Robust Reinforcement Learning for Risk-Sensitive Linear Quadratic Gaussian Control

Leilei Cui, Tamer Başar, *Life Fellow, IEEE,* and Zhong-Ping Jiang, *Fellow, IEEE*

***Abstract*— This paper proposes a novel robust reinforcement learning framework for discrete-time linear systems with model mismatch that may arise from the sim-to-real 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 optimization algorithm is proposed to generate a robust optimal controller. The dual-loop policy optimization algorithm is shown to be globally and uniformly convergent, and robust against disturbances during the learning process. This robustness property is called small-disturbance input-to-state stability and guarantees that the proposed policy optimization algorithm converges to a small neighborhood of the optimal controller as long as the disturbance at each learning step is relatively small. In addition, when the system dynamics is unknown, a novel model-free off-policy policy optimization algorithm is proposed. Finally, numerical examples are provided to illustrate the proposed algorithm.**

***Index Terms*— Robust reinforcement learning, policy optimization, risk-sensitive LQG.**

I. Introduction

By optimizing a specified accumulated performance index, reinforcement learning (RL), as a branch of machine learning, is aimed at learning optimal decisions from data in the absence of model knowledge. Policy optimization (PO) plays a pivotal role in the development of RL algorithms [1, Chapter 13]. The key idea of PO is to parameterize the policy and update the policy parameters along the gradient ascent direction of the performance index for maximization, or gradient descent for minimization. Since the system model is unknown, the policy gradient should be estimated by data-driven methods through sampling and experimentation. Consequently, accurate policy gradient can hardly be obtained because of various errors that may be induced by function approximation, measurement noise, and external disturbance. Therefore, both convergence and robustness properties of PO should be theoretically studied in the presence of gradient estimation error.

This work has been supported in part by the NSF grants CNS-2148309 and ECCS-2210320, and in part by the ARO Grant W911NF-24-1-0085.

L. Cui and Z. P. Jiang are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA (e-mail: l.cui@nyu.edu; zjiang@nyu.edu).

T. Başar is with the Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mail: basar1@illinois.edu).

Since linear quadratic regulator (LQR) is theoretically tractable and widely applied in various engineering fields, it stands out as providing an ideal benchmark for studying RL problems. For the LQR problem, the control policy is parameterized as a linear function of the state, i.e. $u_t = -Kx_t$. The corresponding performance index is $\mathcal{J}_{LQR}(K) = \sum_{t=0}^{\infty} \mathbb{E}(x_t^T Q x_t + u_t^T R u_t)$. PO for the LQR problem aims at solving the constrained optimization problem $\min_{K \in \mathcal{W}} \mathcal{J}_{LQR}(K)$, where $\mathcal{W}$ is the admissible set of stabilizing control policies. Since $\mathcal{J}_{LQR}(K)$ can be expressed in terms of a Lyapunov equation, which depends on K, $\mathcal{J}_{LQR}(K)$ is differentiable in K. Based on this result, standard gradient descent, natural policy gradient, and Newton gradient algorithms have been developed to minimize the performance index $\mathcal{J}_{LQR}(K)$ [2]–[7]. Interestingly, the Newton gradient algorithm with a step size of $\frac{1}{2}$ is equivalent to the celebrated Kleinman's policy iteration algorithm [8]–[11]. By the coercive property of the performance index $\mathcal{J}_{LQR}(K)$ ($\mathcal{J}_{LQR}(K) \to \infty$ as $K \to \partial\mathcal{W}$), the stability of the updated control policy is maintained during PO. Furthermore, the global linear convergence rate of the algorithms is theoretically demonstrated by the gradient dominance property which is shown in [4, Remark 2] and [5, Lemma 3]. One of the reasons for using PO in these model-based approaches (perhaps the most important one) is that it provides a natural pathway to model-free analysis, where the RL techniques come into play. For example, when the system model is unknown, zeroth-order methods are applied to approximate the gradient of the performance index, supported by a sample complexity analysis [4], [5], [7]. Since the estimation error of policy gradient is inevitable at each iteration, whether the error will accumulate and whether the adopted algorithms still converge in the presence of estimation error should be further studied. By considering the PO algorithm as a nonlinear discrete-time system and invoking input-to-state stability [12], the authors of [13], [14] show that the Kleinman's policy iteration algorithm can still find a near-optimal control policy even under the influence of estimation error. A similar robustness property is investigated for the steepest gradient descent algorithm [15].

The aforementioned PO for the LQR problem cannot guarantee the robustness of the closed-loop system. For example, the obtained controller may fail to stabilize the system in the presence of model mismatch that may be induced by the sim-to-real gap and parameter variation. Risk-sensitive linear quadratic Gaussian (LQG) control was first proposed by

[16], [17], which generalizes the risk-neutral optimal control (i.e. LQR) by minimizing the expectation of the accumulative quadratic cost transformed by the exponential function. It was shown in [18] and [19] that the risk-sensitive LQG control is equivalent to the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control and the linear quadratic zero-sum dynamic game. Therefore, it can guarantee the stability of the closed-loop system even under model mismatch. The authors of [20]–[22] proposed RL algorithms for solving the model-free risk-sensitive control, but the methods are only applicable to Markov decision processes whose state and action spaces are finite. In [23], [24], the authors proposed Q-learning algorithms for linear quadratic zero-sum dynamic games. The paper [25] proposed PO methods for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control to guarantee robust stability of the closed-loop system. Through the concept of implicit regularization, it was shown that the proposed PO algorithms can find the globally optimal solution of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control at globally sublinear and locally superlinear rates. As there is a fundamental connection between mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control and linear-quadratic zero-sum dynamic games (LQ ZSDGs), the natural policy gradient and Newton algorithms have been equivalently transformed into provably convergent dual-loop PO algorithms for the LQ ZSDG [26]–[28]. The outer loop is to learn a protagonist under the worst-case adversary while the inner loop is to learn a worst-case adversary. In this way, the protagonist can robustly perform the control tasks under the disturbances created by the adversary. Interestingly, the Newton algorithm with a step size of $\frac{1}{2}$ is equivalent to the policy iteration algorithm for ZSDG [23], [29], [30]. In the aforementioned papers, the convergence of the learning algorithms for the risk-sensitive control has been analyzed under the ideal noise-free case. Besides the convergence, a useful learning algorithm should be robust and is capable of finding a near-optimal solution even in the face of noise that may be induced by noisy experimental data, rounding errors of numerical computation, or the finite-time stopping of the inner loop in the dual-loop learning setup. However, the issues of uniform convergence and robustness of the dual-loop learning algorithm are still unsolved.

A. Our Contributions in this Paper

A fundamental challenge of the convergence of the dual-loop PO 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 increases, 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 has not been theoretically analyzed heretofore.

In addition, PO algorithm cannot be implemented accurately in practical applications, due to the influence of various errors arising from gradient estimation error, sensor noise, external disturbance, and modeling error. Hence, a fundamental question arises: Is the PO algorithm robust to the errors? In particular, does the PO algorithm still converge to a neighbourhood of the optimal solution 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 PO algorithm has not been fully understood in the present literature.

In this paper, we investigate uniform convergence and robustness of the dual-loop iterative algorithm for solving the problem of risk-sensitive linear quadratic Gaussian control. Even though the convergence of the dual-loop iterative algorithm is analyzed separately in [5], [25], [28], uniform convergence and 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 global linear convergence of the inner-loop iteration and find an upperbound on 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) [12] and its latest variant called "small-disturbance ISS" [13], [31] 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 PO algorithm still converges to a small neighbourhood of the optimal solution, as long as the noise is relatively small. Furthermore, based on these results and the technique of approximate dynamic programming [32], [33], 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 in this paper are threefold: 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 learning-based off-policy policy optimization algorithm is proposed.

A shorter and preliminary version of this paper was presented at the conference L4DC 2023 [34]. Compared with the conference paper, in this paper, we provide rigorous proofs for all the theoretical results. In addition, in Section V-A, we propose a method to learn an initial admissible controller. Finally, a benchmark example known as cart-pole system is provided to demonstrate the effectiveness of the proposed dual-loop algorithm.

B. Organization of the Paper

Following this Introduction section, Section II provides some preliminaries on linear exponential quadratic Gaussian (LEQG) control problem, linear quadratic zero-sum dynamic game, and robustness analysis. Section III introduces the model-based dual-loop iterative algorithm to optimize the policy for LEQG control, and the convergence of the algorithm is analyzed. Section IV analyzes robustness of the dual-loop iterative algorithm to various errors in the learning process within the framework of ISS. Section V presents a learning-based policy optimization algorithm for LEQG control. Section VI provides two numerical examples to illustrate the proposed algorithm. The paper ends with the concluding remarks of

Section VII and seven appendices which include proofs of the main results in the main body of the paper.

C. 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 a 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 $\|\cdot\|_2$. ℓ_∞ is the space of bounded sequences equipped with the norm $\|\cdot\|_\infty$. $\bar{\sigma}(\cdot)$ and $\underline{\sigma}(\cdot)$ are respectively the maximum and minimum singular values of a given matrix. 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 $\|G\|_{\mathcal{H}_\infty} := \sup_{u\in\ell_2} \frac{\|Gu\|_2}{\|u\|_2}$.

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 $P \in \mathbb{S}^n$, $\mathrm{vecs}(P) := [p_{1,1}, p_{1,2}, \cdots, p_{1,n}, p_{2,2}, p_{2,3}, \cdots, p_{n,n}]^T$, where $p_{i,j}$ is the ith row and jth column entry of the matrix P. $[X]_i$ dentotes the ith row of X. $[X]_{i,j}$ denotes the submatrix of the matrix X that is comprised of the rows between the ith and jth rows of X. For a vector $a \in \mathbb{R}^n$, $\mathrm{vecv}(a) := [a_1^2, 2a_1a_2, \cdots, 2a_1a_n, a_2^2, 2a_2a_3, \cdots, a_n^2]^T$. I_n denotes the n-dimensional identity matrix.

II. Preliminaries

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

A. Linear Exponential Quadratic Guassian Control

Consider the discrete-time linear time-invariant system

$$x_{t+1} = Ax_t + Bu_t + Dw_t \quad x_0 \sim \mathcal{N}(0, I_n), \tag{1a}$$

$$y_t = Cx_t + Eu_t, \tag{1b}$$

where $x_t \in \mathbb{R}^n$ is the state of the system; $u_t \in \mathbb{R}^m$ is the control input; $x_0 \in \mathbb{R}^n$ is the initial state; $w_t \in \mathbb{R}^q \sim \mathcal{N}(0, I_q)$ is independent and identically distributed random variable; $y_t \in \mathbb{R}^p$ is the controlled output. A, B, C, D, E are constant matrices with compatible dimensions. The LEQG control problem entails finding an input sequence $u := \{u_t\}_{t=0}^\infty$, depending on the current value of the state, that is $\{u_t = \mu_t(x_t)\}_{t=0}^\infty$ where $\mu := \{\mu_t : \mathbb{R}^n \to \mathbb{R}^m\}_{t=0}^\infty$ is a sequence of appropriately defined measurable control policies, such that the following risk-averse exponential quadratic cost is minimized

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

where γ is a positive constant. For proper formulation of the optimization problem, the following two assumptions are standard:

Assumption 1: (A, B) is stabilizable, $C^TC = Q \succ 0$, and $\gamma > \gamma_\infty$, where $\gamma_\infty > 0$ is the minimal value of γ such that for $\gamma > \gamma_\infty$ the solution to (6) given below exists, or equivalently there exists a control policy under which (2) is finite.

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

Assumptions 1 and 2 are used throughout the paper. Assumption 1 ensures the existence of a stabilizing solution to the LEQG control problem. As demonstrated in [19, Theorem 3.8], γ_∞ is finite. In addition, the positive definiteness of Q can be relaxed to $Q \succeq 0$, as long as (A, C) is taken to be detectable. Assumption 2 has two parts. The first, positive definiteness of the weighting matrix on control, is standard even in LQR. The second one is also 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 1, and K is stabilizing if $A - BK$ is stable. A feedback gain K is called admissible if it belongs to the admissible set $\mathcal{W}$ defined in (11).

As investigated by [16] and [35, Lemma 2.1], for any admissible linear control policy $\mu_t(x_t) = -Kx_t$, the cost admits the closed-form:

$$\mathcal{J}_{LEQG}(K) = -\gamma^2 \log\det(I_n - \gamma^{-2} P_K DD^T), \tag{3}$$

where the matrix $P_K = P_K^T \succ 0$ is the unique solution to

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

$$U_K := P_K + P_K D(\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K. \tag{4b}$$

Furthermore, the LEQG problem admits a unique optimal controller $u_t^* = -K^* x_t$, where

$$K^* = (R + B^T U^* B)^{-1} B^T U^* A. \tag{5}$$

with $P^* = (P^*)^T \succ 0$ the unique solution to the generalized algebraic Riccati equation (GARE)

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

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

B. Linear Quadratic Zero-Sum Dynamic Game

The dynamic game can be mathematically formulated as

$$\min_\mu \max_\nu \mathcal{J}_{DG}(\mu, \nu) := \mathbb{E}_{x_0}\left(\sum_{t=0}^{\infty} y_t^T y_t - \gamma^2 w_t^T w_t\right), \quad \text{subject to (1)}, \tag{7}$$

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, generated by state-feedback policies $\mu := \{\mu_t\}_{t=0}^\infty$ and $\nu := \{\nu_t\}_{t=0}^\infty$. Note that in (1a), w is no longer a Gaussian random sequence, but a second control variable, at the disposal of the maximizer.

For any admissible controller $\mu_t(x_t) = -Kx_t$, and with $\gamma > \gamma_\infty$, the closed-form cost is

$$\mathcal{J}_{DG}(K, \nu^*(K)) = \max_\nu \mathcal{J}_{DG}(-Kx_t, \nu) = \mathrm{Tr}(P_K), \tag{8}$$

where P_K is the solution of (4). From [19, Equation 3.51], it follows that the optimizers for the minimax problem are

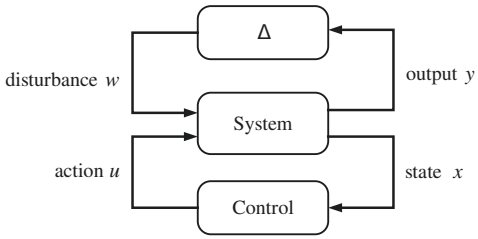

Fig. 1: Robust control design with model mismatch Δ.

