Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ LQ Games for Robust Policy Optimization Under Unknown Dynamics

Leilei Cui and Olalekan Ogunmolu, *Member, IEEE*

***Abstract*—We consider some aspects of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control in a policy optimization setting. We study the convergence and robustness properties of our proposed policy scheme for autonomous systems described by stochastic differential equations with non-trivial additive Brownian motion as a disturbance. We then propose efficiently *learning robustly stabilizing optimal control policies* for such systems when the dynamics is unknown. We evaluate our proposed schemes on two- and three-link kinematic chains. Our evaluations demonstrate robust steady-state convergence to equilibrium under worst-case disturbance and Brownian motion alike. This policy optimization scheme is well-suited to reinforcement learning, and learning-enabled control systems where modeling errors, unknown dynamics, parametric and non-parametric uncertainties typically hamper system operations.**

***Index Terms*—Iterative Learning Control, $\mathcal{H}_\infty$ Control, Robust Control, Machine Learning**

I. Introduction

We are poised with *robustly stabilizing optimal policies* for stochastic dynamical systems (i) with *unknown state transition and control matrix parameters*; (ii) *exhibiting non-parametric uncertainties*; or (iii) *exhibiting parametric uncertainties*. For parametric uncertainties, our policy optimization (PO) scheme *learns stabilizing and optimal policies for systems with imperfect information*. Techniques for systems possessing non-parametric uncertainties in literature typically assume an idealization of the noise as an additive stochastic process with zero correlation time (white noise) – unrealistic for most biological and cyberphysical systems. Here, the noise is an additive stochastic Brownian process with a nontrivial correlation structure. For non-parametric uncertainties that are additive in nature, it learns a robust policy for the control problem. When the system dynamics is altogether unknown, in an iterative fashion it learns the associated system model.

Control design with $\mathcal{H}_2$ or LQG lend many applicability to real-world stochastic control processes. These controllers construct linear systems' feedback compensators by minimizing a quadratic cost in the presence of a fixed noise (covariance) intensity (usually an additive Gaussian noise) that is subject to the system dynamics [1]. In this form, $\mathcal{H}_2$ controllers have found applications in various problem domains since their introduction [2] such as robotics [3], autonomous vehicle [4], and recently in motor control [5] *inter alia*.

Leilei Cui is with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201. Lekan Molu is with Microsoft Research, 300 Lafayette Street, New York, NY 10012, USA.(email: l.cui@nyu.edu, lekanmolu@microsoft.com).

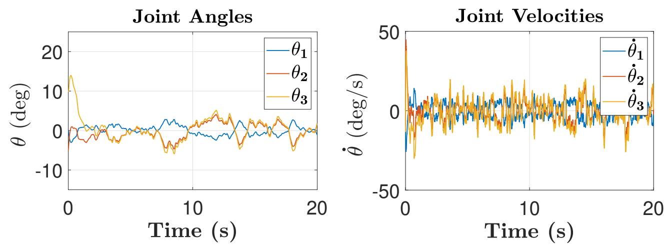

Fig. 1: **Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy optimization** on a 3-link pendulum subject to an additive stochastic Brownian motion. The policy fails to regulate the system around equilibrium.

$\mathcal{H}_2$ control systems provide a few interesting properties. Firstly, the optimal feedback controller is a linear time-varying function of the state variable. As is well-known, linear time-varying systems tend to possess parametric and dynamic uncertainties that must be carefully managed throughout the life-cycle of a control process. Time-varying controllers deployed on systems with parametric or dynamic uncertainties (especially when unknown aforetime) are difficult to stabilize and they notoriously have no formal guarantees (e.g. see the counterexamples of [6]).

In light of these drawbacks, various authors have proposed robust time-domain schemes for $\mathcal{H}_2$ controllers. Of importance is Jacobson's pioneering work on linear exponential quadratic Gaussian control design [7]. Here, by taking the exponent of the LQ cost, a designer obtains stabilizing control laws that provide a measure of risk aversion or risk propensity. This formulation is particularly well-suited to certain economic decision processes. While Jacobson obtained a smooth solution to the associated Hamilton-Jacobi-Bellman equation [7], Duncan [8] generalized an elementary solution using a squares completion and Radon-Nikodym derivative scheme.

Khargonekar et al. [9] and Bernstein et al. [10] proposed an algebraic Riccati equation (ARE) solution under an $\mathcal{H}_\infty$ attenuation constraint of the closed-loop transfer function of a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ system. Mustafa [11], via a minimum entropy approach, showed that a maximum entropy/$\mathcal{H}_\infty$ control is equivalent to a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem. Basar et al. [12] solved mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem in a linear quadratic zero-sum differential game setting. It should be noted that most works require an accurate system model in order to solve the associated nonlinear indefinite ARE in an iterative fashion [13], [14], [15], [16]. Such models are typically difficult to obtain for complex systems.

Notably, reinforcement learning algorithms solve these problems under unknown system models. However, it is not clear what convergence guarantees thay do possess. With policy and value iterations in an adaptive dynamic programming framework, reinforcement learning has found applications in linear, nonlinear, and periodic continuous-time systems particularly when handling optimal stabilization and output regulation problems [17], [18], [19], [20], [21], [22], [23], [24]. Pang et al. [25] studied the robustness of policy iteration under process noise in an input-to-state stability framework and showed that policy iteration finds an approximate solution to the optimal control problem. Utilizing the gradient of the performance index with respect to the parameters of the control policy, policy gradient algorithms were proposed in [26], [27], [28]. Along these lines of work, a *learned controller* optimizes the given performance index in an $\mathcal{H}_2$ sense only. As a result, it is not clear what robustness properties the resulting controller possesses.

Robust reinforcement learning based on mixed $\mathcal{H}_2/\mathcal{H}_\infty$ minimizes the performance index with guaranteed policy robustness to worst-case disturbance. For example, [29], [30], [31], [32] proposed adaptive dynamic programming approaches for zero-sum differential games. The additional $\mathcal{H}_\infty$ norm constraint imposed on the performance index in fact ensures robustness as the learning algorithm approaches infinity. By estimating the gradient with zero-order methods, derivative-free algorithms were used to directly search for optimal policy parameters in [33], [34]. In [35], [14], the authors proposed an on-policy reinforcement learning scheme in a zero-sum LQ differential game setting based on gradients estimated by zeroth-order schemes. Fazel et al. [36] estimated gradients (in a zeroth-order sense) of the cost difference between nominal and perturbed polices. For stochastic systems, zeroth-order methods are a special case of Monte Carlo methods – essentially high variance methods that produce slow learning [37]. Robust *policy optimization* in LQ zero-sum two-player game settings have also found applications in general simulated robotics and RL video game problems [35], [38], [39], [40]. Iteratively updating a controller over a performance index, these *policy optimization* schemes optimize the system's $\mathcal{H}_2$ norm. In the presence of additive noise to the system, however, it is not clear that these frameworks provide robustness, especially under unknown system dynamics. To buttress this point, consider the three-link pendulum under a classical mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy optimization scheme, but with a dynamic uncertainty present in the form of a Brownian disturbance. As seen in Fig. 1, the PO scheme fails to stabilize the pendulum along the equilibrium position (here $(0, 0, 0)$) for all three joint angles.

Classical LQ two-player games require an accurate measurement of the control inputs of the two players – which are rarely a given for many physical, chemical, and biological systems. In addition, disturbance and uncertainties are the norm rather than the exception for many feedback systems. It seems reasonable to place the convergence and robustness analyses of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ LQ two-player zero sum game systems under imperfect information on a rigorous mathematical footing. This is the essence of this article.

In this article, we are concerned with robust stabilization of optimal control problems in the presence of incorrect model assumptions, model parameters, or when there is an unknown model altogether. Revisiting mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control [41], we introduce an iterative solution to the cost matrix of $\mathcal{H}_2$ control problems for two-player zero-sum differential games; we learn robustly stabilizing optimal policies in the presence of a worst-case disturbance in an *iterative optimization scheme*. Our scheme imbues policy optimization schemes with a robustness-preserving metric in a two-player zero-sum linear exponential quadratic Gaussian (LEQG) framework [7]. Our inquiry is motivated by the lack of robustness guarantees of time-domain linear quadratic Gaussian (LQG) frameworks [6], [8] and the well-posedness of $\mathcal{H}_\infty$ control objectives in the presence of a worst-case disturbance for multivariable robust control [42], [43].

The rest of this article is structured as follows: in Section II, we set up notation, and introduce the problem. In Section III, we present an iterative optimization scheme for solving a model-based mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy optimization problem, and a learning-based mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control scheme is presented in Section IV. In Section V, we analyze the convergence properties of our proposed algorithm. The robustness of the iterative algorithm is analyzed in Section VI. Finally, we demonstrate the efficacy of our proposed algorithm with numerical results in Section VII. We discuss our findings and draw conclusions in Section VIII. All theoretical machinery needed for proving our main results are given in the appendices.

II. Background

In this section, we first set up notations that are commonly used throughout this article, give a few preliminary results for some of the machinery needed for proving our main results, then formally introduce the problem formulation.

A. Notations

We adopt vector-matrix notations throughout. Conventions: capital Roman letters are matrices; in lower-case they are vectors. Exceptions: p, q, n, m are matrix or vector indices or dimensions. Unless otherwise stated, optimization iteration indices are denoted by i or j. $A := B$ means that A is defined by B, and $A =: B$ implies that A defines B. We let $\mathbb{R}$ denote the set of real numbers, $\mathbb{N}$ the set of natural numbers, and $\mathbb{N}_+$ the set of positive integers. The set of all symmetric matrices with dimension n is denoted by $\mathbb{S}^n$ while $\|\cdot\|$ denotes the 2-norm of a vector or the induced matrix norm. We let $\|\cdot\|_F$ denote the Frobenius norm of a matrix.

For a matrix $P \in \mathbb{S}^n$, $\text{vecs}(P) := [p_{11}, 2p_{12}, \cdots, 2p_{1n}, p_{22}, \cdots, p_{nn}]^T$, where p_{ij} is the ith row and jth column entry of P. Let the operator $\text{vec}(A) := [a_1^T, \cdots, a_n^T]^T$, where a_i is the ith column of the matrix A; and let $\text{vecv}(x) := [x_1^2, x_1x_2, \cdots, x_1x_n, x_2^2, x_2x_3, \cdots, x_n^2]^T$. The Kronecker product is denoted by $\otimes$. The sub-matrix of the matrix A that is comprised of the rows between the ith and jth rows is denoted by $[A]_{i:j}$. The maximum and minimum singular values of a matrix T are respectively denoted by $\bar{\sigma}(T)$ and $\underline{\sigma}(T)$. For $X \in \mathbb{R}^{m\times n}$ and $\delta > 0$, $\mathcal{B}(X, \delta) := \{Y \in \mathbb{R}^{m\times n} \,|\, \|Y - X\|_F \leq \delta\}$. The identity matrix with dimension n is I_n. For the transfer function

$G(s)$, its $\mathcal{H}_\infty$ norm is a bounded linear operator defined as $\|G\|_{\mathcal{H}_\infty} = \sup_{\omega \in \mathbb{R}} \bar{\sigma}(G(j\omega))$.

B. Mixed Design as the LEQG Problem

Consider the following stochastic autonomous system

$$dx(t) = Ax(t)\,dt + Bu(t)\,dt + Ddw(t),\; x(0) = x_0, \quad \text{(1a)}$$
$$z(t) = Cx(t) + Eu(t), \quad \text{(1b)}$$

where $x(t) \in \mathbb{R}^n$ is the state, $x(0)$ is the initial state, $u(t) \in \mathbb{R}^m$ is the control input, and $w(t) \in \mathbb{R}^q$ is the standard Wiener process defined over the probability space $(\Omega, \mathcal{F}, \mathcal{P})$[1], and the system output is $z(t) \in \mathbb{R}^p$. Matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ while matrix $D \in \mathbb{R}^{n\times q}$. By design, $C \in \mathbb{R}^{p\times n}$ and $E \in \mathbb{R}^{p\times m}$ are known. The transfer function from the disturbance $w(t)$ to $z(t)$ is

$$\mathcal{T}(K)(s) = (C - EK)(sI_n - A + BK)^{-1}D. \quad (2)$$

The linear exponential quadratic Gaussian (LEQG) stochastic optimal control problem (SOCP) [7], [44] is an adaptation of the standard LQG problem to processes that are *risk-sensitive*. Taking an exponent of the standard LQG cost and penalizing it with a scalar term γ^{-2}, the LEQG cost is

$$\mathcal{J}(K) = \limsup_{\tau\to\infty} \frac{2\gamma^2}{\tau} \log \mathbb{E}\left[\exp\left(\frac{1}{2\gamma^2}\int_0^\tau z^T(t)z(t)dt\right)\right], \quad (3)$$

where sign(γ) for a $\gamma \in \mathbb{R}$ dictates a *risk-seeking* (i.e. $\gamma > 0$), *risk-avoidance* (i.e. $\gamma < 0$), or *risk-neutrality* (i.e. $\gamma = 0$, whereupon we have the standard LQP problem) behavior for an ensuing controller to be sought. In [14], the authors pointed out that the standard LEQG problem admits the form of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ *suboptimal* optimal control problem if written in the form (3) for the linear time-invariant problem (1).

The mixed design problem consists in minimizing an upper bound on the $\mathcal{H}_2$ cost (3) subject to the $\mathcal{H}_\infty$ robustness constraint, $\|\mathcal{T}(K)\|_{\mathcal{H}_\infty} \le \gamma$ so that all stabilizing feedback gains, K, for system (1) can be realized within the set

$$\mathcal{K} = \{K | (A - BK) \text{ Hurwitz}, \|\mathcal{T}(K)\|_{\mathcal{H}_\infty} \le \gamma\}. \quad (4)$$

Formally, the control optimization problem is to find the set of stabilizing gains K via

$$\min_K \mathcal{J}(K) \text{ such that } K \in \mathcal{K}. \quad (5)$$

In (3) the noise intensity term must be well-conditioned to guarantee the existence of a solution to the SOCP. Therefore, we make the following standard assumptions.

Assumption 1. *(A, B) is stabilizable, and (C, A) is observable. In addition, $\gamma > \gamma_\infty$, for a $\gamma_\infty = \inf\{\gamma > 0 \mid \min_u \max_w J(0, u, w) \le 0\}$.*

Assumption 2. *$E^T E = R \succ 0$ and $E^T C = 0$.*

Remark 1. *Assumption 1 implies that the indefinite ARE associated with (3) has a unique positive definite solution and hence guarantees the existence of a stabilizing control law u. Assumption 2 simplifies the derivation of the solution to the indefinite ARE.*

[1]Here, Ω is the sample space, $\mathcal{F}$ is the σ-algebra i.e. the natural filtration for the Brownian motion, $\mathcal{P}$ is the probability measure for $t \in [0, T]$ where $T > 0$ is fixed.

