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

**Leilei Cui** L.CUI@NYU.EDU
*Department of Electrical and Computer Engineering, New York University, New York, NY 11201.*

**Tamer Başar** BASAR1@ILLINOIS.EDU
*Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL 61801.*

**Zhong-Ping Jiang** ZJIANG@NYU.EDU
*Department of Electrical and Computer Engineering, New York University, New York, NY 11201.*

**Editors:** N. Matni, M. Morari, G. J. Pappas

## Abstract

In this paper, we propose a robust reinforcement learning method for a class of linear discrete-time systems to handle model mismatches that may be induced by sim-to-real gap. Under the formulation of risk-sensitive linear quadratic Gaussian control, a dual-loop policy optimization algorithm is proposed to iteratively approximate the robust and optimal controller. The convergence and robustness of the dual-loop policy optimization algorithm are rigorously analyzed. It is shown that the dual-loop policy optimization algorithm uniformly converges to the optimal solution. In addition, by invoking the concept of small-disturbance input-to-state stability, it is guaranteed that the dual-loop policy optimization algorithm still converges to a neighborhood of the optimal solution when the algorithm is subject to a sufficiently small disturbance at each step. When the system matrices are unknown, a learning-based off-policy policy optimization algorithm is proposed for the same class of linear systems with additive Gaussian noise. The numerical simulation is implemented to demonstrate the efficacy of the proposed algorithm.

**Keywords:** Robust reinforcement learning, policy optimization (PO), input-to-state stability (ISS).

## 1. Introduction

By interacting continuously with an unknown environment, reinforcement learning (RL) is a branch of machine learning to iteratively learn optimal decisions from data without knowing the system dynamics. Policy optimization (PO) is a fundamental technique for RL algorithm development as introduced in Sutton and Barto (2018). The key strategy of PO is to parameterize the policy and then iteratively update the policy parameters along the gradient direction of the specified performance index. When the system model is unknown, the gradient of the performance index is approximated by learning-based methods through sampling and experimentation. As a result, accurate policy gradient is hard to compute in reality due to measurement noise, immeasurable disturbance of the system, and function approximation errors. Therefore, convergence and robustness of PO are two important properties for practical implementation of RL algorithms.

The linear quadratic regulator (LQR) problem provides a tractable and insightful benchmark for the theoretical study of RL algorithms. For the PO of LQR, the control policy is parameterized as a linear function of the state, and the performance index is a quadratic function of the state as well as of the control. Since the performance index is differentiable with respect to the policy

parameters, several policy gradient descent algorithms, including vanilla gradient descent, natural policy gradient descent, and Gauss-Newton gradient descent algorithms, have been developed in Fazel et al. (2018); Bu et al. (2020); Gravell et al. (2021); Mohammadi et al. (2022); Li et al. (2021); Hu et al. (2022) and Cassel and Koren (2021). It has been shown that the control policy generated at each iteration of PO is stabilizing, and it globally converges to the optimal control policy. The developed PO algorithms for LQR pave a natural pathway to model-free analysis, where the RL techniques come into play. For example, based on zeroth-order methods, the gradient of the performance index can be approximated in the absence of precise model knowledge, and several model-free PO algorithms have been proposed in Fazel et al. (2018); Mohammadi et al. (2022) and Li et al. (2021). Since the gradient of the performance index cannot be estimated accurately, it is important to study the robustness of the PO algorithms subject to errors at each step. In Pang and Jiang (2021); Pang et al. (2022a), by viewing the PO algorithm as a discrete-time nonlinear system, and invoking the concept of input-to-state stability (ISS), the authors have shown that the control policies generated by Kleinman's policy iteration in Kleinman (1968); Hewer (1971) (same as the Gauss-Newton algorithm with a step size of $\frac{1}{2}$) can still converge to a neighborhood of the optimal control policy, even in the presence of sufficiently small disturbance. A similar robustness property was demonstrated in Sontag (2022) for the steepest gradient descent algorithm.

Since the robustness of the closed-loop system is ignored in the aforementioned PO algorithms for LQR, the control policy obtained may fail to stabilize the system in the presence of model mismatches and disturbances, that may be caused by sim-to-real gap and parameter variations of the system, e.g. Cui et al. (2021b). Robust and optimal control theory, particularly the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, is a key strategy to handle model mismatches and disturbances Zhou et al. (1996); Doyle et al. (1989); Mustafa and Bernstein (1991); Apkarian et al. (2008). In Zhang et al. (2021a), utilizing the concept of implicit regularization, the authors have proposed PO algorithms for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control such that the stability and constraint on the $\mathcal{H}_\infty$-norm of the closed-loop system are maintained at each iteration. By the fundamental connection between mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, the risk-sensitive linear-quadratic-Gaussian optimal control (with exponentiated loss), and linear-quadratic zero-sum dynamic games (LQ ZSDGs), the PO algorithms can be transformed into the dual-loop PO algorithms for LQ ZSDGs in Zhang et al. (2020, 2021a,b); Bu et al. (2019). The outer loop is to generate a protagonist under the worst-case adversary while the inner loop is to generate the worst-case adversary. However, issues related to uniform convergence and robustness of the dual-loop algorithm are to be explored.

