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

Leilei Cui and Lekan Molu, 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.

$\mathcal{H}_2$ 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 aforeside) 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 averse 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 they 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 $H_2$ sense only. As a result, it is not clear what robustness properties the resulting controller possesses.

Robust reinforcement learning based on mixed $H_2/H_\infty$ 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 $H_\infty$ 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 $H_2$ 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 $H_2/H_\infty$ 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 $H_2/H_\infty$ 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 $H_2/H_\infty$ control [41], we introduce an iterative solution to the cost matrix of $H_2$ 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 $H_\infty$ 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 $H_2/H_\infty$ policy optimization problem, and a learning-based mixed $H_2/H_\infty$ 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. We let $\| \cdot \|$ denote the 2-norm of a vector or the induced matrix norm. We let $\| \cdot \|_F$ denote the Frobenius norm of a matrix.

The Kronecker product is denoted by $\otimes$. The sub-matrix of the matrix $A$ that is comprised of the rows between the $i$th and $j$th rows is denoted by $[A]_{i:j}$. The maximum and minimum singular values of a matrix $T$ are respectively denoted by $\sigma(T)$ and $\sigma(T)$. For $X \in \mathbb{R}^{m \times n}$ and $\delta > 0$, $B(X, \delta) := \{Y \in \mathbb{R}^{m \times n} \mid \| Y - X \|_F \leq \delta \}$. The identity matrix with dimension $n$ is $I_n$. For the transfer function $G(s)$, its $H_\infty$ norm is a bounded linear operator defined as $\| G \|_{H_\infty} = \sup_{\omega \in \mathbb{R}} \sigma(G(j \omega))$.

A symmetric matrix with dimension $n$ is denoted $\mathbb{S}^n$. For a $P \in \mathbb{S}^n$, $\text{vecs}(P) := [p_{11}, 2p_{12}, \cdots, 2p_{m-1, m}, p_{22}, \cdots, p_{nn}]^T$, where $p_{ij}$ is the $i$th row and $j$th column entry of $P$. Let the operator $\text{vec}(A) := [a_1^T, \cdots, a_n^T]^T$, where $a_i$ is the $i$th column of the matrix $A$, and let $\text{vec}(x) := [x_1^2, x_1 x_2, \cdots, x_1 x_n, x_2^2, x_2 x_3, \cdots, x_n^2]^T$. We let $T_{n}^w$ and $T_{n}s$ respectively denote inverse transformations that represent the matricization of $\text{vecs}(P)$ and $\text{vec}(P)$ for a $P \in \mathbb{S}^n$, i.e.
vecs(P) = T_{et}^u vec(P), \quad vec(P) = T_{et}^u vecs(P). In addition, we let $T_{et}$ denote the matricization of the vectorized $X \in \mathbb{R}^{m \times n}$, so that vec($X^T$) = $T_{et}$ vec(X).

**B. Mixed Design as the LEQG Problem**

Consider the following stochastic autonomous system

\[
\begin{align*}
\dot{x}(t) &= Ax(t) dt + Bu(t) dt + Ddw(t), \quad x(0) = x_0, \quad (1a) \\
\dot{z}(t) &= Cx(t) + Eu(t), \quad (1b)
\end{align*}
\]

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 $(\Omega, \mathcal{F}, P)$, 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

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

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 $\gamma^{-2}$, the LEQG cost is

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

where $\text{sign}(\gamma)$ for a $\gamma \in \mathbb{R}$ dictates a risk-avoidance (i.e. $\gamma > 0$), risk-seeking (i.e. $\gamma < 0$), or risk-neutral (i.e. $\gamma = 0$), whereupon we have the standard LQG problem) behavior for an ensuing controller to be sought. In [14], the authors pointed out that the standard LQG problem admits the form of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ suboptimal 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, $\|T(K)\|_{\mathcal{H}_\infty} \leq \gamma$ so that all stabilizing feedback gains, $K$, for system (1) can be realized within the set

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

Formally, the control optimization problem is to find the set of stabilizing gains $K$ via $\min_K J(K)$ such that $K \in K$. 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.** Because we desire a stabilizing feedback controller, we require $(A, B)$ to be stabilizable. Furthermore, we require that unstable modes of $A$ be observable in order to guarantee a solution to the minimization of $J_K \in K$. Whence, we require that $(C, A)$ be observable.

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

**Remark 1.** Assumption 1 implies that the indefinite ARE associated with (3) has a unique positive definite solution and

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.

**hence guarantees the existence of a stabilizing control law $u$.**

**Assumption 2 simplifies the derivation of the solution to the indefinite ARE.**

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

\[
u^*(x(t)) = -R^{-1}B^T P^* x(t), \quad (5)
\]

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

\[
A^T P^* + P^* A + C^T C - P^* (BR^{-1}B^T - \gamma^{-2}DD^T) P^* = 0. \quad (6)
\]

**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

\[
\begin{align*}
u(x(t)) &= -K_r(x(t)) := -R^{-1}B^T P_r(t)x(t) \quad (7)
\end{align*}
\]

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

\[
J(K) = \mathbb{E} \left[ \exp \left( \frac{1}{2\gamma^2} \int_0^\tau z^T(t)z(t) dt \right) \right], \quad (8)
\]

where $P_r$ is the solution to the following continuous-time closed-loop (generalized) algebraic Riccati equation (GARE)

\[
A^T P_r + P_r A + C^T C - P_r (BR^{-1}B^T - \gamma^{-2}DD^T) P_r = 0 \quad (9)
\]

**A fortiori**, given assumptions 1 and 2, we have $\lim_{\tau \to \infty} P_r(t) = P^*$, and $\lim_{\tau \to \infty} K_r(t) = K^*$ (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,

\[
\begin{align*}
\min_{u \in \mathbb{R}^m} \max_{w \in \mathbb{R}^q} J(\cdot) &= \int_{t=0}^\infty z^T(t)z(t) - \gamma^2 w^T(t)w(t) dt \quad (10)
\end{align*}
\]

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.
Remark 2. Suppose that there exists a nonempty set \( \gamma_\infty = \inf\{\gamma > 0 \mid \min_u \max_w J(0, u, w) \leq 0\} \), then the associated Riccati differential equation to (3) has a solution for a sufficiently large \( \gamma \). i.e. \( \gamma > \gamma_\infty \) on \([0, \tau]\).

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

\[
 u^*(\cdot) = -R^{-1}B^TP^*x(t), \quad w^*(t) = \gamma^{-2}D^TP^*x(t). \tag{11}
\]

In addition, \( P^* > 0 \) is the stabilizing solution of (6).

Proof. This is just a statement of Th. 4.8 and 9.7 in [12]. \( \Box \)

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 (6) and the equivalence of the \( H_\infty \) norm bound on the system transfer function to the Riccati equation and inequality in Lemma A.9, we conclude that the optimal \( K^* \in \mathcal{K} \). Observe: minimizing the performance index (10) under the worst-case disturbance, the optimal controller \( u^* \) can robustly improve the system performance w.r.t the \( H_\infty \) norm penalty (this is demonstrated in our experiments).

Remark 3. 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 (11).

While Duncan [8] proposed the Riccati equation (6) as a means to realizing the optimal controller in (5), no closed-form solution to the ARE exists to our knowledge. Conventional LQG control cannot guarantee the robustness of system (3) or (10) as found by Doyle [6]. In what follows, we first introduce a model-based algorithm followed by a model-free proposal for realizing the policies for \( u \) and \( w \) respectively.

III. MODEL-BASED ITERATIVE ALGORITHM

The ARE (6) is a nonlinear matrix equation that does not have a closed-form solution. We establish the model-based iterative solution to (6) in this section in what follows.

A. Algorithm Description

The procedure for obtaining the optimal \( P^* \) is now described. Let \( i \in \{i\} \) and \( j \in \{j\} \) 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 \( \{i, j\} \in \mathbb{N}^+ \). To aid our derivations, let us first define the following matrices:

\[
 \begin{align*}
 A_{K_i} &= A - BK_i, \\
 A_i &= A_{K_i} + \gamma^{-2}DD^TP_{K_i}, \tag{12a} \\
 A^*_i &= A - BK^* + DL^*_K, \tag{12b} \\
 Q_{K_i} &= C^TC + K_i^2RK_i, \tag{12c}
 \end{align*}
\]

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. \( A^*_i \) is the closed-loop system’s transition matrix with arbitrary gains \( K_i \) and \( L_{K_i}^* \); and \( A^* \) is the closed-loop system’s transition matrix with the optimal gains \( K^* \) and \( L^* \). 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 \( j = \{1, 2, \cdots, j\} \), iteratively solve the Riccati equation (6).
  - Update the maximizing player’s gain \( L_{K_i} \) given \( K_i \).
  - Update the minimizing player’s gain \( K_{i+1} \) given \( P_{K_i} \).