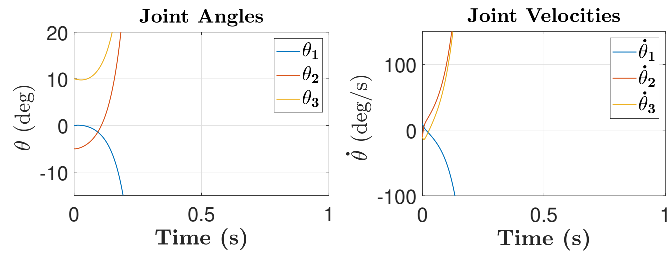

Fig. 2: An LQG controller applied to a 3-link pendulum whose dynamics is corrupted by the worst-case disturbance. The controller fails to drive the system into steady-state.

Proposition 1. *Given a constant $\gamma > 0$, the solution to the LEQG problem (3) is*

$$u^\star(x(t)) = -\underbrace{R^{-1}B^T P^\star}_{K^\star} x(t), \quad (6)$$

where $P^\star$ is the unique, stabilizing symmetric positive definite solution to the ARE

$$A^T P^\star + P^\star A + C^T C - P^\star(BR^{-1}B^T - \gamma^{-2}DD^T)P^\star = 0. \quad (7)$$

Proof. Our Riccati equation is an extension of the Riccati equation introduced by Duncan [8, Th. II.1] for the finite-horizon LEQG problem. The controller

$$u(x(t)) = -K_\tau(t)x(t) := -R^{-1}B^T P_\tau(t)x(t) \quad (8)$$

minimizes the finite-horizon cost (by a slight abuse of notation of (3))

$$\mathcal{J}(K) = \mathbb{E}\left[\exp\left(\frac{1}{2\gamma^2}\int_0^\tau z^T z\,dt\right)\right], \quad (9)$$

where $P_\tau(t)$ is the solution to the following Riccati equation

$$-\dot{P}_\tau = A^T P_\tau + P_\tau A - P_\tau(BR^{-1}B^T - \gamma^{-2}DD^T)P_\tau \quad (10)$$

A fortiori, given assumptions 1 and 2, we have $\lim_{\tau\to\infty} P_\tau(t) = P^\star$, and $\lim_{\tau\to\infty} K_\tau(t) = K^\star$ (This conclusion is a special case of [12, Theorem 9.7]. □

Figure 2 illustrates the ineffectiveness of a designed LQG controller in maintaining a 3-link pendulum system's trajectories at steady state as the system dynamics evolves.

C. Mixed Design as an LQ Zero-Sum Differential Game

Let us now consider the following two-player, zero-sum differential game with the quadratic linear cost,

$$\min_{u\in\mathbb{R}^m} \max_{w\in\mathbb{R}^q} J(\cdot) = \int_{t=0}^\infty z^T(t)z(t) - \gamma^2 w^T(t)w(t)\,dt \quad (11)$$

for system (1). The controller $u(t)$ is minimizing while the disturbance $w(t)$ is maximizing. The solution to this differential game is given by Theorem 4.8 and 9.7 of [12], and it is summarized in Proposition 2. Henceforth, for conciseness we abuse notations, dropping the time arguments in $x(t)$ and $u(x(t))$ when the meaning is not diminished in our notations.

Proposition 2. *The respective optimal controllers for the two players at time step t are*

$$u^\star(\cdot) = -R^{-1}B^T P^\star x(t),\ w^\star(t) = \gamma^{-2}D^T P^\star x(t). \tag{12}$$

In addition, $P^\star \succ 0$ is the stabilizing solution of (7).

Proof. This is just a statement of Th. 4.8 and 9.7 in [12]. □

Enforcing the gains K over the set (4), which is in the frequency domain, in general requires a difficult transformation. However, with the bounded real Lemma A.9, we can express a relationship between a Riccati equation solution and a Riccati inequality. The Lemma is given in Appendix A. Now, given the Riccati equation (7) and the equivalence of the $\mathcal{H}_\infty$ norm bound on the system transfer function to the Riccati equation and inequality in Lemma A.9, we conclude that the optimal $K^\star \in \mathcal{K}$. Observe: minimizing the performance index (11) under the worst-case disturbance, the optimal controller u^* can *robustly* improve the system performance w.r.t the $\mathcal{H}_\infty$ norm penalty (this is demonstrated in our experiments).

Remark 2. *Observing Propositions 1 and 2, we see that both the LEQG and zero-sum differential game generate the same robust and optimal controller for the system, c.f. (3) and (12).*

While Duncan [8] proposed the Riccati equation (7), no closed-form solution exists to our knowledge. Conventional LQG control cannot guarantee the robustness of system (3) or (11) as found by Doyle [6]. Before we introduce the learning-based algorithm, we introduce the model-based algorithm whose convergence and robustness analysis is basically the same as the model-free algorithm, that is the essence of the paper.

III. Model-based Iterative Algorithm

We establish the model-based iterative solution to (7) in this section. The ARE (7) is a nonlinear matrix equation that does not have a closed-form solution. In what follows, we propose a two-loop model-based iterative algorithm for computing P^* from a sequence of linear Lyapunov equations.

A. Algorithm Description

The procedure for obtaining the optimal $P^\star$ is now described. Let $i \in \bar{i}$ and $j \in \bar{j}$ denote the iteration indices for the outer and inner loop stages (i.e. update loops for the controller and disturbance respectively) of the algorithm respectively, where $\{\bar{i}, \bar{j}\} \in \mathcal{N}_+$. To aid our derivations, let us first define the following matrices:

$$A_{K_i} = A - BK_i,\ A_i = A_{K_i} + \gamma^{-2}DD^T P_{K_i}, \tag{13a}$$

$$A_{K_i}^j = A_{K_i} + DL_{K_i}^j,\ A^\star = A - BK^\star + DL^\star, \tag{13b}$$

$$Q_{K_i} = C^T C + K_i^T R K_i, \tag{13c}$$

where A_{K_i} is the first player's closed-loop system transition matrix under an arbitrary feedback gain K_i while A_i is the second player's closed-loop system transition matrix. $A_{K_i}^j$ is the closed-loop system's transition matrix with arbitrary gains K_i and L_i; and $A^\star$ is the closed-loop system's transition matrix with the optimal gains $K^\star$ and $L^\star$. The algorithm, described in Algorithm 1, is explained as follows:

- Starting at iteration $i = 1$, set a $K_1 \in \mathcal{K}$ and an $L_{K_1} = 0$. Iterate for K_i until convergence:
  - For $j = \{1, 2, \cdots, \bar{j}\}$, iteratively solve the Riccati equation (7). Call this solution $P_{K_i}^j$;
  - Increment i by 1; then, update the maximizing player's gain $L_{K_i}^{j+1}$ given $P_{K_i}^j$;
- Update the minimizing player's gain K_{i+1} given $P_{K_i}^{\bar{j}}$.

Note that the system matrices (A, B) are required to successfully run Algorithm 1. We defer treatment of when matrices (A, B) are unknown to section IV. Ours is similar to best-response alternating minimax iterative dynamic games of [45] between the two players in (11).

Algorithm 1: Model-Based Iterative Algorithm

```
Initialize K_1 ∈ 𝒦                                  ▷ e.g. pole placement;
for i ≤ ī do
    Set L^1_{K_i} = 0;
    Q_{K_i} = C^T C + K_i^T R K_i;
    for j ≤ j̄ do
        A^j_{K_i} = A − BK_i + DL^j_{K_i}             ▷ Update A^j_{K_i};
        (A^j_{K_i})^T P^j_{K_i} + P^j_{K_i} A^j_{K_i} + Q_{K_i} − γ^2 (L^j_{K_i})^T L^j_{K_i} = 0,
        ▷ Get P^j_{K_i};
        L^{j+1}_{K_i} = γ^{-2} D^T P^j_{K_i}   ▷ Update disturbance gain L^{j+1}_{K_i};
    end
    K_{i+1} = R^{-1} B^T P^{j̄}_{K_i};          ▷ Update control gain K_{i+1} ;
end
```

IV. Model-free Iterative Algorithm

In the this section, based on the results of the last two sections, we propose a learning-based iterative algorithm. From Proposition 1, we see that matrices (A, B) must be exactly known in order to find a stabilizing controller $K^\star$. However, we are concerned with learning an optimal controller $K^\star$ when the matrices (A, B) are unknown. For this algorithm, only the trajectories of state x and control input u collected along system (1) are required. During the data collection phase, suppose the respective control policy is

$$u = -\hat{K}_1 x + \sigma_u \xi,\ d\xi = -\xi dt + dv \tag{14}$$

where $\xi \in \mathbb{R}^m$ is the exploration noise, $\sigma_u > 0$ is a constant, and v is a standard Brownian motion independent of w.

As shown in Algorithm 1, the cost matrix P plays a pivotal role. In order to obtain the value of the cost matrix directly from the collected trajectory data of the system, the derivative of $x^T P x$ is derived. Along the state trajectory of (1), by Ito's formula [46, Lemma 3.2], the derivative of $x^T P x$, where $P \in$

$\mathbb{S}^n$, can be verified to be

$$\begin{aligned}
d(x^T P x) &= (dx)^T P x + x^T P dx + (dx)^T P dx \\
&= x^T (A^T P + PA) x dt + 2x^T P B u dt + 2x^T P D dw \\
&\quad + \underbrace{(Ax + Bu)^T P (Ax + Bu)(dt)^2}_{=0} \\
&\quad + \underbrace{2(dw)^T D^T P (Ax + Bu)(dt)}_{=0} + \underbrace{(dw)^T D^T P D dw}_{=\mathrm{Tr}(D^T P D dw dw^T)} \\
&= x^T (A^T P + PA) x dt + 2x^T P B u dt \\
&\quad + 2x^T P D dw + \mathrm{Tr}(D^T P D) dt.
\end{aligned} \tag{15}$$

We adopt efficient vectorization of (15) so that

$$\begin{aligned}
&d(\mathrm{vecv}^T(x)) \,\mathrm{vecs}(P) = \mathrm{vecv}^T(x)\, \mathrm{vecs}(A^T P + PA) dt \\
&+ 2(x^T \otimes u^T) dt \,\mathrm{vec}(B^T P) + \mathrm{Tr}(D^T P D) dt + 2x^T P D dw.
\end{aligned} \tag{16}$$

Let $\phi(t) = \left[\mathrm{vecv}^T(x), 2(x^T \otimes u^T), 1\right]^T$. Integrating both sides of (16) from 0 to t_f yields

$$\begin{aligned}
&\frac{1}{t_f} \int_0^{t_f} \phi d(\mathrm{vecv}^T(x)) \,\mathrm{vecs}(P) \\
&= \frac{1}{t_f} \int_0^{t_f} \phi \phi^T dt \begin{bmatrix} \mathrm{vecs}(A^T P + PA) \\ \mathrm{vec}(B^T P) \\ \mathrm{Tr}(D^T P D) \end{bmatrix} + \frac{1}{t_f} \int_0^{t_f} 2x^T P D dw.
\end{aligned} \tag{17}$$

Let

$$\begin{aligned}
\hat{\Phi}(t_f) &= \frac{1}{t_f} \int_0^{t_f} \phi(t) \phi^T(t) dt, \\
\hat{\Xi}(t_f) &= \frac{1}{t_f} \int_0^{t_f} \phi d(\mathrm{vecv}^T(x)).
\end{aligned} \tag{18}$$

By Lemmas A.7 and A.8, the following equations hold *almost surely*

$$\lim_{t_f \to \infty} \frac{1}{t_f} \int_0^{t_f} 2x^T P D dw = 0, \ \lim_{t_f \to \infty} \hat{\Phi}(t_f) = \Phi := \mathbb{E}(\phi \phi^T). \tag{19}$$

If we combine (17) and (19), there exists a constant matrix Ξ, such that the following holds *almost surely*

$$\lim_{t_f \to \infty} \hat{\Xi}(t_f) = \Xi. \tag{20}$$

Therefore, from (17), we have

$$\begin{bmatrix} \mathrm{vecs}(A^T P + PA) \\ \mathrm{vec}(B^T P) \\ \mathrm{Tr}(D^T P D) \end{bmatrix} = \Phi^{-1} \Xi \,\mathrm{vecs}(P). \tag{21}$$

Let $n_1 := (n+1)n/2$ and $n_2 := n_1 + mn$. Then, we have

$$\begin{aligned}
\mathrm{vecs}(A^T P + PA) &= [\Phi^{-1}]_{1:n_1} \Xi \,\mathrm{vecs}(P) \\
\mathrm{vec}(B^T P) &= [\Phi^{-1}]_{n_1+1:n_2} \Xi \,\mathrm{vecs}(P)
\end{aligned} \tag{22}$$

Let T_v^{vs} and T_{vs}^v denote the transformation matrices between $\mathrm{vecs}(\cdot)$ and $\mathrm{vec}(\cdot)$, that is for any $P \in \mathbb{S}^n$, we have

$$\mathrm{vecs}(P) = T_v^{vs} \,\mathrm{vec}(P), \quad \mathrm{vec}(P) = T_{vs}^v \,\mathrm{vecs}(P).$$

Also, let T_{vt} denote the transformation matrix such that for any $X \in \mathbb{R}^{n \times n}$,

$$\mathrm{vec}(X^T) = T_{vt} \,\mathrm{vec}(X).$$

The vectorization of (29) results in

$$\begin{aligned}
&\mathrm{vecs}(A^T P_{K_i}^j + P_{K_i}^j A) - T_v^{vs}(I_n \otimes K_i^T) \,\mathrm{vec}(B^T P_{K_i}^j) \\
&- T_v^{vs}(K_i^T \otimes I_n) T_{vt} \,\mathrm{vec}(B^T P_{K_i}^j) \\
&+ T_v^{vs}(I_n \otimes L_{K_i}^{jT} D^T + L_{K_i}^{jT} D^T \otimes I_n) T_{vs}^v \,\mathrm{vecs}(P_{K_i}^j) \\
&+ \mathrm{vecs}(Q_{K_i} - \gamma^2 L_{K_i}^{jT} L_{K_i}^j) = 0.
\end{aligned} \tag{23}$$