$\mu_t^*(x_t) = -K^*x_t$ and $\nu_t^*(x_t) = L^*x_t$, where K^* is defined in (5) and L^* is given by

$$L^* = (\gamma^2 I_q - D^T P^* D)^{-1} D^T P^* (A - BK^*). \tag{9}$$

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

Therefore, the minimizer of DG shares the same optimal controller as the optimizer in the LEQG problem. Also note that, P_K is critical for determining the closed-from costs of $\mathcal{J}_{LEQG}(K)$ and $\mathcal{J}_{DG}(K, \nu^*)$.

C. Robustness Analysis

With w taken as a deterministic input in (1), and taking any stabilizing feedback $\mu_t(x_t) = -Kx_t$, the discrete-time transfer function from w to y can be expressed as

$$\mathcal{T}(K) := (C - EK)[zI_n - (A - BK)]^{-1}D. \tag{10}$$

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

Now consider the depiction in Fig. 1, where Δ denotes the model mismatch that may be induced by the sim-to-real gap, and satisfies $\|\Delta\|_{\mathcal{H}_\infty} \leq \frac{1}{\gamma}$. Thanks to the small-gain theorem [36]–[38], when subjected to model mismatch, the system remains stable as long as $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} < \gamma$. Consequently, the controller $\mu_t(x_t) = -Kx_t$ is robust to the model mismatch Δ if K lies within the admissible set $\mathcal{W}$ defined as

$$\mathcal{W} := \{K \in \mathbb{R}^{m\times n} | \rho(A - BK) < 1, \|\mathcal{T}(K)\|_{\mathcal{H}_\infty} < \gamma\}. \tag{11}$$

As investigated in [19, Theorem 3.8], the LEQG control in (5) satisfies $K^* \in \mathcal{W}$, and therefore, it is optimal with respect to (2) and robust to the model mismatch. This motivates us to pose the following problem.

Problem 1: Design a learning-based control algorithm such that near-optimal control gains, i.e. approximate values of K^*, can be learned from the input-state data.

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

III. Model-based Policy Optimization

In this section, by resorting to the PO method, a dual-loop iterative algorithm is proposed.

A. Introduction of the Outer Loop

The outer-loop iteration is developed based on the Newton PO algorithm in [35]. For any $K \in \mathcal{W}$, the gradient $\nabla_K \mathcal{J}(K) = \nabla_K \mathcal{J}_{LEQG}(K) = \nabla_K \mathcal{J}_{DG}(K, \nu^*(K))$ can be computed to be

$$\nabla_K \mathcal{J}(K) = 2[(R + B^T U_K B)K - B^T U_K A]\Sigma_K, \tag{12}$$

where

$$\Sigma_K = \sum_{t=0}^{\infty} [(A - BK + DL_{K,*})^T]^t D \tag{13}$$
$$(I_q - \gamma^{-2} D^T P_K D)^{-1} D^T (A - BK + DL_{K,*})^t.$$

and

$$L_{K,*} := (\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K (A - BK), \tag{14}$$

It is noticed that given $u_t = -Kx_t$, $L_{K,*}$ is considered as a worst-case feedback gain for w. [1] By the Newton's method with a step size of $\frac{1}{2}$, to iteratively minimize $\mathcal{J}(K)$, the updated feedback gain is

$$\begin{aligned} K' &= K - \frac{1}{2}(R + B^T U_K B)^{-1} \nabla_K \mathcal{J}(K) \Sigma_K^{-1} \\ &= (R + B^T U_K B)^{-1} B^T U_K A. \end{aligned} \tag{15}$$

Let i denote the iteration index for the outer loop and we introduce the following variables to simplify the notation

$$A_i := A - BK_i, \quad Q_i := Q + K_i^T R K_i, \tag{16a}$$

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, by (4) and (15), the outer-loop iteration can be expressed as

$$A_i^T U_i A_i - P_i + Q_i = 0, \tag{17a}$$
$$U_i := P_i + P_i D(\gamma^2 I_q - D^T P_i D)^{-1} D^T P_i, \tag{17b}$$
$$K_{i+1} = (R + B^T U_i B)^{-1} B^T U_i A. \tag{17c}$$

In (17), from the policy iteration perspective [33], we consider (17a) as the policy evaluation step under the worst-case disturbance and (17c) as the policy improvement step. P_i is the cost matrix of (7) for the controller $u_t = -K_i x_t$ under the worst-case disturbance.

As seen in Lemma 3, for each iteration, the controller K_i generated by (17) preserves robustness to model mismatch, i.e. $K_i \in \mathcal{W}$. P_i converges to P^* with a globally sublinear and locally quadratic rate [25, Theorems 4.3 and 4.4]. We further investigate the convergence rate of (17) and rigorously demonstrate that P_i monotonically converges to the optimal solution P^* with a globally linear convergence rate. The proof of the following theorem can be found in Appendix B.

Theorem 1: Let $\mathcal{H} := \{\mathrm{Tr}(P_K) - \mathrm{Tr}(P^*) | K \in \mathcal{W}\}$. For any $h \in \mathcal{H}$ and $K_1 \in \mathcal{G}_h$, where $\mathcal{G}_h = \{K \in \mathcal{W} | \mathrm{Tr}(P_K) \leq \mathrm{Tr}(P^*) + h\}$, there exists $\alpha(h) \in [0, 1)$, such that

$$\mathrm{Tr}(P_{i+1} - P^*) \leq \alpha(h)\, \mathrm{Tr}(P_i - P^*), \quad \forall i \in \mathbb{Z}_+. \tag{18}$$

[1] Henceforth, we write $\mu_t(x_t) = -Kx_t$ simply as $u_t = -Kx_t$, and likewise for w, as appropriate.

Since $P_{i+1}-P^* \succeq 0$ and $\|P_{i+1}-P^*\|_F \leq \mathrm{Tr}(P_{i+1}-P^*) \leq \sqrt{n}\|P_{i+1}-P^*\|_F$ (Lemma 2), the following inequality holds

$$\|P_{i+1}-P^*\|_F \leq \sqrt{n}\alpha^i(h)\|P_1-P^*\|_F. \tag{19}$$

Since $\mathcal{J}_{DG}(K_i,\nu^*(K_i)) = \mathrm{Tr}(P_i)$ from (8), it follows that

$$\begin{aligned}&\mathcal{J}_{DG}(K_{i+1},\nu^*(K_{i+1})) - \mathcal{J}_{DG}(\mu^*,\nu^*) \leq \\ &\alpha(h)[\mathcal{J}_{DG}(K_i,\nu^*(K_i)) - \mathcal{J}_{DG}(\mu^*,\nu^*)].\end{aligned} \tag{20}$$

Hence, the cost of the dynamic game (under the worst-case disturbance) converges to the saddle point at a linear convergence rate $\alpha(h) \in [0,1)$.

In the following subsection, given K_i, by maximizing $\mathcal{J}_{DG}(K_i,\nu)$ over ν, the inner-loop iteration is developed to get the worst-case disturbance $\nu^*(K_i)$.

B. Introduction of the Inner Loop

Given the feedback gain of the minimizer $K \in \mathcal{W}$, the inner loop iteratively finds the optimal controller for the maximizer $w^*(K)$ by solving [2]

$$\begin{aligned}&\max_w \mathcal{J}_{DG}(K,w) = \sum_{t=0}^{\infty} y_t^T y_t - \gamma^2 w_t^T w_t, \\ &\text{subject to } x_{t+1} = (A-BK)x_t + Dw_t, x_0 \sim \mathcal{N}(0,I_n).\end{aligned} \tag{21}$$

The optimal solution is $w_t^*(K) = L_{K,*}x_t$, where $L_{K,*}$ is defined in (14). For any admissible controller $w_t = Lx_t$ (with $(A-BK+DL)$ stable), by [5] the closed-form cost is

$$\mathcal{J}_{DG}(K,L) = \mathrm{Tr}(P_{K,L}), \tag{22}$$

where $P_{K,L} = P_{K,L}^T \succ 0$ is the solution to

$$\begin{aligned}&(A-BK+DL)^T P_{K,L}(A-BK+DL) \\ &- P_{K,L} + Q + K^T R K - \gamma^2 L^T L = 0.\end{aligned} \tag{23}$$

The gradient of $\nabla_L \mathcal{J}_{DG}(K,L)$ is

$$\begin{aligned}\nabla_L \mathcal{J}_{DG}(K,L) = 2\big[&(\gamma^2 I_q - D^T P_{K,L} D)L \\ &- D^T P_{K,L}(A-BK)\big]\Sigma_{K,L},\end{aligned} \tag{24}$$

where

$$\Sigma_{K,L} = \sum_{t=0}^{\infty}(A-BK+DL)^t[(A-BK+DL)^T]^t. \tag{25}$$

By Newton's method, the updated feedback gain is

$$\begin{aligned}L' &= L - \frac{1}{2}(\gamma^2 I_q - D^T P_{K,L} D)^{-1}\nabla \mathcal{J}_{DG}(K,L)\Sigma_{K,L}^{-1} \\ &= (\gamma^2 I_q - D^T P_{K,L} D)^{-1} D^T P_{K,L}(A-BK).\end{aligned} \tag{26}$$

Let j denote the iteration index of the inner loop, and we introduce the following variables to simplify the notation:

$$A_{i,j} := A - BK_i + DL_{i,j}, Q_i := Q + K_i^T R K_i, \tag{27a}$$

$$A_{i,*} := A - BK_i + DL_{i,*}, A^* := A - BK^* + DL^*. \tag{27b}$$

By (23) and (26), the inner loop is designed as

$$A_{i,j}^T P_{i,j} A_{i,j} - P_{i,j} + Q_i - \gamma^2 L_{i,j}^T L_{i,j} = 0, \tag{28a}$$

$$L_{i,j+1} = (\gamma^2 I_q - D^T P_{i,j} D)^{-1} D^T P_{i,j} A_i. \tag{28b}$$

[2]Here and below, we have used the control w instead of the policy ν, for simplicity of the notation.

From the policy iteration perspective, we consider (28a) as the policy evaluation step and (28b) as the policy improvement step for the inner loop. $P_{i,j}$ is the cost matrix of $\mathcal{J}_{DG}(K_i,L_{i,j})$ with the state-feedback policies $\mu_t(x_t) = -K_i x_t$ and $\nu_t(x_t) = L_{i,j}x_t$. The inner-loop policy iteration possesses the monotonicity property and preserves 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 stated in Lemma 10. We prove that the inner loop globally and linearly converges to the optimal solution P_i. The details of the proof are given in Appendix C. This brings us to the following theorem, whose proof is in Appendix C.

Theorem 2: Given $L_{i,1} = 0$, for any $K_i \in \mathcal{W}$, there exists a constant $\beta(K_i) \in [0,1)$, such that

$$\mathrm{Tr}(P_i - P_{i,j+1}) \leq \beta(K_i)\,\mathrm{Tr}(P_i - P_{i,j}), \quad \forall j \in \mathbb{Z}_+. \tag{29}$$

Based on Theorem 2, we can further obtain that

$$\|P_i - P_{i,j+1}\|_F \leq \sqrt{n}\beta^j(K_i)\|P_i - P_{i,1}\|_F. \tag{30}$$

In addition, since $\mathcal{J}_{DG}(K_i,L_{i,j}) = \mathrm{Tr}(P_{i,j})$ and $\mathcal{J}_{DG}(K_i,L_{i,*}) = \mathrm{Tr}(P_i)$, from Theorem 2, we have

$$\begin{aligned}&\mathcal{J}_{DG}(K_i,L_{i,*}) - \mathcal{J}_{DG}(K_i,L_{i,j+1}) \leq \\ &\beta(K_i)[\mathcal{J}_{DG}(K_i,L_{i,*}) - \mathcal{J}_{DG}(K_i,L_{i,j})].\end{aligned} \tag{31}$$

Hence, the cost of the dynamic game with K_i is monotonically increasing and converges to the maximum at a linear rate $\beta(K_i) \in [0,1)$.

The dual-loop policy iteration algorithm is in Algorithm 1. Ideally, $P_{i,j}$ generated by the inner loop converges to P_i, and then the control gain for the minimizer is updated to K_{i+1} by (17c). In practice, the inner loop stops after $\bar{j}$ iterations, and $P_{i,\bar{j}}$, instead of P_i, is used for updating the control gain at line 13. $K_{i,\bar{j}}$ denotes the control gain updated at the outer loop.

C. Convergence Analysis for the Dual-Loop Algorithm

For the dual-loop algorithm, the inner-loop iteration linearly converges to the optimal solution P_K with the rate dependent on K. Since K is updated iteratively, it is required that the inner loop enters the given neighborhood of P_K within a constant number of steps, regardless of K. Otherwise, as the outer-loop iteration proceeds, the required number of inner-loop iterations may grow explosively, thus making the dual-loop algorithm not practically implementable. The uniform convergence rate of the overall algorithm is given in the following theorem, whose proof is given in Appendix D.

Theorem 3: For any $h \in \mathcal{H}$, $K \in \mathcal{G}_h$, and $\epsilon > 0$, there exists $\bar{j}(h,\epsilon) \in \mathbb{Z}_+$ independent of K, such that for all $j \geq \bar{j}(h,\epsilon)$, $\|P_{K,j} - P_K\|_F \leq \epsilon$.

IV. Robustness Analysis for the Dual-Loop Algorithm

In the previous section, the exact PO algorithm was introduced in the sense that an accurate knowledge of system matrices was required to implement the algorithm. In practice, however, we do not have access to such an accurate model. Therefore, Algorithm 1 has to be implemented in a model-free setting. For example, we can approximate the gradient

Algorithm 1 Model-Based Policy Optimization

Initialize $K_{1,\bar{j}} \in \mathcal{W}$
Set $\bar{i}, \bar{j} \in \mathbb{Z}_+$
for $i \leq \bar{i}$ **do**
  Initialize $j = 1$ and $L_{i,1} = 0$
  $Q_{i,\bar{j}} = C^TC + K_{i,\bar{j}}^T R K_{i,\bar{j}}$
  for $j \leq \bar{j}$ **do**
    $A_{i,j} = A - BK_{i,\bar{j}} + DL_{i,j}$
    Get $P_{i,j}$ by solving (28a)
    Update $L_{i,j+1}$ by (28b)
    $j \leftarrow j + 1$
  end for
  $U_{i,\bar{j}} = P_{i,\bar{j}} + P_{i,\bar{j}} D(\gamma^2 I_q - D^T P_{i,\bar{j}} D)^{-1} D^T P_{i,\bar{j}}$
  $K_{i+1,\bar{j}} = (R + B^T U_{i,\bar{j}} B)^{-1} B^T U_{i,\bar{j}} A$
  $i \leftarrow i + 1$
end for

by zeroth-order method or approximate the value function (parameterized by cost matrix P_K) by approximate dynamic programming. The updates for the controllers in (17c) and (28b) are subjected to noise. 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.

