the convergence of Algorithm 3, we define the optimality gap for (D.5) as

$$\mathrm{Gap}(\widehat{\beta}, \widehat{w}) = \max_{w \in \mathcal{X}_W} G(\widehat{\beta}, w) - \min_{\beta \in \mathcal{X}_B} G(\beta, \widehat{w}), \tag{D.11}$$

for any $(\widehat{\beta}, \widehat{w}) \in \mathcal{X}_B \times \mathcal{X}_W$. We establish the following lemma relating the primal-dual gap defined in (D.11) to the gap between $(\widehat{\beta}, \widehat{w})$ and the saddle point $(\beta^*, 0)$.

Lemma D.4 (Bound Parameter Difference via Objective Difference)**.** For any $(\widehat{\beta}, \widehat{w}) \in \mathcal{X}_B \times \mathcal{X}_W$, let $\widehat{\Lambda} = \mathrm{smat}(\beta^2)$. Then we have

$$\big|\widehat{\beta}^1 - J(K,L)\big|^2 + \|\widehat{\Lambda} - \Lambda_{K,L}\|_{\mathrm{F}}^2 \leq \frac{\mathrm{Gap}(\widehat{\beta}, \widehat{w})}{\sigma_{\min}(\mathfrak{T}_{K,L})^2},$$

where $\mathfrak{T}_{K,L}$ is defined in (D.3).

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

Note that Algorithm 3 is a stochastic gradient based method. We have

$$G(\beta, w) = \mathbb{E}_{z\sim\rho_{K,L}, z'\sim\widetilde{P}(\cdot\,|\,z)}\big[g(\beta, w; z, z')\big]$$

where g is defined in (D.6). The expectation is taken with respect to $x \sim \rho_{K,L}$, where $\widetilde{\rho}_{K,L}$ is the stationary distribution induced by policy $\pi_{K,L}$. But the trajectory $\{z_t\}_{t\geq 0}$ is generated from $z_0 \sim \mathcal{D}, z_{t+1} \sim \widetilde{P}(\cdot\,|\,z)$, which implies that $\{z_t\}$ is not drawn from the stationary distribution $\widetilde{\rho}_{K,L}$. Such a case has been studied in Wang et al. (2017). However, different from the setting in Wang et al. (2017), our function $g(\beta, w; z, z')$ is not Lipschitz continuous in (β, w), which makes the analysis of Wang et al. (2017) not applicable for our setting. To bypass such an issue, we prove that the function g is Lipschitz continuous in (β, w) with high probability, which allows us to apply Theorem 1 in Wang et al. (2017) to establish the following lemma.

Lemma D.5. It holds for the output $(\widehat{\beta}, \widehat{w})$ of Algorithm 3 after T iterations that

$$\mathrm{Gap}(\bar{\beta}, \bar{w}) \leq \mathrm{Poly}\big(J(K,L), \bar{R}_B, \bar{R}_W, \sigma_{\min}(Q - L^\top SL)^{-1}\big) \cdot \frac{(\log T)^4}{(1-\rho)\sqrt{T}},$$

with probability at least $1 - T^{-4}$.

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

Combining Lemma D.3, Lemma D.4 and Lemma D.5, we complete the proof of Theorem 5.2. □

D.2 Proof of Lemma D.1

Recall that we denote by $z = (x, u, v)$ the state action pair. and by $z' = (x', u', v')$ the subsequent state action pair following $\pi_{K,L}$. Then, by (2.5), we have that,

$$z' = Uz + \zeta, \tag{D.12}$$

where U and ζ are defined as follows.

$$U := \begin{pmatrix} I_d \\ -K \\ -L \end{pmatrix} \begin{pmatrix} A & B & C \end{pmatrix}, \quad \zeta_t := \begin{pmatrix} e_t \\ -Ke_t + \sigma_1\eta^1 \\ -Le_t + \sigma_2\eta^2 \end{pmatrix}. \tag{D.13}$$

Here $\eta_1 \sim N(0, I_{m_1}), \eta_2 \sim N(0, I_{m_2})$ are Gaussian noises. By the fact that $\rho(MN) = \rho(NM)$ for any matrices M, N and that $\rho(A - BK - CL) < 1$, we have that $\rho(U) < 1$. We denote the covariance matrix of ζ_t by $\widetilde{\Psi}_\sigma$, which takes the form of

$$\widetilde{\Psi}_\sigma = \begin{pmatrix} I_d \\ -K \\ -L \end{pmatrix} \Psi \begin{pmatrix} I_d \\ -K \\ -L \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_1^2 I_{m_1} & 0 \\ 0 & 0 & \sigma_2^2 I_{m_2} \end{pmatrix}.$$

Recall that we denote by $\widetilde{\rho}_{K,L}$ the stationary distribution of (x, u, v) induced by $\pi_{K,L}$. We denote the covariance matrix of $\widetilde{\rho}_{K,L}$ by $\widetilde{\rho}_{K,L}$, which takes the form of

$$\widetilde{\Sigma}_{K,L} = \begin{pmatrix} I_d \\ -K \\ -L \end{pmatrix} \Sigma_{K,L} \begin{pmatrix} I_d \\ -K \\ -L \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_1^2 I_{m_1} & 0 \\ 0 & 0 & \sigma_2^2 I_{m_2} \end{pmatrix}. \tag{D.14}$$

By (D.14), we upper bound the norms of $\widetilde{\Sigma}_{K,L}$ as follows,

$$\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}} \leq \sigma_1^2 m_1 + \sigma_2^2 m_2 + \big(d + \|K\|_{\mathrm{F}}^2 + \|L\|_{\mathrm{F}}^2\big) \cdot \|\Sigma_{K,L}\|, \tag{D.15}$$

$$\|\widetilde{\Sigma}_{K,L}\| \leq \sigma_1^2 + \sigma_2^2 + \big(1 + \|K\|^2 + \|L\|^2\big) \cdot \|\Sigma_{K,L}\|. \tag{D.16}$$

By the fact that $\rho(U) < 1$ and the definition of $\Sigma_{K,L}$ in (C.2) of Proposition C.1, we conclude that $\widetilde{\Sigma}_{K,L}$ is the unique positive definite solution of the Lyapunov equation,

$$\widetilde{\Sigma}_{K,L} = \widetilde{\Psi}_\sigma + U\widetilde{\Sigma}_{K,L}U^\top. \tag{D.17}$$

We prove a stronger lemma to establish Lemma D.1,

Lemma D.6. For any policy $\pi_{K,L}$ that is stable in the sense that $\rho(A-BK-CL)<1$, it holds for the matrix $\mathfrak{T}_{K,L}$ defined in (D.3) that

$$\mathfrak{T}_{K,L} = (\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L}) - (\widetilde{\Sigma}_{K,L}U^\top)\otimes(\widetilde{\Sigma}_{K,L}U^\top) = (\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L})(I-U^\top\otimes U^\top). \tag{D.18}$$

Moreover, $\mathfrak{T}_{K,L}$ is invertible and it holds that

$$\|\mathfrak{T}_{K,L}\| \le 2\big(\sigma_1^2+\sigma_2^2+(1+\|K\|^2+\|L\|^2)\cdot\|\Sigma_{K,L}\|\big)^2. \tag{D.19}$$

Proof. We first prove (D.18) of Lemma D.6. We denote the right-hand side of (D.18) by

$$\bar{\mathfrak{T}} = (\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L}) - (\widetilde{\Sigma}_{K,L}U^\top)\otimes(\widetilde{\Sigma}_{K,L}U^\top) = (\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L})(I-U^\top\otimes U^\top).$$

In what follows, we show that for any matrices M,N, it holds that $\mathrm{svec}(M)\mathfrak{T}_{K,L}\mathrm{svec}(N) = \mathrm{svec}(M)\bar{\mathfrak{T}}\mathrm{svec}(N)$. By (D.12) and the fact that $\phi(z)=\mathrm{svec}(zz^\top)$ in (D.1), we have that

$$\phi(z)-\phi(z') = \mathrm{svec}\big[zz^\top-(Uz+\zeta)(Uz+\zeta)^\top\big].$$

By the fact that z and ζ are independent and the definition of $\mathfrak{T}_{K,L}$ in (D.3), we have that

$$\mathfrak{T}_{K,L} = \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\mathrm{svec}(zz^\top)\mathrm{svec}(zz^\top-Uzz^\top U^\top-\widetilde{\Psi}_\sigma)^\top\big].$$

Let M and N be two matrices. Then, we have that

$$\begin{aligned}\mathrm{svec}(M)\mathfrak{T}_{K,L}\mathrm{svec}(N) &= \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\langle zz^\top,M\rangle\cdot\langle zz^\top-Uzz^\top U^\top-\widetilde{\Psi}_\sigma,N\rangle\big]\\ &= \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[z^\top Mzz^\top(N-U^\top NU)z\big]-\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[z^\top Mz\big]\cdot\langle\widetilde{\Psi}_\sigma,N\rangle\\ &= 2\langle M,\widetilde{\Sigma}_{K,L}(N-U^\top NU)\widetilde{\Sigma}_{K,L}\rangle+\langle M,\widetilde{\Sigma}_{K,L}\rangle\cdot\big[\langle N-U^\top NU,\widetilde{\Sigma}_{K,L}\rangle-\langle\widetilde{\Psi}_\sigma,N\rangle\big].\end{aligned} \tag{D.20}$$

where the last equality holds because of the following lemma

Lemma D.7. Let $g\sim N(0,I_d)$ be some standard Gaussian random vector in $\mathbb{R}^d$ and let A_1,A_2 be two symmetric matrices. Then we have

$$\mathbb{E}[g^\top A_1 g\cdot g^\top A_2 g] = 2\mathrm{Tr}(A_1A_2)+\mathrm{Tr}(A_1)\cdot\mathrm{Tr}(A_2).$$