Substituting P in (22) for $P_{K_i}^j$ in (23), we have

$$\begin{aligned}
&[\Phi^{-1}]_{1:n_1} \Xi \,\mathrm{vecs}(P_{K_i}^j) \\
&- T_v^{vs}[(I_n \otimes K_i^T) + (K_i^T \otimes I_n) T_{vt}][\Phi^{-1}]_{n_1+1:n_2} \Xi \,\mathrm{vecs}(P_{K_i}^j) \\
&+ T_v^{vs}(I_n \otimes L_{K_i}^{jT} D^T + L_{K_i}^{jT} D^T \otimes I_n) T_{vs}^v \,\mathrm{vecs}(P_{K_i}^j) \\
&+ \mathrm{vecs}(Q_{K_i} - \gamma^2 L_{K_i}^{jT} L_{K_i}^j) = 0.
\end{aligned} \tag{24}$$

Define

$$\begin{aligned}
&\Lambda_i^j := [\Phi^{-1}]_{1:n_1} \Xi \\
&- T_v^{vs}[(I_n \otimes K_i^T) + (K_i^T \otimes I_n) T_{vt}][\Phi^{-1}]_{n_1+1:n_2} \Xi \\
&+ T_v^{vs}(I_n \otimes L_{K_i}^{jT} D^T + L_{K_i}^{jT} D^T \otimes I_n) T_{vs}^v.
\end{aligned} \tag{25}$$

Then, (24) can be rewritten as

$$\mathrm{vecs}(P_{K_i}^j) = -(\Lambda_i^j)^{-1} \,\mathrm{vecs}(Q_{K_i} - \gamma^2 L_{K_i}^{jT} L_{K_i}^j). \tag{26}$$

Observe: equation (26) consists of unknown variables in the cost ($P_{K_i}^j$) and gain matrices (K_i, $L_{K_i}^j$), which can be iteratively computed using Algorithm 1. The other unknown variables from (18) are trajectory variables. Hence, we have essentially transformed the model-based algorithm of Section III into a model-free algorithm that requires only the state x and control signal u at time t in solving for a robustly stabilizing optimal control law. The learning-based algorithm for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ is shown in Algorithm 2. Compared against Algorithm 1 that requires the accurate model to update control policy, Algorithm 2 only requires the input-state data to construct the necessary matrices $\hat{\Phi}(t_f)$ and $\hat{\Xi}(t_f)$. The gain K_1 can be determined via sum of squares means [47] for example, whereupon a control Lyapunov function (CLF) candidate can be found to guarantee global or local asymptotic stability. This is only done once before the algorithm is run. Hence, it does not greatly hamper the time-efficiency of the proposed scheme.

Next, we analyze the convergence of both loops.

V. Convergence Analyses

As seen in Algorithm 1 and 2, the solution P_{K_i} to the Riccati equation must converge to the unique optimal positive-definite solution $P^\star$ so that the gains $L_{K_i}^j$ and K_{i+1} are optimal. In what follows, we provide a rigorous analysis of the convergence of the Riccati equation via a successive substitution scheme that is inspired by Kleinman's iterative Riccati computational scheme [48].

Algorithm 2: Learning-based $\mathcal{H}_2/\mathcal{H}_\infty$ Control

Initialize $\hat{K}_1 \in \mathcal{K}$ ▷ e.g. searching for a valid CLF [47];
Collect data from (1) with exploratory input (14); and
Construct matrices $\hat{\Phi}(t_f)$ and $\hat{\Psi}(t_f)$;
for $i \leq \bar{i}$ **do**
  for $j \leq \bar{j}$ **do**
    Construct the matrices $\hat{\Lambda}_i^j(t_f)$ using (25);
    Calculate $\hat{P}^j_{K_i}$ using (26);
    Update $\hat{L}^j_{K_i} = \gamma^{-2} D^T \hat{P}^j_{K_i}$;
  end
  Form $\mathrm{vec}(\widehat{B^T P^{\bar{j}}_{K_i}})$ as $[\hat{\Phi}^{-1}(t_f)]_{n_1+1:n_2} \hat{\Xi}(t_f)\, \mathrm{vecs}(\hat{P}^{\bar{j}}_{K_i})$;
  Calculate $\hat{K}_{i+1} = R^{-1} \widehat{B^T P^{\bar{j}}_{K_i}}$;
end

A. Control Update (Outer) Loop

The control law in the outer-loop of the algorithms seeks to decrease the cost (11) by iterating the following equations until convergence

$$A^T_{K_i} P_{K_i} + P_{K_i} A_{K_i} + Q_{K_i} + \gamma^{-2} P_{K_i} D D^T P_{K_i} = 0 \quad \text{(27a)}$$
$$K_{i+1} = R^{-1} B^T P_{K_i}, \quad i = 1, 2, \cdots . \quad \text{(27b)}$$

The control sequence (policy) K_i guarantees the system's safety via the stabilizing robust controller (as we show in Theorem B.1). Previous works have shown that the controller update phase i.e. the outer-loop iteration has a global sub-linear convergence rate and local quadratic convergence rate [14, Theorem A.7 and A.8]. We improve upon existing results in literature and demonstrate that the outer-loop iteration has a global linear convergence rate – which improves the sub-linear convergence rate.

Theorem 1. *For any $K_1 \in \mathcal{K}$, the outer-loop iteration has a global linear convergence rate, i.e. there exists $\alpha \in [0, 1)$, such that*

$$\mathrm{Tr}(P_{K_{i+1}} - P^\star) \leq \alpha\, \mathrm{Tr}(P_{K_i} - P^\star) \quad (28)$$

Proof. The proof to this theorem is provided in Appendix B-B. □

Remark 3. *With Theorem 1, we have $\|P_{K_i} - P^\star\|_F \leq \mathrm{Tr}(P_{K_i} - P^\star) \leq \alpha^i\, \mathrm{Tr}(P_{K_1} - P^\star)$.*

B. Disturbance Update (Inner) Loop

In this part, via a successive substitution scheme inspired by Kleiman's iterative Riccati computation scheme [48], we will analyze the monotonic convergence of the inner loop of Algorithm 1 and 2.

Let $P^j_{K_i}$ be the positive definite solution of the associated ARE at iteration j for the player with control $w(t)$, and iteration i for the player with control $u(t)$ so that

$$\left(A^j_{K_i}\right)^T P^j_{K_i} + P^j_{K_i} A^j_{K_i} + Q_{K_i} - \gamma^2 (L^j_{K_i})^T L^j_{K_i} = 0 \quad (29)$$

is recursively solved for $L^{j+1}_{K_i} = \gamma^{-2} D^T P^j_{K_i}$. Notice that we have replaced $K_i^T R K_i + C^T C$ with Q_{K_i}.

Theorem 2. *Given $K \in \mathcal{K}$, the inner-loop iteration has a global linear convergence rate, i.e. for any $j \in \mathbb{N}_+$, there exists $\beta(K) \in [0, 1)$, such that*

$$\mathrm{Tr}(P_K - P^{j+1}_K) \leq \beta(K)\, \mathrm{Tr}(P_K - P^j_K). \quad (30)$$

The proof is established in Appendix B-C.

Remark 4. *As seen from Theorem B.2, $P_K - P^j_K \succeq 0$. From Lemma A.1 and the result of Theorem 2, we have $\|P_K - P^j_K\|_F \leq \mathrm{Tr}(P_K - P^j_K) \leq \beta^{j-1}(K)\, \mathrm{Tr}(P_K)$, i.e. P^j_K exponentially converges to P_K in the sense of Frobenius norm.*

C. Iterative Uniform Convergence

Given our construction so far, for each K_i the inner-loop iteration generates sequences $\{P^j_{K_i}\}^{i=\bar{i}, j=\bar{j}}_{i=1, j=1}$ and $\{L^j_{K_i}\}^{i=\bar{i}, j=\bar{j}}_{i=1, j=1}$ which converge to the worst-case cost matrix and disturbance P_{K_i} and L_{K_i} respectively. We require that $\{P^j_{K_i}\}^{i=\bar{i}, j=\bar{j}}_{i=1, j=1}$ and $\{L^j_{K_i}\}^{i=\bar{i}, j=\bar{j}}_{i=1, j=1}$ enter the given neighborhood of P_{K_i} and L_{K_i} in a constant number of iterations. The following theorem guarantees uniform convergence after an equal number of inner-loop iterations i.e. the sequences generated by $\{P^j_{K_i}\}$ and $\{L^j_{K_i}\}$ enter the vicinity of P_{K_i} and L_{K_i}, irrespective of the different values of K_i.

Theorem 3. *For any $i \in \mathbb{N}_+$, and $\epsilon > 0$, there exists $\bar{j} \in \mathbb{N}_+$ independent of i, such that if $j \geq \bar{j}$,*

$$\|P^j_{K_i} - P_{K_i}\|_F \leq \epsilon. \quad (31)$$

The proof of this theorem is given in Appendix B-D. That is, the iterations converge uniformly to an $\epsilon > 0$.

VI. Robustness Analyses

In the last section, we assumed that the accurate model of the system is known and the iterative algorithm can be implemented exactly. In practice, due to model mismatch and various noise induced by measurements and external disturbance, the proposed two-loop iterative algorithm can hardly be executed precisely. Hence, the robustness of the algorithm to these aforementioned noise and disturbance is critical. The question of the iterative algorithm finding an approximate optimal solution to (3) and (11) under noise uncertainties needs to be answered. In this section, by considering the outer loop and inner loop as two separate discrete nonlinear systems, we will analyze the robustness of inner-loop and outer-loop iterations separately in the sense of input-to-state stability (ISS) [49], [50].

A. Control (Outer) Loop

The exact outer-loop iteration is as (27). At each iteration, K_{i+1} can be updated precisely without the influence from noise and disturbance. Let

$$\hat{A}_{K_i} := A - B\hat{K}_i,\ \hat{A}_i = A - B\hat{K}_i + \gamma^{-2} D D^T \hat{P}_{K_i}, \quad \text{(32a)}$$
$$\hat{Q}_{K_i} = C^T C + \hat{K}_i^T R \hat{K}_i. \quad \text{(32b)}$$

When noise exists and the policy is updated inaccurately, the inexact outer-loop iteration is

$$(\hat{A}_{K_i})^T\hat{P}_{K_i} + \hat{P}_{K_i}\hat{A}_{K_i} + \hat{Q}_{K_i} + \gamma^{-2}\hat{P}_{K_i}DD^T\hat{P}_{K_i} = 0, \quad (33a)$$
$$\hat{K}_{i+1} = R^{-1}B^T\hat{P}_{K_i} + \Delta K_i. \quad (33b)$$

Henceforth, we set $\tilde{K}_i = R^{-1}B^T\hat{P}_{K_i}$. Let us now give a statement of the theorem of the outer loop's robustness to perturbations.

Theorem 4. *There exists an $\underline{l} > 0$, $\hat{\alpha} \in [0,1)$, and $\kappa(\cdot) \in \mathcal{K}_\infty$, such that $\|\hat{P}_{K_i} - P^\star\|_F \leq \hat{\alpha}^{i-1}\operatorname{Tr}(\hat{P}_{K_1} - P^\star) + \kappa(\|\Delta K\|_\infty)$, as long as $\|\Delta K\|_\infty \leq \underline{l}$.*

Proof. The proof is provided in Appendix C-B. □

Remark 5. *That is, as iteration goes to infinity, $\hat{P}_{K_i}$ approaches the optimal value $P^\star$, entering its neighborhood. The radius of the neighbor is proportional to $\|\Delta K\|_\infty^2$. Therefore we conclude that the outer loop of the iteration is robust to noise and uncertainties.*

B. Disturbance (Inner) Loop

The exact inner-loop Iteration is (29), and the control policy $L^j_{K_i}$ can be updated precisely. In reality, due to the influence of disturbance and noise, $L^j_{K_i}$ may be updated inaccurately. Therefore, the inexact inner-loop iteration is

$$(\hat{A}^j_{K_i})^T\hat{P}^j_{K_i} + \hat{P}^j_{K_i}\hat{A}^j_{K_i} + Q_{K_i} - \gamma^2(\hat{L}^j_{K_i})^T\hat{L}^j_{K_i} = 0, \quad (34a)$$
$$\hat{L}^{j+1}_{K_i} = \gamma^{-2}D^T\hat{P}^j_{K_i} + \Delta L^j_{K_i}. \quad (34b)$$

where $\{\hat{L}^j_{K_i}\}_{j=1}^\infty$ and $\{\hat{P}^j_{K_i}\}_{j=1}^\infty$ are sequences generated by the inexact inner-loop iteration (34).

Theorem 5. *Assume $\|\Delta L^j_K\| < e$ for all $j \in \mathbb{N}_+$. There exists $\hat{\beta}(K) \in [0,1)$, and $\lambda(\cdot) \in \mathcal{K}_\infty$, such that*

$$\|\hat{P}^j_K - P_K\|_F \leq \hat{\beta}^{j-1}(K)\operatorname{Tr}(P_K) + \lambda(\|\Delta L\|_\infty). \quad (35)$$

The proof of this Theorem can be found in Appendix B-C. From Theorem 5, as $j \to \infty$, $\hat{P}^j_K$ approaches the solution P_K and enters the ball centered by P_K. The radius of the ball is proportional to $\|\Delta L\|_\infty$. Hence, the proposed inner-loop iterative algorithm finds the approximate value of P_K in the presence of unmodeled noise dynamics.

VII. Numerical Simulations

In this section, we will demonstrate our theoretical results on double and three-link inverted pendulums. The triple inverted pendulum is the base model for humanoid robots[51], [52] with the two upper hinge joints (hip and knee) being actuated while the lowest hinge (ankle) joint is passive. There are several challenges for a stabilizing controller design: the system is inherently unstable being non-minimum phase; it is underactuated system since the degrees of freedom is larger than the number of actuators; and the physical parameters of the humanoid robots are hard to accurately measure. In this section, the triple inverted pendulum is adopted as the numerical setups for our proposed algorithms. Specifically, we will design a learning-based balance PO scheme for this three-link robot with inaccurate system model.

The state of the triple inverted pendulum is $x = [\theta_1, \theta_2, \theta_3, \dot{\theta}_1, \dot{\theta}_2, \dot{\theta}_3]^T$, where θ_1, θ_2, and θ_3 are the angles of the ankle, hip, and knee. With actuator noise and possible installation error of the mechanism (e.g. the base is not securely attached to the ground), the linearized model of the triple inverted pendulum can be depicted by (1), where $A \in \mathbb{R}^{6\times 6}$ and $B \in \mathbb{R}^{6\times 2}$ are as given in [53, Section 3], and $D = \left[0_{3\times 3}, I_3\right]^T$.