In this paper, we propose a dual-loop PO algorithm for solving the risk-sensitive linear quadratic Gaussian control to handle model mismatches and disturbances. It is demonstrated that the dual-loop PO algorithm uniformly converges to the optimal solution with robustness guarantee. Specifically, by showing the linear convergence of the inner-loop iteration and computing the upperbound of the convergence rate, we demonstrate uniform convergence of the dual-loop algorithm. Furthermore, by invoking the concept of ISS Sontag (2008) and its latest variant "small-disturbance ISS" Pang and Jiang (2021), it is demonstrated that the PO algorithm still converges to a small neighborhood of the optimal solution, when the noise is sufficiently small. Based on these results, a learning-based off-policy PO algorithm is proposed when the system is disturbed by an immeasurable Gaussian noise and the system matrices are unknown. Several numerical examples are given to validate the efficacy of our theoretical results.

To sum up, our main contributions are three-fold: 1) the convergence, especially uniform convergence, of the dual-loop PO algorithm has been theoretically analyzed; 2) under the concept of the

small-disturbance ISS, the robustness of the outer and inner loops has been theoretically analyzed; 3) a novel learning-based off-policy PO algorithm has been proposed.

*Notations*: For a matrix $X \in \mathbb{R}^{m\times n}$, $\text{vec}(X) := [x_1^T, \cdots, x_n^T]^T$, where $x_i$ is the $i$th column of $X$. For a real symmetric matrix $P$, $\text{vecs}(P) := [p_{1,1}, 2p_{1,2}, \cdots, 2p_{1,n}, p_{2,2}, 2p_{2,3}, \cdots, p_{n,n}]^T$, where $p_{i,j}$ is the element at the $i$th row and $j$th column. $[X]_{i,j}$ denotes the submatrix of the matrix $X$ that is comprised of the rows between the $i$th and $j$th rows of $X$. For a vector $a \in \mathbb{R}^n$, $\text{vecv}(a) := [a_1^2, a_1a_2, \cdots, a_1a_n, a_2^2, a_2a_3, \cdots, a_n^2]^T$. $\text{Tr}(\cdot)$ denotes the trace of a matrix.

## 2. Preliminaries

In this section, we begin with the formulation of linear exponential quadratic Guassian (LEQG) control problem, followed by the robustness analysis of the closed-loop system.

### 2.1. Linear Exponential Quadratic Guassian Control

Consider the discrete-time linear time-invariant system

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

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

$$\mathcal{J}_{LEQG}(x_0, u) := \lim_{\tau\to\infty} \frac{2\gamma^2}{\tau} \log\left[\mathbb{E}\exp\left(\frac{1}{2\gamma^2}\sum_{t=0}^{\tau} y_t^T y_t\right)\right], \tag{2}$$

where $\gamma$ is a positive constant characterizing the magnitude of the risk sensitivity.

**Assumption 1** *$(A, B)$ is stabilizable, $C^TC = Q \succ 0$, and $\gamma > \gamma_\infty$, where $\gamma_\infty > 0$ is the minimal value of $\gamma$ such that for all $\gamma > \gamma_\infty$, there exists a control under which (2) is finite.*

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

Under Assumptions 1 and 2, as investigated by Jacobson (1973); Başar and Bernhard (1995), the optimal controller of the LEQG problem is $u_t^* = -K^*x_t$, where

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

with $P^* = (P^*)^T$ the unique positive definite solution to the generalized algebraic Riccati equation (GARE)

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

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

### 2.2. Robustness Analysis

By considering $w$ as a disturbance input in (1) and taking any stabilizing feedback control $u_t = -Kx_t$, the discrete-time transfer function from $w$ to $y$ becomes

$$\mathcal{T}(K) := (C - EK)[zI_n - (A - BK)]^{-1}D. \tag{5}$$

where $z \in \mathbb{C}$ is the $z$-transform variable. As shown in Fig. 1, $\Delta$ denotes the model mismatch induced by the sim-to-real gap, and its $\mathcal{H}_\infty$-norm satisfies $\|\Delta\|_{\mathcal{H}_\infty} \leq \frac{1}{\gamma}$. Thanks to the small-gain theorem Zhou et al. (1996); Jiang and Liu (2018); Zames (1966), in the presence of model mismatch, the system is stable if $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} < \gamma$. Consequently, the controller $u_t = -Kx_t$ is robust to the model mismatch $\Delta$ if $K$ lies within the admissible set $\mathcal{W}$ defined as

