A Reinforcement Learning Look at Risk-Sensitive Linear Quadratic Gaussian Control

Leilei Cui and Zhong-Ping Jiang

Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University,
370 Jay Street, Brooklyn, New York, 11201, USA
l.cui@nyu.edu, zjiang@nyu.edu

Abstract

This paper proposes a novel robust reinforcement learning framework for discrete-time systems with model mismatch that may arise from the sim2real gap. A key strategy is to invoke advanced techniques from control theory. Using the formulation of the classical risk-sensitive linear quadratic Gaussian control, a dual-loop policy iteration algorithm is proposed to generate a robust optimal controller. The dual-loop policy iteration algorithm is shown to be globally exponentially and uniformly convergent, and robust against disturbance during the learning process. This robustness property is called small-disturbance input-to-state stability and guarantees that the proposed policy iteration algorithm converges to a small neighborhood of the optimal controller as long as the disturbance at each learning step is small. In addition, when the system dynamics is unknown, a novel model-free off-policy policy iteration algorithm is proposed for the same class of dynamical system with additive Gaussian noise. Finally, numerical examples are provided for the demonstration of the proposed algorithm.

Introduction

By optimizing a specified accumulated performance index, reinforcement learning (RL) is a branch of machine learning that is aimed to learn optimal decisions from data in the absence of precise model knowledge. Recently, it has shown wonderful successes in a wide range of fields, for example video games (Mnih et al. 2015), robotics (Kober, Bagnell, and Peters 2013), and recommendation systems (Shani, Heckerman, and Brafman 2005). During the exploration and training phase of RL, the policy should be randomly sampled to fully excite the system, which cannot be tolerated by real-world engineering systems such as self-driving cars. To alleviate safety concern, the agent is typically trained in a simulator and then transferred to real-world plants. Since the model in the simulator may not be a good representation of the real-world plant, the model mismatch induced by sim2real gap seems unavoidable. Another source of model mismatch arises from the changing operational conditions of plants. For example, as shown in our recent work (Cui et al. 2021), the height of a wheel-legged robot may change during its movement, resulting in a different dynamical model of the robot. These model mismatches can be detrimental for the safe operation of the system (Mankowitz et al. 2019), and it is desirable to train a robust control policy, which is agnostic to model mismatches and disturbance.

The robust and optimal control theory, especially the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, is a key strategy to handle the model mismatches (Zhou, Doyle, and Glover 1996; Doyle et al. 1989; Mustafa and Bernstein 1991; Apkarian, Noll, and Rondepierre 2008). By minimizing an upperbound of the $\mathcal{H}_2$ norm and constraining the $\mathcal{H}_\infty$ norm under a given threshold, mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is capable of guaranteeing the robustness of the closed-loop system to model mismatches without greatly sacrificing the performance of the system. Risk-sensitive linear quadratic Gaussian control (Jacobson 1973; Whittle 1981) and linear quadratic zero-sum differential game (Başar 2008) are two classical approaches for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller design. From the past literature, e.g. (Jacobson 1973; Başar 2008), the optimal and robust controller can be derived from the solution of the related generalized algebraic Riccati equation (GARE). Since GARE is a nonlinear matrix equation, it is hard to solve it directly, which motivates the development of various iterative algorithms to solve the GARE (Li and Gajic 1995; Lanzon et al. 2008; Abu-Khalaf, Lewis, and Huang 2006; Zhang, Yang, and Başar 2019; Zhang, Hu, and Başar 2019; Bu, Ratliff, and Mesbahi 2019). However, there are still unsolved fundamental problems for these iterative algorithms.

These iterative algorithms approximate the solution of the GARE by iteratively solving a series of AREs. However, at each iteration, ARE is still a nonlinear matrix equation and cannot be solved directly. Therefore, the purpose of this paper is to develop a dual-loop iterative algorithm, where the outer loop approximates the solution of GARE by iteratively solving a series of AREs, and the inner loop approximates the solution of each ARE by iteratively solving a series of Lyapunov equations. A fundamental challenge of the convergence of the dual-loop iterative algorithm is to address the uniform convergence issue tied to the inner loop. Specifically, the required number of inner-loop iterations should be independent of the outer-loop iteration. Otherwise, as the outer-loop iteration moving on, the required number of inner-loop iterations may grow explosively, thus making the dual-loop algorithm not practically implementable. To the best of our knowledge, the uniform convergence of the dual-loop algorithm is not theoretically analyzed in the existing literature.

In addition, the iterative algorithm cannot be implemented accurately in reality, due to the influence of various errors arising from state estimation, sensor noise, external disturbance, and modeling error, etc. Hence, a fundamental question arises: Is the iterative algorithm robust to the errors? In particular, does the iterative algorithm still converge to a neighbourhood of the accurate solution of GARE in the presence of various errors and, if yes, what is the size of the neighborhood? For both the outer and inner loops, the iterative process is nonlinear, and the robustness of the iterative algorithm has not been fully understood in the existing literature.

In this paper, we investigate the uniform convergence and robustness of the dual-loop iterative algorithm for solving the risk-sensitive linear quadratic Gaussian control. Even though the convergence of the dual-loop iterative algorithm is analyzed separately in (Zhang, Hu, and Başar 2019; Hewer 1971), the uniform convergence and the robustness of the overall algorithm are still open problems. To analyze the uniform convergence of the dual-loop algorithm, the key idea is to demonstrate the global linear convergence of the inner-loop iteration and find the upperbound of the linear convergence rate. To address the robustness issue, a key strategy of the paper is to invoke techniques from advanced control theory, such as input-to-state stability (ISS) (Sontag 2008) and its latest variant "small-disturbance ISS" (Pang and Jiang 2021) to analyze the robustness of the proposed discrete-time iterative algorithm. In the presence of noise during the learning process, it is demonstrated that the iterative algorithm still converges to a small neighbourhood of the optimal solution, as long as the noise is small and bounded. Furthermore, based on these results, an off-policy data-driven RL algorithm is proposed when the system is disturbed by an immeasurable Gaussian noise. Several numerical examples are given to validate the efficacy of our theoretical results.

To sum up, our main contributions are three-fold: 1) the uniform convergence of the dual-loop iterative algorithm is theoretically analyzed; 2) under the framework of the small-disturbance ISS, the robustness of both the outer and inner loops is theoretically demonstrated; 3) a novel off-policy policy iteration RL algorithm is proposed.

Notations

$\mathbb{R}$ and $\mathbb{C}$ are the sets of real and complex numbers, respectively. $\mathbb{Z}$ ($\mathbb{Z}_+$) is the set of (positive) integers. $\mathbb{S}^n$ is the set of n-dimensional real symmetric matrices. $|a|$ denotes the Euclidean norm of the vector a. $\|\cdot\|$ and $\|\cdot\|_F$ denote the spectral norm and the Frobenius norm of a matrix. ℓ_2 is the space of square-summable sequences equipped with the norm $\|\cdot\|_{\ell_2}$. ℓ_∞ is the space of bounded sequences equipped with the norm $\|\cdot\|_\infty$. $\bar{\sigma}(\cdot)$ and $\underline{\sigma}(\cdot)$ are the maximum and minimum singular values of a given matrix, respectively. For a transfer function $G(z)$, its $\mathcal{H}_\infty$ norm is defined as $\|G\|_{\mathcal{H}_\infty} := \sup_{\omega\in[0,2\pi]} \bar{\sigma}(G(e^{j\omega}))$, which is equivalent to $\|G\|_{\mathcal{H}_\infty} := \sup_{u\in\ell_2} \frac{\|Gu\|_{\ell_2}}{\|u\|_{\ell_2}}$.

For a matrix $X \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(X) := [x_1^T, \cdots, x_n^T]^T$, where x_i is the ith column of X. For a matrix $P \in \mathbb{S}^n$, $\mathrm{vecs}(P) := [p_{1,1}, 2p_{1,2}, \cdots, 2p_{1,n}, p_{2,2}, 2p_{2,3}, \cdots, p_{n,n}]^T$, where $p_{i,j}$ is the ith row and jth column entry of the matrix P. $[X]_{i,j}$ denotes the submatrix of the matrix X that is comprised of the rows between the ith and jth rows of X. For a vector $a \in \mathbb{R}^n$, $\mathrm{vecv}(a) := [a_1^2, a_1a_2, \cdots, a_1a_n, a_2^2, a_2a_3, \cdots, a_n^2]^T$. $\mathcal{B}(X, r) := \{Y \in \mathbb{R}^{m\times n} \mid \|Y - X\|_F \le r\}$ denotes the ball centered at X with the radius r.

Preliminaries

In this section, we begin with the problem formulation of linear exponential quadratic Guassian (LEQG) control. Then, we discuss its relation to linear quadratic zero-sum differential game (DG) and its robustness analysis.

Linear Exponential Quadratic Guassian Control

Consider the discrete-time linear time-invariant system

$$x_{t+1} = Ax_t + Bu_t + Dw_t \quad x_0 = x_{ini}, \tag{1a}$$
$$y_t = Cx_t + Eu_t, \tag{1b}$$

where $x_t \in \mathbb{R}^n$ is the state of the system, $u_t \in \mathbb{R}^m$ is the control input, x_{ini} is the initial state, $w_t \in \mathbb{R}^q \sim \mathcal{N}(0, I_q)$, $t = 0, 1, \cdots$ are independent and identically distributed random variables, and $y_t \in \mathbb{R}^p$ is the controlled output. A, B, C, D, E are constant matrices with the compatible dimensions. The LEQG control aims to find the input sequence $u := \{u_t\}_{t=0}^\infty$ such that the following risk-averse exponential quadratic cost is minimized

$$\mathcal{J}_{LEQG}(x_{ini}, u) := \lim_{\tau\to\infty} \frac{2\gamma^2}{\tau} \log\left[\mathbb{E}\exp\left(\frac{1}{2\gamma^2}\sum_{t=0}^{\tau} y_t^T y_t\right)\right] \tag{2}$$

where γ is a positive constant. The study of LEQG problem is initiated by (Jacobson 1973; Whittle 1981; Speyer, Deyst, and Jacobson 1974) and revisited recently by (Zhang, Hu, and Başar 2019; Roulet et al. 2020) for safe control.

Assumption 1. *(A, B) is stabilizable, $C^TC = Q \succ 0$, and $\gamma \ge \gamma_\infty$, where $\gamma_\infty > 0$ is the minimal value such that the solution to (4) exists.*

Assumption 2. *The matrices in (1b) satisfy $E^TE = R \succ 0$, and $C^TE = 0$.*

Assumption 1 ensures the existence of a stabilizing solution to the LEQG control problem. As demonstrated in (Başar 2008, Theorem 3.8), γ_∞ is finite. Assumption 2 is a standard condition to simplify the LEQG control problem by eliminating the cross term in the cost (2) between the control input u and state x. Stabilizability of the pair (A, B) implies that there exists a feedback gain $K \in \mathbb{R}^{m\times n}$ such that the spectral radius $\rho(A - BK) < 1$. Henceforth, a matrix is stable if its spectral radius is less than one, and K is stabilizing if $A - BK$ is stable.

Under Assumptions 1 and 2, as investigated by (Jacobson 1973) and (Zhang, Hu, and Başar 2019, Lemma C.2), LEQG admits a unique optimal controller $u_t^* = -K^*x_t$, where

$$K^* = (R + B^TU^*B)^{-1}B^TU^*A. \tag{3}$$

with $P^* \in \mathbb{S}^n$ the unique positive definite solution to the generalized algebraic Riccati equation (GARE)

$$(A - BK^*)^T U^*(A - BK^*) - P^* + Q + (K^*)^T RK^* = 0, \tag{4a}$$

$$U^* = P^* + P^* D(\gamma^2 I_q - D^T P^* D)^{-1} D^T P^*. \tag{4b}$$

Linear Quadratic Zero-Sum Differential Game

The differential game can be mathematically formulated as

$$\min_u \max_w \mathcal{J}_{DG}(x_{ini}, u, w) := \sum_{t=0}^{\infty} y_t^T y_t - \gamma^2 w_t^T w_t, \tag{5}$$

subject to (1a),

where $u := \{u_t\}_{t=0}^{\infty}$ and $w := \{w_t\}_{t=0}^{\infty}$ are the input sequences for the minimizer and the maximizer, respectively.

From (Başar 2008, Equation 3.51), it is seen that the optimizers for the min-max problem are $u_t^* = -K^* x_t$ and $w_t^* = L^* x_t$, where K^* is defined in (3) and L^* is written as

$$L^* = (\gamma^2 I_q - D^T P^* D)^{-1} D^T P^*(A - BK^*). \tag{6}$$

Furthermore, $(A - BK^*)$ is stable, $I_q - \gamma^{-2} D^T P^* D \succ 0$, and $(A - BK^* + DL^*)$ is stable.

Therefore, the minimizer of DG shares the same optimal controller as LEQG, and solving GARE (4) is critical for the LEQG and DG problems.

Robustness Analysis

With any stabilizing feedback $u_t = -Kx_t$, the discrete-time transfer function from w to y can be expressed as

$$\mathcal{T}(K) := (C - EK)[zI_n - (A - BK)]^{-1} D. \tag{7}$$

where $z \in \mathbb{C}$ is the z-transform variable.

As shown in Fig. 1, Δ denotes the model mismatch induced by the sim2real gap, and it satisfies $\|\Delta\|_{\mathcal{H}_\infty} < \frac{1}{\gamma}$. Thanks to the small-gain theorem (Zhou, Doyle, and Glover 1996; Jiang and Liu 2018; Zames 1966), when influenced by the model mismatch, the system is stable as long as $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} \le \gamma$. Consequently, the controller $u_t = -Kx_t$ is robust to the model mismatch Δ if K lies within the feasible set $\mathcal{W}$ defined as

$$\mathcal{W} := \{K \in \mathbb{R}^{m\times n} | (A - BK) \text{ is stable, } \|\mathcal{T}(K)\|_{\mathcal{H}_\infty} \le \gamma\}. \tag{8}$$

As investigated in (Başar 2008, Theorem 3.8), the LEQG control in (3) satisfies $K^* \in \mathcal{W}$, and therefore, it is optimal with respect to (2) and robust to the model mismatch.

The GARE (4) is a non-trivial matrix-valued nonlinear equation. Requiring the accurate knowledge of the system matrices (A, B) to solve (4) makes it more difficult. Thus, to overcome this obstacle, in this paper, we propose a learning-based robust optimal control framework without assuming the exact model knowledge.

Problem 1. *Given an initial feasible controller $K_1 \in \mathcal{W}$, design a learning-based control algorithm such that near-optimal control gains, i.e. approximate values of K^*, can be learned from input-state data collected along the trajectories of system* (1a).

We will first introduce the model-based iterative algorithm whose convergence and robustness properties are instrumental for the learning-based algorithm.

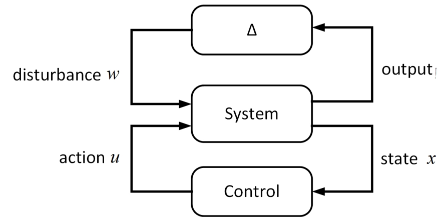

Figure 1: Robust control design with model mismatch Δ.

Model-based Policy Iteration

In this section, by transferring the nonlinear GARE (4) into a series of linear Lyapunov equations, a model-based dual-loop policy iteration algorithm is proposed to solve the GARE (4).

Introduction of the Outer Loop

The outer-loop iteration is developed based on the results of (Zhang, Hu, and Başar 2019, Equation 3.5), and it aims to update the control policy $u_t = -K_i x_t$ under the worst-case disturbance w. Let i denotes the iteration index for the outer loop and the following variables are introduced to simplify the notations

$$A_i := A - BK_i, \quad Q_i := Q + K_i^T RK_i, \tag{9a}$$

where A_i is the closed-loop transition matrix with $u_t = -K_i x_t$, and Q_i is the cost weighting matrix. Then, the outer loop iteration can be expressed as

$$A_i^T U_i A_i - P_i + Q_i = 0, \tag{10a}$$

$$K_{i+1} = (R + B^T U_i B)^{-1} B^T U_i A, \tag{10b}$$

where U_i is

$$U_i := P_i + P_i D(\gamma^2 I_q - D^T P_i D)^{-1} D^T P_i. \tag{11}$$

In (10), we consider (10a) as the policy evaluation step under the worst case disturbance and (10b) as the policy improvement step. P_i is the cost matrix of (5) for the controller $u_t = -K_i x_t$ under the worst case disturbance.