Proof. See e.g. Nagar (1959); Magnus et al. (1978) for a detailed proof. □

By (D.17), we have that

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

Therefore, by (D.20), we have that

$$\begin{aligned}\mathrm{svec}(M)\mathfrak{T}_{K,L}\mathrm{svec}(N) &= 2\langle M,\widetilde{\Sigma}_{K,L}(N-U^\top NU)\widetilde{\Sigma}_{K,L}\rangle\\ &= 2\mathrm{svec}(M)^\top(\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L}-\widetilde{\Sigma}_{K,L}U^\top\otimes\Sigma_{K,L}U^\top)\mathrm{svec}(N)^\top\\ &= 2\mathrm{svec}(M)^\top\big[(\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L})(I-U^\top\otimes U^\top)\big]\mathrm{svec}(N)^\top.\end{aligned} \tag{D.21}$$

Since (D.21) holds for any matrices M,N, we have that

$$\mathfrak{T}_{K,L} = (\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L})(I-U^\top\otimes U^\top),$$

which concludes (D.18). By the fact that $\rho(U)=\rho(A-BK-CL)<1$, we concludes that $\mathfrak{T}_{K,L}$ is invertible.

Finally, we upper bound $\|\mathfrak{T}_{K,L}\|$. By the triangle inequality, we have that

$$\begin{aligned}\|\mathfrak{T}_{K,L}\| &\le \|\widetilde{\Sigma}_{K,L}\otimes\widetilde{\Sigma}_{K,L}\|\cdot(1+\|U^\top\otimes U^\top\|)\\ &\le \|\widetilde{\Sigma}_{K,L}\|^2\cdot(1+\|U\|^2)\\ &\le 2\|\widetilde{\Sigma}_{K,L}\|^2,\end{aligned} \tag{D.22}$$

where the last inequality follows from the fact that $\rho(U)<1$, and the second inequality holds as a result of the following lemma from Alizadeh et al. (1998)

Lemma D.8 (Lemma 7.2 in Alizadeh et al. (1998))**.** Let $A, B \in \mathbb{R}^{m\times m}$ be two matrices that can be diagonalizable simultaneously, and let $\lambda_1, ..., \lambda_m$ and $\mu_1, ..., \mu_m$ be the eigenvalues of A, B respectively. Then, the eigenvalues of $A \otimes B$ are given by $\{1/2 \cdot (\lambda_i\mu_j + \lambda_j\mu_i), i, j \in [m]\}$.

Plugging (D.16) into (D.22), we have that

$$\|\mathfrak{T}_{K,L}\| \le 2\big(\sigma_1^2 + \sigma_2^2 + (1 + \|K\|^2 + \|L\|^2) \cdot \|\Sigma_{K,L}\|\big)^2, ,$$

which completes the proof of Lemma D.6. □

E Proof of Supporting Lemmas

E.1 Proof of Lemma D.3

Proof. We first show that $w(\beta) \in \mathcal{X}_W$ for any $\beta \in \mathcal{X}_B$. By the definition of $G(\beta, w)$ in (D.5) and the property of quadratic function, we have that

$$w(\beta) = \operatorname*{argmax}_w G(\beta, w) = \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}, z'\sim\widetilde{P}(\cdot\,|\,z;\pi_{K,L})}\big[H(z, z'; \beta)\big].$$

where we define $H(z.z'; \beta)$ as follows

$$H^1(z.z'; \beta) = \beta^1 - c(z), \tag{E.1}$$

$$H^2(z, z'; \beta) = \Big\{\beta^1 + \big[\phi(z) - \phi(z')\big]^\top \beta^2 - c(z)\Big\}\phi(z). \tag{E.2}$$

Thus, it suffices to bound $H(z, z'; \beta)$ for any $\beta \in \mathcal{X}_B$. We bound the two components $H^1(z, z'; \beta)$ and $H^2(z, z'; \beta)$ separately. We first bound $H^1(z, z'; \beta)$. By (E.1), we have that

$$|w^1(\beta)| = \Big|\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}, z'\sim\widetilde{P}(\cdot\,|\,z;\pi_{K,L})}\big[H^1(z, z'; \beta)\big]\Big| = \big|\beta^1 - J(K, L)\big| \le J(K, L), \tag{E.3}$$

where the second inequality follows from the fact that $\beta^1 \in [0, J(K, L)]$. We now bound $H^2(z, z'; \beta)$. By (E.2), we have that

$$w^2(\beta) = \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}, z'\sim\widetilde{P}(\cdot\,|\,z;\pi_{K,L})}\big[H^2(z, z'; \beta)\big] = \beta^1 \cdot \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big] + \mathfrak{T}_{K,L}\beta^2 - b_{K,L},$$

where $\mathfrak{T}_{K,L}$ and $b_{K,L}$ are defined in (D.3). By the fact that $\beta \in \mathcal{X}_B$ in (D.7) we have that

$$\|w^2(\beta)\| \le J(K, L) \cdot \big\|\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big]\big\| + \|\mathfrak{T}_{K,L}\| \cdot \bar{R}_B + \|b_{K,L}\|. \tag{E.4}$$

We now bound the three terms on the right-hand side of (E.4). In what follows, we use the same notation $\widetilde{\Sigma}_{K,L}$ defined in (D.14). For the first term on the right-hand side of (E.4), we have that

$$\big\|\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big]\big\| = \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}} \le \sigma_1^2 \cdot m_1 + \sigma_2^2 \cdot m_2 + (d + \|K\|_{\mathrm{F}}^2 + \|L\|_{\mathrm{F}}^2) \cdot \|\Sigma_{K,L}\|. \tag{E.5}$$

For the second term on the right-hand side of (E.4), by Lemma D.6, we have that

$$\|\mathfrak{T}_{K,L}\|_2 \le 2\big(\sigma_1^2 + \sigma_2^2 + (1 + \|K\|^2 + \|L\|^2) \cdot \|\Sigma_{K,L}\|\big)^2. \tag{E.6}$$

For the last term on the right-hand side of (E.4), it suffices to bound $b_{K,L}$, we now study the linear operator induced by $b_{K,L}$. For any positive definite matrix $M \in \mathbb{R}^{(d+m_1)\times(d+m_1)}$, we have

$$\begin{aligned}\langle b_{K,L}, \mathrm{svec}(M)\rangle &= \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\Big[\langle\phi(z), \mathrm{svec}(M)\rangle \cdot \big\langle\phi(z), \mathrm{svec}\big(\mathrm{diag}(Q, R, -S)\big)\big\rangle\Big] \\ &= 2\langle\widetilde{\Sigma}_{K,L}\mathrm{diag}(Q, R, -S)\widetilde{\Sigma}_{K,L}, M\rangle + \langle\widetilde{\Sigma}_{K,L}, \mathrm{diag}(Q, R, -S)\rangle \cdot \langle\widetilde{\Sigma}_{K,L}, M\rangle \end{aligned} \tag{E.7}$$

where the first equality follows from the fact that $c(z) = \langle\phi(z), \mathrm{svec}(\mathrm{diag}(Q, R, S))\rangle$, and the second equality follows form Lemma D.7. Thus, we have that

$$\begin{aligned}\|b_{K,L}\| &\le 3(\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}} + \|S\|_{\mathrm{F}}) \cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^2 \\ &\le 3(\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}} + \|S\|_{\mathrm{F}}) \cdot \big(\sigma_1^2 m_1 + \sigma_2^2 m_2 + (d + \|K\|_{\mathrm{F}}^2 + \|L\|_{\mathrm{F}}^2) \cdot \|\Sigma_{K,L}\|\big)^2, \end{aligned} \tag{E.8}$$

where the last inequality follows from (D.15). We establish the following lemma to bound $\|\Sigma_{K,L}\|$.

Lemma E.1. For any stable policy $\pi_{K,L}$, we have

$$\|\Sigma_{K,L}\| \le \big[J(K,L)+\chi\big]/\sigma_{\min}(Q-L^\top SL), \qquad \|P_{K,L}\| \le \big[J(K,L)+\chi\big]/\sigma_{\min}(\Psi_\sigma).$$

Here χ is defined as $\chi := (-\sigma_1^2\mathrm{Tr}(R)+\sigma_2^2\mathrm{Tr}(S))^+$.

Proof. By (C.7) in Proposition C.1, we have that

$$\begin{aligned}
&J(K,L)+\chi \ge \mathrm{Tr}\big((Q-L^\top SL+K^\top RK)\Sigma_{K,L}\big) \ge \sigma_{\min}(Q-L^\top SL)\cdot\|\Sigma_{K,L}\|,\\
&J(K,L)+\chi \ge \mathrm{Tr}(P_{K,L}\Psi_\sigma) \ge \sigma_{\min}(\Psi_\sigma)\cdot\|P_{K,L}\|.
\end{aligned}$$

Therefore, we conclude the proof. □

Plugging (E.5), (E.6), and (E.8) into (E.4), by Lemma E.1, we get that

$$\|w^2(\beta)\| \le C\cdot(d+\|K\|_{\mathrm{F}}^2+\|L\|_{\mathrm{F}}^2)^2\cdot\bar{R}_B\cdot\sigma_{\min}(Q-L^\top SL)^{-2}\cdot\big(J(K,L)+\chi\big)^2+C_0, \tag{E.9}$$

where $C, C_0 > 0$ are constants. Thus, combining (E.3), (E.9), and the definition of $\mathcal{X}_W$ in (D.8), we concludes that $w(\beta)\in\mathcal{X}_W$ for any $\beta\in\mathcal{X}_B$.