Furthermore, we bound the system's $\mathcal{H}_\infty$ norm by $\gamma = 5$ from above. The initial state is set as

$$x(0) = [0, -5, 10, 10, -10, 10]^T. \quad (36)$$

The matrices related to the controlled output $z(t) = Cx(t) + Eu(t)$ are set as

$$C = \left[I_6, 0_{2\times 6}\right]^T, \quad E = \left[0_{6\times 2}, I_2\right]^T. \quad (37)$$

A. Comparison with LQG Control

Here, we assume the model of the systems are known, and Algorithm 1 is applied to solve for the optimal controller of $u^\star(t) = -K^\star x(t)$ and the worst-case disturbance $w^\star(t) = L^\star x(t)$. We choose the LQG cost function as,

$$J_{LQG} = \int_0^\infty x^T C^T C x + u^T E^T E u dt, \quad (38)$$

and we find the LQG feedback gain as

$$K_{LQG} = \begin{bmatrix} -26.77 & -8.755 & -4.20 & -9.033 & -3.05 & -2.30 \\ -65.10 & -23.79 & -8.24 & -21.60 & -9.27 & -3.93 \end{bmatrix}. \quad (39)$$

While executing Algorithm 1, the numbers of iterations for outer and inner loops are heuristically chosen as $\bar{i} = 20$ and $\bar{j} = 20$ and the initial controller, K_1, was chosen via linear matrix inequality approach. The result is shown in Fig. 3. We see that after around 5 iterations, the controller and the cost matrix converge to the optimal solution. Moreover, the $\mathcal{H}_\infty$ norms of closed-loop system with LQG and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controllers are respectively 8.72 and 4.99. Thus, this algorithm generates an optimal controller and guarantees the $\mathcal{H}_\infty$ norm system constraint, while the LQG controller does violate the constraint.

Carrying along with $x(0)$, we compare the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller with the LQG controller when w is a 1) Brownian motion; and 2) worst-case disturbance respectively.

When the disturbance exhibits Brownian motion in nature, mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller and LQG controller results are illustrated in Figures 1 and 4 respectively. Notice the chattering in joint angles and angular velocities around equilibrium after $5s$. In addition, the magnitude of the joint angle under a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller is within $[-5, +5]$ while that of the LQG controller is within $[-10, +10]$. Thus, for the Brownian motion, the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller does suppress the chattering.

Under a worst-case disturbance, the results LQG and for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ controller are shown in Figures 2 and 5 respectively. The LQG controller fails to satisfy the $\mathcal{H}_\infty$ norm constraint as seen in Fig. 2 i.e. the states becomes unstable after $1s$. As a comparison, consider Fig. 5, the state converges to equilibrium under the influence of the worst-case disturbance.

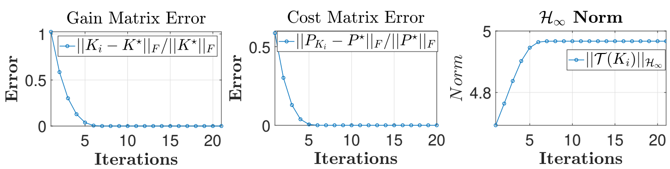

Fig. 3: Using Algorithm 1, the gain matrix and cost matrix converge to the optimal values, and the $\mathcal{H}_\infty$ norm satisfies the constraint.

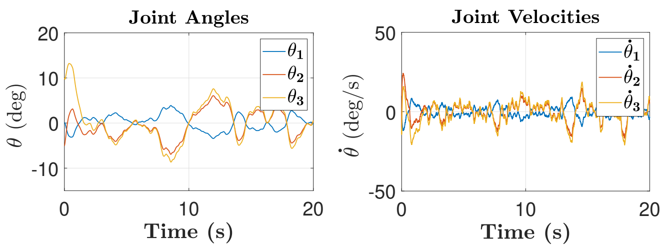

Fig. 4: With *LQG controller*, the evolution of the joint angles and velocities under Brownian motion disturbance.

B. Comparison with Natural Policy Gradient Algorithm

We further compare Algorithm 1 with the natural policy gradient (NPG) algorithm of Zhang et al [14] to test the veracity of the convergence rate and the robustness to process noise, ΔK. At each iteration of the algorithm, a ΔK_i sampled from a standard Gaussian distribution with Frobenius norm normalized to 0.15 is introduced into the algorithm following our derivations in Section VI. The results are shown in Figures 6 and 7. As seen in Fig. 6, the proposed iterative algorithm does approach the optimal solution after the 5th iteration despite the disturbance. At the last iteration, the deviation from the optimal cost matrix, $\frac{\|\hat{P}_{K_{20}} - P^\star\|}{\|P^\star\|_F}$, is 2.9%, while the gain error, $\frac{\|\hat{K}_{20} - K^\star\|_F}{\|K^\star\|_F}$, is 2.6%. In contrary, the natural policy gradient has a cost matrix and controller gain errors that are unbounded as the iteration proceeds. Therefore, our algorithm is more robust than the natural policy gradient algorithm to process noise.

The computational time of Algorithm 1 is compared with that of NPG, and the result is shown in Table I. It is seen that for the double and triple inverted pendulums, the computational time of our algorithm is much less than that of NPG by around 90%. This is in fact a validation of our superior convergence rate (i.e. a global linear and local quadratic rate) compared to NPG's sublinear convergence rate.

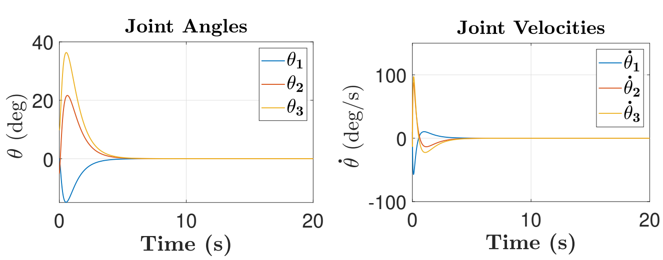

Fig. 5: With our mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy, the evolution of the joint angles and velocities under a worst-case disturbance.

TABLE I: Comparison of Alg. 1 and NPG.

Computational time (sec)					
Double Inverted Pendulum			Triple Inverted Pendulum		
Alg. 1	Alg. 2	NPG	Alg. 1	Alg. 2	NPG
0.0901	0.3061	2.1649	0.1455	0.7829	2.3209

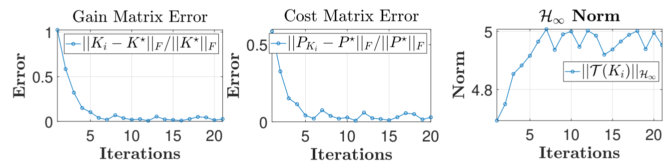

Fig. 6: Robustness of Algorithm 1 under a process noise whose norm is $\|\Delta K\|_\infty = 0.15$.

C. Results on Learning-based Control

For the parameters of Algorithm 2, we set $\bar{i} = 20$ and $\bar{j} = 30$, and collected data for $t_f = 1500s$. The parameters of the A and B matrices are unknown but the initial controller $\hat{K}_1 \in \mathcal{K}$ is known. We run Algorithm 2 to find a near optimal solution of (3) using the input and state data collected from system (1). As seen in Fig. 8, the obtained controller $\hat{K}_i$ at each iteration converges after 5 iterations. The corresponding evaluative matrix $\hat{P}_{K_i}$ also converges. At 20th iteration, the relative error of $\|\hat{K}_{20} - K_*\|/\|K_*\| = 31.5\%$ and $\|\hat{P}_{K_{20}} - P_*\|/\|P_*\| = 31.6\%$. These demonstrate that the proposed algorithm can find an approximate optimal solution using the noisy data.

VIII. Conclusions

In the research effort presented in this article, we have proposed a two-loop iterative algorithm for the robust and optimal control of linear time-invariant systems. Rigorous convergence has been given that demonstrate that the proposed two-loop iterative algorithm has a global linear convergence alongside uniform convergence. Furthermore, by considering the iterative algorithm as a nonlinear system, we have presented novel robustness analyses and evaluation of the iterative algorithm in the sense of small-disturbance input-to-state stability. Based on these premises, a learning-based iterative algorithm has been developed to generate an approximate robust and optimal controller using noisy data collected from the system. The proposed algorithms are evaluated on two- and three-link inverted pendulum testbeds and our results confirm our various hypotheses.

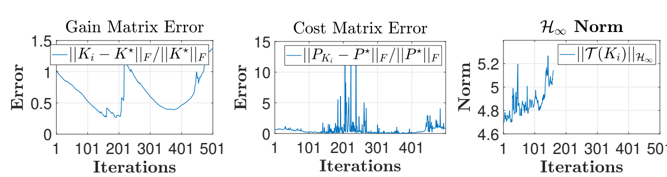

Fig. 7: Robustness of natural policy gradient under a noise norm $\|\Delta K\|_\infty = 0.1$.

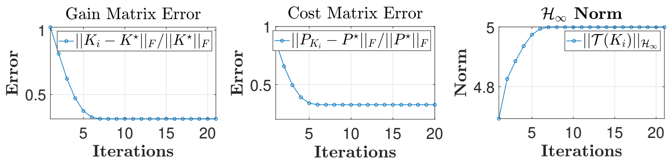

Fig. 8: Algorithm 2 generates an approximation to the robustly optimal controller based on noisy system data.

IX. Acknowledgments

A vote of thanks to Professor Zhong-Ping Jiang of NYU Tandon School of Engineering for bringing this line of research inquisition to our notice and for pointing out helpful reading materials.

Appendix A
A Calatog of Preliminary Lemmas

In this appendix, we provide Lemmas necessary for the construction of our main results. For further reading, readers can consult H_∞ control theory texts such as [12], [54], [55].

Lemma A.1. *For any symmetric and positive semi-definite matrix $P \in \mathbb{S}^n$, $\|P\|_F \le \mathrm{Tr}(P)$, $\|P\| \le \mathrm{Tr}(P)$, and $\mathrm{Tr}(P) \le n\|P\|$. For $x \in \mathbb{R}^n$, $x^T P x \ge \underline{\sigma}(P)\|x\|^2$.*

Proof. Let $\sigma_1 \ge \cdots \ge \sigma_n$ be ordered singular values of P. Since P is symmetric and positive semi-definite, its singular values are equal to its eigenvalues. Then, $\|P\|_F = \sqrt{\sum_{i=1}^n \sigma_i^2}$, $\mathrm{Tr}(P) = \sum_{i=1}^n \sigma_i$, and $\|P\| = \sigma_1(P)$. Since $\sum_{i=1}^n \sigma_i^2 \le (\sum_{i=1}^n \sigma_i)^2$, we have $\|P\|_F \le \mathrm{Tr}(P)$, $\|P\| \le \mathrm{Tr}(P)$, and $\mathrm{Tr}(P) \le n\|P\|$. Using Rayleigh's theorem [56, Theorem 4.2.2], we have $x^T P x \ge \underline{\sigma}(P)\|x\|^2$. □

Lemma A.2. *For $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times p}$, $\|XY\|_F \le \|X\|\|Y\|_F$.*

Proof. Let $Y = [y_1, \cdots, y_p]$, then it follows that $XY = [Xy_1, \cdots, Xy_p]$. This implies that $\|XY\|_F^2 = \sum_{i=1}^p \|Xy_i\|^2$. Furthermore, as the spectral norm is defined by $\|X\| = \max_{x\neq 0} \frac{\|Xx\|}{\|x\|}$, we have $\|Xy_i\|^2 \le \|X\|^2\|y_i\|^2$. Hence, $\|XY\|_F^2 \le \|X\|^2 \sum_{i=1}^p \|y_i\|^2 = \|X\|\|Y\|_F$, which is in fact a proof of the theorem. □

Lemma A.3. *Assume $A \in \mathbb{R}^{n\times n}$ is Hurwitz and satisfies $A^T P + PA + Q = 0$. Then, the following properties hold*

1) *$P = \int_{t=0}^\infty e^{(A^T t)} Q e^{(At)} dt$;*
2) *$P \succ 0$ if $Q \succ 0$, and $P \succeq 0$ if $Q \succeq 0$;*
3) *If $Q \succeq 0$, then (Q, A) is observable iff $P \succ 0$;*
4) *If P' satisfies $A^T P' + P' A + Q' = 0$, and $Q' \preceq Q$, then $P' \preceq P$.*

Proof. The first three statements are proven in [55, Lemma 3.18]. Consequently, P' can be expressed as

$$P' = \int_0^\infty e^{A^T t} Q' e^{At} dt. \tag{A.1}$$

Since $Q' \preceq Q$, $P' \preceq P$. □

Lemma A.4. *Suppose P satisfies $A^T P + PA + Q = 0$, then*

1) *A is Hurwitz if $P \succ 0$ and $Q \succ 0$.*
2) *A is Hurwitz if $P \succeq 0$, $Q \succeq 0$ and (Q, A) is detectable.*

Proof. This Lemma is proven in [55, Lemma 3.19]. □

Lemma A.5. *Let $(X, Y) \in \mathbb{R}^{n\times n}$ and let $Y = Y^T \ge 0$. Then,*

$$-\mu_2(-X)\,\mathrm{Tr}(Y) \le \mathrm{Tr}(XY) \le \mu_2(X)\,\mathrm{Tr}(Y)$$

where $\mu_2(X)$ is the matrix measure, as a function of the spectral norm of the matrix X, i.e. $\frac{1}{2}\lambda_{max}(X + X^T)$.

Proof. This Lemma is proven in [57]. □

Lemma A.6. *Let $U : S \to \mathbb{R}^{m\times r}$ and $V : S \to \mathbb{R}^{r\times p}$ be two matrix functions defined and differentiable on an open set S in $\mathbb{R}^{n\times q}$. Then the simple product UV is differentiable on S and the Jacobian matrix is the $mp \times nq$ matrix*

$$\frac{\partial\,\mathrm{vec}(UV)}{\partial\,\mathrm{vec}(X)} = (V^T \otimes I_m)\frac{\partial\,\mathrm{vec}(U)}{\partial\,\mathrm{vec}(X)} + (I_p \otimes U)\frac{\partial\,\mathrm{vec}(V)}{\partial\,\mathrm{vec}(X)}. \tag{A.2}$$

Proof. This Lemma is proven in [58, Theorem 9]. □

Lemma A.7 (Birkhoff's Ergodic Theorem)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\{T_t \mid t \ge 0\}$ be a measurable semi-group of transformations preserving $(\Omega, \mathcal{F}, P)$. Suppose that the time average operator A_t is so defined*

$$A_t f(w) = \frac{1}{t}\int_0^t f(T_s(w))ds. \tag{A.3}$$

Then, for every $f \in L^1(P)$, there exists $\bar{f} \in L^1(P)$ such that $A_t f \to \bar{f}$ both P-almost surely and in $L^1(P)$ as $t \to \infty$, where $\bar{f} = \mathbb{E}[f]$.