As studied in (Zhang, Hu, and Başar 2019, Theorems 4.3, 4.4, and 4.6), for each iteration, the controller K_i generated by (10) preserves the robustness to model mismatch, i.e. $K_i \in \mathcal{W}$. P_i converges to P^* with a globally sublinear and locally quadratic rate. We further investigate the convergence rate of (10) and rigorously demonstrate that P_i is monotonically decreasing ($P_{i+1} \succeq P_i$), and globally and exponentially converges to the optimal solution P^* (with a linear convergence rate). The proof of the following theorem is postponed to Appendix B.

Theorem 1. *Given $K_1 \in \mathcal{W}$, for any $i \in \mathbb{N}_+$, there exists $\alpha \in [0, 1)$, such that*

$$\mathrm{Tr}(P_{i+1} - P^*) \le \alpha \,\mathrm{Tr}(P_i - P^*). \tag{12}$$

Since $P_{i+1} - P^* \succeq 0$ and $\|P_{i+1} - P^*\|_F \le \mathrm{Tr}(P_{i+1} - P^*) \le \sqrt{n}\|P_{i+1} - P^*\|_F$ (Lemma 7 in Appendix A), the following inequlity holds

$$\|P_{i+1} - P^*\|_F \le \alpha^i \,\mathrm{Tr}(P_1 - P^*) \le \sqrt{n}\alpha^i \|P_1 - P^*\|_F. \tag{13}$$

As we can see, the outer loop iteration (10a) is still a nonlinear matrix-valued equation with respect to P_i. It is nontrival to solve it directly to calculate P_i. In the following subsection, the inner-loop iteration is developed to transform the nonlinear equation into a series of linear Lyapunov equations.

Introduction of the Inner Loop

Let j denote the iteration index of the inner loop. Given the feedback gain of the minimizer K_i, the inner loop iteratively finds the optimal controller for the maximizer w, that is

$$\max_{w} \mathcal{J}_{DG}(x_{ini}, w) = \sum_{t=0}^{\infty} y_t^T y_t - \gamma^2 w_t^T w_t, \quad \text{subject to } x_{t+1} = (A - BK_i)x_t + Dw_t. \tag{14}$$

The optimal solution is $w_{i,*} = L_{i,*}x_t$ where

$$L_{i,*} := (\gamma^2 I_q - D^T P_i D)^{-1} D^T P_i A_i. \tag{15}$$

The following variables are introduced to simplify the notations.

$$A_{i,j} := A - BK_i + DL_{i,j},\ A_{i,*} := A - BK_i + DL_{i,*} \tag{16a}$$
$$A^* := A - BK^* + DL^*. \tag{16b}$$

Inspired by (Hewer 1971), the inner loop is designed as

$$A_{i,j}^T P_{i,j} A_{i,j} - P_{i,j} + Q_i - \gamma^2 L_{i,j}^T L_{i,j} = 0, \tag{17a}$$
$$L_{i,j+1} = (\gamma^2 I_q - D^T P_{i,j} D)^{-1} D^T P_{i,j} A_i. \tag{17b}$$

We consider (17a) as the policy evaluation step and (17b) as the policy improvement step for the inner loop. $P_{i,j}$ is the cost matrix of (5) with the policies $u_t = -K_i x_t$ and $w_t = L_{i,j}x_t$. The inner-loop policy iteration possesses the monotonicity property and preserves the stability, that is the sequence $\{P_{i,j}\}_{j=1}^{\infty}$ is monotonically increasing and upper bounded by P_i, and $A - BK_i + DL_{i,j}$ is stable. These results are shown in Lemma 16. We rigorously show that the inner loop globally and exponentially converges to the optimal solution P_i (with a linear convergence rate). The details of the proof are shown in Appendix C.

Theorem 2. *Given $L_{i,j} = 0$, there exists a constant $\beta(K_i) \in [0, 1)$, such that*

$$\mathrm{Tr}(P_i - P_{i,j+1}) \leq \beta(K_i)\,\mathrm{Tr}(P_i - P_{i,j}). \tag{18}$$

Based on Theorem 2, we can further obtain that

$$\|P_i - P_{i,j+1}\|_F \leq \sqrt{n}\beta^j(K_i)\|P_i - P_{i,1}\|_F. \tag{19}$$

The dual-loop policy iteration algorithm is detailed in Algorithm 1. As we can see, the inner loop is from line 5 to line 10, and the generated sequence $\{P_{i,j}\}$ converges to P_i. Then, the outer loop updates the controller K_{i+1} for the minimizer at line 12. In this algorithm, only the linear Lyapunov equation (17a) is solved, so P^*, the solution of (4), is approximated by a series of linear Lyapunov equations.

Algorithm 1: Model-Based Iterative Algorithm

1 Initialize $K_1 \in \mathcal{W}$;
2 **for** $i \leq \bar{i}$ **do**
3   Initialize $j = 1$ and $L_{i,1} = 0$;
4     $Q_i = C^T C + K_i^T R K_i$;
5   **repeat**
6         $A_{i,j} = A - BK_i + DL_{i,j}$;
7   Get $P_{i,j}$ by solving (17a);
8   Update $L_{i,j+1}$ by (17b);
9         $j \leftarrow j + 1$
10   **until** $\|P_{i,j} - P_{i,j-1}\|_F \leq \epsilon$;
11     $U_{i,\bar{j}} = P_{i,\bar{j}} + P_{i,\bar{j}} D(\gamma^2 I_n - D^T P_{i,\bar{j}} D)^{-1} D^T P_{i,\bar{j}}$;
12   Update K_{i+1} by (10b) ;
13 **end**

Convergence Analysis for the Dual-Loop Algorithm

For the dual-loop algorithm, the inner-loop iteration exponentially converges to the optimal solution P_i with the rate dependent on K_i. Since K_i is updated iteratively, it is required that the inner loop enters the given neighborhood of P_i within a constant number of steps, regardless of different K_i. The uniform convergence rate of the overall algorihtm is given in the following theorem, whose proof is given in Appendix D.

Theorem 3. *For any $i \in \mathbb{N}_+$ and $\epsilon > 0$, there exists $\bar{j} \in \mathbb{N}_+$ independent of i, such that for all $j \geq \bar{j}$, $\|P_{i,j} - P_i\|_F \leq \epsilon$.*

Robustness Analysis for the Dual-Loop Algorithm

In the last section, the exact policy iteration algorithm is introduced in the sense that the accurate knowledge of system matrices (A, B) is required to implement the algorithm. In practice, we cannot access to such an accurate model, and for the outer and inner loops, the updates for the controllers in (10b) and (17b) are subjected to noise. Such noise may be induced by noisy input-state data (Pang and Jiang 2021) and modeling errors in indirect adaptive control (Åström and Wittenmark 2008), system identification, and model-based reinforcement learning (Tu and Recht 2019). In this section, using the well-known concept of input-to-state stability in control theory, we will analyze the robustness of the dual-loop policy iteration algorithm in the presence of disturbance.

Notions of Input-to-State Stability

Consider the general nonlinear discrete-time system

$$\chi_{k+1} = f(\chi_k, v_k). \tag{20}$$

where $\chi_k \in \mathcal{X}$, $v_k \in \mathcal{V}$, and f is continuous. χ_e is the equilibrium of the system, that is $0 = f(\chi_e, 0)$.

Definition 1. *(Hahn 1967) A function $\xi(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a $\mathcal{K}$-function if it is continuous, strictly increasing and vanishes at zero. A function $\kappa(\cdot, \cdot) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is a*

$\mathcal{KL}$-function if for any fixed $t \geq 0$, $\kappa(\cdot, t)$ is a $\mathcal{K}$-function, and for any $r \geq 0$, $\kappa(r, \cdot)$ is decreasing and $\kappa(r,t) \to 0$ as $t \to \infty$.

Definition 2. *(Jiang and Wang 2001) System (20) is input-to-state stable if there exist a $\mathcal{KL}$-function κ and a $\mathcal{K}$-function ξ such that for each input $v \in \ell_\infty$ and initial state $\chi_1 \in \mathcal{X}$ the following holds*

$$\|\chi_k - \chi_e\| \leq \kappa(\|\chi_1 - \chi_e\|, k) + \xi(\|v\|_\infty). \tag{21}$$

for any $k \in \mathbb{N}_+$.

Generally speaking, input-to-state stability characterizes the influence of input v to the evolution of state χ. The deviation of the state χ to the equilibrium is bounded as long as the input v is bounded. Furthermore, the influence of the initial deviation $\|\chi_1 - \chi_e\|$ vanishes as time tends to infinity.

Robustness Analysis for the Outer Loop

The exact outer loop iteration is shown in (10), and in the presence of disturbance, it is modified as

$$\hat{A}_i^T \hat{U}_i \hat{A}_i - \hat{P}_i + \hat{Q}_i = 0, \tag{22a}$$

$$\hat{K}_{i+1} = (R + B^T \hat{U}_i B)^{-1} B^T \hat{U}_i A + \Delta K_{i+1}, \tag{22b}$$

where ΔK_i is the disturbance at the ith iteration, and the hat is used to distinguish the sequences generated by the exact (10) and inexact (22) outer-loop iterations. By considering (22) as a discrete-time nonlinear system with the state $\hat{P}_i$ and input ΔK_i, it is shown that (22) is inherently robust to ΔK_i in the sense of small-disturbance input-to-state stability (Pang and Jiang 2021; Pang, Bian, and Jiang 2022).

Theorem 4. *There exists a constant $d > 0$, such that if $\|\Delta K\|_\infty < d$, (22) is input-to-state stable.*

Robustness Analysis for the Inner Loop

As a counterpart of inexact outer-loop iteration, the inexact inner-loop iteration can be developed as

$$\hat{A}_{i,j}^T \hat{P}_{i,j} \hat{A}_{i,j} - \hat{P}_{i,j} + \hat{Q}_i - \gamma^2 \hat{L}_{i,j}^T \hat{L}_{i,j} = 0, \tag{23a}$$

$$\hat{L}_{i,j+1} = (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T \hat{P}_{i,j} \hat{A}_i + \Delta L_{i,j+1}. \tag{23b}$$

Here, $\Delta L_{i,j+1}$ denotes the disturbance to the inner loop iteration and the hat emphasizes that the corresponding sequences are generated by the inexact iteration. With the inexact inner loop in hand, the following theorem shows that the inner loop iteration (23) is robust to disturbance $\Delta L_{i,j}$ in the sense of small-disturbance input-to-state stability.

Theorem 5. *There exists a constant $e(K_i) > 0$, such that if $\|\Delta L_i\|_\infty < e(K_i)$, (23) is input-to-state stable.*

These two original theorems guarantee the robustness of the dual-loop policy iteration algorithm. Literally speaking, when the dual-loop policy iteration algorithm is implemented in the presence of disturbance, it still finds the near optimal solution, and the deviation between the generated policy and the optimal one is determined by the magnitude of the disturbance. To be more specific, as iteration i (j for inner loop) goes to infinite, the cost matrix $\hat{P}_i$ ($\hat{P}_{i,j}$) enters a small neighborhood of the optimal solution P^* ($\hat{P}_i$).

Learning-based Off-Policy Policy Iteration

For system (1a) with additive Gaussian noise, we will develop a learning-based algorithm to learn from data a robust suboptimal controller (i.e., an approximation of K^*) without requiring the accurate knowledge of (A, B). Suppose that the exploratory control policy is

$$u_t = -\hat{K}_1 x_t + \sigma_u \xi, \quad \xi \sim \mathcal{N}(0, I_m). \tag{24}$$

where $\hat{K}_1 \in \mathcal{W}$ is the initial feasible controller, and $\sigma_u > 0$ is the standard deviation of the exploratory noise.

For any matrix $X \in \mathbb{S}^n$, along the trajectories of system (1a), $X x_{t+1}$ and $x_{t+1}^T X x_{t+1}$ are

$$X x_{t+1} = X A x_t + X B u_t + X D w_t \tag{25a}$$

$$\begin{aligned} x_{t+1}^T X x_{t+1} &= x_t^T A^T X A x_t + u_t^T B^T X B u_t + w_t^T D^T X D w_t \\ &+ 2u_t^T B^T X A x_t + 2 w_t^T D^T X A x_t + 2 u_t^T B^T X D w_t. \end{aligned} \tag{25b}$$

By vectorizing, one can rewrite (25) as

$$\begin{aligned} (x_{t+1}^T \otimes I_n) \operatorname{vec}(X) &= (x_t^T \otimes I_n) \operatorname{vec}(XA) \\ &+ (u_t^T \otimes I_n) \operatorname{vec}(XB) + X D w_t \end{aligned} \tag{26a}$$

$$\begin{aligned} \operatorname{vecv}^T(x_{t+1}) \operatorname{vecs}(X) &= \operatorname{vecv}^T(x_t) \operatorname{vecs}(A^T X A) \\ &+ \operatorname{vecv}^T(u_t) \operatorname{vecs}(B^T X B) + 2(x_t^T \otimes u_t^T) \operatorname{vec}(B^T X A) \\ &+ w_t^T D^T X D w_t + 2 w_t^T D^T X A x_t + 2 u_t^T B^T X D w_t \end{aligned} \tag{26b}$$

Define ϕ_t', ϕ_t, Γ', and Γ as

$$(\phi_t')^T := \left[(x_t^T \otimes I_n), (u_t^T \otimes I_n)\right], \tag{27a}$$

$$\phi_t := \left[\operatorname{vecv}^T(x_t), \operatorname{vecv}^T(u_t), 2(x_t^T \otimes u_t^T), 1\right]^T, \tag{27b}$$

$$\Gamma'(X) := \left[\operatorname{vec}^T(XA), \operatorname{vec}^T(XB)\right]^T, \tag{27c}$$

$$\begin{aligned} \Gamma(X) := [&\operatorname{vecs}^T(A^T X A), \operatorname{vecs}^T(B^T X B), \\ &\operatorname{vec}^T(B^T X A), \operatorname{Tr}(D^T X D)]^T. \end{aligned} \tag{27d}$$

Assumption 3. *$\mathbb{E}_\pi\left[\phi_t \phi_t^T\right]$ and $\mathbb{E}_\pi\left[\phi_t' \phi_t'^T\right]$ are invertible.*

Multiplying (26a) with ϕ_t' and (26b) with ϕ_t, and taking the expectation of both sides yield

$$\mathbb{E}\left[\phi_t' \phi_t'^T \Gamma'(X) - \phi_t'(x_{t+1}^T \otimes I_n) \operatorname{vec}(X) | x_t, u_t\right] = 0, \tag{28a}$$

$$\mathbb{E}\left[\phi_t \phi_t^T \Gamma(X) - \phi_t \operatorname{vecv}^T(x_{t+1}) \operatorname{vecs}(X) | x_t, u_t\right] = 0. \tag{28b}$$

Taking the expectation of (28) with respect to the invariant probability measure π and using Assumption 3, we have

$$\Gamma'(X) = (\Phi')^\dagger \Xi' \operatorname{vec}(X), \quad \Gamma(X) = \Phi^\dagger \Xi \operatorname{vecs}(X), \tag{29}$$

where

$$\Phi' := \mathbb{E}_\pi\left[\phi_t' \phi_t'^T\right], \quad \Xi' := \mathbb{E}_\pi\left[\phi_t'(x_{t+1}^T \otimes I_n)\right] \tag{30a}$$

$$\Phi := \mathbb{E}_\pi\left[\phi_t \phi_t^T\right], \quad \Xi := \mathbb{E}_\pi\left[\phi_t \operatorname{vecv}^T(x_{t+1})\right] \tag{30b}$$

In addition, we use a finite number of trajectory data to approximate Φ', Ξ', Φ, and Ξ, that is

$$\hat{\Phi}_\tau' := \frac{1}{\tau}\sum_{t=1}^{\tau} \phi_t' \phi_t'^T \quad \hat{\Xi}_\tau' := \frac{1}{\tau}\sum_{t=1}^{\tau} \phi_t'(x_{t+1}^T \otimes I_n), \tag{31a}$$

$$\hat{\Phi}_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau} \phi_t \phi_t^T \quad \hat{\Xi}_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau} \phi_t \operatorname{vecv}^T(x_{t+1}). \tag{31b}$$

Since $(A - B\hat{K}_1)$ is stable, by Birkhoff ergodic Theorem in (Koralov and G. Sinai 2007, Theorem 16.2), the following relations hold *almost surely*

$$\lim_{\tau\to\infty} \hat{\Phi}'_\tau = \Phi', \quad \lim_{\tau\to\infty} \hat{\Xi}'_\tau = \Xi' \tag{32a}$$

$$\lim_{\tau\to\infty} \hat{\Phi}_\tau = \Phi, \quad \lim_{\tau\to\infty} \hat{\Xi}_\tau = \Xi. \tag{32b}$$

Then, by (29), $\Gamma'(X)$ and $\Gamma(X)$ are approximated as follows:

$$\hat{\Gamma}'(X) = (\hat{\Phi}'_\tau)^\dagger \hat{\Xi}'_\tau \operatorname{vec}(X), \; \hat{\Gamma}(X) = \hat{\Phi}^\dagger_\tau \hat{\Xi}_\tau \operatorname{vecs}(X). \tag{33}$$

According to the definitions of Γ and Γ' in (27), and their relations to X in (29), the components of Γ' and Γ can be recovered as

$$\operatorname{vec}(XA) = [(\Phi')^\dagger]_{1,n_1} \Xi' D_n \operatorname{vecs}(X) \tag{34a}$$

$$\operatorname{vec}(XB) = [(\Phi')^\dagger]_{n_1+1,n_2} \Xi' D_n \operatorname{vecs}(X) \tag{34b}$$

$$\operatorname{vecs}(A^T XA) = [(\Phi)^\dagger]_{1,n_3} \Xi \operatorname{vecs}(X) \tag{34c}$$

$$\operatorname{vecs}(B^T XB) = [(\Phi)^\dagger]_{n_3+1,n_4} \Xi \operatorname{vecs}(X) \tag{34d}$$

$$\operatorname{vec}(B^T XA) = [(\Phi)^\dagger]_{n_4+1,n_5} \Xi \operatorname{vecs}(X), \tag{34e}$$

where $n_l (l = 1, \cdots, 5)$ can be determined by the dimensions of the matrices A and B, and D_n is the duplication matrix ($\operatorname{vec}(X) = D_n \operatorname{vecs}(X)$) in Lemma 10.