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 (10).

IV. MODEL-FREE ITERATIVE ALGORITHM

We are concerned with learning an optimal controller \( K^* \) when the matrices \( (A, B) \) are unknown. For this algorithm, only the trajectory of state \( x \) and control input \( u \) collected along system (1) are required for computing the policies (11). During the data collection phase, suppose the respective control policy is

\[
 u = -\dot{K}x + \sigma w, \quad \xi = -\xi dt + dw \tag{13}
\]

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 \( w \).

The Itô stochastic differential equation (10) admits equal upper and lower values because the generalized ARE for \( P \) admits a positive definite solution [12, Th 4.8 (iii)] given by \( J^* = x^TPx \). Therefore, along the state trajectories of (1), by Itô’s formula [46, Lemma 3.2], the derivative of \( x^TPx \), where \( P \in \mathbb{S}^n \), can be verified to be

\[
 \begin{align*}
 d(x^TPx) &= (dx)^T(Px + x^TPdx + (dx)^TPdx) \\
 &= x^T(A^TP + PA)xdt + 2x^TPBu + 2x^TPDdw + \underbrace{(Ax + Bu)^T(AX + Bu)(dt)^2}_{=0} \\
 &\quad + (2dw)^TDPx + (Ax + Bu)(dt) + (dw)^TDPDdw \\
 &= x^T(A^TP + PA)xdt + 2x^TPBu + 2x^TPDdw + \underbrace{(Tr(D^TPD)dt)}_{=} \\
 &\quad + (2dw)^TDPx + (Ax + Bu)(dt) + (dw)^TDPDdw. \tag{14}
 \end{align*}
\]

Algorithm 1: Model-Based Iterative Algorithm

\begin{algorithm}
\caption{Model-Based Iterative Algorithm}
\begin{algorithmic}[1]
\State Initialize \( K_1 \in \mathcal{K} \) \\
\For {\( i \leq i \)}
\State Set \( L_{K_i} = 0; \)
\State \( Q_{K_i} = C^TP + K_i^2RK_i; \)
\For {\( j \leq j \)}
\State Update \( A^*_i; \)
\State Update \( P_{K_i}; \)
\State Update \( L_{K_{i+1}}; \)
\EndFor
\State \( K_{i+1} = R^{-1}B^TP^*_i; \)
\EndFor
\end{algorithmic}
\end{algorithm}
Vectorization (14), we have
\[
d(\text{vec}(v^T(x))) \text{vec}(P) = \text{vec}(A^T P + PA) dt + 2(x^T \otimes u^T) dt \text{vec}(B^T P) + \text{Tr}(D^T PD) dt + 2x^T P Ddw.
\]
(15)

Suppose that \( \phi(t) = [\text{vec}(v^T(x)), 2(x^T \otimes u^T), 1]^T \). Integrating both sides of (15) on the interval \([0, t_f]\) yields
\[
\frac{1}{t_f} \int_0^{t_f} \phi d(\text{vec}(v^T(x))) \text{vec}(P) = \left[ \frac{\text{vec}(A^T P + PA)}{\text{vec}(B^T P)} \right] \times \text{vec}(D^T PD)) \times \frac{1}{t_f} \int_0^{t_f} \phi dt + \frac{1}{t_f} \int_0^{t_f} 2\phi x^T P Ddw.
\]
(16)

Furthermore, let
\[
\hat{\Phi}(t_f) = \frac{1}{t_f} \int_0^{t_f} \phi(t) \phi(t) dt, \quad \hat{\Xi}(t_f) = \frac{1}{t_f} \int_0^{t_f} \phi d(\text{vec}(v^T(x))).
\]
(17)

Then, by Lemmas A.7 and A.8, the following hold almost surely
\[
\lim_{t_f \to \infty} \frac{1}{t_f} \int_0^{t_f} 2\phi x^T P Ddw = 0, \quad \lim_{t_f \to \infty} \hat{\Phi}(t_f) = \Phi := \Xi(\phi T).
\]
(18)

Combining (16) and (18), we see that there exists a constant matrix \( \Xi \) such that the following holds almost surely
\[
\lim_{t_f \to \infty} \hat{\Xi}(t_f) = \Xi.
\]
(19)

It follows from (16) that
\[
\left[ \frac{\text{vec}(A^T P + PA)}{\text{vec}(B^T P)} \right] = \Phi^{-1} \Xi \text{vec}(P).
\]
(20)

Suppose we let \( n_1 = n(n + 1)/2 \) and \( n_2 = n_1 + mn \), we find
\[
\text{vec}(A^T P + PA) = \text{vec}(B^T P) = [\Phi^{-1}]_{n_1+n_2} \Xi \text{vec}(P).
\]
(21)

Let the \((i, j)\)th iterate of the GARE be
\[
(A^j_{K_i})^T P^j_{K_i} + P^j_{K_i} A^j_{K_i} + Q K_i - \gamma^2 (L^j_{K_i})^T L^j_{K_i} = 0.
\]
(22)

Suppose that \( P^j_{K_i} \) is the positive definite solution at iteration \( j \) for the maximizing player corresponding to the solution at iteration \( i \) for the minimizing player. Equation (22) is recursively solved for \( L^j_{K_i} = \gamma^{-2} D^T P^j_{K_i} \), using the identities in equation (22). At the \( i \)th iteration, in the vectorization style of (22), becomes
\[
\text{vec}(A^T P^j_{K_i} + P^j_{K_i} A) - T^u v^s (I_n \otimes K^T_i) \text{vec}(B^T P^j_{K_i}) - T^u v^s (K^T_i \otimes I_n) T_v \text{vec}(B^T P^j_{K_i}) + T^u v^s (I_n \otimes L^T_{K_i} D^T + L^T_{K_i} D^T \otimes I_n) T_v \text{vec}(P^j_{K_i}) + \text{vec}(Q K_i - \gamma^2 L^j_{K_i} L^j_{K_i}) = 0.
\]
(23)

Putting the terms in (21) in (23), we find that
\[
[\Phi^{-1}]_{n_1+n_2} \Xi \text{vec}(P^j_{K_i}) - T^u v^s [(I_n \otimes K^T_i) + (K^T_i \otimes I_n) T_v] \text{vec}(P^j_{K_i}) + T^u v^s (I_n \otimes L^T_{K_i} D^T + L^T_{K_i} D^T \otimes I_n) T_v \text{vec}(P^j_{K_i}) + \text{vec}(Q K_i - \gamma^2 L^j_{K_i} L^j_{K_i}) = 0.
\]
(24)

Now, let
\[
\Lambda^j_i := [\Phi^{-1}]_{n_1+n_2} \Xi
\]
\[
- T^u v^s [(I_n \otimes K^T_i) + (K^T_i \otimes I_n) T_v] \text{vec}(P^j_{K_i}) + T^u v^s (I_n \otimes L^T_{K_i} D^T + L^T_{K_i} D^T \otimes I_n) T_v \text{vec}(P^j_{K_i}) + \text{vec}(Q K_i - \gamma^2 L^j_{K_i} L^j_{K_i}) = 0.
\]
(25)

Then, the equation to be iterated upon becomes
\[
\text{vec}(P^j_{K_i}) = -[\Lambda^j_i]^{-1} \text{vec}(Q K_i - \gamma^2 L^j_{K_i} L^j_{K_i}).
\]
(26)

Observe: equation (26) consists of unknown variables in the cost \( (P^j_{K_i}) \) and gain matrices \( (K_i, L_{K_i}) \), which can be iteratively computed using Algorithm 1. The other unknown variables from (17) 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 \( H_2/H_{\infty} \) 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.

**Algorithm 2: Learning-based \( H_2/H_{\infty} \) Control**

1. Initialize \( K_1 \in K \) \( \triangleright \) e.g. searching for a valid CLF [47];
2. Collect data from (1) with exploratory input (13); and
3. Construct matrices \( \hat{\Phi}(t_f) \) and \( \hat{\Xi}(t_f) \);
4. for \( i \leq j \) do
5. for \( i \leq j \) do
6. Construct the matrices \( \hat{\Phi}(t_f) \) using (25);
7. Calculate \( \hat{P}^j_{K_i} \) using (26);
8. Update \( \Lambda^j_i = \gamma^{-2} D^T P^j_{K_i} \);
9. Form \( \text{vec}(\hat{B}^T P^j_{K_i}) \) as \( [\hat{\Phi}(t_f)]_{n_1+n_2} \hat{\Xi}(t_f) \text{vec}(\hat{P}^j_{K_i}) \);
10. \( \text{Calculate } K_{i+1} = R^{-1} B^T P^j_{K_i} \);
11. \( \text{end} \)
12. \( \text{end} \)

Next, we analyze the convergence of both loops.