Proof. This Lemma is proven in [59, Theorem16.14] and [60, Theorem 1.5.9]. □

Lemma A.8. *Let $a(x(t))$ be a vector function such that $\mathbb{E}[a^T(x(t))a(x(t))] < \infty$ where $\{x(t) | t \ge 0\}$ is the solution of process $dx(t) = \mu(x(t))dt + dw$, and w is a standard independent Brownian motion. It follows that*

$$\lim_{t_f\to\infty} \frac{1}{t_f}\int_0^{t_f} a(x(t))dw = 0. \tag{A.4}$$

Proof. This lemma is proven in [61, pp. 530]. □

Lemma A.9 (Bounded Real Lemma)**.** *For the stabilizing gain matrix K, the following conditions are equivalent:*

- *$||\mathcal{T}(K)||_{\mathcal{H}_\infty} \le \gamma$;*
- *The Riccati equation*

$$(A - BK)^T P_K + P_K(A - BK) + C^T C + K^T R K + \gamma^{-2} P_K D D^T P_K = 0 \tag{A.5}$$

  admits a unique stabilizing solution with a Hurwitz $A - BK + \gamma^{-2} D D^T P_K$;
- *There exists some $P_K \succ 0$ such that*

$$(A - BK)^T P_K + P_K(A - BK) + C^T C + K^T R K + \gamma^{-2} P_K D D^T P_K \preceq 0. \tag{A.6}$$

Proof. This Lemma is a restatement of [54], [12]. □

Appendix B
Proof of Convergence of Iterations

In this appendix, we set up the proofs for the convergence of the inner and outer loops of the two algorithms.

A. Outer Loop Convergence

Let us first introduce the following results.

Theorem B.1. *Consider Assumptions 1 and 2. If $K_1 \in \mathcal{K}$, then for any $i \geq 1$*

1) $K_i \in \mathcal{K}$;
2) $P_{K_1} \succeq \cdots P_{K_i} \cdots \succeq P^\star$, *for* $K_i \succeq K_1$;
3) $\lim_{i\to\infty} ||K_i - K^*||_F = 0$, $\lim_{i\to\infty} ||P_{K_i} - P^*||_F = 0$.

We here establish the proof to convergence of K_i per iteration i to the optimal gain $K^\star$.

Proof. Statements 1) and 3) and the corresponding proof are given in [30, Theorems A.6 and A.7]. To make the paper self-contained, the method in [48] is adopted to prove the statements. The first statement will be proven by induction. When $i = 1$, $K_1 \in \mathcal{K}$, and it satisfies 1). When $i > 1$, assume $K_i \in \mathcal{K}$. Therefore, by the second condition of Lemma A.9, $P_{K_i} \succ 0$ is the unique stabilizing solution to (27a). Now consider the identities,

$$
\begin{aligned}
(A - BK_i)^T &= (A - BK_{i+1})^T P_{K_i} + (K_{i+1} - K_i)^T B^T P_{K_i} \\
P_{K_i}(A - BK_i) &= P_{K_i}(A - BK_{i+1}) + P_{K_i} B (K_{i+1} - K_i).
\end{aligned}
\tag{B.1}
$$

It follows that equation (27a) can be rewritten as

$$
\begin{aligned}
&= A_{K_{i+1}}^T P_{K_i} + P_{K_i} A_{K_{i+1}} + C^T C + \gamma^{-2} P_{K_i} DD^T P_{K_i} \\
&+ K_{i+1}^T R K_{i+1} + (K_{i+1} - K_i)^T R (K_{i+1} - K_i) = 0.
\end{aligned}
\tag{B.2}
$$

As $(K_{i+1} - K_i)^T R (K_{i+1} - K_i) \succeq 0$, following the third condition of Lemma A.9, we find that $K_{i+1} \in \mathcal{K}$. *A fortiori*, we establish that $K_i \in \mathcal{K}$. We proceed as follows for statement 2). Writing out (27a) for the $(i+1)$'th iteration, and subtracting the resulting equation from (27a), we find that

$$
\begin{aligned}
&A_{i+1}^T (P_{K_i} - P_{K_{i+1}}) + (P_{K_i} - P_{K_{i+1}}) A_{i+1} \\
&+ (K_i - K_{i+1})^T R (K_i - K_{i+1}) \\
&+ \gamma^{-2} (P_{K_i} - P_{K_{i+1}}) DD^T (P_{K_i} - P_{K_{i+1}}) = 0.
\end{aligned}
\tag{B.3}
$$

Given statement 1), it follows that $K_{i+1} \in \mathcal{K}$. Furthermore, by the second condition of Lemma A.9, A_{i+1} is Hurwitz. As $(K_i - K_{i+1})^T R (K_i - K_{i+1}) \succeq 0$ and $\gamma^{-2}(P_{K_i} - P_{K_{i+1}}) DD^T (P_{K_i} - P_{K_{i+1}}) \succ 0$, by Lemma A.3, we find that $P_{K_i} - P_{K_{i+1}} \succeq 0$. Because $K_{i+1} \in \mathcal{K}$, A_{i+1} is Hurwitz by Lemma A.9. Hence, $P_{K_i} \succeq P_{K_{i+1}}$ i.e., the sequence $\{P_{K_i}\}_{i=1}^\infty$ is decreasing, and lower bounded by 0. *A fortiori*, $\{P_{K_i}\}_{i=1}^\infty$ convergences to P_{K_∞}, which also satisfies (7). Due to the uniqueness of the solution to (7), $P_{K_\infty} = P^\star$. □

B. Global Linear Convergence of the Outer (Control) Loop

Before proving Theorem 1, we first establish the following cost matrix' quadratic convergence result.

Proposition 3. *For any $i \in \mathbb{N}_+$, there exists an $a > 0$, such that* $\mathrm{Tr}(P_{K_{i+1}} - P^\star) \leq a \left[\mathrm{Tr}(P_{K_i} - P^\star)\right]^2$.

Proof. For the $(i+1)'th$ iteration, (27a) can be rewritten as

$$
\begin{aligned}
&A_{i+1}^T P_{K_{i+1}} + P_{K_{i+1}} A_{i+1} + C^T C + K_{i+1}^T R K_{i+1} \\
&- \gamma^{-2} P_{K_{i+1}} DD^T P_{K_{i+1}} = 0.
\end{aligned}
\tag{B.4}
$$

Also, (7) can be rewritten as

$$
\begin{aligned}
&A_{i+1}^T P^\star + P^\star A_{i+1} + C^T C - K^{\star T} R K^\star \\
&+ K_{i+1}^T R K^\star + K^{\star T} R K_{i+1} - \gamma^{-2} P_{K_{i+1}} DD^T P^\star \\
&- \gamma^{-2} P^\star DD^T P_{K_{i+1}} + \gamma^{-2} P^\star DD^T P^\star = 0.
\end{aligned}
\tag{B.5}
$$

Subtracting (B.5) from (B.4), and completing squares, we have

$$
\begin{aligned}
&A_{i+1}^T (P_{K_{i+1}} - P^\star) + (P_{K_{i+1}} - P^\star) A_{i+1} \\
&+ (K_{i+1} - K^\star)^T R (K_{i+1} - K^\star) \\
&- \gamma^{-2} (P_{K_{i+1}} - P^\star) DD^T (P_{K_{i+1}} - P^\star) = 0.
\end{aligned}
\tag{B.6}
$$

From Theorem B.1 and Lemma A.9, we see that A_{i+1} is Hurwitz. From Lemma A.3, it follows that

$$
\begin{aligned}
&P_{K_{i+1}} - P^\star \\
&\preceq \int_0^\infty e^{A_{i+1}^T t} (P_{K_i} - P^\star) BR^{-1} B^T (P_{K_i} - P^\star) e^{A_{i+1} t} dt.
\end{aligned}
\tag{B.7}
$$

Using the cyclic property of matrix trace,

$$
\begin{aligned}
&\mathrm{Tr}(P_{K_{i+1}} - P^\star) \leq \\
&\mathrm{Tr}\left[(P_{K_i} - P^\star) BR^{-1} B^T (P_{K_i} - P^\star) \int_0^\infty e^{A_{i+1} t} e^{A_{i+1}^T t} dt\right].
\end{aligned}
\tag{B.8}
$$

Let us define $M_{i+1} := \int_0^\infty e^{A_{i+1} t} e^{A_{i+1}^T t} dt$. Because A_{i+1} is Hurwitz, M_{i+1} satisfies

$$
A_{i+1} M_{i+1} + M_{i+1} A_{i+1}^T + I_n = 0. \tag{B.9}
$$

From Theorem B.1, it follows that as $i \to \infty$, P_{K_i} converges $P^\star$ and A_i converges to $A^\star$. Consequently, M_i converges to $M^\star \in \mathbb{S}^n$, which is the solution to

$$
A^\star M^\star + M^\star (A^\star)^T + I_n = 0. \tag{B.10}
$$

Thus, $\lim_{i\to\infty} \|M_i\| = \|M^\star\|$ and $\bar{m} := \sup_{i\in\mathbb{N}_+} \|M_i\| < \infty$. As a consequence,

$$
\begin{aligned}
&\mathrm{Tr}(P_{K_{i+1}} - P^\star) \leq \mathrm{Tr}\left[(P_{K_i} - P^\star) BR^{-1} B^T (P_{K_i} - P^\star) M_i\right] \\
&\mathrm{Tr}(P_{K_{i+1}} - P^\star) \leq \bar{m} \|BR^{-1} B^T\| \left[\mathrm{Tr}(P_{K_i} - P^\star)\right]^2.
\end{aligned}
\tag{B.11}
$$

Setting $a := \bar{m} \|BR^{-1} B^T\|$, we see that the outer-loop's cost matrix convergences in a quadratic manner in the vicinity of $P^\star$. □

Proposition 4. *For any $i \in \mathbb{N}_+$, there exists a scalar $b > 0$, such that $\mathrm{Tr}(P_{K_i} - P^\star) \leq b\, \mathrm{Tr}(P_{K_{i-1}} - P_{K_i})$. In addition, there exists $b' > 0$, such that $\|K_{i+1} - K^\star\| \leq b' \|K_i - K_{i+1}\|$.*

Proof. From Proposition 3, for any $\epsilon > 0$, there exists $\hat{i} \in \mathbb{N}_+$, such that if $i \geq \hat{i}$,

$$\operatorname{Tr}(P_{K_{i-1}} - P^\star) \leq \frac{1}{a(1+\epsilon)}, \tag{B.12}$$

where $a > 0$ is as given in Proposition 3. We have by Proposition 3 for any $i \geq \hat{i}$ that

$$\frac{\operatorname{Tr}(P_{K_{i-1}} - P_{K_i})}{\operatorname{Tr}(P_{K_i} - P^\star)} = \frac{\operatorname{Tr}(P_{K_{i-1}} - P^\star)}{\operatorname{Tr}(P_{K_i} - P^\star)} - 1 \tag{B.13a}$$

$$\geq \frac{1}{a\operatorname{Tr}(P_{K_{i-1}} - P^\star)} - 1 \geq \epsilon. \tag{B.13b}$$

Now, for $i < \hat{i}$, from Theorem B.1 we have

$$0 < \operatorname{Tr}(P_{K_i} - P^\star) \leq \operatorname{Tr}(P_{K_1} - P^\star). \tag{B.14}$$

Suppose that we let $b_1 = \min_{i<\hat{i}} \operatorname{Tr}(P_{K_{i-1}} - P_{K_i})$. Then, by Theorem B.1, $b_1 > 0$ so that

$$\frac{\operatorname{Tr}(P_{K_{i-1}} - P_{K_i})}{\operatorname{Tr}(P_{K_i} - P^\star)} \geq \frac{b_1}{\operatorname{Tr}(P_{K_1} - P^\star)}. \tag{B.15}$$

Let

$$b := \max\left[\frac{1}{\epsilon}, \frac{\operatorname{Tr}(P_{K_1} - P^\star)}{b_1}\right]. \tag{B.16}$$

We see that $\operatorname{Tr}(P_{K_i} - P^\star) \leq b\operatorname{Tr}(P_{K_{i-1}} - P_{K_i})$ for all $i \in \mathbb{N}_+$. From Lemma A.1, we have

$$\begin{aligned} \|K_{i+1} - K^\star\| &\leq \|R^{-1}B^T\|\|P_{K_i} - P^\star\| \\ &\leq \|R^{-1}B^T\|\operatorname{Tr}(P_{K_i} - P^\star) \\ &\leq b\|R^{-1}B^T\|\operatorname{Tr}(P_{K_{i-1}} - P_{K_i}) \\ &\leq bn\|R^{-1}B^T\|\|P_{K_{i-1}} - P_{K_i}\|. \end{aligned} \tag{B.17}$$

Furthermore,

$$\begin{aligned} \|K_i - K_{i+1}\| &= \|R^{-1}B^T(P_{K_i} - P_{K_{i-1}})\| \\ &\geq \underline{\sigma}(R^{-1}B^T)\|P_{K_i} - P_{K_{i-1}}\|. \end{aligned} \tag{B.18}$$

As B is full column rank, $\underline{\sigma}(R^{-1}B^T) > 0$. Setting $b' = bn\|R^{-1}B^T\|/\underline{\sigma}(R^{-1}B^T)$, we establish the second statement. □

Proposition 5. *For any $i \in \mathbb{N}_+$, we have*

$$(K_i - K^\star)^T R(K_i - K^\star) \succeq \gamma^{-2}(P_{K_i} - P^\star)DD^T(P_{K_i} - P^\star). \tag{B.19}$$

Proof. Similar to (B.6), we have

$$A_i^T(P_{K_i} - P^\star) + (P_{K_i} - P^\star)A_i + (K_i - K^\star)^T R(K_i - K^\star) - \gamma^{-2}(P_{K_i} - P^\star)DD^T(P_{K_i} - P^\star) = 0. \tag{B.20}$$

Using Theorem B.1, A_i is Hurwitz, and $P_{K_i} - P^\star \succeq 0$. Therefore, by Lemma A.3, we arrive at the required inequality. □

Lemma B.1. *Let $E_{K_i} = (K_i - K_{i+1})^T R(K_i - K_{i+1})$. For the sequences $\{P_{K_i}\}_{i=1}^\infty$ and $\{K_i\}_{i=1}^\infty$ obtained by the control loop update, the following inequality holds*

$$\operatorname{Tr}(P_{K_i} - P_*) \leq c\|E_{K_i}\|, \text{ ***where,*** }$$

$$c = [\underline{\sigma}(R)]^{-1}(2+2b')\|R\|\operatorname{Tr}\left(\int_0^\infty e^{A^\star t}e^{A^{\star T}t}dt\right). \tag{B.21}$$