It remains to show that $\beta^*\in\mathcal{X}_B$. By the definition of β^* in (D.2) and the definition of $\mathcal{X}_B$ in (D.7), it suffices to bound $\|\Lambda_{K,L}\|_{\mathrm{F}}$. By the definition of $\Lambda_{K,L}$ in (4.4), we have that

$$\begin{aligned}
\Lambda_{K,L} &= \begin{pmatrix} Q+A^\top P_{K,L}A & A^\top P_{K,L}B & A^\top P_{K,L}C\\ B^\top P_{K,L}A & R+B^\top P_{K,L}B & B^\top P_{K,L}C\\ C^\top P_{K,L}A & C^\top P_{K,L}B & -S+C^\top P_{K,L}C\end{pmatrix}\\
&= \begin{pmatrix} Q & & \\ & R & \\ & & -S\end{pmatrix} + \begin{pmatrix} A & B & C\end{pmatrix}^\top P_{K,L}\begin{pmatrix} A & B & C\end{pmatrix}.
\end{aligned}$$

which implies that

$$\|\Lambda_{K,L}\|_{\mathrm{F}} \le \|Q\|_{\mathrm{F}}+\|R\|_{\mathrm{F}}+\|S\|_{\mathrm{F}}+\big(\|A\|_{\mathrm{F}}^2+\|B\|_{\mathrm{F}}^2+\|C\|_{\mathrm{F}}^2\big)\cdot\|P_{K,L}\|. \tag{E.10}$$

Apply Lemma E.1 to (E.10), we get that

$$\begin{aligned}
\|\Lambda_{K,L}\|_{\mathrm{F}} &\le \|Q\|_{\mathrm{F}}+\|R\|_{\mathrm{F}}+\|S\|_{\mathrm{F}}+\big(\|A\|_{\mathrm{F}}^2+\|B\|_{\mathrm{F}}^2+\|C\|_{\mathrm{F}}^2\big)\cdot\big(J(K,L)+\chi\big)/\sigma_{\min}(\Psi_\sigma)\\
&= C_0+C\cdot\big(J(K,L)+\chi\big)/\sigma_{\min}(\Psi_\sigma).
\end{aligned} \tag{E.11}$$

By (E.11) and definition of $\bar{R}_B$ in (D.9), we conclude that $\beta^*\in\mathcal{X}_B$. Thus, we complete the proof. □

E.2 Proof of Lemma D.4

Proof. First, by the definition of $G(\beta,w)$ in (D.5), we have for any $\widehat{\beta}\in\mathcal{X}_B, \widehat{w}\in\mathcal{X}_W$ that

$$|\widehat{\beta}^1-J(K,L)|^2+\|\widehat{\beta}^1\cdot\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big]+\mathfrak{T}_{K,L}\widehat{\beta}^2-b_{K,L}\|_2^2 = \max_{w\in\mathcal{X}_W} G(\widehat{\beta},w) \le \mathrm{Gap}(\widehat{\beta},\widehat{w}), \tag{E.12}$$

where the last inequality follows from the following inequality

$$\min_{\beta\in\mathcal{X}_B} G(\beta,\widehat{w}) \le \min_{\beta\in\mathcal{X}_B} G(\beta,w(\beta)) = G(\beta^*,0)=0.$$

By the definition of β^* in (D.2), we rewrite the left-hand side of (E.12) as follows

$$\left\|\begin{pmatrix} 1 & 0\\ \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\phi(z)\big] & \mathfrak{T}_{K,L}\end{pmatrix}(\widehat{\beta}-\beta^*)\right\|^2 \ge \sigma_{\min}(\mathfrak{T}_{K,L})^2\cdot\big(|\widehat{\beta}^1-J(K,L)|^2+\|\widehat{\Lambda}-\Lambda_{K,L}\|^2\big). \tag{E.13}$$

Combining (E.12) and (E.13), we have that

$$(|\widehat{\beta}^1-J(K,L)|^2+\|\widehat{\Lambda}-\Lambda_{K,L}\|^2) \le 1/\sigma_{\min}(\mathfrak{T}_{K,L})^2\cdot\mathrm{Gap}(\widehat{\beta},\widehat{w}), \tag{E.14}$$

which completes the proof of Lemma D.4. □

E.3 Proof of Lemma D.5

Proof. First, let us characterize the geometrically $\boldsymbol{\beta}$-mixing property of our problem via the following lemma.

Lemma E.2 (Geometrically $\boldsymbol{\beta}$-mixing)**.** Consider a linear transition $z_{t+1} = Uz_t + e_t$, where $z_t \in \mathbb{R}^D$ is the Markov Chain and $e_t \sim N(0, \Psi)$ is the Gaussian noise independent of $\{z_t\}_{t\geq 0}$, and $U \in \mathbb{R}^{D\times D}$ satisfies that $\rho(U) < 1$. We denote by ν_t the distribution of z_t and by $N(0, \Sigma_\infty)$ the stationary distribution of z_t. We define the k-th mixing time coefficient $\boldsymbol{\beta}(k)$ as follows

$$\boldsymbol{\beta}(k) = \sup_t \mathbb{E}_{z\sim\nu_t}\big[\|\mathbb{P}_{z_k}(\cdot \,|\, z_0 = z) - \mathbb{P}_{N(0,\Sigma_\infty)}(\cdot)\|_{\mathrm{TV}}\big].$$

Then, it holds that

$$\boldsymbol{\beta}(k) \leq C_{\rho,U}[\mathrm{Tr}(\Sigma_\infty) + m(1-\rho)^{-2}]^{1/2}\rho^k,$$

where $\rho \in (\rho(U), 1)$, and $C_{\rho,U}$ is a constant depending on U and ρ.

We obtain the following theorem from Wang et al. (2017) to establish Lemma D.5.

Theorem E.3 (Theorem 1 in Wang et al. (2017))**.** Let $\mathcal{X}_W$ and $\mathcal{X}_B$ be two convex, closed, and bounded sets and the radius of $\mathcal{X}_W \times \mathcal{X}_W$ is $D > 0$. Consider the following minimax stochastic optimization problem

$$\min_{\beta\in\mathcal{X}_B} \max_{w\in\mathcal{X}_W} G(\beta, w) = \mathbb{E}_{\xi\sim\Xi}[g(\beta, w; \xi)]$$

where ξ is a random variable drawn from Ξ. The function g is convex in β and concave in w. We assume Ξ is the stationary distribution of a Markov Chain $\{\xi_t\}_{t=0}^\infty$ with uniform mixing time $\tau(\eta)$ for any $\eta > 0$. Meanwhile, we assume that for almost every ξ, the function $g(\beta, w; \xi)$ is L_1 Lipschitz continuous for every $\beta \in \mathcal{X}_B, w \in \mathcal{X}_W$. The stochastic gradient $\nabla_\beta g(\beta, w; \xi)$ and $\nabla_w g(\beta, w; \xi)$ are L_2-Lipschitz continuous for every $\beta \in \mathcal{X}_B, w \in \mathcal{X}_W$ and almost every ξ. For any non-increasing step size α_t, the projected primal dual stochastic gradient method updates β and w by

$$\beta_{t+1} = \Pi_{\mathcal{X}_B}\big[\beta_t - \alpha_t \cdot \nabla_\beta g(\beta_t, w_t; \xi_t)\big], \qquad w_{t+1} = \Pi_{\mathcal{Y}}\big[w_t + \alpha_t \cdot \nabla_w g(\beta_t, w_t; \xi_t)\big].$$

for $t \in [T-1]$. Let

$$\bar{\beta} = \frac{\sum_{t=0}^T \alpha_t\beta_t}{\sum_{t=0}^T \alpha_t}, \qquad \bar{w} = \frac{\sum_{t=0}^T \alpha_t w_t}{\sum_{t=0}^T \alpha_t}.$$

Then, for any $\delta, \eta > 0$ such that $\tau(\eta) \leq T/2$, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\max_{w\in\mathcal{X}_W} G(\bar{\beta}, w) - \min_{\beta\in\mathcal{X}_B} G(\beta, \bar{w}) \leq \Big(\sum_{t=0}^T \alpha_t\Big)^{-1}\Big\{A_0 + A_1\cdot\eta\cdot\sum_{t=0}^T \alpha_t + A_2\sum_{t=0}^T \alpha_t^2 \\ +16DL_1\cdot\Big[2\tau(\eta)\cdot\log\big(\tau(\eta)/\delta\big)\cdot\big(\sum_{t=0}^T \alpha_t^2 + \tau(\eta)\alpha_1\big)\Big]^{1/2}\Big\},\end{aligned} \tag{E.15}$$

where we define

$$A_0 = D^2 + 12D\alpha_1\cdot\tau(\eta), \quad A_1 = 4L_1D, \quad A_2 = 10L_1 + \big(24L_1^2 + 8L_1L_2D\big)\cdot\tau(\eta). \tag{E.16}$$

Specifically, for geometrically $\boldsymbol{\beta}$-mixing process defined in Lemma E.2, we have the following corollary.

