Robust Policy Optimization in Continuous-time Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Stochastic Control

Leilei Cui and Lekan Molu (*Member, IEEE*)

***Abstract*—An optimal control problem for a linear stochastic system possessing additive Brownian motion together with a cost that is an exponent of the quadratic form of the state, input, and disturbance terms is solved within the context of recent resurgence in establishing theoretical benchmarks for reinforcement leaning-based policy optimization for complex dynamical systems with continuous state and action spaces. While recent efforts in deterministic LQ regulator and additive Gaussian noise settings abound, our analyses are distinguished by many natural systems characterized by additive Wiener process, amenable to Îto's stochastic differential calculus in dynamic game settings. Since convergence to the *near-optimal* policy and robustness to the inherent noise statistics is crucial for assuring *system validation*, we provide rigorous convergence and robustness analyses followed by numerical experiments. Our approach provides basic insight into designing robust data-driven policy optimization for state feedback control policies for linear systems with a Wiener process as an additive disturbance.**

***Index Terms*—Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control, Policy Optimization, Robust Reinforcement Learning**

I. Introduction

Lately, various system-theoretic results analyzing the global convergence [1] and computational complexity [2] of nonconvex, constrained [3] gradient-based [4] and derivative-free [5], [6] policy optimization [7] in sampling-based reinforcement learning (RL) when the complete set of decision (or state feedback) variables are not previously known have appeared as control benchmarks [1], [8], [9], [10]. The most basic setting consists in optimizing over a decision variable K which must be determined from a (restricted) class of controllers $\mathcal{K}$ i.e. $\min_{K\in\mathcal{K}} J(K)$ where $J(K)$ is an objective (e.g. tracking error, safety assurance, goal-reaching measure of performance e.t.c.) required to be satisfied. In principle, K can be realized as a linear controller, a linear-in-the-parameters polynomial, or as a nonlinear kernel in the form of a radial basis function, or neural network.

These policy optimization (PO) schemes apply to a broad range of problems and have enjoyed wide success in complex systems where analytic models are difficult to derive [11], [12]. While they have become a popular tool for modern learning-based control [13], [14], the theoretical underpinning of their convergence, sample complexity, and robustness guarantees are little understood *in the large*. Only recently have rigorous analyses tools emerged [9], [15] for benchmarking RL with deterministic and additive Gaussian disturbance linear quadratic (LQ) controllers [16], [17], [18].

Tools for analyzing the convergence, sample complexity, or robustness of RL-based PO largely fall into one of infinite-horizon (i) discrete-time LQ regulator (LQR) settings i.e.

$$\min_{K\in\mathcal{K}} \mathbb{E}\sum_{t=0}^{\infty}(x_t^TQx_t+u_t^TRu_t) \text{ s.t. } x_{t+1}=Ax_t+Bu_t, x_0\sim\mathcal{P}_0$$

where $A, B, Q,$ and R are standard LQR matrices for state x_t, control input u_t and x_0 is drawn from a random distribution P_0 [18]; (ii) discrete-time LQ problems under multiplicative noise i.e. $\min_{\pi\in\Pi}\mathbb{E}_{x_0,\{\delta_i\},\{\gamma_i\}\}}\sum_{t=0}^{\infty}(x_t^TQx_t+u_t^TRu_t)$ subject to $x_{t+1}=(A+\sum_{i=1}^{p}\delta_{ti}A_i)x_t+(B+\sum_{i=1}^{q}\gamma_{ti}B_i)u_t$ and A, B, Q, R are the standard LQR matrices with δ_{ti} and γ_{tj} serving as the i.i.d zero-mean and mutually independent multiplicative noise terms [4]; or (iii) Zames' risk-sensitive [19] $\mathcal{H}_\infty$-control [20], [21] and discrete- and continuous-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design [22], [23], [3], [9] where the upper bound on the $\mathcal{H}_2$ cost is minimized subject to satisfying a set of risk-sensitive (often $\mathcal{H}_\infty$) constraints that *attenuate* [24], [25], [26] an unknown disturbance i.e. $\min_{K\in\mathcal{K}} J(K) := Tr(P_KDD^T)$ subject to $\mathcal{K} := \{K|\rho(A-BK)<1, \|T_{zw}(K)\|_\infty<\gamma\}$ where P_K is the solution to the generalized algebraic Riccati equation (GARE), A, B, D, K are standard closed-loop system matrices, $\|T_{zw}(K)\|_\infty$ denotes the $\mathcal{H}_\infty$-norm of the closed-loop transfer function from a disturbance input w to its output z, and $\gamma>0$, Here, $\gamma>0$, upper-bounded by $\gamma^\star$, a scalar measure of system risk-sensitivity [20].

We focus on continuous-time linear systems in which disturbances enter additively as random stochastic Wiener processes following contingent upon recent efforts on policy optimization for LQ regulator problems [18]); these systems may be modeled more accurately with uncertain additive Brownian noise. Here, diffusion processes modeled with Îto's stochastic calculus are the theoretical machinery for analysis. The prominent examples featuring additive Wiener processes occur in economics and finance [27], [28], stock options trading [29], protein kinetics, population growth models, dynamics of murmurations [30], and models involving computations with round-off error in floating point arithmetic calculations such as overparameterized neural network dynamics.

Our chief goal is to keep a controlled process, z, small in an infinite-horizon constrained optimization setting under a minimizing policy $\pi\in\Pi$ in spite of unforeseen *additive vector-valued stochastic Brownian process* $w(t)\in\mathbb{R}^q$ — which may be of large noise intensity. In terms of the L_2 norm, we can

L. Cui is with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA. (email: l.cui@nyu.edu).

L. Molu is with the Reinforcement Learning group at Microsoft Research, 300 Lafayette Street, New York, NY 10012, USA. (email: lekanmolu@microsoft.com).

write $\|z\|_2 = \left(\int |z(t)|^2 dt\right)^{1/2}$. The associated performance criteria can be realized as minimizing the expected value of the risk-sensitive linear exponential functions of positive definite quadratic forms state and control variables

$$\min_{\pi\in\Pi} \mathcal{J}_{exp}(x_0,\pi) := \mathbb{E}\Big|_{x_0\in\mathcal{P}_0} \exp\left[\frac{\alpha}{2}\int_0^\infty z^T(t)z(t)\mathrm{d}t\right],\ \alpha > 0$$
$$\text{subject to } \mathrm{d}x(t) = Ax(t)\mathrm{d}t + Bu(t)\mathrm{d}t + D\mathrm{d}w(t),$$
$$z(t) = Cx(t) + Eu(t) \tag{1}$$

with state process $x \in \mathbb{R}^n$, output process $z \in \mathbb{R}^p$ to be controlled, and control input $u \in \mathbb{R}^m$. The derivative of $w(t) \in \mathbb{R}^v$ *i.e.* dw/dt is a zero-mean Gaussian white noise with variance W, and $x(0)$ is a zero-mean Gaussian random vector independent of $w(t)$, $z(0) = 0$, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}, D \in \mathbb{R}^{n\times q}$, and $E \in \mathbb{R}^{p\times m}$ are constant matrix functions. The random signal $x(0)$ and the process $w(t)$ are defined over a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

Suppose that we carry out a Taylor series expansion about $\alpha = 0$ in (1), the variance term, $\alpha^2\texttt{var}(\int_0^\infty z^T z)$, will be small after minimization. Then α can be seen as a measure of *risk-aversion* if $\alpha > 0$. It is important to note that in this paper, we only consider state feedback. In particular when noise is present in the system, the value of α signifies the level of noise attenuation that penalizes the covariance matrix of the system's noise.

We adopt an adaptive policy iteration (PI) method in a continuous PO scheme. This can be seen as an instance of the actor-critic (AC) configuration in RL-based neural network *online* policy optimization schemess. Without explicit access to internal dynamics (system matrices), we must iterate between steps of policy evaluation and policy improvement. Mimicking the actor in RL AC settings, a parameterized controller must be evaluated relative to a parameterized cost function (the critic). The new policy is then used to improve the erstwhile (actor) policy by aiming to drive the cost to an extremal on the overall.

A. Contributions

We focus on the more sophisticated case of optimizing an *unknown stochastic linear policy* class $\mathcal{K}$ in an *infinite-horizon* LQ cost setting such that optimization iterates enjoy the *implicit regularization (IR) property* [10]—satisfying $\mathcal{H}_\infty$ robustness constraints. We place PO for *continuous-time linear stochastic controllers* on a rigorous *global convergence* and *robustness* footing. This is a distinguishing feature of our work. Our contributions are stated below.

- Reiterating the usual connection between risk-sensitive policy design and $\mathcal{H}_\infty$ zero-sum two-person dynamic differential games, we propose a *two-loop iterative alternating best-response procedure* for computing the optimal mixed-design policy – that accelerates the optimization scheme's convergence better than [4], [18], [8], [10] – in model- and sampling-based (i.e. model-free) cases;
- Rigorous local, global, and uniform convergence analyses follow for the loop updates in both settings;
- We further provide rigorous robustness, formalized in an input-to-state (ISS) framework, analyses to perturbations and uncertainties for the loop updates.
- Lastly, we benchmark our results against the natural policy gradient of [**?**] (which enjoys the IR property [10]) in the spirit of recent system-theoretic analysis works [18], [8], [31], [3].

Within the limits of our knowledge, we are not aware of other rigorous convergence and robustness analyses for the PO algorithms we that present.

B. Notations

The set of all symmetric matrices with dimension n is $\mathbb{S}^n$ and $\mathbb{R}$ ($\mathbb{N}_+$) is the set of real numbers (positive integers). The Kronecker product is denoted by $\otimes$. The Euclidean (Frobenius) norm of a vector or the spectral norm of a matrix is $\|\cdot\|$ ($\|\cdot\|_F$). Let $\|\cdot\|_\infty$ denote the supremum norm of a matrix-valued signal, i.e. $\|\Delta\|_\infty = \sup_{s\in\mathbb{Z}_+}\|\Delta_s\|_F$. The open ball of radius δ is $\mathcal{B}_\delta(X) = \{Y \in \mathbb{R}^{m\times n} | \|Y - X\|_F < \delta\}$. The maximum and minimum singular values (eigenvalues) of a matrix A are respectively denoted by $\bar{\sigma}(A)$ ($\bar{\lambda}(A)$) and $\sigma_{min}(A)$ ($\lambda_{min}(A)$). The eigenvalues of $A \in \mathbb{R}^{n\times n}$ are $\lambda_i(A)$ for $i = 1, 2, \cdots, n$. For the transfer function $G(s)$, its $\mathcal{H}_\infty$ norm is $\|G\|_{\mathcal{H}_\infty} = \sup_{\omega\in\mathbb{R}} \bar{\sigma}(G(j\omega))$.

The n-dimensional identity matrix is I_n. Denote by x_{ij} the (ij)'th entry of $X \in \mathbb{R}^{m\times n}$ and by x_i the i'th element of $x \in \mathbb{R}^n$. The full vectorization of $X \in \mathbb{R}^{m\times n}$ is the $mn \times 1$ vector, $\texttt{vec}(X) = [x_{11}, x_{21}, \cdots, x_{m1}, x_{12}, \cdots, x_{m2}, \cdots, x_{mn}]^T$. Let $P \in \mathbb{S}^n$, then the half-vectorization of P is the $n(n+1)/2$ column vector as a result of a vectorization of upper-triangular part of P i.e. $\texttt{svec}(P) = \left[p_{11}, \sqrt{2}p_{12}, \cdots, \sqrt{2}p_{1n}, p_{22}, \cdots, \sqrt{2}p_{n-1,n}, p_{nn}\right]^T$. The vectorization of the dot product $\langle x, x^T\rangle$, where $x \in \mathbb{R}^n$, is $\texttt{vecv}(x) = [x_1^2, \cdots, x_1x_n, x_2^2, x_2x_3, \cdots, x_n^2]^T$. The inverse of $\texttt{vec}(x)$ and $\texttt{svec}(y)$ are respectively the full and symmetric matricizations: $\texttt{mat}_{m\times n}(x) = \left(\texttt{vec}(I_n)^T \otimes I_m\right)$ $(I_n \otimes x)$, and $\texttt{smat}_m(P)$ so that $\texttt{smat}(\texttt{svec}(p)) = P$. Here, $x \in \mathbb{R}^{mn}$ and $y \in \mathbb{R}^{m(m+1)/2}$ for $n, m \in \mathbb{R}_{\geq 0}$. Finally, we denote by $T_{\texttt{vec}}(A)$ the vectorization of A^T i.e. $\texttt{vec}(A^T) = T_{\texttt{vec}}(\texttt{vec}(A))$.

C. Paper Structure

The rest of this article is structured as follows. LEQG connections to dynamic games are briefly established in Section II. In Section III, we present a double-loop PI procedure for robust policy recovery in a mixed $\mathcal{H}_2$/$\mathcal{H}_\infty$ PO settings (both in model-free and model-based optimization settings); this is followed by a rigorous analysis of their convergence and robustness properties. We demonstrate the efficacy of our proposed algorithm on numerical examples, discuss findings, and draw conclusions in Section IV.