By vectorizing (17a), we have

$$\begin{aligned}
&\operatorname{vecs}(A^T P_{i,j} A) - D_n^\dagger[(K_i^T \otimes I_n)T_{mn} + I_n \otimes K_i^T] \operatorname{vec}(B^T P_{i,j} A) \\
&+ D_n^\dagger[(L_{i,j}^T D^T \otimes I_n)T_{nn} + I_n \otimes L_{i,j}^T D^T] \operatorname{vec}(P_{i,j} A) \\
&- D_n^\dagger \left[(L_{i,j}^T D^T \otimes K_i^T)T_{nm} + K_i^T \otimes L_{i,j}^T D^T\right] \operatorname{vec}(P_{i,j} B) \\
&+ D_n^\dagger (K_i^T \otimes K_i^T) D_m \operatorname{vecs}(B^T P_{i,j} B) \\
&+ D_n^\dagger (L_{i,j}^T D^T \otimes L_{i,j}^T D^T) D_n \operatorname{vecs}(P_{i,j}) - \operatorname{vecs}(P_{i,j}) \\
&+ \operatorname{vecs}(Q_i - \gamma^2 L_{i,j}^T L_{i,j}) = 0.
\end{aligned} \tag{35}$$

Replacing X in (34) with $P_{i,j}$ and substituting it into (35) yields a linear equation with respect to $P_{i,j}$

$$\begin{aligned}
&\Big\{[(\Phi)^\dagger]_{1,n_3} \Xi - D_n^\dagger[(K_i^T \otimes I_n)T_{mn} + I_n \otimes K_i^T][(\Phi)^\dagger]_{n_4+1,n_5} \Xi \\
&+ D_n^\dagger[(L_{i,j}^T D^T \otimes I_n)T_{nn} + I_n \otimes L_{i,j}^T D^T][(\Phi')^\dagger]_{1,n_1} \Xi' D_n \\
&- D_n^\dagger \left[(L_{i,j}^T D^T \otimes K_i^T)T_{nm} + K_i^T \otimes L_{i,j}^T D^T\right] [(\Phi')^\dagger]_{n_1+1,n_2} \Xi' D_n \\
&+ D_n^\dagger (K_i^T \otimes K_i^T) D_m [(\Phi)^\dagger]_{n_3+1,n_4} \Xi \\
&+ D_n^\dagger (L_{i,j}^T D^T \otimes L_{i,j}^T D^T) D_n - I_{(1+n)n/2} \Big\} \operatorname{vecs}(P_{i,j}) \\
&+ \operatorname{vecs}(Q_i - \gamma^2 L_{i,j}^T L_{i,j}) = 0,
\end{aligned} \tag{36}$$

where T_{mn}, T_{nn}, and T_{nm} are commutation matrices defined in Lemma 10. Consequently, $P_{i,j}$ can be obtained by solving the linear equation (36). The details are shown in Algorithm 2. Since the data matrices $\hat{\Phi}'_\tau$, $\hat{\Xi}'_\tau$, $\hat{\Phi}_\tau$, and $\hat{\Xi}_\tau$ are reused throughout the policy iteration, the proposed algorithm is called off-policy.

Numerical Simulation

We apply Algorithms 1 and 2 to the system studied in (Zhang et al. 2021). The system matrices are

$$A = \begin{bmatrix} 1 & 0 & -5 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 3 & 1 \\ -1 & 0 & 2 \end{bmatrix}, D = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} \tag{37}$$

Algorithm 2: Learning-based Policy Iteration

Initialize $\hat{K}_1 \in \mathcal{W}$, the length of the sampled trajectory τ, and the exploration variance σ_u^2 ;
Collect data from (1a) with exploratory input (24) ;
Construct $\hat{\Phi}'_\tau$, $\hat{\Phi}_\tau$, $\hat{\Xi}'_\tau$, and $\hat{\Xi}_\tau$ defined in (31) ;
for $i \leq \bar{i}$ **do**
  Set $L_{i,j} = 0$;
  $\hat{Q}_i = C^T C + \hat{K}_i^T R \hat{K}_i$;
  for $j \leq \bar{j}$ **do**
    Get $\hat{P}_{i,j}$ by solving (36);
    Get $\widehat{P_{i,j}A}$ and $\widehat{P_{i,j}B}$ by (34a) and (34b).
    Update $\hat{L}_{i,j+1} = (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T (\widehat{P_{i,j}A} - \widehat{P_{i,j}B}K_i)$;
  end
  $\hat{U}_{i,\bar{j}} = \hat{P}_{i,\bar{j}} + \hat{P}_{i,\bar{j}} D(\gamma^2 I_n - D^T \hat{P}_{i,\bar{j}} D)^{-1} D^T \hat{P}_{i,\bar{j}}$;
  Get $\widehat{B^T U_{i,\bar{j}} B}$ and $\widehat{B^T U_{i,\bar{j}} A}$ by (34d) and (34e) ;
  $\hat{K}_{i+1} = (R + \widehat{B^T U_{i,\bar{j}} B})^{-1} \widehat{B^T U_{i,\bar{j}} A}$;
end

The matrices related to the output are $C = [I_3, 0_{3\times3}]^T$ and $E = [0_{3\times3}, I_3]^T$. The $\mathcal{H}_\infty$ norm threshold is $\gamma = 5$. The simulation is implemented on a desktop computer with a CPU Intel i7-9700K CPU @ 3.60GHz. The computer has two 16GB 3200MHz DDR4 RAMs and the numeric computing platform is MATLAB 2020b. $\bar{i} = 20$ and $\bar{j} = 20$.

The robustness of Algorithm 1 in the presence of disturbance during the learning process is validated first. For each outer and inner loop iteration, the entries of the disturbances ΔK_i and $\Delta L_{i,j}$ are samples from a standard Gaussian distribution and then their Frobenius norms are normalized to 0.1. The relative errors of the gain matrix $\hat{K}_i$ and cost matrix $\hat{P}_i$, and the $\mathcal{H}_\infty$ norm of the closed-loop system with $\hat{K}_i$ are shown in Fig. 2. It is seen that with the disturbance at each outer and inner loop iteration, the generated controller and the corresponding cost matrix approach the optimal solution and finally enters a neighborhood of the optimal controller K^* and cost matrix P^*. The $\mathcal{H}_\infty$ norm of the closed-loop system is smaller than the threshold throughout the learning process. These numerical results are consistent with the developed theoretical results in Theorems 4 and 5. The robustness of Algorithm 1 is compared with the natural policy gradient (NPG) algorithm in (Zhang et al. 2021). In Fig. 3, it is shown that in the presence of disturbance, the algorithm diverges and the closed-loop system is unstable. This further shows that Algorithm 1 is more robust to disturbance.

Algorithm 2 is implemented to learn a robust suboptimal controller for system (1a). The length of the sampled trajectory is $\tau = 1000$, i.e. 1000 data are collected in total to train the robust optimal controller. The standard deviation of the exploratory noise is $\sigma_u = 5$. The relative errors of the gain matrix and cost matrix are shown in Fig. 4. The algorithm converges at the 3th iteration. At the 20th iteration, $\|\hat{K}_{20} - K^*\|_F / \|K^*\|_F = 7.37\%$ and

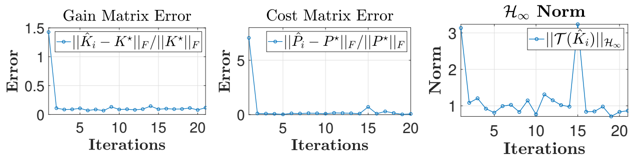

Figure 2: Robustness of Algorithm 1 when $\|\Delta K\|_\infty = 0.1$ and $\|\Delta L_i\|_\infty = 0.1$.

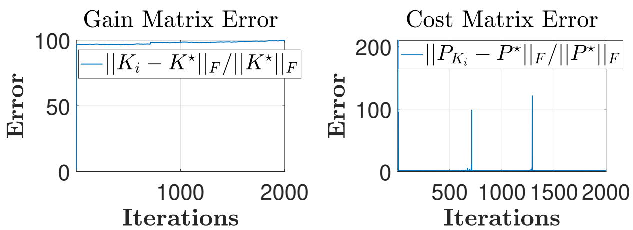

Figure 3: Robustness of the natural policy gradient algorithm in (Zhang et al. 2021) when $\|\Delta K\|_\infty = 0.1$ and $\|\Delta L_i\|_\infty = 0.1$.

$\|\hat{P}_i - P^*\|_F/\|P^*\|_F = 4.29\%$. Therefore, the proposed off-policy RL algorithm can still approximate the optimal solution when the system is disturbed by the additive Gaussian noise. It should be noticed that in (Zhang et al. 2021), around $10^5 \sim 10^6$ trajectories are sampled to approximate the policy gradient at each iteration. Compared with (Zhang et al. 2021), Algorithm 2 is more data-efficient.

Related Work

By adopting the idea of robust control, robust RL is aimed at optimizing the control policy considering the effects of model mismatch (Morimoto and Doya 2005; Donti et al. 2020; Zhang et al. 2021; Zhang, Hu, and Başar 2019; Pinto et al. 2017; Han et al. 2019). In detail, by considering the disturbance as an adversarial player, the robust RL is transferred to a minimax problem, and an actor-critic framework is proposed to find the minimax solution (Morimoto and Doya 2005). Along this line, Donti *et al.* (Donti et al. 2020) adopt the model-based robust control as a guidance for learning a nonlinear neural network control policy. The authors of (Pinto et al. 2017) propose a framework to learn policies for the minimizer and maximizer simultaneously. In (Han et al. 2019), a robust Lyapunov-based actor-critic algorithm is proposed to simultaneously find the Lyapunov function and robust policy. The authors of (Zhang, Hu, and Başar 2019; Zhang et al. 2021) provide the policy gradient algorithm to solve the corresponding minimax optimal control problem. In these works, the robustness of the RL algorithms to noise in the learning process is not studied. In addition, compared with (Zhang et al. 2021), the proposed algorithm is off-policy, and as demonstrated, it is more data-efficient.

Via solving the zero-sum differential game, the robustness of the obtained control policy to model mismatch can be guaranteed, and recently, many algorithms are proposed

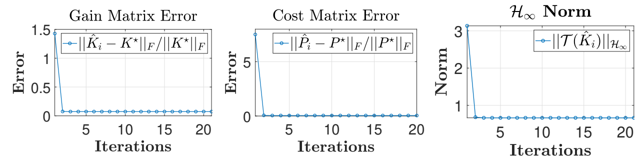

Figure 4: Using Algorithm 2, the solutions of each iteration approach the optimal solution, and the $\mathcal{H}_\infty$ norm is smaller than the threshold.

to solve the differential game. For example, the authors of (Abu-Khalaf, Lewis, and Huang 2006) propose the model-based policy iteration algorithm to solve the Hamilton-Jacobi-Isaacs equation for continuous-time nonlinear systems. For continuous-time linear time-invariant systems, the iterative algorithms are proposed to solve the corresponding GARE in (Lanzon et al. 2008; Wu and Luo 2013). By policy optimization, gradient decent algorithms are proposed in (Bu, Ratliff, and Mesbahi 2019). Based on these model-based iterative algorithms, learning-based approaches are proposed to solve the differential game (Vrabie and Lewis 2011; Luo, Yang, and Liu 2021; Al-Tamimi, Lewis, and Abu-Khalaf 2007; Vrabie, Lewis, and Vamvoudakis 2012; Wei et al. 2017). In these paper, the system is deterministic, and the inputs of two players are measurable. In contrast, Algorithm 2 is designed for a stochastic system and only the input of the player u is available for the algorithm.

Recently, the learning-based linear quadratic regulator (LQR) problem is resurgent in the machine learning field. The developed methods to address the learning-based LQR problem fall into three categories. The first one is a model-based RL method. It tries to explicitly estimate the model and then get the control policy by model-based controller design methods (Dean et al. 2018; Mhammedi et al. 2020; Cassel, Cohen, and Koren 2020; Umenberger and Schön 2020). The second category for solving learning-based LQR is model-free RL. Adopting the value function, it directly generates the approximation of the optimal controller using the data from the system (Tu and Recht 2019; Jiang, Bian, and Gao 2020; Krauth, Tu, and Recht 2019). The third category uses the gradient decent method to update the control policy such that the cost can be minimized iteratively (Fazel et al. 2018; Bu et al. 2019; Bu, Mesbahi, and Mesbahi 2020; Mohammadi et al. 2022; Li et al. 2021). However, LQR cannot guarantee the $\mathcal{H}_\infty$ constraint and consequently, it may be unstable with certain model mismatch.

Conclusion

In this paper, we have proposed a novel dual-loop policy iteration algorithm for risk-sensitive linear quadratic Gaussian control. The convergence and robustness properties of the algorithm is rigorously analyzed. It is proved that the iterative algorithm possesses the property of small-disturbance input-to-state stability, that is, starting from an initial feasible controller, the solutions of the proposed policy iteration algorithm ultimately enters a neighbourhood of the optimal solution, given that the disturbance is small. Based on these model-based theoretical results, when the accurate system knowledge is unavailable, we have also proposed a novel off-policy policy iteration RL algorithm to learn from data robust optimal controllers. Numerical examples are provided and the efficacy of the proposed methods are demonstrated.

References

Abu-Khalaf, M.; Lewis, F. L.; and Huang, J. 2006. Policy iterations on the Hamilton–Jacobi–Isaacs equation for H_∞ state feedback control with input saturation. *IEEE Transactions on Automatic Control*, 51(12): 1989–1995.

Al-Tamimi, A.; Lewis, F. L.; and Abu-Khalaf, M. 2007. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H_∞ control. *Automatica*, 43(3): 473–481.

Apkarian, P.; Noll, D.; and Rondepierre, A. 2008. Mixed $\mathcal{H}\_2/\mathcal{H}\_\infty$ Control via Nonsmooth Optimization. *SIAM Journal on Control and Optimization*, 47(3): 1516–1546.

Åström, K.; and Wittenmark, B. 2008. *Adaptive Control*. Dover Books on Electrical Engineering. Dover Publications. ISBN 9780486462783.

Başar, T. 2008. *H_∞-optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer.

Bu, J.; Mesbahi, A.; Fazel, M.; and Mesbahi, M. 2019. LQR through the lens of first order methods: Discrete-time case. *arXiv e-prints*, arXiv:1907.08921.

Bu, J.; Mesbahi, A.; and Mesbahi, M. 2020. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv e-prints*, arXiv:2006.09178.

Bu, J.; Ratliff, L. J.; and Mesbahi, M. 2019. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv e-prints*, arXiv:1911.04672.