Proof. For the ith iteration, (27a) can be rewritten as

$$\begin{aligned} &A^{\star T}P_{K_i} + P_{K_i}A^\star + (K^\star - K_i)^T RK_{i+1} \\ &+ K_{i+1}^T R(K^\star - K_i) + \gamma^{-2}(P_{K_i} - P^\star)DD^T P_{K_i} \\ &+ \gamma^{-2}P_{K_i}DD^T(P_{K_i} - P^\star) + C^T C + K_i^T RK_i \\ &- \gamma^{-2}P_{K_i}DD^T P_{K_i} = 0, \end{aligned} \tag{B.22}$$

and (7) can be rewritten as

$$\begin{aligned} &A^{\star T}P^\star + P^\star A^\star + C^T C + K^{\star T}RK^\star \\ &- \gamma^{-2}P^\star DD^T P^\star = 0. \end{aligned} \tag{B.23}$$

Subtracting (B.23) from (B.22) and completing squares, we have

$$\begin{aligned} &A^{\star T}(P_{K_i} - P^\star) + (P_{K_i} - P^\star)A^\star + E_{K_i} \\ &+ \gamma^{-2}(P_{K_i} - P^\star)DD^T(P_{K_i} - P^\star) \\ &- (K_{i+1} - K^\star)^T R(K_{i+1} - K^\star) = 0. \end{aligned} \tag{B.24}$$

Using Proposition 5, and completing the squares, (B.24) becomes

$$\begin{aligned} &A^{\star T}(P_{K_i} - P^\star) + (P_{K_i} - P^\star)A^\star + 2E_{K_i} \\ &+ (K_i - K_{i+1})^T R(K_{i+1} - K^\star) \\ &+ (K_{i+1} - K^\star)^T R(K_i - K_{i+1}) \succeq 0. \end{aligned} \tag{B.25}$$

Now, using Lemma A.3, we have

$$\begin{aligned} &\operatorname{Tr}(P_{K_i} - P^\star) \leq \operatorname{Tr}\left\{\int_0^\infty e^{A^{\star T}t}\left[2E_{K_i}\right.\right. \\ &+ (K_i - K_{i+1})^T R(K_{i+1} - K^\star) \\ &\left.\left.+ (K_{i+1} - K^\star)^T R(K_i - K_{i+1})\right]e^{A^\star t}dt\right\} \end{aligned} \tag{B.26a}$$

$$\begin{aligned} &\leq \operatorname{Tr}\left\{\left[2E_{K_i} + (K_i - K_{i+1})^T R(K_{i+1} - K^\star)\right.\right. \\ &\left.\left.+ (K_{i+1} - K^\star)^T R(K_i - K_{i+1})\right]\int_0^\infty e^{A^\star t}e^{A^{\star T}t}dt\right\} \end{aligned} \tag{B.26b}$$

$$\begin{aligned} &\leq [2\|E_{K_i}\| + 2\|K_i - K_{i+1}\|\|R\|\|K_{i+1} - K_*\|] \\ &\quad \operatorname{Tr}\left(\int_0^\infty e^{A^\star t}e^{(A^{\star T})t}dt\right) \end{aligned} \tag{B.26c}$$

$$\begin{aligned} &\leq [2\|E_{K_i}\| + 2b'\|K_i - K_{i+1}\|\|R\|\|K_i - K_{i+1}\|] \\ &\quad \operatorname{Tr}\left(\int_0^\infty e^{A^\star t}e^{A^{\star T}t}dt\right), \end{aligned} \tag{B.26d}$$

where the last expression is as a result of Proposition 4. And by Lemma A.1, we arrive at the required inequality (B.21), i.e.

$$\begin{aligned} &\operatorname{Tr}(P_{K_i} - P^\star) \leq \\ &\underbrace{(2+2b')\frac{\|R\|}{\underline{\sigma}(R)}}_{:=c}\operatorname{Tr}\left(\int_0^\infty e^{A^\star t}e^{A^{\star T}t}dt\right)\|E_{K_i}\|. \end{aligned} \tag{B.27}$$

□

Lemma B.2. *Given that $E \in \mathbb{S}^n$ is positive semi-definite, and $W \in \mathbb{R}^{n\times n}$ is Hurwitz. Let $F := \int_0^\infty e^{W^T t}Ee^{Wt}dt$, and $d(W) = \log(5/4)/\|W\|$. Then, $\|F\| \geq \frac{1}{2}d(W)\|E\|$.*

Proof. A Taylor expansion of e^{Wt} yields

$$e^{Wt} = I_n + \underbrace{\left[\sum_{k=1}^\infty (Wt)^k/k!\right]}_{:=S(t)}, \tag{B.28}$$

so that $\|S(t)\| \leq e^{\|W\|t} - 1$. For an $x_0 \neq 0$ satisfying $x_0^T E x_0 = \|E\|\|x_0\|^2$, we have,

$$\begin{aligned} x_0^T F x_0 &\geq \int_0^{d(W)} x_0^T e^{W^T t} E e^{Wt} x_0 dt \\ &= \int_0^{d(W)} x_0^T (I_n + S(t)) E (I_n + S(t)) x_0 dt \\ &\geq \int_0^{d(W)} \|E\|\|x_0\|^2 - 2\|S(t)\|\|E\|\|x_0\|^2 dt \\ &\geq \frac{1}{2} d(W)\|E\|\|x_0\|^2. \end{aligned} \tag{B.29}$$

From (B.29), we see that $\|F\| \geq \frac{1}{2}d(W)\|E\|$. This proves the Lemma. □

Proof of Theorem 1. The proof of Theorem 1 is now straightforward. Let us write

$$(P_{K_i} - P_{K_{i+1}}) \succeq \int_0^\infty e^{A_{i+1}^T t} E_{K_i} e^{A_{i+1} t} dt =: F_{K_i},$$

following (B.3). Recall that $P_{K_i} \succeq P_{K_{i+1}}$ and A_{i+1} is Hurwitz. Thus, because $P_{K_1} \succeq P_{K_i}$, we have

$$\|A_{i+1}\| \leq \|A\| + (\|BR^{-1}B^T\| + \gamma^{-2}\|DD^T\|)\|P_{K_1}\|. \tag{B.30}$$

Let us set

$$d = \frac{\log(5/4)}{\|A\| + (\|BR^{-1}B^T\| + \gamma^{-2}\|DD^T\|)\|P_{K_1}\|}. \tag{B.31}$$

It follows that (Lemma B.2) $\|F_{K_i}\| \geq \frac{1}{2}d\|E_{K_i}\|$, so that by Lemma B.1, we can write

$$\begin{aligned} \mathrm{Tr}(P_{K_{i+1}} - P^*) &\leq \mathrm{Tr}(P_{K_i} - P^*) - \frac{1}{2}d\|E_{K_i}\| && \text{(B.32a)} \\ &\leq \mathrm{Tr}(P_{K_i} - P^*) - \frac{d}{2c}\mathrm{Tr}(P_{K_i} - P^*) && \text{(B.32b)} \\ &\leq \underbrace{\left(1 - \frac{d}{2c}\right)}_{:=\alpha} \mathrm{Tr}(P_{K_i} - P^*) && \text{(B.32c)} \end{aligned}$$

Equation (B.32c) is in fact the required inequality for (28), that is, $\mathrm{Tr}(P_{K_{i+1}} - P^*) \leq \alpha\, \mathrm{Tr}(P_{K_i} - P^*)$. □

Next, we establish the proof of the convergence of the inner loop of the iteration.

C. Proof of Inner Loop Convergence (Theorem 2)

To establish this theorem, we first introduce a few results.

Theorem B.2. *Assume $L_{K_i}^1 = 0$, then for any $i, j \in \mathbb{N}_+$, the following holds*

1) *$A_{K_i}^j$ is Hurwitz,*
2) *$P_{K_i} \succeq \cdots \succeq P_{K_i}^{j+1} \succeq P_{K_i}^j \succeq \cdots \succeq P_{K_i}^1$,*
3) *$\lim_{j\to\infty} \|P_{K_i}^j - P_{K_i}\|_F = 0$, where P_{K_i} is the solution to (27a).*

Proof. The first statement will be proven by induction. By Theorem B.1, $K_i \in \mathcal{K}$. According to Lemma A.9, there exists a unique stabilizing solution $P_{K_i} \succ 0$ to (27a) and A_i is Hurwitz. When $j = 1$, $L_{K_i}^1 = 0$, then, (29) can be rewritten as

$$A_{K_i}^T P_{K_i}^1 + P_{K_i}^1 A_{K_i} + C^T C + K_i^T R K_i = 0. \tag{B.33}$$

Since A_{K_i} is Hurwitz and (C, A) is observable, by Lemma A.3, $A_i^1 = A_{K_i}$ is Hurwitz, and $P_{K_i}^1 \succ 0$. Subtracting (29) from (27a) yields

$$\begin{aligned} &(A_{K_i}^j)^T (P_{K_i} - P_{K_i}^j) + (P_{K_i} - P_{K_i}^j) A_{K_i}^j \\ &\quad + \gamma^2 (L_{K_i}^j - L_{K_i})^T (L_{K_i}^j - L_{K_i}) = 0. \end{aligned} \tag{B.34}$$

When $j = 1$, since $A_{K_i}^1$ is Hurwitz and $\gamma^2(L_{K_i}^j - L_{K_i})^T(L_{K_i}^j - L_{K_i}) \succeq 0$, according to Lemma A.3, (B.34) results in $P_{K_i} \succeq P_{K_i}^1 \succ 0$. We can rewrite (27a) as

$$A_i^T P_{K_i} + P_{K_i} A_i + Q_{K_i} - \gamma^{-2} P_{K_i} D D^T P_{K_i} = 0. \tag{B.35}$$

Since A_i is Hurwitz and $P_{K_i} \succ 0$, by Lemma A.3, the following inequality holds

$$Q_{K_i} - \gamma^{-2} P_{K_i} D D^T P_{K_i} \succ 0. \tag{B.36}$$

Assume A_i^j is Hurwitz, and from (B.34), we have $P_{K_i} \succeq P_{K_i}^j$. Following (27a), we have

$$\begin{aligned} &(A_i^{j+1})^T P_{K_i} + P_{K_i} A_i^{j+1} + Q_{K_i} - \gamma^2 P_{K_i}^j D D^T P_{K_i}^j \\ &\quad + \gamma^2 (L_{K_i}^{j+1} - L_{K_i})^T (L_{K_i}^{j+1} - L_{K_i}) = 0. \end{aligned} \tag{B.37}$$

By (B.36) and $P_{K_i} \succeq P_{K_i}^j$,

$$Q_{K_i} - \gamma^2 P_{K_i}^j D D^T P_{K_i}^j \succ 0. \tag{B.38}$$

Hence, by Lemma A.4 and (B.37), A_i^{j+1} is Hurwitz. As a consequence, the proof of statement 1) is completed.

Rewriting (29) for the $(j+1)'th$ iteration and subtracting the resulting equation from (29) results in

$$\begin{aligned} &(A_{K_i}^{j+1})^T (P_{K_i}^{j+1} - P_{K_i}^j) + (P_{K_i}^{j+1} - P_{K_i}^j) A_{K_i}^{j+1} \\ &\quad + \gamma^2 (L_{K_i}^{j+1} - L_{K_i}^j)^T (L_{K_i}^{j+1} - L_{K_i}^j) = 0. \end{aligned} \tag{B.39}$$

As $\gamma^2(L_{K_i}^{j+1} - L_{K_i}^j)^T(L_{K_i}^{j+1} - L_{K_i}^j) \succeq 0$ and $A_{K_i}^{j+1}$ is Hurwitz, by (B.39) and Lemma A.3, we have $P_{K_i}^{j+1} \succeq P_{K_i}^j$. As a result, the proof of statement 2) is completed.

Statement 2) implies that the sequence $\{P_{K_i}^j\}_{j=1}^\infty$ is monotonically increasing and upper-bounded by P_{K_i}. Hence, $P_{K_i}^\infty$ exists, and is the solution to (27a). Due to the uniqueness of the solution, we have $P_{K_i}^\infty = P_{K_i}$, which proves the statement 3). □

As shown in [48], policy iteration has a quadratic convergence rate in the vicinity of the solution P_{K_i}. Next, we will show that the inner loop iteration has a global linear convergence rate.

Lemma B.3. *Suppose that*

$$E_K^j := \gamma^{-2}(D^T P_K^j - \gamma^2 L_K^j)^T (D^T P_K^j - \gamma^2 L_K^j), \tag{B.40}$$

and $K \in \mathcal{K}$. Then, for the sequences $\{P_K^j\}_{j=0}^\infty$ and $\{L_K^j\}_{j=1}^\infty$ obtained by the inner-loop iteration (29), the following inequality holds

$$\mathrm{Tr}(P_K - P_K^j) \leq \|E_K^j\| c(K), \tag{B.41}$$

where

$$c(K) = \mathrm{Tr}\left(\int_0^\infty e^{(A_K+DL_K)t} e^{(A_K+DL_K)^T t} dt\right). \tag{B.42}$$

Proof. Subtracting (29) from (27a) and completing squares, we have

$$\begin{aligned}(A_K+DL_K)^T(P_K-P_K^j)+(P_K-P_K^j)(A_K+DL_K)\\ -\gamma^{-2}(D^TP_K-D^TP_K^j)^T(D^TP_K-D^TP_K^j)\\ +\gamma^{-2}(D^TP_K^j-\gamma^2L_K^j)^T(D^TP_K^j-\gamma^2L_K^j)=0.\end{aligned} \tag{B.43}$$

Because $K \in \mathcal{K}$, $A_K + DL_K$ is Hurwitz by Lemma A.9. Using Lemma A.3, we have

$$P_K - P_K^j \preceq \int_0^\infty e^{(A_K+DL_K)^T t} E_K^j e^{(A_K+DL_K)t} dt.$$

By Lemma A.5 and the cyclic property of the trace, we have

$$\mathrm{Tr}(P_K - P_K^j) \leq \|E_K^j\| \underbrace{\mathrm{Tr}\left(\int_0^\infty e^{(A_K+DL_K)t} e^{(A_K+DL_K)^T t} dt\right)}_{:=c(K)}. \tag{B.44}$$

Therefore, $\mathrm{Tr}(P_K - P_K^j) \leq \|E_K^j\| c(K)$ holds. □

The proof of the theorem given in Theorem 2 is now given.