**V. CONVERGENCE ANALYSES**

As seen in Algorithms 1 and 2, the solution \( P_{K_i} \) to the Riccati equation must converge to the unique optimal positive-definite solution \( P^* \) so that the gains \( L^j_{K_i} \) 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].
A. Control Update (Outer) Loop

The control law in the outer-loop of the algorithms seeks to decrease the cost (10) by iterating the following equations until convergence

\[ A^T K_i P_i + P_i A_i + Q_i + \gamma^{-2} P_i D D^T P_i = 0 \] (27a)

\[ K_{i+1} = R^{-1} B^T P_i \quad i = 1, 2, \cdots . \] (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 \( K_i \in K \), the outer-loop iteration has a global linear convergence rate, i.e. there exists \( \alpha \in [0, 1) \), such that

\[ \text{Tr}(P_{K_{i+1}} - P^*) \leq \alpha \text{Tr}(P_k - P^*) \] (28)

**Proof.** The proof to this theorem is provided in Appendix B-B. \( \square \)

**Remark 4.** With Theorem 1, we have \( \| P_{K_i} - P^* \|_F \leq \text{Tr}(P_{K_i} - P^*) \leq \alpha^i \text{Tr}(P_{K_1} - P^*) \).

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.

**Theorem 2.** Given \( K \in K \), 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

\[ \text{Tr}(P_{K_j} - P_{K_{j+1}}) \leq \beta(K) \text{Tr}(P_{K_1} - P_{K_{j+1}}) \] (29)

The proof is established in Appendix B-C.

**Remark 5.** As seen from Theorem B.2, \( P_{K_j} - P_{K_{j+1}} \geq 0 \). From Lemma A.1 and the result of Theorem 2, we have \( \| P_{K_j} - P_{K_{j+1}} \|_F \leq \text{Tr}(P_{K_j} - P_{K_{j+1}}) \leq \beta^{-1}(K) \text{Tr}(P_{K_1}) \), i.e. \( P_{K_j} \) exponentially converges to \( P_{K_1} \) 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 \( \{ P_{K_i} \}_{i=1}^{j=3} \) and \( \{ L_{K_i} \}_{i=1}^{j=2} \) which converge to the worst-case cost matrix and disturbance \( P_{K_i} \) and \( L_{K_i} \) respectively. We require that \( \{ P_{K_i} \}_{i=1}^{j=3} \) and \( \{ L_{K_i} \}_{i=1}^{j=2} \) enter the given neighborhood of \( P_{K_1} \) and \( L_{K_1} \), 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 \( \{ P_{K_i} \} \) and \( \{ L_{K_i} \} \) enter the vicinity of \( P_{K_1} \) and \( L_{K_1} \), irrespective of the different values of \( K_i \).

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

\[ \| P_{K_{i+1}} - P_{K_i} \|_F \leq \epsilon . \] (30)

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 (10) 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

\[ \hat{A}_{K_i} := A - B K_i, \quad \hat{A}_i = A - B \hat{K}_i + \gamma^{-2} D D^T \hat{P}_{K_i}, \] (31a)

\[ \hat{Q}_{K_i} = C^T C + \hat{K}_i^T \hat{R} \hat{K}_i \] (31b)

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

\[ (\hat{A}_{K_i})^T \hat{P}_{K_i} + \hat{P}_{K_i} \hat{A}_{K_i} + \hat{Q}_{K_i} + \gamma^{-2} \hat{P}_{K_i} D D^T \hat{P}_{K_i} = 0, \] (32a)

\[ \hat{K}_{i+1} = R^{-1} B^T \hat{P}_{K_i} + \Delta \hat{K}_i \] (32b)

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

**Theorem 4.** There exists an \( \ell > 0 \), \( \alpha \in [0, 1) \), and \( \kappa(\cdot) \in K_{\infty} \), such that \( \| \hat{P}_{K_i} - P^* \|_F \leq \alpha^{j-1} \text{Tr}(\hat{P}_{K_1} - P^*) + \kappa(\| \Delta \hat{K} \|_{\infty}) \), as long as \( \| \Delta \hat{K} \|_{\infty} \leq \ell \).

**Proof.** The proof is provided in Appendix C-B. \( \square \)

**Remark 6.** That is, as iteration goes to infinity, \( \hat{P}_{K_i} \) approaches the optimal value \( P^* \), entering its neighborhood. The radius of the neighborhood is proportional to \( \| \Delta \hat{K} \|_{\infty} \). 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 (22), and the control policy \( L_{K_i} \) can be updated precisely. In reality, due to the influence of disturbance and noise, \( L_{K_i} \) may be updated inaccurately. Therefore, the inexact inner-loop iteration is

\[ (\hat{A}_{K_i})^T \hat{P}_{K_i} + \hat{P}_{K_i} \hat{A}_{K_i} + \hat{Q}_{K_i} - \gamma^2 \hat{L}_{K_i} \hat{L}_{K_i} = 0, \] (33a)

\[ \hat{L}_{K_i} = \gamma^{-2} D^T \hat{P}_{K_i} + \Delta \hat{L}_{K_i} \] (33b)
where \( \{ \tilde{L}^j_K \}_{j=1}^\infty \) and \( \{ \tilde{P}^j_K \}_{j=1}^\infty \) are sequences generated by the inexact inner-loop iteration (33).

**Theorem 5.** Assume \( \| \Delta L_K \|_\infty < \epsilon \) for all \( j \in \mathbb{N}_+ \). There exists \( \beta(K) \in [0, 1) \), and \( \lambda(\cdot) \in K_{\infty} \), such that

\[
\| \hat{P}^j_K - P_K \|_F \leq \beta^{-1}(K) \text{Tr}(P_K) + \lambda(\| \Delta L \|_\infty). \tag{34}
\]

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 \( \theta_1, \theta_2, \) and \( \theta_3 \) 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 \( D = [0_{3 \times 3}, I_3]^T \).

Furthermore, we bound the system’s \( H_\infty \) norm by \( \gamma = 5 \) from above. The initial state is set as

\[
x(0) = [0, -5, 10, 10, -10, 10]^T. \tag{35}
\]

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

\[
C = [J_6, 0_{2 \times 6}]^T, \quad E = [0_{6 \times 2}, I_2]^T. \tag{36}
\]

**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 \( u^*(t) = -K^*x(t) \) and the worst-case disturbance \( w^*(t) = L^*x(t) \). We choose the LQG cost function as,

\[
J_{LQG} = \int_{0}^{\infty} x^TC^TCx + u^TE^TEudt, \tag{37}
\]

and we find the LQG feedback gain as

\[
K_{LQG} = \begin{bmatrix}
-26.77 & -8.755 & -4.20 & -9.033 & -3.05 & -2.30 \\
-65.10 & -23.79 & -8.24 & -21.60 & -9.27 & -3.93
\end{bmatrix} \tag{38}
\]

While executing Algorithm 1, the numbers of iterations for outer and inner loops are heuristically chosen as \( i = 20 \) and \( j = 20 \) 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 \( H_\infty \) norms of closed-loop system with LQG and mixed \( H_2/H_\infty \) controllers are respectively 8.72 and 4.99. Thus, this algorithm generates an optimal controller and guarantees the \( H_\infty \) norm system constraint, while the LQG controller does violate the constraint.

Carrying along with \( x(0) \), we compare the mixed \( H_2/H_\infty \) 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 \( H_2/H_\infty \) 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 \( H_2/H_\infty \) controller is within \([-5, +5]\) while that of the LQG controller is within \([-10, +10]\). Thus, for the Brownian motion, the mixed \( H_2/H_\infty \) controller does suppress the chattering.

Under a worst-case disturbance, the results LQG and for mixed \( H_2/H_\infty \) controller are shown in Figures 2 and 5 respectively. The LQG controller fails to satisfy the \( H_\infty \) 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.

**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, \( \Delta K \). At each iteration of the algorithm, a \( \Delta K \) 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, $\|P_6 - P^*\|_F$, is 2.9%, while the gain error, $\|K_{20} - K^*\|_F$, 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.

| Computational time (sec) |
|--------------------------|
| Alg. 1 Alg. 2 NPG Alg. 1 Alg. 2 NPG |
| Double Inverted Pendulum | 0.0901 0.3064 2.1649 | 0.1455 0.7829 2.3209 |
| Triple Inverted Pendulum | | |

Fig. 5: With our mixed $H_2/H_\infty$ policy, the evolution of the joint angles and velocities under a worst-case disturbance.