A. Notions of Input-to-State Stability

Consider the general nonlinear discrete-time system

$$\chi_{k+1} = f(\chi_k, \rho_k), \tag{32}$$

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

Definition 1: [39] 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: [40] System (32) is ISS if there exist a $\mathcal{KL}$-function κ and a $\mathcal{K}$-function ξ such that for each input $\rho \in \ell_\infty$ and initial state $\chi_1 \in \mathcal{X}$, the following holds

$$\|\chi_k - \chi_e\| \leq \kappa(\|\chi_1 - \chi_e\|, k) + \xi(\|\rho\|_\infty). \tag{33}$$

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

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

B. Robustness Analysis for the Outer Loop

The exact outer loop iteration is shown in (17), and in the presence of disturbance, it is modified as

$$\hat{A}_i^T \hat{U}_i \hat{A}_i - \hat{P}_i + \hat{Q}_i = 0, \tag{34a}$$
$$\hat{U}_i = \hat{P}_i + \hat{P}_i D(\gamma^2 I_q - D^T \hat{P}_i D)^{-1} D^T \hat{P}_i, \tag{34b}$$
$$\hat{K}_{i+1} = (R + B^T \hat{U}_i B)^{-1} B^T \hat{U}_i A + \Delta K_{i+1}, \tag{34c}$$

where ΔK_i is the disturbance at the ith iteration, and "hat" is used to distinguish the sequences generated by the exact (17) and inexact (34) outer-loop iterations. By considering (34) as a discrete-time nonlinear system with the states $\hat{P}_i$ and inputs ΔK_i, it can be shown that (34) is inherently robust to ΔK in the sense of small-disturbance ISS [13], [14].

Theorem 4: Given $K_1 \in \mathcal{G}_h$, there exists a constant $d(h) > 0$, such that if $\|\Delta K\|_\infty < d(h)$, (34) is small-disturbance ISS. That is, there exist a $\mathcal{KL}$-function $\kappa_1(\cdot,\cdot)$ and a $\mathcal{K}$-function $\xi_1(\cdot)$, such that

$$\|\hat{P}_i - P^*\|_F \leq \kappa_1(\|\hat{P}_1 - P^*\|_F, i) + \xi_1(\|\Delta K\|_\infty). \tag{35}$$

We note that $\Delta K = \{\Delta K_i\}_{i=1}^\infty$ is a sequence of disturbance signals, and its ℓ_∞-norm is $\|\Delta K\|_\infty = \sup_{i \in \mathbb{Z}_+} \|\Delta K_i\|_F$.

If the outer-loop update in (17) can be computed exactly, which requires P_i to be exact, then the iterations of the exact update will not leave the admissible set $\mathcal{W}$. In contrast, the dual-loop algorithm (Algorithm 1) has access only to $P_{i,\bar{j}}$, which is close to P_i but not exact. There is the possibility that this inaccuracy could drive the outer-loop iteration away from the optimal solution or even beyond the admissible region $\mathcal{W}$. As a direct corollary to Theorems 3 and 4, we state below that Algorithm 1 can still find a near-optimal solution if $\bar{i}$ and $\bar{j}$ are large enough.

Corollary 1: For any $h \in \mathcal{H}$, $K_1 \in \mathcal{G}_h$, and $\epsilon > 0$, there exist $\bar{i}(h,\epsilon) \in \mathbb{Z}_+$ and $\bar{j}(h,\epsilon) \in \mathbb{Z}_+$, such that $\|P_{\bar{i},\bar{j}} - P^*\|_F < \epsilon$.

C. Robustness Analysis for the Inner Loop

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

$$\hat{A}_{i,j}^T \hat{P}_{i,j} \hat{A}_{i,j} - \hat{P}_{i,j} + \hat{Q}_i - \gamma^2 \hat{L}_{i,j}^T \hat{L}_{i,j} = 0, \tag{36a}$$
$$\hat{L}_{i,j+1} = (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T \hat{P}_{i,j} \hat{A}_i + \Delta L_{i,j+1}. \tag{36b}$$

Here, $\Delta L_{i,j+1}$ denotes the disturbance to the inner loop iteration and "hat" emphasizes that the corresponding sequences are generated by the inexact iteration. With the inexact inner loop at hand, the following theorem shows that the inner-loop iteration (36) is robust to disturbance ΔL_i in the sense of small-disturbance input-to-state stability [13], [14].

Theorem 5: For any $\hat{K}_i \in \mathcal{W}$, there exists a constant $e(\hat{K}_i) > 0$, such that if $\|\Delta L_i\|_\infty < e(\hat{K}_i)$, (36) is small-disturbance ISS. That is, there exist a $\mathcal{KL}$-function $\kappa_2(\cdot,\cdot)$ and a $\mathcal{K}$-function $\xi_2(\cdot)$, such that

$$\|\hat{P}_{i,j} - \hat{P}_i\|_F \leq \kappa_2(\|\hat{P}_{i,1} - \hat{P}_i\|_F, j) + \xi_2(\|\Delta L_i\|_\infty). \tag{37}$$

We note that $\Delta L_i = \{\Delta L_{i,j}\}_{j=1}^\infty$ is a sequence of disturbance signals, and $\|\Delta L_i\|_\infty = \sup_{j \in \mathbb{Z}_+} \|\Delta L_{i,j}\|_F$. The results of these last two theorems guarantee robustness of the dual-loop PO algorithm. Literally speaking, when the dual-loop PO 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^* ($\hat{P}_i$).

V. Learning-based Off-Policy Policy Optimization

We will develop a learning-based algorithm to learn from data a robust suboptimal controller (i.e. an approximation of K^*) without requiring any accurate knowledge of (A, B, D) under the setting of zero-sum dynamic game with additive Gaussian noise. The input-state data from the following system will be utilized for learning:

$$x_{t+1} = Ax_t + Bu_t + Dw_t + v_t, \tag{38}$$

where $v_t \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \Sigma^T \succeq 0$ denotes independent and identically distributed noise.

A. Learning-Based Policy Optimization

Suppose that the exploratory policies for the minimizer and maximizer are

$$u_t = -\hat{K}_{exp}x_t + \sigma_1\xi_1, \qquad \xi_1 \sim \mathcal{N}(0, I_m), \tag{39a}$$
$$w_t = \hat{L}_{exp}x_t + \sigma_2\xi_1, \qquad \xi_2 \sim \mathcal{N}(0, I_q), \tag{39b}$$

where $\hat{K}_{exp} \in \mathbb{R}^{m\times n}$ and $\hat{L}_{exp} \in \mathbb{R}^{q\times n}$ are exploratory feedback gains, and $\sigma_1, \sigma_2 > 0$ are the standard deviations of the exploratory noise. Let $z_t := [x_t^T, u_t^T, w_t^T]^T$. For any given matrices $X \in \mathbb{S}^n$, let

$$\begin{aligned}\Gamma(X) &:= \begin{bmatrix}\Gamma_{xx}(X) & \Gamma_{xu}(X) & \Gamma_{xw}(X)\\ \Gamma_{ux}(X) & \Gamma_{uu}(X) & \Gamma_{uw}(X)\\ \Gamma_{wx}(X) & \Gamma_{wu}(X) & \Gamma_{ww}(X)\end{bmatrix}\\ &= \begin{bmatrix}A^TXA+Q & A^TXB & A^TXD\\ B^TXA & B^TXB+R & B^TXD\\ D^TXA & D^TXB & D^TXD-\gamma^2I_q\end{bmatrix}.\end{aligned} \tag{40}$$

Along the trajectories of system (38), $x_{t+1}^TXx_{t+1}$ can be computed to be

$$\begin{aligned}x_{t+1}^TXx_{t+1} = z_t^T\Gamma(X)z_t - r_t + v_t^TXv_t\\ + 2v_t^TX(Ax_t + Bu_t + Dw_t).\end{aligned} \tag{41}$$

where $r_t := x_t^TQx_t + u_t^TRu_t - \gamma^2w_t^Tw_t$ is the stage cost of the zero-sum dynamic game. Taking the expectation of (41) results in

$$\mathbb{E}\left[z_t^T\Gamma(X)z_t + \mathrm{Tr}(\Sigma X) - r_t - x_{t+1}^TXx_{t+1}|z_t\right] = 0. \tag{42}$$

To simplify the notations, let

$$\bar{z}_t := [\mathrm{vecv}(z_t)^T, 1]^T \tag{43}$$
$$\theta(X) := [\mathrm{vecs}(\Gamma(X))^T, \mathrm{Tr}(\Sigma X)]^T. \tag{44}$$

By pre-multiplying (42) with $\bar{z}_t$, one can obtain

$$\mathbb{E}\left[\bar{z}_t\bar{z}_t^T\theta(X) - \bar{z}_tr_t - \bar{z}_t\,\mathrm{vecv}(x_{t+1})^T\,\mathrm{vecs}(X)|z_t\right] = 0. \tag{45}$$

Assumption 3: There exists an ergodic stationary probability measure π on $\mathbb{R}^{n+m+q}$ for system (38) with controller (39).

Remark 1: Assumption 3 is widely used in approximate dynamic programming and in the RL literature [41], [42].

Assumption 4: $\mathbb{E}_\pi\left[z_tz_t^T\right]$ and $\mathbb{E}_\pi\left[\bar{z}_t\bar{z}_t^T\right]$ are invertible.

Remark 2: Assumption 4 is reminiscent of the persistent excitation (PE) condition [43], [44]. As in the literature of data-driven control [10], [45], one can satisfy it by means of added exploration noise, such as sinusoidal signals or random noise.

Under Assumptions 3 and 4, taking the expectation of (45) with respect to the invariant probability measure π, we have

$$\theta(X) = \Phi^\dagger\Xi\,\mathrm{vecs}(X) + \Phi^\dagger\Psi. \tag{46}$$

where

$$\begin{aligned}&\Phi := \mathbb{E}_\pi\left[\bar{z}_t\bar{z}_t^T\right], \quad \Xi := \mathbb{E}_\pi\left[\bar{z}_t\,\mathrm{vecv}(x_{t+1})^T\right],\\ &\Psi := \mathbb{E}_\pi\left[\bar{z}_tr_t\right].\end{aligned} \tag{47}$$

Therefore, $\Gamma_{ww}(X)$ and $\Gamma(X)$ can be reconstructed as

$$\mathrm{vecs}(\Gamma_{ww}(X)) = [\Phi^\dagger]_{n_1,n_2}\Xi\,\mathrm{vecs}(X) + [\Phi^\dagger]_{n_1,n_2}\Psi, \tag{48a}$$
$$\mathrm{vecs}(\Gamma(X)) = [\Phi^\dagger]_{1,n_2}\Xi\,\mathrm{vecs}(X) + [\Phi^\dagger]_{1,n_2}\Psi, \tag{48b}$$

where

$$\begin{aligned}n_1 &= \frac{(n+m+q)(n+m+q+1)}{2} + 1 - \frac{(q+1)q}{2},\\ n_2 &= \frac{(n+m+q)(n+m+q+1)}{2}.\end{aligned} \tag{49}$$

In practice, we use a finite number of trajectory samples to estimate Φ, Ξ, and Ψ, that is,

$$\begin{aligned}&\hat{\Phi}_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau}\bar{z}_t\bar{z}_t^T, \quad \hat{\Xi}_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau}\bar{z}_t\,\mathrm{vecv}(x_{t+1})^T,\\ &\hat{\Psi}_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau}\bar{z}_tr_t.\end{aligned} \tag{50}$$

By the Birkhoff Ergodic Theorem [46, Theorem 16.2], the following relations hold *almost surely*

$$\lim_{\tau\to\infty}\hat{\Phi}_\tau = \Phi, \quad \lim_{\tau\to\infty}\hat{\Xi}_\tau = \Xi, \quad \lim_{\tau\to\infty}\hat{\Psi}_\tau = \Psi. \tag{51a}$$

Then, by (48b), $\Gamma(X)$ is estimated by a data-driven approach as follows:

$$\mathrm{vecs}(\hat{\Gamma}(X)) = [\hat{\Phi}_\tau^\dagger]_{1,n_2}\hat{\Xi}_\tau\,\mathrm{vecs}(X) + [\hat{\Phi}_\tau^\dagger]_{1,n_2}\hat{\Psi}_\tau, \tag{52}$$

With the data-driven estimate $\hat{\Gamma}(X)$, we will transform model-based PO in Algorithm 1 to a learning-based algorithm. Considering (40), we can rewrite (28a) as

$$[I_n, -K_i^T, L_{i,j}^T]\Gamma(P_{i,j})[I_n, -K_i^T, L_{i,j}^T]^T - P_{i,j} = 0. \tag{53}$$

Vectorizing (53) and plugging (46) into (53) result in

$$\begin{aligned}&\left\{\left([I_n, -K_i^T, L_{i,j}^T]\otimes[I_n, -K_i^T, L_{i,j}^T]\right)T_{n+m+q}\right.\\ &\left.[\Phi^\dagger]_{1,n_1}\Xi - T_n\right\}\mathrm{vecs}(P_{i,j}) =\\ &-\left([I_n, -K_i^T, L_{i,j}^T]\otimes[I_n, -K_i^T, L_{i,j}^T]\right)T_{n+m+q}[\Phi^\dagger]_{1,n_1}\Psi,\end{aligned} \tag{54}$$

where T_n and T_{n+m+q} are the duplication matrices defined in [47, pp. 56]. One can view (54) as a linear equation with respect to $\mathrm{vecs}(P_{i,j})$. Hence, at each inner-loop iteration, $\mathrm{vecs}(P_{i,j})$ can be computed by solving (54). With $P_{i,j}$, according to (28b) and (40), the feedback gain of the maximizer can be updated as

$$L_{i,j+1} = -\Gamma_{ww}(P_{i,j})^{-1}(\Gamma_{wx}(P_{i,j}) - \Gamma_{wu}(P_{i,j})K_i). \tag{55}$$