Cassel, A.; Cohen, A.; and Koren, T. 2020. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv e-prints*, arXiv:2002.08095.

Chen, C.-T. 1999. *Linear System Theory and Design*. Oxford University Press, 3rd edition edition.

Cui, L.; Wang, S.; Zhang, J.; Zhang, D.; Lai, J.; Zheng, Y.; Zhang, Z.; and Jiang, Z. P. 2021. Learning-based balance control of wheel-legged robots. *IEEE Robotics and Automation Letters*, 6(4): 7667–7674.

Dean, S.; Mania, H.; Matni, N.; Recht, B.; and Tu, S. 2018. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Donti, P. L.; Roderick, M.; Fazlyab, M.; and Zico Kolter, J. 2020. Enforcing robust control guarantees within neural network policies. *arXiv e-prints*, arXiv:2011.08105.

Doyle, J.; Glover, K.; Khargonekar, P.; and Francis, B. 1989. State-space solutions to standard H/sub 2/ and H/sub infinity / control problems. *IEEE Transactions on Automatic Control*, 34(8): 831–847.

Fazel, M.; Ge, R.; Kakade, S.; and Mesbahi, M. 2018. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1467–1476. PMLR.

Hahn, W. 1967. *Stability of Motion*. Springer Berlin, Heidelberg.

Han, M.; Tian, Y.; Zhang, L.; Wang, J.; and Pan, W. 2019. H_∞ model-free reinforcement learning with robust stability guarantee. *arXiv e-prints*, arXiv:1911.02875.

Hespanha, J. P. 2018. *Linear Systems Theory*. Princeton, New Jersey: Princeton Press. ISBN13: 9780691179575.

Hewer, G. 1971. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4): 382–384.

Jacobson, D. 1973. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2): 124–131.

Jiang, Z. P.; Bian, T.; and Gao, W. 2020. Learning-based control: A tutorial and some recent results. *Foundations and Trends® in Systems and Control*, 8(3): 176–284.

Jiang, Z. P.; and Liu, T. 2018. Small-gain theory for stability and control of dynamical networks: A survey. *Annual Reviews in Control*, 46: 58–79.

Jiang, Z. P.; and Wang, Y. 2001. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6): 857–869.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274.

Koralov, L.; and G. Sinai, Y. 2007. *Theory of Probability and Random Processes*. Springer Berlin, Heidelberg, 2nd ed. edition.

Krauth, K.; Tu, S.; and Recht, B. 2019. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lanzon, A.; Feng, Y.; Anderson, B. D. O.; and Rotkowitz, M. 2008. Computing the positive stabilizing solution to algebraic Riccati equations with an indefinite quadratic term via a recursive method. *IEEE Transactions on Automatic Control*, 53(10): 2280–2291.

Li, T.-Y.; and Gajic, Z. 1995. Lyapunov Iterations for Solving Coupled Algebraic Riccati Equations of Nash Differential Games and Algebraic Riccati Equations of Zero-Sum Games. In Olsder, G. J., ed., *New Trends in Dynamic Games and Applications*, 333–351. Boston, MA: Birkhäuser Boston. ISBN 978-1-4612-4274-1.

Li, Y.; Tang, Y.; Zhang, R.; and Li, N. 2021. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 1–1.

Luo, B.; Yang, Y.; and Liu, D. 2021. Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems. *IEEE Transactions on Cybernetics*, 51(7): 3630–3640.

Magnus, J. R.; and Neudecker, H. 1985. Matrix differential calculus with applications to simple, Hadamard, and Kronecker products. *Journal of Mathematical Psychology*, 29(4): 474–492.

Magnus, J. R.; and Neudecker, H. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Mankowitz, D. J.; Levine, N.; Jeong, R.; Shi, Y.; Kay, J.; Abdolmaleki, A.; Springenberg, J. T.; Mann, T.; Hester, T.; and Riedmiller, M. 2019. Robust reinforcement learning for continuous control with model misspecification. *arXiv e-prints*, arXiv:1906.07516.

Mhammedi, Z.; Foster, D. J.; Simchowitz, M.; Misra, D.; Sun, W.; Krishnamurthy, A.; Rakhlin, A.; and Langford, J. 2020. Learning the linear quadratic regulator from nonlinear observations. In *Advances in Neural Information Processing Systems*, volume 33, 14532–14543. Curran Associates, Inc.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529—533.

Mohammadi, H.; Zare, A.; Soltanolkotabi, M.; and Jovanović, M. R. 2022. Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5): 2435–2450.

Mori, T. 1988. Comments on "A matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation" by J. M. Saniuk and I.B. Rhodes. *IEEE Transactions on Automatic Control*, 33(11): 1088–.

Morimoto, J.; and Doya, K. 2005. Robust reinforcement learning. *Neural Computation*, 17(2): 335–359.

Mustafa, D.; and Bernstein, D. S. 1991. LQG cost bounds in discrete-time H2/H-infinity control. *Transactions of the Institute of Measurement and Control*, 13(5): 269–275.

Pang, B.; Bian, T.; and Jiang, Z. P. 2022. Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, 67(1): 504–511.

Pang, B.; and Jiang, Z. P. 2021. Robust reinforcement learning: a case study in linear quadratic regulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10): 9303–9311.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2817–2826. PMLR.

Roulet, V.; Fazel, M.; Srinivasa, S.; and Harchaoui, Z. 2020. On the convergence of the iterative linear exponential quadratic Gaussian algorithm to stationary points. In *2020 American Control Conference (ACC)*, 132–137.

Shani, G.; Heckerman, D.; and Brafman, R. I. 2005. An MDP-based recommender system. *Journal of Machine Learning Research*, 6(43): 1265–1295.

Sontag, E. 2008. *Input to state stability: Basic concepts and results*, 163–220. Lecture Notes in Mathematics. Germany: Springer Verlag. ISBN 9783540776444.

Speyer, J.; Deyst, J.; and Jacobson, D. 1974. Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria. *IEEE Transactions on Automatic Control*, 19(4): 358–366.

Tu, S.; and Recht, B. 2019. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, 3036–3083. PMLR.

Umenberger, J.; and Schön, T. B. 2020. Optimistic robust linear quadratic dual control. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, 550–560. PMLR.

Vrabie, D.; and Lewis, F. 2011. Adaptive dynamic programming for online solution of a zero-sum differential game. *J. Contr. Theory Appl.*, 9: 353–360.

Vrabie, D.; Lewis, F. L.; and Vamvoudakis, K. G. 2012. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. IET Press.

Wei, Q.; Liu, D.; Lin, Q.; and Song, R. 2017. Adaptive dynamic programming for discrete-time zero-sum games. *IEEE Transactions on Neural Networks and Learning Systems*, PP: 1–13.

Whittle, P. 1981. Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, 13(4): 764–777.

Wu, H.-N.; and Luo, B. 2013. Simultaneous policy update algorithm for learning the solution of linear continuous-time $\mathcal{H}_\infty$ state feedback control. *Information Sciences*, 222: 472–485.

Zames, G. 1966. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity. *IEEE Transactions on Automatic Control*, 11(2): 228–238.

Zhang, K.; Hu, B.; and Başar, T. 2019. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence. *arXiv e-prints*, arXiv:1910.09496.

Zhang, K.; Yang, Z.; and Başar, T. 2019. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. *arXiv e-prints*, arXiv:1906.00729.

Zhang, K.; Zhang, X.; Hu, B.; and Basar, T. 2021. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. In *Advances in Neural Information Processing Systems*, volume 34, 2949–2964. Curran Associates, Inc.

Zhou, K.; Doyle, J. C.; and Glover, K. 1996. *Robust and Optimal Control*. Princeton, New Jersey: Prentice Hall.

Appendix A: Useful Auxiliary Results

In this section, some fundamental lemmas from discrete-time linear systems and robust control theory are introduced to assist in the development of the proposed method.

Lemma 1. *Consider a stable matrix $A \in \mathbb{R}^{n\times n}$ satisfying*

$$A^T PA - P + Q = 0. \tag{A.1}$$

The following statements hold

1. $P = \sum_{t=0}^{\infty}(A^T)^t QA^t$ *is the unique solution to* (A.1) *and* $P = P^T \succeq 0$ *if* $Q \succeq 0$*;*
2. *If* $P' = P'^T \succeq 0$ *satisfies* $A^T P' A - P' + Q' = 0$ *with* $Q' \succeq Q$*, then* $P' \succeq P$*.*

Proof. The first statement follows directly from (Chen 1999, Theorem 5.D6). The second statement follows from the fact that $P' - P = \sum_{t=0}^{\infty}(A^T)^t(Q' - Q)A^t \succeq 0$. □

Lemma 2. *If there exists $P = P^T \succ 0$ amd $Q = Q^T \succ 0$ such that* (A.1) *holds, then A is stable.*

Proof. See (Hespanha 2018, Theorem 8.4) □

Lemma 3. *For any $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times m}$, we have*

$$det(I_m + XY) = det(I_n + Y X). \tag{A.2}$$

Proof. Using the fact that $\det(AB) = \det(A)\det(B)$ for any matrices $A, B \in R^{(m+n)\times(m+n)}$, we obtain

$$\det \begin{bmatrix} I_n & -Y \\ X & I_m \end{bmatrix} \det \begin{bmatrix} I_n & Y \\ 0 & I_m \end{bmatrix} = \det \begin{bmatrix} I_n & 0 \\ X & I_m + XY \end{bmatrix} = \det(I_m + XY), \tag{A.3a}$$

$$\det \begin{bmatrix} I_n & Y \\ 0 & I_m \end{bmatrix} \det \begin{bmatrix} I_n & -Y \\ X & I_m \end{bmatrix} = \det \begin{bmatrix} I_n + YX & 0 \\ X & I_m \end{bmatrix} = \det(I_n + YX). \tag{A.3b}$$

Therefore, (A.2) holds. □

Lemma 4. *For any $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times m}$ such that $(I_m + XY)$ and $(I_n + Y X)$ are invertible, we have*

$$(I_m + XY)^{-1}X = X(I_n + Y X)^{-1}. \tag{A.4}$$

Proof.

$$(I_m + XY)^{-1}X = (I_m + XY)^{-1}X(I_n + Y X)(I_n + Y X)^{-1} = X(I_n + Y X)^{-1}. \tag{A.5}$$

□

Lemma 5 (Real Bounded Lemma)**.** *Consider a stabilizing matrix K and the transfer function $\mathcal{T}(K)$, the following statements are equivalent:*

1. $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} \leq \gamma$*;*
2. *The following algebraic Riccati equation*

$$(A - BK)^T U_K(A - BK) - P + Q + K^T RK = 0 \tag{A.6a}$$

$$U_K = P + PD(\gamma^2 I_q - D^T PD)^{-1}D^T P. \tag{A.6b}$$

admits a unique stabilizing solution $P_K \succ 0$ such that i) $I_q - \gamma^{-2}D^T P_K D \succ 0$; ii) $[A - BK + D(\gamma^2 I_q - D^T P_K D)^{-1}D^T P_K(A - BK)]$ is stable;

3. *There exists $P \succ 0$ such that*

$$I_q - \gamma^{-2}D^T PD \succ 0, \quad (A - BK)^T U_K(A - BK) - P + Q + K^T RK \preceq 0. \tag{A.7}$$

Proof. See (Zhang, Hu, and Başar 2019, Lemma 2.7). □

Lemma 6. *Let X, $Y \in \mathbb{S}^n$ and $Y \succeq 0$. Then,*

$$\lambda_{min}(X)\,\mathrm{Tr}(Y) \leq \mathrm{Tr}(XY) \leq \lambda_{max}(X)\,\mathrm{Tr}(Y).$$

Proof. The proof can be found in (Mori 1988). □

Lemma 7. *For any positive semi-definite matrix $P \in \mathbb{S}^n$, $\|P\|_F \leq \mathrm{Tr}(P) \leq \sqrt{n}\|P\|_F$ and $\|P\| \leq \mathrm{Tr}(P) \leq n\|P\|$.*

Proof. Let $\sigma_1 \geq \cdots \geq \sigma_n$ denote the descending sequence of P's singular values. Then, $\|P\|_F = \sqrt{\sum_{i=1}^n \sigma_i^2}$, $\mathrm{Tr}(P) = \sum_{i=1}^n \sigma_i$, and $\|P\| = \sigma_1(P)$. Hence, $\|P\| \leq \mathrm{Tr}(P) \leq n\|P\|$. $\|P\|_F \leq \mathrm{Tr}(P) \leq \sqrt{n}\|P\|_F$ can be proven by the fact that $\sum_{i=1}^n \sigma_i^2 \leq (\sum_{i=1}^n \sigma_i)^2 \leq n \sum_{i=1}^n \sigma_i^2$. □

Lemma 8. *For $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times p}$, $\|XY\|_F \leq \|X\|\|Y\|_F$.*

Proof. Let $Y = [y_1, \cdots, y_p]$. Then, it follows that $XY = [Xy_1, \cdots, Xy_p]$. It implies that $\|XY\|_F^2 = \sum_{i=1}^p |Xy_i|^2 \leq \|X\|^2 \sum_{i=1}^p |y_i|^2 = \|X\|^2\|Y\|_F^2$. □

Lemma 9. *Let $U : \mathcal{S} \to \mathbb{R}^{m\times r}$ and $V : \mathcal{S} \to \mathbb{R}^{r\times p}$ be two matrix-valued functions defined and differentiable on an open set $\mathcal{S} \subseteq \mathbb{R}^{n\times q}$. Then the simple product UV is differentiable on $\mathcal{S}$ and the Jacobian matrix is the $mp \times nq$ matrix*

$$\frac{\partial \operatorname{vec}(UV)}{\partial \operatorname{vec}(X)} = (V^T \otimes I_m)\frac{\partial \operatorname{vec}(U)}{\partial \operatorname{vec}(X)} + (I_p \otimes U)\frac{\partial \operatorname{vec}(V)}{\partial \operatorname{vec}(X)}. \tag{A.8}$$

Proof. The proof can be found in (Magnus and Neudecker 1985, Theorem 9). □

Lemma 10. *For $X \in \mathbb{S}^n$, there exists a unique matrix $D_n \in \mathbb{R}^{n^2\times\frac{1}{2}n(n+1)}$ with full column rank, such that*

$$\operatorname{vec}(X) = D_n \operatorname{vecs}(X), \quad \operatorname{vecs}(X) = D_n^\dagger \operatorname{vec}(X). \tag{A.9}$$

For $Y \in \mathbb{R}^{m\times n}$, there exists a unique nonsingular matrix $T_{mn} \in \mathbb{R}^{mn\times mn}$ such that

$$T_{mn} \operatorname{vec}(A) = \operatorname{vec}(A^T). \tag{A.10}$$

Proof. D_n is called duplication matrix and T_{mn} is called commutation matrix. See (Magnus and Neudecker 2007, pp. 54 and pp. 56). □

Appendix B: Proof of Theorem 1

Lemma 11. *For any $K, K' \in \mathcal{W}$, we have*

$$(I_n - \gamma^{-2}P_{K'}DD^T)^{-1}(P_K - \gamma^{-2}P_{K'}DD^TP_{K'})(I_n - \gamma^{-2}DD^TP_{K'})^{-1} \preceq U_K, \tag{B.1}$$

where P_K is the stabilizing solution to $(A-BK)^TU_K(A-BK) - P_K + Q + K^TRK = 0$ and $U_K = P_K + P_KD(\gamma^2 I_q - D^TP_KD)^{-1}D^TP_K$.

Proof. See (Zhang, Hu, and Başar 2019, Lemma B.1). □

The following lemma shows that the cost matrix P_i generated by the outer-loop iteration (10) is monotonically decreasing and all the updated feedback gains are admissible given an initial admissible feedback gain.

Lemma 12. *Under Assumptions 1 and 2, if $K_1 \in \mathcal{W}$, then for any $i \in \mathbb{Z}_+$,*

1. *$K_i \in \mathcal{W}$;*
2. *$P_1 \succeq \cdots \succeq P_i \succeq P_{i+1} \succeq \cdots \succeq P^*$;*
3. *$\lim_{i\to\infty}\|K_i - K^*\|_F = 0$ and $\lim_{i\to\infty}\|P_i - P^*\|_F = 0$.*

Proof. The statements 1) and 3) are shown in (Zhang, Hu, and Başar 2019, Theorem 4.3 and Theorem 4.6). To make the paper self-contained, we apply the method from (Hewer 1971) to prove the statements.