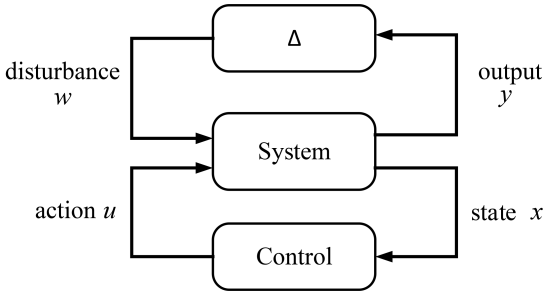


Figure 1: Robust control design with model mismatch $\Delta$.

$$\mathcal{W} := \{K \in \mathbb{R}^{m\times n} | (A - BK) \text{ is stable}, \|\mathcal{T}(K)\|_{\mathcal{H}_\infty} < \gamma\}. \tag{6}$$

As investigated in Theorem 3.8 of Başar and Bernhard (1995), the LEQG control in (3) satisfies $K^* \in \mathcal{W}$, and therefore, it is optimal with respect to (2) and robust to the model mismatch. Given the aforementioned preliminaries, in this paper, we investigate the following learning-based PO problem.

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

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

## 3. Model-Based Policy Optimization

In this section, a model-based dual-loop PO algorithm is proposed to solve the LEQG problem (2).

### 3.1. Introduction of the Outer Loop

The outer-loop iteration is developed based on the results of equation (3.5) in Zhang et al. (2019), and it aims to update the control policy $u_t = -K_i x_t$ under the worst-case disturbance. Let $i$ denote the iteration index for the outer loop and introduce the following variables:

$$A_i := A - BK_i, \quad Q_i := Q + K_i^T R K_i. \tag{7}$$

Then, the outer-loop iteration is

$$A_i^T U_i A_i - P_i + Q_i = 0, \tag{8a}$$
$$K_{i+1} = (R + B^T U_i B)^{-1} B^T U_i A, \tag{8b}$$

where $U_i$ is

$$U_i := P_i + P_i D(\gamma^2 I_q - D^T P_i D)^{-1} D^T P_i. \tag{9}$$

The following theorem now says that $P_i$ is monotonically decreasing ($P_{i+1} \succeq P_i$) and converges to $P^*$ with a linear convergence rate. The proof of the theorem can be found in Appendix B of Cui and Jiang (2022).

**Theorem 1** *Given $K_1 \in \mathcal{W}$, for any $i \geq 1$, there exists $\alpha(K_1) \in [0, 1)$, such that*

$$\mathrm{Tr}(P_{i+1} - P^*) \leq \alpha(K_1)\,\mathrm{Tr}(P_i - P^*). \tag{10}$$

Since $\|P_i - P^*\|_F \leq \mathrm{Tr}(P_i - P^*) \leq \sqrt{n}\|P_i - P^*\|_F$, it follows from Theorem 1 that $\|P_{i+1} - P^*\|_F \leq \alpha^i(K_1)\sqrt{n}\|P_1 - P^*\|_F$.

### 3.2. Introduction of the Inner Loop

Let $j$ denote the iteration index of the inner loop. The following variable is introduced to simplify the notation

$$A_{i,j} := A - BK_i + DL_{i,j}. \tag{11}$$

Given the control gain of the minimizer, $K_i$, the inner loop iteratively finds the optimal control gain for the maximizer $w$, that is

$$A_{i,j}^T P_{i,j} A_{i,j} - P_{i,j} + Q_i - \gamma^2 L_{i,j}^T L_{i,j} = 0, \tag{12a}$$

$$L_{i,j+1} = (\gamma^2 I_q - D^T P_{i,j} D)^{-1} D^T P_{i,j} A_i. \tag{12b}$$

The inner-loop PO 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. The following theorem says that the inner loop globally converges to $P_i$ with a linear convergence rate, whose proof can be found in Appendix C of Cui and Jiang (2022).

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

$$\mathrm{Tr}(P_i - P_{i,j+1}) \leq \beta(K_i)\,\mathrm{Tr}(P_i - P_{i,j}). \tag{13}$$

It follows from Theorem 2 that $\|P_i - P_{i,j+1}\|_F \leq \beta^j(K_i)\sqrt{n}\|P_i - P_{i,1}\|_F$.

### 3.3. Uniform Convergence of the Dual-Loop Algorithm

For the dual-loop algorithm, the inner-loop iteration linearly converges to the optimal solution $P_i$ with the rate dependent on $K_i$. Since $K_i$ is updated iteratively, it is required that the inner loop enters the given neighborhood of $P_i$ within a constant number of steps, independent of $K_i$. The uniform convergence guarantees that the required number of inner-loop iterations do not grow explosively as the outer-loop iteration increases to infinity. The uniform convergence is given in the following theorem, whose proof is given in Appendix D of Cui and Jiang (2022).