Next, we will develop a data-driven approach for the outer-loop iteration. According to the expression of $\Gamma_{ww}(X)$ in (40), and the duplication matrices defined in [47, pp. 56], we have

$$\operatorname{vecs}(\Gamma_{ww}(X)) = T_q^\dagger (D^T \otimes D^T) T_n \operatorname{vecs}(X) - \gamma^2 T_q^\dagger \operatorname{vec}(I_q). \tag{56}$$

Since (48a) and (56) hold for any $X \in \mathbb{S}^n$, it follows that

$$T_q^\dagger (D^T \otimes D^T) T_n = [\Phi^\dagger]_{n_1,n_2} \Xi. \tag{57}$$

Then, from (57), we obtain

$$(T_q^T T_q)[\Phi^\dagger]_{n_1,n_2} \Xi (T_n^T T_n)^{-1} = T_q^T (D^T \otimes D^T)(T_n^\dagger)^T \tag{58}$$

For any $Y \in \mathbb{S}^q$, let

$$\Omega(Y) = DYD^T. \tag{59}$$

According to the duplication matrices defined in [47, pp. 56] and (58), it holds

$$\begin{aligned} \operatorname{vecs}(\Omega(Y)) &= T_n^\dagger (D \otimes D) T_q \operatorname{vecs}(Y) \\ &= (T_n^T T_n)^{-1} \Xi^T [\Phi^\dagger]_{n_1,n_2}^T (T_q^T T_q) \operatorname{vecs}(Y). \end{aligned} \tag{60}$$

Now, $\Omega(Y)$ can be computed by (60) without knowing D. By (17c), the feedback gain for the minimizer is updated as

$$\begin{aligned} U_i &= P_i - P_i \Omega(\Gamma_{ww}(P_i)^{-1}) P_i, \\ K_{i+1} &= \Gamma_{uu}(U_i)^{-1} \Gamma_{ux}(U_i). \end{aligned} \tag{61}$$

The learning-based PO algorithm is given in the table labeled as Algorithm 2. It should be noticed that in Algorithm 2, the system matrices (A, B, and D) are not involved in computing $\operatorname{vecs}(P_{i,j})$, $L_{i,j+1}$ and K_{i+1}. In addition, the updated controller K_i is not applied for the data collection. Therefore, it is a learning-based off-policy algorithm for PO.

B. Learning an Initial Admissible Controller

In Algorithm 2, an initial admissible feedback gain is required to start the learning-based PO algorithm. In this section, we will develop a data-driven method for learning such an initial feedback gain.

Taking the expectation of (38) with respect to the invariant probability measure π in Assumption 3, we have

$$\mathbb{E}_\pi \left[x_{t+1}^T - z_t^T [A, B, D]^T \right] = 0. \tag{62}$$

Pre-multiplying (62) by z_t and using Assumption 4 yield

$$[A, B, D]^T = (\Phi')^\dagger \Xi', \tag{63}$$

where

$$\Phi' = \mathbb{E}_\pi \left[z_t z_t^T \right], \quad \Xi' = \mathbb{E}_\pi \left[z_t x_{t+1} \right]. \tag{64}$$

In practice, we can utilize a finite number of trajectory samples to estimate Φ' and Ξ', i.e.

$$\hat{\Phi}'_\tau := \frac{1}{\tau} \sum_{t=1}^{\tau} z_t z_t^T, \quad \hat{\Xi}'_\tau := \frac{1}{\tau} \sum_{t=1}^{\tau} z_t (x_{t+1})^T. \tag{65}$$

By the Birkhoff Ergodic Theorem [46, Theorem 16.2], the following relations hold *almost surely*

$$\lim_{\tau \to \infty} \hat{\Phi}'_\tau = \Phi', \quad \lim_{\tau \to \infty} \hat{\Xi}'_\tau = \Xi'. \tag{66}$$

Algorithm 2 Learning-Based Policy Optimization

1: Initialize $\hat{K}_1 \in \mathcal{W}$
2: Set $\bar{i}, \bar{j} \in \mathbb{Z}_+$.
3: Set the length of the sampled trajectory τ, and the exploration variances σ_1^2 and σ_2^2
4: Collect data from (38) with exploratory input (39)
5: Construct $\hat{\Phi}_\tau$, $\hat{\Xi}_\tau$, and $\hat{\Psi}_\tau$ defined in (50)
6: Get the expressions of $\hat{\Gamma}(X)$ and $\hat{\Omega}(X)$ by (52) and (60).
7: **for** $i \le \bar{i}$ **do**
8: Set $\hat{L}_{i,1} = 0$
9: **for** $j \le \bar{j}$ **do**
10: Get $\hat{P}_{i,j}$ by solving (54)
11: Update $\hat{L}_{i,j+1}$ by (55)
12: $j \leftarrow j + 1$
13: **end for**
14: Update $\hat{K}_{i+1}$ by (61)
15: $i \leftarrow i + 1$
16: **end for**

By Assumption 4, (66) and the definition of limit, if τ is large enough, $\hat{\Phi}'_\tau$ is invertible *almost surely*. As a result, we obtain the estimates of the system matrices

$$[\hat{A}_\tau, \hat{B}_\tau, \hat{D}_\tau]^T = (\hat{\Phi}'_\tau)^\dagger \hat{\Xi}_\tau. \tag{67}$$

With the identified system matrices, the linear matrix inequalities (LMIs) in (68) can be solved for an initial admissible controller design. The admissibility of the obtained initial controller is guaranteed by Theorem 6 below, which is proved in Appendix G.

Theorem 6: There exist $\epsilon > 0$, $\mu > 0$, and $\tau^*(\epsilon, \mu) > 0$, such that for any $\tau > \tau^*(\epsilon, \mu)$, the following LMIs

$$\begin{bmatrix} -W & * & * & * \\ 0 & -\gamma^2 I_q & * & * \\ \hat{A}_\tau W - \hat{B}_\tau V & \hat{D}_\tau & -W & * \\ CW - EV & 0 & 0 & -I_p \end{bmatrix} \prec -\epsilon I, \tag{68a}$$

$$- \begin{bmatrix} I_n & * & * \\ \mu W & I_n & * \\ \mu V & 0 & I_m \end{bmatrix} \prec 0, \tag{68b}$$

have a solution and $K = VW^{-1}$ that belongs to $\mathcal{W}$ *almost surely*.

The following corollary shows that whenever the LMIs (68) have a solution, the controller derived is admissible, which can be directly obtained from the proof of Theorem 6.

Corollary 2: For any $\epsilon > 0$, any $\mu > 0$, and any $\tau > \tau^*(\epsilon, \mu)$, if the LMIs (68) have a solution, then $K = VW^{-1}$ belongs to $\mathcal{W}$ *almost surely*.

VI. Numerical Simulations

A. An Illustrative Example

We apply Algorithms 1 and 2 to the system studied in [48] with the same A and B matrices as there. The matrices related to the output are $C = [I_3, 0_{3\times 3}]^T$ and $E = [0_{3\times 3}, I_3]^T$. The $\mathcal{H}_\infty$ norm threshold is $\gamma = 5$. $\bar{i} = 10$ and $\bar{j} = 20$.

The robustness of Algorithm 1 in the presence of disturbance is validated first. For each outer and inner loop

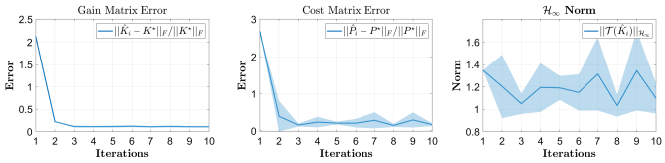

Fig. 2: Robustness of Algorithm 1 when $\|\Delta K\|_\infty = 0.09$ and $\|\Delta L_i\|_\infty = 0.09$.

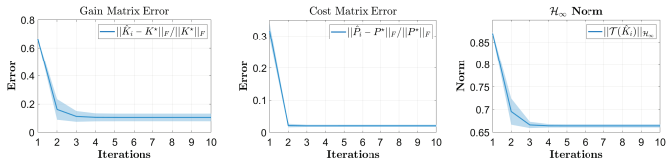

Fig. 3: Using Algorithm 2, the solutions of each iteration approach the optimal solution, and the $\mathcal{H}_\infty$ norm is smaller than the threshold.

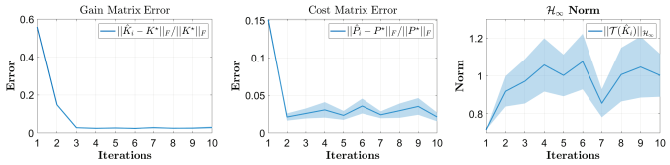

Fig. 4: For the cart-pole system, the robustness of Algorithm 1 when $\|\Delta K\|_\infty = 0.7$ and $\|\Delta L_i\|_\infty = 0.1$.

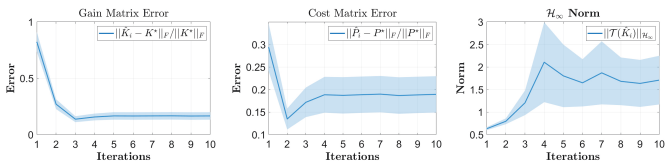

Fig. 5: For the cart-pole system, as the iteration of Algorithm 2 proceeds, the gain and cost matrices approach the optimal solution, and the $\mathcal{H}_\infty$ norm is smaller than the threshold.

iteration, the entries of the disturbances ΔK_i and $\Delta L_{i,j}$ are taken as samples from a standard Gaussian distribution and then their Frobenius norms are normalized to 0.09. The algorithm is run independently for 50 times. The results are shown in Fig. 2, where the bold line is the mean of the trials and the shaded region denotes the variance. It is seen that with the disturbances at both the outer and inner loop iterations, the generated controller and the corresponding cost matrix approach to the optimal solution and finally enter a neighborhood of the optimal controller K^* and cost matrix P^*, respectively. The $\mathcal{H}_\infty$ norm of the closed-loop system is smaller than the threshold throughout the policy optimization. These numerical results are consistent with the developed theoretical results in Theorems 4 and 5.

Algorithm 2 is implemented independently for 50 trials to validate its performance. The length of the trajectory samples is $\tau = 5000$. The standard deviation of the system additive noise is $\Sigma = I_3$. The standard deviation of the exploratory noise is $\sigma_1 = \sigma_2 = 1$. The relative errors of the gain matrix and cost matrix are shown in Fig. 3. The proposed off-policy RL algorithm can still approximate the optimal solution when the system is disturbed by an additive Gaussian noise.

B. Cart-Pole Example

We next consider the cart-pole system in [49], where the inverted pendulum is hinged to the top of a wheeled cart that moves along a straight line. The system is discretized under the sampling period $\Delta t = 0.01sec$. Considering the noise for the system, the system can be described by a linear system with $x = [s, \dot{s}, \phi, \dot{\phi}]^T$,

$$A = \begin{bmatrix} 1 & 0.01 & 0 & 0 \\ 0 & 1 & -0.01 & 0 \\ 0 & 0 & 1 & 0.01 \\ 0 & 0 & 0.16 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0.01 \\ 0 \\ -0.015 \end{bmatrix},$$

$C = [I_4, 0_{1\times 4}]^T$, $D = 0.001 I_4$, $E = [0_{1\times 4}, 1]$. The $\mathcal{H}_\infty$ norm threshold is $\gamma = 10$.

The robustness evaluation of Algorithm 1 is shown in Fig. 4, where the mean-variance curves are plotted for 50 independent trials. At each iteration, the entries of the disturbances ΔK_i and $\Delta L_{i,j}$ are randomly sampled from a standard Gaussian distribution, and then $\|\Delta K_i\|_F$ and $\|\Delta L_{i,j}\|_F$ are normalized to 0.7 and 0.1, respectively. The relative errors approach zero even in the presence of the disturbance, demonstrating the small-disturbance ISS properties of the outer and inner loops in Theorems 4 and 5. In addition, the $\mathcal{H}_\infty$ norm of the system is below the given threshold, and the robustness of the closed-loop system is guaranteed during the iteration.

When the matrices (A, B, D) are unknown, Algorithm 2 is implemented independently for 50 times. The length of the sampled trajectory is $\tau = 10000$. The standard deviation of the system additive noise is $\Sigma = 0.1 I_4$. The standard deviation of the exploratory noise is $\sigma_1 = \sigma_2 = 20$. It is seen from Fig. 5 that using the noisy data, the learning-based PO developed in Algorithm 2 still finds a near-optimal solution.

VII. Conclusions

In this paper, we have proposed a novel dual-loop policy optimization algorithm for data-driven risk-sensitive linear quadratic Gaussian control whose convergence and robustness properties have been analyzed. We have shown that the iterative algorithm possesses the property of small-disturbance input-to-state stability, that is, starting from any initial admissible controller, the solutions of the proposed policy optimization algorithm ultimately enter a neighbourhood of the optimal solution, given that the disturbance is relatively small. Based on these model-based theoretical results, when the accurate system knowledge is unavailable, we have also proposed a novel off-policy policy optimization RL algorithm to learn from data robust optimal controllers. Numerical examples are provided and the efficacy of the proposed methods are demonstrated by an academic example and a benchmark cart-pole system. Future work will be directed toward investigating the input-to-state stability of standard gradient descent and natural gradient descent algorithms. In addition, by invoking small-gain theory, the application of learning-based risk-sensitive LQG control to decentralized control design of complex systems will be studied.

APPENDIX A: USEFUL AUXILIARY RESULTS

Some fundamental lemmas are introduced to assist in the development of the results in the main body of the paper.