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

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.

C. Results on Learning-based Control

For the parameters of Algorithm 2, we set $i = 20$ and $j = 30$, and collected data for $t_f = 1500s$. The parameters of the $A$ and $B$ matrices are unknown but the initial controller $K_1 \in K$ 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 $K_i$ at each iteration converges after 5 iterations. The corresponding evaluative matrix $P_{K_f}$ also converges. At 20th iteration, the relative error of $\|K_{20} - K^*\|/\|K^*\| = 31.5\%$ and $\|\hat{P}_{K_{20}} - P_\star\|/\|P_\star\| = 31.6\%$. 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.

IX. ACKNOWLEDGMENTS

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

APPENDIX A

A Catalog of Preliminary Lemmas

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

Lemma A.1. For any symmetric and positive semi-definite matrix $P \in S^n$, $\|P\| \leq \text{Tr}(P)$, $\|P\| \leq \text{Tr}(P)$, and $\text{Tr}(P) \leq n\|P\|$. For $x \in \mathbb{R}^n$, $x^T P x \geq \sigma(P) \|x\|^2$.

Proof. Let $\sigma_1 \geq \cdots \geq \sigma_n$ be ordered singular values of $P$. Since $P$ is symmetric and positive semi-definite, its singular
values are equal to its eigenvalues. Then, \( \| P \|^F = \sqrt{\sum_{i=1}^n \sigma_i^2} \), \( \text{Tr}(P) = \sum_{i=1}^n \sigma_i \), and \( \| P \|^2 = \sigma_1^2 \). Since \( \sum_{i=1}^n \sigma_i^2 \leq (\sum_{i=1}^n \sigma_i)^2 \), we have \( \| P \|^F \leq \text{Tr}(P) \), \( \| P \| \leq \text{Tr}(P) \), and \( \text{Tr}(P) \leq n \| P \| \). Using Rayleigh’s theorem [56, Theorem 4.2.2], we have \( x^T P x \geq \mathcal{g}(P) \| x \|^2 \).

\[ \tag{A.2} \]

**Lemma A.2.** For \( X \in \mathbb{R}^{m \times n} \) and \( Y \in \mathbb{R}^{p \times n} \), \( \| XY \|^F \leq \| X \| \| Y \| \).

**Proof.** Let \( Y = [y_1, \ldots, y_p] \), then it follows that \( XY = [Xy_1, \ldots, Xy_p] \). This implies that \( \| XY \|^2 = \sum_{i=1}^p \| Xy_i \|^2 \).

Furthermore, as the spectral norm is defined by \( \| X \| = \max_{x \neq 0} \frac{\| Xx \|}{\| x \|} \), we have \( \| Xy_i \|^2 \leq \| X \|^2 \| y_i \|^2 \). Hence, \( \| XY \|^2 \leq \| X \|^2 \sum_{i=1}^p \| y_i \|^2 = \| X \| \| Y \| \), 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. \( P = \int_{t=0}^\infty e^{(A^T + A) t} Q e^{(A^T + A) t} dt \);
2. \( P > 0 \) if \( Q > 0 \), and \( P \geq 0 \) if \( Q \geq 0 \);
3. \( P \geq 0 \) when \( Q \geq 0 \) and \( Q, A \) is observable if \( P > 0 \);
4. \( P \) satisfies \( A^T P + P A + Q = 0 \), \( Q \leq 0 \), then \( P \leq P \).

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

\[ \tag{A.1} \]

Since \( Q \leq 0 \), \( P \leq P \).

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

1. \( A \) is Hurwitz if \( P > 0 \) and \( Q > 0 \).
2. \( A \) is Hurwitz if \( P \geq 0 \), \( Q \geq 0 \) 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 \( Y = Y^T \).

Then,

\[ \mu_2(-X) \text{Tr}(Y) \leq \text{Tr}(XY) \leq \mu_2(X) \text{Tr}(Y) \]

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 \( U: S \rightarrow \mathbb{R}^{m \times r} \) and \( V: S \rightarrow \mathbb{R}^{p \times q} \) 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

\[ \frac{\partial \text{vec}(UV)}{\partial \text{vec}(X)} = (V^T \otimes I_m) \frac{\partial \text{vec}(U)}{\partial \text{vec}(X)} + (I_p \otimes U) \frac{\partial \text{vec}(V)}{\partial \text{vec}(X)} \].

**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 \( \{T_t \mid t \geq 0\} \) be a measurable semi-group of transformations preserving \( (\Omega, \mathcal{F}, P) \). Suppose that the time average operator \( A_t \) is so defined

\[ A_t f(w) = \frac{1}{t} \int_0^t f(T_s(w)) ds \].

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

**Proof.** This Lemma is proven in [59, Theorem 16.14] and [60, Theorem 1.5.9].

**Lemma A.8.** Let \( a(x(t)) \) be a vector function such that

\[ \mathbb{E}[a'(x(t)) a(x(t))] < \infty \]

where \( \{x(t) \mid t \geq 0\} \) is the solution of process \( dx(t) = \mu(x(t)) dt + dw \), and \( w \) is a standard independent Brownian motion. It follows that

\[ \lim_{t \rightarrow \infty} T_f = 0 \]

\[ \mathbb{E} \left[ \int_0^t a(x(t)) dw \right] = 0. \]

**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:

1. \( \| T(K) \|_{\infty} \leq \gamma \);
2. The Riccati equation

\[ (A - BK)^T P_K + P_K (A - BK) + CT C \]

\[ + K^T R K + \gamma^{-2} P_K DD^T P_K = 0 \]

admits a unique stabilizing solution with a Hurwitz \( A - BK + \gamma^{-2} DD^T P_K \);
3. There exists some \( P_K > 0 \) such that

\[ (A - BK)^T P_K + P_K (A - BK) + CT C \]

\[ + K^T R K + \gamma^{-2} P_K DD^T P_K \leq 0. \]

**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 \( K_i \in K \), then for any \( i \geq 1 \)

1. \( K_i \in K \);
2. \( P_{K_i} \geq P_{K_{i+1}} \geq \cdots \geq P^* \), for \( K_i \geq K_{i+1} \);
3. \( \lim_{i \rightarrow \infty} \| K_i - K^* \|_{F} = 0 \), \( \lim_{i \rightarrow \infty} \| P_{K_i} - P^* \|_{F} = 0 \).

We here establish the proof to convergence of \( K_i \) per iteration \( i \) to the optimal gain \( K^* \).

**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 \( i = 1 \), \( K_1 \in K \), and it satisfies 1). When \( i > 1 \), assume \( K_{i-1} \in K \). Therefore, by the second condition of Lemma A.9, \( P_{K_i} > 0 \) is the unique stabilizing solution to (27a). Now consider the identities,

\[ (A - BK_i)^T = (A - BK_{i+1})^T P_{K_i} + (K_{i+1} - K_i)^T B^T P_{K_i} \]

\[ P_{K_i} (A - BK_i) = P_{K_i} (A - BK_{i+1}) + P_{K_i} B (K_{i+1} - K_i) \].

(B.1)
It follows that equation (27a) can be rewritten as
\[ A_{i+1}^T P_{K_{i+1}} + P_{K_{i+1}} + C^T C + \gamma^{-2} P_{K_{i+1}} DD^T P_{K_{i+1}} \]
+ \[ K_{i+1}^T RK_{i+1} + (K_{i+1} - K)^T R (K_{i+1} - K) = 0. \]
(B.2)

As \((K_{i+1} - K)^T R (K_{i+1} - K) \geq 0\), following the third condition of Lemma A.9, we find that \(K_{i+1} \in K\). A fortiori, we establish that \(K_i \in K\). 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
\[ A_{i+1}^T (P_{K_i} - P_{K_{i+1}}) + (P_{K_{i+1}} - P_{K_{i+1}}) A_{i+1} \]
+ \[ (K_{i+1} - K_i)^T R (K_{i+1} - K_i) \]
+ \[ \gamma^{-2} (P_{K_i} - P_{K_{i+1}}) DD^T (P_{K_i} - P_{K_{i+1}}) = 0. \]
(B.3)

Given statement 1), it follows that \(K_{i+1} \in K\). Furthermore, by the second condition of Lemma A.9, \(A_{i+1} \in H\). As \((K_i - K_{i+1})^T R (K_i - K_{i+1}) \geq 0\) and \(\gamma^{-2} (P_{K_i} - P_{K_{i+1}}) DD^T (P_{K_i} - P_{K_{i+1}}) > 0\), by Lemma A.3, we find that \(P_{K_i} - P_{K_{i+1}} \geq 0\). Because \(K_{i+1} \in K\), \(A_{i+1} \in H\) by Lemma A.9. Hence, \(P_{K_i} \geq P_{K_{i+1}}\) i.e., the sequence \(\{P_{K_i}\}_{i=1}^\infty\) is decreasing, and lower bounded by 0. A fortiori, \(\{P_{K_i}\}_{i=1}^\infty\) converges to \(P_{K}\), which also satisfies (6). Due to the uniqueness of the solution to (6), \(P_{K} = P^*\).