We will prove the first statement by induction. When $i = 1$, $K_1 \in \mathcal{W}$ as it is an initial admissible controller. Therefore, the statement 1) holds for the case of $i = 1$. Now assume $K_i \in \mathcal{W}$ for some $i \geq 1$. By Lemma 5, $P_i \succ 0$ and $I_q - \gamma^{-2}D^TP_iD \succ 0$. We can rewritten (10a) as

$$\begin{aligned} &A_{i+1}^TU_iA_{i+1} + (K_{i+1}-K_i)^TB^TU_iA + A^TU_iB(K_{i+1}-K_i) + K_i^TB^TU_iBK_i - K_{i+1}^TB^TU_iBK_{i+1} \\ &\quad - P_i + Q + K_i^TRK_i = 0. \end{aligned} \tag{B.2}$$

Since $(R + B^TU_iB)K_{i+1} = B^TU_iA$ from (10b), by completing the squares, we have

$$A_{i+1}^TU_iA_{i+1} - P_i + Q + K_{i+1}^TRK_{i+1} + (K_{i+1}-K_i)^T(R + B^TU_iB)(K_{i+1}-K_i) = 0. \tag{B.3}$$

Since the last term in (B.3) are positive semi-definite, by Lemma 5, we have $K_{i+1} \in \mathcal{W}$. *A fortiori*, $K_i \in \mathcal{W}$ for any $i \in \mathbb{Z}_+$.

Writing out (10a) for the $(i+1)$th iteration, subtracting it from (B.3), we can obtain that

$$A_{i+1}^T(U_i - U_{i+1})A_{i+1} - (P_i - P_{i+1}) + (K_{i+1}-K_i)^T(R + B^TU_iB)(K_{i+1}-K_i) = 0. \tag{B.4}$$

From (11), the expression of U_{i+1} is derived as

$$\begin{aligned}
U_{i+1} &= [I_n + P_{i+1}D(\gamma^2 I_q - D^T P_{i+1} D)^{-1} D^T] P_{i+1} = [I_n + P_{i+1} D D^T (\gamma^2 I_n - P_{i+1} D D^T)^{-1}] P_{i+1} \\
&= (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} P_{i+1} = (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} P_{i+1} (I_n - \gamma^{-2} D D^T P_{i+1})(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} \\
&= (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} (P_i - \gamma^{-2} P_{i+1} D D^T P_{i+1})(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} \\
&\quad - (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} (P_i - P_{i+1})(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} \\
&\preceq U_i - (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} (P_i - P_{i+1})(I_n - \gamma^{-2} D D^T P_{i+1})^{-1},
\end{aligned} \tag{B.5}$$

where the second equation is from Lemma 4 and the last inequality is derived by Lemma 11. Combining (B.4) and (B.5), we have

$$\begin{aligned}
&A_{i+1}^T (I_n - \gamma^{-2} P_{i+1} D D^T)^{-1} (P_i - P_{i+1})(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} A_{i+1} - (P_i - P_{i+1}) \\
&\quad + (K_{i+1} - K_i)^T (R + B^T U_i B)(K_{i+1} - K_i) \preceq 0.
\end{aligned} \tag{B.6}$$

Considering the expression of $L_{i+1,*}$ in (15), $(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} A_{i+1}$ can be rewritten as

$$\begin{aligned}
(I_n - \gamma^{-2} D D^T P_{i+1})^{-1} A_{i+1} &= \left[I_n + \gamma^{-2} D D^T P_{i+1} (I_n - \gamma^{-2} D D^T P_{i+1})^{-1} \right] A_{i+1} \\
&= A_{i+1} + \gamma^{-2} D (I_q - \gamma^{-2} D^T P_{i+1} D)^{-1} D^T P_{i+1} A_{i+1} \\
&= A_{i+1} + D L_{i+1,*} = A_{i+1,*},
\end{aligned} \tag{B.7}$$

where the second equality comes from Lemma 4. As a consequence, (B.6) can be rewritten as

$$A_{i+1,*}^T (P_i - P_{i+1}) A_{i+1,*} - (P_i - P_{i+1}) + (K_{i+1} - K_i)^T (R + B^T U_i B)(K_{i+1} - K_i) \preceq 0. \tag{B.8}$$

Since $K_{i+1} \in \mathcal{W}$, by Lemma 5, $A_{i+1,*}$ is stable. Therefore, via Lemma 1, we have $P_i - P_{i+1} \succeq 0$. Hence, $\{P_i\}_{i=1}^{\infty}$ is monotonically decreasing.

As $\{P_i\}_{i=1}^{\infty}$ is monotonically decreasing and lower bounded by the null matrix, there exists $P_\infty \succeq 0$ such that P_i converges to P_∞. As $K_\infty = (R + B^T U_\infty B)^{-1} B^T U_\infty A$, plugging it into (10a), it is seen that P_∞ satisfies GARE (4). Since (4) has a unique positive definite solution P^*, we have $P_\infty = P^*$. Furthermore, comparing (3) and (10b), K_i converges to K^*. ☐

The following lemma gives the expression of the difference between $K \in \mathcal{W}$ and the updated controller K'.

Lemma 13. *For any $K \in \mathcal{W}$ and $K' := (R + B^T U_K B)^{-1} B^T U_K A$, we have*

$$K' - K^* = R^{-1} B^T (\Lambda_K)^{-T} (P_K - P^*) A^*, \tag{B.9}$$

where

$$\Lambda_K := I_n + B R^{-1} B^T P_K - \gamma^{-2} D D^T P_K \tag{B.10}$$

Proof. By Lemma 4 and the expression of U_K in (A.6b), we have

$$\begin{aligned}
(R + B^T U_K B)^{-1} B^T U_K A &= R^{-1} (I_m + B^T U_K B R^{-1})^{-1} B^T U_K A \\
&= R^{-1} B^T U_K (I_n + B R^{-1} B^T U_K)^{-1} A \\
&= R^{-1} B^T U_K [I_n + B R^{-1} B^T P_K + B R^{-1} B^T P_K D (\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K]^{-1} A \\
&= R^{-1} B^T U_K \{ I_n + B R^{-1} B^T P_K [I_n + D D^T P_K (\gamma^2 I_n - D D^T P_K)^{-1}] \}^{-1} A \\
&= R^{-1} B^T U_K [I_n + B R^{-1} B^T P_K (I_n - \gamma^{-2} D D^T P_K)^{-1}]^{-1} A \\
&= R^{-1} B^T U_K (I_n - \gamma^{-2} D D^T P_K)(I_n + B R^{-1} B^T P_K - \gamma^{-2} D D^T P_K)^{-1} A \\
&= R^{-1} B^T P_K [I_n + D (\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K](I_n - \gamma^{-2} D D^T P_K) \Lambda_K^{-1} A \\
&= R^{-1} B^T P_K [I_n + \gamma^{-2} D D^T P_K (I_n - \gamma^{-2} D D^T P_K)^{-1}](I_n - \gamma^{-2} D D^T P_K) \Lambda_K^{-1} A = R^{-1} B^T P_K \Lambda_K^{-1} A.
\end{aligned} \tag{B.11}$$

By the expressions of K' and K^* in (3), we have

$$\begin{aligned}
K' - K^* &= R^{-1} B^T P_K \Lambda_K^{-1} A - R^{-1} B^T P^* (\Lambda^*)^{-1} A \\
&= R^{-1} B^T (P_K - P^*)(\Lambda^*)^{-1} A + R^{-1} B^T P_K \Lambda_K^{-1} A - R^{-1} B^T P_K (\Lambda^*)^{-1} A \\
&= R^{-1} B^T (P_K - P^*)(\Lambda^*)^{-1} A + R^{-1} B^T P_K \left[\Lambda_K^{-1} \Lambda^* (\Lambda^*)^{-1} - \Lambda_K^{-1} \Lambda_K (\Lambda^*)^{-1} \right] A \\
&= R^{-1} B^T (P_K - P^*)(\Lambda^*)^{-1} A - R^{-1} B^T P_K \Lambda_K^{-1} (B R^{-1} B^T - \gamma^{-2} D D^T)(P_K - P^*)(\Lambda^*)^{-1} A \\
&= R^{-1} B^T [I_n - P_K \Lambda_K^{-1} (B R^{-1} B^T - \gamma^{-2} D D^T)](P_K - P^*) A^* \\
&= R^{-1} B^T [I_n - (I_n + P_K B R^{-1} B^T - \gamma^{-2} P_K D D^T)^{-1} P_K (B R^{-1} B^T - \gamma^{-2} D D^T)](P_K - P^*) A^* \\
&= R^{-1} B^T (I_n + P_K B R^{-1} B^T - \gamma^{-2} P_K D D^T)^{-1} (P_K - P^*) A^* = R^{-1} B^T (\Lambda_K)^{-T} (P_K - P^*) A^*.
\end{aligned} \tag{B.12}$$

It is noticed that the fifth equation comes from the equality

$$\begin{aligned}A^* &= [I_n + D(\gamma^2 I_q - D^T P^* D)^{-1} D^T P^*](A - BK^*) = (I_n - \gamma^{-2} DD^T P^*)^{-1}(A - BK^*)\\&= (I_n - \gamma^{-2} DD^T P^*)^{-1}[I_n - B(R + B^T U^* B)^{-1} B^T U^*]A\\&= (I_n - \gamma^{-2} DD^T P^*)^{-1}(I_n + BR^{-1} B^T U^*)^{-1} A\\&= (I_n - \gamma^{-2} DD^T P^*)^{-1}[I_n + BR^{-1} B^T P^*(I_n - \gamma^{-2} DD^T P^*)^{-1}]^{-1} A = (\Lambda^*)^{-1} A.\end{aligned} \tag{B.13}$$

□

The following lemma gives the expression of the difference between U_K and U^*.

Lemma 14. *For any $K \in \mathcal{W}$, $(U_K - U^*)$ satisfies*

$$\begin{aligned}U_K - U^* &= (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}\\&\quad + (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - P^*)D(\gamma^2 I_q - D^T P_K D)^{-1} D^T (P_K - P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}\end{aligned} \tag{B.14}$$

Proof. By recalling the expression of U^* in (4), we can further derive it as

$$\begin{aligned}U^* &= [I_n + P^* D(\gamma^2 I_q - D^T P^* D)^{-1} D^T] P^* = [I_n + P^* DD^T(\gamma^2 I_n - P^* DD^T)^{-1}] P^*\\&= (I_n - \gamma^{-2} P^* DD^T)^{-1} P^* = (I_n - \gamma^{-2} P^* DD^T)^{-1} P^* (I_n - \gamma^{-2} DD^T P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}\\&= (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - \gamma^{-2} P^* DD^T P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}\\&\quad - (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}.\end{aligned} \tag{B.15}$$

To simplify the notation, define S_K as

$$S_K := U_K - (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - \gamma^{-2} P^* DD^T P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}. \tag{B.16}$$

Then, $(I_n - \gamma^{-2} P^* DD^T) S_K (I_n - \gamma^{-2} DD^T P^*)$ is

$$\begin{aligned}(I_n - \gamma^{-2} P^* DD^T) S_K (I_n - \gamma^{-2} DD^T P^*) &= (I_n - \gamma^{-2} P^* DD^T) U_K (I_n - \gamma^{-2} DD^T P^*) - (P_K - \gamma^{-2} P^* DD^T P^*)\\&= U_K - P_K - \gamma^{-2} P^* DD^T U_K - \gamma^{-2} U_K DD^T P^* + \gamma^{-4} P^* DD^T U_K DD^T P^* + \gamma^{-2} P^* DD^T P^*.\end{aligned} \tag{B.17}$$

Following the expression of U_K in (A.6b) and Lemma 4, we have

$$\gamma^{-2} P^* DD^T U_K = \gamma^{-2} P^* DD^T (I_n - \gamma^{-2} P_K DD^T)^{-1} P_K = \gamma^{-2} P^* D (I_q - \gamma^{-2} D^T P_K D)^{-1} D^T P_K. \tag{B.18}$$

Since $U_K - P_K = P_K D(\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K$, plugging (B.18) into (B.17) and completing the squares yield

$$\begin{aligned}(I_n - \gamma^{-2} P^* DD^T) S_K (I_n - \gamma^{-2} DD^T P^*) &= (P_K - P^*) D(\gamma^2 I_q - D^T P_K D)^{-1} D^T (P_K - P^*)\\&\quad - P^* D(\gamma^2 I_q - D^T P_K D)^{-1} D^T P^* + \gamma^{-4} P^* DD^T U_K DD^T P^* + \gamma^{-2} P^* DD^T P^*.\end{aligned} \tag{B.19}$$

By matrix inversion lemma,

$$(\gamma^2 I_q - D^T P_K D)^{-1} = \gamma^{-2} I_q + \gamma^{-4} D^T (I_n - \gamma^{-2} P_K DD^T)^{-1} P_K D = \gamma^{-2} I_q + \gamma^{-4} D^T U_K D, \tag{B.20}$$

where the last equality is from $U_K = (I_n - \gamma^{-2} P_K DD^T)^{-1} P_K$. Plugging (B.20) into (B.19), we have

$$S_K = (I_n - \gamma^{-2} P^* DD^T)^{-1}(P_K - P^*) D(\gamma^2 I_q - D^T P_K D)^{-1} D^T (P_K - P^*)(I_n - \gamma^{-2} DD^T P^*)^{-1}. \tag{B.21}$$

Therefore, (B.14) is obtained by (B.15) and (B.21). □

The following lemma plays a pivotal role in the proof of Theorem 1, and it states that for any $K \in \mathcal{W}$, the distance between the cost matrix P_K and the optimal cost matrix P^* is linearly upperbounded by $\|(K' - K)^T (R + B^T U_K B)(K' - K)\|_F$.

Lemma 15. *For any $K \in \mathcal{W}$, let $K' := (R + B^T U_K B)^{-1} B^T U_K A$, and $E_K := (K' - K)^T (R + B^T U_K B)(K' - K)$. Then, there exists $a(K) > 0$, such that*

$$\|P_K - P^*\|_F \leq a(K) \|E_K\|_F. \tag{B.22}$$

Proof. From (A.6), it is seen that

$$\begin{aligned}&(A - BK^*)^T U_K (A - BK^*) - P_K + Q + K^T R K + A^T U_K B(K^* - K)\\&+ (K^* - K)^T B^T U_K A + K^T B^T U_K B K - (K^*)^T B^T U_K B K^* = 0.\end{aligned} \tag{B.23}$$

Subtracting (4a) from (B.23), and considering $B^TU_KA = (R + B^TU_KB)K'$, we have

$$
\begin{aligned}
&(A - BK^*)^T(U_K - U^*)(A - BK^*) - (P_K - P^*) + K^T(R + B^TU_KB)K \\
&(K')^T(R + B^TU_KB)(K^* - K) + (K^* - K)^T(R + B^TU_KB)K' - (K^*)^T(R + B^TU_KB)K^* = 0.
\end{aligned} \tag{B.24}
$$

Competing the squares in (B.24) yields

$$
(A - BK^*)^T(U_K - U^*)(A - BK^*) - (P_K - P^*) + E_K - (K' - K^*)^T(R + B^TU_KB)(K' - K^*) = 0. \tag{B.25}
$$

It follows from $(I_n - \gamma^{-2}DD^TP^*)^{-1}(A - BK^*) = (A - BK^* + DL^*) = A^*$ and Lemma 14 that

$$
\begin{aligned}
&(A^*)^T(P_K - P^*)(A^*) - (P_K - P^*) + E_K - (K' - K^*)^T(R + B^TU_KB)(K' - K^*) \\
&+ (A^*)^T(P_K - P^*)^TD(\gamma^2I_q - D^TP_KD)^{-1}D^T(P_K - P^*)A^* = 0.
\end{aligned} \tag{B.26}
$$

To simplify notation, let $\Delta P_K := P_K - P^*$. By Lemma 13, we have

$$
\begin{aligned}
&(K' - K^*)^T(R + B^TU_KB)(K' - K^*) = (A^*)^T\Delta P_K\Lambda_K^{-1}BR^{-1}(R + B^TU_KB)R^{-1}B^T(\Lambda_K)^{-T}\Delta P_KA^* \\
&= (A^*)^T\Delta P_K\Lambda_K^{-1}BR^{-1}B^T(I_n + U_KBR^{-1}B^T)(\Lambda_K)^{-T}\Delta P_KA^* \\
&= (A^*)^T\Delta P_K\Lambda_K^{-1}BR^{-1}B^T[I_n + (I_n - \gamma^{-2}P_KDD^T)^{-1}P_KBR^{-1}B^T](\Lambda_K)^{-T}\Delta P_KA^* \\
&= (A^*)^T\Delta P_K\Lambda_K^{-1}BR^{-1}B^T(I_n - \gamma^{-2}P_KDD^T)^{-1}\Delta P_KA^*.
\end{aligned} \tag{B.27}
$$