Lemma 1 (Bounded Real Lemma): Given a matrix $K \in \mathbb{R}^{m\times n}$ and the transfer function $\mathcal{T}(K)$ for the linear system (1) under Assumptions 1 and 2, the following statements are equivalent:

1) K is stabilizing and $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} < \gamma$;

2) The algebraic Riccati equation (4) admits a unique stabilizing solution $P_K = P_K^T \succ 0$ 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 $P_K = P_K^T \succ 0$ such that

$$I_q - \gamma^{-2} D^T P_K D \succ 0, \tag{A.1}$$
$$(A - BK)^T U_K (A - BK) - P_K + Q + K^T R K \prec 0.$$

4) There exist $W = W^T \succ 0$ and $V = KW$, such that

$$\begin{bmatrix} -W & * & * & * \\ 0 & -\gamma^2 I_q & * & * \\ AW - BV & D & -W & * \\ CW - EV & 0 & 0 & -I_p \end{bmatrix} \prec 0. \tag{A.2}$$

Proof: The first three statements are from [35, Lemma 2.7]. The last one follows from Schur complement lemma. ■

Lemma 2: For any positive semi-definite matrix $P \in \mathbb{S}^n$, $\|P\|_F \le \mathrm{Tr}(P) \le \sqrt{n}\|P\|_F$ and $\|P\| \le \mathrm{Tr}(P) \le n\|P\|$.

Proof: Let $\sigma_1 \ge \cdots \ge \sigma_n$ denote 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, the lemma holds by noting that $\sum_{i=1}^n \sigma_i^2 \le (\sum_{i=1}^n \sigma_i)^2 \le n\sum_{i=1}^n \sigma_i^2$ and $\sigma_1 \le \sum_{i=1}^n \sigma_i \le n\sigma_1$. ■

APPENDIX B: PROOF OF THEOREM 1

The following lemma shows that the cost matrix P_i generated by the outer-loop iteration (17) is monotonically decreasing and all the updated feedback gains are admissible given an initial admissible feedback gain.

Lemma 3: 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 [35, Theorem 4.3 and Theorem 4.6]. The statement 2) follows from [35, Equation (5.27)]. ■

Lemmas 4-8 provide the preliminaries for Theorem 1, and further details on their proofs can be found in [50].

Lemma 4: For any $K \in \mathcal{W}$ and $K' := (R + B^T U_K B)^{-1} B^T U_K A$, we have

$$K' - K^* = R^{-1} B^T (\Lambda_K)^{-T} (P_K - P^*) A^*, \tag{B.1}$$
$$\Lambda_K := I_n + BR^{-1}B^T P_K - \gamma^{-2} D D^T P_K. \tag{B.2}$$

Proof: By the expression of U_K in (4b), we have

$$(R + B^T U_K B)^{-1} B^T U_K A = R^{-1} B^T P_K \Lambda_K^{-1} A. \tag{B.3}$$

Using the expressions of K' and K^* in (5), and noting that $A^* = (\Lambda^*)^{-1} A$, we can obtain (B.1). ■

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

Lemma 5: For any $K \in \mathcal{W}$, $(U_K - U^*)$ satisfies

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

Proof: The variable U^* in (6) can be rewritten as

$$\begin{aligned} U^* &= -(I_n - \gamma^{-2} P^* D D^T)^{-1} (P_K - P^*)(I_n - \gamma^{-2} D D^T P^*)^{-1} \\ &+ (I_n - \gamma^{-2} P^* D D^T)^{-1} (P_K - \gamma^{-2} P^* D D^T P^*) \\ &(I_n - \gamma^{-2} D D^T P^*)^{-1} \end{aligned} \tag{B.5}$$

To simplify the notation, define S_K as

$$\begin{aligned} S_K &:= U_K - (I_n - \gamma^{-2} P^* D D^T)^{-1} (P_K - \gamma^{-2} P^* D D^T P^*) \\ &\quad (I_n - \gamma^{-2} D D^T P^*)^{-1}. \end{aligned} \tag{B.6}$$

By noting $U_K - P_K = P_K D (\gamma^2 I_q - D^T P_K D)^{-1} D^T P_K$, completing the squares, and using the matrix inversion lemma, it can be verified that

$$\begin{aligned} &(I_n - \gamma^{-2} P^* D D^T) S_K (I_n - \gamma^{-2} D D^T P^*) \\ &= (P_K - P^*) D (\gamma^2 I_q - D^T P_K D)^{-1} D^T (P_K - P^*). \end{aligned} \tag{B.7}$$

The proof is thus completed by following (B.5) and (B.7). ■

Lemma 6: For any $K \in \mathcal{W}$, P_K is continuous with respect to K, where P_K is the unique positive-definite solution to (4).

Proof: Consider the function

$$\begin{aligned} F(K, P_K) &:= (A - BK)^T U_K (A - BK) - P_K \\ &\quad + Q + K^T R K. \end{aligned} \tag{B.8}$$

Let $\mathcal{F}(\mathrm{vec}(K), \mathrm{vec}(P_K)) := \mathrm{vec}(F(K, P_K))$. By following [51, Theroem 9] and [35, Equation (B.10)], we have

$$\begin{aligned} \frac{\partial \mathcal{F}(\mathrm{vec}(K), \mathrm{vec}(P_K))}{\partial\, \mathrm{vec}(P_K)} &= [(A - BK + DL_{K,*})^T \\ \otimes (A - BK + DL_{K,*})^T] - I_{n^2}. \end{aligned} \tag{B.9}$$

Since $K \in \mathcal{W}$, by Lemma 1, $(A - BK + DL_{K,*})$ is stable. Since $\bar{\sigma}(A - BK + DL_{K,*}) < 1$, $\frac{\partial \mathcal{F}(\mathrm{vec}(K), \mathrm{vec}(P_K))}{\partial\, \mathrm{vec}(P_K)}$ is invertible. By the implicit function theorem, P_K is continuous with respect to K for any $K \in \mathcal{W}$. ■

Lemma 7: Let $K \in \mathcal{W}$, P_K be the positive-definite solution of (4a), and $K' = (R + B^T U_K B)^{-1} B^T U_K A$. Then, $(K - K')^T R (K - K') = 0$ is equivalent to $K = K^*$.

Proof: ($\Rightarrow$): $(K - K')^T R (K - K') = 0$ implies $K = K'$. It follows from (4a) that $P_K = P_K^T \succ 0$ is the solution of (6). Due to the uniqueness of the positive-definite solution to (6), it is deduced that $P_K = P^*$ and $K = K^*$.

($\Leftarrow$): Since $K = K^*$ and $P_K = P^*$, it follows that $K' = (R + B^T U_K B)^{-1} B^T U_K A = K$. Hence, we have $K = K'$. ■

Lemma 8: For any $h \in \mathcal{H}$, $\mathcal{G}_h := \{K \in \mathcal{W} | \mathrm{Tr}(P_K) \le \mathrm{Tr}(P^*) + h\}$ is compact.

Proof: It can be checked that $\mathcal{G}_h$ is bounded and closed. The compactness of $\mathcal{G}_h$ is demonstrated using the Heine–Borel theorem [52, p. 81]. ■

Lemma 9: For any $h \in \mathcal{H}$ and $K \in \mathcal{G}_h$, 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 $a(h) > 0$, such that

$$\|P_K - P^*\|_F \le a(h) \|E_K\|_F. \tag{B.10}$$

Proof: It follows from (4) that

$$\begin{aligned}&(A-BK^*)^TU_K(A-BK^*)-P_K+Q+K^TRK\\&+A^TU_KB(K^*-K)+(K^*-K)^TB^TU_KA\\&+K^TB^TU_KBK-(K^*)^TB^TU_KBK^*=0.\end{aligned}\tag{B.11}$$

Subtracting (6a) from (B.11), and considering $B^TU_KA=(R+B^TU_KB)K'$, we have

$$\begin{aligned}&(A-BK^*)^T(U_K-U^*)(A-BK^*)-(P_K-P^*)+E_K\\&-(K'-K^*)^T(R+B^TU_KB)(K'-K^*)=0.\end{aligned}\tag{B.12}$$

It follows from $(I_n-\gamma^{-2}DD^TP^*)^{-1}(A-BK^*)=(A-BK^*+DL^*)=A^*$ and Lemma 5 that

$$\begin{aligned}&(A^*)^T\Delta P_K(A^*)-\Delta P_K+E_K\\&-(K'-K^*)^T(R+B^TU_KB)(K'-K^*)\\&+(A^*)^T\Delta P_KD(\gamma^2I_q-D^TP_KD)^{-1}D^T\Delta P_KA^*=0,\end{aligned}\tag{B.13}$$

where $\Delta P_K:=P_K-P^*$. According to [53, Theorem 5.D6], we have

$$\begin{aligned}\Delta P_K\preceq\sum_{t=0}^{\infty}(A^*)^{T,t}\big[&E_K+(A^*)^T\Delta P_KD\\&(\gamma^2I_q-D^TP_KD)^{-1}D^T\Delta P_KA^*\big](A^*)^t.\end{aligned}\tag{B.14}$$

Taking the trace of (B.14) and using the cyclic property of trace and trace inequality in [54, Lemma 1] yield

$$\begin{aligned}&\operatorname{Tr}(\Delta P_K)\le\operatorname{Tr}\left[\sum_{t=0}^{\infty}(A^*)^t(A^*)^{T,t}E_K\right]+\operatorname{Tr}\left[\sum_{t=1}^{\infty}(A^*)^t\right.\\&\left.(A^*)^{T,t}\Delta P_KD(\gamma^2I_q-D^TP_KD)^{-1}D^T\Delta P_K\right]\\&\le a_1\operatorname{Tr}(E_K)+a_2\gamma^{-2}\operatorname{Tr}[\Delta P_K\\&(I_n-\gamma^{-2}DD^TP_K)^{-1}DD^T\Delta P_K],\end{aligned}\tag{B.15}$$

where

$$a_1=\|\sum_{k=0}^{\infty}(A^*)^k(A^*)^{T,k}\|,\ a_2=\|\sum_{k=1}^{\infty}(A^*)^k(A^*)^{T,k}\|.$$

By Neumann series, for any $X\in\mathbb{R}^{n\times n}$ with $\|X\|<1$, $\|(I-X)^{-1}\|\le\frac{1}{1-\|X\|}$. Therefore, for small ΔP_K, we have

$$\begin{aligned}&\|(I_n-\gamma^{-2}DD^TP_K)^{-1}\|\\&=\|(I_n-\gamma^{-2}DD^TP^*-\gamma^{-2}DD^T\Delta P_K)^{-1}\|\\&\le\frac{\|(I_n-\gamma^{-2}DD^TP^*)^{-1}\|}{1-a_3\|\Delta P_K\|}.\end{aligned}\tag{B.16}$$

where

$$a_3=\gamma^{-2}\|(I_n-\gamma^{-2}DD^TP^*)^{-1}\|\|DD^T\|.\tag{B.17}$$

Using the trace inequality in [54, Lemma 1], it follows from (B.15) and (B.16) that

$$\left(1-\frac{a_2a_3\|\Delta P_K\|}{1-a_3\|\Delta P_K\|}\right)\operatorname{Tr}(\Delta P_K)\le a_1\operatorname{Tr}(E_K).\tag{B.18}$$

Therefore, if

$$\|\Delta P_K\|\le\frac{1}{a_3+2a_2a_3}=:h_1,\tag{B.19}$$

it follows from Lemma 2 that

$$\|\Delta P_K\|_F\le2a_1\operatorname{Tr}(E_K)\le2a_1\sqrt{n}\|E_K\|_F.\tag{B.20}$$

When $\|\Delta P_K\|\ge h_1$, if follows from Lemma 7 that $\|E_K\|_F\neq0$. Since E_K is continuous with respect to K (Lemma 6) and the set $\mathcal{G}_h\cap\{K\in\mathcal{W}|\|\Delta P_K\|\ge h_1\}$ is compact (Lemma 8), there exists $a_4(h)>0$, such that $\|E_K\|_F\ge a_4(h)$. Hence, $\|\Delta P_K\|_F\le\operatorname{Tr}(\Delta P_K)\le\frac{h}{a_4(h)}\|E_K\|_F$. By taking $a(h)=\max(2a_1\sqrt{n},\frac{h}{a_4(h)})$, we obtain that $\|\Delta P_K\|_F\le a(h)\|E_K\|_F$. ∎

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. We can rewrite (17a) as

$$\begin{aligned}&A_{i+1}^TU_iA_{i+1}+K_i^TB^TU_iBK_i-K_{i+1}^TB^TU_iBK_{i+1}\\&+(K_{i+1}-K_i)^TB^TU_iA+A^TU_iB(K_{i+1}-K_i)\\&-P_i+Q+K_i^TRK_i=0.\end{aligned}\tag{B.21}$$

Since $(R+B^TU_iB)K_{i+1}=B^TU_iA$ from (17c), by completing the squares, we have

$$A_{i+1}^TU_iA_{i+1}-P_i+Q+K_{i+1}^TRK_{i+1}+E_i=0,\tag{B.22}$$

where $E_i=E_{K_i}=(K_{i+1}-K_i)^T(R+B^TU_iB)(K_{i+1}-K_i)$. Writing out (17a) for the $(i+1)$th iteration, subtracting it from (B.22), we can obtain that

$$A_{i+1}^T(U_i-U_{i+1})A_{i+1}-(P_i-P_{i+1})+E_i=0.\tag{B.23}$$

From (17b) and using [35, Lemma B.1], it holds

$$\begin{aligned}U_{i+1}&=(I_n-\gamma^{-2}P_{i+1}DD^T)^{-1}P_{i+1}\\&\preceq U_i-(I_n-\gamma^{-2}P_{i+1}DD^T)^{-1}(P_i-P_{i+1})\\&\quad(I_n-\gamma^{-2}DD^TP_{i+1})^{-1},\end{aligned}\tag{B.24}$$

Combining (B.23) and (B.24), we have

$$\begin{aligned}&A_{i+1}^T(I_n-\gamma^{-2}P_{i+1}DD^T)^{-1}(P_i-P_{i+1})\\&(I_n-\gamma^{-2}DD^TP_{i+1})^{-1}A_{i+1}-(P_i-P_{i+1})+E_i\preceq0.\end{aligned}\tag{B.25}$$

Considering the expression of $L_{i+1,*}$ in (14), we have

$$\begin{aligned}&(I_n-\gamma^{-2}DD^TP_{i+1})^{-1}A_{i+1}\\&=\left[I_n+\gamma^{-2}DD^TP_{i+1}(I_n-\gamma^{-2}DD^TP_{i+1})^{-1}\right]A_{i+1}\\&=A_{i+1,*},\end{aligned}\tag{B.26}$$

As a consequence, (B.25) can be rewritten as

$$A_{i+1,*}^T(P_i-P_{i+1})A_{i+1,*}-(P_i-P_{i+1})+E_i\preceq0,\tag{B.27}$$

By Lemma 3, it follows that $\{P_i\}$ is monotonically decreasing and $\operatorname{Tr}(P_i)\le\operatorname{Tr}(P_1)$ for any $i\in\mathbb{Z}_+$. Hence, given $K_1\in\mathcal{G}_h$, $K_i\in\mathcal{G}_h$ for any $i\in\mathbb{Z}_+$. Following (B.27) and [53, Theorem 5.D6], we have

$$(P_i-P_{i+1})\succeq\sum_{t=0}^{\infty}(A_{i+1,*}^T)^tE_iA_{i+1,*}^t\tag{B.28}$$

Subtracting P^* from both sides of (B.28) and taking the trace of (B.28), we have

$$\begin{aligned}&\operatorname{Tr}(P_{i+1}-P^*)\le\operatorname{Tr}(P_i-P^*)-\operatorname{Tr}(E_i)\\&\le\operatorname{Tr}(P_i-P^*)-\|E_i\|_F,\end{aligned}\tag{B.29}$$

where the last inequality is from Lemma 2. Considering Lemmas 2 and 9, (B.29) can be further derived as

$$\mathrm{Tr}(P_{i+1}-P^*) \le \left(1-\frac{1}{\sqrt{n}a(h)}\right)\mathrm{Tr}(P_i-P^*). \quad \text{(B.30)}$$

The theorem is thus proved by setting $\alpha(h)=1-\frac{1}{\sqrt{n}a(h)}$. From Lemma 3, $P^*\preceq P_{i+1}\preceq P_i$, and thus $0\le \mathrm{Tr}(P_{i+1}-P^*)\le \mathrm{Tr}(P_i-P^*)$. As a result, $\alpha(h)\in[0,1)$.