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 \(\text{Tr}(P_{K_{i+1}} - P^*) \leq a \left[\text{Tr}(P_{K_i} - P^*)\right]^2\).

**Proof.** For the \((i + 1)\)th iteration, (27a) can be rewritten as
\[ A_{i+1}^T P_{K_{i+1}} + P_{K_{i+1}} A_{i+1} + C^T C + K_{i+1}^T RK_{i+1} \]
\[ - \gamma^{-2} P_{K_{i+1}} DD^T P_{K_{i+1}} = 0. \]
(B.4)

Also, (6) can be rewritten as
\[ A_{i+1}^T P^* + P^* A_{i+1} + C^T C + K^{*T} RK^* \]
\[ + K_{i+1}^T RK_{i+1} - \gamma^{-2} P_{K_{i+1}} DD^T P^* \]
\[ - \gamma^{-2} P^* DD^T P_{K_{i+1}} + \gamma^{-2} P^* DD^T P^* = 0. \]
(B.5)

Subtracting (B.5) from (B.4), and completing squares, we have
\[ A_{i+1}^T (P_{K_{i+1}} - P^*) + (P_{K_{i+1}} - P^*) A_{i+1} \]
\[ + (K_{i+1} - K^*)^T R (K_{i+1} - K^*) \]
\[ - \gamma^{-2} (P_{K_{i+1}} - P^*) DD^T (P_{K_{i+1}} - P^*) = 0. \]
(B.6)

From Theorem B.1 and Lemma A.9, we see that \(A_{i+1} \in H\). From Lemma A.3, it follows that \(P_{K_{i+1}} - P^*\)
\[ \leq \int_0^\infty e^{\lambda_{i+1} t} (P_{K_i} - P^*) BR^{-1} B^T (P_{K_i} - P^*) e^{\lambda_{i+1} t} dt. \]
(B.7)

Using the cyclic property of matrix trace,
\[ \text{Tr}(P_{K_{i+1}} - P^*) \leq \]
\[ \text{Tr} \left[(P_{K_i} - P^*) BR^{-1} B^T (P_{K_i} - P^*) \int_0^\infty e^{\lambda_{i+1} t} e^{\lambda_{i+1} t} dt \right]. \]
(B.8)

Let us define \(M_{i+1} := \int_0^\infty e^{\lambda_{i+1} t} e^{\lambda_{i+1} t} dt. \) Because \(A_{i+1} \) is Hurwitz, \(M_{i+1}\) satisfies
\[ A_{i+1} M_{i+1} + M_{i+1} A_{i+1}^T + I_n = 0. \]
(B.9)

From Theorem B.1, it follows that as \(i \to \infty\), \(P_{K_i}\) converges to \(P^*\) and \(A_i\) converges to \(A\). Consequently, \(M_i\) converges to \(M^* \in \mathbb{S}^n\), which is the solution to
\[ A^* M^* + M^* (A^*)^T + I_n = 0. \]
(B.10)

Thus, \(\lim_{i \to \infty} \|M_i\| = \|M^*\|\) and \(\hat{m} := \sup_{i \in \mathbb{N}} \|M_i\| < \infty\). As a consequence,
\[ \text{Tr}(P_{K_{i+1}} - P^*) \leq \text{Tr} \left[(P_{K_i} - P^*) BR^{-1} B^T (P_{K_i} - P^*) M_i \right] \]
\[ \text{Tr}(P_{K_{i+1}} - P^*) \leq \hat{m} \|BR^{-1} B^T\| \|\text{Tr}(P_{K_i} - P^*)\|^2. \]
(B.11)

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

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

**Proof.** From Proposition 3, for any \(\epsilon > 0\), there exists \(i \in \mathbb{N}_+,\) such that if \(i \geq i^*\),
\[ \text{Tr}(P_{K_{i-1}} - P^*) \leq \frac{1}{a(1 + \epsilon)}, \]
(B.12)

where \(a > 0\) is as given in Proposition 3. We have by Proposition 3 for any \(i \geq i^*\) that
\[ \frac{\text{Tr}(P_{K_{i-1}} - P_{K_i})}{\text{Tr}(P_{K_i} - P^*)} = \frac{\text{Tr}(P_{K_{i-1}} - P^*)}{\text{Tr}(P_{K_i} - P^*)} - 1 \]
\[ \geq \frac{1}{a \text{Tr}(P_{K_{i-1}} - P^*)} - 1 \geq \epsilon. \]
(B.13a)

Now, for \(i < i^*\), from Theorem B.1 we have
\[ 0 < \text{Tr}(P_{K_i} - P^*) \leq \text{Tr}(P_{K_i} - P^*). \]
(B.14)

Suppose that we let \(b_1 = \min_{i < i^*} \text{Tr}(P_{K_{i-1}} - P_{K_i})\). Then, by Theorem B.1, \(b_1 > 0\) so that
\[ \frac{\text{Tr}(P_{K_{i-1}} - P_{K_i})}{\text{Tr}(P_{K_i} - P^*)} \geq \frac{b_1}{\text{Tr}(P_{K_i} - P^*)}. \]
(B.15)

Let
\[ b := \max \left[1, \frac{\text{Tr}(P_{K_{i-1}} - P_{K_i})}{\text{Tr}(P_{K_i} - P^*)} \right]. \]
(B.16)

We see that \(\text{Tr}(P_{K_i} - P^*) \leq b \text{Tr}(P_{K_{i-1}} - P_{K_i})\) for all \(i \in \mathbb{N}_+\).

From Lemma A.1, we have
\[ \|K_{i+1} - K^*\| \leq \|R^{-1} B^T\| \|P_{K_i} - P^*\| \leq \|R^{-1} B^T\| \|P_{K_i} - P^*\| \leq b \|R^{-1} B^T\| \|\text{Tr}(P_{K_{i-1}} - P_{K_i})\| \leq b \|R^{-1} B^T\| \|P_{K_{i-1}} - P_{K_i}\|\].
(B.17)

Furthermore,
\[ \|K_i - K_{i+1}\| \geq \|R^{-1} B^T\| \|P_{K_i} - P_{K_{i-1}}\| \geq \|R^{-1} B^T\| \|P_{K_i} - P_{K_{i-1}}\|. \]
(B.18)
As $B$ is full column rank, $\varrho(R^{-1}B^T) > 0$. Setting $b' = \varrho(R^{-1}B^T)/\varrho(R^{-1}B^T)$, we establish the second statement. $\square$

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

$$(K_i - K^*)^T R(K_i - K^*) \geq 
\gamma^{-2}(P_{K_i} - P^*) D D^T (P_{K_i} - P^*). \tag{B.19}$$

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

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

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

**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

$$\text{Tr}(P_{K_i} - P_i) \leq c \|E_{K_i}\|, \text{ where,}
\begin{align*}
\gamma = \varrho(R) \quad &\in (0, 1) \quad \text{and} \quad E_{K_i} = (K_i - K_{i+1})^T R(K_i - K_{i+1}), \tag{B.21}\n\end{align*}$$

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

$$A^T P_{K_i} + P_{K_i} A^* + (K^* - K_i)^T R K_{i+1} 
+ K_{i+1}^T R(K^* - K_i) + \gamma^{-2}(P_{K_i} - P^*) D D^T P_{K_i} 
+ \gamma^{-2}P_{K_i} D D^T (P_{K_i} - P^*) + CT C + K_i^T R K_i 
- \gamma^{-2}P_{K_i} D D^T P_{K_i} = 0, \tag{B.22}$$

and (6) can be rewritten as

$$A^T P^* + P^* A^* + CT C + K^*^T R K^* 
- \gamma^{-2}P^* D D^T P^* = 0. \tag{B.23}$$

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

$$A^T (P_{K_i} - P^*) + (P_{K_i} - P^*) A^* + E_{K_i} 
+ \gamma^{-2}(P_{K_i} - P^*) D D^T (P_{K_i} - P^*) 
- (K_{i+1} - K^*)^T R(K_{i+1} - K^*) = 0. \tag{B.24}$$

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

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

Now, using Lemma A.3, we have

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

$$\leq \text{Tr}\left\{ \int_0^\infty e^{A^T t} e^{A^* t} dt \right\} \tag{B.26b}$$

$$\leq \|E_{K_i}\| + 2 \|K_i - K_{i+1}\| \|R\| \|K_i - K_{i+1}\| \tag{B.26c}$$

$$\text{Tr}\left\{ \int_0^\infty e^{A^T t} e^{A^* t} dt \right\} \tag{B.26d}$$

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.

$$\text{Tr}(P_{K_i} - P^*) \leq \left(2 + 2b'\|R\| \right) \text{Tr}\left( \int_0^\infty e^{A^T t} e^{A^* t} dt \right) \|E_{K_i}\|. \tag{B.27}$$

**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 $F := \int_0^\infty e^{W t} E e^{W^T t} dt$, and $d(W) = \log(5/4)/\|W\|$. Then, $\|F\| \geq \frac{1}{2} d(W) \|E\|$.  