By plugging (B.27) and $D(\gamma^2I_q - D^TP_KD)^{-1}D^T = \gamma^{-2}DD^T(I_n - \gamma^{-2}P_KDD^T)^{-1}$ into (B.26), one can obtain

$$
(A^*)^T\Delta P_K(A^*) - \Delta P_K + E_K + (A^*)^T\Delta P_K\underbrace{(\gamma^{-2}DD^T - \Lambda_K^{-1}BR^{-1}B^T)(I_n - \gamma^{-2}P_KDD^T)^{-1}}_{:=V_K}\Delta P_KA^* = 0. \tag{B.28}
$$

It follows from (B.28) that

$$
\begin{aligned}
\Lambda_KV_K(I_n - \gamma^{-2}P_KDD^T) &= \gamma^{-2}(I_n + BR^{-1}B^TP_K - \gamma^{-2}DD^TP_K)DD^T - BR^{-1}B^T \\
&= (\gamma^{-2}DD^T - BR^{-1}B^T) + \gamma^{-2}(BR^{-1}B^T - \gamma^{-2}DD^T)P_KDD^T \\
&= (\gamma^{-2}DD^T - BR^{-1}B^T)(I_n - \gamma^{-2}P_KDD^T).
\end{aligned} \tag{B.29}
$$

Consequently, V_K is expressed as

$$
V_K = \Lambda_K^{-1}(\gamma^{-2}DD^T - BR^{-1}B^T). \tag{B.30}
$$

By plugging (B.30) into (B.28), one can obtain

$$
(A^*)^T\Delta P_K(A^*) - \Delta P_K + E_K - (A^*)^T\Delta P_K\Lambda_K^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)\Delta P_KA^* = 0. \tag{B.31}
$$

Considering the equality

$$
\begin{aligned}
&I_n - (I_n + BR^{-1}B^TP_K - \gamma^{-2}DD^TP_K)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)\Delta P_K \\
&= (I_n + BR^{-1}B^TP_K - \gamma^{-2}DD^TP_K)^{-1}(I_n + BR^{-1}B^TP^* - \gamma^{-2}DD^TP^*) \\
&= [(I_n + BR^{-1}B^TP^* - \gamma^{-2}DD^TP^*) + (BR^{-1}B^T - \gamma^{-2}DD^T)\Delta P_K]^{-1}(I_n + BR^{-1}B^TP^* - \gamma^{-2}DD^TP^*) \\
&= [I_n + (I_n + BR^{-1}B^TP^* - \gamma^{-2}DD^TP^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)\Delta P_K]^{-1} \\
&= [I_n + (\Lambda^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)\Delta P_K]^{-1}.
\end{aligned} \tag{B.32}
$$

we can finally derive (B.31) as

$$
(A^*)^T[I_n + \Delta P_K(\Lambda^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)]^{-1}\Delta P_KA^* + E_K - \Delta P_K = 0. \tag{B.33}
$$

Vectorizing both sides of (B.33) yields

$$
[(A^*)^T \otimes (A^*)^T]\left\{I_n \otimes [I_n + \Delta P_K(\Lambda^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)]^{-1}\right\}\mathrm{vec}(\Delta P_K) - \mathrm{vec}(\Delta P_K) = -\mathrm{vec}(E_K). \tag{B.34}
$$

Let $K' := (R + B^TU_KB)^{-1}B^TU_KA$ and $L_{K',*} := (\gamma^2I_q - D^TP_KD)^{-1}D^TP_K(A - BK')$. Then,

$$
\begin{aligned}
A - BK' + DL_{K',*} &= [I_n + D(\gamma^2I_q - D^TP_KD)^{-1}D^TP_K](A - BK') \\
&= (I_n - \gamma^{-2}DD^TP_K)^{-1}(A - BK') \\
&= (I_n - \gamma^{-2}DD^TP_K)^{-1}[I_n - B(R + B^TU_KB)^{-1}B^TU_K]A \\
&= (I_n - \gamma^{-2}DD^TP_K)^{-1}(I_n + BR^{-1}B^TU_K)^{-1}A \\
&= (I_n - \gamma^{-2}DD^TP_K)^{-1}[I_n + BR^{-1}B^TP_K(I_n - \gamma^{-2}DD^TP_K)^{-1}]^{-1}A \\
&= (I_n + BR^{-1}B^TP_K - \gamma^{-2}DD^TP_K)^{-1}A.
\end{aligned} \tag{B.35}
$$

Since $K \in \mathcal{W}$, by Lemma 12, $K' \in \mathcal{W}$. Consequently, $A - BK' + DL_{K',*}$ is stable by Lemma 5. Since $A^* = (\Lambda^*)^{-1}A$ and $\Lambda^* = I_n + BR^{-1}B^TP^* - \gamma^{-2}DD^TP^*$, we have

$$
\begin{aligned}
(A^*)^T[I_n + \Delta P_K(\Lambda^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)]^{-1} &= A^T(\Lambda^*)^{-T}[I_n + \Delta P_K(\Lambda^*)^{-1}(BR^{-1}B^T - \gamma^{-2}DD^T)]^{-1} \\
&= A^T(\Lambda^*)^{-T}[I_n + \Delta P_K(BR^{-1}B^T - \gamma^{-2}DD^T)(\Lambda^*)^{-T}]^{-1} = A^T[I_n + P_K(BR^{-1}B^T - \gamma^{-2}DD^T)]^{-1} \\
&= (A - BK' + DL_{K',*})^T,
\end{aligned}
\tag{B.36}
$$

where the second equation comes from Lemma 4, and the last equation is from (B.35). Plugging (B.36) into (B.34) results in

$$
[(A^*)^T \otimes (A - BK' + DL_{K',*})^T]\operatorname{vec}(\Delta P_K) - \operatorname{vec}(\Delta P_K) = -\operatorname{vec}(E_K). \tag{B.37}
$$

Since $(A - BK' + DL_{K',*})$ and A^* are stable, we have $\bar{\sigma}[(A^*)^T \otimes (A - BK' + DL_{K',*})^T] < 1$. As a result, we have

$$
\operatorname{vec}(\Delta P_K) = -[(A^*)^T \otimes (A - BK' + DL_{K',*})^T - I_{n^2}]^{-1}\operatorname{vec}(E_K). \tag{B.38}
$$

Taking the norm on both sides of (B.38), we have

$$
\|\Delta P_K\|_F \leq a(K)\|E_K\|_F, \tag{B.39a}
$$

$$
a(K) = \underline{\sigma}^{-1}\left[(A^*)^T \otimes (A - BK' + DL_{K',*})^T - I_{n^2}\right]. \tag{B.39b}
$$

□

Now, we are ready to prove the statement of Theorem 1.
Proof of Theorem 1. Following (B.8) and Lemma 1, we have

$$
(P_i - P_{i+1}) \succeq \sum_{t=0}^{\infty}(A_{i+1,*}^T)^tE_iA_{i+1,*}^t. \tag{B.40}
$$

where $E_i = E_{K_i} = (K_{i+1} - K_i)^T(R + B^TU_iB)(K_{i+1} - K_i)$. Subtracting P^* from both sides of (B.40) and taking trace of (B.40), we have

$$
\operatorname{Tr}(P_{i+1} - P^*) \leq \operatorname{Tr}(P_i - P^*) - \operatorname{Tr}\left[\sum_{t=0}^{\infty}(A_{i+1,*}^T)^tE_iA_{i+1,*}^t\right] \leq \operatorname{Tr}(P_i - P^*) - \operatorname{Tr}(E_i) \leq \operatorname{Tr}(P_i - P^*) - \|E_i\|_F, \tag{B.41}
$$

where the last inequality comes from Lemma 7. Considering Lemmas 7 and 15, (B.41) can be further derived as

$$
\operatorname{Tr}(P_{i+1} - P^*) \leq \left(1 - \frac{1}{\sqrt{n}a(K_i)}\right)\operatorname{Tr}(P_i - P^*). \tag{B.42}
$$

From Lemma 12, it is seen that K_i converges to K^*, and as a result, $\lim_{i\to\infty} a(K_i) = a(K^*)$. Consequently, $\bar{a} = \sup_{i\in\mathbb{Z}_+} a(K_i) < \infty$. The theorem is proved by setting $\alpha = 1 - \frac{1}{\sqrt{n}\bar{a}}$. Since $P^* \preceq P_{i+1} \preceq P_i$ from Lemma 12, $0 \leq \operatorname{Tr}(P_{i+1} - P^*) \leq \operatorname{Tr}(P_i - P^*)$. Hence, $\alpha \in [0, 1)$.

Appendix C: Proof of Theorem 2

The following lemma states the monotonic convergence of the inner-loop iteration.

Lemma 16. *Assume the inner loop starts from the initial condition $L_{i,1} = 0$. For any $i, j \in \mathbb{Z}_+$, the following statements hold*

1. *$A_{i,j}$ is stable;*
2. *$P_i \succeq \cdots \succeq P_{i,j} \succeq P_{i,j+1} \succeq \cdots \succeq P_{i,1}$;*
3. *$\lim_{j\to\infty}\|P_{i,j} - P_i\|_F = 0$ and $\lim_{j\to\infty}\|L_{i,j} - L_{i,*}\|_F = 0$.*

Proof. Considering the equalities $U_i = (I_n - \gamma^{-2}P_iDD^T)^{-1}P_i$, $A_{i,*} = (I_n - \gamma^{-2}DD^TP_i)^{-1}A_i$ from (B.7), and $L_{i,*}$ defined in (15), we can rewrite (10a) as

$$
A_{i,*}^TP_iA_{i,*} - P_i + Q_i - \gamma^2L_{i,*}^TL_{i,*} = 0. \tag{C.1}
$$

Since $P_i \succ 0$ and $\|A_{i,*}\| < 1$, $P_i - A_{i,*}^TP_iA_{i,*} \succ 0$. Therefore, from (C.1), we have

$$
Q_i - \gamma^2L_{i,*}^TL_{i,*} \succ 0. \tag{C.2}
$$

Considering the equality $(\gamma^2 I_q - D^T P_i D)L_{i,*} = D^T P_i A_i$ in (15) and completing the square of $(L_{i,*} - L_{i,j})^T(\gamma^2 I_q - D^T P_i D)(L_{i,*} - L_{i,j})$, we can rewrite (C.1) as

$$A_{i,j}^T P_i A_{i,j} - P_i + Q_i - \gamma^2 L_{i,j}^T L_{i,j} + (L_{i,*} - L_{i,j})^T(\gamma^2 I_q - D^T P_i D)(L_{i,*} - L_{i,j}) = 0. \tag{C.3}$$

Subtracting (17a) from (C.3) and completing the squares yield

$$A_{i,j}^T(P_i - P_{i,j})A_{i,j} - (P_i - P_{i,j}) + (L_{i,*} - L_{i,j})^T(\gamma^2 I_q - D^T P_i D)(L_{i,*} - L_{i,j}) = 0. \tag{C.4}$$

We prove the first statement by induction. When $j = 1$, the inner loop starts from $L_{i,j} = 0$. Since $K_i \in \mathcal{W}$, $A_{i,1} = A - BK_i + DL_{i,1} = A - BK_i$ is stable. Now, assume $A_{i,j}$ is stable for some $j \geq 1$. Since $A_{i,j}$ is stable and $(L_{i,*} - L_{i,j})^T(\gamma^2 I_q - D^T P_i D)(L_{i,*} - L_{i,j}) \succeq 0$, Lemma 1 and (C.4) result in $P_i \succeq P_{i,j} \succ 0$. In addition, following the derivation of (C.3), (C.1) can be rewritten as

$$A_{i,j+1}^T P_i A_{i,j+1} - P_i + Q_i - \gamma^2 L_{i,j+1}^T L_{i,j+1} + (L_{i,*} - L_{i,j+1})^T(\gamma^2 I_q - D^T P_i D)(L_{i,*} - L_{i,j+1}) = 0. \tag{C.5}$$

As $P_i \succeq P_{i,j}$, $(\gamma^2 I_q - D^T P_i D)^{-1} \succeq (\gamma^2 I_q - D^T P_{i,j} D)^{-1}$. As a consequence, $L_{i,j}^T L_{i,j} \preceq L_{i,*}^T L_{i,*}$ is obtained by comparing (15) with (17b). Then, $Q_i - \gamma^2 L_{i,j+1}^T L_{i,j+1} \succ 0$ follows from (C.2). From (C.5) and Lemma 2, we see that $A_{i,j+1}$ is stable. *A fortiori*, the first statement holds.

For $(j+1)$th iteration, the policy evaluation step in (17a) is

$$A_{i,j+1}^T P_{i,j+1} A_{i,j+1} - P_{i,j+1} + Q_i - \gamma^2 L_{i,j+1}^T L_{i,j+1} = 0. \tag{C.6}$$

Subtracting (17a) from (C.6), considering $(\gamma^2 I_q - D^T P_{i,j} D)L_{i,j+1} = D^T P_{i,j} A_i$ in (17b), and completing the square, we see

$$A_{i,j+1}^T(P_{i,j+1} - P_{i,j})A_{i,j+1} - (P_{i,j+1} - P_{i,j}) + (L_{i,j+1} - L_{i,j})^T(\gamma^2 I_q - D^T P_{i,j} D)(L_{i,j+1} - L_{i,j}) = 0. \tag{C.7}$$

As $K_i \in \mathcal{W}$ and by Lemma 5, $\gamma^2 I_q - D^T P_i D \succ 0$. By the fact that $P_i \succeq P_{i,j}$, we have $\gamma^2 I_q - D^T P_{i,j} D \succ 0$. Since $A_{i,j+1}$ is stable, by Lemma 1, $P_{i,j+1} - P_{i,j} \succeq 0$. Hence, the second statement holds.

From the second state, the sequence $\{P_{i,j}\}_{j=1}^{\infty}$ is monotonically increasing and bounded by P_i. Therefore, there exists a constant matrix $P_{i,\infty} \succeq 0$ such that $\lim_{j\to\infty} P_{i,j} = P_{i,\infty}$. It can be verified that $P_{i,\infty}$ satisfies (10a). Due to the uniqueness of the positive definite solution of (10a), we have $P_{i,\infty} = P_i$ and the third statement holds. $\square$

Lemma 17. *Let $E_{i,j} := (L_{i,j} - L_{i,j+1})^T(\gamma^2 I_q - D^T P_{i,j} D)(L_{i,j} - L_{i,j+1})$. Then, for any $i, j \in \mathbb{Z}_+$, there exists a constant $b(K_i) > 0$, such that*

$$\mathrm{Tr}(P_i - P_{i,j}) \leq b(K_i)\|E_{i,j}\|, \tag{C.8}$$

and

$$b(K_i) := \mathrm{Tr}\left[\sum_{t=0}^{\infty}(A_{i,*})^t(A_{i,*}^T)^t\right]. \tag{C.9}$$

Proof. Considering $(\gamma^2 I_q - D^T P_{i,j} D)L_{i,j+1} = D^T P_{i,j} A_i$ in (17b) and completing the squares, (17a) can be rewritten as

$$\begin{aligned} &A_{i,*}^T P_{i,j} A_{i,*} - P_{i,j} + Q_i - (L_{i,j} - L_{i,j+1})^T(\gamma^2 I_q - D^T P_{i,j} D)(L_{i,j} - L_{i,j+1}) \\ &+ (L_{i,*} - L_{i,j+1})^T(\gamma^2 I_q - D^T P_{i,j} D)(L_{i,*} - L_{i,j+1}) - \gamma^2 L_{i,*}^T L_{i,*} = 0 \end{aligned} \tag{C.10}$$

Subtracting it from (C.1) results in

$$A_{i,*}^T(P_i - P_{i,j})A_{i,*} - (P_i - P_{i,j}) + E_{i,j} - (L_{i,*} - L_{i,j+1})^T(\gamma^2 I_q - D^T P_{i,j} D)(L_{i,*} - L_{i,j+1}) = 0. \tag{C.11}$$

As $A_{i,*}$ is stable, by Lemma 1, we have

$$P_i - P_{i,j} \preceq \sum_{t=0}^{\infty}(A_{i,*}^T)^t E_{i,j}(A_{i,*})^t. \tag{C.12}$$

The following inequality can be derived by the cyclic property of trace and Lemma 6

$$\mathrm{Tr}(P_i - P_{i,j}) \leq \mathrm{Tr}\left[E_{i,j}\sum_{t=0}^{\infty}(A_{i,*})^t(A_{i,*}^T)^t\right] \leq \|E_{i,j}\| \underbrace{\mathrm{Tr}\left[\sum_{t=0}^{\infty}(A_{i,*})^t(A_{i,*}^T)^t\right]}_{b(K_i)}. \tag{C.13}$$

$\square$

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. By (C.7), Lemma 1 and the definition of $E_{i,j}$ in Lemma 17, we have

$$\operatorname{Tr}(P_{i,j+1} - P_{i,j}) = \operatorname{Tr}\left[\sum_{t=0}^{\infty}(A_{i,j+1}^T)^t E_{i,j}(A_{i,j+1})^t\right]. \tag{C.14}$$

Consequently,

$$\begin{aligned}\operatorname{Tr}(P_i - P_{i,j+1}) &= \operatorname{Tr}(P_i - P_{i,j}) - \operatorname{Tr}\left[\sum_{t=0}^{\infty}(A_{i,j+1}^T)^t E_{i,j}(A_{i,j+1})^t\right] \leq \operatorname{Tr}(P_i - P_{i,j}) - \operatorname{Tr}(E_{i,j}) \\ &\leq \operatorname{Tr}(P_i - P_{i,j}) - \|E_{i,j}\| \leq \underbrace{(1 - \frac{1}{b(K_i)})}_{\beta(K_i)} \operatorname{Tr}(P_i - P_{i,j}),\end{aligned} \tag{C.15}$$

where the last inequality comes from Lemma 17. Since $P_i \succeq P_{i,j}$ for any $j \in \mathbb{Z}_+$, $\operatorname{Tr}(P_i - P_{i,j+1}) \geq 0$. Hence, $\beta(K_i) \in [0, 1)$.