Appendix C: Proof of Theorem 2

Given an admissible feedback $K\in\mathcal{W}$, and starting from $L_{K,1}=0$, the inner-loop iteration is

$$A_{K,j}^T P_{K,j}A_{K,j}-P_{K,j}+Q_K-\gamma^2 L_{K,j}^T L_{K,j}=0 \quad \text{(C.1a)}$$
$$L_{K,j+1}=(\gamma^2 I_q-D^T P_{K,j}D)^{-1}D^T P_{K,j}A_K \quad \text{(C.1b)}$$

Recall that $Q_K=Q+K^TRK$, $A_K=A-BK$ and $A_{K,j}=A-BK+L_{K,j}$. The following lemma states the monotonic convergence of the inner-loop iteration.

Lemma 10: Suppose that the inner loop starts from the initial condition $L_{K,1}=0$. For any $K\in\mathcal{W}$, and $j\in\mathbb{Z}_+$, the following statements hold

1) $A_{K,j}:=A-BK+L_{K,j}$ is stable;
2) $P_K\succeq\cdots\succeq P_{K,j+1}\succeq P_{K,j}\succeq\cdots\succeq P_{K,1}$;
3) $\lim_{j\to\infty}\|P_{K,j}-P_K\|_F=0$ and $\lim_{j\to\infty}\|L_{K,j}-L_{K,*}\|_F=0$.

Proof: The lemma can be proved by following [8]. ■

Lemma 11: Given $K\in\mathcal{W}$, let L be admissible, i.e. $A-BK+DL$ is stable, and recall from (26) that $L'=(\gamma^2 I_q-D^TP_{K,L}D)^{-1}D^TP_{K,L}A_K$, where $P_{K,L}$ is defined in (23) . Define $E_{K,L}:=(L-L')^T(\gamma^2 I_q-D^TP_{K,L}D)(L-L')$. Then, there exists a constant $b(K)>0$, such that

$$\mathrm{Tr}(P_K-P_{K,L})\le b(K)\|E_{K,L}\|, \quad \text{(C.2)}$$

and

$$b(K):=\mathrm{Tr}\left(\sum_{t=0}^{\infty}(A_{K,*})^t(A_{K,*}^T)^t\right). \quad \text{(C.3)}$$

Proof: Subtracting (23) from (4a) results in

$$A_{K,*}^T(P_K-P_{K,L})A_{K,*}-(P_K-P_{K,L})+E_{K,L}-(L_{K,*}-L')^T(\gamma^2 I_q-D^TP_{K,L}D)(L_{K,*}-L')=0. \quad \text{(C.4)}$$

As $A_{K,*}$ is stable, by [53, Theorem 5.D6] we have

$$P_K-P_{K,L}\preceq\sum_{t=0}^{\infty}(A_{K,*}^T)^tE_{K,L}(A_{K,*})^t. \quad \text{(C.5)}$$

Taking the trace of (C.5) and using the cyclic property of the trace and [54, Lemma 1], we can obtain (C.2). ■

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. Let $E_{K,j}=E_{K,L_j}=(L_{K,j+1}-L_{K,j})^T(\gamma^2 I_q-D^TP_{K,j}D)(L_{K,j+1}-L_{K,j})$. Subtracting the jth iteration of (C.1a) from (C.1a) at the $(j+1)$th iteration, considering $(\gamma^2 I_q-D^TP_{K,j}D)L_{K,j+1}=D^TP_{K,j}A_K$ in (28b), and completing the squares, we have

$$A_{K,j+1}^T(P_{K,j+1}-P_{K,j})A_{K,j+1}-(P_{K,j+1}-P_{K,j})+E_{K,j}=0. \quad \text{(C.6)}$$

By (C.6) and [53, Theorem 5.D6], we have

$$\mathrm{Tr}(P_{K,j+1}-P_{K,j})=\mathrm{Tr}\left[\sum_{t=0}^{\infty}(A_{K,j+1}^T)^tE_{K,j}(A_{K,j+1})^t\right]. \quad \text{(C.7)}$$

Consequently,

$$\mathrm{Tr}(P_K-P_{K,j+1})\le\mathrm{Tr}(P_K-P_{K,j})-\mathrm{Tr}(E_{K,j}) \quad \text{(C.8)}$$
$$\le\mathrm{Tr}(P_K-P_{K,j})-\|E_{K,j}\|\le\beta(K)\,\mathrm{Tr}(P_K-P_{K,j}),$$

where $\beta(K)=1-1/b(K)$ and the last inequality comes from Lemma 11. Since $P_K\succeq P_{K,j}$, $\mathrm{Tr}(P_K-P_{K,j+1})\ge 0$. Hence, $\beta(K)\in[0,1)$.

Appendix D: Proof of Theorem 3

We will first show that $b(K)$ is continuous in K. Let $M_K:=\sum_{t=0}^{\infty}(A_{K,*})^t(A_{K,*}^T)^t$, where $A_{K,*}=A-BK+DL_{K,*}$ and $L_{K,*}=(\gamma^2 I_q-D^TP_KD)^{-1}D^TP_K(A-BK)$. Since $K\in\mathcal{W}$, $A_{K,*}$ is stable by Lemma 1. By [53, Theorem 5.D6], M_K is the unique solution to

$$A_{K,*}M_KA_{K,*}^T-M_K+I_n=0. \quad \text{(D.1)}$$

Since P_K is continuous in $K\in\mathcal{W}$ (Lemma 6), $A_{K,*}$ is continuous in $K\in\mathcal{W}$. Hence, M_K is continuous in $K\in\mathcal{W}$, and $b(K)=\mathrm{Tr}(M_K)$ is continuous in K. In addition, the set $\mathcal{G}_h:=\{K\in\mathcal{W}|\,\mathrm{Tr}(P_K)\le\mathrm{Tr}(P^*)+h\}$ is compact (Lemma 8). Therefore, the upperbound of $b(K)=\mathrm{Tr}(M_K)$ exists on $K\in\mathcal{G}_h$, that is $b(K)\le\bar{b}(h)$ for any $K\in\mathcal{G}_h$. Consequently, for any $K\in\mathcal{G}_h$, $\beta(K)\le\bar{\beta}(h)$, which is defined as

$$\bar{\beta}(h):=1-\frac{1}{\bar{b}(h)}. \quad \text{(D.2)}$$

Given $K\in\mathcal{G}_h$, following Lemma 2 and Theorem 2, we have

$$\|P_K-P_{K,j}\|_F\le\bar{\beta}^{j-1}(h)\,\mathrm{Tr}(P_K-P_{K,1})$$
$$\le\bar{\beta}^{j-1}(h)\,\mathrm{Tr}(P_K)\le\bar{\beta}^{j-1}(h)(\mathrm{Tr}(P^*)+h). \quad \text{(D.3)}$$

Therefore, for any $K\in\mathcal{G}_h$ and $\epsilon>0$, if $j\ge\bar{j}(h,\epsilon)=\log_{\bar{\beta}}^{\frac{\epsilon}{\mathrm{Tr}(P^*)+h}}+1$, $\|P_{K,j}-P_K\|_F\le\epsilon$. Noticing that $\bar{j}(h,\epsilon)$ is independent of K, the uniform convergence of the dual-loop algorithm follows readily.

Appendix E: Proof of Theorem 4

Recall that $\mathcal{G}_h:=\{K\in\mathcal{W}|\,\mathrm{Tr}(P_K)\le\mathrm{Tr}(P^*)+h\}$. The following lemma ensures that for $K\in\mathcal{G}_h$ and small perturbation $\Delta K\in\mathbb{R}^{m\times n}$, the updated policy becomes $K'+\Delta K$ that still belongs to $\mathcal{G}_h$. In other words, $\mathcal{G}_h$ is an invariant set under small disturbance.

Lemma 12: Let $K\in\mathcal{G}_h$, $K':=(R+B^TU_KB)^{-1}B^TU_KA$, and $\hat{K}':=K'+\Delta K$. Then, there exists $d(h)>0$, such that $\hat{K}'\in\mathcal{G}_h$ if $\|\Delta K\|_F<d(h)$.

Proof: Since $K\in\mathcal{G}_h$, it follows from Lemma 3 that $K'\in\mathcal{W}$. Suppose that $\hat{K}'\in\mathcal{W}$. According to Lemma 1, there exists a unique solution $\hat{P}_{K'}=\hat{P}_{K'}^T\succ 0$ to

$$(A-B\hat{K}')^T\hat{U}_{K'}(A-B\hat{K}')-\hat{P}_{K'}+Q+(\hat{K}')^TR\hat{K}'=0, \quad \text{(E.1a)}$$
$$\hat{U}_{K'}=\hat{P}_{K'}+\hat{P}_{K'}D(\gamma^2 I_q-D^T\hat{P}_{K'}D)^{-1}D^T\hat{P}_{K'}. \quad \text{(E.1b)}$$

In addition, we can rewrite (4a) as

$$\begin{aligned}&(A-B\hat{K}')^TU_K(A-B\hat{K}')-P_K+Q\\&+(\hat{K}'-K)^TB^TU_KA+A^TU_KB(\hat{K}'-K)\\&+K^T(R+B^TU_KB)K-(\hat{K}')^TB^TU_KB\hat{K}'=0.\end{aligned}\tag{E.2}$$

Noticing $(R+B^TU_KB)K'=B^TU_KA$ and $\hat{K}'=K'+\Delta K$, (E.2) implies

$$\begin{aligned}&(A-B\hat{K}')^TU_K(A-B\hat{K}')-P_K+Q\\&+(K'-K)^T(R+B^TU_KB)K'\\&+(K')^T(R+B^TU_KB)(K'-K)\\&+\Delta K^T(R+B^TU_KB)K'+(K')^T(R+B^TU_KB)\Delta K\\&+K^T(R+B^TU_KB)K-(\hat{K}')^TB^TU_KB\hat{K}'=0.\end{aligned}\tag{E.3}$$

Subtracting (E.1a) from (E.3) and completing the squares yield

$$\begin{aligned}&(A-B\hat{K}')^T(U_K-\hat{U}_{K'})(A-B\hat{K}')\\&-(P_K-\hat{P}_{K'})+E_K-\Delta K^T(R+B^TU_KB)\Delta K=0.\end{aligned}\tag{E.4}$$

From Lemma 9, $E_K=(K'-K)^T(R+B^TU_KB)(K'-K)$. Using [35, Lemma B.1] and following the derivation of (B.27), we have

$$\begin{aligned}&(A-B\hat{K}'+DL_{\hat{K}',*})^T(P_K-\hat{P}_{K'})(A-B\hat{K}'+DL_{\hat{K}',*})\\&-(P_K-\hat{P}_{K'})+E_K-\Delta K^T(R+B^TU_KB)\Delta K\preceq 0,\end{aligned}\tag{E.5}$$

where $L_{\hat{K}',*}=(\gamma^2I_q-D^T\hat{P}_{K'}D)^{-1}D^T\hat{P}_{K'}(A-B\hat{K}')$. Since $\hat{K}'\in\mathcal{W}$, by Lemma 1, $(A-B\hat{K}'+DL_{\hat{K}',*})$ is stable. Using [53, Theorem 5.D6], we have

$$\begin{aligned}\mathrm{Tr}(P_K-\hat{P}_{K'})\geq\mathrm{Tr}\Big\{\sum_{t=0}^{\infty}\Big\{&(A-B\hat{K}'+DL_{\hat{K}',*})^{T,t}[E_K\\&-\Delta K^T(R+B^TU_KB)\Delta K](A-B\hat{K}'+DL_{\hat{K}',*})^t\Big\}\Big\}.\end{aligned}\tag{E.6}$$

Let

$$\begin{aligned}c(\hat{K}')=\|\sum_{t=0}^{\infty}&(A-B\hat{K}'+DL_{\hat{K}',*})^t\\&(A-B\hat{K}'+DL_{\hat{K}',*})^{T,t}\|\end{aligned}\tag{E.7}$$

and $d_1(h)=\sup_{K\in\mathcal{G}_h}\|R+B^TU_KB\|$. Then, by Lemmas 2 and 9, and [54, Lemma 1], (E.6) implies

$$\begin{aligned}\mathrm{Tr}(\hat{P}_{K'}-P^*)\leq(1-\frac{1}{\sqrt{n}a(h)})\,&\mathrm{Tr}(P_K-P^*)\\&+c(\hat{K}')d_1(h)\|\Delta K\|_F^2.\end{aligned}\tag{E.8}$$

Therefore, if $\|\Delta K\|_F^2\leq\frac{h}{c(\hat{K}')d_1(h)\sqrt{n}a(h)}$, it is ensured that $\mathrm{Tr}(\hat{P}_{K'}-P^*)\leq h$, i.e. $\hat{K}'\in\mathcal{G}_h$. Let $\bar{c}(h)=\sup_{K\in\mathcal{G}_h}c(K)$. Since P_K is continuous in K (Lemma 6) and $L_{K,*}$ defined in (14) is continuous in P_K and K, $c(K)$ is continuous with respect to K and $\bar{c}(h)<\infty$. Hence, if

$$\|\Delta K\|_F\leq\left(\frac{h}{\bar{c}(h)d_1(h)\sqrt{n}a(h)}\right)^{\frac{1}{2}}=:d(h),\tag{E.9}$$

it is ensured that $\hat{K}'\in\mathcal{G}_h$. In other words, $\mathcal{B}(K',d(h))=\{K\in\mathbb{R}^{m\times n}|\|K-K'\|_F\leq d(h)\}\subset\mathcal{G}_h$.