**Theorem 3** *For any $i \geq 1$ and $\epsilon > 0$, there exists $\bar{j}$ independent of $i$, such that for all $j \geq \bar{j}$, $\|P_{i,j} - P_i\|_F \leq \epsilon$.*

**Algorithm 1:** Model-Based PO Algorithm

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

## 4. Robustness Analysis for the Dual-Loop Algorithm

In the previous section, the exact PO algorithm was introduced in the sense that the accurate knowledge of system matrices $(A, B)$ is required to implement the algorithm. In practice, however, we cannot access such an accurate model, and for the outer and inner loops, the updates of the controllers in (8b) and (12b) are subjected to noise. Such noise may be induced by noisy input-state data Pang and Jiang (2021) and/or modeling errors Åström and Wittenmark (1997); Tu and Recht (2019). In this section, using the well-known concept of ISS in nonlinear control, we will analyze the robustness of the dual-loop PO algorithm in the presence of disturbance.

### 4.1. Robustness Analysis for the Outer Loop

The exact outer loop iteration is shown in (8), and in the presence of disturbance it is modified as

$$\hat{A}_i^T \hat{U}_i \hat{A}_i - \hat{P}_i + \hat{Q}_i = 0, \tag{14a}$$

$$\hat{K}_{i+1} = (R + B^T \hat{U}_i B)^{-1} B^T \hat{U}_i A + \Delta K_{i+1}, \tag{14b}$$

where

$$\hat{A}_i = A - B\hat{K}_i, \;\; \hat{Q}_i = Q + \hat{K}_i^T R \hat{K}_i, \tag{15}$$

$\Delta K_i$ is the disturbance at the $i$th iteration, and the "hat" is used to distinguish the sequences generated by the exact (8) and inexact (14) outer-loop iterations. By considering (14) as a discrete-time nonlinear system with the state $\hat{P}_i$ and input $\Delta K_i$, the following theorem says that (14) is inherently robust to $\Delta K_i$ in the sense of small-disturbance ISS Pang and Jiang (2021); Pang et al. (2022a). See Appendix E in Cui and Jiang (2022) for the proof.

**Theorem 4** *For any $\hat{K}_1 \in \mathcal{W}$, there exists $d(\hat{K}_1) > 0$, such that if $\|\Delta K\|_\infty < d(\hat{K}_1)$, system (14) is ISS.*

### 4.2. Robustness Analysis for the Inner Loop

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

$$\hat{A}_{i,j}^T \hat{P}_{i,j} \hat{A}_{i,j} - \hat{P}_{i,j} + \hat{Q}_i - \gamma^2 \hat{L}_{i,j}^T \hat{L}_{i,j} = 0, \tag{16a}$$

$$\hat{L}_{i,j+1} = (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T \hat{P}_{i,j} \hat{A}_i + \Delta L_{i,j+1}. \tag{16b}$$

Here, $\hat{A}_{i,j} = A - B\hat{K}_i + D\hat{L}_{i,j}$, $\Delta L_{i,j+1}$ denotes the disturbance to the inner loop iteration, and the "hat" emphasizes that the corresponding sequences are generated by the inexact iteration. With the inexact inner loop at hand, the following theorem shows that the inner loop iteration (16) is robust to disturbance $\Delta L_{i,j}$ in the sense of small-disturbance ISS. See Appendix F of Cui and Jiang (2022) for the proof.

**Theorem 5** *There exists $e(\hat{K}_i) > 0$, such that if $\|\Delta L_i\|_\infty < e(\hat{K}_i)$, system (16) is ISS.*

## 5. Learning-Based Policy Optimization

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

$$u_t = -\hat{K}_1 x_t + \sigma_u \eta_t, \quad \eta_t \sim \mathcal{N}(0, I_m). \tag{17}$$

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

For any $n$-dimensional real symmetric matrix $X$, along the trajectories of system (1), we have

$$\begin{aligned} x_{t+1}^T X x_{t+1} = {} & x_t^T A^T X A x_t + u_t^T B^T X B u_t + w_t^T D^T X D w_t \\ & + 2u_t^T B^T X A x_t + 2w_t^T D^T X A x_t + 2u_t^T B^T X D w_t, \end{aligned} \tag{18a}$$

$$X x_{t+1} = X A x_t + X B u_t + X D w_t. \tag{18b}$$

By vectorizing, one can rewrite (18) as

$$\begin{aligned} & \operatorname{vecv}^T(x_{t+1}) \operatorname{vecs}(X) = \operatorname{vecv}^T(x_t) \operatorname{vecs}(A^T X A) + \operatorname{vecv}^T(u_t) \operatorname{vecs}(B^T X B) \\ & + 2(x_t^T \otimes u_t^T) \operatorname{vec}(B^T X A) + w_t^T D^T X D w_t + 2w_t^T D^T X A x_t + 2u_t^T B^T X D w_t, \end{aligned} \tag{19a}$$

$$(x_{t+1}^T \otimes I_n) \operatorname{vec}(X) = (x_t^T \otimes I_n) \operatorname{vec}(XA) + (u_t^T \otimes I_n) \operatorname{vec}(XB) + XDw_t. \tag{19b}$$