Appendix D: Proof of Theorem 3

Let $M_i := \sum_{t=0}^{\infty}(A_{i,*})^t(A_{i,*}^T)^t$. Since $A_{i,*}$ is stable, by Lemma 1, M_i is the unique solution to

$$A_{i,*}M_iA_{i,*}^T - M_i + I_n = 0. \tag{D.1}$$

From Theorem 1, as K_i converges to K^*, $A_{i,*} = A - BK_i + D(\gamma^2 I_q - D^T P_i D)^{-1} D^T P_i (A - BK_i)$ converges to A^*. As a result, M_i converges to M^*, which is the unique solution to

$$(A^*)M^*(A^*)^T - M^* + I_n = 0. \tag{D.2}$$

Since $b(K_i) = \operatorname{Tr}(M_i)$ in Lemma 17 converges to $b(K^*) = \operatorname{Tr}(M^*)$, the supreme of $b(K_i)$ exists and bounded, i.e. $\bar{b} = \sup_{i\in\mathbb{Z}_+} b(K_i) < \infty$. Hence, for any $i \in \mathbb{Z}_+$, $\beta(K_i) \leq \bar{\beta}$, which is defined as

$$\bar{\beta} := 1 - \frac{1}{\bar{b}}. \tag{D.3}$$

By Theorem 2 and the fact that P_i is monotonically decreasing, for any $i \in \mathbb{Z}_+$, we have

$$\|P_i - P_{i,j}\|_F \leq \operatorname{Tr}(P_i - P_{i,j}) \leq \bar{\beta}^{j-1}\operatorname{Tr}(P_i - P_{i,1}) \leq \bar{\beta}^{j-1}\operatorname{Tr}(P_i) \leq \bar{\beta}^{j-1}\operatorname{Tr}(P_1). \tag{D.4}$$

Therefore, for any $i \in \mathbb{Z}_+$ and $\epsilon > 0$, if $j \geq \bar{j} = \log_{\bar{\beta}}^{\frac{\epsilon}{\operatorname{Tr}(P_1)}} + 1$, $\|P_{i,j} - P_i\|_F \leq \epsilon$. It is noticed that $\bar{j}$ is independent on i, and therefore, the uniform convergence of the dual-loop algorithm is demonstrated.

Appendix E: Proof of Theorem 4

The following lemma tells us that for any $K \in \mathcal{W}$, when it is disturbed by a small ΔK, $K + \Delta K \in \mathcal{W}$.

Lemma 18. *For any $K \in \mathcal{W}$, there exists $c(K) > 0$, such that $K + \Delta K \in \mathcal{W}$, whenever $\|\Delta K\|_F \leq c(K)$.*

Proof. Since $K \in \mathcal{W}$, by Lemmas 4 and 5, there exists a positive definite matrix $P_K \in \mathbb{S}^n$ satisfying

$$(A - BK)^T P_K (A - BK) - P_K + Q + K^T RK + (A - BK)^T \gamma^{-2}(I_n - \gamma^{-2}P_K DD^T)^{-1} P_K DD^T P_K (A - BK) = 0. \tag{E.1}$$

Define the matrix-valued functions $F(\Delta K, \Delta P)$ and $S(\Delta P)$ as

$$S(\Delta P) := \gamma^{-2}[I_n - \gamma^{-2}(P_K + \Delta P)DD^T]^{-1}(P_K + \Delta P)DD^T(P_K + \Delta P) \tag{E.2a}$$

$$\begin{aligned}F(\Delta K, \Delta P) := {}& [A - B(K + \Delta K)]^T(P_K + \Delta P)[A - B(K + \Delta K)] + Q + (K + \Delta K)^T R(K + \Delta K) \\ & - (P_K + \Delta P) + [A - B(K + \Delta K)]^T S(\Delta P)[A - B(K + \Delta K)].\end{aligned} \tag{E.2b}$$

It is seen that when, when K is disturbed by ΔK, P_K will turn into $P_K + \Delta P$ to guarantee the left hand side of (E.1) equal to zero, i.e. $F(\Delta K, \Delta P) = 0$. Therefore, ΔP can be considered as a implicit function of ΔK. In addition, $\Delta K = 0$ implies

$\Delta P = 0$. Let $\mathcal{F}(\Delta K, \text{vec}(\Delta P)) := \text{vec}[F(\Delta K, \Delta P)]$ and $\mathcal{S}[\text{vec}(\Delta P)] := \text{vec}[S(\Delta P)]$. Using Lemma 9, we have

$$
\begin{aligned}
\frac{\partial \mathcal{S}[\text{vec}(\Delta P)]}{\partial \,\text{vec}(\Delta P)} &= \gamma^{-2}\left[(P_K+\Delta P)DD^T(P_K+\Delta P)\otimes I_n\right]\frac{\partial\,\text{vec}\{[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\}}{\partial\,\text{vec}(\Delta P)} \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\}[(P_K+\Delta P)D\otimes I_n]\frac{\partial\,\text{vec}[(P_K+\Delta P)D]}{\partial\,\text{vec}(\Delta P)} \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\}[I_n\otimes(P_K+\Delta P)D]\frac{\partial\,\text{vec}[D^T(P_K+\Delta P)]}{\partial\,\text{vec}(\Delta P)} \qquad \text{(E.3a)} \\
&= \gamma^{-2}\left[(P_K+\Delta P)DD^T(P_K+\Delta P)\otimes I_n\right]\left\{(\gamma^{-2}DD^T)[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\} \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\}[(P_K+\Delta P)D\otimes I_n](D^T\otimes I_n) \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\}[I_n\otimes(P_K+\Delta P)D](I_n\otimes D^T) \qquad \text{(E.3b)} \\
&= \gamma^{-2}\left\{(P_K+\Delta P)DD^T(P_K+\Delta P)(\gamma^{-2}DD^T)[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\} \\
&\quad + \gamma^{-2}\left\{(P_K+\Delta P)DD^T\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\} \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}(P_K+\Delta P)DD^T\right\} \qquad \text{(E.3c)} \\
&= \gamma^{-2}\left\{(P_K+\Delta P)DD^T[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}\right\} \\
&\quad + \gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}(P_K+\Delta P)DD^T]^{-1}(P_K+\Delta P)DD^T\right\}. \qquad \text{(E.3d)}
\end{aligned}
$$

and

$$
\frac{\partial \mathcal{F}(\Delta K, \text{vec}(\Delta P))}{\partial\,\text{vec}(\Delta P)} = \left\{[A-B(K+\Delta K)]^T\otimes[A-B(K+\Delta K)]^T\right\}\left(I_{n^2}+\frac{\partial \mathcal{S}[\text{vec}(\Delta P)]}{\partial\,\text{vec}(\Delta P)}\right)-I_{n^2}. \qquad \text{(E.4)}
$$

Equation (E.3b) is derived from (Zhang, Hu, and Başar 2019, Equation (B.10)). When $\Delta P = 0$ and $\Delta K = 0$, from (E.3d), we have

$$
\begin{aligned}
I_{n^2}+\frac{\partial \mathcal{S}[\text{vec}(\Delta P)]}{\partial\,\text{vec}(\Delta P)} &= I_n\otimes I_n+\gamma^{-2}\left\{I_n\otimes[I_n-\gamma^{-2}P_KDD^T]^{-1}P_KDD^T\right\} \\
&\quad + \gamma^{-2}\left\{P_KDD^T[I_n-\gamma^{-2}P_KDD^T]^{-1}\otimes[I_n-\gamma^{-2}P_KDD^T]^{-1}\right\} \qquad \text{(E.5)} \\
&= I_n\otimes[I_n-\gamma^{-2}P_KDD^T]^{-1}+\gamma^{-2}\left\{P_KDD^T[I_n-\gamma^{-2}P_KDD^T]^{-1}\otimes[I_n-\gamma^{-2}P_KDD^T]^{-1}\right\} \\
&= [I_n-\gamma^{-2}P_KDD^T]^{-1}\otimes[I_n-\gamma^{-2}P_KDD^T]^{-1},
\end{aligned}
$$

where the second equality comes from $W\otimes X+W\otimes Y=W\otimes(X+Y)$. Following the derivation of (B.7) and the expression of $L_{K,*}=(\gamma^2 I_q-D^TP_KD)^{-1}D^TP_K(A-BK)$, we have

$$
(I_n-\gamma^{-2}DD^TP_K)^{-1}(A-BK)=A-BK+DL_{K,*} \qquad \text{(E.6)}
$$

Plugging (E.5) and (E.6) into (E.4) and considering $(W\otimes X)(Y\otimes Z)=(WY\otimes XZ)$, we have

$$
\frac{\partial \mathcal{F}(\Delta K, \text{vec}(\Delta P))}{\partial\,\text{vec}(\Delta P)}|_{\Delta P=0,\Delta K=0} = \left\{(A-BK+DL_{K,*})^T\otimes(A-BK+DL_{K,*})^T\right\}-I_{n^2}. \qquad \text{(E.7)}
$$

As $K\in\mathcal{K}$, from Lemma 5, it is seen that $(A-BK+DL_{K,*})$ is stable and $\bar{\sigma}(A-BK+DL_{K,*})<1$. Consequently, $\frac{\partial \mathcal{F}(\Delta K, \text{vec}(\Delta P))}{\partial\,\text{vec}(\Delta P)}|_{\Delta P=0,\Delta K=0}$ is invertible. By the implicit function theorem, there exists $c_1(K)>0$, such that ΔP is continuously differentiable with respect to $\Delta K\in\mathcal{B}(0,c_1(K))$. Since $P_K\succ 0$, there exists $c(K)\le c_1(K)$, such that $P_K+\Delta P\succ 0$, for any $\Delta K\in\mathcal{B}(0,c(K))$. Hence, according to Lemma 5 and (E.2b), $K+\Delta K\in\mathcal{W}$. □

For any $H\in\mathcal{W}$, we can define a compact set $\mathcal{G}_H$ consisting of all the controllers $K\in\mathcal{W}$ with smaller cost matrix P_K. Then, the following lemma ensures that when disturbed by ΔK, the updated controller $K'+\Delta K\in\mathcal{G}_H$ given that $K\in\mathcal{G}_H$ and ΔK is small.

Lemma 19. *For any $H\in\mathcal{W}$, define $\mathcal{G}_H:=\{K\in\mathcal{W}|\,\text{Tr}(P_K)\le\text{Tr}(P_H)\}$. Let $K\in\mathcal{G}_H$, $K':=(R+B^TU_KB)^{-1}B^TU_KA$, and $\hat{K}':=K'+\Delta K$. Then, exists $d(H)>0$, such that $\hat{K}'\in\mathcal{G}_H$ if $\|\Delta K\|_F<d(H)$.*

Proof. Since $K \in \mathcal{W}$, from Lemma 12, it is seen that $K' \in \mathcal{W}$. Let $\bar{c}(H) = \inf_{K\in\mathcal{G}_H} c(K)$. By Lemma 18, if $\|\Delta K\|_F \leq \bar{c}(H)$, $\hat{K}' \in \mathcal{W}$. Consequently, according to Lemma 5, there exists the unique positive definite solution $\hat{P}_{K'}$ to

$$(A - B\hat{K}')^T \hat{U}_{K'}(A - B\hat{K}') - \hat{P}_{K'} + Q + (\hat{K}')^T R\hat{K}' = 0, \tag{E.8a}$$

$$\hat{U}_{K'} = \hat{P}_{K'} + \hat{P}_{K'}D(\gamma^2 I_q - D^T\hat{P}_{K'}D)^{-1}D^T\hat{P}_{K'}. \tag{E.8b}$$

Furthermore, (A.6a) is rewritten as

$$\begin{aligned}&(A - B\hat{K}')^T U_K(A - B\hat{K}') - P_K + Q + (\hat{K}' - K)^T B^T U_K A + A^T U_K B(\hat{K}' - K)\\ &\quad + K^T(R + B^T U_K B)K - (\hat{K}')^T B^T U_K B\hat{K}' = 0.\end{aligned} \tag{E.9}$$

Considering the equality $(R + B^T U_K B)K' = B^T U_K A$ and $\hat{K}' = K' + \Delta K$, (E.9) is further derived as

$$\begin{aligned}&(A - B\hat{K}')^T U_K(A - B\hat{K}') - P_K + Q + (K' - K)^T(R + B^T U_K B)K' + (K')^T(R + B^T U_K B)(K' - K)\\ &\quad + \Delta K^T(R + B^T U_K B)K' + (K')^T(R + B^T U_K B)\Delta K + K^T(R + B^T U_K B)K - (\hat{K}')^T B^T U_K B\hat{K}' = 0.\end{aligned} \tag{E.10}$$

Subtracting (E.8a) from (E.10) and completing the squares yield

$$(A - B\hat{K}')^T(U_K - \hat{U}_{K'})(A - B\hat{K}') - (P_K - \hat{P}_{K'}) + E_K - \Delta K^T(R + B^T U_K B)\Delta K = 0. \tag{E.11}$$

Recall that $E_K = (K' - K)^T(R + B^T U_K B)(K' - K)$ in Lemma 15. Following Lemma 11, we have

$$(A - B\hat{K}' + D\hat{L}_{K',*})^T(P_K - \hat{P}_{K'})(A - B\hat{K}' + D\hat{L}_{K',*}) - (P_K - \hat{P}_{K'}) + E_K - \Delta K^T(R + B^T U_K B)\Delta K \preceq 0, \tag{E.12}$$

where $\hat{L}_{K',*} = (\gamma^2 I_q - D^T\hat{P}_{K'}D)^{-1}D^T\hat{P}_{K'}(A - B\hat{K}')$. Since $\hat{K}' \in \mathcal{W}$, $(A - B\hat{K}' + D\hat{L}_{K',*})$ is stable by Lemma 5. Therefore, according to Lemma 1, we have

$$\mathrm{Tr}(P_K - \hat{P}_{K'}) \geq \mathrm{Tr}\left\{\sum_{t=0}^{\infty}\left\{((A - B\hat{K}' + D\hat{L}_{K',*})^T)^t[E_K - \Delta K^T(R + B^T U_K B)\Delta K](A - B\hat{K}' + D\hat{L}_{K',*})^t\right\}\right\} \tag{E.13}$$

Let $\bar{a}(H) = \sup_{K\in\mathcal{G}_H} a(K)$, where $a(K)$ is defined in Lemma 15. In addition, define $d_1(H) = \sup_{K\in\mathcal{G}_H}\|\sum_{t=0}^{\infty}(A - BK + DL_{K,*})^t((A - BK + DL_{K,*})^T)^t\|$, and $d_2(H) = \sup_{K\in\mathcal{G}_H}\|R + B^T U_K B\|$. Then, by Lemmas 6 and 15, (E.13) implies

$$\mathrm{Tr}(\hat{P}_{K'} - P^*) \leq (1 - \frac{1}{\sqrt{n}\bar{a}(H)})\,\mathrm{Tr}(P_K - P^*) + d_1(H)d_2(H)\|\Delta K\|_F^2. \tag{E.14}$$

Therefore, if $\|\Delta K\|_F^2 \leq \frac{\mathrm{Tr}(P_H - P^*)}{d_1(H)d_2(H)\sqrt{n}\bar{a}(H)}$, it is ensured that $\mathrm{Tr}(\hat{P}_{K'} - P^*) \leq \mathrm{Tr}(\hat{P}_H - P^*)$. In summary, $d(H) := \min\left\{\bar{c}(H), \sqrt{\frac{\mathrm{Tr}(P_H - P^*)}{d_1(H)d_2(H)\sqrt{n}\bar{a}(H)}}\right\}$, and if $\|\Delta K\|_F \leq d(H)$, $\hat{K}' \in \mathcal{G}_H$ is ensured. □

Now, we are ready to prove the statement of Theorem 4.