**Proof.** A Taylor expansion of $e^{W t}$ yields

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

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,

$$x_0^T F x_0 \geq \int_0^{d(W)} x_0^T e^{W^T t} E e^{W t} x_0 dt 
= \int_0^{d(W)} x_0^T (I_n + S(t)) E (I_n + S(t)) x_0 dt 
\geq \int_0^{d(W)} \|E\|\|x_0\|^2 - 2\|S(t)\|\|E\|\|x_0\|^2 dt 
\geq \frac{1}{2} d(W) \|E\|\|x_0\|^2.$$

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

$$(P_{K_i} - P_{K_{i+1}}) \geq \int_0^\infty e^{A_{i+1}^T t} E_{K_i} e^{A_i t} dt =: F_{K_i}, \tag{B.29}$$

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

$$\|A_{i+1}\| \leq \|A\| + (\|BR^{-1} B^T\| + \gamma^{-2}\|DD^T\|)\|P_{K_i}\|. \tag{B.30}$$
Let us set
\[ d = \frac{\log(5/4)}{\|A\| + (\|BR^{-1}B^T\| + \gamma^{-2}\|DD^T\|)}\|P_{K_i}\| \quad \text{(B.31)} \]

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
\[ \text{Tr}(P_{K_{i+1}} - P^*) \leq \text{Tr}(P_{K_i} - P^*) - \frac{d}{2c}\|E_{K_i}\| \quad \text{(B.32a)} \]
\[ \leq \text{Tr}(P_{K_i} - P^*) - \frac{d}{2c}\text{Tr}(P_{K_i} - P^*) \quad \text{(B.32b)} \]
\[ \leq \left(1 - \frac{d}{2c}\right)\text{Tr}(P_{K_i} - P^*) \quad \text{(B.32c)} \]

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

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 \( L^{1}_{K_i} = 0 \), then for any \( i, j \in \mathbb{N}_+ \), the following holds

1) \( A^i_{K_i} \) is Hurwitz,
2) \( P_{K_i} \geq \ldots \geq P_{K_{i+1}} \geq \ldots \geq P_{K_1} \),
3) \( \lim_{i \to \infty} \|P_{K_i} - P_{K_i}\|_F = 0 \), 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 \( P_{K_0} > 0 \) to (27a) and \( A_i \) is Hurwitz. When \( j = 1, L^{1}_{K_i} = 0 \), then (22) can be rewritten as
\[ A^T_{K_i}P_{K_i} + P^{1}_{K_i}A_{K_i} + C^T_{i}C + K^T_{i}RK_i = 0. \quad \text{(B.33)} \]

Since \( A_{K_i} \) is Hurwitz and \((C,A)\) is observable, by Lemma A.3, \( A^1_{K_i} = A_{K_i} \), is Hurwitz, and \( P_{K_i} > 0 \). Subtracting (22) from (27a) yields
\[ (A^i_{K_i})^T(P_{K_i} - P^{1}_{K_i}) + (P_{K_i} - P^{1}_{K_i})A^i_{K_i} \]
\[ + \gamma^2(L^{i}_{K_i} - L_{K_i})^T(L^{i}_{K_i} - L_{K_i}) = 0. \quad \text{(B.34)} \]

When \( j = 1 \), since \( A_{K_i} \) is Hurwitz and \( \gamma^2(L^{1}_{K_i} - L_{K_i})^T(L^{1}_{K_i} - L_{K_i}) \geq 0 \), according to Lemma A.3, (B.34) results in \( P_{K_i} > 0 \). We can rewrite (27a) as
\[ A^T_{K_i}P_{K_i} + P^{1}_{K_i}A_{K_i} + Q_{K_i} - \gamma^{-2}P_{K_i}DD^T_{K_i} = 0. \quad \text{(B.35)} \]

Since \( A_i \) is Hurwitz and \( P_{K_i} > 0 \), by Lemma A.3, the following inequality holds
\[ Q_{K_i} - \gamma^{-2}P_{K_i}DD^T_{K_i} > 0. \quad \text{(B.36)} \]

Assume \( A^1_{K_i} \) is Hurwitz, and from (B.34), we have \( P_{K_i} > P^{1}_{K_i} \). Following (27a), we have
\[ (A^{i+1}_{K_i})^T P_{K_i} + P_{K_i} A^{i+1}_{K_i} + Q_{K_i} - \gamma^2 P_{K_i} DD^T_{K_i} \]
\[ + \gamma^2(L^{i+1}_{K_i} - L_{K_i})^T(L^{i+1}_{K_i} - L_{K_i}) = 0. \quad \text{(B.37)} \]

By (B.36) and \( P_{K_i} \geq P^{1}_{K_i} \),
\[ Q_{K_i} - \gamma^2 P^{1}_{K_i} DD^T_{K_i} P_{K_i} = 0. \quad \text{(B.38)} \]

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

Rewriting (22) for the \((j+1)^{th}\) iteration and subtracting the resulting equation from (22) results in
\[ (A^{i+1}_{K_i})^T(P^{i+1}_{K_i} - P^1_{K_i}) + (P^{i+1}_{K_i} - P^1_{K_i})A^{i+1}_{K_i} \]
\[ + \gamma^2(L^{i+1}_{K_i} - L^1_{K_i})^T(L^{i+1}_{K_i} - L^1_{K_i}) = 0. \quad \text{(B.39)} \]

As \( \gamma^2(L^{i+1}_{K_i} - L^1_{K_i})^T(L^{i+1}_{K_i} - L^1_{K_i}) \geq 0 \) and \( A^{i+1}_{K_i} \) is Hurwitz, by (B.39) and Lemma A.3, we have \( P^{i+1}_{K_i} \geq P^i_{K_i} \). As a result, the proof of statement 2) is completed.

Statement 2) implies that the sequence \( \{P^j_{K_i}\}_{j=1}^{\infty} \) is monotonically increasing and upper-bounded by \( P_{K_i} \). Hence, \( P^\infty_{K_i} \) exists, and is the solution to (27a). Due to the uniqueness of the solution, we have \( P^\infty_{K_i} = P_{K_i} \), 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
\[ E^i_K := \gamma^{-2}(D^TP^{i}_{K_i} - \gamma^2 L^1_{K_i})^T(D^TP^i_{K} - \gamma^2 L^1_{K_i}), \quad \text{(B.40)} \]

and \( K \in \mathcal{K} \). Then, for the sequences \( \{P^j_{K_i}\}_{j=1}^{\infty} \) and \( \{L^j_{K_i}\}_{j=1}^{\infty} \) obtained by the inner-loop iteration (22), the following inequality holds
\[ \text{Tr}(P_{K_i} - P^j_{K}) \leq \|E^j_K\|_c(K), \quad \text{(B.41)} \]

where
\[ c(K) = \text{Tr}\left(\int_0^\infty e^{(A_{K} + DL_K)t}e^{(A_{K} + DL_K)^T}dt\right). \quad \text{(B.42)} \]

**Proof.** Subtracting (22) from (27a) and completing squares, we have
\[ (A_{K} + DL_K)^T(P_{K_i} - P^j_{K_i}) + (P_{K_i} - P^j_{K_i})(A_{K} + DL_K) \]
\[ - \gamma^{-2}(D^TP_{K_i} - D^TP^j_{K})(D^TP^j_{K} - D^TP_{K}) \]
\[ + \gamma^{-2}(D^TP^j_{K} - \gamma^2 L^1_{K_i})^T(D^TP^j_{K} - \gamma^2 L^1_{K_i}) = 0. \quad \text{(B.43)} \]

Because \( K \in \mathcal{K}, A_{K} + DL_K \) is Hurwitz by Lemma A.9. Using Lemma A.3, we have
\[ P_{K_i} - P^j_{K} \leq \int_0^\infty e^{(A_{K} + DL_K)t}E^j_K e^{(A_{K} + DL_K)^T}dt. \]

By Lemma A.5 and the cyclic property of the trace, we have
\[ \text{Tr}(P_{K_i} - P^j_{K}) \leq \|E^j_K\| \text{Tr}\left(\int_0^\infty e^{(A_{K} + DL_K)t}e^{(A_{K} + DL_K)^T}dt\right). \]

Therefore, \( \text{Tr}(P_{K_i} - P^j_{K}) \leq \|E^j_K\|_c(K) \) holds. \( \square \)
The proof of the theorem given in Theorem 2 is now given.