Corollary E.4. Use the same setting as in Theorem E.3. We assume that the k-th mixing time satisfies $\boldsymbol{\beta}(k) \leq C_\xi\rho^k$ for some $C_\xi > 0, \rho < 1$. We set step size $\alpha_t = \alpha/\sqrt{t}$ for some $\alpha > 0$. Then, we have that, with probability at least $1-\delta$,

$$\max_{w\in\mathcal{X}_W} G(\bar{\beta}, w) - \min_{\beta\in\mathcal{X}_B} G(\beta, \bar{w}) \leq \frac{C\cdot D^2 + L_1^2 + L_1L_2D}{\log(1/\rho)}\cdot\frac{\log^2 T + \log(1/\delta)}{\sqrt{T}} + \frac{C\cdot C_\xi L_1 D}{T}.$$

Proof. The proof follows direct computations. By the property of geometrically β-mixing processes, we define mixing time by $\tau(\eta) = \log(\eta/C_\xi)/\log\rho$ and $\eta = C_\xi/T$. We set $\alpha_t = \alpha/\sqrt{T}$ and the constants A_0, A_1, A_2 as (E.16). Note that $\sum_{t=1}^T 1/t \sim \log T, \sum_{t=1}^T 1/\sqrt{t} \sim 2\sqrt{T}$. Thus, we bound the first three terms of the right-hand side of (E.15) as follows

$$
\begin{aligned}
&A_0 \le C \cdot D^2 \cdot \log T/\log(1/\rho),\\
&A_1 \cdot \eta \cdot \sum_{t=0}^T \alpha_t \le C \cdot C_\xi L_1 D/T,\\
&A_2 \sum_{t=0}^T \alpha_t^2 \le C \cdot (L_1 + L_1^2 + L_1 L_2 D) \cdot \log^2 T/\sqrt{T}.
\end{aligned}
$$

For the last term, we have

$$
\begin{aligned}
&16DL_1\Big[2\tau(\eta)\log\big(\tau(\eta)/\delta\big)\big(\sum_{t=0}^T \alpha_t^2 + \tau(\eta)\alpha_1\big)\Big]^{1/2}\\
&\le C \cdot DL_1 \log T/\log(1/\rho)\sqrt{\log\log T + \log(1/\delta)}.
\end{aligned}
$$

Thus, we conclude the proof of Corollary E.4. □

However, since $z \sim \widetilde{\rho}_{K,L}$ is not bounded, the function $f(\beta, w; z, z')$ defined in (D.6) is not Lipschitz continuous with probability bigger than 0, which prevents us to apply Theorem 1 in Wang et al. (2017) directly. Instead, the following lemma confirms that z is bounded with high probability, so f is Lipschitz continuous with high probability.

Lemma E.5 (Hansen-Wright inequality)**.** Let $A \in \mathbb{R}^{m\times m}$ be a matrix and let $\eta \sim N(0, I_m)$ is a Gaussian random variable. Then there exists some constant $C > 0$ such that for any $t \ge 0$, it holds that

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

Proof. See Rudelson et al. (2013) for a detailed proof. □

Applying Lemma E.5 to $z \sim \widetilde{\rho}_{K,L} = N(0, \widetilde{\Sigma}_{K,L})$, where $\widetilde{\Sigma}_{K,L}$ is defined in (D.17), we have that

$$
\mathbb{P}\big(|\|z\|^2 - \mathrm{Tr}(\widetilde{\Sigma}_{K,L})| > t\big) \le 2\exp\Big(-C\cdot\min\big\{t^2\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^{-2}, t\cdot\|\widetilde{\Sigma}_{K,L}\|^{-1}\big\}\Big), \tag{E.17}
$$

for an absolute constant C. We set $t = C_1 \cdot \|\widetilde{\Sigma}_{K\,|\,L}\| \cdot \log T$ for sufficiently large C_1 such that $C \cdot C_1 \ge 6$. Then, we have that

$$
t^2 \cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^{-2} = C_1^2 \cdot \log^2 T \cdot \|\widetilde{\Sigma}K, L\|^2 \cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^{-2} \ge C_1^2 \cdot (d+m_1)^{-1}\log^2 T \ge t \cdot \|\widetilde{\Sigma}_{K,L}\|^{-1}, \tag{E.18}
$$

where the first inequality holds as a result of the relation between operator norms and Frobenius norms, and the second inequality holds for sufficiently large T such that $C_1(d+m_1)^{-1}\log T \ge 1$. Define

$$
\mathcal{E}_t = \big\{\|z_t\|^2 - \mathrm{Tr}(\widetilde{\Sigma}_{K,L}) \le C_1 \cdot \|\widetilde{\Sigma}_{K,L}\| \cdot \log T\big\}, \tag{E.19}
$$

for $t \in [T]$. We wite $\mathcal{E} = \cap_{1\le t\le T}\mathcal{E}_t$. Combining (E.17) and (E.18) we get $\mathbb{P}(\mathcal{E}_t) \ge 1 - 2T^{-6}$. Applying union bound to $1 \le t \le T$, we have that $\mathbb{P}(\mathcal{E}) \ge 1 - 2T^{-5}$. On event $\mathcal{E}$, by (E.19), we have that

$$
\max_{1\le t\le T}\|z_t\|^2 \le C_1 \cdot \|\widetilde{\Sigma}_{K,L}\| \cdot \log T + \mathrm{Tr}(\widetilde{\Sigma}_{K,L}) \le C \cdot \log T \cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}, \tag{E.20}
$$

which illustrates an upper bound on the feature vectors $\phi(z_t), 1 \le t \le T$, since we have $\big\|\phi(z)\big\|_2 = \|z\|_2^2$.

To prove Theorem 5.2, we consider the minimax optimization problem (D.5) restricted on the set $\mathcal{E}$. For any z, we define

$$
\mathcal{E}_z = \big\{|\|\phi_z\|_2 - \mathrm{Tr}(\widetilde{\Sigma}_{K,L})| < C_1 \cdot \|\widetilde{\Sigma}_{K,L}\| \cdot \log T\big\}. \tag{E.21}
$$

Thus, by (E.19), we have $\mathbb{P}(\mathcal{E}_z) \geq 1 - 2T^{-6}$. For the feature vector ϕ, vector function H, and objective function g and G, we define

$$\begin{aligned} \widetilde{\phi}(z) &= \phi(z) \cdot \mathbb{1}_{\mathcal{E}_z}, \\ \widetilde{H}(z, z'; \beta) &= H(z, z'; \beta) \cdot \mathbb{1}_{\mathcal{E}_z} \cdot \mathbb{1}_{\mathcal{E}_{z'}}, \\ \widetilde{c}(z) &= c(z) \cdot \mathbb{1}_{\mathcal{E}_z}, \\ \widetilde{g}(\beta, w; z, z') &= g(\beta, w; z, z') \cdot \mathbb{1}_{\mathcal{E}_z} \cdot \mathbb{1}_{\mathcal{E}_{z'}}, \\ \widetilde{G}(\beta, w) &= \mathbb{E}_{z \sim \rho_{K,L}, z' \sim \widetilde{P}(\cdot \,|\, z; \pi_{K,L})} \big[\widetilde{g}(\beta, w; z, z')\big]. \end{aligned} \tag{E.22}$$

We note that when conditioned on $\mathcal{E}$, the output of the primal dual stochastic update applied on $\widetilde{G}$ is the same as that on G. We denote the duality gap for $\widetilde{G}$ by $\widetilde{\mathrm{Gap}}(\widehat{\beta}, \widehat{w}) = \max_{w \in \mathcal{X}_W} \widetilde{G}(\widehat{\beta}, w) - \min_{\beta \in \mathcal{X}_B} \widetilde{G}(\beta, \widehat{w})$. Since $\widetilde{g}$ is Lipschitz continuous, we apply Corollary E.4 to bound the primal dual gap $\widetilde{\mathrm{Gap}}(\widehat{\beta}, \widehat{w})$.