Next, we prove that $\hat{K}'\in\mathcal{W}$ by contradiction. If $\hat{K}'\notin\mathcal{W}$, it follows that $\hat{K}'\notin\mathcal{B}(K',d(h))$. Hence, $\|\Delta K\|_F>d(h)$, which contradicts with the condition $\|\Delta K\|<d(h)$. ■

Now, we prove Theorem 4 and Corollary 1.

Proof of Theorem 4. From Lemma 12 and given an initial admissible policy $\hat{K}_1\in\mathcal{G}_h$, it is seen that if $\|\Delta K\|_\infty\leq d(h)$, $\hat{K}_i<\mathcal{G}_h$ for any $i\in\mathbb{Z}_+$. In (E.8), considering $\hat{P}_i$ and $\hat{P}_{i+1}$ as P_K and $\hat{P}_{K'}$, respectively, we have

$$\begin{aligned}\mathrm{Tr}(\hat{P}_{i+1}-P^*)\leq(1-\frac{1}{\sqrt{n}a(h)})\,&\mathrm{Tr}(\hat{P}_i-P^*)\\&+\bar{c}(h)d_1(h)\|\Delta K_{i+1}\|_F^2.\end{aligned}\tag{E.10}$$

Repeating the above inequality from $i=1$ yields

$$\begin{aligned}\mathrm{Tr}(\hat{P}_{i+1}-P^*)\leq(1-\frac{1}{\sqrt{n}a(h)})^i\,&\mathrm{Tr}(\hat{P}_1-P^*)\\&+\sqrt{n}a(h)\bar{c}(h)d_1(h)\|\Delta K\|_\infty^2\end{aligned}\tag{E.11}$$

From Lemma 2, it follows that

$$\begin{aligned}\|\hat{P}_i-P^*\|_F\leq(1-\frac{1}{\sqrt{n}a(h)})^{i-1}&\sqrt{n}\|\hat{P}_1-P^*\|_F\\&+\sqrt{n}a(h)\bar{c}(h)d_1(h)\|\Delta K\|_\infty^2.\end{aligned}\tag{E.12}$$

Thus, $\kappa_1(\cdot,\cdot)$ defined by $\kappa_1(\|\hat{P}_1-P^*\|_F,i):=(1-\frac{1}{\sqrt{n}a(h)})^{i-1}\sqrt{n}\|\hat{P}_1-P^*\|_F$ is a $\mathcal{KL}$-function, and $\xi_1(\cdot)$ defined by $\xi_1(\|\Delta K\|_\infty)=\sqrt{n}a(h)\bar{c}(h)d_1(h)\|\Delta K\|_\infty^2$ is a $\mathcal{K}$-function. Therefore, the inexact outer-loop iteration is small-disturbance ISS.

Proof of Corollary 1. For each outer-loop iteration of Algorithm 1, $P_{i,\bar{j}}$ instead of P_i is used to update the policy. This leads to the disturbance at each iteration

$$\begin{aligned}\Delta K_{i+1}=&(R+B^TU_{i,\bar{j}}B)^{-1}B^TU_{i,\bar{j}}A\\&-(R+B^TU_iB)^{-1}B^TU_iA.\end{aligned}\tag{E.13}$$

As $K'=(R+B^TU_KB)^{-1}B^TU_KA$ is continuously differentiable in U_K, and U_K is continuously differentiable in P_K, K' is Lipschitz continuous in $P_K\in\{P\succ 0|\,\mathrm{Tr}(P)\leq\mathrm{Tr}(P^*)+h\}$. Consequently, there exists $d_2(h)>0$,

$$\|\Delta K_{i+1}\|_F\leq d_2(h)\|P_i-P_{i,\bar{j}}\|_F.\tag{E.14}$$

By Theorem 4, for any $\epsilon>0$, there exist $d_3(h,\epsilon)>0$ and $\bar{i}\in\mathbb{Z}_+$, such that, if $K_1\in\mathcal{G}_h$ and $\|\Delta K\|_\infty\leq d_3(h,\epsilon)$, $K_i\in\mathcal{G}_h$ for all $i\in\mathbb{Z}_+$ and

$$\|P_{\bar{i}}-P^*\|_F\leq(1/2)\epsilon.\tag{E.15}$$

By Theorem 3, there exists $\bar{j}(h,\epsilon)\in\mathbb{Z}_+$, such that for any $i\in\mathbb{Z}_+$

$$\|P_i-P_{i,\bar{j}}\|_F\leq\min[d_3/d_2,(1/2)\epsilon].\tag{E.16}$$

Therefore, (E.14) and (E.16) imply $\|\Delta K\|_\infty\leq d_3(h,\epsilon)$. By norm's triangle inequality, (E.15) and (E.16), we have

$$\begin{aligned}&\|P_{\bar{i},\bar{j}}-P^*\|_F\leq\\&\|P_{\bar{i},\bar{j}}-P_{\bar{i}}\|_F+\|P_{\bar{i}}-P^*\|_F\leq\epsilon.\end{aligned}\tag{E.17}$$

APPENDIX F: PROOF OF THEOREM 5

Lemma 13: 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 simplify the notation, the following variables are defined to denote the inner-loop update without disturbance

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

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 1. By induction, assume that $\hat{A}_{i,j} = A - B\hat{K}_i + D\hat{L}_{i,j}$ is stable for some $j \in \mathbb{Z}_+$. We can rewrite (34a) as

$$\begin{aligned} &(A - B\hat{K}_i + D\hat{L}_{i,*})^T \hat{P}_i (A - B\hat{K}_i + D\hat{L}_{i,*}) \\ &- \hat{P}_i + \hat{Q}_i - \gamma^2 \hat{L}_{i,*}^T \hat{L}_{i,*} = 0. \end{aligned} \tag{F.2}$$

Subtracting the jth iteration of (36a) from (F.2) and completing the squares, we have

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

Since $\hat{A}_{i,j}$ is stable, from [53, Theorem 5.D6], $\hat{P}_i \succeq \hat{P}_{i,j}$, where $\hat{P}_i$ is from (34a). By completing the squares, (F.2) implies

$$\begin{aligned} 0 &= \hat{A}_{i,j+1}^T \hat{P}_i \hat{A}_{i,j+1} - \hat{P}_i + \hat{Q}_i - \gamma^2 \tilde{L}_{i,j+1}^T \tilde{L}_{i,j+1} - \hat{\Omega}_{i,j+1} \\ &+ (\tilde{L}_{i,j+1} - \hat{L}_{i,*})^T (\gamma^2 I_q - D^T \hat{P}_i D)(\tilde{L}_{i,j+1} - \hat{L}_{i,*}), \end{aligned} \tag{F.4}$$

where

$$\begin{aligned} \hat{\Omega}_{i,j+1} &= \Delta L_{i,j+1}^T D^T \hat{P}_i D \tilde{L}_{i,j+1} + \tilde{L}_{i,j+1}^T D^T \hat{P}_i D \Delta L_{i,j+1} \\ &+ \Delta L_{i,j+1}^T D^T \hat{P}_i D \Delta L_{i,j+1} \\ &+ \hat{A}_i^T \hat{P}_i D \Delta L_{i,j+1} + \Delta L_{i,j+1}^T D^T \hat{P}_i \hat{A}_i. \end{aligned} \tag{F.5}$$

Since $\hat{P}_i \succeq \hat{P}_{i,j}$, it follows that $\hat{L}_{i,*}^T \hat{L}_{i,*} \succeq \tilde{L}_{i,j+1}^T \tilde{L}_{i,j+1}$ and $\|\hat{L}_{i,*}\| \geq \|\tilde{L}_{i,j+1}\|$. As a consequence,

$$\|\hat{\Omega}_{i,j+1}\| \leq e_1(\hat{K}_i)\|\Delta L_{i,j+1}\| + e_2(\hat{K}_i)\|\Delta L_{i,j+1}\|^2. \tag{F.6}$$

where

$$\begin{aligned} e_1(\hat{K}_i) &= (2\|D^T \hat{P}_i D\|\|\hat{L}_{i,*}\| + 2\|D^T \hat{P}_i \hat{A}_i\|), \\ e_2(\hat{K}_i) &= \|D^T \hat{P}_i D\|. \end{aligned}$$

Following Lemma 1, we know that $\hat{P}_i \succ 0$ and $\hat{A}_{i,*} = A - B\hat{K}_i + D\hat{L}_{i,*}$ is stable. Therefore, by [53, Theorem 5.D6] and (F.2), $\hat{Q}_i - \gamma^2 \hat{L}_{i,*}^T \hat{L}_{i,*} \succ 0$, 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

$$\|\Delta L_{i,j+1}\| \leq \frac{-e_1 + \sqrt{e_1^2 + 2e_2 e_3}}{2e_2} := e(\hat{K}_i), \tag{F.7}$$

we have

$$\hat{Q}_i - \gamma^2 \hat{L}_{i,*}^T \hat{L}_{i,*} - \hat{\Omega}_{i,j+1} \succ \frac{1}{2} e_3(\hat{K}_i) I_n. \tag{F.8}$$

As $\hat{L}_{i,*}^T \hat{L}_{i,*} \succeq \tilde{L}_{i,j+1}^T \tilde{L}_{i,j+1}$, $\hat{Q}_i - \gamma^2 \tilde{L}_{i,j+1}^T \tilde{L}_{i,j+1} - \hat{\Omega}_{i,j+1} \succ 0$. $\hat{A}_{i,j+1}$ is stable as a result of (F.4) and [55, Theorem 8.4]. Therefore, for any $j \in \mathbb{Z}_+$, $\hat{A}_{i,j}$ is stable. ■

Proof of Theorem 5. Rewrite the jth iteration of (36a) as

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

Subtracting (F.9) from the $(j+1)$th iteration of (36) yields

$$\begin{aligned} &\hat{A}_{i,j+1}^T (\hat{P}_{i,j+1} - \hat{P}_{i,j}) \hat{A}_{i,j+1} - (\hat{P}_{i,j+1} - \hat{P}_{i,j}) + \hat{E}_{i,j} \\ &- \Delta L_{i,j+1}^T (\gamma^2 I_q - D^T \hat{P}_{i,j} D) \Delta L_{i,j+1} = 0. \end{aligned} \tag{F.10}$$

where

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

When $\|\Delta L_i\|_\infty < e(\hat{K}_i)$, by Lemma 13, $\hat{A}_{i,j+1}$ is stable. Following [53, Theorem 5.D6], we have

$$\begin{aligned} \operatorname{Tr}(\hat{P}_{i,j+1} - \hat{P}_{i,j}) = \operatorname{Tr} \Bigg\{ \sum_{t=0}^{\infty} (\hat{A}_{i,j+1}^T)^t \Big[\hat{E}_{i,j} \\ - \Delta L_{i,j+1}^T (\gamma^2 I_q - D^T \hat{P}_{i,j} D) \Delta L_{i,j+1} \Big] (\hat{A}_{i,j+1})^t \Bigg\}. \end{aligned} \tag{F.11}$$

Consequently, by Lemma 11,

$$\begin{aligned} &\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j+1}) \leq (1 - 1/b(\hat{K}_i)) \operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) \\ &+ \gamma^2 \|\Delta L_i\|_\infty^2 \operatorname{Tr}(\hat{M}_{i,j+1}). \end{aligned} \tag{F.12}$$

where $\hat{M}_{i,j+1} := \sum_{t=0}^{\infty} (\hat{A}_{i,j+1}^T)^t (\hat{A}_{i,j+1})^t$. By [53, Theorem 5.D6], $\hat{M}_{i,j+1}$ satisfies

$$\hat{A}_{i,j+1}^T \hat{M}_{i,j+1} \hat{A}_{i,j+1} - \hat{M}_{i,j+1} + I_n = 0. \tag{F.13}$$

Multiplying both sides of (F.13) by $\frac{1}{2} e_3(\hat{K}_i)$ and subtracting it from (F.4), we have

$$\begin{aligned} &\hat{A}_{i,j+1}^T (\hat{P}_i - \frac{1}{2} e_3 \hat{M}_{i,j+1}) \hat{A}_{i,j+1} - (\hat{P}_i - \frac{1}{2} e_3 \hat{M}_{i,j+1}) \\ &+ \hat{Q}_i - \gamma^2 \tilde{L}_{i,j+1}^T \tilde{L}_{i,j+1} - \hat{\Omega}_{i,j+1} - \frac{1}{2} e_3 I_n \\ &+ (\tilde{L}_{i,j+1} - \hat{L}_{i,*})^T (\gamma^2 I_q - D^T \hat{P}_i D)(\tilde{L}_{i,j+1} - \hat{L}_{i,*}) = 0. \end{aligned} \tag{F.14}$$

By (F.8) and [53, Theorem 5.D6], $\hat{P}_i - \frac{1}{2} e_3(\hat{K}_i) \hat{M}_{i,j+1} \succeq 0$. As a consequence $\operatorname{Tr}(\hat{M}_{i,j+1}) \leq 2/e_3 \operatorname{Tr}(\hat{P}_i)$.