II. Dynamic Games' Natural Connection

In this section, we connect PO under linear controllers to the theory of two-person dynamic games. Let us first impose conditions to make the problem introduced in (1) amenable to our analysis.

Assumption 1 : *We take $C^TC \triangleq Q \succ 0$, $E^T\,(C,\,E) = (0,\,R)$ for some matrix-valued function $R \succ 0$; and since in (1), we want $w(t)$ to be statistically independent, we take $DD^T = 0$. Seeing we are seeking a linear feedback controller for (1), we require that the pair (A, B) be stabilizable. We expect to compute solutions via an optimization process, therefore we require that unstable modes of A must be observable through Q. Whence $(\sqrt{Q}, A)$ must be detectable.*

Given Assumption 1, the LEQG cost functional now becomes

$$\mathcal{J}_{exp}(x_0, \pi) = \mathbb{E}\Big|_{x_0 \in \mathcal{P}_0} \exp\left[\frac{\alpha}{2}\int_0^\infty \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right)\mathrm{d}t\right], \tag{2}$$

for a fixed $\alpha > 0$ and the closed-loop transfer function is

$$T_{zw}(K) = (C - EK)\,(sI - A + BK)^{-1}D. \tag{3}$$

The set of all *suboptimal* control policies that robustly stabilizes the linear system against all (finite gain) stable perturbations Δ, interconnected to the system by $w = \Delta z$, such that $\|\Delta\|_\infty \le 1/\gamma$ is

$$\mathcal{K} = \{\, K : \; \lambda_i(A - B_1K) < 0, \; \|T_{zw}(K)\|_\infty < \gamma\}. \tag{4}$$

Observe: Cost (2), without the log term introduced in [6], mitigates against gradient bias in PG-based derivative-free optimization as we will show in subsequent sections — this objective provably converges to the optimal solution. We now establish the optimal control gain matrix for the LEQG problem.

Proposition 1 : *[32, Th. II.1] The optimal control to the LEQG optimization problem (1) and cost functional (2) under π in the infinite-horizon setting is of a linear-in-the-data form i.e. $u^\star(t) = -K^\star_{leqg}\hat{x}(t)$ where gain $K^\star_{leqg} = R^{-1}B^\top P_\tau$, and P_τ is the unique, symmetric, positive definite solution to the algebraic Riccati equation (ARE)*

$$A^\top P_\tau + P_\tau A - P_\tau(BR^{-1}B^\top - \alpha^{-2}DD^\top)P_\tau = -Q. \tag{5}$$

Corollary 1 : *In the infinite-horizon time-invariant case with constant system matrices and a stabilizable (A, B), by the theorem on "limit of monotonic operators" [33] and [20, Theorem 9.7], we find that $P^\star \triangleq P_\infty = \lim_{\tau\to\infty} P_\tau$, and $K^\star_{leqg} \triangleq K_\infty = \lim_{\tau\to\infty} K_\tau$.*

Remark 1 : *It is well-known by now that directly solving the linear exponential quadratic Gaussian problem (1) in policy-gradient frameworks incurs biased gradient estimates during iterations; this may affect the preservation of risk-sensitivity in infinite-horizon LTI settings (see [6], [10]). As such, we introduce a workaround with an equivalent dynamic game formulation to the stochastic LQ PO control problem in what follows.*

Lemma 1 (Closed-loop Two-Player Game Connection): *Consider the parameterized soft-constrained upper value, with a stochastically perturbed noise process $w(t)$, which enters the system dynamics as an additive bounded Gaussian with known statistics*[1],

$$\min_{u\in\mathcal{U}}\max_{\xi\in W} \bar{\mathcal{J}}_\gamma(x_0, u, \xi) := \mathbb{E}\Big|_{x_0\sim\mathcal{P}_0,\,\xi(t)} \int_0^\infty \left[x^\top(t)Qx(t) + u^\top(t)Ru(t) - \gamma^2\xi^\top(t)\xi(t)\right]\mathrm{d}t$$
$$\text{subject to } \mathrm{d}x(t) = Ax(t)\mathrm{d}t + Bu(t)\mathrm{d}t + D\xi(t),$$
$$z(t) = Cx(t) + Eu(t) \tag{6}$$

with $\xi(= dw)$ is the zero-mean Gaussian noise with variance W(equivalent to dw/dt in (1)), scalar $\gamma > 0$ denoting the level of disturbance attenuation, and x_0 an arbitrary initial state. Suppose that there exists a non-negative definite (nnd) solution of (5) (with α replaced by γ), then its minimal realization, $P^\star$, exists. If $(A, Q^{\frac{1}{2}})$ is observable, then every nnd solution $P^\star$ of (5) is positive definite. For a nnd $P^\star$, there exists a common upper and lower value for the game and if $\bar{\mathcal{J}}_\gamma$ is finite for some $\gamma := \hat{\gamma} > 0$, then $\bar{\mathcal{J}}_\gamma$ is bounded (if and only if the pair (A, B) is stabilizable) and equivalent to the lower value $\underline{\mathcal{J}}_\gamma$[2]. *In addition, for a bounded $\bar{\mathcal{J}}_\gamma$ for some $\gamma = \hat{\gamma}$ and for optimal gain matrices, $K^\star = R^{-1}B^\top P_{K,L}$, $L^\star = \gamma^{-2}D^\top P_{K,L}$, $\bar{\mathcal{J}}_\gamma$ admits the following Hurwitz feedback matrices for all $\gamma > \hat{\gamma}$*

$$A^\star_K = A - BK^\star,\; A^\star_{K,L} = A^\star_K + DL^\star \tag{7}$$

where the nnd $P_{K,L}$ is the unique solution of (5) for $\gamma > \hat{\gamma}$ in the class of nnd matrices if it renders $A^\star_{K,L}$ Hurwitz. Whence, the saddle-point optimal controllers are

$$u^\star(x(t)) = -K^\star x(t),\; \xi^\star(x(t)) = L^\star x(t). \tag{8}$$

Proof. The proof follows that in [20, Th. 9.7] exactly if we preserve the γ^{-1} term in the ARE of equation 9.31 in [20] and replace it by γ^{-2} as we have here. □

Remark 2 (Connection to the LEQG problem): *Suppose that we equate α to γ in (5), then for a finite scalar $\gamma^\star > 0$ such that $\gamma > \gamma^\star$, the LEQG gain matrix K_{leqg} is equivalent to the saddle-point gain matrix for the minimizing player in (6). In addition, the problem in (6) is the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem when $x_0 = 0$, and the linear system is time-invariant for an infinite-horizon problem [10]. This allows us to circumvent directly optimizing the LEQG policy in (1). We refer readers to [10, Lemma C.2] for the justification of using LTI control for LEQG.*

It follows that for any stabilizing control pair (K, L), if (8) is applied to the system in (6), the resulting cost from (6) is $\bar{J}_\gamma = (x_0^\top P_{K,L}x_0)$ [34]. Furthermore, let the state correlation matrix be defined as $\Sigma_{K,L} = \mathbb{E}_{x_0\sim\mathcal{D}}(x^\top(t)P_{K,L}x(t))$. In what follows, we present a double-loop iterative solver for the gains K and L in (8) for finding the saddle point (equivalently Nash Equilibrium) policies (8)

[1] Since the time derivative of a Brownian process $w(t)$ is $dw(t)$, we maximize over the Gaussian $dw(t)$, rather than the unbounded stochastic noise $w(t)$.

[2] The lower value is constructed by reversing the order of play in the value defined in (6).

III. Policy Iteration: Convergence, and Robustness

We resort to policy iteration in a double loop iterative scheme for obtaining the optimal feedback control policies (8) for the dynamic game (6) when (i) exact models are known to provide a barometer for our later analysis when (ii) exact models are unknown.

Let p and q be indices of nested iterations between updating the closed-loop minimizing player's controller K_p (in an outer loop) and the maximizing player's controller $L_q(K_p)$ (in an inner-loop) for $p = 1, 2, \ldots, \bar{p}$ and $q = 1, 2, \ldots, \bar{q}$. Note that $(\bar{p}, \bar{q}) \in \mathbb{N}_+$. Furthermore, let the iterates' identities cref. (7) be

$$\begin{aligned} A_K^p &= A - BK_p, \quad A_{K,L}^{p,q} = A_K^p + DL_q(K_p), \\ Q_K^p &= Q + K_p^\top RK_p, \; A_K^\gamma = A_K^p + \gamma^{-2}DD^\top P_K^p. \end{aligned} \tag{9}$$

For the soft-constrained value functional (6) at the p'th iterate of the minimizing controller K we have the following value iteration form for (5),

$$A_K^{p\top} P_K^p + P_K^p A_K^p + Q_K^p + \gamma^{-2} P_K^p DD^\top P_K^p = 0, \tag{10a}$$
$$K_{p+1} = R^{-1}B^T P_K^p \tag{10b}$$

where P_K^p is the p'th iterate's solution to (10). Similarly, for the maximizing controller, $L_q(K_p)$, the following closed-loop ARE iteration applies

$$A_{K,L}^{(p,q)^\top} P_{K,L}^{p,q} + P_{K,L}^{p,q} A_{K,L}^{p,q} + Q_K^p - \gamma^2 L_q^\top(K_p) L_q(K_p) = 0 \tag{11a}$$

$$K_{p+1} = R^{-1}B^T P_K^{p,q}, \; L_{q+1}(K_p) = \gamma^{-2}D^\top P_{K,L}^{p,q} \tag{11b}$$

where $P_{K,L}^{p,q}$ is the solution to (11) for arbitrary gains $[K_p, L_q(K_p)]$ — updated recursively in a nested loop (discussed shortly). Choosing a stabilizing minimizing player control gain, we evaluate u's performance by solving (10). The policy is then improved in a following iteration by solving (11b). Observe: P_K^p and $P_{K,L}^{p,q}$ admit an equivalence arising from 2.

Problem 1 (Model-Based Policy Iteration): *After the last update of the inner loop maximizing player gain $L_{\bar{q}}(K_p)$, the outer-loop update of the minimizing controller gain K_p robustly stabilizes the closed-loop transfer function from w to z i.e. $T_{zw}(K, L)$ under gains K_p and $L_{\bar{q}}(K_p)$ against all (finite gain) disturbances Δ interconnected to system (6) (by $w = \Delta z$) such that $\|\Delta\|_\infty \le \gamma^{-2}$. The set of all such suboptimal controllers is denoted*

$$\begin{aligned} \breve{\mathcal{K}} = \{(K, L) : \lambda_i(A_K^p) < 0, \lambda_i(A_{K,L}^{p,q}) < 0, \\ \|T_{zw}(K, L)\|_\infty < \gamma \textit{ for all } (p, q) \in \mathbb{N}\}. \end{aligned} \tag{12}$$

A. Double Loop (DL) Successive Substitution

The procedure for obtaining the optimal $P^\star$ in Problem 1 is described in Algorithm 1. It finds a global Nash Equilibrium (NE) (or equivalently a saddle-point equilibrium) [8] of the LQ zero-sum game (6) by solving the nonlinear ARE (11) in a nested two-loop policy iteration (PI) scheme.

Algorithm 1: Model-Based Policy Iteration

```
  Input: Fix max number of loop iterations $\bar{p}$ and $\bar{q}$;
  Input: Fix desired level of risk attenuation $\gamma > 0$;
  Input: Compute $(K_0, L_0) \in \mathcal{K}$;    ▷ From [35, Alg. 1];
  Input: Set $P_{K,L}^{0,0} = Q_K^0$;              ▷ See (9);
  Input: Fix a control penalty matrix $R \succ 0$.
1 for $p = 0, \ldots, \bar{p}$ do
2     Compute $Q_K^p$ and $A_K^p$                  ▷ See (9);
3     while $q < \bar{q}$ do
4         Compute $L_{q+1}(K_p) := \gamma^{-2} D^\top P_{K,L}^{p,q}$;
5         GARE_solver($\gamma, Q_K^p, A_K^p, P_{K,L}^{p,q}, L_{q+1}(K_p)$)
           ▷ Solve (11) recursively for $P_{K,L}^{p,q}$ until $q \equiv \bar{q}$;
6     end
7     Compute $K_{p+1} = R^{-1} B^\top P_{K,L}^{p,\bar{q}}$     ▷ See (11b) ;
8 end
```

An initial (K_0, L_0) control pair that guarantees the iterates' feasibility upon projection onto the set $\mathcal{K}$ is first determined in order to enforce the condition (12). We refer readers to our recent conference paper [35] where this procedure is elucidated[3]. Afterwards, the Riccati equation's (11) solution i.e. $P_{K,L}^{0,0} \triangleq Q_K^0$ must be computed and the gains $[K_p, L_q(K_p)]$ are updated as in (11b). *As an RL-based PI procedure, Line 7 in Alg. 1 can be seen as a reinforcement over the time interval p to $p+1$.* In our implementation, we recursively solved for $P_{K,L}^{p,q+1}$ until a user-specified error tolerance is numerically achieved.

B. Outer Loop Stability, Optimality, and Convergence

We now discuss the convergence guarantees of the iterations under perfect dynamics and then analyze the fidelity of projecting the gains to the feasible set under inexact loop updates. Let us introduce the following preliminary Lemma to guide the establishment of our later results in this section.