To apply Corollary E.4, we first establish upper bounds of the $\nabla_{(\beta,w)} \widetilde{g}$, $\nabla^2_{\beta,\beta} \widetilde{g}$ and $\nabla^2_{w,w} \widetilde{g}$ to show that $\widetilde{g}$ and $\nabla \widetilde{g}$ are Lipschitz continuous, where $\widetilde{g}$ is defined in (E.22). By direct computation, we have that

$$\nabla_{\beta^1} \widetilde{g} = w^1 + \widetilde{\phi}(z)^\top w^2, \qquad \nabla_{\beta^2} \widetilde{g} = [\widetilde{\phi}(z)^\top w^2]^\top [\widetilde{\phi}(z) - \widetilde{\phi}(z')] \tag{E.23}$$

$$\nabla_{w^1} \widetilde{g} = \beta^1 - \widetilde{c}(z) - w^1, \qquad \nabla_{w^2} \widetilde{g} = \widetilde{H}^2(z, z'; \beta) - w^2. \tag{E.24}$$

By (E.20) and (E.23), we have that

$$\|\nabla_\beta \widetilde{g}\|_2 \leq C \cdot (\log T)^2 \cdot \|\widetilde{\Sigma}\|_{\mathrm{F}}^2. \tag{E.25}$$

For the gradient with respect to w, we first bound $\widetilde{H}^2$. By (E.2),

$$\|\widetilde{H}^2\| \leq \big(|\beta^1| + (\|\widetilde{\phi}(z)\| + \|\widetilde{\phi}(z')\|) \cdot \bar{R}_B + (\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}} + \|S\|_{\mathrm{F}}) \cdot \|\widetilde{\phi}(z)\|\big) \cdot \|\widetilde{\phi}(z)\|,$$

which concludes that

$$\begin{aligned} \|\nabla_w \widetilde{g}\| &\leq \sqrt{|\beta^1 - \widetilde{c}(z) + w^1|^2 + (|w^2| + \|\widetilde{H}^2(z, z'; \beta)\|)^2} \\ &\leq C \cdot (\log T)^3 \cdot \big(J(K, L) + \bar{R}_B^2 + \bar{R}_W^2 + \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^2\big) \cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}. \end{aligned} \tag{E.26}$$

Meanwhile, we have $\nabla^2_{\beta\beta} \widetilde{g} = 0$ and that $\nabla^2_{ww} \widetilde{g} = -I$, which, combined with (E.25) and (E.26), implies that $\widetilde{g}$ and $\nabla \widetilde{g}$ are Lipschitz continuous with respect to β and w. Let $\widetilde{\beta}, \widetilde{w}$ be the output of GTD (Algorithm 3) applied on $\widetilde{G}$. Then, by applying Corollary E.4, we have with probability at least $1 - T^{-5}$ that

$$\begin{aligned} \widetilde{\mathrm{Gap}}(\widetilde{\beta}, \widetilde{w}) &\leq \mathrm{Poly}\big(J(K, L), \bar{R}_B, \bar{R}_W, \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}\big) \cdot \frac{(\log T)^4}{(1 - \rho)\sqrt{T}} \\ &\leq \mathrm{Poly}\big(J(K, L), \bar{R}_B, \bar{R}_W, \sigma_{\min}(Q - L^\top S L)^{-1}\big) \cdot \frac{(\log T)^4}{(1 - \rho)\sqrt{T}}, \end{aligned} \tag{E.27}$$

where the second inequality holds as a result of Lemma E.1, and $\rho \in (\rho(A - BK - CL), 1)$.

We note that when the event $\mathcal{E} = \cap_{1 \leq t \leq T} \mathcal{E}_t$ holds, it holds that $(\bar{\beta}, \bar{w}) = (\widetilde{\beta}, \widetilde{w})$. Thus, it remains to bound the difference $|\mathrm{Gap} - \widetilde{\mathrm{Gap}}|$. For any $\beta \in \mathcal{X}_B, w \in \mathcal{X}_W$, we have that

$$\begin{aligned} |\widetilde{G}(\beta, w) - G(\beta, w)| &= \big|\big\langle \mathbb{E}_{z,z'} \big[\widetilde{H}(z, z'; \beta) - H(z, z'; \beta)\big], w \big\rangle\big| \\ &\leq \big|\mathbb{E}_{z,z'} \big[\widetilde{H}^1(z, z'; \beta) - H^1(z, z'; \beta)\big]\big| \cdot J(K, L) \\ &\quad + \big\|\mathbb{E}_{z,z'} \big[\widetilde{H}^2(z, z'; \beta) - H^2(z, z'; \beta)\big]\big\| \cdot \bar{R}_W, \end{aligned} \tag{E.28}$$

where the expectation is with respect to $z \sim \widetilde{\rho}_{K,L}, z' \sim \widetilde{P}(\cdot \,|\, z; \pi_{K,L})$. For the two terms on the right-hand side of (E.28), we have that

$$\widetilde{H}^1(z, z'; \beta) - H^2(z, z'; \beta) = c(z) \cdot \mathbb{1}_{\mathcal{E}_z^c}, \tag{E.29}$$

$$\widetilde{H}^2(z, z'; \beta) - H^2(z, z'; \beta) = H^2(z, z'; \beta) \mathbb{1}_{\mathcal{E}_z^c} + \phi(z')^\top \beta^2 \cdot \phi(z) \cdot \mathbb{1}_{\mathcal{E}_z} \cdot \mathbb{1}_{\mathcal{E}_{z'}^c}, \tag{E.30}$$

We first bound the right-hand side of (E.29). Applying Cauchy-Schwarz inequality to the first term of the right-hand side of (E.28) and (E.29), we have that

$$|\widetilde{H}^1(z,z';\beta) - H^1(z,z';\beta)| \le \sqrt{\mathbb{P}(\mathcal{E}_z^c)\cdot \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[c(z)^2\big]}. \tag{E.31}$$

Since $z, z' \sim N(0, \widetilde{\Sigma}_{K,L})$, we have that

$$\mathbb{E}_{z,z'}\big[\|\phi(z)\|^4\big] = \mathbb{E}_{z,z'}\big[\|\phi(z')\|^4\big] \le C\cdot \|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^4, \tag{E.32}$$

for an absolute constant C. Thus, we have that

$$\begin{aligned}\mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[c(z)^2\big] &= \mathbb{E}_{z\sim\widetilde{\rho}_{K,L}}\big[\big(\phi(z)^\top \mathrm{svec}(\mathrm{diag}(Q,R,-S))\big)^2\big]\\ &\le C\cdot(\|Q\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2)\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^2\end{aligned} \tag{E.33}$$

By plugging (E.33) into (E.31) and the fact that $\mathbb{P}(\mathcal{E}_z) > 1 - 2T^{-6}$, we have that

$$\big|\mathbb{E}_{z,z'}\big[\widetilde{H}^1(z,z';\beta) - H^1(z,z';\beta)\big]\big| \le C\cdot T^{-3}\cdot(\|Q\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2)^{1/2}\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}. \tag{E.34}$$

We now bound the right-hand side of (E.30). By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\big\|\mathbb{E}_{z,z'}\big[\widetilde{H}^2(z,z';\beta) - H^2(z,z';\beta)\big]\big\|_2\\ &\quad\le \sqrt{\mathbb{P}(\mathcal{E}_z^c)}\cdot\sqrt{\mathbb{E}_{z,z'}\big[\|H^2(z,z';\beta)\|^2\big]} + \sqrt{\mathbb{P}(\mathcal{E}_{z'}^c)}\cdot\sqrt{\mathbb{E}_{z.z'}\big[\|\phi(z')^\top\beta^2\cdot\phi(z)\|_2^2\big]}.\end{aligned} \tag{E.35}$$

For the first term on the right-hand side of (E.35), by the definition of H^2 in (E.2) and the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}&\mathbb{E}_{z,z'}\big[\|H^2(z,z';\beta)\|^2\big]\\ &\quad\le 4\Big(|\beta^1|^2\cdot\mathbb{E}_{z,z'}\big[\|\phi(z)\|^2\big] + \sqrt{\mathbb{E}_{z,z'}\big[|c(z)|^4\big]\cdot\mathbb{E}_{z,z'}\big[\|\phi(z)\|^4\big]} +\\ &\qquad + \sqrt{\mathbb{E}_{z,z'}\big[|\phi(z)^\top\beta^2|^4\big]\cdot\mathbb{E}_{z,z'}\big[\|\phi(z)\|^4\big]} + \sqrt{\mathbb{E}_{z,z'}\big[|\phi(z')^\top\beta^2|^4\big]\cdot\mathbb{E}_{z,z'}\big[\|\phi(z)\|^4\big]}\Big).\end{aligned} \tag{E.36}$$

Plugging (E.32) into (E.36), by (E.33), we have for an absolute constant C that

$$\mathbb{E}_{z,z'}\big[\|H^2(z,z';\beta)\|^2\big] \le C\cdot\big(J(K,L)^2 + \bar{R}_B^2 + \|Q\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2\big)\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^4. \tag{E.37}$$

Plugging (E.37) into (E.35), by $\mathbb{P}(\mathcal{E}_z) \ge 1 - 2T^{-6}$, we have that

$$\begin{aligned}&\big\|\mathbb{E}_{z,z'}\big[\widetilde{H}^2(z,z';\beta) - H^2(z,z';\beta)\big]\big\|_2\\ &\quad\le C\cdot T^{-3}\cdot\big(J(K,L)^2 + \bar{R}_B^2 + \|Q\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2\big)^{1/2}\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^2.\end{aligned} \tag{E.38}$$

Plugging (E.33) and (E.38) into (E.28), we have that

$$|\widetilde{G}(\beta,w) - G(\beta,w)| \le C\cdot T^{-3}\cdot\big(J(K,L)^2 + \bar{R}_B^2 + \bar{R}_W^2\big)\cdot\|\widetilde{\Sigma}_{K,L}\|_{\mathrm{F}}^2, \tag{E.39}$$

for an absolute constant C. So for sufficiently large T, we have that $|\widetilde{G}(\beta,w) - G(\beta,w)| < T^{-1}$, which implies that

$$|\mathrm{Gap}(\widehat{\beta},\widehat{w}) - \widetilde{\mathrm{Gap}}(\widehat{\beta},\widehat{w})| \le \max_{w\in\mathcal{X}_W}|G(\widehat{\beta},w) - \widetilde{G}(\widehat{\beta},w)| + \max_{\beta\in\mathcal{X}_B}|G(\beta,\widehat{w}) - \widetilde{G}(\beta,\widehat{w})| \le 2T^{-1}. \tag{E.40}$$