Proof of Theorem 4. From Lemma 19 and given an initial feasible controller $\hat{K}_1$, it is seen that if $\|\Delta K\|_\infty \leq d(\hat{K}_1)$, $\hat{K}_i \in \mathcal{G}_{\hat{K}_1}$ for any $i \in \mathbb{Z}_+$. In (E.14), considering $\hat{P}_i$ and $\hat{P}_{i+1}$ as P_K and $\hat{P}_{K'}$, respectively, we have

$$\mathrm{Tr}(\hat{P}_{i+1} - P^*) \leq (1 - \frac{1}{\sqrt{n}\bar{a}(\hat{K}_1)})\,\mathrm{Tr}(\hat{P}_i - P^*) + d_1(\hat{K}_1)d_2(\hat{K}_1)\|\Delta K_{i+1}\|_F^2. \tag{E.15}$$

Repeating the above inequality from $i = 1$ yields

$$\mathrm{Tr}(\hat{P}_{i+1} - P^*) \leq (1 - \frac{1}{\sqrt{n}\bar{a}(\hat{K}_1)})^i\,\mathrm{Tr}(\hat{P}_1 - P^*) + \sum_{s=1}^{i}(1 - \frac{1}{\sqrt{n}\bar{a}(\hat{K}_1)})^{s-1}d_1(\hat{K}_1)d_2(\hat{K}_1)\|\Delta K\|_\infty^2 \tag{E.16a}$$

$$\leq (1 - \frac{1}{\sqrt{n}\bar{a}(\hat{K}_1)})^i\,\mathrm{Tr}(\hat{P}_1 - P^*) + \sqrt{n}\bar{a}(\hat{K}_1)d_1(\hat{K}_1)d_2(\hat{K}_1)\|\Delta K\|_\infty^2 \tag{E.16b}$$

It is seen that $\kappa_1(\mathrm{Tr}(\hat{P}_1 - P^*), i) = (1 - \frac{1}{\sqrt{n}\bar{a}(\hat{K}_1)})^i\,\mathrm{Tr}(\hat{P}_1 - P^*)$ is a $\mathcal{KL}$-function, and $\xi_1(\|\Delta K\|_\infty) = \sqrt{n}\bar{a}(\hat{K}_1)d_1(\hat{K}_1)d_2(\hat{K}_1)\|\Delta K\|_\infty^2$ is a $\mathcal{K}$-function. Therefore, it is concluded that the inexact outer loop iteration is small-disturbance input-to-state stable.

Appendix F: Proof of Theorem 5

The following lemma ensures the stability of the closed-loop system with the feedback gain $\hat{L}_{i,j}$ generated from the inexact inner loop.

Lemma 20. *Given $\hat{K}_i \in \mathcal{W}$, there exists a constant $e(\hat{K}_i) > 0$, such that $\hat{A}_{i,j} = A - B\hat{K}_i + D\hat{L}_{i,j}$ is stable for all $j \in \mathbb{Z}_+$, as long as $\|\Delta L_i\|_\infty < e(\hat{K}_i)$.*

Proof. This lemma is proven by induction. To simply the notations, the following variable is defined to denote the inner-loop update without disturbance

$$\tilde{L}_{i,j+1} := (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T \hat{P}_{i,j} \hat{A}_i. \tag{F.1}$$

Then, $\hat{L}_{i,j+1} = \tilde{L}_{i,j+1} + \Delta L_{i,j+1}$. Since $\hat{K}_i \in \mathcal{W}$ and $\hat{L}_{i,1} = 0$, $\hat{A}_{i,1} = A - B\hat{K}_i + D\hat{L}_{i,1}$ is stable by Lemma 5. Assume $\hat{A}_{i,j} = A - B\hat{K}_i + D\hat{L}_{i,j}$ is stable for some $j \in \mathbb{Z}_+$. Following the derivation of (C.4), we have

$$\hat{A}_{i,j}^T(\hat{P}_i - \hat{P}_{i,j})\hat{A}_{i,j} - (\hat{P}_i - \hat{P}_{i,j}) + (\hat{L}_{i,*} - \hat{L}_{i,j})^T(\gamma^2 I_q - D^T\hat{P}_i D)(\hat{L}_{i,*} - \hat{L}_{i,j}) = 0. \tag{F.2}$$

Since $\hat{A}_{i,j}$ is stable by assumption, from Lemma 1, $\hat{P}_i \succeq \hat{P}_{i,j}$, where $\hat{P}_i$ is from (22a). By completing the square and $(\gamma^2 I_q - D^T\hat{P}_i D)\hat{L}_{i,*} = D^T\hat{P}_i(A - B\hat{K}_i)$, (22a) can be rewritten as

$$\begin{aligned}&\hat{A}_{i,j+1}^T\hat{P}_i\hat{A}_{i,j+1} - \hat{P}_i + \hat{Q}_i - \gamma^2\tilde{L}_{i,j+1}^T\tilde{L}_{i,j+1} + (\tilde{L}_{i,j+1} - \hat{L}_{i,*})^T(\gamma^2 I_q - D^T\hat{P}_i D)(\tilde{L}_{i,j+1} - \hat{L}_{i,*})\\ &\underbrace{-\Delta L_{i,j+1}^T D^T\hat{P}_i D\tilde{L}_{i,j+1} - \tilde{L}_{i,j+1}^T D^T\hat{P}_i D\Delta L_{i,j+1} - \Delta L_{i,j+1}^T D^T\hat{P}_i D\Delta L_{i,j+1} - \hat{A}_i^T\hat{P}_i D\Delta L_{i,j+1} - \Delta L_{i,j+1}^T D^T\hat{P}_i\hat{A}_i}_{:=-\hat{\Omega}_{i,j+1}} = 0.\end{aligned} \tag{F.3}$$

Since $\hat{P}_i \succeq \hat{P}_{i,j}$, $\hat{L}_{i,*}^T\hat{L}_{i,*} \succeq \tilde{L}_{i,j+1}^T\tilde{L}_{i,j+1}$ and $\|\hat{L}_{i,*}\| \geq \|\tilde{L}_{i,j+1}\|$. As a consequence, the upperbound of $\|\hat{\Omega}_{i,j+1}\|$ is

$$\|\hat{\Omega}_{i,j+1}\| \leq \underbrace{(2\|D^T\hat{P}_i D\|\|\hat{L}_{i,*}\| + 2\|D^T\hat{P}_i\hat{A}_i\|)}_{=:e_1(\hat{K}_i)}\|\Delta L_{i,j+1}\| + \underbrace{\|D^T\hat{P}_i D\|}_{:=e_2(\hat{K}_i)}\|\Delta L_{i,j+1}\|^2. \tag{F.4}$$

Following Lemma 5, we know that $\hat{P}_i \succ 0$ and $\hat{A}_{i,*} = A - B\hat{K}_i + D\hat{L}_{i,*}$ is stable. The equation in outer loop (22a) can be rewritten as

$$\hat{A}_{i,*}^T\hat{P}_i\hat{A}_{i,*} - \hat{P}_i + \hat{Q}_i - \gamma^2\hat{L}_{i,*}^T\hat{L}_{i,*} = 0. \tag{F.5}$$

Therefore, by Lemma 1, $\hat{Q}_i - \gamma^2\hat{L}_{i,*}^T\hat{L}_{i,*} \succ 0$, and $e_3(\hat{K}_i) := \underline{\sigma}(\hat{Q}_i - \gamma^2\hat{L}_{i,*}^T\hat{L}_{i,*}) > 0$. Hence, if $\|\Delta L_{i,j+1}\|$ satisfies

$$\|\Delta L_{i,j+1}\| < \frac{-e_1 + \sqrt{e_1^2 + 4e_2e_3}}{2e_2} := e(\hat{K}_i), \tag{F.6}$$

we have

$$\hat{Q}_i - \gamma^2\hat{L}_{i,*}^T\hat{L}_{i,*} - \hat{\Omega}_{i,j+1} \succ 0. \tag{F.7}$$

As $\hat{L}_{i,*}^T\hat{L}_{i,*} \succeq \hat{L}_{i,j+1}^T\hat{L}_{i,j+1}$, $\hat{Q}_i - \gamma^2\hat{L}_{i,j+1}^T\hat{L}_{i,j+1} - \hat{\Omega}_{i,j+1} \succ 0$. $\hat{A}_{i,j+1}$ is stable as a result of (F.3) and Lemma 2. By induction, for any $j \in \mathbb{Z}_+$, $\hat{A}_{i,j}$ is stable.

□

Proof of Theorem 5. For the $(j+1)$th iteration of (23), we have

$$\hat{A}_{i,j+1}^T\hat{P}_{i,j+1}\hat{A}_{i,j+1} - \hat{P}_{i,j+1} + \hat{Q}_i - \gamma^2\hat{L}_{i,j+1}^T\hat{L}_{i,j+1} = 0 \tag{F.8}$$

For the jth iteration, (23a) can be rewritten as

$$\begin{aligned}&\hat{A}_{i,j+1}^T\hat{P}_{i,j}\hat{A}_{i,j+1} - \hat{P}_{i,j} + \hat{Q}_i - \gamma^2\tilde{L}_{i,j+1}^T\tilde{L}_{i,j+1} - (\hat{L}_{i,j} - \tilde{L}_{i,j+1})^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)(\hat{L}_{i,j} - \tilde{L}_{i,j+1})\\ &\quad - \gamma^2\Delta L_{i,j+1}^T\tilde{L}_{i,j+1} - \gamma^2\tilde{L}_{i,j+1}^T\Delta L_{i,j+1} - \Delta L_{i,j+1}^T D^T\hat{P}_{i,j}D\Delta L_{i,j+1} = 0.\end{aligned} \tag{F.9}$$

Subtracting (F.9) from (F.8) yields

$$\hat{A}_{i,j+1}^T(\hat{P}_{i,j+1} - \hat{P}_{i,j})\hat{A}_{i,j+1} - (\hat{P}_{i,j+1} - \hat{P}_{i,j}) + \underbrace{(\hat{L}_{i,j} - \tilde{L}_{i,j+1})^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)(\hat{L}_{i,j} - \tilde{L}_{i,j+1})}_{\hat{E}_{i,j}} \tag{F.10}$$

$$- \Delta L_{i,j+1}^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)\Delta L_{i,j+1} = 0. \tag{F.11}$$

When $\|\Delta L_i\|_\infty < e(\hat{K}_i)$, by Lemma 20, $\hat{A}_{i,j+1}$ is stable. Following Lemma 1, we have

$$\operatorname{Tr}(\hat{P}_{i,j+1} - \hat{P}_{i,j}) = \operatorname{Tr}\left\{\sum_{t=0}^{\infty}(\hat{A}_{i,j+1}^T)^t \left[\hat{E}_{i,j} - \Delta L_{i,j+1}^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)\Delta L_{i,j+1}\right](\hat{A}_{i,j+1})^t\right\}. \tag{F.12}$$

Consequently,

$$\begin{aligned}\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j+1}) &= \operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) - \operatorname{Tr}\left\{\sum_{t=0}^{\infty}(\hat{A}_{i,j+1}^T)^t \left[\hat{E}_{i,j} - \Delta L_{i,j+1}^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)\Delta L_{i,j+1}\right](\hat{A}_{i,j+1})^t\right\} \\ &\leq \operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) - \|\hat{E}_{i,j}\| + \operatorname{Tr}\left\{\sum_{t=0}^{\infty}(\hat{A}_{i,j+1}^T)^t\Delta L_{i,j+1}^T(\gamma^2 I_q - D^T\hat{P}_{i,j}D)\Delta L_{i,j+1}(\hat{A}_{i,j+1})^t\right\} \\ &\leq (1 - \frac{1}{c(\hat{K}_i)})\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) + \gamma^2\|\Delta L_i\|_\infty \operatorname{Tr}\left[\sum_{t=0}^{\infty}(\hat{A}_{i,j+1}^T)^t(\hat{A}_{i,j+1})^t\right],\end{aligned} \tag{F.13}$$

where the second line comes from the fact that $\operatorname{Tr}(\hat{E}_{i,j}) \geq \|\hat{E}_{i,j}\|$ of Lemma 7, and the third line is derived by the lowerbound of $\|\hat{E}_{i,j}\|$ in Lemma 17 and the upperbound of trace in Lemma 6.

When $\|\Delta L_i\|_\infty < e(\hat{K}_i)$, from (F.7), there exists a constant $\epsilon > 0$, such that $\|\hat{Q}_i - \gamma^2\hat{L}_{i,*}^T\hat{L}_{i,*} - \hat{\Omega}_{i,j+1}\| \geq \epsilon$. Let $\hat{M}_{i,j+1} := \sum_{t=0}^{\infty}(\hat{A}_{i,j+1}^T)^t(\hat{A}_{i,j+1})^t$, and by Lemma 1, $\hat{M}_{i,j+1}$ satisfies

$$\hat{A}_{i,j+1}^T\hat{M}_{i,j+1}\hat{A}_{i,j+1} - \hat{M}_{i,j+1} + I_n = 0. \tag{F.14}$$

Multiplying the above equation by ϵ and subtracting it from (F.3) yields

$$\begin{aligned}&\hat{A}_{i,j+1}^T(\hat{P}_i - \epsilon\hat{M}_{i,j+1})\hat{A}_{i,j+1} - (\hat{P}_i - \epsilon\hat{M}_{i,j+1}) + \hat{Q}_i - \gamma^2\tilde{L}_{i,j+1}^T\tilde{L}_{i,j+1} - \hat{\Omega}_{i,j+1} - \epsilon I_n \\ &+ (\tilde{L}_{i,j+1} - \hat{L}_{i,*})^T(\gamma^2 I_q - D^T\hat{P}_iD)(\tilde{L}_{i,j+1} - \hat{L}_{i,*}) = 0.\end{aligned} \tag{F.15}$$

Hence, by Lemma 1, $\hat{P}_i - \epsilon\hat{M}_{i,j+1} \succeq 0$, and as a consequence $\operatorname{Tr}(\hat{M}_{i,j+1}) \leq 1/\epsilon\operatorname{Tr}(\hat{P}_i)$.

From (F.13), we have

$$\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j+1}) \leq \beta(\hat{K}_i)\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) + 1/\epsilon\operatorname{Tr}(\hat{P}_i)\gamma^2\|\Delta L_i\|_\infty^2 \tag{F.16}$$

Recall that $\beta(\cdot)$ is defined in (C.15). Repeating the above argument for $j, j-1, \cdots, 1$ gives

$$\begin{aligned}\|\hat{P}_i - \hat{P}_{i,j}\|_F \leq \operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) &\leq \beta^{j-1}(\hat{K}_i)\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,1}) + \sum_{s=1}^{j-1}\beta^{s-1}(\hat{K}_i)1/\epsilon\operatorname{Tr}(\hat{P}_i)\gamma^2\|\Delta L_i\|_\infty^2 \\ &\leq \beta^{j-1}(\hat{K}_i)\sqrt{n}\|\hat{P}_i - \hat{P}_{i,1}\|_F + \frac{1}{1-\beta(\hat{K}_i)}1/\epsilon\operatorname{Tr}(\hat{P}_i)\gamma^2\|\Delta L_i\|_\infty^2\end{aligned} \tag{F.17}$$

Therefore, we can conclude that the inexact inner loop iteration is input-to-state stable.