Define $\phi_t$, $\phi_t'$, $\Gamma$, and $\Gamma'$ as

$$\phi_t := \left[\operatorname{vecv}^T(x_t), \operatorname{vecv}^T(u_t), 2(x_t^T \otimes u_t^T), 1\right]^T, \quad \phi_t' := \left[x_t^T, u_t^T\right]^T, \tag{20a}$$

$$\Gamma(X) := [\operatorname{vecs}^T(A^T X A), \operatorname{vecs}^T(B^T X B), \operatorname{vec}^T(B^T X A), \operatorname{Tr}(D^T X D)]^T, \tag{20b}$$

$$\Gamma'(X) := \left[\operatorname{vec}^T(XA), \operatorname{vec}^T(XB)\right]^T. \tag{20c}$$

Multiplying (19a) with $\phi_t$ and (19b) with $\phi_t' \otimes I_n$, and taking the expectation of both sides yield

$$\mathbb{E}\left[\phi_t \phi_t^T \Gamma(X) - \phi_t \operatorname{vecv}^T(x_{t+1}) \operatorname{vecs}(X) | x_t, u_t\right] = 0, \tag{21a}$$

$$\mathbb{E}\left[(\phi_t' \phi_t'^T \otimes I_n)\Gamma'(X) - (\phi_t' x_{t+1}^T \otimes I_n) \operatorname{vec}(X) | x_t, u_t\right] = 0. \tag{21b}$$

We now need the following assumption:

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

**Remark 1** *Assumption 3 is reminiscent of the persistent excitation (PE) condition in adaptive control Jiang et al. (2021); Åström and Wittenmark (1997). Similar PE conditions can be found in the literature on learning-based control Jiang and Jiang (2017); Lewis and Liu (2013); Liu et al. (2021); Pang et al. (2022b); Cui et al. (2021a).*

Taking the expectation of (21) with respect to the invariant probability measure $\pi$ and using Assumption 3, we have

$$\Gamma(X) = \Phi^\dagger \Xi \,\mathrm{vecs}(X), \quad \Gamma'(X) = (\Phi')^\dagger \Xi' \,\mathrm{vec}(X). \tag{22}$$

where

$$\Phi = \mathbb{E}_\pi\left[\phi_t\phi_t^T\right], \Xi = \mathbb{E}_\pi\left[\phi_t \,\mathrm{vecv}^T(x_{t+1})\right], \Phi' = \mathbb{E}_\pi\left[\phi_t'\phi_t'^T \otimes I_n\right], \Xi' = \mathbb{E}_\pi\left[\phi_t' x_{t+1}^T \otimes I_n\right]. \tag{23}$$

In addition, we use a finite number of trajectory data to approximate $\Phi$, $\Xi$, $\Phi'$, and $\Xi'$, that is

$$\hat{\Phi}_T = \frac{1}{T}\sum_{t=1}^T \phi_t\phi_t^T, \hat{\Xi}_T = \frac{1}{T}\sum_{t=1}^T \phi_t \,\mathrm{vecv}^T(x_{t+1}), \hat{\Phi}'_T = \frac{1}{T}\sum_{t=1}^T \phi_t'\phi_t'^T \otimes I_n, \hat{\Xi}'_T = \frac{1}{T}\sum_{t=1}^T \phi_t' x_{t+1}^T \otimes I_n. \tag{24}$$

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

$$\lim_{T\to\infty} \hat{\Phi}'_T = \Phi', \quad \lim_{T\to\infty} \hat{\Xi}'_T = \Xi', \quad \lim_{T\to\infty} \hat{\Phi}_T = \Phi, \quad \lim_{T\to\infty} \hat{\Xi}_T = \Xi. \tag{25}$$

Then, by (22), $\Gamma'(X)$ and $\Gamma(X)$ are approximated as follows:

$$\hat{\Gamma}(X) = \hat{\Phi}_T^\dagger \hat{\Xi}_T \,\mathrm{vecs}(X), \quad \hat{\Gamma}'(X) = (\hat{\Phi}'_T)^\dagger \hat{\Xi}'_T \,\mathrm{vec}(X). \tag{26}$$