Finally, note that $\mathcal{E}$ holds with probability $1 - T^{-5}$. Thus, by (E.27), we have with probability at least $1 - 2T^{-5} > T^{-4}$ that

$$\begin{aligned}\mathrm{Gap}(\bar{\beta},\bar{w}) &\le \mathrm{Poly}\big(J(K,L),\bar{R}_B,\bar{R}_W,\sigma_{\min}(Q - L^\top SL)^{-1}\big)\cdot\frac{(\log T)^4}{(1-\rho)\sqrt{T}} + 2T^{-1}\\ &\le \mathrm{Poly}\big(J(K,L),\bar{R}_B,\bar{R}_W,\sigma_{\min}(Q - L^\top SL)^{-1}\big)\cdot\frac{(\log T)^4}{(1-\rho)\sqrt{T}},\end{aligned} \tag{E.41}$$

where $\rho \in (\rho(A - BK - CL), 1)$. Thus, we complete the proof of Lemma D.5. □

E.4 Proof of Lemma B.2

Proof. Note that $P_{K,L}$ and $P_{K',L'}$ satisfy the Bellman equation in (C.1). Using the operator $\mathcal{T}^\top$ defined in (C.14), we have that $P_{K',L'} = \mathcal{T}^\top_{K',L'}\big(Q + (K')^\top R(K') - (L')^\top S(L')\big)$, which implies that

$$\begin{aligned}
&x^\top P_{K',L'} x \\
&\quad = \sum_{t\geq 0} x^\top \big[(A - BK' - CL')^t\big]^\top \big(Q + (K')^\top R(K') - (L')^\top S(L')\big)(A - BK' - CL')^t \\
&\quad = \sum_{t\geq 0} (x_t')^\top \big(Q + (K')^\top R(K') - (L')^\top S(L')\big) x_t'.
\end{aligned}$$

So we have that

$$\begin{aligned}
&x^\top P_{K',L'} x - x^\top P_{K,L} x \\
&\quad = \sum_{t\geq 0} \Big\{ {x_t'}^\top \big(Q + {K'}^\top R K' - {L'}^\top S L'\big) x_t' + {x_{t+1}'}^\top P_{K,L} x_{t+1}' - {x_t'}^\top P_{K,L} x_t' \Big\} \\
&\quad = \sum_{t\geq 0} A_{K,L,K',L'}(x_t),
\end{aligned}$$

where

$$\begin{aligned}
&A_{K,L,K',L'}(x) \\
&\quad = {x'}^\top \Big\{ \big(Q + {K'}^\top R K' - {L'}^\top S L'\big) + (A - BK' - CL')^\top P_{K,L}(A - BK' - CL') - P_{K,L} \Big\} x' \\
&\quad = 2x^\top (K' - K)^\top E_{K,L} x + x^\top (K' - K)^\top (R + B^\top P_{K,L} B)(K' - K) x \\
&\qquad + 2x^\top (L' - L) F_{K,L} x + x^\top (L' - L)^\top (-S + C^\top P_{K,L} C)(L' - L) x \\
&\qquad + 2x^\top (L' - L)^\top C^\top P_{K,L} B (K' - K) x,
\end{aligned}$$

which completes the proof. □

E.5 Proof of Lemma B.3

Proof. First, for the upper bound in (B.6), by (C.7) and Lemma B.2, we have that

$$J(K,L) - J\big(K(L), L\big) = \mathrm{Tr}\big((P_{K,L} - P_L^*)\Psi_\sigma\big) = -\mathbb{E}^{\pi_{K(L),L}}_{x_0^* \sim N(0,\Psi_\sigma)}\big[A^L_{K,K(L)}(x_t^*)\big], \tag{E.42}$$

where $x_{t+1}^* = \big(A - BK(L) - CL\big)x_t^*$. By completing the square, for any K, K', we have

$$\begin{aligned}
A^L_{K,K(L)}(x) &= 2x^\top (K' - K)^\top E_{K,L} x + x^\top (K' - K)(R + B^\top P_{K,L} B)(K' - K) x \\
&= \mathrm{Tr}\Big[x x^\top \big[K' - K + (R + B^\top P_{K,L} B)^{-1} E_{K,L}\big]^\top (R + B^\top P_{K,L} B) \\
&\qquad \cdot \big[K' - K + (R + B^\top P_{K,L} B)^{-1} E_{K,L}\big] \Big] \\
&\qquad - \mathrm{Tr}\big(x x^\top E_K^\top (R + B^\top P_{K,L} B)^{-1} E_K\big) \\
&\geq -\mathrm{Tr}\big(x x^\top E_K^\top (R + B^\top P_{K,L} B)^{-1} E_K\big),
\end{aligned} \tag{E.43}$$

where the equality holds when $K' = K - (R + B^\top P_{K,L} B)^{-1} E_{K,L}$. By (E.42) and (E.43), we have that

$$\begin{aligned}
&J(K,L) - J\big(K(L), L\big) \\
&\quad \leq \|(R + B^\top P_{K,L} B)^{-1}\| \cdot \mathrm{Tr}\Big(\mathbb{E}_{x_0^* \sim N(0,\Psi_\sigma)}\big[\sum_{t\geq 0} x_t^* {x_t^*}^\top\big]\Big) \cdot \mathrm{Tr}(E_{K,L}^\top E_{K,L}).
\end{aligned} \tag{E.44}$$

On the other hand, we have

$$\mathbb{E}_{x_0^* \sim N(0,\Psi_\sigma)}\big[\sum_{t\geq 0} x_t^* {x_t^*}^\top\big] = \mathbb{E}_{x_0^* \sim N(0,\Psi_\sigma)}\Big[\mathcal{T}_{K(L),L}(x x^\top)\Big] = \Sigma_{K(L),L},$$

which combining with $\sigma_{\min}(R + B^\top P_{K,L} B) \geq \sigma_{\min}(R)$ and (E.44), we have that

$$J(K, L) - J(K(L), L) \leq 1/\sigma_{\min} \cdot \|\Sigma_L^*\| \cdot \mathrm{Tr}(E_{K,L}^\top E_{K,L}).$$

Meanwhile, to establish the lower bound, note that in (E.43) the equality is obtained when $K' = K - (R + B^\top P_{K,L} B)^{-1} E_{K,L}$, so we have

$$\begin{aligned} J(K, L) - J(K(L), L) &\geq J(K, L) - J(K', L) \\ &= -\mathbb{E}_{x_0' \sim N(0, \Psi_\sigma)}[\sum_{t \geq 0} A_{K,K'}^L(x_t')] \\ &= \mathrm{Tr}\Big(E_{K,L}^\top (R + B^\top P_{K,L} B)^{-1} E_{K,L} \Sigma_{K',L}\Big) \\ &\geq 1/\sigma_{\min}(\Psi_\sigma) \cdot \|R + B^\top P_{K,L} B\|^{-1} \mathrm{Tr}(E_{K,L}^\top E_{K,L}). \end{aligned}$$

□

E.6 Proof of Lemma B.4

Proof. According to Polyak (1963), we need to show that $|J(K_{t+1}', L) - J(K_t, L)|$ is bounded by $O\big(\mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L})\big)$. By Lemma B.2, we have

$$\begin{aligned} &J(K_{t+1}', L) - J(K_t, L) = \mathrm{Tr}((P_{K_{t+1}',L} - P_{K_t,L})\Psi_\sigma) \\ &= -2\gamma \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}) + \gamma^2 \cdot \mathrm{Tr}(E_{K_t,L}^\top (R + B^\top P_{K,L} B) E_{K_t,L} \cdot \Sigma_{K_t,L}) \\ &\leq -2\gamma \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}) + \gamma^2 \|R + B^\top P_{K,L} B\| \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}) \\ &\leq -2\gamma \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}) \\ &+ \gamma^2 (\|R\| + \sigma_{\min}(\Psi_\sigma) \cdot \|B\|^2 \cdot (J(K_0, L) + \chi)) \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}), \end{aligned} \tag{E.45}$$

where the last inequality is a result of Lemma E.1 and the induction assumption $J(K_t, L) \leq J(K_0, L)$. Set $\gamma > 0$ to be sufficiently small such that

$$\gamma \leq \big[\|R\| + \sigma_{\min}(\Psi_\sigma) \cdot \|B\|^2 \cdot (J(K_0, L) + \chi)\big]^{-1}. \tag{E.46}$$

Combining (E.45) and (E.46), applying Lemma B.3 and the fact $\Sigma_{K_t,L} \succ \Psi_\sigma$, we get that

$$\begin{aligned} J(K_{t+1}', L) - J(K_t, L) &\leq -\gamma \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L} \cdot \Sigma_{K_t,L}) \leq -\gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \mathrm{Tr}(E_{K_t,L}^\top E_{K_t,L}) \\ &\leq -\gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K(L),L}\|^{-1} \cdot (J(K_t, L) - J(K(L), L)). \end{aligned} \tag{E.47}$$

This inequality also implies that $J(K_{t+1}', L) \leq J(K_t, L) \leq J(K_0, L)$. Further, (E.47) implies that

$$\begin{aligned} J(K_{t+1}', L) - J(K(L), L) \leq &\big[1 - \gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \sigma_{\min}(R) \cdot \|\Sigma_{K(L),L}\|^{-1}\big] \\ &\cdot (J(K_t, L) - J(K(L), L)) \end{aligned}$$

□

E.7 Proof of Lemma B.5

Proof. We now bound the difference of $J(K_{t+1}', L)$ and $J(K_{t+1}, L)$. By (C.7) in Proposition 4.2, we have that

$$|J(K_{t+1}, L) - J(K_{t+1}', L)| = \big|\mathrm{Tr}\big((P_{K_{t+1},L} - P_{K_{t+1}',L})\Psi_\sigma\big)\big| \leq \|\Psi_\sigma\|_{\mathrm{F}} \cdot \|P_{K_{t+1},L} - P_{K_{t+1}',L}\|. \tag{E.48}$$

We modify Lemma 24 in Fazel et al. (2018) to construct an upper bound of $\|P_{K_{t+1},L} - P_{K_{t+1}',L}\|$, which is established in the following lemma.