From (F.12), we have

$$\begin{aligned} &\operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j+1}) \leq (1 - 1/b(\hat{K}_i)) \operatorname{Tr}(\hat{P}_i - \hat{P}_{i,j}) \\ &+ 2/e_3(\hat{K}_i) \operatorname{Tr}(\hat{P}_i) \gamma^2 \|\Delta L_i\|_\infty^2. \end{aligned} \tag{F.15}$$

Using Lemma 2 and repeating the above argument for $j, j-1, \cdots, 1$, it follows that

$$\begin{aligned} \|\hat{P}_i - \hat{P}_{i,j}\|_F &\leq (1 - \frac{1}{b(\hat{K}_i)})^{j-1} \sqrt{n} \|\hat{P}_i - \hat{P}_{i,1}\|_F \\ &+ \frac{2}{e_3(\hat{K}_i)} b(\hat{K}_i) \operatorname{Tr}(\hat{P}_i) \gamma^2 \|\Delta L_i\|_\infty^2. \end{aligned} \tag{F.16}$$

Clearly, $\kappa_2(\|\hat{P}_i - \hat{P}_{i,1}\|_F, j) = (1 - \frac{1}{b(\hat{K}_i)})^{j-1} \sqrt{n} \|\hat{P}_i - \hat{P}_{i,1}\|_F$ is a $\mathcal{KL}$-function, and $\xi_2(\|\Delta L_i\|_\infty) = \frac{2}{e_3(\hat{K}_i)} b(\hat{K}_i) \operatorname{Tr}(\hat{P}_i) \gamma^2 \|\Delta L_i\|_\infty^2$ is a $\mathcal{K}$-function. Therefore, we can conclude that the inexact inner-loop iteration is ISS.

APPENDIX G: PROOF OF THEOREM 6

Since when $\epsilon = \mu = 0$ and $\tau \to \infty$, the feasible set of the LMIs (68a) and (68b) is nonempty, by continuity, (68a) and (68b) has a solution for sufficiently small ϵ and μ and sufficiently large τ. Equation (68a) implies that

$$\text{L.H.S. of (A.1)} + \begin{bmatrix} \epsilon I_n & * & * & * \\ 0 & \epsilon I_q & * & * \\ \tilde{A}_\tau W - \tilde{B}_\tau V & \tilde{D}_\tau & \epsilon I_n & * \\ 0 & 0 & 0 & \epsilon I_p \end{bmatrix} \prec 0, \tag{G.1}$$

where $\tilde{A}_\tau = \hat{A}_\tau - A$, $\tilde{B}_\tau = \hat{B}_\tau - B$, and $\tilde{D}_\tau = \hat{D}_\tau - D$. By Schur complement lemma, it follows from (68b) that

$$I_n - \mu^2 [W, -V^T] \begin{bmatrix} W \\ -V \end{bmatrix} \succ 0. \tag{G.2}$$

Since the following relation holds *almost surely*

$$\lim_{\tau\to\infty} \tilde{A}_\tau = 0, \quad \lim_{\tau\to\infty} \tilde{B}_\tau = 0, \quad \lim_{\tau\to\infty} \tilde{D}_\tau = 0, \tag{G.3}$$

by the definition of limit, there exists $\tau^*(\epsilon, \mu) > 0$, such that for all $\tau > \tau^*(\epsilon, \mu)$, the following holds *almost surely*:

$$\|\tilde{A}_\tau\|_F \le \frac{\epsilon\mu}{\sqrt{2}}, \quad \|\tilde{B}_\tau\|_F \le \frac{\epsilon\mu}{\sqrt{2}}, \quad \|\tilde{D}_\tau\|_F \le \frac{\epsilon}{2}. \tag{G.4}$$

Consequently, $\|[\tilde{A}_\tau, \tilde{B}_\tau]\| \le \frac{\epsilon\mu}{\sqrt{2}}$, and

$$\begin{aligned} &[0, W\tilde{A}_\tau^T - V^T\tilde{B}_\tau^T] \begin{bmatrix} \epsilon I_q & \tilde{D}^T \\ \tilde{D} & \epsilon I_n \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \tilde{A}_\tau W - \tilde{B}_\tau V \end{bmatrix} \\ &= [W, -V^T] \begin{bmatrix} \tilde{A}_\tau^T \\ \tilde{B}_\tau^T \end{bmatrix} (\epsilon I_n - \frac{1}{\epsilon}\tilde{D}\tilde{D}_\tau^T)^{-1} \begin{bmatrix} \tilde{A}_\tau & \tilde{B}_\tau \end{bmatrix} \begin{bmatrix} W \\ -V \end{bmatrix} \\ &\preceq \epsilon\mu^2 [W, -V^T][W, -V^T]^T. \end{aligned} \tag{G.5}$$

Therefore, by combining (G.2), (G.5), and Schur complement lemma, the second term in (G.1) is positive semi-definite, and the first term in (G.1) is negative definite. By Lemma 1, it follows that $K = VW^{-1} \in \mathcal{W}$.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2nd ed. ed., 2018.

[2] J. Bu, A. Mesbahi, and M. Mesbahi, "Policy gradient-based algorithms for continuous-time linear quadratic control," *arXiv preprint arXiv:2006.09178*, 2020.

[3] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[4] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2435–2450, 2022.

[5] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 1467–1476, PMLR, 10–15 Jul 2018.

[6] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Trans. Autom. Control*, vol. 66, no. 11, pp. 5283–5298, 2021.

[7] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6429–6444, 2022.

[8] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Trans. Autom. Control*, vol. 16, no. 4, pp. 382–384, 1971.

[9] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Control*, vol. 13, no. 1, pp. 114–115, 1968.

[10] Y. Jiang and Z. P. Jiang, *Robust Adaptive Dynamic Programming*. Wiley-IEEE Press, 2017.

[11] F. L. Lewis, D. Vrabie, and V. L. Syros, *Optimal Control*. John Wiley & Sons, 2012.

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

[13] B. Pang and Z. P. Jiang, "Robust reinforcement learning: A case study in linear quadratic regulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9303–9311, May 2021.

[14] B. Pang, T. Bian, and Z. P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Trans. Autom. Control*, vol. 67, no. 1, pp. 504–511, 2022.

[15] E. D. Sontag, "Remarks on input to state stability of perturbed gradient flows, motivated by model-free feedback control learning," *Systems & Control Letters*, vol. 161, p. 105138, 2022.

[16] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Trans. Autom. Control*, vol. 18, no. 2, pp. 124–131, 1973.

[17] P. Whittle, "Risk-sensitive linear/quadratic/Gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.

[18] K. Glover and J. C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to risk sensitivity," *Systems & control letters*, vol. 11, no. 3, pp. 167–172, 1988.

[19] T. Başar and P. Bernhard, *H_∞-optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. New York, USA: Springer, 1995.

[20] V. Borkar, "A sensitivity formula for risk-sensitive cost and the actor–critic algorithm," *Systems & Control Letters*, vol. 44, no. 5, pp. 339–346, 2001.

[21] V. S. Borkar, "Q-learning for risk-sensitive control," *Mathematics of Operations Research*, vol. 27, no. 2, pp. 294–311, 2002.

[22] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, pp. 267–290, 2002.

[23] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to $\mathcal{H}_\infty$ control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[24] Y. Zhang, Z. Yang, and Z. Wang, "Provably efficient actor-critic for risk-sensitive and robust adversarial RL: A linear-quadratic case," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (A. Banerjee and K. Fukumizu, eds.), vol. 130 of *Proceedings of Machine Learning Research*, pp. 2764–2772, PMLR, 13–15 Apr 2021.

[25] K. Zhang, B. Hu, and T. Başar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence," *SIAM J. Control Optim.*, vol. 59, no. 6, pp. 4081–4109, 2021.

[26] K. Zhang, Z. Yang, and T. Başar, "Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games," in *Advances in Neural Information Processing Systems*, vol. 32, p. 11602–11614, 2019.

[27] J. Bu, L. J. Ratliff, and M. Mesbahi, "Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games," *arXiv preprint arXiv:1911.04672*, 2019.

[28] K. Zhang, B. Hu, and T. Başar, "On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 22056–22068, 2020.

[29] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Policy iterations on the Hamilton–Jacobi–Isaacs equation for H_∞ state feedback control with input saturation," *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1989–1995, 2006.

[30] K. G. Vamvoudakis and F. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *Int. J. Robust Nonlinear Control*, vol. 22, no. 13, pp. 1460–1483, 2012.

[31] L. Cui, Z. P. Jiang, and E. D. Sontag, "Small-disturbance input-to-state stability of perturbed gradient flows: Applications to LQR problem," *arXiv preprint arXiv:2310.02930*, 2023.

[32] Z. P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Found. Trends Syst. Control*, vol. 8, no. 3, pp. 176–284, 2020.

[33] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[34] L. Cui, T. Başar, and Z. P. Jiang, "A reinforcement learning look at risk-sensitive linear quadratic Gaussian control," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference* (N. Matni, M. Morari, and G. J. Pappas, eds.), vol. 211 of *Proceedings of Machine Learning Research*, pp. 534–546, PMLR, 15–16 Jun 2023.
[35] K. Zhang, B. Hu, and T. Başar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence," *arXiv:1910.09496*, 2019.
[36] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Princeton, NJ: Prentice Hall, 1996.
[37] Z. P. Jiang and T. Liu, "Small-gain theory for stability and control of dynamical networks: A survey," *Annual Reviews in Control*, vol. 46, pp. 58–79, 2018.
[38] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity," *IEEE Trans. Autom. Control*, vol. 11, no. 2, pp. 228–238, 1966.
[39] W. Hahn, *Stability of Motion*. Berlin, Germany: Springer, 1967.
[40] Z. P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, 2001.
[41] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.
[42] J. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Autom. Control*, vol. 42, no. 5, pp. 674–690, 1997.
[43] Z. P. Jiang, C. Prieur, and A. Astolfi (Editors), *Trends in Nonlinear and Adaptive Control: A Tribute to Laurent Praly for His 65th Birthday*. Cham, Switzerland: Springer Nature, 2022.
[44] K. J. Åström and B. Wittenmark, *Adaptive control*. Mineola, NY: Dover Publications, 2nd ed., 2008.
[45] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: Wiley-IEEE Press, 2013.
[46] L. Koralov and Y. G. Sinai, *Theory of Probability and Random Processes*. Berlin/Heidelberg, Germany: Springer-Verlag, 2nd ed., 2007.
[47] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Hoboken, NJ: John Wiley & Sons, 2007.
[48] K. Zhang, X. Zhang, B. Hu, and T. Başar, "Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 2949–2964, 2021.
[49] C. Anderson, "Learning to control an inverted pendulum using neural networks," *IEEE Control Syst. Mag.*, vol. 9, no. 3, pp. 31–37, 1989.
[50] L. Cui, T. Başar, and Z. P. Jiang, "Robust reinforcement learning for risk-sensitive linear quadratic Gaussian control," *arXiv preprint arXiv:2212.02072*, 2023.
[51] J. R. Magnus and H. Neudecker, "Matrix differential calculus with applications to simple, Hadamard, and Kronecker products," *Journal of Mathematical Psychology*, vol. 29, no. 4, pp. 474–492, 1985.
[52] C. C. Pugh, *Real Mathematical Analysis*. Switzerland: Springer, 2nd ed., 2015.
[53] C.-T. Chen, *Linear System Theory and Design*. New York, NY: Oxford University Press, 3rd ed., 1999.
[54] S.-D. Wang, T.-S. Kuo, and C.-F. Hsu, "Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equation," *IEEE Trans. Autom. Control*, vol. 31, no. 7, pp. 654–656, 1986.
[55] J. P. Hespanha, *Linear Systems Theory*. Princeton, NJ: Princeton Press, Feb. 2018.

Leilei Cui received the B.Sc. degree in Automation from Northwestern Polytechnical University, Xian, China, in 2016, and the M.Sc. degree in Control Science and Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently a Ph.D. candidate in the Control and Networks Lab, Tandon School of Engineering, New York University. His research interests include optimal control, reinforcement learning, and adaptive dynamic programming.

Tamer Başar has been with the University of Illinois Urbana Champaign since 1981, where he is currently Swanlund Endowed Chair Emeritus and Center for Advanced Study (CAS) Professor Emeritus of Electrical and Computer Engineering, with also affiliations with the Coordinated Science Laboratory, Information Trust Institute, and Mechanical Science and Engineering. At Illinois, he has also served as Director of CAS (2014-2020), Interim Dean of Engineering (2018), and Interim Director of the Beckman Institute (2008-2010). He received B.S.E.E. from Robert College, Istanbul, and M.S., M.Phil, and Ph.D. from Yale University, from which he received in 2021 the Wilbur Cross Medal. He is a member of the US National Academy of Engineering, and Fellow of the American Academy of Arts and Sciences, AAIA, IEEE, IFAC, and SIAM. He has served as presidents of IEEE CSS (Control Systems Society), ISDG (International Society of Dynamic Games), and AACC (American Automatic Control Council). He has received several awards and recognitions over the years, including the highest awards of IEEE CSS, IFAC, AACC, and ISDG, the IEEE Control Systems Award, and a number of international honorary doctorates and professorships. He has around 1000 publications in systems, control, communications, optimization, networks, and dynamic games, including books on non-cooperative dynamic game theory, robust control, network security, wireless and communication networks, and stochastic networked control. He was the Editor-in-Chief of Automatica between 2004 and 2014, and is currently editor of several book series. His current research interests include stochastic teams, games, and networks; multi-agent systems and learning; data-driven distributed optimization; epidemics modeling and control over networks; security and trust; energy systems; and cyber-physical systems.

Zhong-Ping Jiang received the M.Sc. degree in statistics from the University of Paris XI, France, in 1989, and the Ph.D. degree in automatic control and mathematics from ParisTech-Mines, France, in 1993, under the direction of Prof. Laurent Praly.

Currently, he is a professor in the Department of Electrical and Computer Engineering and an affiliate professor in the Department of Civil and Urban Engineering at the Tandon School of Engineering, New York University. His main research interests include stability theory, robust/adaptive/distributed nonlinear control, robust adaptive dynamic programming, reinforcement learning and their applications to information, mechanical and biological systems. Prof. Jiang is a recipient of the prestigious Queen Elizabeth II Fellowship Award from the Australian Research Council, CAREER Award from the U.S. National Science Foundation, JSPS Invitation Fellowship from the Japan Society for the Promotion of Science, Distinguished Overseas Chinese Scholar Award from the NSF of China, and several best paper awards. He has served as Deputy Editor-in-Chief, Senior Editor and Associate Editor for numerous journals, and is among the Clarivate Analytics Highly Cited Researchers and Stanford's Top 2% Most Highly Cited Scientists. In 2022, he received the Excellence in Research Award from the NYU Tandon School of Engineering. Prof. Jiang is a foreign member of the Academia Europaea (Academy of Europe) and the European Academy of Sciences and Arts, and also is a Fellow of the IEEE, IFAC, CAA and AAIA.