Lemma 2 : *Under Assumption 1 and for the ARE (10), if $K_0 \in \mathcal{K}$[4], then for any $p \in \mathbb{N}_+$, we must have the following conditions for the optimal $K^\star$ and $P^\star$,*

(1) $K_p \in \mathcal{K}$;
(2) $P_K^0 \succeq P_K^1 \succeq \cdots P_K^p \succeq \cdots \succeq P^\star$;
(3) $\lim_{p\to\infty} \|K_p - K^*\|_F = 0$, $\lim_{p\to\infty} \|P_K^p - P^*\|_F = 0$.

We provide proof (with the robustness operator γ of (10)) in below. An alternate proof is given in [34].

Proof of Lemma 2. When $p = 0$, $K_0 \in \mathcal{K}$, and it satisfies (1) (See [35, Alg. 1].) For $p > 0$, introduce the identities,

$$RK_{p+1} = B^\top P_K^p, \qquad K_{p+1}^\top R = P_K^p B, \tag{13a}$$
$$A_K^{p\top} P_K^p = A_K^{(p+1)\top} P_K^p + (K_{p+1} - K_p)^\top B^\top P_K^p, \tag{13b}$$
$$P_K^p A_K^p = P_K^p A_K^{(p+1)} + P_K^p B (K_{p+1} - K_p). \tag{13c}$$

[3] We remark that this can also be found via linear matrix inequality approaches [36], [37].

[4] The $\mathcal{K}$ defined here refers to the one defined in (4).

Therefore, equation (10) becomes

$$A_K^{(p+1)\top} P_K^p + P_K^p A_K^{(p+1)} + \gamma^{-2} P_K^p D D^\top P_K^p + C^\top C \quad (14)$$
$$+ K_{p+1}^\top R K_{p+1} + (K_{p+1} - K_p)^\top R (K_{p+1} - K_p) = 0.$$

Thus, for a stabilizing $K_{p+1} (\neq K_p)$ we must have $(K_{p+1} - K_p)^\top R (K_{p+1} - K_p) \succ 0$ so that

$$A_K^{(p+1)\top} P_K^p + P_K^p A_K^{(p+1)} + \gamma^{-2} P_K^p D D^\top P_K^p + Q_K^{p+1} \prec 0. \tag{15}$$

If (read: since) the inequality (15) holds, the bounded real Lemma [10, Lemma A.1, statement 3] stipulates that a $P_K^p \succ 0$ exists; by [10, Lemma A.1, statement 1], $\|T_{zw}(K)\|_\infty < \gamma$ given that $\lambda_i(A_K^{(p+1)}) < 0$ in (15). *A fortiori*, $K_p \in \mathcal{K}$ for $p > 0$ by the bounded real Lemma. This proves the first statement.

The proof for statement (2) now follows. At the $(p+1)$'th iteration, it can be verified that (10) admits the form

$$A_K^{(p+1)\top} P_K^{p+1} + P_K^{p+1} A_K^{(p+1)} + C^\top C + K_{p+1}^\top R K_{p+1}$$
$$+ \gamma^{-2} P_K^{p+1} D D^\top P_K^{p+1} = 0 \tag{16}$$

so that subtracting (16) from (14) (i.e. at the p'th iteration) and using the statistical independence property of the noise term $w(t)$ (from Ass. 1) i.e. $DD^\top = 0$, we have

$$A_K^{(p+1)\top} \left[P_K^p - P_K^{p+1} \right] + \left[P_K^p - P_K^{p+1} \right] A_K^{(p+1)}$$
$$(K_{p+1} - K_p)^\top R (K_{p+1} - K_p) = 0. \tag{17}$$

Observe: (17) is a Lyapunov equation of the form

$$A_K^{(p+1)\top} \tilde{P}_K^{p+1} + \tilde{P}_K^{p+1} A_K^{(p+1)} + \tilde{Q}_K^{p+1} = 0.$$

If we let $\tilde{P}_K^{p+1} = (P_K^p - P_K^{p+1})$, $\tilde{K}_{p+1} = (K_{p+1} - K_p)$, and $\tilde{Q}_K^{p+1} = \tilde{K}_{p+1}^\top R \tilde{K}_{p+1}$. Since $A_K^{(p+1)}$ is Hurwitz and satisfies the above Lyapunov equation, we must have $\tilde{P}_K^{p+1} \succeq 0$ because $\tilde{Q}_K^{p+1} \succeq 0$ by Lemma 13. Whence, $\tilde{P}_K^{p+1} \succeq 0$ implies that $P_K^p \succeq P_K^{p+1}$ and $\tilde{Q}_K^{p+1} \succeq 0$ implies $K_{p+1} \geq K_p$. This proves the second statement. In this sentiment, the sequence $\{P_K^p\}_{p=1}^\infty$ is decreasing, bounded below by 0 and has a finite norm so that $\{P_K^p\}_{p=1}^\infty$ converges to P_K^∞. This satisfies (5). Observe from equation (10) that P_K^p is self-adjoint so that from the "limit of monotonic positive operators theorem" [33, p. 189], $\lim_{p\to\infty} \|P_K^p - P^*\|_F = 0$. By a similar argument for decreasing operators sequences [33, p. 190], the sequence $\{K_K^p\}_{p=0}^\infty$ is increasing and upper bounded by K_K^∞. Hence, $\lim_{p\to\infty} \|K_p - K^*\|_F = 0$. The third statement is thus proven. □

In [10, Theorem A.7 and A.8], the authors showed that the controller update phase in the outer-loop has a global sub-linear and local quadratic convergence rates. We now demonstrate that the outer-loop iteration has a global linear convergence rate. Let us first establish a few preliminary results that we will need in the proof of our main result i.e. Theorem 1.

Lemma 3 : *Let* $\Psi = \tilde{Q}_K^{p+1}$ *so that* $\Psi = \Psi^\top \succeq 0$. *Furthermore, let* $\Phi \in \mathbb{R}^{n \times n}$ *be Hurwitz,* $\Theta = \int_0^\infty e^{(\Phi^\top t)} \Psi e^{(\Phi t)} dt$, *and* $a(\Phi) = \log(5/4) \|\Phi\|^{-1}$. *Then,* $\|\Theta\| \geq \frac{1}{2} a(\Phi) \|\Psi\|$.

Proof. Define $S(t) = \sum_{k=1}^\infty (\Phi t)^k / k!$ so that $e^{\Phi t} = I_n + \sum_{k=1}^\infty (\Phi t)^k / k! = I_n + S(t)$ after a Taylor series expansion. Whence $\|S(t)\| = \sum_{k=1}^\infty (\|\Phi\| t)^k / k!$ or $\|S(t)\| \geq e^{\|\Phi\| t} - 1$. For $x_0 \neq 0$ satisfying $x_0^\top \Psi x_0 = \|\Psi\| \|x_0\|^2$, we have

$$\begin{aligned} x_0^\top \Theta x_0 &\geq \int_0^{a(\Phi)} x_0^\top e^{\Phi^\top} \Psi e^{\Phi} x_0 \mathrm{d}t, \\ &\geq \int_0^{a(\Phi)} x_0^\top \left(I_n + S(t) \right)^\top \Psi \left(I_n + S(t) \right) x_0 \mathrm{d}t, \\ &\geq \int_0^{a(\Phi)} \|\Psi\| \|x_0\|^2 - 2(e^{(\|\Phi\| a(\Phi))} - 1) \|\Psi\| \|x_0\|^2 \mathrm{d}t, \\ &\geq \int_0^{a(\Phi)} \frac{1}{2} \|\Psi\| \|x_0\|^2 \mathrm{d}t \geq \frac{1}{2} a(\Phi) \|\Psi\| \|x_0\|^2. \end{aligned} \tag{18}$$

A fortiori, Lemma 3's proof follows from (18). □

Remark 3 : *For* $A_K = A - BK$, *we know from the bounded real Lemma [10, Lemma A.1] that the Riccati equation*

$$A_K^\top P_K + P_K A_K + Q_K + \gamma^{-2} P_K D D^\top P_K = 0 \tag{19}$$

admits a unique positive definite solution $P_K \succ 0$ *with a Hurwitz* $(A_K + \gamma^{-2} D D^\top P_K)$.

Lemma 4 (Optimality of the iteration): *Consider any* $K \in \mathcal{K}$, *let* $K' = R^{-1} B^\top P_K$ *(where* P_K *is the solution to (19)), and* $E_K = (K - K')^\top R (K - K')$. *If* $E_K = 0$, *then* $K = K^\star$.

Lemma 5 (Bounds on Gain Difference Matrix): *For any* $h > 0$, *define* $\mathcal{K}_h := \{K \in \mathcal{K} | Tr(P_K - P^\star) \leq h\}$. *For any* $K \in \mathcal{K}_h$, *let* $K' := R^{-1} B^\top P_K$, *where* P_K *is the solution to (19), and* $E_K := (K - K')^\top R (K - K')$. *Then, there exists* $b(h) > 0$, *such that* $\|P_K - P^\star\|_F \leq b(h) \|E_K\|_F$.

Theorem 1 : *For any* $h > 0$ *and* $K_0 \in \mathcal{K}_h$, *there exists* $\alpha(h) \in [0, 1)$, *such that* $Tr(P_K^{p+1} - P^\star) \leq \alpha(h) Tr(P_K^p - P^\star)$. *That is,* $P^\star$ *is an exponentially stable equilibrium.*

Proof. For a Hurwitz $A_K^{(p+1)}$ and an $E_K^p = (K_p - K_{p+1})^\top R (K_p - K_{p+1})$, Lemma 13 and equation (17) imply that

$$P_K^p - P_K^{p+1} \succeq \int_0^\infty e^{A_K^{(p+1)\top} t} E_K^p e^{A_K^{(p+1)} t} \mathrm{d}t =: H_K^p. \tag{20}$$

From Lemma 2, we have for $p > 0$ that $P_K^0 \succeq P_K^p$ so that

$$\|A_K^{p+1}\| \leq \|A\| + (\|BR^{-1}B^T\| + \gamma^{-2} \|DD^T\|) h. \tag{21}$$

Set $c(h) = \log(5/4) / \|A\| + (\|BR^{-1}B^T\| + \gamma^{-2} \|DD^T\|) h$ so that we have $\|H_K^p\| \geq \frac{1}{2} c(h) \|E_K^p\|$ from Lemma 3. Using Lemmas 15 and 5, and taking the trace of (20) we find that

$$\begin{aligned} &Tr(P_K^{p+1} - P^\star) \leq Tr(P_K^p - P^\star) - Tr(H_K^p) \\ &\leq Tr(P_K^p - P^\star) - c(h) \|E_K^p\| / 2 \\ &\leq Tr(P_K^p - P^\star) - \frac{c(h)}{2\sqrt{n}} \|E_K^p\|_F \\ &\leq Tr(P_K^p - P^\star) - \frac{c(h)}{2\sqrt{n} b(h)} \|P_K^p - P^\star\|_F \\ &\leq \left(1 - \frac{c(h)}{2 n b(h)} \right) Tr(P_K^p - P^\star). \end{aligned} \tag{22}$$

The proof follows by setting $\alpha(h) = 1 - c(h) / 2nb(h)$. □

C. Inner Loop Stability, Optimality, and Convergence

We now analyze the monotonic convergence rate of the inner loop. Given arbitrary gains $K_p \in \mathcal{K}$ and $L_q(K_p)$, let $P_{K,L}^{p,q}$ be the positive definite solution of the associated Lyapunov equation (11). The following lemma shows that the cost matrix $P_{K,L}^{p,q}$ monotonically converges to (11)'s solution.

Lemma 6 : *Suppose that $L_0(K_0)$ is stabilizing, then for any $q \in \mathbb{N}_+$ (with $P_{K,L}^{p,\bar{q}}$ as the solution to (11)),*

1) $A_{K,L}^{p,q}$ is Hurwitz;
2) $P_{K,L}^{p,\bar{q}} \succeq \cdots \succeq P_K^{(p,q+1)} \succeq P_K^{p,q} \succeq \cdots \succeq P_{K,L}^{p,0}$; and
3) $\lim_{q\to\infty} \|P_{K,L}^{p,q} - P_{K,L}^{p,\bar{q}}\|_F = 0$.

Proof. To prove the first statement, we proceed by induction. For a $p \geq 0$ we have $K_p \in \mathcal{K}$ by Theorem 1. Subtracting (11) from (10) yields

$$0 = A_K^{p\top}(P_K^p - P_{K,L}^{p,q}) + (P_K - P_{K,L}^{p,q})A_K^p + \\ \gamma^2 \left[L_{q+1}(K_p) - L_q(K_p)\right]^\top \left[L_{q+1}(K_p) - L_q(K_p)\right]. \quad (23)$$

In equation (23), we have that $\left[L_{q+1}(K_p) - L_q(K_p)\right]^\top \left[L_{q+1}(K_p) - L_q(K_p)\right] \succeq 0$ so that (23) admits a Lyapunov equation form. Following statement 2 of Lemma 13, we must have $(P_K^p - P_{K,L}^{p,q}) \succeq 0$. *A fortiori*, we must have $A_{K,L}^{(p,q)}$ Hurwitz in (23) following Lemma 13. This proves the first statement.