Lemma E.6 (Perturbation of $P_{K,L}$ for fixed L)**.** Suppose that K' is a small perturbation of K in the sense that

$$\|K - K'\| \leq \frac{\sigma_{\min}(\Psi_\sigma)}{4 \cdot \|\Sigma_{K,L}\| \cdot \|B\| \cdot \big(\|A - BK - CL\| + 1\big)}. \tag{E.49}$$

Then, we have that

$$\|P_{K',L} - P_{K,L}\| \leq C\sigma_{\min}^{-1}(\Psi_\sigma) \cdot \|\Sigma_{K,L}\| \cdot \|K\|^2 \cdot \|R\| \cdot \|B\| \cdot \|K - K'\|. \tag{E.50}$$

Proof. See a detailed proof in Fazel et al. (2018). Note that for fixed L, the problem is equivalent to single agent LQR problem. □

To verify condition (E.49) of Lemma E.6, in what follows, we establish an upper bound of

$$4 \cdot \|\Sigma_{K_{t+1},L}\| \cdot \|B\| \cdot \big(\|A - BK_{t+1} - CL\| + 1\big) \cdot \|K_{t+1} - K'_{t+1}\|.$$

By the fact that $\|A - BK - CL\| < 1$, and Lemma E.1, we have that

$$\begin{aligned} &4 \cdot \|\Sigma_{K_{t+1},L}\| \cdot \|B\| \cdot \big(\|A - BK_{t+1} - CL\| + 1\big) \cdot \|K_{t+1} - K'_{t+1}\| \\ &\quad \leq 8\|B\| \cdot \big(J(K_{t+1}, L) + \chi\big) \cdot \|K_{t+1} - K'_{t+1}\| \end{aligned} \tag{E.51}$$

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

$$\|K_{t+1} - K'_{t+1}\| \leq \gamma \cdot (1 + \|K_t\| + \|L\|) \cdot \|\widehat{\Lambda}_{K_t,L} - \Lambda_{K_t,L}\|. \tag{E.52}$$

Combining (E.51) and (E.52), to ensure condition (E.49), it suffices to make the following inequality holds

$$8\gamma \cdot \|B\| \cdot \big(J(K_{t+1}, L) + \chi\big) \cdot (1 + \|K_t\| + \|L\|) \cdot \|\widehat{\Lambda}_{K_t,L} - \Lambda_{K_t,L}\|) \leq \sigma_{\min}(\Psi_\sigma). \tag{E.53}$$

On the other hand, to establish Lemma B.5, we need to ensure the following inequality

$$\begin{aligned} &C \cdot \sigma_{\min}^{-1}(\Psi_\sigma) \cdot \|\Sigma_{K_{t+1},L}\| \cdot \|K_{t+1}\|^2 \cdot \|R\| \cdot \|B\| \cdot \|K_{t+1} - K'_{t+1}\| \\ &\quad \leq 1/2 \cdot \gamma \cdot \sigma_{\min}(\Psi_\sigma) \cdot \sigma_{\min}(R) \cdot \|\Sigma_L^*\|^{-1} \cdot \big(J(K_t, L) - J(K(L), L)\big), \end{aligned}$$

which is implied by

$$C \cdot J(K_0(L), L)^2 \cdot \|K_{t+1}\|^2 \cdot (1 + \|K_{t+1}\| + \|L\|) \cdot \|\widehat{\Lambda}_{K_t,L} - \Lambda_{K_t,L}\| \leq \big(J(K_t, L) - J(K(L), L)\big). \tag{E.54}$$

Here we apply Lemma E.1 and assume that $\|R\|, \|B\|, \sigma_{\min}(\Psi_\sigma)$ are constants. Thus, to ensure (E.54) and (E.53), we need to ensure

$$\|\widehat{\Lambda}_{K_t,L} - \Lambda_{K_t,L}\| \leq \frac{J(K_t, L) - J(K(L), L)}{C \cdot J(K_0(L), L)^2 \cdot \|K_{t+1}\|^2 \cdot (1 + \|K_{t+1}\| + \|L\|)}. \tag{E.55}$$

By Theorem 5.2, (E.55) holds with probability $1 - T_{\mathrm{TD}}^{-4}$ by setting

$$T \geq \left(\frac{\mathrm{Poly}\big(J(K_t, L), \|K_t\|_{\mathrm{F}}, \|L\|_{\mathrm{F}}, \kappa^{-1}\big)}{\sigma_{\min}(\mathfrak{T}_{K_t,L}) \cdot \big(1 - \rho(A - BK_t - CL)\big) \cdot \big(J(K_t, L) - J(K(L), L)\big)} \right)^5,$$

which completes the proof.

□

E.8 Proof of Lemma B.7

Proof. Applying Lemma B.2, we have that

$$x^\top P_{K',L'}x - x^\top P_L^* x = \sum_{t\geq 0} A_{K(L),L,K',L'}(x_t'),$$

where $x_0' = x, x_{t+1}' = (A - BK' - CL')x_t'$. Moreover, by completing the square and that $E_L^* = 0$, we have

$$\begin{aligned}
A_{K(L),L,K',L'}(x) &= x^\top \big(K' - K(L)\big)^\top (R + B^\top P_L^* B)\big(K' - K(L)\big)x \\
&\quad + 2x^\top (L' - L)F_L^* x + x^\top (L' - L)^\top (-S + C^\top P_L^* C)(L' - L)x \\
&\quad + 2x^\top (L' - L)^\top C^\top P_L^* B\big(K' - K(L)\big)x \\
&= 2x^\top (L' - L)^\top F_L^* x + x^\top (L' - L)(-S + C^\top P_K^* C)(L' - L)x \\
&\quad + x^\top \big[K' - K(L) - (R + B^\top P_L^* B)^{-1} B^\top P_L^* C(L' - L)\big]^\top (R + B^\top P_L^* B) \\
&\quad \cdot \big[K' - K(L) - (R + B^\top P_L^* B)^{-1} B^\top P_L^* C(L' - L)\big]x \\
&\quad - x^\top (L' - L)^\top C^\top P_L^* B(R + B^\top P_L^* B)^{-1} B^\top P_L^* C(L' - L)x \\
&\geq 2x^\top (L' - L)^\top F_L^* x - x^\top (L' - L)^\top W_L (L' - L)x.
\end{aligned}$$

By Woodbury matrix identity, we have that

$$\begin{aligned}
W_L &= S - C^\top \big[P_L^* - P_L^* B(R + B^\top P_L^* B)^{-1} B^\top P_L^*\big]C \\
&= S - C^\top \big[(P_L^*)^{-1} + BS^{-1}B^\top\big]^{-1} C
\end{aligned}$$

which means that W_L^* is monotonically decreasing when P_L^* is increasing. □

E.9 Proof of Lemma B.8

Proof. By applying Lemma B.7, we have that

$$x^\top P_{\bar{L}^{n+1}}^* x - x^\top P_{L^n}^* x \geq 2\iota \cdot \sum_{s\geq 0} {x_s'}^\top \widetilde{F}_{L^n}^\top F_{L^n}^* x_s' - \iota^2 \cdot \sum_{s\geq 0} {x_s'}^\top \widetilde{F}_{L^n}^\top W_{L^n} \widetilde{F}_{L^n} x_s',$$

where $x_0' = x, x_{s+1}' = \big(A - BK(\bar{L}^{n+1}) - CL\big)x_s'$.

After taking expectation with respect to $x \sim N(0, \Psi_\sigma)$, combining with (C.7) and that $\mathcal{T}_{\bar{L}^{n+1}}^*(\Psi_\sigma) = \Sigma_{\bar{L}^{n+1}}^*$, we have that

$$J^*(\bar{L}^{n+1}) - J^*(L^n) \geq 2\iota \cdot \mathrm{Tr}\big(\Sigma_{\bar{L}^{n+1}}^* \widetilde{F}_{L^n}^\top F_{L^n}^*\big) - \iota^2 \cdot \mathrm{Tr}\big(\Sigma_{\bar{L}^{n+1}}^* \widetilde{F}_{L^n}^\top W_{L^n} \widetilde{F}_{L^n}\big). \tag{E.56}$$

We now bound the two terms on the right-hand side of (E.56) separately.

First, we bound the first term on the right of (E.56). Applying Lemma B.6 and tangle inequality, we have

$$\mathrm{Tr}(\Sigma_{\bar{L}^{n+1}}^* \widetilde{F}_{L^n}^\top F_{L^n}^*) \geq \mathrm{Tr}(\Sigma_{L^n}^* \widetilde{F}_{L^n}^\top \widetilde{F}_{L^n}) - \|\Sigma_{\bar{L}^{n+1}}^* - \Sigma_{L^n}^*\| \cdot \big|\mathrm{Tr}(\widetilde{F}_{L^n}^\top F_{L^n}^*)\big|. \tag{E.57}$$

To bound the second term of (E.57), we obbtain the following lemma from Zhang et al. (2019b).