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

$$P^{j+1}_K - P^j_K = \int_0^\infty e^{(A^{j+1}_K)^T t} E_K e^{A^{j+1}_K t} dt.$$  \hspace{1cm} (B.45)

Therefore,

$$P_K - P^{j+1}_K = P_K - P^j_K - P^j_K.$$ \hspace{1cm} (B.46)

By Theorem B.2, $P_K \geq P^j_K$ so that

$$\|A^{j+1}_K\| = \|A - BK + \gamma^{-2}DD^T P^j_K\|$$
$$\leq \|A - BK\| + \gamma^{-2}\|DD^T\|\|P_K\|.$$ \hspace{1cm} (B.47)

Define

$$d(K) := \log(5/4) \frac{\log(5/4)}{\|A - BK\| + \gamma^{-2}\|DD^T\|\|P_K\|}.$$ \hspace{1cm} (B.48)

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

$$\text{Tr}(P_K - P^{j+1}_K) = \text{Tr}(P_K - P^j_K) - \text{Tr}(F^j_K) \leq \text{Tr}(P_K - P^j_K) - \|F^j_K\|$$
$$\leq \text{Tr}(P_K - P^j_K) - \frac{1}{2}d(K)E_K^2.$$ \hspace{1cm} (B.49a)

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

$$\text{Tr}(P_K - P^{j+1}_K) \leq \text{Tr}(P_K - P^j_K) - \frac{1}{2}d(K) \frac{\log(5/4)}{c(K)} \text{Tr}(P_K - P^j_K)$$
$$\leq \left(1 - \frac{1}{2}d(K) \frac{\log(5/4)}{c(K)} \right) \text{Tr}(P_K - P^j_K).$$ \hspace{1cm} (B.50)

Equation (B.50) is equivalent to (29), and gives us the required result. \hspace{1cm} \Box

**D. Proof: Uniform Convergence**

We now establish the proof of Theorem 3.

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

$$A^T M_i + M_i A_i + I_n = 0.$$ \hspace{1cm} (B.51)

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

Hence, we have

$$c(K_i) = \text{Tr}(M_i) \leq \bar{c},$$ \hspace{1cm} (B.52)

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

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

Hence,

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

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

$$\|P^{j+1}_K - P^j_K\| \leq \bar{\beta} \text{Tr}(P_{K_i}) \leq \bar{\beta}^{j-1} \text{Tr}(P_{K_i}).$$ \hspace{1cm} (B.55)

We see that for any $i \in \mathbb{N}_+$ and $\epsilon > 0$, there exists $\bar{j} > 0$, such that if $j \geq \bar{j}$, $\|P^{\bar{j}}_K - P^j_K\| \leq \epsilon$. A fortiori, the iteration in fact converges uniformly to an $\epsilon > 0$. \hspace{1cm} \Box

**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 $\tilde{A}_K^j := A - BK + D\tilde{L}_K^j$ denote the two-player transition matrix, and $\Delta L_K^j$ 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 $K \in K$, there exists an $l(K) > 0$ such that for a perturbation of $K$ by $\Delta K$, $K + \Delta K \in K$, as long as $\Delta K \|I\| < l(K)$.

**Proof.** Let us introduce the perturbations $\Delta P \in S^n$ 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 $\Delta P$ and $\Delta K$ as follows

$$F(\Delta P, \Delta K) = (A_K + \gamma^{-2}DD^T P_K)T \Delta P$$
$$+ \Delta P(A_K + \gamma^{-2}DD^T P_K) - \Delta K^T RB + K^T R \Delta K$$
$$+ \Delta K^T R \Delta K + \gamma^{-2}\Delta PDD^T \Delta P.$$ \hspace{1cm} (C.1)

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

$$F(\text{vec}(\Delta P), \Delta K) = [I_n \otimes (A_K + \gamma^{-2}DD^T P_K)]T$$
$$+ (A_K + \gamma^{-2}DD^T P_K)T \otimes I_n \text{vec}(\Delta P)$$
$$- (P_K B \otimes I_n) \text{vec}(\Delta K) - (I_n \otimes P_K B) \text{vec}(\Delta K)$$
$$- (I_n \otimes \Delta K^T) - (I_n \otimes \Delta K^T) \text{vec}(\Delta P)$$
$$+ (K^T R \otimes I_n) \text{vec}(\Delta K) + (I_n \otimes K^T R) \text{vec}(\Delta K)$$
$$+ \text{vec}(\Delta K^T R \Delta K) + \gamma^{-2}\text{vec}(\Delta PDD^T \Delta P).$$ \hspace{1cm} (C.2)

Using Lemma A.6, we have

$$\frac{\partial \text{vec}(\Delta PDD^T \Delta P)}{\partial \text{vec}(\Delta P)} = (\Delta P \otimes I_n) \frac{\partial \text{vec}(\Delta PDD^T)}{\partial \text{vec}(\Delta P)}$$
$$+ (I_n \otimes \Delta PDD^T) \frac{\partial \text{vec}(\Delta P)}{\partial \text{vec}(\Delta P)}$$
$$= (\Delta P \otimes I_n)(DD^T \otimes I_n) + I_n \otimes \Delta PDD^T$$
$$= \Delta PDD^T \otimes I_n + I_n \otimes \Delta PDD^T.$$ \hspace{1cm} (C.3)
Therefore,
\[
\frac{\partial F}{\partial \text{vec}(\Delta P)}(\text{vec}(\Delta P), \text{vec}(\Delta K))
= I_n \otimes [\left(A_K + \gamma^{-2}DD^T P_K\right) - B\Delta K]^T
+ [(A_K + \gamma^{-2}DD^T P_K) - B\Delta K]^T \otimes I_n
+ \Delta PDD^T \otimes I_n + I_n \otimes \Delta PDD^T.
\]  
(C.4)

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

As \(\Delta P\) and \(\Delta K\) satisfy \(F(\Delta P, \Delta K) = 0\), we have
\[
(A - BK - B\Delta K)^T (P_K + \Delta P) + (P_K + \Delta P) + CT C
+ (K + \Delta K)^T R (K + \Delta K)
+ \gamma^{-2}(P_K + \Delta P)DD^T (P_K + \Delta P) = 0.
\]  
(C.5)

As \(P_K + \Delta P \succeq 0\) when \(\|\Delta K\| \leq l(K)\), 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 \(\Delta K\), it is still robust and stabilizing. Let us now show that the inexact outer loop does converge to the optimal gain \(K^*\) despite perturbations.

**Proof of Theorem 4.** Assume \(\bar{K}_i \in \mathcal{K}\) and \(\|\Delta K_{i+1}\|_F \leq l(\bar{K}_{i+1})\). By Lemma C.1, \(\bar{K}_{i+1} \in \mathcal{K}\). Rewriting (32a) for the \((i+1)\)th iteration and subtracting (33a) from it yields
\[
\dot{\Delta}_{i+1} = (\hat{P}_K - \hat{P}_{K,i})
+ (R\hat{K}_i - B^T \hat{P}_K)^T R^{-1}(R\hat{K}_i - B^T \hat{P}_K)
+ \gamma^{-2}(\hat{P}_K - \hat{P}_{K,i})DD^T(\hat{P}_K - \hat{P}_{K,i})
- \Delta K_{i+1}^T R \Delta K_{i+1} = 0.
\]  
(C.6)

Therefore, we have
\[
\hat{P}_K - \hat{P}_{K,i} \succeq \hat{F}_K - \int_0^\infty e^{A_{\Delta Ki}^T t}(\Delta K_{i+1}^T R \Delta K_{i+1})e^{A_{\Delta Ki} t} dt,
\]  
(C.7)

where \(\hat{E}_K = (R\hat{K}_i - B^T \hat{P}_K)^T R^{-1}(R\hat{K}_i - B^T \hat{P}_K)\), and \(\hat{F}_K := \int_0^\infty e^{A_{\Delta Ki}^T t} \hat{E}_K e^{A_{\Delta Ki} t} dt\). Let \(h_i := \|\hat{A}_i\|\) and \(s_i := \text{Tr}(e^{A_{\Delta Ki} t} A_{\Delta Ki}^T t)\). By Lemma B.4, \(\|\hat{F}_K\| \geq \frac{\log(5/4)}{2h_ie^c}\|\hat{E}_K\|\). By Lemma B.1, \(\|\hat{E}_K\| \geq \frac{1}{\nu} \text{Tr}(\hat{P}_K - P^*)\).