To prove the second statement, we proceed thus. Abusing notation by dropping the templated argument in $L_q(K_p)$, let us consider the identities,

$$A_{K,L}^{(p,q)^\top} P_{K,L}^{p,q} = A_{K,L}^{(p,q+1)^\top} P_{K,L}^{p,q} - \gamma^2 \left[L_{q+1} - L_q\right]^\top L_{q+1} \\ P_{K,L}^{p,q} A_{K,L}^{(p,q)} = P_{K,L}^{p,q} A_{K,L}^{(p,q+1)} - \gamma^2 L_{q+1}^\top \left[L_{q+1} - L_q\right]. \quad (24)$$

We now rewrite (11) in light of (24) as

$$A_{K,L}^{(p,q+1)^\top} P_{K,L}^{p,q} + P_{K,L}^{p,q} A_{K,L}^{(p,q+1)} - \gamma^2 \left[L_{q+1} - L_q\right]^\top L_{q+1} \\ + Q_K - \gamma^2 L_{q+1}^\top \left[L_{q+1} - L_q\right] - \gamma^2 (L_q^\top L_q) = 0. \quad (25)$$

At the $(q+1)$'th iteration, we have (11) as

$$A_{K,L}^{(p,q+1)^\top} P_{K,L}^{p,q+1} + P_{K,L}^{p,q+1} A_{K,L}^{(p,q+1)} + Q_K \\ - \gamma^2 L_{q+1}^\top(K_p) L_{q+1}(K_p) = 0. \quad (26)$$

Subtracting (25) from (26), we have

$$A_{K,L}^{(p,q+1)^\top} \left[P_{K,L}^{p,q+1} - P_{K,L}^{p,q}\right] + \left[P_{K,L}^{p,q+1} - P_{K,L}^{p,q}\right] A_{K,L}^{(p,q+1)} + \\ + \gamma^2 \left[L_{q+1} - L_q\right]^\top \left[L_{q+1} - L_q\right] = 0. \quad (27)$$

Since $\left[L_{q+1} - L_q\right]^\top \left[L_{q+1} - L_q\right] \succeq 0$, (27) is indeed a Lyapunov equation so that $P_{K,L}^{p,q+1} \succeq P_{K,L}^{p,q}$ holds following Lemma 13. Whence, we must have $A_{K,L}^{(p,q+1)\top}$ Hurwitz. Following the argument for all $(q, q') \in \bar{q}$ with $q \neq q'$, statement 2) holds.

Observe: $P_{K,L}^{p,q}$ is self-adjoint by reason of (10). By the theorem on the "limit of monotonically decreasing operators" [33, pp. 190], statement 2) implies that the sequence $\{P_{K,L}^{p,\bar{q}}, \cdots, P_{K,L}^{p,q=0}\}$ is monotonically decreasing and bounded from above by $P_{K,L}^{p,\bar{q}} \equiv P_{K,L}^{p,\star} \triangleq P_K^p$. That is, P_K^p exists and is the solution of (10) and $P_{K,L}^{p,\bar{q}}$ is the unique positive definite solution to (11). *A fortiori*, we must have $\lim_{q\to\infty} P_{K,L}^{p,q} = P_{K,L}^{p,\infty}$. This establishes the third statement. □

1) Convergence Rate Analysis: We now analyze the monotonic convergence of the inner loop of the nested double loop algorithm. Let us first discuss a preliminary result.

Lemma 7 (Monotonic Convergence of the Inner-Loop): *For any $K \in \mathcal{K}$, let $L(K)$ be the control gain for the player w such that $A_K + DL(K)$ is Hurwitz. In addition, let P_K^L be the solution of*

$$\left(A_K + DL(K)\right)^\top P_K^L + P_K^L \left(A_K + DL(K)\right) + Q_K \\ - \gamma^2 L(K)^\top L(K) = 0. \quad (28)$$

And define $L'(K) = \gamma^{-2} D^\top P_K^L$ and $E_K^L = \gamma^{-2}(L'(K) - L(K))^\top (L'(K) - L(K))$. Then, for a $c(K) = Tr\left(\int_0^\infty e^{(A_K + DL(K^\star))t} e^{(A_K+DL(K^\star))^\top t} \mathrm{d}t\right)$, the following inequality holds $Tr(P_K - P_K^L) \leq \|E_K^L\| c(K)/$

Theorem 2 : *For a $K \in \mathcal{K}$, and for any $q \in \mathbb{N}_+$, there exists $\beta(K) \in [0, 1)$, such that*

$$Tr(P_K^p - P_{K,L}^{p,q+1}) \leq \beta(K) Tr(P_K^p - P_{K,L}^{p,q}). \quad (29)$$

That is, the inner loop has a global linear convergence rate.

Proof. By Lemma 6, $A_{K,L}^{p,q}$ is Hurwitz. It follows from Lemma 13 and (14) that

$$P_{K,L}^{p,q+1} - P_{K,L}^{p,q} = \underbrace{\int_0^\infty e^{(A_{K,L}^{(p,q+1)\top})t} E_K^q e^{A_{K,L}^{(p,q+1)t}} \mathrm{d}t}_{F_K^q}. \quad (30)$$

By Lemma 6, $P_K^p \succeq P_{K,L}^{p,q}$ so that

$$\|A_{K,L}^{(p,q+1)}\| \leq \|A - BK_p\| + \gamma^{-2}\|DD^\top\|\|P_K\|. \quad (31)$$

$$\text{Let } d(K) = \log(5/4)/\left(\|A_K\| + \gamma^{-2}\|DD^T\|\|P_K^p\|\right), \quad (32)$$

so that from the trace of (31), we find that

$$Tr(P_K^p - P_{K,L}^{p,q+1}) = Tr(P_K^p - P_{K,L}^{p,q}) - Tr(F_K^q), \quad (33a)$$
$$\leq Tr(P_K^p - P_{K,L}^{p,q}) - \|F_K^q\|, \quad (33b)$$
$$\leq Tr(P_K^p - P_{K,L}^{p,q}) - \frac{1}{2} d(K) \|E_K^j\|, \quad (33c)$$

where we have used Lemma 15 to arrive at the inequality in (33b), and Lemma 3 for (33c). Furthermore, from Lemma 7

$$Tr(P_K^p - P_{K,L}^{p,q}) \leq \left(1 - \frac{1}{2}\frac{d(K)}{c(K)}\right) Tr(P_K^p - P_{K,L}^{p,q}). \quad (34)$$

The proof follows if we set $\beta(K) = 1 - d(K)/2c(K)$. □

Remark 4 : *As seen from Lemma 6, $P_K - P_K^j \succeq 0$. From Lemma 15 and the result of Theorem 2, we have $\|P_K - P_{K,L}^{p,q}\|_F \leq Tr(P_K - P_{K,L}^{p,q}) \leq \beta^j(K) Tr(P_K)$, i.e. $P_{K,L}^{p,q}$ exponentially converges to P_K in the Frobenius norm.*

The following lemma guarantees uniform convergence after an equal number of inner-loop iterations so that $P_{K,L}^{p,q}$ and $L_q(K)$ enter the vicinity of P_K and $L(K^\star)$ irrespective of the different values of K.

Lemma 8 (Uniform Convergence of Iterates): *For any $h > 0$, $K \in \mathcal{K}_h$, and $\epsilon > 0$, there exists $q'(h) \in \mathbb{N}_+$ independent of K, such that if $q \geq q'(h)$, $\|P_{K,L}^{p,q} - P_K\|_F \leq \epsilon$.*

Proof of Lemma 8. This Lemma is an immediate outcome of Theorems 1 and 2. □

D. Sampling-based (Hybrid) Nonlinear System and PI

In practice, the exact knowledge of the system matrices are unavailable so that the policy evaluation step will result in biased estimates. When errors are present from using I/O or state data for the PO procedure in Alg. 1, residuals from early termination of numerically solving Line 5 in Alg. 1, or using an approximate cost function owing to inexact values of Q and R, the algorithm may fail to converge. In a data sampling-based scheme, we must guarantee the stability of the closed-loop system and its robustness because there is the possibility for a divergence from the stability-robustness feasibility set $\check{\mathcal{K}}$ since the inner loop is computed in a finite number of steps. The problem is stated in Problem 2.

Problem 2 (Sampling-based Policy Iteration): *If A, B, C, D, E, P are all replaced by approximate matrices $\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}, \hat{P}$, under what conditions will the sequences $\{P_{K,L}^{p,q}\}_{(p,q)=1}^{(p,q)=\infty}$, $\{\hat{K}_p\}_{p=0}^{\infty}$, $\{\hat{L}_q\}_{q=1}^{\infty}$ converge to a small neighborhood of the optimal values $P^\star$, $K^\star$, and $L^\star$.*

1) Discrete-Time Nonlinear System Interpretation: From Assumption 1, a $P_K^0 \in \mathbb{S}^n$ exists such that when applied to find K_0 i.e. $K_0 = R^{-1}B^\top P_K^0$, such a K_0 will be stabilizing. Now, factoring in approximation errors between the policy evaluation and improvement structures, we end up with a hybrid system consisting of a continuous-time policy gain pair $(\hat{K}_p, \hat{L}_q(\hat{K}_p))$ and a learning algorithm that is essentially a discrete sampled data from a nonlinear system (owing to errors from various sources lumped together as disturbance input). We will leverage Lemmas 2 and 6 to show that under inexact loop updates, an online PI scheme converges to the optimal solution and closed-loop dynamic stability is guaranteed in an input-to-state stability framework (ISS) [38]. Hence the loops are discrete-time nonlinear systems.

2) Online (Model-Free)Nested Loop Reparameterization: Consider (10b) and suppose that $\hat{P}_K^0 \in \mathbb{S}^n$ is chosen following Assumption 1. It follows that a $\hat{K}_k^1 = R^{-1}B^\top \hat{P}_K^0$ will be stabilizing since $\tilde{K}_k^1 \equiv \hat{K}_k^1 - K_k^1 \triangleq 0$. The same argument applies for L_0. For $(p,q) > 0$, we must show that for $\tilde{K}_k^p \equiv \hat{K}_k^p - K_k^p \triangleq 0$ so that the sequence $\{P_{K,L}^{p,q}\}_{(p,q)=0}^{\infty}$ will converge to the locally exponentially stable $\hat{P}_{K,L}^\star$ going by Lemma 2 and 6.

Under inexact outer loop update, the iterate K_{p+1} becomes inaccurate so that the inexact outer-loop iteration involves the recursions

$$\hat{A}_K^{p\top}\hat{P}_K^p + \hat{P}_K^p\hat{A}_K^p + \hat{Q}_K^p + \gamma^{-2}\hat{P}_K^p DD^\top \hat{P}_K^p = 0, \tag{35a}$$

$$\hat{K}_{p+1} = R^{-1}B^\top \hat{P}_K^p \tag{35b}$$

where $\hat{A}_K^p = A - B\hat{K}_p$ and $\hat{Q}_K^p = Q + \hat{K}_p^\top R\hat{K}_p$. Similar argument applies to the inner loop updates so that the inexact inner loop update is

$$\hat{A}_{K,L}^{p,q\top}\hat{P}_{K,L}^{p,q} + \hat{P}_{K,L}^{p,q}\hat{A}_{K,L}^{p,q} + \hat{Q}_K^p - \gamma^2\hat{L}_q^\top\hat{L}_q(\hat{K}_p) = 0 \tag{36a}$$

$$\hat{K}_{p+1} = R^{-1}B^\top\hat{P}_K^{p,q},\ \hat{L}_{q+1}(\hat{K}_p) = \gamma^{-2}D^\top\hat{P}_{K,L}^{p,q} \tag{36b}$$

Consider the transformation of the infinite-dimensional stochastic differential equation (1) in light of the identities (9) under inexact updates for $(p,q) > 0$

$$\mathrm{d}x = [\hat{A}_{K,L}^{p,q}x + B(\hat{K}_p x - D\hat{L}_q(K_p) + u)]\mathrm{d}t + D\mathrm{d}w. \tag{37}$$

On a time interval $[s, s+\delta s]$, it follows from Itô's stochastic calculus and the Hamilton-Jacobi-Bellman equation that

$$\begin{aligned} &d\left[x^\top(s+\delta s)\hat{P}_{K,L}^{p,q}(s+\delta s) - x^\top(s)\hat{P}_{K,L}^{p,q}x(s)\right] = \\ &(\mathrm{d}x)^\top\hat{P}_{K,L}^{p,q}x + x^\top\hat{P}_{K,L}^{p,q}\mathrm{d}x + (\mathrm{d}x)^\top\hat{P}_{K,L}^{p,q}(\mathrm{d}x). \end{aligned} \tag{38}$$