Lemma E.7 (Lemma 6.8 in Zhang et al. (2019b))**.** Under Assumption 4.1, for any $L, L' \in \mathcal{L}$, there exists some constants $\mathcal{B}_{\mathcal{L}}^L, \mathcal{B}_{\mathcal{L}}^P, \mathcal{B}_{\mathcal{L}}^K > 0$, such that if

$$\begin{aligned}
\|L - L'\| \leq \min\Bigg\{ &\mathcal{B}_{\mathcal{L}}^L, \frac{\|B\|\big[\mathcal{B}_{\mathcal{L}}^P \|A - BK(L) - CL\| + \|P_L^*\|\|C\|\big]}{\mathcal{B}_{\mathcal{L}}^P \|B\|\|C\|}, \\
&\frac{2\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^K \|B\| + \|C\|\big)}{(\mathcal{B}_{\mathcal{L}}^K)^2\|B\|^2 + \|C\|^2 + 2\mathcal{B}_{\mathcal{L}}^K \|B\|\|C\|} \Bigg\},
\end{aligned} \tag{E.58}$$

it holds that

$$\|\Sigma_L^* - \Sigma_{L'}^*\| \leq 4\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^K \|B\| + \|C\|\big) \cdot \|L - L'\|.$$

Thus, if we set $\mathcal{K}_{\mathcal{L}} > 0$ to denote the infimum taken over all of the constraints of $\|L' - L\|$ in (E.58), i.e.,

$$\mathcal{K}_{\mathcal{L}} = \inf_{L\in\mathcal{L}} \Bigg\{ \mathcal{B}_{\mathcal{L}}^{L}, \frac{\|B\|\big[\mathcal{B}_{\mathcal{L}}^{P}\|A - BK(L) - CL\| + \|P_L^*\|\|C\|\big]}{\mathcal{B}_{\mathcal{L}}^{P}\|B\|\|C\|}, \\ \frac{2\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^{K}\|B\| + \|C\|\big)}{(\mathcal{B}_{\mathcal{L}}^{K})^2\|B\|^2 + \|C\|^2 + 2\mathcal{B}_{\mathcal{L}}^{K}\|B\|\|C\|} \Bigg\}$$

combining with (E.57), we have

$$\begin{aligned} \mathrm{Tr}(\Sigma_{\bar{L}^{n+1}}^{*}\widetilde{F}_{L^n}^{\top}F_{L^n}^{*}) \geq\ & \sigma_{\min}(\Psi_\sigma)^{-1}\cdot\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2 \\ & - 16\iota\cdot\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^{K}\|B\| + \|C\|\big)\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2\|F_{L^n}^{*}\|_{\mathrm{F}}, \end{aligned} \tag{E.59}$$

provided that

$$\|\bar{L}^{n+1} - L^n\| = \iota\cdot\|\widetilde{F}_{L^n}\| \leq \mathcal{K}_{\mathcal{L}}. \tag{E.60}$$

Note that we can bound $\|\widetilde{F}_{L^n}\|$ by the following lemma from Zhang et al. (2019b).

Lemma E.8 (Lemma 6.9 in Zhang et al. (2019b))**.** Let $\|\widetilde{F}_L\|$ be the projected policy gradient defined in (B.11). Then it holds that

$$\begin{aligned} \frac{2}{\sqrt{p}}\sigma_{\min}(\Psi_\sigma)\cdot\|\widetilde{F}_L\|_{\mathrm{F}} \leq\ & \|\nabla_L J^*(L)\| \\ \leq\ & \frac{2J^*(L) + 2\chi}{\kappa}\cdot\sqrt{\frac{\|W_L\|\big[J(K^*, L^*) - J^*(L)\big]}{\sigma_{\min}(\Psi_\sigma)}}, \end{aligned}$$

where $p = \min\{d, m_2\}$.

By Lemma E.8, to ensure (E.60), we have the following requirement.

$$\iota \leq \bar{\iota}_1^L = \frac{\kappa\cdot\mathcal{K}_{\mathcal{L}}}{2J\big(K(L^n), L^n\big) + 2\chi}\cdot\sqrt{\frac{\sigma_{\min}(\Psi_\sigma)}{\|W_{L^n}\|\big[J(K^*, L^*) - J\big(K(L^n), L^n\big)\big]}}$$

For the second term on the right side of (E.56), noticing that by Lemma E.1 and Lemma B.7, it holds that

$$\Sigma_{\bar{L}^{n+1}}^{*}\widetilde{F}_{L^n}^{\top}W_{L^n}\widetilde{F}_{L^n} \leq \|\Sigma_{L^n}^{*}\|_{\mathrm{F}}\cdot\|W_{L^n}\|_{\mathrm{F}}\cdot\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2 \leq \sqrt{m_2}\frac{J(K(L^n), L^n)\cdot\|S\|_{\mathrm{F}}}{\kappa}\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2 \tag{E.61}$$

Thus, substituting (E.59) and (E.61) to (E.56), we have that

$$\begin{aligned} J^*(\bar{L}^{n+1}) - J^*(L^n) \geq\ & \iota\cdot\sigma_{\min}(\Psi_\sigma)^{-1}\cdot\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2\Big(1 - \iota\cdot\sqrt{m_2}\frac{J(K(L^n), L^n)\cdot\|S\|_{\mathrm{F}}\cdot\sigma_{\min}(\Psi_\sigma)}{\kappa} \\ & - 16\iota\cdot\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^{K}\|B\| + \|C\|\big)\|F_{L^n}^{*}\|_{\mathrm{F}}\Big), \end{aligned} \tag{E.62}$$

which gives the other requirement on ι

$$\begin{aligned} \iota \leq \bar{\iota}_2^L = \frac{1}{2}\cdot\Big[\sqrt{m_2}\frac{J(K(L^n), L^n)\cdot\|S\|_{\mathrm{F}}\cdot\sigma_{\min}(\Psi_\sigma)}{\kappa} \\ + 16\cdot\big(\|A - BK(L) - CL\| + 1\big)\big(\mathcal{B}_{\mathcal{L}}^{K}\|B\| + \|C\|\big)\|F_{L^n}^{*}\|_{\mathrm{F}}\Big]^{-1} \end{aligned}$$

Note that if we denote $\bar{\iota} = \inf_{L\in\mathcal{L}}\{\bar{\iota}_1^L, \bar{\iota}_2^L\}$, then $\bar{\iota} > 0$. Further, by (E.62), for any $\iota \in (0, \bar{\iota})$, it holds that

$$J^*(\bar{L}^{n+1}) - J^*(L^n) \geq \iota/2\cdot\sigma_{\min}(\Psi_\sigma)^{-1}\cdot\|\widetilde{F}_{L^n}\|_{\mathrm{F}}^2,$$

which completes the proof. □

E.10 Proof of Lemma B.10

Proof. By the expression of $F_{K,L}$ in (C.11), we have

$$\begin{aligned}\|F_{K,L} - F_L^*\| &= \|C^\top P_{K,L}\big(CL + BK - A\big) - C^\top P_L^*\big(CL + BK(L) - A\big)\| \\ &\le \|C\| \cdot \big[\|P_L^*\| \cdot \|B\| \cdot \|K - K(L)\| + \|P_{K,L} - P_L^*\| \cdot \|A - BK - CL\|\big] \\ &\le \|C\| \cdot \big[\|B\| \cdot \|P^*\| \cdot \|K - K(L)\| + \|P_{K,L} - P_L^*\|\big], \end{aligned} \tag{E.63}$$

where the last inequality holds as a result of the fact that $\rho(A - BK - CL) < 1$, and $P^* \succeq P_L^*$. For the first term on the right-hand side of the last inequality, by Lemma B.2, we have that

$$\begin{aligned} J(K,L) - J^*(L) &= \mathrm{Tr}\Big((K - K(L))^\top (R + B^\top P_L^* B)(K - K(L)) \cdot \Psi_\sigma\Big) \\ &\ge \sigma_{\min}(\Psi_\sigma) \cdot \mathrm{Tr}((K - K(L))^\top (R + B^\top P_L^* B)(K - K(L))) \\ &\ge \sigma_{\min}(R) \cdot \sigma_{\min}(\Psi_\sigma) \cdot \|K - K(L)\|_{\mathrm{F}}^2. \end{aligned} \tag{E.64}$$

For the second term on the right-hand side of the last inequality, by (C.7), we have that

$$\|P_{K,L} - P_L^*\| \le \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \big[J(K,L) - J^*(L)\big]. \tag{E.65}$$

Thus, plugging (E.64) and (E.65) into (E.63), we have that

$$\begin{aligned}\|F_{K,L} - F_L^*\| &\le \sigma_{\min}(R)^{-1/2} \cdot \sigma_{\min}(\Psi_\sigma)^{-1/2} \cdot \|C\| \cdot \|B\| \cdot \|P^*\| \cdot \big[J(K,L) - J^*(L)\big]^{1/2} \\ &\quad + \sigma_{\min}(\Psi_\sigma)^{-1} \cdot \|C\| \cdot \big[J(K,L) - J^*(L)\big] \\ &\le C \cdot \big[J(K,L) - J(K(L),L)\big], \end{aligned}$$

for an absolute constant C, which completes the proof. □