According to the definitions of $\Gamma$ and $\Gamma'$ in (20), and their relations to $X$ in (22), the components of $\Gamma$ and $\Gamma'$ can be recovered as

$$\mathrm{vecs}(A^T X A) = [(\Phi)^\dagger]_{1,n_1}\Xi \,\mathrm{vecs}(X),\ \mathrm{vecs}(B^T X B) = [(\Phi)^\dagger]_{n_1+1,n_2}\Xi \,\mathrm{vecs}(X) \tag{27a}$$

$$\mathrm{vec}(B^T X A) = [(\Phi)^\dagger]_{n_2+1,n_3}\Xi \,\mathrm{vecs}(X), \tag{27b}$$

$$\mathrm{vec}(XA) = [(\Phi')^\dagger]_{1,n_4}\Xi' D_n \,\mathrm{vecs}(X),\ \mathrm{vec}(XB) = [(\Phi')^\dagger]_{n_4+1,n_5}\Xi' D_n \,\mathrm{vecs}(X) \tag{27c}$$

where $n_l (l = 1, \cdots, 5)$ can be determined by the dimensions of the matrices $A$ and $B$, and $D_n$ is the duplication matrix ($\mathrm{vec}(X) = D_n \,\mathrm{vecs}(X)$) in (Magnus and Neudecker, 2007, pp. 56).

Replacing $X$ in (27) by $P_{i,j}$ and substituting it into the vectorization of (12a) yield a linear equation for $P_{i,j}$

$$\begin{aligned}
&\Big\{[(\Phi)^\dagger]_{1,n_1}\Xi - D_n^\dagger[(K_i^T \otimes I_n)T_{mn} + I_n \otimes K_i^T][(\Phi)^\dagger]_{n_2+1,n_3}\Xi \\
&+ D_n^\dagger[(L_{i,j}^T D^T \otimes I_n)T_{nn} + I_n \otimes L_{i,j}^T D^T][(\Phi')^\dagger]_{1,n_4}\Xi' D_n \\
&- D_n^\dagger\left[(L_{i,j}^T D^T \otimes K_i^T)T_{nm} + K_i^T \otimes L_{i,j}^T D^T\right][(\Phi')^\dagger]_{n_4+1,n_5}\Xi' D_n \\
&+ D_n^\dagger(K_i^T \otimes K_i^T)D_m[(\Phi)^\dagger]_{n_1+1,n_2}\Xi \\
&+ D_n^\dagger(L_{i,j}^T D^T \otimes L_{i,j}^T D^T)D_n - I_{(1+n)n/2}\Big\}\,\mathrm{vecs}(P_{i,j}) + \mathrm{vecs}(Q_i - \gamma^2 L_{i,j}^T L_{i,j}) = 0,
\end{aligned} \tag{28}$$

**Algorithm 2:** Learning-Based PO Algorithm

Initialize $\hat{K}_1 \in \mathcal{W}$, the length of the sampled trajectory $T$, and the exploration variance $\sigma_u^2$ ;
Collect data from (1) with exploratory input (17) ;
Construct $\hat{\Phi}'_T$, $\hat{\Phi}_T$, $\hat{\Xi}'_T$, and $\hat{\Xi}_T$ defined in (24) ;
**for** $i \leq \bar{i}$ **do**
  Set $L_{i,j} = 0$;
  $\hat{Q}_i = C^T C + \hat{K}_i^T R \hat{K}_i$;
  **for** $j \leq \bar{j}$ **do**
    Get $\hat{P}_{i,j}$ by solving (28);
    Get $\widehat{P_{i,j}A}$ and $\widehat{P_{i,j}B}$ by (27c). ;
    Update $\hat{L}_{i,j+1} = (\gamma^2 I_q - D^T \hat{P}_{i,j} D)^{-1} D^T (\widehat{P_{i,j}A} - \widehat{P_{i,j}B} K_i)$;
  **end**
  $\hat{U}_{i,\bar{j}} = \hat{P}_{i,\bar{j}} + \hat{P}_{i,\bar{j}} D (\gamma^2 I_q - D^T \hat{P}_{i,\bar{j}} D)^{-1} D^T \hat{P}_{i,\bar{j}}$;
  Get $\widehat{B^T U_{i,\bar{j}} B}$ and $\widehat{B^T U_{i,\bar{j}} A}$ by (27a) and (27b) ;
  $\hat{K}_{i+1} = (R + \widehat{B^T U_{i,\bar{j}} B})^{-1} \widehat{B^T U_{i,\bar{j}} A}$ ;
**end**

where $T_{mn}$, $T_{nn}$, and $T_{nm}$ are commutation matrices defined in (Magnus and Neudecker, 2007, pp. 54). Consequently, $P_{i,j}$ can be approximated by solving the linear equation (28). The details are shown in Algorithm 2. Since the data matrices $\hat{\Phi}'_T$, $\hat{\Xi}'_T$, $\hat{\Phi}_T$, and $\hat{\Xi}_T$ are reused throughout the policy iteration, the proposed algorithm is off-policy.