Along the trajectories of equation (37) and using the gains in (11), the r.h.s. in the foregoing becomes

$$\begin{aligned} &x^\top\left[\hat{A}_{K,L}^{p,q\top}\hat{P}_{K,L}^{p,q} + \hat{P}_{K,L}^{p,q}\hat{A}_{K,L}^{p,q}\right]x\mathrm{d}t + 2x^\top\hat{P}_{K,L}^{p,q}D\mathrm{d}w \qquad (39)\\ &+ 2x^\top\hat{P}_{K,L}^{p,q}B(K_p x - D\hat{L}_q(K_p) + u)dt + Tr(D^\top PD),\\ &= -x^\top\hat{Q}_K^p x\mathrm{d}t - \gamma^{-2}x^\top\hat{P}_{K,L}^{p,q}DD^\top\hat{P}_{K,L}^{p,q}xdt + Tr(D^\top\hat{P}_{K,L}^{p,q}D)\\ &+ 2x^\top\hat{P}_{K,L}^{p,q}B\left[\hat{K}_p x - D\hat{L}_q(K_p) + u\right]\mathrm{d}t + 2x^\top\hat{P}_{K,L}^{p,q}D\mathrm{d}w, \end{aligned}$$

so that $x^\top(s+\delta s)\hat{P}_{K,L}^{p,q}(s+\delta s) - x^\top(s)\hat{P}_{K,L}^{p,q}x(s)$

$$\begin{aligned} &= \int_s^{s+\delta s}\left[(-x^\top\hat{Q}_K^p x - \gamma^2 w^\top w)\mathrm{d}t + 2\gamma^2 x^\top\hat{L}_{q+1}^\top(K_p)\mathrm{d}w\right]\\ &+ \int_s^{s+\delta s} 2x^\top\hat{K}_{p+1}^\top R\left[\hat{K}_p x - D\hat{L}_q(\hat{K}_p) + u\right]\mathrm{d}t\\ &\quad + \int_s^{s+\delta s} Tr(D^\top\hat{P}_{K,L}^{p,q}D)\mathrm{d}t. \end{aligned} \tag{40}$$

Observe: The system dependent matrices $\hat{A}_{K,L}^{p,q}, B, D$ from equation (39) are now replaced by input and state terms including $\hat{Q}_K^p$, $\hat{K}_{p+1}$, and $\hat{L}_{q+1}$ which are all retrievable via online measurements. *We essentially end up with an input-to-state system.* The price we pay is that the noise feedthrough matrix D must be known precisely. In this article, as is common in many linear stochastic system with Brownian motion, D is taken to be identity [39], [40][5].

3) Sampling PI Scheme: Our goal is to explore the system model until exact equality of $\hat{A}_{K,L}^{p,q}, P_{K,L}^{p,q}$ and $K_{p+1}, L_{q+1}(K_p)$ to the corresponding terms in (11). To this end, (40) allows us to explore with the controls $u = -K_0 x + \eta_i$ and $w = -L_0 x + \eta_i$ where η_i is drawn uniformly at random over matrices with a Frobenius norm r for a smoothing

[5] We defer analysis for when D is not an identity matrix to a future contribution.

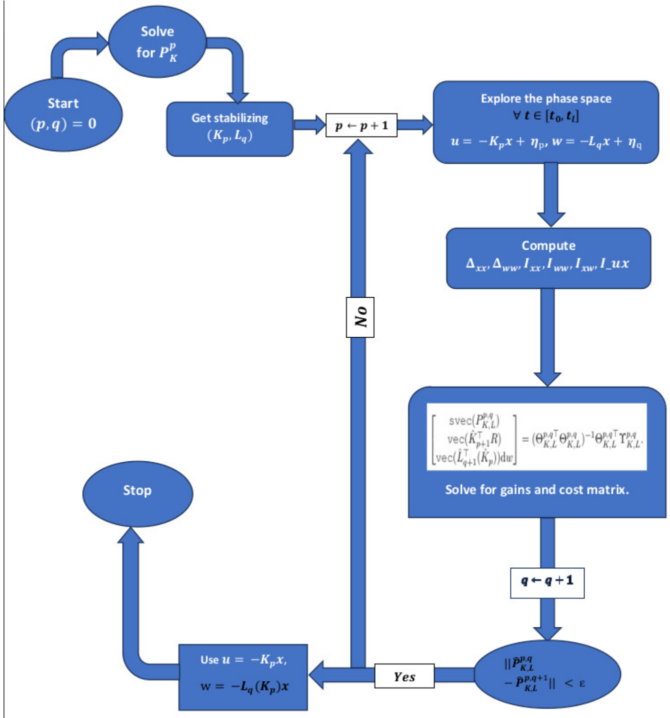

Start
$(p,q)=0$
Solve for P_K^p
Get stabilizing (K_p, L_q)
$p \leftarrow p+1$
Explore the phase space $\forall\, t \in [t_0, t_l]$
$u = -K_p x + \eta_p,\ w = -L_q x + \eta_q$
Compute $\Delta_{xx}, \Delta_{ww}, I_{xx}, I_{ww}, I_{xw}, I\_ux$
$\begin{bmatrix} \mathrm{svec}(P^{p,q}_{K,L}) \\ \mathrm{vec}(\hat{K}^\top_{p+1}R) \\ \mathrm{vec}(\hat{L}^\top_{q+1}(\hat{K}_p))\mathrm{d}w \end{bmatrix} = (\Theta^{p,q\top}_{K,L}\Theta^{p,q}_{K,L})^{-1}\Theta^{p,q\top}_{K,L}\Upsilon^{p,q}_{K,L}.$
Solve for gains and cost matrix.
$q \leftarrow q+1$
$\|\tilde{P}^{p,q}_{K,L} - \tilde{P}^{p,q+1}_{K,L}\| < \varepsilon$
Yes
No
Use $u = -K_p x$, $w = -L_q(K_p)x$
Stop

Fig. 1: Flowchart for Sampling-based Stochastic PO in Continuous-time Mixed $\mathcal{H}_2/\mathcal{H}_\infty$.

parameter r [4], [18]. Let us now introduce the following identities,

$$
\begin{aligned}
&x^\top \hat{Q}^p_K x = (x^\top \otimes x^\top)\,\mathrm{vec}(\hat{Q}^p_K),\\
&\gamma^2 w^\top w = \gamma^2 (w^\top \otimes w^\top)\ \mathrm{vec}(I_v),\\
&2\gamma^2 x^\top \hat{L}^\top_{q+1}(\hat{K}_p)\mathrm{d}w = 2\gamma^2 (I_n \otimes x^\top)\mathrm{d}w\ \mathrm{vec}(\hat{L}^\top_{q+1}(\hat{K}_p)),\\
&2x^\top \hat{K}^\top_{p+1} R\hat{K}_p x = 2(x^\top \otimes x^\top)(I_n \otimes \hat{K}^\top_p)\,\mathrm{vec}(\hat{K}^\top_{p+1}R),\\
&2x^\top \hat{K}^\top_{p+1} RD\hat{L}_q(\hat{K}_p) = 2(\hat{L}^\top_q(\hat{K}_p)D^\top \otimes x^\top)\,\mathrm{vec}(\hat{K}^\top_{p+1}R),\\
&2x^\top \hat{K}^\top_{p+1} Ru = 2(u^\top \otimes x^\top)\,\mathrm{vec}(\hat{K}^\top_{p+1}R),\\
&Tr(D^\top \hat{P}^{p,q}_{K,L} D) = \mathrm{vec}^\top(D)\,\mathrm{vec}(\hat{P}^{p,q}_{K,L} D). \qquad (41)
\end{aligned}
$$

Furthermore, consider the matrices $\delta_{xx} \in \mathbb{R}^{\frac{n(n+1)}{2}l}$, $\delta_{ww} \in \mathbb{R}^{\frac{v(v+1)}{2}l}$, $I_{xx} \in \mathbb{R}^{l\times n^2}$, and $I_{ux} \in \mathbb{R}^{l\times nm}$ for $l \in \mathbb{N}_+$ so that

$$
\begin{aligned}
\Delta_{xx} &= \left[\mathrm{vecv}(x_1), \ldots, \mathrm{vecv}(x_l)\right]^\top,\ x_l = x_{l+1} - x_l,\\
\Delta_{ww} &= \left[\mathrm{vecv}(w_1), \ldots, \mathrm{vecv}(w_l)\right]^\top,\ w_l = w_{l+1} - w_l,\\
I_{xx} &= \left[\int_{s_0}^{s_1} x \otimes x\,\mathrm{d}t, \ldots, \int_{s_{l-1}}^{s_l} x \otimes x\,\mathrm{d}t\right]^\top,\\
I_{ww} &= \left[\int_{s_0}^{s_1} w \otimes w\,\mathrm{d}t, \ldots, \int_{s_{l-1}}^{s_l} w \otimes w\,\mathrm{d}t\right]^\top,\\
I_{xw} &= \left[\int_{s_0}^{s_1} (I_n \otimes x)\mathrm{d}w, \ldots, \int_{s_{l-1}}^{s_l} (I_n \otimes x)\mathrm{d}w\right]^\top,\\
I_{ux} &= \left[\int_{s_0}^{s_1} u \otimes x\,\mathrm{d}t, \ldots, \int_{s_{l-1}}^{s_l} u \otimes x\,\mathrm{d}t\right]^\top. \qquad (42)
\end{aligned}
$$

Next, set

$$
\begin{aligned}
\Theta^{p,q}_{K,L} &= \Big[\Delta_{xx}, -2I_{xx}(I_n \otimes \hat{K}^\top_p) + 2(\hat{L}^\top_q(\hat{K}_p)D^\top \otimes x^\top)\\
&\qquad -2I_{ux}, -2\gamma^2 I_{xw}, -\mathrm{vec}^\top(D)\mathrm{vec}(\hat{P}^{p,q}_{K,L}D)\Big], \qquad (43a)\\
\Upsilon^{p,q}_{K,L} &= \Big[-I_{xx}\mathrm{vec}(\hat{Q}^p_K),\ -\gamma^2 I_{ww}\mathrm{vec}(I_v)\Big]. \qquad (43b)
\end{aligned}
$$

Define $\mathbf{1}_{q^2}$ as one-vector with dimension q^2. Thus,

$$
\Theta^{p,q}_{K,L}\begin{bmatrix} \mathrm{svec}(P^{p,q}_{K,L}) \\ \mathrm{vec}(\hat{K}^\top_{p+1}R) \\ \mathrm{vec}(\hat{L}^\top_{q+1}(\hat{K}_p)) \\ \mathbf{1}_{q^2} \end{bmatrix} = \Upsilon^{p,q}_{K,L}. \qquad (44)
$$

Suppose that $\Theta^{p,q}_{K,L}$ is of full rank, then we can retrieve the unknown matrices via least squares estimation i.e.

$$
\begin{bmatrix} \mathrm{svec}(P^{p,q}_{K,L}) \\ \mathrm{vec}(\hat{K}^\top_{p+1}R) \\ \mathrm{vec}(\hat{L}^\top_{q+1}(\hat{K}_p))\mathrm{d}w \end{bmatrix} = (\Theta^{p,q\top}_{K,L}\Theta^{p,q}_{K,L})^{-1}\Theta^{p,q\top}_{K,L}\Upsilon^{p,q}_{K,L}. \qquad (45)
$$

We thus end up with a scheme for retrieving the system matrices provided that the algorithm is robust to perturbations upon iterating through (45) for each (p,q). We next state the condition under which $\Theta^{p,q}_{K,L}$ is of full rank.

Lemma 9 : *[41, Lemma 6] If there exists an integer $l_0 > 0$ such that for all $l \geq l_0$, rank$(I_{xx}, I_{ux}, I_{xw}, \mathbf{1}_{q^2}) = n(n+1) + mn + nq + q^2$, then $\Theta^{p,q}_{K,L}$ has full rank for all $(p,q) \in (\bar{p}, \bar{q})$.*

Remark 5 : *Lemma 9 allows the convergence assurance of Fig. 1 under the condition that the rank condition be fulfilled.*

4) Robustness of Minimizing Controller to Perturbations: We now analyze the robustness of the sampling-based scheme as a hybrid nonlinear discrete time system gains with continuous-time dynamics. Let $\tilde{P} = P_K - \hat{P}_K$ and $\tilde{K} = K - \hat{K}$ denote errors arising from the inexact updates.