Proof. By Theorem B.2, A_K^{j+1} is Hurwitz. By Lemma A.3 and (B.39), we have

$$P_K^{j+1} - P_K^j = \underbrace{\int_0^\infty e^{(A_K^{j+1})^T t} E_K^j e^{A_K^{j+1} t} dt}_{:=F_K^j}. \tag{B.45}$$

Therefore,

$$P_K - P_K^{j+1} = P_K - P_K^j - F_K^j. \tag{B.46}$$

By Theorem B.2, $P_K \succeq P_K^j$ so that

$$\begin{aligned}\|A_K^{j+1}\| &= \|A - BK + \gamma^{-2}DD^TP_K^j\|\\ &\leq \|A-BK\| + \gamma^{-2}\|DD^T\|\|P_K\|.\end{aligned} \tag{B.47}$$

Define

$$d(K) := \frac{\log(5/4)}{\|A-BK\| + \gamma^{-2}\|DD^T\|\|P_K\|}. \tag{B.48}$$

Taking the trace of both sides of (B.46), we find that

$$\begin{aligned}\mathrm{Tr}(P_K - P_K^{j+1}) &= \mathrm{Tr}(P_K - P_K^j) - \mathrm{Tr}(F_K^j) && \text{(B.49a)}\\ &\leq \mathrm{Tr}(P_K - P_K^j) - \|F_K^j\| && \text{(B.49b)}\\ &\leq \mathrm{Tr}(P_K - P_K^j) - \frac{1}{2}d(K)\|E_K^j\|, && \text{(B.49c)}\end{aligned}$$

where we have used Lemma A.1 to arrive at the inequality in (B.49b), and we have used Lemma B.2 to arrive at the inequality in (B.49c). Furthermore, using Lemma B.3, we find that

$$\begin{aligned}\mathrm{Tr}(P_K - P_K^{j+1}) &\leq \mathrm{Tr}(P_K - P_K^j) - \frac{1}{2}\frac{d(K)}{c(K)}\mathrm{Tr}(P_K - P_K^j)\\ &\leq \underbrace{\left(1 - \frac{1}{2}\frac{d(K)}{c(K)}\right)}_{:=\beta(K)} \mathrm{Tr}(P_K - P_K^j).\end{aligned} \tag{B.50}$$

Equation (B.50) is equivalent to (30), and gives us the required result. □

D. Proof: Uniform Convergence

We now establish the proof of Theorem 3.

Proof. Let $M_i := \int_0^\infty e^{A_i^T t} e^{A_i t} dt$. Following Theorem B.1 and Lemma A.9, we see that A_i is Hurwitz. Hence, by Lemma A.3 for any $i \in \mathbb{N}_+$,

$$A_i^T M_i + M_i A_i + I_n = 0. \tag{B.51}$$

From Theorem B.1, we get that as $i \to \infty$, P_{K_i} converges $P^\star$ and A_i converges to $A^\star$. Consequently, M_i converges to $M^\star$, which is the solution of (B.10). Consequently, $\bar{c} := \sup_{i \in \mathbb{R}_+} \mathrm{Tr}(M_i) < \infty$.

Hence, we have

$$c(K_i) = \mathrm{Tr}(M_i) \leq \bar{c}, \tag{B.52}$$

where $c(K_i)$ is defined in (B.41). Recall from Theorem B.1 that

$$d(K_i) \geq \frac{\log(5/4)}{\|A\| + (\|BR^{-1}B^T\| + \gamma^{-2}\|DD^T\|)\|P_{K_1}\|} =: \underline{d}. \tag{B.53}$$

Hence,

$$\beta(K_i) = 1 - \frac{d(K_i)}{2c(K_i)} \leq 1 - \frac{\underline{d}}{2\bar{c}} =: \bar{\beta}. \tag{B.54}$$

Therefore, by Theorem 2, for any $i \in \mathbb{N}_+$, we have

$$\|P_{K_i}^j - P_{K_i}\|_F \leq \bar{\beta}^{j-1}\mathrm{Tr}(P_{K_i}) \leq \bar{\beta}^{j-1}\mathrm{Tr}(P_{K_1}). \tag{B.55}$$

We see that for any $i \in \mathbb{N}_+$ and $\epsilon > 0$, there exists $\bar{j} > 0$, such that if $j \geq \bar{j}$, $\|P_{K_i}^j - P_{K_i}\|_F \leq \epsilon$. *A fortiori*, the iteration in fact converges uniformly to an $\epsilon > 0$. □

Appendix C
Robustness To Perturbations

As in the foregoing appendices, we again introduce a few preliminary results before we establish our main result.

A. Preliminaries

Let $\hat{A}_{K_i}^j := A - BK_i + D\hat{L}_{K_i}^j$ denote the two-player transition matrix, and $\Delta L_{K_i}^j$ denote the influence from the noise. In the rest of this appendix, the subscript i will be discarded for notational simplicity.

B. Control (Outer) Loop

In order to prove Theorem 4, let us first introduce the following preliminary Lemma.

Lemma C.1. *For any $K \in \mathcal{K}$, there exists an $l(K) > 0$ such that for a perturbation of K by ΔK, $K + \Delta K \in \mathcal{K}$, as long as $\|\Delta K\|_F \leq l(K)$.*

Proof. Let us introduce the perturbations $\Delta P \in \mathbb{S}^n$ and $\Delta K \in \mathbb{R}^{m\times n}$ to P and K respectively. Next, we introduce a matrix-valued function, $F(\Delta P, \Delta K)$, in ΔP and ΔK as follows

$$\begin{aligned} &F(\Delta P,\Delta K) = (A_K+\gamma^{-2}DD^TP_K)^T\Delta P \\ &+\Delta P(A_K+\gamma^{-2}DD^TP_K) - \Delta K^TB^T(P_K+\Delta P) \\ &-(P_K+\Delta P)B\Delta K + \Delta K^TRK + K^TR\Delta K \\ &+\Delta K^TR\Delta K + \gamma^{-2}\Delta PDD^T\Delta P. \end{aligned} \tag{C.1}$$

Let $\mathcal{F}(\mathrm{vec}(\Delta P),\Delta K) := \mathrm{vec}(F(\Delta P,\Delta K))$, then

$$\begin{aligned} &\mathcal{F}(\mathrm{vec}(\Delta P),\Delta K) = \big[I_n\otimes(A_K+\gamma^{-2}DD^TP_K)^T \\ &+(A_K+\gamma^{-2}DD^TP_K)^T\otimes I_n\big]\,\mathrm{vec}(\Delta P) \\ &-(P_KB\otimes I_n)\,\mathrm{vec}(\Delta K^T) - (I_n\otimes P_KB)\,\mathrm{vec}(\Delta K) \\ &-(I_n\otimes\Delta K^TB^T+\Delta K^TB^T\otimes I_n)\,\mathrm{vec}(\Delta P) \\ &+(K^TR\otimes I_n)\,\mathrm{vec}(\Delta K^T) + (I_n\otimes K^TR)\,\mathrm{vec}(\Delta K) \\ &+\mathrm{vec}(\Delta K^TR\Delta K) + \gamma^{-2}\,\mathrm{vec}(\Delta PDD^T\Delta P). \end{aligned} \tag{C.2}$$

Using Lemma A.6, we have

$$\begin{aligned} \frac{\partial\,\mathrm{vec}(\Delta PDD^T\Delta P)}{\partial\,\mathrm{vec}(\Delta P)} &= (\Delta P\otimes I_n)\frac{\partial\,\mathrm{vec}(\Delta PDD^T)}{\partial\,\mathrm{vec}(\Delta P)} \\ &+(I_n\otimes\Delta PDD^T)\frac{\partial\,\mathrm{vec}(\Delta P)}{\partial\,\mathrm{vec}(\Delta P)} \\ &=(\Delta P\otimes I_n)(DD^T\otimes I_n) + I_n\otimes\Delta PDD^T \\ &=\Delta PDD^T\otimes I_n + I_n\otimes\Delta PDD^T. \end{aligned} \tag{C.3}$$

Therefore,

$$\begin{aligned} &\frac{\partial\mathcal{F}}{\partial\,\mathrm{vec}(\Delta P)}(\mathrm{vec}(\Delta P),\mathrm{vec}(\Delta K)) \\ &= I_n\otimes[(A_K+\gamma^{-2}DD^TP_K)-B\Delta K]^T \\ &+[(A_K+\gamma^{-2}DD^TP_K)-B\Delta K]^T\otimes I_n \\ &+\Delta PDD^T\otimes I_n + I_n\otimes\Delta PDD^T. \end{aligned} \tag{C.4}$$

Observe that $\mathcal{F}(0,0) = 0$. Moreover, since $(A_K+\gamma^{-2}DD^TP_K)$ is Hurwitz, $\frac{\partial\,\mathrm{vec}(\Delta PDD^T\Delta P)}{\partial\,\mathrm{vec}(\Delta P)}|_{\Delta P=0,\Delta K=0}$ is invertible. From implicit function theorem, that there exists a scalar $l_1>0$, such that $\mathrm{vec}(\Delta P)$ is continuously differentiable with respect to ΔK for any $\Delta K\in\mathcal{B}(0,l_1)$. Thus, $\|\Delta P\|_F\to 0$ as $\|\Delta K\|_F\to 0$. Since $K\in\mathcal{K}$, by Lemma A.9, $P_K\succ 0$. Therefore, there exists $l(K)>0$, such that $\sigma_{\max}(\Delta P)\le\sigma_{\min}(P_K)$, i.e. $P_K-\Delta P\succeq 0$, as long as $\|\Delta K\|_F\le l(K)$.

As ΔP and ΔK satisfy $F(\Delta P,\Delta K)=0$, we have

$$\begin{aligned} &(A-BK-B\Delta K)^T(P_K+\Delta P)+(P_K+\Delta P)+C^TC \\ &\quad+(K+\Delta K)^TR(K+\Delta K) \\ &\quad+\gamma^{-2}(P_K+\Delta P)DD^T(P_K+\Delta P)=0. \end{aligned} \tag{C.5}$$

As $P_K+\Delta P\succeq 0$ when $\|\Delta K\|\le l(K)$, by Lemma A.9, $K+\Delta K\in\mathcal{K}$. □

From Lemma C.1, we see that for a robust and stabilizing controller $K\in\mathcal{K}$, after a small perturbation ΔK, it is still robust and stabilizing. Let us now show that the inexact outer loop does converge to the optimal gain $K^\star$ despite perturbations.

Proof of Theorem 4. Assume $\hat K_i\in\mathcal{K}$, and $\|\Delta K_{i+1}\|_F\le l(\tilde K_{i+1})$. By Lemma C.1, $\hat K_{i+1}\in\mathcal{K}$. Rewriting (33a) for the $(i+1)$th iteration and subtracting (33a) from it yields

$$\begin{aligned} &\hat A_{i+1}^T(\hat P_{K_i}-\hat P_{K_{i+1}})+(\hat P_{K_i}-\hat P_{K_{i+1}})\hat A_{i+1} \\ &\quad+(R\hat K_i-B^T\hat P_{K_i})^TR^{-1}(R\hat K_i-B^T\hat P_{K_i}) \\ &\quad+\gamma^{-2}(\hat P_{K_i}-\hat P_{K_{i+1}})DD^T(\hat P_{K_i}-\hat P_{K_{i+1}}) \\ &\quad-\Delta K_{i+1}^TR\Delta K_{i+1}=0. \end{aligned} \tag{C.6}$$

Therefore, we have

$$\hat P_{K_i}-\hat P_{K_{i+1}}\succeq\hat F_{K_i}-\int_0^\infty e^{\hat A_{i+1}^Tt}(\Delta K_{i+1}^TR\Delta K_{i+1})e^{\hat A_{i+1}t}dt, \tag{C.7}$$

where $\hat E_{K_i} = (R\hat K_i-B^T\hat P_{K_i})^TR^{-1}(R\hat K_i-B^T\hat P_{K_i})$, and $\hat F_{K_i} := \int_0^\infty e^{\hat A_{i+1}^Tt}\hat E_{K_i}e^{\hat A_{i+1}t}dt$. Let $h_i := \|\hat A_i\|$ and $s_i := \mathrm{Tr}(\int_0^\infty e^{\hat A_{i+1}t}e^{\hat A_{i+1}^Tt}dt)$. By Lemma B.2, $\|\hat F_{K_i}\|\ge\frac{\log(5/4)}{2h_i}\|\hat E_{K_i}\|$. By Lemma B.1, $\|\hat E_{K_i}\|\ge\frac{1}{c}\mathrm{Tr}(\hat P_{K_i}-P^\star)$. As a consequence,

$$\begin{aligned} \mathrm{Tr}(\hat P_{K_{i+1}}-P^\star) &\le (1-\frac{\log(5/4)}{2h_ic})\,\mathrm{Tr}(\hat P_{K_i}-P^\star) \\ &\quad+s_i\bar\sigma(R)\|\Delta K_{i+1}\|^2. \end{aligned} \tag{C.8}$$

Hence, if $\|\Delta K_i\|_F\le l(\tilde K_i)$ for all $i\le i'$,

$$\begin{aligned} \mathrm{Tr}(\hat P_{K_i}-P^\star) &\le \prod_{i=1}^{i'-1}\left(1-\frac{\log(5/4)}{2h_ic}\right)\mathrm{Tr}(\hat P_{K_1}-P^\star) \\ &\quad+\sum_{i=1}^{i'-1}s_i\bar\sigma(R)\|\Delta K_{i+1}\|^2. \end{aligned} \tag{C.9}$$

Let $\underline{l} := \inf_{i\in\mathbb{N}_+}l(\tilde K_i)$, $\bar h := \sup_{i\in\mathbb{N}_+}h_i$, $\bar s := \sup_{i\in\mathbb{N}_+}s_i$, $\hat\alpha := (1-\frac{\log(5/4)}{2\bar hc})$, and $\kappa(\|\Delta K\|_\infty) = \frac{\bar s\bar\sigma(R)}{1-\hat\alpha}\|\Delta K\|_\infty^2$. And if $\|\Delta K\|_\infty\le\underline{l}$, for any $i\in\mathbb{N}_+$

$$\mathrm{Tr}(\hat P_{K_i}-P^\star)\le\hat\alpha^{i-1}\,\mathrm{Tr}(\hat P_{K_1}-P^\star)+\kappa(\|\Delta K\|_\infty).$$

As $\|\hat P_{K_i}-P^\star)\|_F\le\mathrm{Tr}(\hat P_{K_i}-P^\star)$ by Lemma A.1, the Theorem 4 holds. □

We next prove the robustness of the disturbance (inner) loop to perturbations.