As a consequence,
\[
\text{Tr}(\hat{P}_{K,i} - P^*) \leq (1 - \frac{\log(5/4)}{2h_ie^c}) \text{Tr}(\hat{P}_K - P^*)
+ s_i \delta(R)\|\Delta K_{i+1}\|^2.
\]  
(C.8)

Hence, if \(\|\Delta K_i\|_F \leq l(\bar{K}_i)\) for all \(i \leq i'\),
\[
\text{Tr}(\hat{P}_{K,i} - P^*) \leq \sum_{i=1}^{i'-1} \left(1 - \frac{\log(5/4)}{2h_ie^c}\right) \text{Tr}(\hat{P}_{K,i} - P^*)
+ \sum_{i=1}^{i'-1} s_i \delta(R)\|\Delta K_{i+1}\|^2.
\]  
(C.9)

Let \(l := \inf_{i \in \mathbb{N}_+} l(\bar{K}_i)\), \(\hat{h} := \sup_{i \in \mathbb{N}_+} h_i\), \(\hat{s} := \sup_{i \in \mathbb{N}_+} s_i\), \(\hat{\delta} := (1 - \frac{\log(5/4)}{2h_i e^c})\). From Lemma A.9, \(\|\Delta K\|_F \leq l(\bar{K}_i)\) for any \(i \in \mathbb{N}_+\).

\[
\text{Tr}(\hat{P}_{K,i} - P^*) \leq \hat{\delta}^{i-1} \text{Tr}(\hat{P}_K - P^*) + \kappa(\|\Delta K\|_F).
\]

As \(\|\hat{P}_K - P^*\|_F \leq \text{Tr}(\hat{P}_K - P^*)\) by Lemma A.1, the Theorem 4 holds.

**C. Disturbance (Inner) Loop**

**Proof of Theorem 5.** When \(\|\Delta L_K\| < c\), we have \(\hat{A}_{Ki}\) as Hurwitz, owing to Lemma C.2. Rewriting (33a) for the \((i+1)\)th iteration and subtracting (33a) from it, we have
\[
(\hat{A}_{Ki}^{i+1})^T (\hat{P}_{Ki+1} - \hat{P}_K) + (\hat{P}_{Ki+1} - \hat{P}_K)(\hat{A}_{Ki}^{i+1})
+ \gamma^{-2}(\hat{P}_K - \hat{P}_{Ki+1})DD^T(\hat{P}_K - \hat{P}_{Ki+1})
- \Delta L_{Ki+1}^T \Delta L_{Ki+1} = 0.
\]  
(C.10)

As \(\hat{A}_{Ki}^{i+1}\) is Hurwitz, we have
\[
\hat{P}_{Ki+1} - \hat{P}_K
= \int_0^\infty e^{(\hat{A}_{Ki}^{i+1})^T t} \left[\hat{E}_{Ki} - \gamma^2 (\Delta L_{Ki})^T \Delta L_{Ki}\right] e^{(\hat{A}_{Ki}^{i+1}) t} dt
\]  
(C.11)

where \(\hat{E}_{Ki} := \gamma^{-2}(\hat{P}_K - D^T \hat{P}_K)^T (\hat{P}_K - D^T \hat{P}_K)^T\).

Define \(\tilde{E}_{Ki} := \int_0^\infty e^{(\hat{A}_{Ki}^{i+1})^T t} \hat{E}_{Ki} e^{(\hat{A}_{Ki}^{i+1})^T t} dt\) so that
\[
\hat{P}_K - \hat{P}_{Ki+1} = \hat{P}_K - \hat{P}_K
+ \int_0^\infty e^{(\hat{A}_{Ki}^{i+1})^T t} \left[\gamma^2 (\Delta L_{Ki})^T \Delta L_{Ki}\right] e^{(\hat{A}_{Ki}^{i+1})^T t} dt.
\]  
(C.13)

Let \(F_K = \sup_{i \in \mathbb{N}_+} \|\hat{A}_{Ki}^{i+1}\|\). From Lemma B.2, we can write, \(-\|\hat{F}_K\| \leq -\frac{\log(5/4)}{c(\mathcal{K})}\|\hat{E}_K\|\). Furthermore, by Lemma B.3, we can write \(-\|\hat{E}_K\| \leq -\frac{1}{c(\mathcal{K})} \text{Tr}(P_K - \hat{P}_K)\), where \(c(K) = \text{Tr}(\int_0^\infty e^{(\hat{A}_{Ki}+DL_K) t} e^{(\hat{A}_{Ki}+DL_K) t} dt)\) is defined in (B.41). Therefore, the trace of (C.13) can be written as
\[
\text{Tr}(P_K - \hat{P}_{Ki+1}) \leq \left(1 - \frac{\log(5/4)}{2h_i e^c}\right) \text{Tr}(P_K - \hat{P}_K)
+ \gamma^2 \|\Delta L_{Ki}\|^2.
\]  
(C.14)
Setting $g$ and $\hat{\beta}(K)$ as

$$g := \sup_{j \in \mathbb{N}_+} \text{Tr} \left( \int_0^\infty e^{A_k^{j+1} t} \mathbf{P}_k e^{A_k^{j+1} t}^T dt \right)$$  \hspace{1cm} (C.15a)

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

we find that

$$\text{Tr}(P_k - \hat{P}_k) \leq \hat{\beta}^{-1}(K) \text{Tr}(P_k) + \lambda(\|\Delta L\|_\infty),$$

where

$$\lambda(\|\Delta L\|_\infty) := \frac{1}{1 - \hat{\beta}(K)} \gamma_2^2 \|\Delta L\|_2^2.$$  \hspace{1cm} (C.16)

As $\|P_k - \hat{P}_k\|_F \leq \text{Tr}(P_k - \hat{P}_k)$, we thus establish the statement of Theorem 5.

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

**Proof.** Assume $\hat{A}_k$ is Hurwitz. From (27a), we have

$$(\hat{A}_k^{j+1})^T P_k + P_k \hat{A}_k^{j+1} + Q_k - \gamma^{-2} \hat{P}_k D D^T \hat{P}_k^j + \gamma^{-2} (P_k - \hat{P}_k) D D^T (P_k - \hat{P}_k)$$

$$- (\Delta L_k^j)^T D^T P_k - P_k D \Delta L_k^j = 0.$$  \hspace{1cm} (C.17)

If $\|\Delta L_k\|_F < \sigma_{\min}(Q_k - \gamma^{-2} P_k D D^T P_k) / 2\|D^T P_k\|_2 =: e$, it follows from (B.36) and the inequality $P_k \succeq \hat{P}_k$ that

$$Q_k - \gamma^{-2} \hat{P}_k D D^T \hat{P}_k^j + \gamma^{-2} (P_k - \hat{P}_k) D D^T (P_k - \hat{P}_k)$$

$$- (\Delta L_k^j)^T D^T P_k - P_k D \Delta L_k^j > 0.$$  \hspace{1cm} (C.18)

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, $\hat{A}_k$ is Hurwitz for all $j \in \mathbb{N}_+$ as long as $\|\Delta L_k\|_F \leq e$.

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.

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

[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.

[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.

[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, 2002.

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

[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.

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

[9] P. Khargonekar, I. Petersen, and M. Rotea, “H_\infty optimal control with state-feedback,” *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 786–788, 1988.

[10] D. Bernstein and W. Haddad, “LQG control with an H_\infty performance bound: a Riccati equation approach,” *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 293–305, 1989.
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’Alch´e-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 $H_2/H_\infty$ 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 $H_2$ and $H_\infty$ 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. 2
[45] O. Ogunnolu, 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. 2
[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 California, Berkeley, Berkeley, CA, USA, 2006. 5
[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, 9, 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. 8, 9
[55] K. Zhou, J. C. Doyle, and K. Glover, Robust and Optimal Control. Prentice hall Upper Saddle River, NJ, 1996. 8, 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. Sanituk 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