Lemma 10 (Outer-Loop Robustness to Perturbations): *For any $K \in \mathcal{K}$, there exists an $e(K) > 0$ such that for a perturbation $\tilde{K}$, $K + \tilde{K} \in \mathcal{K}$, as long as $\|\tilde{K}\| < e(K)$.*

Proof of Lemma 10. Let $F(\tilde{P}, \tilde{K})$ be

$$
\begin{aligned}
&(A_K + \gamma^{-2}DD^\top P_K)^\top \tilde{P} + \tilde{P}(A_K + \gamma^{-2}DD^\top P_K)\\
&- \tilde{K}^\top B^\top (P_K + \tilde{P}) - (P_K + \tilde{P})B\tilde{K} + \tilde{K}^\top RK + K^\top R\tilde{K}\\
&+ \tilde{K}^\top R\tilde{K} + \gamma^{-2}\tilde{P}DD^\top \tilde{P}. \qquad (46)
\end{aligned}
$$

Observe: the pair $(P_K + \tilde{P}, K + \tilde{K})$ satisfies (19) iff $F(\tilde{P}, \tilde{K}) = 0$ and that $F(\tilde{P}, \tilde{K}) = 0$ implies an implicit function of $\tilde{P}$ with respect to $\tilde{K}$ since if $\tilde{P} \in \mathbb{S}^n$ exists, $\tilde{K}$ must exist under controllability and observability assumptions. Let $\mathcal{F}(\tilde{P}, \tilde{K}) = \mathrm{vec}(F(\tilde{P}, \tilde{K}))$ so that,

$$
\begin{aligned}
\mathcal{F}(\tilde{P}, \tilde{K}) &= \big[I_n \otimes (A_K + \gamma^{-2}DD^\top P_K)^\top\\
&+ (A_K + \gamma^{-2}DD^\top P_K)^\top \otimes I_n\big]\,\mathrm{vec}(\tilde{P})\\
&- (P_K B \otimes I_n)\mathrm{vec}(\tilde{K}^\top) - (I_n \otimes P_K B)\mathrm{vec}(\tilde{K})\\
&- (I_n \otimes \tilde{K}^\top B^\top + \tilde{K}^\top B^\top \otimes I_n)\mathrm{vec}(\tilde{P})\\
&+ (K^\top R \otimes I_n)\mathrm{vec}(\tilde{K}^\top) + (I_n \otimes K^\top R)\mathrm{vec}(\tilde{K})\\
&+ \mathrm{vec}(\tilde{K}^\top R\tilde{K}) + \gamma^{-2}\mathrm{vec}(\tilde{P}DD^\top \tilde{P}). \qquad (47)
\end{aligned}
$$

Thus,

$$\begin{aligned}\frac{\partial \mathcal{F}(\tilde{P},\tilde{K})}{\partial \mathtt{vec}(\tilde{P})} &= I_n \otimes [(A_K+\gamma^{-2}DD^\top P_K)-B\tilde{K}]^\top \\ &+ [(A_K+\gamma^{-2}DD^\top P_K)-B\Delta K]^\top \otimes I_n \\ &+ \tilde{P}DD^\top \otimes I_n + I_n \otimes \tilde{P}DD^\top \\ &- (P_K B\otimes I_n)\mathtt{vec}(\tilde{K}^\top) - (I_n \otimes P_K B)\mathtt{vec}(\tilde{K}) \\ &+ (K^\top R\otimes I_n)\mathtt{vec}(\tilde{K}^\top) + (I_n\otimes K^\top R)\mathtt{vec}(\tilde{K}),\end{aligned} \tag{48}$$

where we have used [42, Theorem 9], to obtain $\partial\mathtt{vec}(\tilde{P}DD^\top\tilde{P})/\partial\mathtt{vec}(\tilde{P}) = \tilde{P}DD^\top\otimes I_n + I_n\otimes\tilde{P}DD^\top$. Since $\mathcal{F}(0,0)=0$, $(A_K+\gamma^{-2}DD^\top P_K)$ is Hurwitz, hence $\partial\mathcal{F}(\tilde{P},\tilde{K})/\partial\mathtt{vec}(\tilde{P})|_{\tilde{P}=0,\tilde{K}=0}$ is invertible. From the implicit function theorem, there must exist an $e_1(K)>0$, such that $\tilde{P}$ is continuously differentiable with respect to $\tilde{K}$ for any $\tilde{K}\in\mathcal{B}_{e_1(K)}(0)$. Thus, $\|\tilde{P}\|\to 0$ as $\|\tilde{K}\|\to 0$. Since $K\in\mathcal{K}$ by [10, Lemma A.1], we must have $P_K\succ 0$. Therefore, there exists $e(K)>0$, such that $\sigma_{\max}(\tilde{P})<\sigma_{\min}(P_K)$, i.e. $P_K-\tilde{P}\succ 0$, as long as $\|\tilde{K}\|<e(K)$.

Since $\tilde{P}$ and $\tilde{K}$ satisfy $F(\tilde{P},\tilde{K})=0$, we have

$$\begin{aligned}&(A-BK-B\tilde{K})^T(P_K+\tilde{P})+(P_K+\tilde{P})+Q+ \\ &(K+\tilde{K})^T R(K+\tilde{K})+\gamma^{-2}(P_K+\tilde{P})DD^T(P_K+\tilde{P})=0\end{aligned} \tag{49}$$

Since $P_K+\tilde{P}\succ 0$ when $\|\tilde{K}\|<e(K)$, by [10, Lemma A.1], $K+\tilde{K}\in\mathcal{K}$. That is, as long as $\tilde{K}$ is small, if we start the PI with a robustly stabilizing $K\in\mathcal{K}$, we can guarantee the feasibility of the iterates. □

Lemma 11 : *For any $h>0$ and $K\in\mathcal{K}_h$, let $K'=R^{-1}B^\top P_K$, where P_K is the solution of (19), and $\hat{K}'=K'+\tilde{K}$. Then, there exists $f(h)>0$, such that $\hat{K}'\in\mathcal{K}_h$ as long as $\|\tilde{K}\|<f(h)$.*

Proof. Since $\mathcal{K}_h$ is compact, it follows from Lemma 10 that $\underline{e}(h) := \inf_{K\in\mathcal{K}_h} e(K) > 0$. In addition, $\hat{K}'\in\mathcal{K}$ when $\|\tilde{K}\|<\underline{e}(h)$. By [10, Lemma A.1], $P_{\hat{K}'}=P_{\hat{K}'}^\top\succ 0$ is the solution of

$$A_{\hat{K}'}^\top P_{\hat{K}'} + P_{\hat{K}'}A_{\hat{K}'} + Q_{\hat{K}'} + \gamma^{-2}P_{\hat{K}'}DD^\top P_{\hat{K}'} = 0, \tag{50}$$

where $A_{\hat{K}'}=A-B\hat{K}'$ and $Q_{\hat{K}'}=Q+(\hat{K}')^\top R\hat{K}'$. Let $A^\star_{\hat{K}'}=A-B\hat{K}'+\gamma^{-2}DD^TP_{\hat{K}'}$. It follows from [10, Lemma A.1] that $A^\star_{\hat{K}'}$ is Hurwitz. Subtracting (50) from (19), using $K'=R^{-1}B^TP_K$, and completing the squares,

$$\begin{aligned}&A^{\star\top}_{\hat{K}'}(P_K-P_{\hat{K}'})+(P_K-P_{\hat{K}'})A^\star_{\hat{K}'} \\ &+(K'-K)^\top R(K'-K)-\tilde{K}^\top R\tilde{K} \\ &+\gamma^{-2}(P_K-P_{\hat{K}'})DD^\top(P_K-P_{\hat{K}'})=0.\end{aligned} \tag{51}$$

From Lemma 13, we have $(P_K-P_{\hat{K}'})\succeq$

$$\int_0^\infty e^{A^{\star\top}_{\hat{K}'}t}E_K e^{A^\star_{\hat{K}'}t}\mathrm{d}t - \int_0^\infty e^{A^{\star\top}_{\hat{K}'}t}\tilde{K}^TR\tilde{K}e^{A^\star_{\hat{K}'}t}\mathrm{d}t, \tag{52}$$

so that taking the trace, using Lemma 3 and [43, Theorem 2],

$$\begin{aligned}Tr(P_K-P_{\hat{K}'}) &\geq \frac{\log(5/4)}{2\|A^\star_{\hat{K}'}\|}\|E_K\| \\ &- Tr\left(\int_0^\infty e^{A^{\star\top}_{\hat{K}'}t}e^{A^\star_{\hat{K}'}t}\mathrm{d}t\right)\|R\|\|\tilde{K}\|^2.\end{aligned} \tag{53}$$

It follows from Lemmas 5 and 15 that

$$\begin{aligned}Tr(P_{\hat{K}'}-P^\star) &\leq \left(1-\underbrace{\frac{\log(5/4)b(h)}{2n\|A^\star_{\hat{K}'}\|}}_{f_1(\hat{K}')}\right)Tr(P_K-P^\star) \\ &+\underbrace{Tr\left(\int_0^\infty e^{A^{\star\top}_{\hat{K}'}t}e^{A^\star_{\hat{K}'}t}\mathrm{d}t\right)}_{f_2(\hat{K}')}\|R\|\|\tilde{K}\|^2.\end{aligned} \tag{54}$$

Since $f_1(\hat{K}')$ and $f_2(\hat{K}')$ are continuous with respect to $\hat{K}'$,

$$\underline{f}_1(h) = \inf_{\hat{K}'\in\mathcal{K}_h} f_1(\hat{K}')>0, \bar{f}_2(h) = \sup_{\hat{K}'\in\mathcal{K}_h} f_2(\hat{K}')<\infty. \tag{55}$$

It follows from (54) that if $\|\tilde{K}\|<\sqrt{\frac{\underline{f}_1(h)h}{\bar{f}_2(h)\|R\|}}$, then $Tr(P_{\hat{K}'}-P^\star)<h$. In summary, if

$$\|\tilde{K}\| < \min\left\{\underline{e}(h), \sqrt{\frac{\underline{f}_1(h)h}{\bar{f}_2(h)\|R\|}}\right\} =: f(h), \tag{56}$$

we have $\hat{K}'\in\mathcal{K}_h$. □

Theorem 3 : *The inexact outer loop is small-disturbance ISS. That is, for any $h>0$ and $\hat{K}_0\in\mathcal{K}_h$, if $\|\tilde{K}\|<f(h)$, there exist a $\mathcal{KL}$-function $\beta_1(\cdot,\cdot)$ and a $\mathcal{K}_\infty$-function $\gamma_1(\cdot)$ such that*

$$\|P^p_{\hat{K}}-P^\star\| \leq \beta_1(\|P^0_{\hat{K}}-P^*\|,p)+\gamma_1(\|\tilde{K}\|). \tag{57}$$

Proof. From Lemma 11, $\hat{K}^p_K\in\mathcal{K}_h$ for any $p\in\mathbb{N}_+$. From (54), at the p’th iteration, we have

$$\begin{aligned}&Tr(P^p_{\hat{K}}-P^\star)\leq(1-\underline{f}_1(h))Tr(P^{p-1}_{\hat{K}}-P^\star) \\ &+\bar{f}_2(h)\|R\|\|\tilde{K}^p_K\|^2.\end{aligned} \tag{58}$$

Repeating (58) for $p, p-1, \cdots, 1$,

$$Tr[P^p_{\hat{K}}-P^\star] \leq (1-\underline{f}_1)^p Tr(P^1_{\hat{K}}-P^\star) + \frac{\bar{f}_2\|R\|\|\tilde{K}\|^2_\infty}{\underline{f}_1(h)}. \tag{59}$$

It follows from (59) and [43, Theorem 2] that

$$\|P^p_{\hat{K}}-P^\star\|_F \leq (1-\underline{f}_1)^p\sqrt{n}\|P^1_{\hat{K}}-P^\star\|_F + \frac{\bar{f}_2\|R\|\|\tilde{K}\|^2_\infty}{\underline{f}_1}. \tag{60}$$

As $p\to\infty$, $P^p_{\hat{K}}\to P^\star$. The radius of the neighbor of $P^\star$ is proportional to $\|\tilde{K}\|^2_\infty$. Thus, the proof follows. □

5) Robustness of Maximizing Controller to Perturbations: The perturbed inner-loop iteration (36) has inexact matrix $\hat{A}^{p,q}_{K,L}$, and sequences $\{\hat{L}_{q+1}(K_p)\}^\infty_{q=0}$, and $\{\hat{P}^{p,q}_{K,L}\}^\infty_{q=0}$. We next analyze its robustness to perturbations when it differs from the exact loop matrices and sequences.