C. Disturbance (Inner) Loop

Proof of Theorem 5. When $\|\Delta L_K^j\|<e$, we have $\hat A_K^j$ as Hurwitz, owing to Lemma C.2. Rewriting (34a) for the $(i+1)$th iteration and subtracting (34a) from it, we have

$$\begin{aligned} &(\hat A_K^{j+1})^T(\hat P_K^{j+1}-\hat P_K^j)+(\hat P_K^{j+1}-\hat P_K^j)(\hat A_K^{j+1}) \\ &\quad+\gamma^{-2}(\gamma^2\hat L_K^j-D^T\hat P_K^j)^T(\gamma^2\hat L_K^j-D^T\hat P_K^j) \\ &\quad-\gamma^2\Delta {L_K^j}^T\Delta L_K^j=0. \end{aligned} \tag{C.10}$$

As $\hat A_K^{j+1}$ is Hurwitz, we have

$$\begin{aligned} &\hat P_K^{j+1}-\hat P_K^j \\ &=\int_0^\infty e^{(\hat A_K^{j+1})^Tt}\left[\hat E_K^j-\gamma^2\Delta(L_K^j)^T\Delta L_K^j\right]e^{(\hat A_K^{j+1})t}dt \end{aligned} \tag{C.11}$$

where

$$\hat{E}_K^j := \gamma^{-2}(\gamma^2\hat{L}_K^j - D^T\hat{P}_K^j)^T(\gamma^2\hat{L}_K^j - D^T\hat{P}_K^j). \quad \text{(C.12)}$$

Define $\hat{F}_K^j := \int_0^\infty e^{(\hat{A}_K^{j+1})^T t}\hat{E}_K^j e^{(\hat{A}_K^{j+1})t}dt$ so that

$$\begin{aligned}&\hat{P}_K - \hat{P}_K^{j+1} = \hat{P}_K - \hat{P}_K^j - \hat{F}_K^j \\ &+ \int_0^\infty e^{(\hat{A}_K^{j+1})^T t}\left(\gamma^2 {\Delta L_K^j}^T \Delta L_K^j\right)e^{(\hat{A}_K^{j+1})t}dt.\end{aligned} \quad \text{(C.13)}$$

Let $f_K = \sup_{j\in\mathbb{N}_+}\|\hat{A}_K^{j+1}\|$. From Lemma B.2, we can write, $-\|\hat{F}_K^j\| \leq -\frac{\log(5/4)}{2f_K}\|\hat{E}_K^j\|$. Furthermore, by Lemma B.3, we can write $-\|\hat{E}_K^j\| \leq -\frac{1}{c(K)}\operatorname{Tr}(P_K - \hat{P}_K^j)$, where $c(K) = \operatorname{Tr}(\int_0^\infty e^{(A_K+DL_K)t}e^{(A_K+DL_K)^T t}dt)$ is defined in (B.41). Therefore, the trace of (C.13) can be written as

$$\begin{aligned}&\operatorname{Tr}(P_K - \hat{P}_K^{j+1}) \leq \left(1 - \frac{\log(5/4)}{2f_K c(K)}\right)\operatorname{Tr}(P_K - \hat{P}_K^j) \\ &+ \operatorname{Tr}\left(\int_0^\infty e^{(\hat{A}_K^{j+1})t}e^{(\hat{A}_K^{j+1})^T t}dt\right)\gamma^2\|\Delta L_K^j\|^2.\end{aligned} \quad \text{(C.14)}$$

Setting g and $\hat{\beta}(K)$ as

$$g := \sup_{j\in\mathbb{N}_+}\operatorname{Tr}(\int_0^\infty e^{(\hat{A}_K^{j+1})t}e^{(\hat{A}_K^{j+1})^T t}dt) \quad \text{(C.15a)}$$

$$\hat{\beta}(K) := 1 - \frac{\log(5/4)}{2f_K c(K)}, \quad \text{(C.15b)}$$

we find that

$$\operatorname{Tr}(P_K - \hat{P}_{K_i}^j) \leq \hat{\beta}^{j-1}(K)\operatorname{Tr}(P_K) + \lambda(\|\Delta L\|_\infty), \quad \text{(C.16)}$$

where $\lambda(\|\Delta L\|_\infty) := \frac{1}{1-\hat{\beta}(K)}\gamma^2 g\|\Delta L\|_\infty^2$. As $\|P_K - \hat{P}_{K_i}^j\|_F \leq \operatorname{Tr}(P_K - \hat{P}_{K_i}^j)$, we thus establish the statement of Thereom 5. $\square$

Lemma C.2. *Given a $K \in \mathcal{K}$, there exists an $e > 0$, such that if $\|\Delta L_K^j\|_F \leq e$, $\hat{A}_K^j$ is Hurwitz for all $j \in \mathbb{N}_+$.*

Proof. Assume $\hat{A}_K^j$ is Hurwitz. From (27a), we have

$$\begin{aligned}&(\hat{A}_K^{j+1})^T P_K + P_K\hat{A}_K^{j+1} + Q_K - \gamma^{-2}\hat{P}_K^j DD^T\hat{P}_K^j \\ &\quad + \gamma^{-2}(P_K - \hat{P}_K^j)DD^T(P_K - \hat{P}_K^j) \\ &\quad - (\Delta L_K^j)^T D^T P_K - P_K D\Delta L_K^j = 0.\end{aligned} \quad \text{(C.17)}$$

If $\|\Delta L_K^j\| < \sigma_{\min}(Q_K - \gamma^{-2}P_K DD^T P_K)/2\|D^T P_K\| =: e$, it follows from (B.36) and the the inequality $P_K \succeq \hat{P}_K^j$ that

$$\begin{aligned}&Q_K - \gamma^{-2}\hat{P}_K^j DD^T\hat{P}_K^j \\ &\quad + \gamma^{-2}(P_K - \hat{P}_K^j)DD^T(P_K - \hat{P}_K^j) \\ &\quad - (\Delta L_K^j)^T D^T P_K - P_K D\Delta L_K^j \succ 0.\end{aligned}$$

Consequently, $\hat{A}_K^{j+1}$ is Hurwitz. Since $\hat{L}_K^1 = 0$ and $K \in \mathcal{K}$, $\hat{A}_K^1 = A - BK$ is Hurwitz. As a result, $\hat{A}_K^j$ is Hurwitz for all $j \in \mathbb{N}_+$ as long as $\|\Delta L_K^j\|_F \leq e$. $\square$

References

[1] G. Stein and M. Athans, "The LQG/LTR procedure for multivariable feedback control design," *IEEE Transactions on Automatic Control*, vol. 32, no. 2, pp. 105–114, 1987. 1

[2] M. Athans, "The Role and Use of The Stochastic Linear-Quadratic-Gaussian Problem in Control System Design," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 529–552, 1971. 1

[3] N. Papanikolopoulos, P. Khosla, and T. Kanade, "Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 1, pp. 14–35, 1993. 1

[4] W. Xu, J. Pan, J. Wei, and J. M. Dolan, "Motion planning under uncertainty for on-road autonomous driving," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2507–2512, 2014. 1

[5] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience*, vol. 5, no. 11, p. 1226–1235, 1993. 1

[6] J. Doyle, "Guaranteed margins for LQG regulators," *IEEE Transactions on Automatic Control*, vol. 23, no. 4, pp. 756–757, 1978. 1, 2, 4

[7] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 124–131, 1973. 1, 2, 3

[8] T. E. Duncan, "Linear-Exponential-Quadratic Gaussian control," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2910–2911, 2013. 1, 2, 3, 4

[9] P. Khargonekar, I. Petersen, and M. Rotea, "$\mathcal{H}_\infty$ optimal control with state-feedback," *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 786–788, 1988. 1

[10] D. Bernstein and W. Haddad, "LQG control with an $\mathcal{H}_\infty$ performance bound: a Riccati equation approach," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 293–305, 1989. 1

[11] D. Mustafa, "Relations between maximum-entropy/$\mathcal{H}_\infty$ control and combined $\mathcal{H}_\infty$/LQG control," *Systems and Control Letters*, vol. 12, no. 3, pp. 193–203, 1989. 1

[12] T. Başar, *H_∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer, 2008. 1, 3, 4, 9

[13] A. Lanzon, Y. Feng, B. D. O. Anderson, and M. Rotkowitz, "Computing the positive stabilizing solution to algebraic riccati equations with an indefinite quadratic term via a recursive method," *IEEE Transactions on Automatic Control*, vol. 53, no. 10, pp. 2280–2291, 2008. 1

[14] 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 e-prints*, p. arXiv:1910.09496, Oct. 2019. 1, 2, 3, 6, 8

[15] H.-N. Wu and B. Luo, "Simultaneous policy update algorithm for learning the solution of linear continuous-time $\mathcal{H}_\infty$ state feedback control," *Information Sciences*, vol. 222, p. 472–485, 02 2013. 1

[16] J. Bu, L. J. Ratliff, and M. Mesbahi, "Global Convergence of Policy Gradient for Sequential Zero-Sum Linear Quadratic Dynamic Games," *arXiv e-prints*, Nov. 2019. 1

[17] Z.-P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, pp. 176–284, 2020. 2

[18] Y. Jiang and Z. P. Jiang, *Robust Adaptive Dynamic Programming*. NJ, USA: Wiley-IEEE Press, 2017. 2

[19] Y. Jiang and Z. P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012. 2

[20] T. Bian and Z. P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348–360, 2016. 2

[21] T. Bian and Z. P. Jiang, "Reinforcement learning and adaptive optimal control for continuous-time nonlinear systems: a value iteration approach," *IEEE Transactions on Neural Networks and Learning Systems*, in press, 2021. 2

[22] B. Pang and Z. P. Jiang, "Adaptive optimal control of linear periodic systems: an off-policy value iteration approach," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 888–894, 2021. 2

[23] W. Gao and Z. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 4164–4169, 2016. 2

[24] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. NJ, USA: Wiley-IEEE Press, 2013. 2

[25] B. Pang, T. Bian, and Z.-P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Transactions on Automatic Control*, vol. 67, no. 1, pp. 504–511, 2022. 2

[26] J. Bu, A. Mesbahi, and M. Mesbahi, "Policy Gradient-based Algorithms for Continuous-time Linear Quadratic Control," *arXiv e-prints*, p. arXiv:2006.09178, June 2020. 2

[27] 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 Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2022. 2

[28] G. Qu, C. Yu, S. Low, and A. Wierman, "Combining Model-Based and Model-Free Methods for Nonlinear Control: A Provably Convergent Policy Gradient Approach," *arXiv e-prints*, June 2020. 2

[29] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for h_∞ control design," *IEEE Transactions on Cybernetics*, vol. 45, no. 1, pp. 65–76, 2015. 2

[30] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, 2011. 2, 10

[31] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 706–714, 2014. 2

[32] D. Vrabie and F. Lewis, "Adaptive dynamic programming for online solution of a zero-sum differential game," *J. Contr. Theory Appl.*, vol. 9, pp. 353–360, 08 2011. 2

[33] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *IEEE Transactions on Automatic Control*, pp. 1–1, 2021. 2

[34] H. Mania, A. Guy, and B. Recht, "Simple random search of static linear policies is competitive for reinforcement learning," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018. 2

[35] K. Zhang, Z. Yang, and T. Basar, "Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019. 2

[36] 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. 2

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

[38] B. Luo, Y. Yang, and D. Liu, "Policy iteration q-learning for data-based two-player zero-sum game of linear discrete-time systems," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3630–3640, 2021. 2

[39] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 2817–2826, PMLR, 06–11 Aug 2017. 2

[40] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural Computation*, vol. 17, no. 2, pp. 335–359, 2005. 2

[41] M. Rotea and P. Khargonekar, "Mixed $H2/H_\infty$ Control: A Convex Optimization Approach," *IEEE Trans. Automat. Control*, vol. 36, no. 5, pp. 824–837, 1991. 2

[42] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 301–320, 1981. 2

[43] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems," *IEEE Transactions on Automatic Control*, vol. 34, no. 8, pp. 831–847, 1989. 2

[44] P. Whittle, "Risk-sensitive linear/quadratic/gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981. 3

[45] O. Ogunmolu, N. Gans, and T. Summers, "Minimax iterative dynamic game: Application to nonlinear robot control tasks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6919–6925, IEEE, 2018. 4

[46] G. A. Pavliotis, *Stochastic Processes and Applications*. Springer, 2014. 4

[47] W. Tan, *Nonlinear Control Analysis and Synthesis using Sum-of-Squares Programming*. PhD thesis, Dept. of Engineering-Mechanical Engineering, University of Califonia, Berkeley, Berkeley, CA, USA, 2006. 5, 6

[48] D. Z. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, pp. 114–115, 1968. 5, 6, 10, 12

[49] E. D. Sontag, *Input to State Stability: Basic Concepts and Results*, pp. 163–220. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 6

[50] Z.-P. Jiang, Y. Lin, and Y. Wang, "Nonlinear Small-gain Theorems for Discrete-time Feedback Systems and Applications," *Automatica*, vol. 40, no. 12, pp. 2129–2136, 2004. 6

[51] M. González-Fierro, C. Balaguer, N. Swann, and T. Nanayakkara, "A humanoid robot standing up through learning from demonstration using a multimodal reward function," in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 74–79, 2013. 7

[52] R. D. Pristovani, D. R. Sanggar, and P. Dadet, "Implementation of push recovery strategy using triple linear inverted pendulum model in "t-FloW" humanoid robot," *Journal of Physics: Conference Series*, vol. 1007, p. 012068, apr 2018. 7

[53] K. Furut, T. Ochiai, and N. Ono, "Attitude control of a triple inverted pendulum," *International Journal of Control*, vol. 39, no. 6, pp. 1351–1365, 1984. 7

[54] K. Zhou and J. C. Doyle, *Essentials of robust control*, vol. 104. Prentice hall Upper Saddle River, NJ, 1998. 9

[55] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice hall Upper Saddle River, NJ, 1996. 9

[56] R. A. Horn and C. R. Johnson, *Matrix Analysis, second edition*. Cambridge University Press, 2013. 9

[57] T. Mori, "Comments on "a matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation" by J. M. Saniuk and I.B. Rhodes," *IEEE Transactions on Automatic Control*, vol. 33, no. 11, pp. 1088–, 1988. 9

[58] 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. 9

[59] L. Koralov and Y. G. Sinai, *Theory of Probability and Random Processes*. Springer Berlin, Heidelberg, 2nd ed. ed., 2007. 9

[60] A. Arapostathis, V. S. Borkar, and M. K. Ghosh, *Ergodic Control of Diffusion Processes*. Cambridge University Press, 2011. 9

[61] T. S. Lee and F. Kozin, "Almost sure asymptotic likelihood theory for diffusion processes," *Journal of Applied Probability*, vol. 14, no. 3, p. 527–537, 1977. 9