## 6. Simulation

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

$$A = \begin{bmatrix} 1 & 0 & -5 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 3 & 1 \\ -1 & 0 & 2 \end{bmatrix}, D = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix} \tag{29}$$

The matrices related to the controlled 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 $\gamma = 5$. The simulation is implemented on a desktop computer with a CPU Intel i7-9700K CPU @ 3.60GHz. The computer has two 16GB 3200MHz DDR4 RAMs and the numeric computing platform is MATLAB 2020b. $\bar{i} = 20$ and $\bar{j} = 20$.

The robustness of Algorithm 1 in the presence of disturbance at each iteration is validated first. For each outer and inner loop iteration, the entries of the disturbances $\Delta K_i$ and $\Delta L_{i,j}$ are samples from a standard Gaussian distribution and then their Frobenius norms are normalized to 0.1. In Fig. 2, it is seen that with the disturbance at each outer and inner loop iteration, the generated controller and the corresponding cost matrix approach the optimal solution and finally enters a neighborhood of the optimal controller $K^*$ and cost matrix $P^*$. The $\mathcal{H}_\infty$-norm of the closed-loop system is smaller than the threshold throughout the PO process. These numerical results are consistent with the developed theoretical results in Theorems 4 and 5.

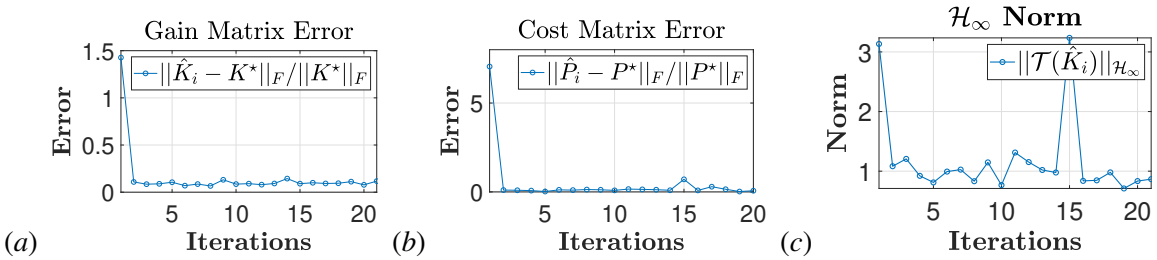


Figure 2: Robustness of Algorithm 1 when $\|\Delta K\|_\infty = 0.1$ and $\|\Delta L_i\|_\infty = 0.1$.

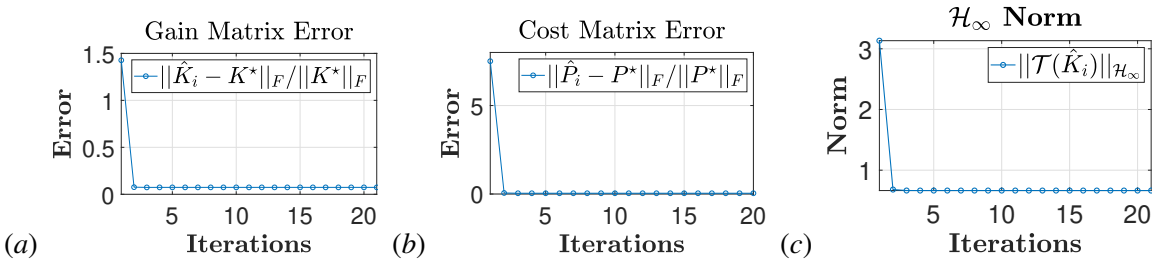


Figure 3: Using Algorithm 2, the solutions of each iteration approach the optimal solution, and the $\mathcal{H}_\infty$-norm is smaller than the threshold.

Algorithm 2 is implemented to learn a robust suboptimal controller for system (1). The length of the sampled trajectory is $T = 1000$, i.e. 1000 data are collected in total to train the robust optimal controller. The standard deviation of the exploratory noise is $\sigma_u = 5$. In Fig. 3 the algorithm converges at the 3rd iteration. At the 20th iteration, $\|\hat{K}_{20} - K^*\|_F/\|K^*\|_F = 7.37\%$ and $\|\hat{P}_i - P^*\|_F/\|P^*\|_F = 4.29\%$. Therefore, the proposed off-policy PO algorithm can still approximate the optimal solution when the system is disturbed by additive Gaussian noise.

## 7. Conclusion

In this paper, we have proposed a novel robust dual-loop PO algorithm for a class of linear discrete-time systems to handle model mismatches and disturbances arising from sim-to-real gap. It is demonstrated that the dual-loop algorithm uniformly converges to the optimal solution. When the algorithm is subject to disturbances, it is proved that the algorithm possesses the property of small-disturbance ISS. Specifically, given an initial admissible control policy, the control policies generated by the proposed PO algorithm ultimately enter a small neighborhood of the optimal solution, given that the disturbance is sufficiently small. Based on these model-based theoretical results, and without knowing the accurate system matrices, we have also proposed a novel learning-based PO algorithm to learn the optimal controllers directly from data. Numerical examples have been provided and the efficacy of the proposed methods is demonstrated.