Lemma 12 (Inner-Loop Robustness to Perturbations): *Given $K\in\mathcal{K}$, there exists a $g\in\mathbb{R}_+$, such that if $\|\tilde{L}_{q+1}(K_p)\|_F\leq g$, $\hat{A}^{p,q}_{K,L}$ is Hurwitz for all $q\in\mathbb{N}_+$.*

Proof. Define $\tilde{L}_q^\top(K_p) = L_q^\top(K_p) - \hat{L}_q^\top(K_p)$ and $\tilde{P}_{K,L}^{p,q} = P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q}$. Further, assume $\hat{A}_{K,L}^{p,q}$ is Hurwitz. From (10),

$$\hat{A}_{K,L}^{p,q\top} P_{K,L}^{p,q} + P_{K,L}^{p,q}\hat{A}_{K,L}^{p,q+1} + Q_K - \gamma^{-2}\hat{P}_{K,L}^{p,q}DD^\top\hat{P}_{K,L}^{p,q} + \gamma^{-2}(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q})DD^\top(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q}) - \tilde{L}_q^\top(K_p)D^\top P_{K,L}^{p,q} - P_{K,L}^{p,q}D\tilde{L}_q(K_p) = 0. \tag{61}$$

Set $\|\tilde{L}_q^\top(K_p)\| < \sigma_{\min}(Q_K - \gamma^{-2}P_{K,L}^{p,q}DD^\top P_{K,L}^{p,q})/2\|D^\top P_{K,L}^{p,q}\| \triangleq e$, it follows from (26) that $Q \succ \gamma^2 L_{q+1}^\top(K_p)L_{q+1}(K_p)$ by reason of it being admissible as a Lyapunov equation and the inequality $P_{K,L}^{p,q} \succeq \hat{P}_{K,L}^{p,q}$ that

$$-\gamma^{-2}\hat{P}_{K,L}^{p,q}DD^\top\hat{P}_{K,L}^{p,q} + \gamma^{-2}(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q})DD^\top(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q}) - (\tilde{L}_K^j)^\top D^\top P_{K,L}^{p,q} - P_{K,L}^{p,q}D\tilde{L}_q(K_p) + Q_K \succ 0.$$

Consequently, $\hat{A}_{K,L}^{p,q+1}$ is Hurwitz. Since $\hat{L}_q(K_0) = 0$ and $K \in \mathcal{K}$, $\hat{A}_{K,L}^{p,0} = A - BK$ is Hurwitz. Hence, $\hat{A}_{K,L}^{p,q}$ is Hurwitz for all $q \in \mathbb{N}_+$ as long as $\|\tilde{L}_q(K_p)\|_F \leq e$. □

Theorem 4 : *Assume $\|\tilde{L}_q(K_p)\| < e$ for all $q \in \mathbb{N}_+$. There exists $\hat{\beta}(K) \in [0,1)$, and $\lambda(\cdot) \in \mathcal{K}_\infty$, such that*

$$\|\hat{P}_{K,L}^{p,q} - P_{K,L}^{p,q}\|_F \leq \hat{\beta}^{j-1}(K)Tr(P_{K,L}^{p,q}) + \lambda(\|\tilde{L}\|_\infty). \tag{62}$$

Proof. When $\|\tilde{L}_q(K_p)\| < e$, we have an Hurwitz $\hat{A}_{K,L}^{p,q}$ going by 12. Rewriting (36) for the $(p+1)$th iteration and subtracting it from (36), we have

$$\hat{A}_{K,L}^{(p,q+1)\top}(\hat{P}_{K,L}^{(p,q+1)} - \hat{P}_{K,L}^{p,q}) + (\hat{P}_{K,L}^{(p,q+1)} - \hat{P}_{K,L}^{p,q})\hat{A}_{K,L}^{(p,q+1)\top} + \gamma^{-2}(\gamma^2\hat{L}_q(K_p) - D^\top\hat{P}_{K,L}^{p,q})^\top(\gamma^2\hat{L}_q(K_p) - D^\top\hat{P}_{K,L}^{p,q}) - \gamma^2\tilde{L}_q^\top(K_p)\tilde{L}_q(K_p) = 0. \tag{63}$$

Suppose that $\hat{E}_{K,L}^{p,q} = \gamma^{-2}(\gamma^2\hat{L}_q(K_p) - D^\top\hat{P}_{K,L}^{p,q})^\top(\gamma^2\hat{L}_q(K_p) - D^\top\hat{P}_K^j)$. It follows that since $\hat{A}_{K,L}^{p,q+1}$ is Hurwitz, $\hat{P}_{K,L}^{(p,q+1)} - \hat{P}_{K,L}^{p,q}$ becomes

$$\int_0^\infty e^{\hat{A}_{K,L}^{(p,q+1)\top}t}\left[\hat{E}_{K,L}^{p,q} - \gamma^2\tilde{L}_q^\top(K_p)\tilde{L}_q^\top(K_p)\right]e^{\hat{A}_{K,L}^{(p,q+1)}t}dt. \tag{64}$$

Now let $\hat{F}_K^q = \int_0^\infty e^{\hat{A}_{K,L}^{(p,q+1)\top}t}\hat{E}_{K,L}^{p,q}e^{\hat{A}_{K,L}^{(p,q+1)}t}dt$ so that

$$P_{K,L}^{p,q+1} - \hat{P}_{K,L}^{p,q+1} = P_{K,L} - \hat{P}_{K,L}^{p} - \hat{F}_K^q + \int_0^\infty e^{\hat{A}_{K,L}^{(p,q+1)\top}t}\left(\gamma^2\tilde{L}_q^\top(K_p)\tilde{L}_q(K_p)\right)e^{\hat{A}_{K,L}^{p,q+1}t}dt. \tag{65}$$

Let $f_K = \sup_{q\in\mathbb{N}_+}\|\hat{A}_{K,L}^{p,q+1}\|$. From Lemma 7, we can write $-\|\hat{F}_K^q\| \leq -\frac{\log(5/4)}{2f_K}\|\hat{E}_{K,L}^{p,q}\|$. Furthermore, by Lemma 7, we can write $-\|\hat{E}_{K,L}^{p,q}\| \leq -\frac{1}{c(K)}Tr(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q})$, where $c(K) = Tr(\int_0^\infty e^{(A_K+DL_q(K_p))t}e^{(A_K+DL_q(K_p))^\top t}dt)$. Therefore, the trace of (65) becomes

$$Tr(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q+1}) \leq \left(1 - \frac{\log(5/4)}{2f_Kc(K)}\right)Tr(P_K - \hat{P}_{K,L}^{p,q}) + Tr\left(\int_0^\infty e^{(\hat{A}_{K,L}^{p,q+1})t}e^{(\hat{A}_{K,L}^{p,q+1})^T t}dt\right)\gamma^2\|\tilde{L}_q(K_p)\|^2. \tag{66}$$

$$\text{Let } g = \sup_{q\in\mathbb{N}_+} Tr\left(\int_0^\infty e^{(\hat{A}_{K,L}^{p,q+1})t}e^{(\hat{A}_{K,L}^{p,q+1})^\top t}dt\right), \tag{67}$$

and $\hat{\beta}(K) = 1 - \frac{\log(5/4)}{2f_Kc(K)}$, so that

$$Tr(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q}) \leq \hat{\beta}^{j-1}(K)Tr(P_{K,L}^{p,q}) + \lambda(\|\tilde{L}\|_\infty), \tag{68}$$

where $\lambda(\|\tilde{L}\|_\infty) := \frac{1}{1-\hat{\beta}(K)}\gamma^2 g\|\tilde{L}\|_\infty^2$. As $\|P_K - \hat{P}_{K,L}^{p,q}\|_F \leq Tr(P_{K,L}^{p,q} - \hat{P}_{K,L}^{p,q})$, we establish the theorem. □

From Theorem 4, as $j \to \infty$, $\hat{P}_{K,L}^{p,q}$ approaches the solution P_K and enters the ball centered by $P_{K,L}^{p,q}$. The radius of ball is proportional to $\|\tilde{L}\|_\infty$. Hence, the proposed inner-loop iterative algorithm approximates $P_{K,L}^{p,q}$ well despite the perturbation.

IV. Numerical Experiments

We consider a humanoid robot model [44], [45] as a three-link kinematic chain. The system is non-minimum phase, underactuated, and possesses badly damped poles. In addition, owing to the passive joint, there exists inherent (Wiener process) noise that additively perturbs the system dynamics.

This model has three states: two upper hinge (the hip and knee) *actuated joints* and a lower hinge (the ankle) passive joint. The state's velocity dynamics 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 respectively. The linearized model of the triple inverted pendulum admits a form of the infinite dimensional linear PDE in (1), where $A \in \mathbb{R}^{6\times 6}$ and $B \in \mathbb{R}^{6\times 2}$(see [46, Section 3]), and $D = \left[0_{3\times 3}, I_3\right]^\top$. We impose an $\mathcal{H}_\infty$ norm bound of $\gamma = 5$ on the robot and set the initial state to $x(0) = [0, -5, 10, 10, -10, 10]^\top$ and set $C = \left[I_6, 0_{2\times 6}\right]^\top$, $E = \left[0_{6\times 2}, I_2\right]^\top$. Throughout, $w(t)$ is set to a Wiener process such that dw is drawn from $\mathcal{N}(0, I)$ and we chose a time step size, $dt = 0.0001$ for the numerical integration.

In addition, we consider a double pendulum and compare the efficacy of the algorithms we have presented thus far against NPG. In what follows, we report various findings when running the model-based and model-free algorithm versus the natural policy gradient algorithm (NPG) [47], which shares similar character with our PI-based PO scheme. For other numerical experiment reports, we refer readers to our recent conference paper [35].

A. Model-based Mixed Design vs. NPG

Let us describe numerical experiments on the algorithms described so far. At each iteration, $\tilde{K}_p$ is sampled from a uniform Gaussian distribution whose Frobenius norm is 0.15. We chose

$$\hat{K}_0 = \begin{bmatrix} -203.3 & -74.2 & -31.4 & -67.7 & -28.4 & -16.5 \\ -529.5 & -198.8 & -77.8 & -175.5 & -78.7 & -39.0 \end{bmatrix}.$$

The results are shown in Figures 2 and 3. The robust mixed design PI scheme approaches the optimal solution after the 5'th iteration (See Fig. 2). At the last iteration, the deviation from the optimal cost matrix[6] is 2.9%, while the gain error[7] is 2.6%. In contrast, NPG exhibits cost matrix and controller gain errors that are unbounded as the iteration lengthens.

[6] Calculated as $\|\hat{P}_K^{20} - P^\star\|/\|P^\star\|_F$.

[7] Calculated as $\|\hat{K}_K^{20} - K^\star\|_F/\|K^\star\|_F$.

TABLE I: Computational Time of Alg. 1 vs. NPG.

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

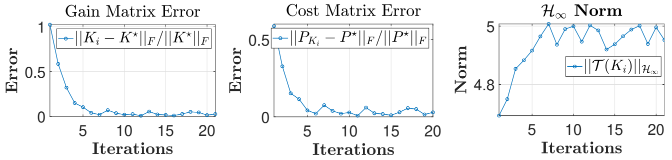


Fig. 2: Alg. 1 with $\|\tilde{K}\|_\infty = 0.15$.

We compared the time it takes to compute the optimal policies in Alg. 1 against NPG in Table I. We see 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.

B. Sampling-based Mixed Design vs. NPG

For the parameters of the algorithm in Fig. 1, we set $\bar{p} = 20$ and $\bar{q} = 30$, and found the maximum data collection time before attaining the full rank condition of Lemma 9 to be $t_l = 1500s$. The parameters of the A and B matrices are unknown but the initial controller $\hat{K}_1 \in \mathcal{K}$ is searched for following [35, Alg. 1].

We run the algorithm in Fig. 1 on (1). As seen in Fig. 4, the controller $\hat{K}_p$ found at each iteration converges after 5 iterations alongside $\hat{P}_{K_i}$ 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.

Appendix A: Proofs to Lemmas

In this appendix, we introduce a series of lemmas to guide our problem description and proposed solution.

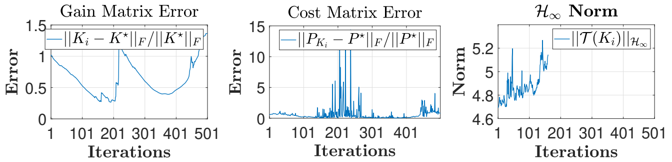


Fig. 3: Model-based Alg. 1 vs. NPG with $\|\tilde{K}\|_\infty = 0.1$.

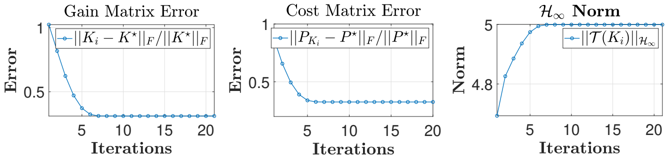


Fig. 4: Sampling-based Scheme Results.

Lemma 13 : *Assume $A \in \mathbb{R}^{n\times n}$ is Hurwitz and satisfies $A^TP + PA + Q = 0$. Then, the following properties hold*

1) *$P = \int_0^\infty e^{A^Tt}Qe^{At}\mathrm{d}t$;*
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) *For $P' \in \mathbb{S}^n$ satisfying $A^TP' + P'A + Q' = 0$, where $Q' \preceq Q$, we have $P' \preceq P$.*

Lemma 14 : *[48, Lemma 3.19] Suppose that P satisfies $A^TP + PA + Q = 0$, then the following statements hold:*

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 of Lemma 13. The first three statements are proven in [48, Lemma 3.18]. Consequently, P' can be expressed as

$$P' = \int_0^\infty e^{A^Tt}Q'e^{At}\mathrm{d}t. \tag{A.1}$$

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

Proof of Lemma 4. We give the proof to Lemma 4. Since $R \succ 0$, $E_K = 0$ implies $K = K'$. Therefore at $E_K = 0$, we must have $K = K'$ which implies that $P_K = P'_K$. If $K = K'$ and $P_K = P'_K$, it suffices to conclude that $K' = K \equiv K^\star$ where $K^\star = R^{-1}B^\top P^\star$. Hence, $E_K = 0$ is tantamount to $P_K = P^\star$ and $K = K^\star$. □

Lemma 15 (Norm of a Matrix Trace): *For any positive semi-definite matrix $P \in \mathbb{S}^n$, $\|P\|_F \leq Tr(P) \leq \sqrt{n}\|P\|_F$, and $\|P\| \leq Tr(P) \leq n\|P\|$. For any $x \in \mathbb{R}^n$, $x^TPx \geq \underline{\lambda}(P)\|x\|^2$.*

Proof of Lemma 15. We now give the proof of Lemma 15. 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, $\|P\|_F = \sqrt{\sum_{i=1}^n \sigma_i^2}$, $Tr(P) = \sum_{i=1}^n \sigma_i$, and $\|P\| = \sigma_1(P)$. Since $\sum_{i=1}^n \sigma_i^2 \leq (\sum_{i=1}^n \sigma_i)^2 \leq n\sum_{i=1}^n \sigma_i^2$, we have $\|P\|_F \leq Tr(P) \leq \sqrt{n}\|P\|_F$, and $\|P\| \leq Tr(P) \leq n\|P\|$. Using Rayleigh's theorem [49, Theorem 4.2.2], we have $x^TPx \geq \underline{\lambda}(P)\|x\|^2$. □

Proof of Lemma 5. Define $A^\star = A - BR^{-1}B^TP^\star + \gamma^{-2}DD^TP^\star$ so that (19) becomes

$$\begin{aligned} &A^{\star\top}P_K + P_KA^\star + Q_K + (K^\star - K)^\top RK' \\ &+ K'^\top R(K^\star - K) - \gamma^{-2}P^\star DD^\top P_K - \gamma^{-2}P_KDD^\top P^\star \\ &+ \gamma^{-2}P_KDD^\top P_K = 0, \end{aligned} \tag{A.2}$$

In addition, (5) can be rewritten (replacing α with γ)as

$$\begin{aligned} &A^{\star\top}P^\star + P^\star A^\star + Q + K^{\star\top}RK^\star \\ &- \gamma^{-2}P^\star DD^\top P^\star = 0. \end{aligned} \tag{A.3}$$

Subtracting (A.3) from (A.2) and completing squares, we have

$$\begin{aligned} &A^{\star\top}(P_K - P^\star) + (P_K - P^\star)A^\star + E_K \\ &+ \gamma^{-2}(P_K - P^\star)DD^\top(P_K - P^\star) \\ &- (K' - K^\star)^\top R(K' - K^\star) = 0. \end{aligned} \tag{A.4}$$

Let $\Delta P_K := P_K - P^\star$. It follows from $K^\star = R^{-1}B^\top P^\star$ and (A.4) that

$$\begin{aligned} A^{\star\top}\Delta P_K + \Delta P_K(A - BR^{-1}B^\top P_K + \gamma^{-2}DD^\top P_K) \\ + E_K = 0, \end{aligned} \tag{A.5}$$

whereupon $\mathcal{A}(K)\mathtt{vect}(\Delta P_K) = -\mathtt{vect}(E_K)$ with $\mathcal{A}(K)$ being

$$\begin{aligned} \mathcal{A}(K) := \big\{&(I_n \otimes A^{\star\top}) \\ &+[(A - BR^{-1}B^\top P_K + \gamma^{-2}DD^\top P_K)^\top \otimes I_n]\big\}\,. \end{aligned} \tag{A.6}$$

From (19) and the implicit function theorem, P_K is a continuously differentiable function of $K \in \mathcal{K}$. Since $A^\star$ is Hurwitz, there exists a ball $\mathcal{B}(K^\star, \delta) := \{K \in \mathcal{K} | \|K - K^\star\|_F \leq \delta\}$, such that $\mathcal{A}(K)$ is invertible for any $K \in \mathcal{K}_h \cap \mathcal{B}(K^\star, \delta)$. Therefore, for any $K \in \mathcal{K}_h \cap \mathcal{B}(K^\star, \delta)$, it follows that

$$\|\Delta P_K\|_F \leq \underline{\sigma}^{-1}(\mathcal{A}(K))\|E_K\|_F. \tag{A.7}$$

On the other hand, for any $K \in \mathcal{K}_h \cap \mathcal{B}^c(K^\star, \delta)$, where $\mathcal{B}^c$ is the complement of $\mathcal{B}$, $E_K \neq 0$, and there exists a constant $b_1 > 0$, such that $\|E_K\|_F \geq b_1$. Thus, by Lemma 15, we have

$$\|\Delta P_K\|_F \leq Tr(P_K) \leq \frac{h + Tr(P^\star)}{b_1}\|E_K\|_F. \tag{A.8}$$

Suppose that $b_2 := \max_{K \in \mathcal{K}_h \cap \mathcal{B}(K^\star, \delta)} \underline{\sigma}^{-1}(\mathcal{A}(K))$ and $b(h) := \max\{b_2, \frac{h+Tr(P^\star)}{b_1}\}$, then the proof follows from (A.7) and the foregoing. □

Proof of Lemma 7. Subtracting (28) from (19), and using $L(K^\star) = \gamma^{-2}D^\top P_K$, we find that

$$\begin{aligned} (A_K + DL(K^\star))^T(P_K - P_K^L) + (P_K - P_K^L)(A_K + \\ DL(K^\star)) + E_K^L - \gamma^{-2}(P_K^L - P_K)DD^\top(P_K^L - P_K) = 0. \end{aligned} \tag{A.9}$$

Since $A_K + DL(K^\star)$ is Hurwitz, it follows from [48, Lemma 3.18] that we must have

$$P_K - P_K^L \preceq \int_0^\infty e^{(A_K + DL(K^\star))^\top t} E_K^L e^{(A_K + DL(K^\star))t} \mathrm{d}t. \tag{A.10}$$

Taking the trace of the lhs, using [43, Theorem 2], and the cyclic property of the trace, the proof follows. □

Lemma 16 : *For $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times p}$, $\|XY\|_F \leq \|X\|\|Y\|_F$.*

Proof. Let $Y = [y_1, \cdots, y_p]$, then it follows that $XY = [Xy_1, \cdots, Xy_p]$. 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 $\|Xy_i\|^2 \leq \|X\|^2\|y_i\|^2$. Hence, $\|XY\|_F^2 \leq \|X\|^2 \sum_{i=1}^p \|y_i\|^2 = \|X\|\|Y\|_F$, which is in fact a proof of the lemma. □

Lemma 17 : *[50, Theorem 16.14] and [51, Theorem 1.5.9] Let $\{T_t : t \geq 0\}$ be a semi-group of measure preserving transformations on a probability space $(\Omega, \mathcal{F}, P)$. Define the time averages as $A_t f = \frac{1}{t}\int_0^t f(T^s w)\mathrm{d}s$. Then, for any $f \in L^1(\Omega, \mathcal{F}, P)$, there exists $\bar{f} \in L^1(\Omega, \mathcal{F}, P)$ such that $\lim_{t\to\infty} A_t f = \bar{f}$ almost surely, where $\bar{f} = \mathbb{E}(f)$.*

Lemma 18 : *[52, p. 530] Let $a(x(t))$ be a vector function such that $\mathbb{E}[a^T(x(t))a(x(t))] < \infty$ where $\{x(t)|t \geq 0\}$ is the solution of the process $\mathrm{d}x(t) = \mu(x(t))\mathrm{d}t + \mathrm{d}w$, and w is a standard independent Brownian motion. The relation $\lim_{t_f\to\infty} \frac{1}{t_f}\int_0^{t_f} a(x(t))\mathrm{d}w = 0$. holds almost surely*

Lemma 19 (Bounded Real Lemma, [20], [53]): *Under Assumption 1, for a stabilizing gain K, the following conditions are equivalent*

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

$$\begin{aligned} A_K^\top P_K + P_K A_K + C^\top C + K^\top RK + \\ \gamma^{-2}P_K DD^\top P_K = 0 \end{aligned} \tag{A.11}$$

  admits a unique positive definite solution $P_K \succeq 0$ for a Hurwitz matrix $(A_K + \gamma^{-2}DD^\top P_K)$;
- *There exists $P_K \succ 0$ such that*

$$A_K^\top P_K + P_K A_K + Q + K^\top RK + \gamma^{-2}P_K DD^\top P_K \prec 0. \tag{A.12}$$

References

[1] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International conference on machine learning*, pp. 1467–1476, PMLR, 2018. 1

[2] 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. 1

[3] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 123–158, 2023. 1, 2

[4] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2021. 1, 2, 8

[5] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd international conference on artificial intelligence and statistics*, pp. 2916–2925, PMLR, 2019. 1

[6] K. Zhang, X. Zhang, B. Hu, and T. Basar, "Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2949–2964, 2021. 1, 3

[7] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 1. Athena scientific, 2012. 1

[8] 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. 1, 2, 4

[9] K. Zhang, A. Koppel, H. Zhu, and T. Başar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020. 1

[10] 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, 5, 9

[11] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016. 1

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015. 1

[13] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019. 1

[14] "A Survey on Policy Search for Robotics," *Foundations and Trends in Robotics*, vol. 2, no. 1, pp. 1–142, 2011. 1

[15] 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. 1

[16] B. D. Anderson and J. B. Moore, *Optimal control: linear quadratic methods*. Courier Corporation, 2007. 1

[17] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020. 1

[18] 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. 1, 2, 8

[19] 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. 1

[20] P. B. Tamer Başar, *H_∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer, 2008. 1, 3, 12

[21] K. Glover, "Minimum entropy and risk-sensitive control: the continuous time case," in *Proceedings of the 28th IEEE Conference on Decision and Control,*, pp. 388–391 vol.1, 1989. 1

[22] 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

[23] 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. 1

[24] 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. 1

[25] 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

[26] T. Basar, "Minimax disturbance attenuation in ltv plants in discrete time," in *1990 American Control Conference*, pp. 3112–3113, IEEE, 1990. 1

[27] N. Gârleanu and L. H. Pedersen, "Dynamic portfolio choice with frictions," *Journal of Economic Theory*, vol. 165, pp. 487–516, 2016. 1

[28] J. M. Steele, *Stochastic calculus and financial applications*, vol. 1. Springer, 2001. 1

[29] B. Øksendal and B. Øksendal, *Stochastic differential equations*. Springer, 2003. 1

[30] L. Molu, "LevelSetPy: A GPU-Accelerated Package for Hyperbolic Hamilton-Jacobi Partial Differential Equations' Solubility," 2023. 1

[31] 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. 2

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

[33] *Functional Analysis in Normed Spaces*. New York: MacMillan, 1964. 3, 5, 6

[34] D. Z. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, pp. 114–115, 1968. 3, 4

[35] L. Molu, "Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ policy synthesis.," in *The International Federation of Automatic Control, 22nd World Congress*, July 2023. https://scriptedonachip.com/downloads/Papers/ifac.pdf. 4, 10, 11

[36] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994. 4

[37] P. Gahinet and P. Apkarian, "A linear matrix inequality approach to $\mathcal{H}_\infty$ control," *International Journal of Robust and Nonlinear Control*, vol. 4, no. 4, pp. 421–448, 1994. 4

[38] E. D. Sontag, *Input to State Stability: Basic Concepts and Results*, pp. 163–220. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 7

[39] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, "Control of some linear stochastic systems in a hilbert space with fractional brownian motions," in *2011 16th International Conference on Methods & Models in Automation & Robotics*, pp. 107–110, IEEE, 2011. 7

[40] T. E. Duncan and B. Pasik-Duncan, "Stochastic linear-quadratic control for systems with a fractional brownian motion," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 6163–6168, IEEE, 2010. 7

[41] Y. Jiang and Z.-P. Jiang, "Computational Adaptive Optimal Control for Continuolus-Time Linear Systems With COmpletely Unknown Dynamics," vol. 48, pp. 2699–2704, 2023. 8

[42] 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

[43] 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, 12

[44] 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. 10

[45] 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. 10

[46] 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. 10

[47] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001. 10

[48] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice hall Upper Saddle River, NJ, 1996. 11, 12

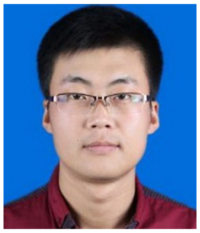
[49] R. A. Horn and C. R. Johnson, *Matrix Analysis, second edition*. Cambridge University Press, 2013. 11

[50] L. Koralov and Y. G. Sinai, *Theory of Probability and Random Processes*. Springer Berlin, Heidelberg, 2nd ed. ed., 2007. 12

[51] A. Arapostathis, V. S. Borkar, and M. K. Ghosh, *Ergodic Control of Diffusion Processes*. Cambridge University Press, 2011. 12

[52] 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. 12

[53] K. Zhou and J. C. Doyle, *Essentials of robust control*, vol. 104. Prentice hall Upper Saddle River, NJ, 1998. 12



Leilei Cui received a B.Sc. in Automation from Northwestern Polytechnical University, Xian, China, in 2016, and an M.Sc. degree in Control Science and Engineering at Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently a Ph.D. candidate in the Control and Networks Lab, Tandon School of Engineering, New York University. He was a Microsoft Research Intern during Summer of 2022. His research interests include robot control, reinforcement learning, adaptive dynamic programming, and optimal control.

Lekan Molu is a Researcher at Microsoft Research, NYC. He got his Master of Science in Engineering in Control Systems from the University of Sheffield, South Yorkshire, United Kingdom in 2013. He was subsequently a PhD student at the University of Texas at Dallas from 2014 through 2019 where he was concurrently a visiting student researcher in the Medical Physics and Engineering Division of the University of Texas Southwestern Medical Center. He was recently a postdoctoral scholar at the University of Pennsylvania in Philadelphia, PA, USA. His research interests span machine learning, optimal, adaptive, and nonlinear control with emphasis on their applications to robotics and complex systems. He was Associate Editor for IEEE ICRA, was a member of the New York Academy of Sciences, and American Association of Physicists in Medicine. Currently, he is a member of IEEE Robotics and Automation, and Control Systems Societies.