## Acknowledgments

This work has been supported in part by the NSF Grants CNS-2148304 and ECCS-2210320, and in part by the AFOSR Grant FA9550-19-1-0353.

## References

Pierre Apkarian, Dominikus Noll, and Aude Rondepierre. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control via nonsmooth optimization. *SIAM Journal on Control and Optimization*, 47(3):1516–1546, 2008.

Tamer Başar and Pierre Bernhard. *$H_\infty$-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer, New York, USA, 1995.

Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint, arXiv:1911.04672*, 2019.

Jingjing Bu, Afshin Mesbahi, and Mehran Mesbahi. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint, arXiv:2006.09178*, 2020.

Asaf B Cassel and Tomer Koren. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1304–1313. PMLR, 2021.

Leilei Cui and Zhong-Ping Jiang. A reinforcement learning look at risk-sensitive linear quadratic gaussian control. *arXiv preprint, arXiv:2212.02072*, 2022.

Leilei Cui, Kaan Ozbay, and Zhong-Ping Jiang. Combined longitudinal and lateral control of autonomous vehicles based on reinforcement learning. In *American Control Conference (ACC)*, pages 1929–1934, 2021a.

Leilei Cui, Shuai Wang, Jingfan Zhang, Dongsheng Zhang, Jie Lai, Yu Zheng, Zhengyou Zhang, and Zhong-Ping Jiang. Learning-based balance control of wheel-legged robots. *IEEE Robotics and Automation Letters*, 6(4):7667–7674, 2021b.

John Doyle, Keith Glover, Pramod Khargonekar, and Bruce Francis. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, 1989.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2021.

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

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Towards a theoretical foundation of policy optimization for learning control policies. *arXiv preprint arXiv:2210.04810*, 2022.

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

Zhong-Ping Jiang and Yu Jiang. *Robust Adaptive Dynamic Programming*. Wiley-IEEE Press, 2017.

Zhong-Ping Jiang and Tengfei Liu. Small-gain theory for stability and control of dynamical networks: A survey. *Annual Reviews in Control*, 46:58–79, 2018. ISSN 1367-5788.

Zhong-Ping Jiang, C. Prieur, and A. Astolfi (Editors). *Trends in Nonlinear and Adaptive Control: A Tribute to Laurent Praly for His 65th Birthday,*. Springer Nature, NY, USA, 2021.

David L. Kleinman. On an iterative technique for riccati equation computations. *IEEE Transactions on Automatic Control*, 13(1):114–115, 1968.

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

Frank L. Lewis and Derong Liu. *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Wiley-IEEE Press, NJ, USA, 2013.

Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67(12):6429–6444, 2021.

Tong Liu, Leilei Cui, Bo Pang, and Zhong-Ping Jiang. Data-driven adaptive optimal control of mixed-traffic connected vehicles in a ring road. In *60th IEEE Conference on Decision and Control (CDC)*, pages 77–82, 2021. doi: 10.1109/CDC45484.2021.9683024.

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

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

Denis Mustafa and Dennis S. Bernstein. LQG cost bounds in discrete-time $\mathcal{H}_2/\mathcal{H}_\infty$ control. *Transactions of the Institute of Measurement and Control*, 13(5):269–275, 1991.

Bo Pang and Zhong-Ping Jiang. Robust reinforcement learning: a case study in linear quadratic regulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9303–9311, May 2021.

Bo Pang, Tao Bian, and Zhong-Ping Jiang. Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, 67(1):504–511, 2022a.

Bo Pang, Leilei Cui, and Zhong Ping Jiang. Human motor learning is robust to control-dependent noise. *Biological Cybernetics*, 116(3):307–325, 2022b.

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

Eduardo D Sontag. Remarks on input to state stability of perturbed gradient flows, motivated by model-free feedback control learning. *Systems & Control Letters*, 161:105138, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2nd ed. edition, 2018.

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

George 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 Transactions on Automatic Control*, 11(2):228–238, 1966.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence. *arXiv e-prints*, art. arXiv:1910.09496, October 2019.

Kaiqing Zhang, Bin Hu, and Tamer 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*, volume 33, pages 22056–22068, 2020.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: implicit regularization and global convergence. *SIAM J. Control Optim.*, 59(6):4081–4109, 2021a.

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer 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*, volume 34, pages 2949–2964. Curran Associates, Inc., 2021b.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, Princeton, New Jersey, 1996.

Karl J. Åström and Bjorn Wittenmark. *Adaptive control, 2nd Edition*. Addison-Wesley, MA, USA, 1997.